Predicting yield traits of individual field-grown *Brassica napus* plants from rosette-stage leaf gene expression

Sam De Meyer[a,b,§], Daniel Felipe Cruz[a,b,§], Tom De Swaef[c], Peter Lootens[c], Jolien De Block[a,b], Kevin Bird[a,b,d], Heike Sprenger[a,b], Michael Van de Voorde[a,b], Stijn Hawinkel[a,b], Tom Van Hautegem[a,b], Dirk Inzé [a,b], Hilde Nelissen[a,b], Isabel Roldán-Ruiz[a,c] and Steven Maere [a,b,*]

[a] Department of Plant Biotechnology and Bioinformatics, Ghent University, Technologiepark 71, 9052 Ghent, Belgium

[b] VIB Center for Plant Systems Biology, Technologiepark 71, 9052 Ghent, Belgium

[c] Plant Sciences Unit, Flanders Research Institute for Agriculture, Fisheries and Food (ILVO), Caritasstraat 39, 9090 Melle, Belgium

[d] present address : Department of Plant Sciences, University of California-Davis, Davis, California 95616, USA

[§] shared first authors

* corresponding author

Email addresses : Sam De Meyer : samey@psb.vib-ugent.be ; Daniel Cruz : dacru@psb.vib-ugent.be ; Tom De Swaef : tom.deswaef@ilvo.vlaanderen.be ; Peter Lootens : peter.lootens@ilvo.vlaanderen.be ; Jolien De Block : joblo@psb.vib-ugent.be ; Kevin Bird : kabird@ucdavis.edu ; Heike Sprenger : hsprenger@mailbox.org ; Michael Van de Voorde : mivoo@psb.vib-ugent.be ; Stijn Hawinkel : sthaw@psb.vib-ugent.be ; Tom Van Hautegem : tohau@psb.vib-ugent.be ; Dirk Inzé : diinz@psb.vib-ugent.be ; Hilde Nelissen hinel@psb.vib-ugent.be ; Isabel Roldán-Ruiz : isabel.roldan-ruiz@ilvo.vlaanderen.be ; Steven Maere : stmae@psb.vib-ugent.be

ABSTRACT

Background : In the plant sciences, results of laboratory studies often do not translate well to the field because lab growth conditions are very different from field conditions. To help close this lab-field gap, we developed a new strategy for studying the wiring of plant traits directly in the field, based on molecular profiling and phenotyping of individual plants of the same genetic background grown in the same field. This single-plant omics strategy leverages uncontrolled micro-environmental variation across the field and stochastic variation among the individual plants as information sources, rather than controlled perturbations. Here, we use single-plant omics on winter-type *Brassica napus* (rapeseed) plants to investigate to what extent rosette-stage gene expression profiles can be linked to the early and late phenotypes of individual field-grown plants.

Results : We find that rosette leaf gene expression in autumn has substantial predictive power for both autumnal leaf phenotypes and final yield in spring. Many of the top predictor genes are linked to developmental processes known to occur in autumn in winter-type *B. napus* accessions, such as the juvenile-to-adult and vegetative-to-reproductive phase transitions, indicating that the yield potential of winter-type *B. napus* is influenced by autumnal development.

Conclusions : Our results show that profiling individual plants under uncontrolled field conditions is a valid strategy for identifying genes and processes influencing crop yield in the field.

KEYWORDS

single-plant omics ; *Brassica napus* ; field trial ; machine learning ; transcriptome-based phenotype prediction.

BACKGROUND

One of the major aims of molecular biology research is to unravel how genes influence phenotypes. This usually involves applying perturbations to the genome or growth environment of an organism of interest and analyzing the ensuing molecular and phenotypic responses. Generally, well-chosen perturbations are applied in a controlled experimental setting, and technical and biological replicates are performed to allow for sufficiently powerful analyses despite noise in the data. Noise in this context may refer to measurement errors, noise due to uncontrolled factors in the experimental setup, or noise due to cellular or environmental stochasticity. The main purpose of avoiding or averaging out such noise is to facilitate causal interpretation of the link between a perturbation and its molecular and phenotypic effects.

It is becoming increasingly clear however that data noise caused by uncontrolled experimental factors and even purely stochastic effects can be a valuable source of information, instead of merely a nuisance. Several studies have shown that stochastic gene expression noise in single cells can be used to infer regulatory influences (1-3). Gene networks are also increasingly inferred from single-cell gene expression datasets in which differences among cells are not purely due to stochastic effects in an otherwise homogeneous cell

71 population, but reflect additional uncontrolled heterogeneity among cells, e.g. in the
72 temporal progression of a cell differentiation program (4-10).
73
74 In addition, several studies have investigated the information content of 'noise' datasets in
75 which the profiled entities are multicellular individuals rather than single cells. Bhosale, Jewell
76 et al. (11) found that gene expression noise among individual *Arabidopsis thaliana* plants
77 grown under the same conditions harbored as much information on the function of genes as
78 gene expression responses to controlled perturbations. The dataset analyzed by Bhosale,
79 Jewell et al. (11) was however not ideal because it contained data on plants of three different
80 accessions grown in six different labs (12), causing lab and accession effects that had to be
81 removed computationally to uncover the individual plant noise of interest. Recently, a study
82 on a cleaner *A. thaliana* seedling dataset confirmed that gene expression noise among
83 individuals of the same background grown under the same lab conditions contains useful
84 information on gene functions and regulatory relationships (13).
85
86 A common denominator in the aforementioned studies is that even under controlled
87 conditions, each cell or individual is subject to a set of stochastic or other perturbations that
88 escape experimental control, and that these uncontrolled perturbations, like any
89 perturbations, generate responses that contain valuable information on the wiring of gene
90 networks. Although most studies to date focused on the information content of noise under
91 controlled lab conditions, there is no reason to believe that 'noise' datasets generated under
92 less controlled conditions would be less valuable. On the contrary, studies performed in a
93 more natural setting in which organisms are subject to uncontrolled perturbations may yield

information that cannot easily be recovered from experiments under controlled lab conditions.

In the plant sciences for instance, controlled growth conditions in a laboratory are generally very different from field conditions, in which plants are subject to a plethora of highly variable environmental cues that often have non-additive phenotypic effects (14-21). Results obtained in the laboratory therefore often translate poorly to the field (14, 22-26). Narrowing this lab-field gap is essential to speed up the development of new crop varieties and optimized agricultural practices, both of which are direly needed in view of the current challenges posed by world population growth, land use and climate change. One option to narrow the lab-field gap is to make lab conditions more field-like (22), but the decreased experimental control this implies challenges traditional experimental design practices to e.g. ensure reproducibility. Another option is to perform interventional experiments in the field rather than the lab, but controlled interventions in a field may be costly and the level of control that can be achieved is often limited (22). Observational 'uncontrolled perturbation' studies on the other hand can easily be set up in the field. Observational data come with their own array of challenges however, e.g. that many of the perturbations influencing the study subjects may remain unobserved and hence unknown, and that it is generally much more challenging to establish cause-effect relationships from observational data (27). Nevertheless, even purely correlational data generated in the field may help narrow the lab-field gap in plant sciences.

To assess the information content of plant molecular responses to uncontrolled perturbations occurring in a field environment, we previously generated transcriptome and metabolome data on the primary ear leaf of 60 individual *Zea mays* (maize) plants of the same genetic

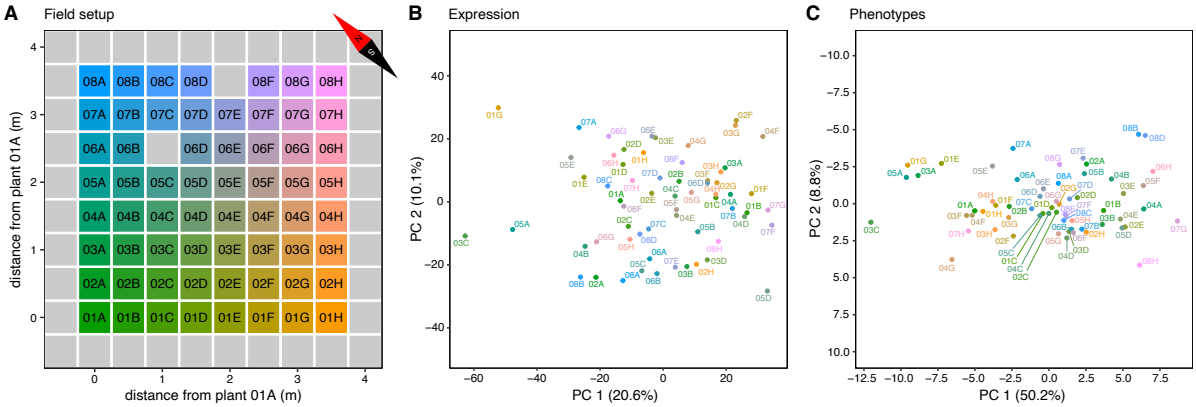
118 background grown in the same field (28). Similar to what was found for lab-grown *A. thaliana*
119 plants (11), the transcriptomes of the individual field-grown maize plants were found to
120 contain as much information on maize gene function as transcriptomes profiling the response
121 of maize plants to controlled perturbations in the lab. In addition, we found that the single-
122 plant transcriptome and metabolome data had better-than-random predictive power for
123 several phenotypes that were measured for the individual plants, and the prediction models
124 also produced sensible candidate genes for these phenotypes (28). However, only a few
125 phenotypes were measured in this study, and they were either closely associated with the
126 material sampled for molecular profiling, not fully developed or both.
127
128 Here, we investigate in more detail how much phenotype information can be extracted from
129 the transcriptomes of single plants subject to uncontrolled perturbations under field
130 conditions. To this end, we profiled the rosette-stage leaf transcriptome of individual field-
131 grown plants of the winter-type accession Darmor of *Brassica napus* (rapeseed), an important
132 oilseed crop (29). Additionally, a wide range of phenotypes was measured for all plants
133 throughout the growing season. We find that the autumnal leaf transcriptomes of the
134 individual plants do not only have predictive power for autumnal leaf phenotypes but also for
135 yield phenotypes measured more than 5 months later, such as silique count and total seed
136 weight. Furthermore, we find that many of the genes that feature prominently in our
137 predictive models are related to developmental processes known to occur in autumn in
138 winter-type *Brassica napus*, in particular the juvenile-to-adult and vegetative-to-reproductive
139 phase transitions. Our results suggest that micro-environmental variations across the field
140 cause a gradual buildup of developmental differences among plants that ultimately result in
141 yield differences at the end of the growing season.

RESULTS

Field trial, expression profiling and phenotyping

One hundred *Brassica napus* plants of the winter-type accession Darmor were grown in a field in a 10x10 equispaced grid pattern with 0.5 m distance between rows and columns (**Fig. 1**). On November 28, 2016, the eighth rosette leaf (leaf 8) of 62 non-border plants was harvested, and the harvested leaves were expression-profiled individually (see Methods and **Additional File 1: Table S1**). After leaf sampling, the plants were allowed to overwinter and set seed in spring. 62 phenotypes were recorded for all plants, ranging from rosette areas and individual leaf measurements in autumn to root and shoot measurements at harvest the following spring (**Additional File 1: Table S1**). Likely because of the low planting density, many of the plants developed one or more secondary inflorescence stems at ground level, which is not usually observed for *B. napus* grown under lab conditions or in the field at agronomically relevant planting densities. These secondary stems (further referred to as side stems) were harvested separately from the primary inflorescence stem with its cauline secondary inflorescences (further referred to as stem 1). Several yield phenotypes were measured for both stem 1 and the entire shoot (i.e. stem 1 plus side stems), including dry weight, seed weight, seed count and silique count. Cauline secondary inflorescence stems on stem 1 and tertiary inflorescence stems on the side stems (both further referred to as branches) were also counted, and branch counts are reported for both stem 1 and the entire shoot (the latter being the sum of branch counts on stem 1 and the side stems). Shoot growth phenotypes such as the time of maximum shoot growth, the maximum shoot growth rate and the end of shoot growth were derived from plant height time series data through curve fitting (see Methods). Several phenotypes were defined as ratios of other phenotypes, e.g. the ratio of

165 total seed weight to shoot dry weight and the ratio of the total number of seeds to the total
166 number of siliques per plant.



167

168 **Fig. 1** Field trial layout and PCA plots for gene expression and phenotypes. **A** Plants were sown
169 on a 10x10 equispaced grid with 0.5 m between rows and columns. Plant identifiers combine
170 a number indicating the row (01-08) and a letter indicating the column (A-H) in which the
171 plant was sown. Only plants with leaf 8 gene expression and phenotype profiles are labeled,
172 border plants and grid positions at which no plants emerged are indicated by grey squares. **B**
173 Plot of the first two principal components of the leaf 8 gene expression dataset, after
174 normalization and RNA-seq batch correction (see Methods). **C** Plot of the first two principal
175 components in the phenotype dataset. Individual plants in **B** and **C** are colored according to
176 the color gradient in **A**, with similar coloring of plants indicating spatial proximity in the field.

177

178 ## Exploratory data analysis

179 Principal component analysis (PCA) suggests that there are no subpopulations of plants with
180 distinct expression or phenotype profiles (**Fig. 1**). A few relative outliers are visible however,
181 e.g. plant 04G in the phenotype PCA plot (**Fig. 1C**), a very small plant that yielded barely any
182 seeds. Single-nucleotide polymorphism (SNP) analysis of the RNA-seq data (see Methods) did

not uncover signs of substantial genetic substructure in the plant population (**Additional File 2: Fig. S1**).

Mapping of the field coordinates on the expression and phenotype PCA plots on the other hand suggests that there is spatial structure in the data (**Fig. 1**). However, the levels of only 169 out of 76,808 transcripts and 1 out of 41 phenotypes (root system width) were found to be significantly spatially autocorrelated across the field (Moran's I, Benjamini-Hochberg (BH) adjusted permutation test *p*-values (*q*-values) < 0.05, **Additional File 3: Table S2, Additional File 2: Fig. S2**). In a previous study on a similar number of field-grown maize plants (28), 14.17% of transcripts were found to be significantly spatially autocorrelated at $q \leq 0.01$, which is considerably more than the 0.22% recovered here at $q \leq 0.05$. This may be due to differences in the way Moran's I values and their significance were calculated in Cruz, De Meyer et al. (28) versus the present study (see Methods). To assess whether some functional classes of genes have on average a stronger or weaker spatial autocorrelation signal than other classes, regardless of the statistical significance of the Moran's I values, two-sided Mann-Whitney U (MWU) tests (30) were performed on the transcript list ranked in order of decreasing Moran's I value. Genes involved in e.g. photosynthesis, translation, the response to abiotic stimuli, response to cytokinin, regulation of circadian rhythm, photoperiodism and the vegetative to reproductive phase transition were found to have a significantly higher Moran's I on average than other genes (MWU $q \leq 0.05$, **Additional File 3: Table S2**). This suggests that there is spatial patterning in the data, but that its discovery may be hampered by a lack of statistical power due to the small size of the field trial.

206 Most continuous phenotypes and high-count discrete phenotypes (e.g. seed and silique
207 counts) are at least approximately normally distributed (Anderson-Darling and Shapiro-Wilk
208 normality tests, $p > 0.01$, **Additional File 4: Table S3**), with the exception of five ratio
209 phenotypes (seeds per silique, seeds per silique stem 1, seed weight/dry weight stem 1, total
210 seed weight/shoot dry weight and branches per stem), leaf count (74 DAS) and two shoot
211 growth phenotypes (time of max shoot growth and end of shoot growth). Many of these
212 phenotypes exhibit relative outliers that may influence normality testing results (**Additional**
213 **File 2: Fig. S3**). When removing outliers (see Methods), four additional phenotypes (seeds per
214 silique, seeds per silique stem 1, stem 1 seed weight/ stem 1 dry weight, total seed
215 weight/shoot dry weight) were found to be approximately normally distributed (Anderson-
216 Darling and Shapiro-Wilk normality test, $p > 0.01$, **Additional File 4: Table S3**).
217
218 Some phenotypes were found to be more variable across the field than others. Dry weight,
219 seed and silique phenotypes at harvest are the most variable, with coefficients of variation
220 (CVs) between 43.7% and 51.9% (**Additional File 4: Table S3**). Taproot length also has a high
221 CV (42.8%). Plant height (278 DAS) and shoot growth parameters exhibit the lowest CV values
222 (< 7%). Most ratio phenotypes also have relatively low CV values (≤ 20.3%), with the exception
223 of siliques per branch (35.3%), siliques per branch stem 1 (35.0%) and branches per stem
224 (33.6%). When removing outliers, the CV of some of these ratio phenotypes is further
225 reduced, notably for seed weight stem 1/dry weight stem 1 (20.3% → 9.5%), total seed
226 weight/shoot dry weight (18.1% → 8.9%), seeds per silique (19.5% → 14.7%) and seeds per
227 silique stem 1 (19.4% → 14.6%). Leaf and branch phenotypes generally exhibit intermediate
228 CVs. Whereas leaf 8 fresh weight (81 DAS), leaf 8 area (81 DAS), total branch count and rosette
229 area (42 DAS) have a CV ≥ 30%, other leaf 8 and leaf 6 phenotypes and branch count stem 1

exhibit a CV in the range 19.1%-23.2%, and leaf 8 chlorophyll content (81 DAS) has a CV of only 12.5%.

Gene expression also exhibits substantial variability across the field. Ignoring genes expressed in less than 10 samples, the median gene has an expression CV of 34.2% (**Additional File 4: Table S3**). To investigate whether some classes of genes vary more in expression than others across the field, we ranked *B. napus* genes based on a normalized version of their expression CV (*normCV*, see Methods and **Additional File 4: Table S3**). MWU tests (30) were performed to assess whether any Gene Ontology (GO) biological processes are represented more at the top or bottom of the *normCV*-ranked gene list than expected by chance (**Additional File 4: Table S3**). As observed in earlier studies on populations of lab-grown *Arabidopsis thaliana* Col-0 plants (31) and field-grown *Zea mays* B104 plants (28), genes involved in photosynthesis and responses to biotic and abiotic stimuli were found to be on average more variably expressed than other genes, while genes involved in housekeeping functions related to protein, RNA and DNA metabolism were found to be on average more stably expressed across the field (**Additional File 4: Table S3**). To what extent high gene expression variability is due to either variability in the levels of external stimuli experienced by the individual plants or due to a higher intrinsic noisiness of a gene's expression levels (on the scale of entire leaves) is unclear. Some categories of genes with more variable expression across the field, such as genes involved in photosynthesis or response to abiotic stimuli, also exhibit higher Moran's I values on average, suggesting that their variability may be linked to external stimuli that are spatially patterned. On the other hand, most genes with highly variable expression do not exhibit strong spatial patterns (**Additional File 2: Fig. S4**), which indicates that their expression

253 variability may be caused by intrinsic stochastic factors, or alternatively by extrinsic factors
254 that are not spatially autocorrelated at the field sampling resolution employed.
255

256 ## Linking phenotypes to the leaf 8 expression profiles of single genes

257 To assess how much information leaf 8 gene expression profiles contain on the phenotypes
258 of individual plants, we used linear mixed-effect (lme) models to associate plant phenotypes
259 with the autumnal leaf 8 expression profile of single genes, taking into account spatial
260 autocorrelation effects (see Methods). Between 11,986 and 14,032 gene expression profiles,
261 out of 76,808, were found to be significantly associated ($q \leq 0.05$) with leaf 8 phenotypes such
262 as leaf 8 length, width, area and fresh weight (**Table 1 , Additional File 5: Table S4**). That leaf
263 8 phenotypes yield more associated genes than other phenotypes is not surprising, given that
264 leaf 8 was used for gene expression profiling. Next to leaf 8 phenotypes, also other leaf and
265 rosette phenotypes feature more associated genes than non-leaf phenotypes, except for leaf
266 6 length (74 DAS). The gene sets associated with leaf phenotypes are generally significantly
267 enriched (hypergeometric test, $q \leq 0.05$) in genes involved in e.g. response to biotic and
268 abiotic stimuli (salt), photosynthesis, circumnutation, cell wall biogenesis, amino acid
269 metabolism and response to sulfate and nitrogen starvation (**Additional File 6: Table S5**).
270 Additionally, leaf phenotype-related gene lists show significant enrichment, notably among
271 transcription factors, in genes involved in dorsal/ventral, adaxial/abaxial and radial pattern
272 formation, phloem, xylem and procambium histogenesis, and meristem development
273 (**Additional File 6: Table S5**).
274

275 Interestingly, appreciable numbers of gene-phenotype associations were found as well for
276 several phenotypes that are only distantly related in space and time to the leaf 8 material

277 profiled for RNA-seq. In particular seed, silique and shoot dry weight phenotypes yielded high
278 numbers of associated genes, ranging from 1,859 genes for total shoot dry weight to 1,248
279 genes for the silique count on stem 1 at harvest (**Table 1, Additional File 5: Table S4**). Many
280 of the gene sets associated with these phenotypes are enriched in genes involved in nitrate
281 assimilation, superoxide metabolism, circumnutation, circadian rhythm, response to biotic
282 and abiotic stimuli (cold, salt, water deprivation), response to nutrient levels (nitrogen,
283 sulphate and phosphate starvation), and, in particular among transcription factors, phosphate
284 ion homeostasis, histone modification, regulation of the vegetative to reproductive phase
285 transition and floral organ morphogenesis (**Additional File 6: Table S5**). 1,110 genes were
286 found associated with the branch count on stem 1, with GO enrichments similar to those
287 obtained for dry weight, silique and seed phenotypes (**Additional File 6: Table S5**). In
288 contrast, the total branch count phenotype only yields a set of 89 associated genes (**Table 1,**
289 **Additional File 5: Table S4**), which is however also strongly enriched in e.g. superoxide
290 metabolism and salt stress genes. The fact that the total branch count is composed of cauline
291 secondary inflorescence stems on stem 1 and tertiary inflorescence stems on the side stems
292 may render this phenotype less relevant.
293
294 Phenotypes with very low CV such as leaf 8 chlorophyll content (81 DAS), the maximum shoot
295 growth rate and end of shoot growth yielded no significantly associated genes, suggesting
296 that the biological variation of these phenotypes is limited and that the observed variation
297 may be dominated by technical noise (**Table 1, Additional File 5: Table S4**). The phenotype
298 with the lowest CV on the other hand, the time of maximal shoot growth (CV=0.6%), features
299 3,498 significant leaf 8 gene expression correlates. The associated gene set is strongly
300 enriched in genes involved in e.g. cell wall biogenesis and response to biotic stimuli

(**Additional File 6: Table S5**). For plant height (278 DAS) (CV = 6.8%), 99 associated genes are found with minor GO enrichments.

	All genes			Transcription factors		
Phenotype	**# Significant**	**Most significant**	***q***	**# Significant**	**Most significant**	***q***
leaf 8 length (76 DAS)	14032	BnaC07g39340D	4.35E-14	453	BnaA05g33840D	4.44E-09
leaf 8 width (76 DAS)	13695	BnaC07g39340D	1.62E-15	429	BnaA05g33840D	6.98E-09
leaf 8 width (81 DAS)	13605	BnaA02g18860D	1.37E-12	420	BnaA05g33840D	1.36E-08
leaf 8 fresh weight (81 DAS)	12989	BnaC07g39340D	5.79E-12	412	BnaAnng02740D	4.17E-08
leaf 8 area (81 DAS)	12569	BnaC07g39340D	1.29E-13	408	BnaAnng02740D	1.64E-08
leaf 8 length (81 DAS)	11986	BnaA01g14450D	6.39E-14	383	BnaA05g27750D	5.25E-08
leaf count (74 DAS)	10442	BnaC04g49060D	1.86E-07	313	BnaA05g33840D	2.16E-06
rosette area (42 DAS)	7196	BnaC09g39140D	7.14E-06	212	BnaA06g39930D	1.06E-04
leaf 6 width (74 DAS)	5386	BnaA09g04980D	2.05E-05	184	BnaA05g27750D	2.76E-04
time of max shoot growth	3498	BnaC06g28860D	7.29E-07	89	BnaC04g03950D	5.42E-06
total shoot dry weight	1859	BnaA05g29010D	1.37E-06	76	BnaCnng05590D	2.36E-04
total shoot dry weight (w/o seeds)	1802	BnaA05g29010D	8.88E-07	68	BnaAnng37500D	4.89E-04
dry weight stem 1	1612	BnaA05g29010D	2.79E-06	72	BnaC01g37260D	5.87E-05
dry weight stem 1 (w/o seeds)	1611	BnaA05g29010D	6.82E-06	75	BnaA08g12050D	3.13E-04
total seed weight	1598	BnaA06g35450D	1.02E-05	66	BnaCnng05590D	6.05E-05
total seed count	1545	BnaA06g35450D	9.10E-06	63	BnaCnng05590D	1.12E-04
seed weight stem 1	1539	BnaA05g29010D	1.65E-05	66	BnaC01g37260D	1.65E-05
total silique count	1520	BnaA06g35450D	3.92E-05	64	BnaCnng05590D	7.12E-04
seed count stem 1	1449	BnaC01g37260D	2.58E-05	67	BnaC01g37260D	2.58E-05
leaf 6 length (74 DAS)	1345	BnaA09g04980D	2.37E-04	34	BnaA02g18720D	8.39E-03
silique count stem 1	1248	BnaA05g29010D	6.32E-05	56	BnaC01g37260D	6.32E-05
branch count stem 1	1110	BnaA06g35450D	2.52E-04	39	BnaC01g37260D	1.95E-03
siliques per branch stem 1	593	BnaA01g17100D	2.24E-03	29	BnaC01g37260D	2.48E-03
total seed weight/shoot dry weight	458	BnaC09g50070D	2.24E-05	13	BnaC04g55440D	4.81E-04
seed weight stem 1/dry weight stem 1	280	BnaA03g50380D	8.90E-04	6	BnaC03g62970D	3.39E-03
branch count stem 1/length stem 1	240	BnaAnng11300D	8.22E-03	5	BnaC01g37260D	4.04E-02
seeds per silique stem 1	233	BnaC05g45470D	8.73E-04	9	BnaC04g03950D	4.42E-03
seeds per silique	112	BnaA08g07570D	9.01E-04	3	BnaC04g03950D	1.43E-02
plant height (278 DAS)	99	BnaA06g34140D	7.20E-04	5	BnaC01g37260D	2.48E-02
total branch count	89	BnaA06g35450D	2.39E-03	1	BnaCnng05590D	5.82E-03
root system width	4	BnaA01g06800D	7.53E-03	0	-	-
siliques per branch	3	BnaAnng39720D	2.22E-02	0	-	-
branches per stem	0	-	-	0	-	-
taproot length	0	-	-	0	-	-
leaf 8 chlorophyll content (81 DAS)	0	-	-	0	-	-
max shoot growth rate	0	-	-	0	-	-
end of shoot growth	0	-	-	0	-	-
rosette lesions (74 DAS)	0	-	-	0	-	-
leaf 6 lesions (74 DAS)	0	-	-	0	-	-
leaf 8 lesions (76 DAS)	0	-	-	0	-	-
stem count	0	-	-	0	-	-

Table 1 Numbers of significant gene expression-phenotype associations. For any given phenotype, results are reported on the complete gene set (*n*=76,808; 'All genes' columns) and on the set of transcription factors (*n*=2,521; 'Transcription factors' columns). In both cases, the results shown include (from left to right) the total number of significant gene expression-phenotype associations ($q \leq 0.05$), the most significant gene and its *q*-value.

No genes were found associated ($q \leq 0.05$) with taproot length and only four with root system width, suggesting that autumnal leaf 8 gene expression may not contain a lot of information on root phenotypes. On the other hand, given the difficulty of recovering intact root systems from the soil, it is not excluded that root measurement errors may have influenced these results.

Ratio phenotypes exhibit between 0 and 593 associated genes. In particular the branches per stem and siliques per branch ratios do poorly (0 and 3 associated genes, respectively). Both involve the total branch count, which is itself only associated with 89 genes. Ratios involving the branch count on stem 1 on the other hand yield between 240 and 593 associated genes. One potential reason for ratio phenotypes having at most a few hundred gene associations is that ratios suffer from increased error levels due to the propagation of measurement errors from both the numerator and denominator. This may be particularly problematic for ratios of highly correlated variables such as the seeds per silique and seed weight/dry weight phenotypes (both for stem 1 and the entire shoot), which exhibit a low CV and likely have even lower true biological variation. No genes were found associated at $q \leq 0.05$ with qualitative or low-count discrete phenotypes such as rosette lesions (74 DAS), leaf 6 lesions (74 DAS), leaf 8 lesions (76 DAS) and stem count (i.e. stem 1 plus the number of side stems).

Leaf and final yield phenotypes of individual field-grown *B. napus* plants can be predicted to a considerable extent from their autumnal leaf 8 transcriptome

We built random forest (RF) and elastic net (enet) models to predict the phenotypes of individual plants from their autumnal leaf 8 transcriptome, using either all genes or only transcription factors (TFs) as potential features and using three different feature selection techniques (see Methods). For each combination of phenotype, model type (RF or enet), potential feature set (all genes or TFs) and feature selection technique, 9 repeat models were learned, each time using 10-fold cross-validation with different splits (see Methods), resulting per combination in a total of 90 test sets and 9 test set predictions per plant. The best model for a given phenotype and potential feature set was taken to be the one with the highest median test R^2 value across all 90 test sets for continuous and high-count phenotypes (see Methods), or the highest median test accuracy for qualitative or low-count discrete phenotypes (**Table 2, Additional File 7: Table S6**).

Not surprisingly, leaf 8 phenotypes, which are most closely related in space and time to the material sampled for transcriptome profiling, are the most predictable. Except for the leaf 8 chlorophyll content at sampling time (81 DAS), which features very poor prediction performance, the median test R^2 scores for leaf 8 phenotypes range from 0.48 to 0.70 when using all genes as potential features. Other leaf-related phenotypes such as leaf 6 width (74 DAS, median test R^2 = 0.38), rosette area (42 DAS, median test R^2 = 0.23) and leaf 6 length (74 DAS, median test R^2 = 0.07) are comparatively less predictable.

Surprisingly, many of the final seed, silique and shoot dry weight phenotypes are more predictable from the autumnal leaf 8 transcriptome than leaf 6 and rosette phenotypes, with seed weight on stem 1 rivaling the leaf 8 phenotypes in terms of median test R^2 value (**Table 2, Additional File 7: Table S6**, **Fig. 2, Additional File 2: Fig. S5**). All seed weight, seed and

356 silique count and shoot dry weight phenotypes have median test R^2 values in the range 0.35
357 - 0.51 for the 'all genes' models, which is in all cases higher than the 95th percentile of test R^2
358 values obtained from single train-test splits on 90 datasets in which the phenotype values
359 were permuted (**Additional File 7: Table S6**). In other words, the model for the real data train-
360 test split with median test R^2 outperforms 95% of the models for comparable train-test splits
361 on randomized data. Note that this serves only as an indication of model performance on real
362 versus randomized data, not as a formal test assessing whether the median test R^2 score on
363 real data is significantly higher than expected at random. The latter would require the 9 times
364 repeated 10-fold cross-validation setup used on the real data to be used on each of the
365 permuted datasets as well (instead of the single train-test split per permutation used here),
366 which is computationally prohibitive.
367
368 Interestingly, yield phenotypes measured for stem 1 are generally slightly more predictable
369 than the corresponding phenotypes measured for the entire shoot, with median test R^2 score
370 differences between stem 1 and total shoot phenotypes in the range 0.02-0.09 for the 'all
371 genes' models and 0.05-0.13 for the 'transcription factors' models. This suggests that gene
372 expression levels in leaf 8 of the rosette may be more predictive for phenotypes of stem 1
373 (i.e. the primary inflorescence stem and its cauline secondary inflorescences) than for
374 phenotypes measured on the whole shoot (i.e. including the secondary inflorescence stems
375 branching at ground level).
376
377 Root phenotypes, branching phenotypes, final plant height (278 DAS) and shoot growth
378 phenotypes are generally poorly predictable (**Table 2**). Plant height and shoot growth
379 phenotypes are likely poorly predictable because they show little variation across the field

(**Additional File 4: Table S3**), increasing the risk that measurement error outweighs biological variation. Also taproot length and root system width may suffer from measurement errors. The total branch count and branch count stem 1 phenotypes on the other hand have a high CV and likely limited measurement error, suggesting that leaf 8 gene expression profiles may contain less information on these branching phenotypes than on leaf, seed, silique and dry weight phenotypes.

Most phenotypes calculated as ratios of other phenotypes are very poorly predictable, even if the constituent phenotypes have high prediction performance values. For instance, the median test R^2 value for seeds per silique (total seed count divided by total silique count) is negative (-0.14), whereas both total seed count and total silique count have median test R^2 values ≥ 0.38. In many cases however, the numerator and denominator phenotypes of a ratio are highly correlated, leading to a derived phenotype with a small range that may be dominated by noise propagated from measurement errors in the constituent phenotypes rather than biological variability. The number of siliques per branch on stem 1 and the entire shoot are notable exceptions with high CV values and reasonable prediction performance (**Table 2**). The latter ratio phenotypes are highly correlated with the number of siliques on stem 1 (PCC = 0.92) and the entire shoot (PCC = 0.72), respectively, indicating that the number of siliques per branch is an important determinant of silique count, in addition to the number of branches (PCC between total branch count and total silique count = 0.87, PCC between branch count stem 1 and silique count stem 1 = 0.82).

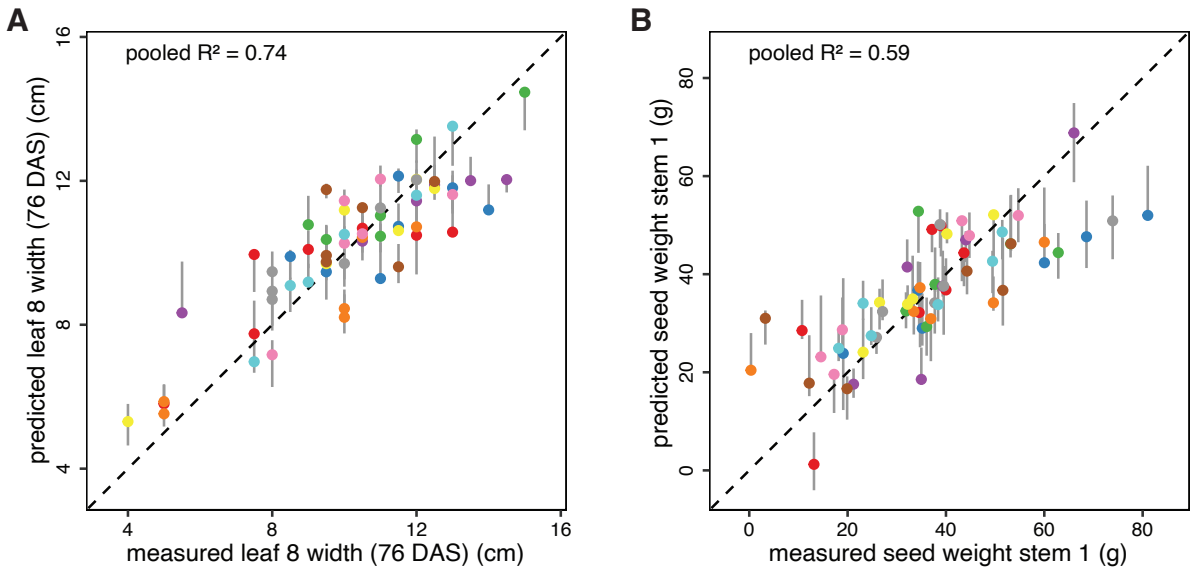
	All genes				Transcription factors				Single gene			
Continuous and high-count phenotypes	**Feature sel.**	**Model type**	**Median test R2**	**Median pooled PCC**	**Feature sel.**	**Model type**	**Median test R2**	**Median pooled PCC**	**Top gene**	**Median test R2**	**Median pooled PCC**	**CV**
leaf 8 width (76 DAS)	median	enet	0.70 *	0.87	median	enet	0.64 *	0.84	BnaC04g39580D	0.67	0.83	2.32E-01
leaf 8 width (81 DAS)	median	enet	0.65 *	0.86	median	enet	0.65 *	0.84	BnaA02g18860D	0.62	0.83	2.32E-01
leaf 8 area (81 DAS)	median	enet	0.63 *	0.83	median	enet	0.58 *	0.81	BnaCnng33420D	0.60	0.81	3.70E-01
leaf 8 fresh weight (81 DAS)	median	enet	0.59 *	0.81	median	enet	0.53 *	0.78	BnaCnng33420D	0.58	0.79	3.88E-01
seed weight stem 1	spearman	enet	0.51 *	0.77	median	enet	0.53 *	0.78	BnaA05g29010D	0.42	0.67	4.51E-01
leaf 8 length (76 DAS)	spearman	rf	0.51 *	0.80	median	enet	0.53 *	0.79	BnaC07g39340D	0.58	0.80	2.21E-01
leaf 8 length (81 DAS)	spearman	rf	0.48 *	0.78	spearman	enet	0.51 *	0.78	BnaC01g17020D	0.52	0.81	2.11E-01
seed count stem 1	spearman	enet	0.47 *	0.72	median	enet	0.43 *	0.73	BnaA06g20870D	0.38	0.61	4.71E-01
silique count stem 1	median	enet	0.46 *	0.74	median	enet	0.45 *	0.72	BnaA05g29010D	0.37	0.66	4.37E-01
total seed count	spearman	enet	0.45 *	0.73	median	enet	0.38 *	0.71	BnaC03g60710D	0.39	0.61	4.78E-01
dry weight stem 1	spearman	enet	0.44 *	0.73	median	enet	0.39 *	0.70	BnaA05g29010D	0.40	0.70	4.83E-01
dry weight stem 1 (w/o seeds)	hsic-5000	enet	0.42 *	0.69	spearman	enet	0.35 *	0.64	BnaA05g29010D	0.39	0.69	5.19E-01
total seed weight	spearman	enet	0.42 *	0.74	median	enet	0.40 *	0.70	BnaA06g35450D	0.41	0.69	4.69E-01
total shoot dry weight	spearman	enet	0.41 *	0.71	median	enet	0.31 *	0.68	BnaA09g48720D	0.41	0.67	4.89E-01
leaf 6 width (74 DAS)	median	enet	0.38 *	0.68	hsic-5000	rf	0.07	0.51	BnaC03g15540D	0.35	0.61	1.91E-01
total silique count	median	enet	0.38 *	0.68	median	enet	0.36 *	0.68	BnaC04g21390D	0.40	0.63	4.56E-01
siliques per branch stem 1	hsic-5000	enet	0.37 *	0.67	hsic-5000	rf	0.14	0.51	BnaC04g21390D	0.25	0.60	3.50E-01
total shoot dry weight (w/o seeds)	spearman	enet	0.35 *	0.66	spearman	enet	0.29	0.62	BnaA06g05150D	0.40	0.69	5.13E-01
leaf count (74 DAS)	median	rf	0.24 *	0.66	median	rf	0.40 *	0.72	BnaA01g34700D	0.38	0.70	1.12E-01
rosette area (42 DAS)	median	enet	0.23 *	0.59	median	enet	0.36 *	0.68	BnaC05g30740D	0.33	0.64	3.00E-01
branch count stem 1	spearman	rf	0.19	0.56	median	enet	0.11 *	0.52	BnaA10g29560D	0.38	0.56	2.00E-01
siliques per branch	spearman	enet	0.16	0.49	spearman	enet	-0.01	0.36	BnaA08g09860D	0.12	0.53	3.53E-01
plant height (278 DAS)	median	enet	0.12 *	0.47	spearman	enet	0.16	0.51	BnaC07g25920D	0.34	0.63	6.81E-02
total branch count	median	rf	0.10	0.40	median	enet	0.17 *	0.57	BnaA09g48720D	0.26	0.59	3.42E-01
branch count stem 1/length stem 1	median	rf	0.07	0.39	median	rf	-0.06	0.23	BnaC05g15590D	0.22	0.55	1.58E-01
leaf 6 length (74 DAS)	median	enet	0.07 *	0.45	median	rf	0.06	0.43	BnaA09g04980D	0.31	0.64	1.96E-01
max shoot growth rate	median	rf	0.03	0.41	hsic-5000	rf	-0.02	0.34	BnaA10g21770D	0.15	0.50	6.75E-02
root system width	median	rf	0.01	0.36	median	rf	0.17 *	0.56	BnaC07g01150D	0.17	0.56	2.16E-01
time of max shoot growth	median	rf	-0.02	0.38	hsic-5000	rf	0.20	0.57	BnaA05g08250D	0.15	0.56	6.45E-03
taproot length	spearman	rf	-0.02	0.26	median	enet	-0.09	0.18	BnaA04g17830D	0.16	0.50	4.28E-01
branches per stem	median	rf	-0.09	0.26	hsic-5000	rf	-0.12	0.21	BnaC03g42190D	0.13	0.51	3.36E-01
leaf 8 chlorophyll content (81 DAS)	median	enet	-0.14	-0.35	median	enet	-0.12	-0.40	BnaA03g40350D	0.09	0.46	1.25E-01
seeds per silique	median	rf	-0.14	-0.18	median	enet	-0.17	-0.31	BnaC03g38990D	-0.09	0.22	1.95E-01
seeds per silique stem 1	median	enet	-0.15	-0.31	median	enet	-0.15	-0.08	BnaC01g44890D	-0.05	0.40	1.94E-01
seed weight stem 1/dry weight stem 1	median	enet	-0.18	-0.09	median	enet	-0.16	-0.39	BnaC09g50070D	-0.13	0.58	2.03E-01
total seed weight/shoot dry weight	median	enet	-0.19	-0.08	median	enet	-0.17	-0.39	BnaA02g15500D	-0.12	0.33	1.81E-01
end of shoot growth	hsic-5000	rf	-0.22	0.02	median	rf	-0.29	-0.17	BnaA05g09440D	0.04	0.45	1.02E-02
Qualitative and low-count phenotypes	**Feature sel.**	**Model type**	**Median test accuracy**		**Feature sel.**	**Model type**	**Median test accuracy**					
leaf 8 lesions (76 DAS)	hsic-5000	enet	0.67		hsic-5000	rf	0.67					
rosette lesions (74 DAS)	median	rf	0.50		spearman	enet	0.46					
leaf 6 lesions (74 DAS)	median	enet	0.33		spearman	enet	0.50					
stem count	median	enet	0.33		hsic-5000	enet	0.50					

401

402 **Table 2** Best-performing multi-gene and single-gene models for each phenotype. Results are

403 shown for models including all genes as potential features ('All genes' columns), models

including only TFs as potential features ('Transcription factors' columns) and models using a single gene as feature ('Single gene' columns). For the best all-gene and TF models for continuous or high-count discrete phenotypes, columns from left to right indicate the feature selection technique used (median = selection of features with median *rlog* gene expression > 0, spearman = Spearman correlation, hsic-5000 = HSIC lasso, see Methods), the model type (enet = elastic net, rf = random forest), the median test R^2 and the median pooled Pearson correlation coefficient (PCC, see Methods). Stars in the median test R^2 column indicate that the median test R^2 score on real data is higher than the 95th percentile of test R^2 scores on permuted data (**Additional File 7: Table S6**). For qualitative and low-count phenotypes, the median test accuracy was used as a performance metric instead of the median test R^2 (see Methods). Single-gene model columns include the best-performing gene and the corresponding median test R^2 and median pooled PCC. All single-gene models are cross-validated lme models with spatial error structure. The CV column contains the coefficients of variation for the phenotypes.



421

422 **Fig. 2** Predictions versus observations for the best-scoring leaf and yield phenotypes. **A**
423 Predicted versus measured values for leaf 8 width (76 DAS), using the 'all genes' model with
424 the best median test R^2 score (enet + median feature selection, **Table 2**). **B** Predicted versus
425 measured values for seed weight stem 1, using the 'all genes' model with the best median
426 test R^2 score (enet + Spearman feature selection, **Table 2**). Vertical grey lines range from the
427 minimum to the maximum predicted value for a given plant across all model repeats, and
428 colored dots represent predictions for the repeat with the median pooled R^2 score (i.e. the R^2
429 score of the pooled test set predictions in the repeat concerned). Different marker colors
430 indicate the 10 different test sets in this repeat. Perfect predictions are located on the dashed
431 diagonal line in each panel.

432

433 To compare multi-gene models to single-gene models in terms of phenotype prediction
434 performance, we used the same repeated cross-validation setup as used for the multi-gene
435 models to calculate median test R^2 scores and median pooled PCC values for single-gene
436 models (lme models with spatial structure, see previous section). Cross-validation scores were

437 calculated for each of the 100 genes most significantly associated with a given phenotype
438 (lowest *q*-value for gene coefficient in lme model, **Additional File 7: Table S6**).
439
440 For leaf 8 phenotypes, the best multi-gene models generally have only slightly better median
441 test R^2 scores than the best single-gene models (**Table 2**). In other words, multi-gene models
442 offer little benefit over single-gene models for quantitative prediction of leaf 8 phenotypes.
443 For leaf 8 length at 76 DAS and 81 DAS, the best single-gene models even outperform the
444 multi-gene models. Single-gene models also outperform multi-gene models for several other
445 phenotypes, sometimes with a wide margin, e.g. for plant height (278 DAS), branch count on
446 stem 1, leaf count (74 DAS), leaf 6 length (74 DAS) and rosette area (42 DAS). This suggests
447 that the multi-gene models are vulnerable to overfitting. In particular phenotypes with low
448 single-gene model performance tend to exhibit a multi-gene model performance that is even
449 lower, suggesting that the extent of multi-gene model overfitting is inversely correlated with
450 the proportion of trait variance explained by single genes. An alternative explanation for the
451 observation that the best single-gene models sometimes outperform the corresponding
452 multi-gene models may be the 'winner's curse' effect, also known as selection bias (32),
453 whereby the apparently best-performing single-gene models may overestimate prediction
454 performance.
455
456 Most of the phenotypes with comparatively high single-gene model performance scores
457 however exhibit a modest increase of multi-gene model performance over the best single-
458 gene model. Like most of the leaf 8-associated traits, total seed weight, total seed count and
459 most of the shoot dry weight traits are modestly better predicted by multi-gene models than
460 by single-gene models. Many of the seed and silique traits related to stem 1 on the other hand

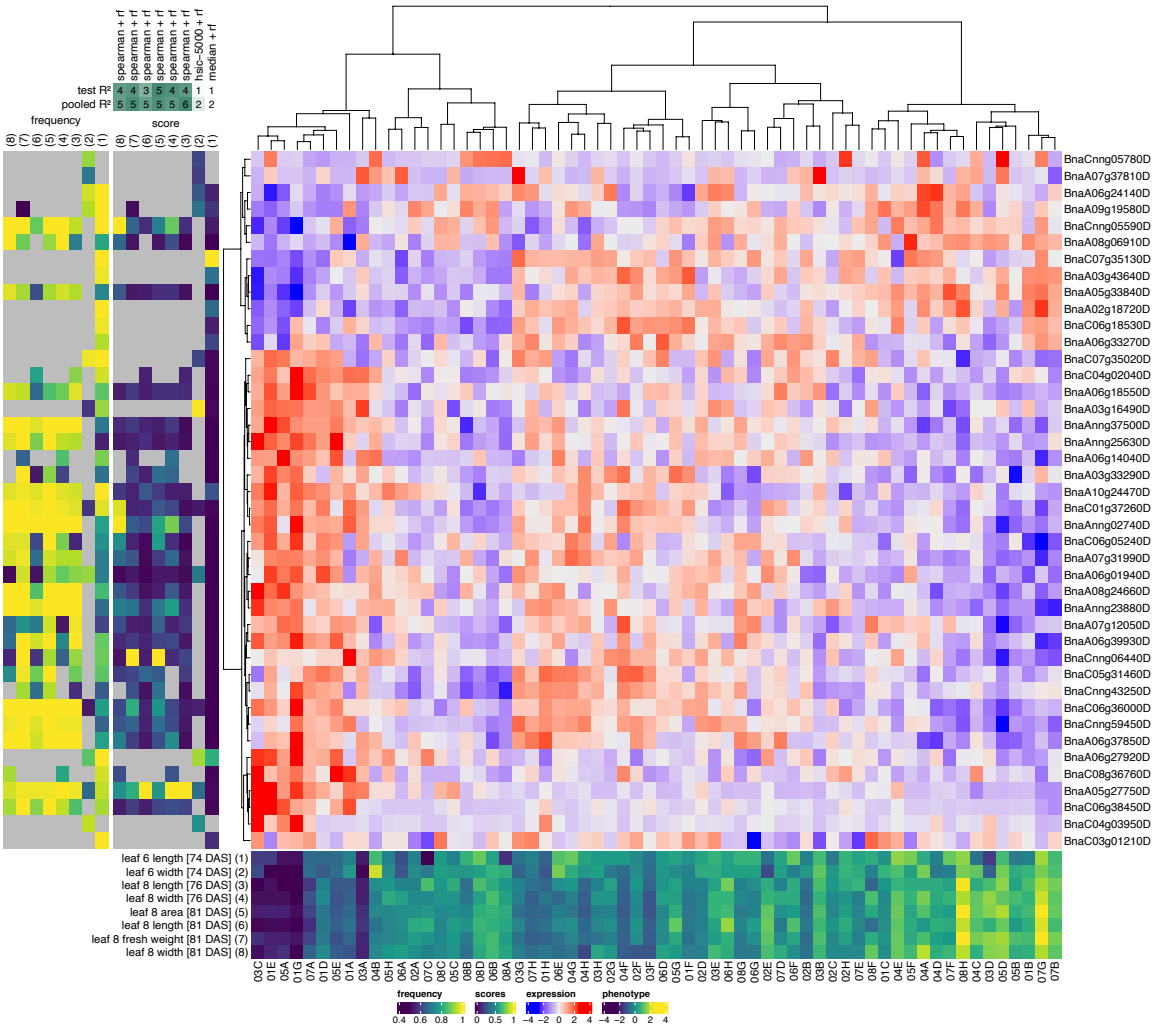
(seed weight, seed count and silique count on stem 1, the number of siliques per branch on stem 1) are substantially better predicted by multi-gene models than by single-gene models. This indicates that several distinct gene expression patterns are likely relevant for quantitative prediction of stem 1 seed and silique traits.

For most ratio phenotypes, both the multi-gene and single-gene models have very poor prediction performance, in particular when the numerator and denominator phenotypes that make up the ratio are very highly correlated. In these cases, the denominator is essentially already a good predictor of the numerator. To assess whether any gene expression profiles contain additional information on the numerator given knowledge of the denominator, we used alternative single-gene models with a log link (see Methods) to predict the numerators of the seeds per silique ratio on stem 1 and the branches per stem ratio (seed count stem 1 and total branch count, respectively) conditioned on their denominator (silique count stem 1 and stem count, respectively). These models are not suited for making predictions in practice, given the need to know the denominator, but they may indicate whether prediction of the ratio based on gene expression is at all feasible and if so, which genes may be important. If no genes are found to be predictive for the numerator (and hence the ratio) conditioned on the denominator, then attempts to predict the ratio phenotype unconditionally are likely to be unsuccessful. For both seeds per silique stem 1 and branches per stem, the fitted coefficients and residuals look reasonable for the best predictor genes (**Additional File 2: Fig. S6, Additional File 2: Fig. S7**). The corresponding models succeed in suppressing a few of the more extreme residuals of the base model (without gene expression effect), without improving predictions for most other plants. However, no gene coefficients were found to be significantly different from zero for any phenotype after BH correction (q<0.05), neither in

485 models assuming constant error variance nor in models with heteroscedastic and/or spatially
486 covarying error structures (see Methods). This indicates that the poor performance of the
487 original multi-gene and single-gene models for these phenotypes is to be expected.
488

489 ## Top predictors for leaf phenotypes

490 The best multi-gene prediction performance scores were obtained for leaf 8 phenotypes. To
491 assess whether the genes featuring most prominently in the multi-gene models for leaf
492 phenotypes make biological sense, we focused on the top-10 predictor lists of the TF-based
493 models for leaf 8 length and width (76 DAS and 81 DAS), fresh weight (81 DAS) and area (81
494 DAS), and leaf 6 length and width (74 DAS) (**Additional File 7: Table S6**). As these leaf
495 phenotypes are generally highly correlated (PCC between leaf 8 phenotype in the range
496 [0.78,0.97], PCC between leaf 8 and leaf 6 phenotypes in range [0.45, 0.60]), many of the most
497 important predictors (TFs) in the random forest and elastic net models are shared among
498 phenotypes. We therefore grouped the top-10 predictor lists for the different phenotypes in
499 two sets, one for the random forest (RF) models (**Fig. 3**) and one for the elastic net (enet)
500 models (**Additional File 2: Fig. S8**). The rationale for looking at the TF models instead of the
501 models built on all genes is that TFs are more likely than the average gene to have been
502 functionally characterized to some extent, and are more likely to be causally involved in
503 phenotype regulation (although it needs to be stressed that our analysis remains entirely
504 correlational). Given the relative lack of experimentally determined gene functions in *B.*
505 *napus*, most of the functional interpretation given below for *B. napus* genes is based on the
506 experimentally determined functions of likely orthologs in *A. thaliana* (see Methods).
507



508

509 **Fig. 3** Top predictor genes in random forest models of leaf phenotypes. A clustered heatmap
510 of the z-scored gene expression profiles of the top genes for predicting leaf phenotypes is
511 shown centrally (blue-red color scale, Ward.D2 hierarchical clustering). The leaf phenotypes
512 concerned and their z-scored profiles across plants are shown at the bottom (dark blue-yellow
513 heatmap with plant identifiers at the bottom). For each of these phenotypes, the top-10 most
514 important genes (highest median gini importance across all 90 cross-validation splits) of the
515 RF model with the highest median test R^2 score are included on the figure (gene identifiers
516 are shown at right). The mostly dark blue score panel to the left of the expression heatmap
517 shows the median gini importance scores of the selected genes in each of the selected
518 phenotype models, normalized to the maximum importance score per model to make the

519 color scales of the different models (columns) comparable. The mostly yellow frequency panel
520 to the left of the score panel shows the frequencies at which genes were selected as features
521 across all 90 cross-validation splits of a given model. Grey squares in the score and frequency
522 panels indicate that a given gene was not selected as a feature in a given model. The
523 phenotypes in the score and frequency panels are identified by numbers (1-8) on top of the
524 panels, corresponding to the numbers associated with the phenotypes in the bottom
525 phenotype panel. On top of the score panel, the feature selection techniques used in the best-
526 scoring RF models for each phenotype are shown (median = selection of features with median
527 *rlog* gene expression > 0, spearman = Spearman correlation, hsic-5000 = HSIC lasso, see
528 Methods), as well as the corresponding test and pooled R^2 scores rounded to the nearest 0.1
529 and then multiplied by ten (e.g. a test R^2 score of 0.38 would be denoted as 4).
530
531 Many of the top TF predictors for leaf phenotypes have *A. thaliana* orthologs with known
532 functions in leaf development. One TF with high importance scores in both the RF and enet
533 models is *BnaCnng05590D*, a putative ortholog of the homeodomain leucine zipper class I
534 (*HD-ZIP I*) gene *ARABIDOPSIS THALIANA HOMEOBOX 1* (*AtHB1/AT3G01470*). Both the RF and
535 enet top predictor sets additionally contain *BnaA05g33840D*, another putative ortholog of
536 *AtHB1*. Ectopic *AtHB1* overexpression in tobacco seedlings was previously shown to lead to
537 de-etiolated phenotypes in the dark, including true leaf development (33). Mutation of an
538 upstream open reading frame in the *AtHB1* 5' untranslated region that normally represses
539 *AtHB1* translation was shown to lead to smaller, more serrated leaves, smaller rosettes, a
540 delay of the vegetative-to-reproductive phase transition and siliques containing fewer seeds
541 in *A. thaliana* (34). Similarly, *AtHB1* overexpression in a silencing-deficient *rdr6-12* mutant
542 background resulted in plants with shorter and more serrated leaves (35).

The enet top predictor list for leaf phenotypes also contains another *HD-ZIP I* gene, *BnaC02g43700D*, which is putatively orthologous to *AtHB5* (*AT5G65310*) or *AtHB16* (*AT4G40060*). Similar to *AtHB1*, overexpression of *AtHB16* leads to smaller, more serrated leaves exhibiting reduced cell expansion, smaller rosettes and siliques containing fewer seeds (36). Additionally, *AtHB16* overexpression was reported to reduce the flowering time sensitivity to differences in photoperiod in *A. thaliana* (36).

Next to *HD-ZIP I* genes, the RF and enet top predictor lists contain several other *HD-ZIP* genes. *BnaA06g18550D* in the RF top predictor list is putatively orthologous to the *A. thaliana* gene *REVOLUTA* (*AtREV*/*AT5G60690*), which encodes a HD-ZIP III transcription factor known to regulate postembryonic meristem initiation (37) and several polarity-associated growth processes in *A. thaliana*, including abaxial-adaxial patterning in leaves (38). Loss-of-function *atrev-1* mutant plants were shown to exhibit overgrowth and deformation of rosette and cauline leaves after bolting (39). The RF top predictor list also contains two additional *HD-ZIP III* gene family members, *BnaC06g05240D* and *BnaA06g01940D*, that are putatively orthologous to *AtHB8* (*AT4G32880*) or *AtHB15* (*AT1G52150*). *AtHB8* and *AtHB15* are thought to have effects on postembryonic meristem initiation that are antagonistic to the effects of *AtREV* (40). On the other hand, gain-of-function mutations in *AtHB15*, like gain-of-function mutations in *AtREV*, have been shown to result in adaxialized leaves (41). Both *AtHB8* and *AtHB15* are thought to function prominently in vascular development, possibly antagonistically (41-44). Furthermore, the enet top predictor list includes *BnaC03g02700D*, a *HD-ZIP II* gene putatively orthologous to *AtHAT3* (*AT3G60390*), *AtHAT14* (*AT5G06710*),

566 *AtHB17* (*AT2G01430*) or *AtHB18* (*AT1G70920*). *AtHAT3* is known to be involved in leaf
567 abaxial/adaxial patterning (45), and to be regulated by AtREV (46).
568
569 Both the RF and enet top predictor lists prominently feature putative orthologs of *A. thaliana*
570 *WUSCHEL RELATED HOMEOBOX* (*AtWOX*) genes: *BnaA05g27750D* (RF and enet) and
571 *BnaC05g41930D* (enet). Both genes are putatively orthologous to *AtWOX5* (*AT3G11260*) or
572 *AtWOX7* (*AT5G05770*). Next to roles in root development, *AtWOX5* was reported to act
573 redundantly with *AtWOX1* and *AtWOX3* to control leaf shape by promoting lateral leaf growth
574 (47). AtWOX1, 2, 3 and 5 were also shown to regulate the expression of *REVOLUTA* (*AtREV*)
575 and other *HD-ZIP III* genes in the shoot apical meristem (48), and AtWOX1 and AtWOX3 are
576 thought to regulate *HD-ZIP III* genes in lateral leaf regions, thereby contributing to the
577 maintenance of adaxial/abaxial patterning at the margin of growing leaves (49).
578
579 Not all transcription factors in the RF and enet models are equally important for all leaf
580 phenotypes. *BnaCnng06440D* (*AtMYB60*/*AT1G08810*) for instance has higher RF and (to a
581 lesser extent) enet importance scores for leaf 8 area (81 DAS) and leaf 8 fresh weight (81 DAS)
582 than for other leaf phenotypes. Its likely *A. thaliana* ortholog *AtMYB60* is involved in
583 regulating stomatal opening, and its expression is downregulated under drought (50).
584 *atmyb60-1* null mutant plants exhibit a constitutive reduction of stomatal opening and
585 decreased transpirational water loss under drought (50). A second TF in the RF models with
586 higher importance for leaf 8 area (81 DAS) and leaf 8 fresh weight (81 DAS) than for other leaf
587 phenotypes is *BnaC06g36000D* (*AtHB33*/*AtZHD5*/*AT1G75240*). Its likely ortholog *AtHB33*
588 codes for a zinc-finger homeodomain TF downregulated in response to abscisic acid (ABA),
589 which e.g. induces stomatal closure (51). Constitutive *AtHB33* overexpression in *A. thaliana*

590 resulted in accelerated growth, larger leaves and larger epidermal cells (52). A third TF in the
591 RF models with mildly higher importance for leaf 8 area (81 DAS) and leaf 8 fresh weight (81
592 DAS) is *BnaCnn59450D*. Overexpression of its putative orthologs *AtSHN2* (*AT5G11190*) and
593 *AtSHN3* (*AT5G25390*) in *A. thaliana* resulted in folded and twisted leaves, shiny green leaf
594 surfaces with increased levels and altered composition of cuticular wax, increased cuticular
595 permeability, larger pavement cells, reduced trichome number and stomatal density, and
596 increased drought tolerance (53).
597
598 Both the RF and enet top-10 lists feature several orthologs of *A. thaliana NUCLEAR FACTOR*
599 *Y, SUBUNIT A* (*AtNF-YA*) genes (putative *A. thaliana* orthologs in parentheses):
600 *BnaAnng02740D* (*AtNF-YA2/10*, *AT3G05690/AT5G06510*, RF), *BnaA10g24470D* (*AtNF-*
601 *YA2/10*, *AT3G05690/AT5G06510*, RF and enet), *BnaC06g33980D* (*AtNF-YA3/8*,
602 *AT1G72830/AT1G17590*, enet) and *BnaC01g37260D* (*AtNF-YA5/6*, *AT1G54160/AT3G14020*,
603 RF). All four genes are negatively correlated with leaf phenotypes in the field expression
604 dataset. NF-Y transcription factor complexes are heterotrimers, consisting of A, B and C
605 subunits, that function in various developmental programs and abiotic stress responses in
606 plants (54). *AtNF-YA2* and *AtNF-YA10* were previously found to regulate leaf size in *A.*
607 *thaliana*, with their overexpression promoting cell expansion (55). *AtNF-YA5* was found to
608 promote drought resistance, with *atnf-ya5* knockout plants and *AtNF-YA5*-overexpressing
609 plants displaying increased and reduced leaf water loss, respectively, relative to wild-type
610 plants (56). AtNF-YA8 was recently found to negatively regulate the juvenile-to-adult
611 (vegetative) phase change by activating the transcription of *AtMIR156* genes (57).
612 Overexpression of *AtNF-YA8* resulted in a delay of the juvenile-to-adult transition and thereby
613 reduced leaf sizes and altered leaf shapes (57).

Interestingly, several of the top-TFs recovered in the multi-gene models for leaf phenotypes are linked to the regulation of flowering. Plant NF-Y complexes for instance are known to also function in the regulation of flowering time (54). Overexpression of the aforementioned *AtNF-YA8* gene was found to delay flowering time (57), and similar observations were made for other *AtNF-YA* genes such as *AtNF-YA1* (*AtHAP2A*, *AT5G12840*), *AtNF-YA2, AtNF-YA3*, *AtNF-YA4* (*At2g34720*), *AtNF-YA7* (*At1g30500*) and *AtNF-YA10* (58, 59). It has been suggested that the photoperiodic flowering regulator CONSTANS (AtCO) may compete with NF-YA subunits in the NF-Y complex to form an alternative complex activating *FLOWERING LOCUS T* (*FT*) expression in *A. thaliana*, thereby promoting flowering (58). Additionally, AtNF-YA2 has been suggested to function as a negative regulator of flowering in an alternative, stress-mediated flowering pathway (60). On the other hand, AtNF-YA2 was recently suggested to positively regulate flowering by directly influencing *AtFT* expression (61).

The *A. thaliana* orthologs of several of the aforementioned *HD-ZIP* genes (*BnaCnng05590D*/*AtHB1*, *BnaA05g33840D*/*AtHB1*, *BnaC02g43700D*/*AtHB16*, *BnaC06g05240D*/*AtHB15*, *BnaA06g01940D*/*AtHB15*) have also been linked to regulation of the juvenile-to-adult and/or vegetative-to-reproductive phase changes (34, 36, 41). Furthermore, both the enet and RF predictor lists contain *BnaA06g39930D*, a putative ortholog of *EARLY FLOWERING MYB PROTEIN* (*AtEFM*/*AT2G03500*) in *A. thaliana*. AtEFM is known to directly repress the expression of *FLOWERING LOCUS T* (*AtFT*, *AT1G65480*) in the leaf vasculature, and is thought to mediate the effects of temperature and light cues on the timing of the floral transition (62). The RF predictor list additionally contains *BnaC05g31460D*, a putative ortholog of *AtJMJD5* (*AtJMJ30*, *AT3G20810*), the protein product of which interacts

638 with AtEFM to repress AtFT (62). The RF and enet top predictor lists also contain
639 *BnaA07g12050D*, a putative ortholog of the floral homeotic gene *APETALA2*
640 (*AtAP2*/*AT4G36920*) or the related *euAPETALA2* gene *AtTOE3* (*AT5G67180*). Both AtAP2 and
641 AtTOE3 are known to repress *AGAMOUS* (*AtAG*) expression during floral patterning (63).
642
643 In summary, 15/42 and 11/35 transcription factors in the RF and enet lists of top leaf
644 phenotype predictors, respectively, have putative *A. thaliana* orthologs linked to leaf
645 development and patterning, the juvenile-to-adult phase change, the floral transition or
646 drought response.
647

648 Top predictors for seed, silique and shoot dry weight phenotypes

649 Next to leaf 8 phenotypes, also the seed, silique and shoot dry weight phenotypes (further
650 referred to as 'yield' phenotypes) of the individual plants at harvest (late spring) could be
651 predicted to a considerable extent from autumnal leaf 8 transcriptome data (see above).
652 Similar to the leaf phenotypes, the yield phenotypes are highly correlated (PCC range [0.84-
653 0.99]) and hence have a lot of high-scoring RF and enet predictors in common (**Additional File**
654 **7: Table S6**, **Fig. 4**, **Additional File 2: Fig. S9**). Furthermore, these phenotypes are also
655 significantly correlated with leaf phenotypes (PCC range [0.47, 0.74]), leading to a substantial
656 overlap between the top-10 predictor lists of yield and leaf phenotypes.
657

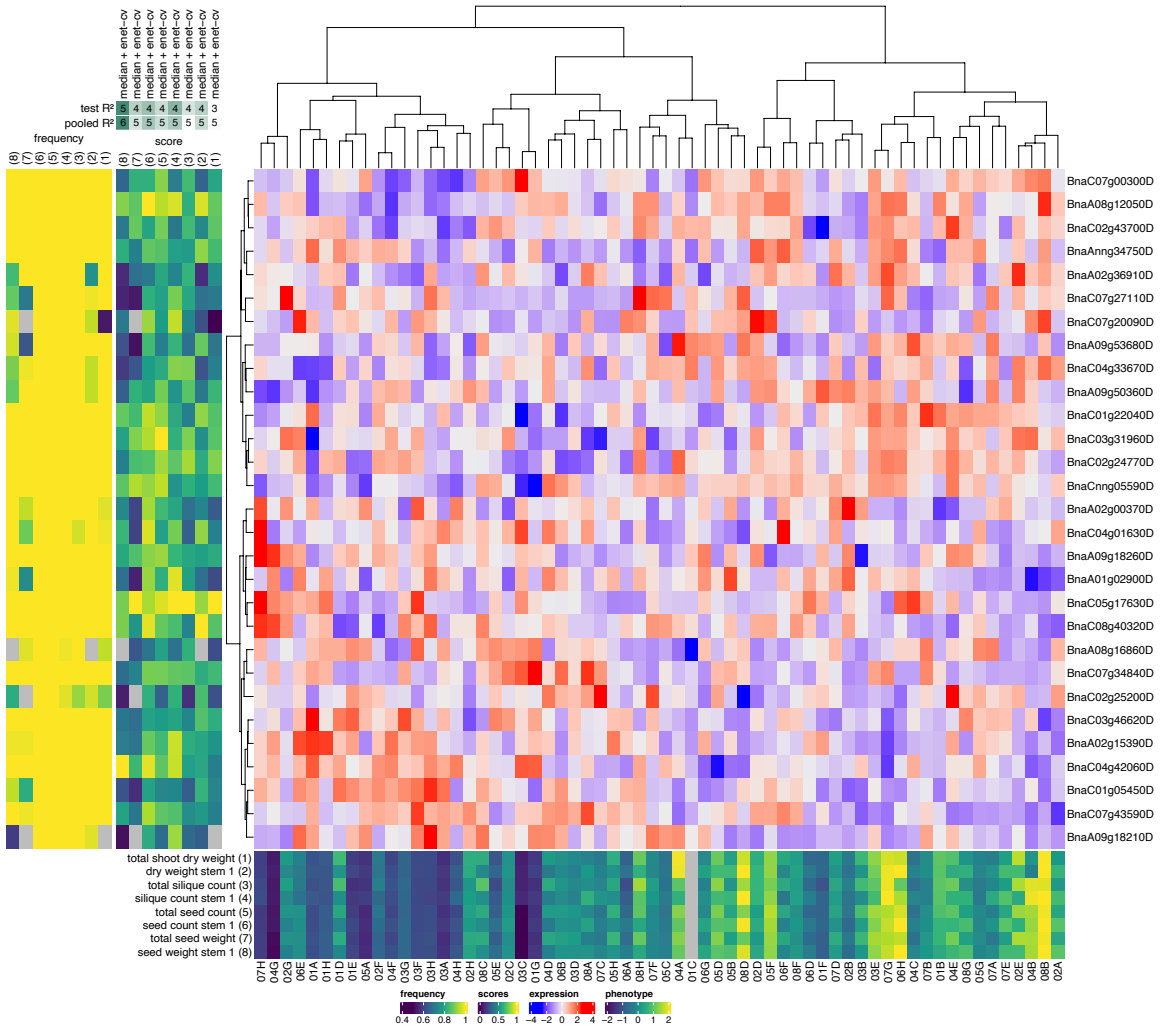


Fig. 4 Top predictor genes in elastic net models of yield phenotypes. A clustered heatmap of the z-scored gene expression profiles of the top genes for predicting yield phenotypes is shown centrally (blue-red color scale, Ward.D2 hierarchical clustering). The yield phenotypes concerned and their z-scored profiles across plants are shown at the bottom (dark blue-yellow heatmap with plant identifiers at the bottom). For each of these phenotypes, the top-10 most important genes (highest median elastic net coefficients across all 90 cross-validation splits) of the enet model with the highest median test R^2 score are included on the figure (gene identifiers are shown at right). The mostly green-blue score panel to the left of the expression heatmap shows the median elastic net coefficients of the selected genes in each of the selected phenotype models, normalized to the maximum coefficient per model to make the

color scales of the different models (columns) comparable. The mostly yellow frequency panel to the left of the score panel shows the frequencies at which genes were selected as features across all 90 cross-validation splits of a given model. Grey squares in the score and frequency panels indicate that a given gene was not selected as a feature in a given model. The phenotypes in the score and frequency panels are identified by numbers (1-8) on top of the panels, corresponding to the numbers associated with the phenotypes in the bottom phenotype panel. On top of the score panel, the feature selection techniques used in the best-scoring enet models for each phenotype are shown (median = selection of features with median *rlog* gene expression > 0, spearman = Spearman correlation, hsic-5000 = HSIC lasso, see Methods), as well as the corresponding test and pooled R^2 scores rounded to the nearest 0.1 and then multiplied by ten (e.g. a test R^2 score of 0.38 would be denoted as 4).

In particular, virtually all TF genes in the leaf top-10 predictor lists with links to the juvenile-to-adult or vegetative-to-reproductive phase changes and flowering also feature prominently in the RF or enet top-10 predictor lists for yield phenotypes, including *BnaCnng05590D* (*AtHB1*/*AT3G01470*), *BnaC02g43700D* (*AtHB5*/*AT5G65310* or *AtHB16*/*AT4G40060*), *BnaA07g12050D* (*AtAP2*/*AT4G36920*), *BnaAnng02740D* (*AtNF-YA2*/*10*, *AT3G05690*/*AT5G06510*), *BnaA10g24470D* (*AtNF-YA2*/*10*, *AT3G05690*/*AT5G06510*), *BnaC06g33980D* (*AtNF-YA3*/*8*, *AT1G72830*/*AT1G17590*) and *BnaC01g37260D* (*AtNF-YA5*/*6*, *AT1G54160*/*AT3G14020*). Furthermore, like the top predictor lists for leaf phenotypes, the enet top predictor list for yield phenotypes contains a putative ortholog of the *A. thaliana* gene *EARLY FLOWERING MYB PROTEIN* (*AtEFM*/*AT2G03500*), but a different one (*BnaAnng34750D*).

693 Furthermore, many of the top predictor TF genes for yield phenotypes that are absent from
694 the top-10 predictor lists for leaf phenotypes also have *A. thaliana* orthologs involved in
695 processes related to the floral transition and flowering. In the combined set of top-10 enet
696 predictors for shoot dry weight, seed and silique phenotypes (**Fig. 4**, n=29), five genes code
697 for AGAMOUS-LIKE MADS-box transcription factors (best candidate *A. thaliana* orthologs and
698 associated AGI codes are given between parentheses): *BnaC05g17630D*
699 (*AtAGL104*/*AT1G22130*), *BnaA02g15390D* (*AtAGL12*/*AT1G71692*), *BnaA02g00370D*
700 (*BnFLC.A2*, *AtFLC*/*AT5G10140*), *BnaA01g02900D* (*AtAGL16*/*AT3G57230*), and
701 *BnaA09g53680D* (*AtAGL30*/*AT2G03060*). *BnFLC.A2* is orthologous to *A. thaliana* FLOWERING
702 LOCUS C (*AtFLC*), a key repressor of the floral transition (64, 65). Two *AGAMOUS-LIKE* genes
703 feature in the combined set of top-10 RF predictors for yield phenotypes (**Additional File 2:**
704 **Fig. S9**, n=21): *BnaA02g15390D* (*AtAGL12*/*AT1G71692*) and *BnaA09g05500D*
705 (*AtAGL8*/*AtFUL*/*FRUITFULL*/*AT5G60910*). *AtFUL* is thought to regulate the floral transition
706 downstream of *AtFT* in the shoot apical meristem, partially redundantly with *AtSOC1*
707 (*AtAGL20*, *AT2G45660*) (66, 67). AtAP2 (APETALA2) and AtFUL are thought to form a bistable
708 switch mechanism through mutual repression that regulates early stages of the floral
709 transition at the shoot apical meristem (68). Negative regulation of *AtAP2* and several *AP2-*
710 *LIKE* genes by AtFUL was also found to contribute to meristem arrest at the end of flowering
711 (69). *ful* mutants were found to exhibit a delayed floral transition (68) and increased flower
712 production, but decreased seed set (69).
713
714 The enet top predictor list also features *BnaA09g18260D*, a *HD-ZIP II* gene putatively
715 orthologous to *JAIBA* (*AtJAB*/*AtHAT1*/*AT4G17460*) or *AtHAT2* (*AT5G47370*). *AtJAB* was shown
716 to be involved in male and female reproductive development and floral meristem

717 determination in *A. thaliana,* and *jab* loss-of-function mutants exhibit an increased number
718 of floral buds per inflorescence but a reduced number of seeds per silique (70), not unlike *ful*
719 mutants. The enet top predictor list also contains two *HD-ZIP IV* genes, *BnaA09g50360D*
720 (*AtHDG2*/ *AT1G05230*) and *BnaC03g31960D* (*AtANL2*/ *AT4G00730*). A combination of *hdg2*
721 and *pdf2* null mutant alleles in *A. thaliana* was shown earlier to produce flowers with sepaloid
722 petals and carpeloid stamens (71). *BnaC03g31960D* also features in the enet top predictor list
723 for leaf phenotypes, but less prominently (**Additional File 2: Fig. S8**).
724
725 Furthermore, the gene *BnaA08g12050D* is ranked highly in both the enet and RF top predictor
726 lists. The best candidate ortholog of this gene in *A. thaliana* is *AtMYB3R1* (*AT4G32730*), coding
727 for a regulator of cell proliferation that acts in a module with AtTSO1 to balance cell
728 proliferation with differentiation in developing roots and shoots (72). Loss-of-function
729 mutations in *AtMYB3R1* suppress all phenotypes of the *tso1-1* mutant, among others a lack
730 of floral organ differentiation (72). *BnaA08g12050D* also features as a predictor for leaf 6
731 length (74 DAS) and leaf 8 area (81 DAS) in **Additional File 2: Fig. S8**.
732
733 *BnaC07g27110D* and *BnaC01g22040D* in the enet predictor list are putative orthologs of
734 *AtGATA16* (*AT5G49300*) and *AtGATA17*(*AT3G16870*) or *AtGATA17L*(*AT4G16141*),
735 respectively. Evidence suggests these and other LLM-domain B-GATA transcription factors
736 are involved (at least partially redundantly) in the regulation of flowering time, silique length,
737 seed set and other developmental processes (73). The enet top predictor list also contains
738 *BnaC04g33670D* and *BnaA08g16860D*, *BZIP* genes putatively orthologous to the *A. thaliana*
739 genes *DRINK ME* (*AtDKM*/*AtBZIP30*/*AT2G21230*) and *DRINK ME-LIKE*
740 (*AtDKML*/*AtBZIP29*/*AT4G38900*), respectively. *AtDKM* and *AtDKML* are negative regulators of

reproductive development and growth (74). *AtDKM* overexpression in *A. thaliana* results in smaller plants with fewer floral buds and shorter siliques, while a *dkm* mutant show the opposite phenotype (74). *dkml* mutant plants also exhibited increased silique length but slightly fewer floral buds than wild-type plants (74). AtDKM was shown to interact *in planta* with several regulators of meristem development, including WUSCHEL (AtWU), HECATE1 (AtHEC1), the aforementioned JAIBA and NGATHA1 (AtNGA1) (74). Interestingly, the RF top predictor list contains a putative ortholog of *NGATHA1* (*AtNGA1*/*AT2G46870*) or *NGATHA2* (*AtNGA2*/*AT3G61970*), namely *BnaA09g39540D*. *AtNGA1* and *AtNGA2* are known to be involved in gynoecium development and were recently shown to also have a function in regulating shoot apical meristem development (75). Another likely regulator of meristem development, *BnaC07g43590D*, is found in the enet predictor list. *BnaC07g43590D* is most likely an ortholog of *ARABIDOPSIS RESPONSE REGULATOR 10* (*AtARR10*/*AT4G31920*) or *12* (*AtARR12*/*AT2G25180*), both known to directly activate the expression of *WUSCHEL* and to play a role in shoot apical meristem regeneration and maintenance (76).

In summary, 17/29 and 11/21 TF genes in the enet and RF lists of top yield predictors, respectively, have putative *A. thaliana* orthologs linked to the juvenile-to-adult phase change, the floral transition, flowering or regulation of meristem development.

Predicting final yield phenotypes from early growth phenotypes

As a baseline to assess the prediction performance of the molecular models, we trained models predicting plant phenotypes in spring (mostly phenotypes at harvest) from single or multiple autumnal leaf and rosette phenotypes. For these single- and multi-phenotype

764 models, the same modeling approaches were used as for the single- or multi-gene models,
765 respectively (see Methods).
766
767 Interestingly, many of the mature plant phenotypes can be predicted to a considerable extent
768 from phenotypes measured earlier in the growing season (**Table 3, Additional File 8: Table**
769 **S7**). In particular the models for phenotypes measured on the entire shoot (total seed, silique
770 and branch count, total seed weight, total shoot dry weight) perform surprisingly well. For
771 most of these phenotypes, the performance of the early-phenotype models is only slightly
772 less than that of the best single-gene or multi-gene model, and the early-phenotype models
773 for total seed weight and total branch count even outperform the molecular models (in the
774 case of total branch count even substantially so). Also for branching phenotypes related to
775 stem 1 (branch count stem 1, branch count stem 1/length stem 1), the best early-phenotype
776 models feature high prediction performance scores. For other stem 1 phenotypes however
777 (seed weight, seed count, silique count and siliques per branch on stem 1, stem 1 dry weight
778 with and without seeds), the molecular models clearly outperform the early-phenotype
779 models.
780
781 Most multi-phenotype models with appreciable prediction performance (median test R^2 >
782 0.10), both for whole-shoot and stem 1 phenotypes, feature leaf 8 area (81 DAS) as the top
783 predictor (**Table 3**). Leaf 8 area (81 DAS) is generally also the most predictive early phenotype
784 in the corresponding sets of single-phenotype models. The multi-phenotype models with the
785 best prediction performance scores, i.e. those for whole-shoot phenotypes and stem 1
786 branching phenotypes (but not the other stem 1 phenotypes), generally also feature rosette
787 area (42 DAS) as a predictor of some importance (**Additional File 8: Table S7**). For total branch

788 count, branch count stem 1 and branch count stem 1/length stem 1, rosette area (42 DAS) is
789 even the top predictor in either the RF or enet model, or both (**Additional File 8: Table S7**).
790 Rosette area (42 DAS) itself is only moderately predictable from the leaf 8 molecular data,
791 which may explain why multi-phenotype models are better at predicting these branching
792 phenotypes than multi-gene models.
793

	All early phenotypes			Single early phenotypes		
Mature plant phenotypes	**Model type**	**Median test R2**	**Median pooled PCC**	**Top phenotype**	**Median test R2**	**Median pooled PCC**
seed weight stem 1	enet	0.33	0.59	leaf 8 area (81 DAS)	0.32	0.63
seed count stem 1	rf	0.25	0.57	leaf 8 area (81 DAS)	0.26	0.61
silique count stem 1	enet	0.24	0.55	leaf 6 width (74 DAS)	0.22	0.53
total seed count	rf	0.40	0.66	leaf 8 area (81 DAS)	0.44	0.71
dry weight stem 1	enet	0.26	0.57	leaf 8 area (81 DAS)	0.35	0.62
dry weight stem 1 (w/o seeds)	enet	0.18	0.54	leaf 8 area (81 DAS)	0.30	0.59
total seed weight	enet	0.45	0.68	leaf 8 area (81 DAS)	0.46	0.72
total shoot dry weight	enet	0.38	0.67	leaf 8 area (81 DAS)	0.44	0.71
total silique count	rf	0.36	0.63	leaf 8 area (81 DAS)	0.41	0.70
siliques per branch stem 1	enet	0.14	0.47	leaf 8 area (81 DAS)	0.07	0.48
total shoot dry weight (w/o seeds)	enet	0.29	0.63	leaf 8 area (81 DAS)	0.37	0.68
branch count stem 1	enet	0.35	0.64	leaf 8 area (81 DAS)	0.34	0.65
siliques per branch	enet	-0.04	0.32	leaf 6 width (74 DAS)	-0.07	0.36
plant height	enet	0.23	0.58	leaf 8 length (81 DAS)	0.25	0.61
total branch count	rf	0.40	0.69	rosette area (42 DAS)	0.38	0.65
branch count stem 1/length stem 1	rf	0.33	0.63	leaf 8 area (81 DAS)	0.22	0.56
max shoot growth rate	enet	0.04	0.40	leaf 8 width (81 DAS)	0.04	0.41
root system width	rf	0.04	0.42	leaf 8 length (81 DAS)	0.05	0.39
time of max shoot growth	enet	-0.01	0.53	leaf 8 width (81 DAS)	0.08	0.53
taproot length	rf	-0.02	0.33	leaf 8 width (81 DAS)	0.01	0.33
branches per stem	enet	-0.14	-0.22	leaf 8 lesions (76 DAS)	-0.15	0.14
seeds per silique	enet	-0.17	-0.18	leaf 8 length (81 DAS)	-0.07	0.18
seeds per silique stem 1	enet	-0.15	-0.06	leaf 8 length (81 DAS)	-0.04	0.23
seed weight stem 1/dry weight stem 1	enet	-0.15	-0.40	leaf 8 lesions (76 DAS)	-0.19	-0.19
total seed weight/shoot dry weight	enet	-0.16	-0.39	leaf 8 lesions (76 DAS)	-0.18	-0.06
end of shoot growth	enet	-0.15	0.20	leaf 8 width (81 DAS)	-0.12	0.30

794

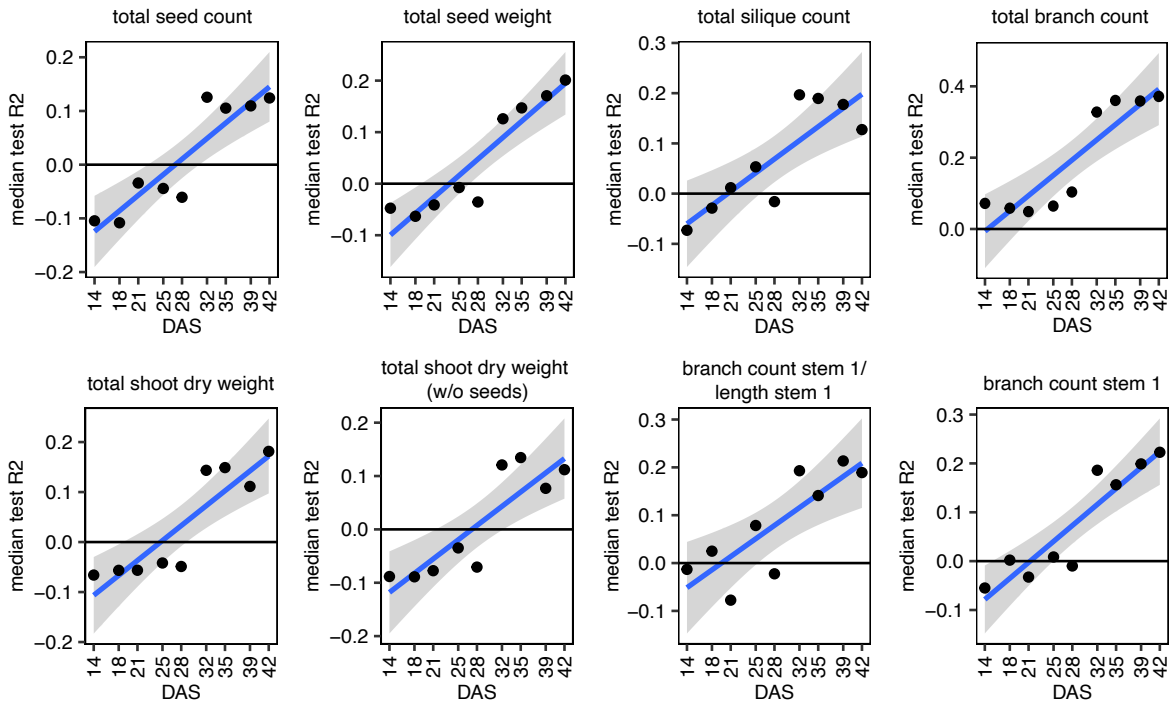
795 **Table 3 Best-performing multi-phenotype and single-phenotype models for mature plant**
796 **phenotypes.** Results are shown for models including all early phenotypes as potential
797 features (multi-phenotype models) and models using a single early phenotype as feature
798 (single-phenotype models). For the best multi-phenotype models, columns from left to right
799 indicate the model type used (enet = elastic net, rf = random forest), the median test R^2 and
800 the median pooled PCC (see Methods). Single-phenotype columns include the best-

performing early phenotype ('Top phenotype' column) and the corresponding median test R^2 and median pooled PCC. All single-phenotype models are cross-validated lme models with spatial error structure.

Our results indicate that the leaf 8 molecular data offer little benefit over early-phenotype measurements for quantitative prediction of mature phenotypes measured on the entire plant. On the other hand, the leaf 8 molecular data yields substantially better models than the early-phenotype data for most mature stem 1 phenotypes. Often, the multi-gene models for stem 1 phenotypes are also slightly better than the multi-gene models for the corresponding whole-plant phenotypes (see previous section). This suggests that the molecular makeup of the 8th rosette leaf at the time of sampling contained more information on the development of the primary flowering stem and its cauline secondary inflorescences than on the development of side stems at ground level. Early phenotypes on the other hand may contain more information on whole-plant yield phenotypes than on phenotypes specifically related to stem 1.

Given that even the earliest of the autumnal phenotypes considered thus far, the rosette area at 42 DAS, still has some predictive power for several yield phenotypes (median test $R^2 > 0.10$ for total branch count, seed count, seed weight and silique count, total dry weight with and without seeds, branch count stem 1 and branch count stem 1/length stem 1), we assessed whether earlier rosette areas (v2, see Methods) are also predictive for these phenotypes (**Fig. 5**, **Additional File 9: Table S8**). Median test R^2 scores were found to decrease when using earlier rosette areas as predictors, with rosette areas measured ≤ 28 DAS generally yielding low (< 0.10) and in many cases negative median test R^2 scores. When using the earliest

825 rosette area (14 DAS) as predictor, the median pooled R^2 and PCC scores are however still in
826 the ranges [0.05, 0.20] and [0.27, 0.45], respectively, indicating that even the earliest rosette
827 area measurements contain some information on final yield phenotypes.
828



829

830 **Fig. 5** Predictive power of early rosette areas for yield phenotypes. In each subplot, median
831 test R^2 values are plotted for lme models predicting the given phenotype from early rosette
832 areas v2 (14-42 DAS, x-axis). Only mature phenotypes that can be predicted from rosette area
833 (42 DAS) with a median test $R^2 > 0.1$ are shown. Blue lines are ordinary least-squares linear
834 regressions, with shaded areas indicating 95% confidence intervals on the trendline. Most
835 phenotypes exhibit a rather dichotomous median test R^2 profile with rosette areas v2 from
836 14 to 28 DAS yielding substantially lower median test R^2 values than rosette areas v2 from 32
837 to 42 DAS. Accordingly, linear model fits at 28 and 32 DAS are often poor.
838
839

DISCUSSION

In this study, we used machine learning models to predict the phenotypes of individual *B. napus* Darmor plants grown in the same field from rosette-stage leaf gene expression data. Our results show that many plant phenotypes can be predicted to a substantial extent from leaf 8 gene expression. Phenotypes closely related in time and space to the material sampled for RNA-seq, in particular leaf 8 phenotypes, generally feature good prediction performance, in accordance with results obtained earlier in a similar setup for maize (28). Interestingly however, also many of the phenotypes measured at the end of the growing season, ~5.5 months after leaf sampling for RNA-seq, feature high prediction performance. In particular seed yield, silique and dry weight traits exhibit prediction performance scores in the same range as the autumnal leaf and rosette phenotypes.

Azodi et al. (77) predicted several agronomically relevant mature plant traits (plant height, grain yield and flowering time) in a population of maize inbred lines from genetic marker data, whole-seedling transcriptome data and combinations thereof. Their transcriptome-based models exhibited PCC scores between predicted and measured values in the range [0.50, 0.61] for flowering time, [0.42, 0.51] for plant height and [0.47, 0.55] for 300 kernel weight (77). In the present study, the transcriptome-based models for mature plant traits in *B. napus* (ignoring ratio phenotypes) exhibit median pooled PCC scores in the range [0.57, 0.77] for seed phenotypes, [0.51, 0.74] for silique phenotypes, [0.56, 0.73] for shoot dry weight phenotypes, [0.40, 0.56] for branch count phenotypes, [0.40, 0.53] for plant height (278 DAS) and [0.07, 0.36] for root phenotypes (**Table 2**, **Additional File 7: Table S6**). Comparing the observed PCC ranges of both studies suggests that mature traits of individual plants of the same line grown in the same field are as predictable from early-stage transcriptome data as

average mature traits in a diversity panel. However, direct comparison of the PCC values across studies is complicated by differences in the phenotypes predicted, prediction and scoring methodology and factors affecting model training, scoring and overfitting potential such as the study population size (388 lines in the maize study versus 62 *B. napus* plants in the present study) and the number of potential model features (31,238 genes in the maize study versus 76,808 genes in the *B. napus* dataset). Also the species difference and the tissue and developmental time point sampled for RNA-seq (whole seedlings at the V1 stage for maize versus rosette leaf 8, 81 DAS, for *B. napus*) may impact how well a transcriptome can predict a given phenotype. The most comparable models are likely the whole-transcriptome-based random forest model for maize plant height, with a PCC of 0.42 (77), and the median-filter random forest model for the height of individual *B. napus* plants (without feature selection other than removing genes with *rlog* expression >0 in less than half of the samples, reducing the feature set to 55,166 genes), with a median pooled PCC of 0.43 (**Additional File 7: Table S6**).

Given that the single-plant transcriptome data can quantitatively predict many plant phenotypes better than expected by chance, the top predictor genes may shed light on biological processes that impact phenotypes in the field. Many of the top predictors in the TF models for seed, silique and dry weight phenotypes for instance are known to function in the floral transition. From the perspective of our experimental setup, it makes sense that such genes are recovered, as it is known that the floral transition starts in autumn in winter-type *B. napus* accessions (78, 79), i.e. around the time that rosette leaves were harvested for RNA-seq in the present field trial, and is set in motion to a large extent by systemic signals emanating from leaves in Brassicaceae and other plant families (80-82).

888

889 Mechanistic interpretation of the correlational links between top predictor genes and
890 phenotypes is however not straightforward. Putative orthologs of *AtHB1* and *AtHB16* are for
891 instance found among the top predictors positively correlated with both leaf and yield
892 phenotypes (**Fig. 4**, **Additional File 2: Fig. S9**), but upregulation of these genes in *A. thaliana*
893 was previously found to lead to smaller and more serrated leaves (35, 36), to delay the
894 vegetative-to-reproductive phase transition and to result in siliques bearing fewer seeds (34,
895 36). Some top predictors that correlate negatively with yield phenotypes have putative *A.*
896 *thaliana* orthologs that are thought to function primarily as negative regulators of the floral
897 transition in leaves, e.g. *AtNF-YA* genes (58, 59), but others are putatively orthologous to a
898 positive regulator of the floral transition, such as *AtFUL*. Other floral transition regulators
899 recovered as predictors in our yield models, e.g. orthologs of *AtFLC* and *AtEFM*, do not by
900 themselves exhibit a significant positive or negative correlation with yield phenotypes.

901

902 Most likely, the associations recovered between individual plant phenotypes and autumnal
903 leaf gene expression patterns are due to developmental timing differences among the plants,
904 rather than reflecting the effects of upregulation or downregulation of specific regulators. In
905 the *A. thaliana* developmental gene expression atlas of Klepikova et al. (83), orthologs of
906 predictors positively correlated with leaf size such as *AtHB1* and *AtHB16* are more highly
907 expressed in mature *A. thaliana* leaves (at flowering), while orthologs of predictors negatively
908 correlated with leaf size such as *AtREV*, *AtWOX5* and *AtHAT3* are more highly expressed in
909 young leaves. This suggests that plants with low expression of *AtHB1*/*16* orthologs and high
910 expression of *AtREV*/*AtWOX5*/*AtHAT3* orthologs had a more juvenile (and hence smaller) leaf
911 8 at the autumnal sampling time point, which explains the observed gene expression-leaf

912 phenotype correlations. That autumnal leaf phenotypes and final yield phenotypes have
913 several developmental predictors in common (e.g. *AtHB1*) and that the autumnal leaf
914 phenotypes themselves are also predictive of yield indicates that the developmental
915 differences in autumn impacted final yield. These differences were not limited to differences
916 in leaf development, as evidenced by the fact that the predictor sets for both leaf and yield
917 phenotypes also contain regulators of plant-wide developmental phase transitions occurring
918 in autumn (juvenile-to-adult, vegetative-to-reproductive).
919
920 In summary, our results indicate that the yield potential of the individual plants was already
921 determined to a large extent by their developmental state at the time of leaf sampling in
922 autumn. Mendham and Scott (84) previously found that the size of winter-type *B. napus*
923 plants at the time of inflorescence initiation affects their yield potential, in the context of an
924 experiment assessing sowing date effects on yield. Our results show that even when sown on
925 the same date in the same field, individual winter-type *B. napus* plants of the same line display
926 developmental differences in autumn that correlate with yield differences in spring. Even if
927 only part of the variability in e.g. total seed weight (CV = 46.9%) observed in our trial is due to
928 autumnal effects on plant growth and development, the gains of mitigating such effects could
929 be substantial.
930
931 The question remains however what could have caused the developmental differences
932 among plants in the present field trial. One potential cause is differences in seed germination
933 and seedling emergence across the field. In wheat, it was established previously that relative
934 differences in seedling emergence date are strongly correlated with differences in final yield
935 (85). Next to seed quality, many environmental factors are known to impact the timing of

936 seed germination and seedling emergence, including soil structure (86), soil temperature (87),
937 sowing depth (85, 87), soil water potential, oxygenation and light quality (88), and soil
938 nutrients (89). The seedling emergence date was not recorded in the present field trial, but
939 the closest proxy that was measured, namely rosette area at 14 DAS, was found to be a bad
940 predictor for yield, indicating that variation in seed germination and seedling emergence
941 across the field did not by themselves have a major impact on yield. The observation that
942 later rosette areas are progressively better at predicting yield rather suggests that
943 developmental differences among plants accumulated over time.
944
945 The observation that genes involved in the regulation of circadian rhythm, photoperiodism
946 and the vegetative-to-reproductive phase transition are on average more spatially
947 autocorrelated in the autumnal gene expression dataset than the average gene suggests that
948 spatially patterned micro-environmental factors may be linked to the variability of
949 developmental gene expression in autumn, and ultimately yield variability in spring. That the
950 phenotypes are influenced by environmental factors is also suggested by the observation that
951 the sets of genes associated with leaf and yield phenotypes are heavily enriched in genes
952 involved in responses to abiotic and biotic stimuli and nutrient levels (**Additional File 6: Table**
953 **S5**). The finding that developmental processes feature more prominently in the TF-based
954 phenotype prediction models than responses to environmental stimuli indicates that micro-
955 environmental variations among plants in the present field trial may have influenced plant
956 phenotypes mainly by influencing development. More work is needed however to establish
957 whether and how micro-environmental variability impacts the growth and development of
958 individual plants in the same field. To address this, additional field trials need to be performed
959 in which, next to the gene expression and phenotypes of individual plants, also a range of

960 environmental parameters is measured on the single-plant level (e.g. soil structure and
961 chemistry).
962
963 Additional single-plant field trials are also needed to assess to what extent the predictive
964 models, gene-phenotype and process-phenotype associations learned from the present field
965 trial generalize to other soils and meteorological conditions, other time points or tissues
966 sampled for RNA-seq, and other cultivars. Given the developmental nature of many of the top
967 predictors in the current models, it is likely that our current prediction models, based on leaf
968 gene expression data for a single field trial at a single time point, will not perform well when
969 applied on follow-up field trials, even when using the same cultivar in a similar field under
970 roughly the same climate conditions. Differences in weather conditions and other
971 environmental factors across trials may for instance influence the timing of developmental
972 phase transitions, making it all but impossible to sample the exact same developmental time
973 point in follow-up trials. If leaf gene expression were to be profiled at a slightly earlier or later
974 developmental time point than in the present trial, the current top predictors may no longer
975 be adequate phenotype proxies and other genes that function earlier or later in e.g. the floral
976 transition may become relevant instead. The construction of robust prediction models will
977 therefore likely require single-plant data generated under a wide variety of field conditions
978 and sampling schemes. We want to emphasize however that quantitative prediction of single-
979 plant phenotypes is not the primary goal we envision for single-plant omics experiments.
980 Rather, the primary aim is to identify which biological processes, environmental factors and
981 associated genes may influence plant phenotypes in the field. In this respect, any additional
982 genes and processes identified in follow-up trials would add to our overall knowledge on how
983 rapeseed plants grow in a field.

It is worth pointing out that the dataset generated in this study may also serve other purposes than gene-phenotype association. Earlier, we have shown that field-generated single-plant transcriptomics data can also be used efficiently to predict the function of genes (28). Given the complex genome duplication history of *B. napus* (90), the combination of gene function prediction and gene-phenotype association may be particularly useful to shed light on which *B. napus* genes in a (long) list of paralogs are most likely functionally orthologous to a given *A. thaliana* gene, and how paralogs have diverged in function. This knowledge may in turn be useful in the context of genetic engineering and breeding efforts to optimize yield and stress tolerance in *B. napus*.

CONCLUSIONS

We have shown that individual *B. napus* plants of the same background grown in the same field exhibit considerable variation in gene expression and phenotypes, and that the plants' autumnal gene expression profiles have predictive power for their yield in spring. Many of the top yield predictor genes are associated with developmental processes occurring in autumn in winter-type *B. napus*, such as the juvenile-to-adult and floral transitions. Together, this indicates that autumnal development has a major influence on the yield potential of winter-type *B. napus* plants. In summary, our data show that profiling individual plants under uncontrolled field conditions is a valid strategy for identifying genes and processes influencing crop yield in the field.

METHODS

Field trial setup

Seeds from the winter-type *Brassica napus* accession Darmor (BnASSYST-120) were sown in a field in Merelbeke, Belgium (50°58'24.9"N 3°46'49.1"E) on September 8, 2016. Three seeds were sown at ~2 cm depth at each of 100 points arranged in a 10x10 grid with 0.5 m spacing within and between rows. Seedlings were thinned out to leave one seedling growing at each grid point. Early- and late-emerging seedlings were pruned preferentially (based on visual assessment) to make the remaining seedling population as homogeneous as possible. At two points, no seedlings emerged.

Plots of *Miscanthus sinensis*, *M. sacchariflorus* and *Miscanthus* hybrids were grown to the northeast and southeast of the *B. napus* field trial, and maize was grown to the northwest, at distances >5 m. The field plot was surrounded by chicken wire and covered by netting to keep out birds and large herbivores. The netting was removed in spring when plants grew taller than ~1 m. Additionally, perimeter fencing was used to protect the field trial and the mobile weather station on site (see **Additional File 1: Table S1** for weather station data).

After germination, individual plant images were taken twice a week between September 22 and October 20, 2016 (9 time points) to assess the projected leaf area of the growing rosettes. Nadir images were taken using a D90 camera (Nikon Inc., USA) equipped with a 35 mm lens (AF-S DX Nikkor 35 mm F1.8G, Nikon Inc., USA) set at iso 200, f/8. The shutter speed could vary to allow for a proper exposure, determined by the camera. The camera/tripod was positioned away from the sun to avoid shadows in the images taken. For each time point a grey calibration card (Novoflex grey card 15 x 20 cm, NOVOFLEX Präzisionstechnik GmbH,

Germany) was used to correct the white balance. This card was also used as reference to convert pixels to areas in cm^2 (see below). The ground sampling difference (GSD) was 0.015 cm/pixel.

At 74 DAS, the length and width of leaf 6 (counting upward from the first true leaf) were measured non-destructively, leaf 6 lesion and total rosette lesion severity were scored and the number of fully emerged rosette leaves (area > ~2 cm^2) was recorded. At 76 DAS, leaf 8 length and width were measured and leaf 8 lesions were scored. The width of leaf blades was measured at the widest point. Leaf lengths were measured from the leaf tip to the point where the petiole first lacked conspicuous laminar tissue (looking from the leaf tip toward the base). Lesion severity was scored qualitatively on a scale from 0 (lesions cover at most five percent of the leaf blade or rosette) to 2 (more than half of the leaf or rosette eaten).

At 81 DAS, on November 28, 2016, the eighth rosette leaves of 62 non-border plants (i.e. the plants at all non-border locations where seedlings emerged) were harvested for RNA-sequencing in a time span of ~1 hour (13:25-14:27). Leaves were cut off where the petiole first lacked conspicuous laminar tissue (looking from the leaf tip toward the base) and washed with DEPC-treated and sterilized water. The chlorophyll content of each leaf was measured at four different positions on the leaf with a CCM-200 chlorophyll content meter (Opti-Sciences, Inc., Hudson, USA), and the average of these measurements was used in the analyses. Leaves were then photographed twice against a white background with a piece of millimeter paper to assess the image scale and perspective, a ruler, and color and greyscale references, the second time covered with a glass plate to flatten them. Next, the midvein of every leaf was cut out using scissors, and the residual leaf material was folded into a pre-weighed 50 ml tube. The filled 50 ml tube and the midvein were weighed together to measure

leaf fresh weight, after which the tube was stored in liquid nitrogen on the field. The entire leaf processing pipeline, from cutting a leaf to storing it in liquid nitrogen, was completed for each leaf in less than 5 minutes.

After leaf sampling, the plants were left to overwinter and set seed in spring. After bolting, plant height was measured from ground level to the top of the primary flowering stem at 13 time points between 189 and 231 DAS (**Additional File 1: Table S1**). One of the plants sampled in autumn for RNA-seq, 01C, did not survive until the end of the growth season. The remaining 61 non-border plants were harvested on June 13, 2017 (278 DAS), at which time ~50% of seeds had started changing color from green to black but no significant pod shattering or seed predation had occurred. Final plant height at 278 DAS was measured on the field, from ground level to the top of the primary flowering stem. Afterwards, shoots were cut off and the root systems were dug up. Taproot length was measured from ground level to the deepest root tip. Root system width was measured perpendicular to the taproot at the root system's widest point.

For each harvested plant, the primary flowering stem plus its cauline secondary inflorescences (stem 1) and the secondary inflorescence stems branching at ground level (side stems) were dried in two separate bags in a well-ventilated, dry attic. The number of branches and siliques per stem, the total shoot dry weight and the dry weight of stem 1 were measured on dried plants. Seeds were recovered manually from the dried-out pods for stem 1 and the side stems separately, and separated from dust and small pod debris using a customized seed aspirator with vibration channel (Baumann Saatzuchtbedarf GmbH, Waldenburg, Germany). The resulting seed batches for stem 1 and the side stems were weighed and counted using an

elmor C3 seed counter (elmor AG, Schwyz, Switzerland). Seed counts and weights are reported for stem 1 and the entire plant (i.e. the sum of stem 1 and the side stems).

Determination of shoot growth parameters

Shoot growth parameters (time of maximum shoot growth t_m, maximum shoot growth rate and the end of shoot growth t_e) were derived by fitting a beta-sigmoid growth curve to the time series of 14 plant height measurements between 189 and 278 DAS (91) :

$$h(t) = h_0 + (h_{max} - h_0) * \left(1 + \frac{t_e - t}{t_e - t_m}\right) * \left(\frac{t}{t_e}\right)^{\frac{t_e}{t_e - t_m}} \qquad t < t_e$$

$$h(t) = h_{max} \qquad t \geq t_e \qquad \textbf{(Eq. 1)}$$

With $h(t)$ the plant height at plant age t, h_0 and h_{max} the initial and final plant height at $t = 0$ and $t = t_e$, respectively, t_e the plant age at the end of growth and t_m the plant age at the moment of maximal growth. Before curve fitting, the time points in day of year (DOY) at which the plant heights were measured were translated to plant ages t in growing degree days (GDD), i.e. $t(i) = \sum_{j=0}^{j=i} max(T_j - T_b, 0)$ with i the time point in DOY, T_j the average air temperature at j DOY (**Additional File 1: Table S1**) and $T_b = 5$ °C a base temperature below which no growth is assumed to occur (79, 92). Optimization of the parameters h_0, h_{max}, t_e and t_m was done with the *nls* function in R using the 'port' algorithm. The maximum shoot growth rate was obtained by calculating the derivative of $h(t)$ (**Eq. 1**) at t_m. After curve fitting, the values obtained for t_m and t_e were converted back from GDD to DOY and subsequently to DAS.

Image-based phenotyping

Leaf 8 areas (81 DAS) were estimated by segmenting the flattened leaf images taken at the time of leaf harvest. The millimeter grid scale on each image was used to correct for perspective distortion and to create a uniform spatial resolution across the entire image of 100 pixels per cm. Images were cropped to remove the grid scale and sample label. Segmentation was done by training a U-Net convolutional neural network (93) on a small dataset of 25 images for which random patches of foreground and background were annotated using VGG Image Annotator (via) v:2.0.7 (94). Random cropping, resizing, rotating (by multiples of 90 degrees), mirroring, color-jittering and gaussian blurring were applied to artificially increase the training dataset size. The training was done using the Adam optimizer (95) in Pytorch v:1.7.1 (96) with default settings. The pixel-wise cross-entropy loss was back-propagated only for annotated regions of each image. The learning rate was initially set to 1e-3 and was automatically halved as soon as the minimal training loss stagnated for more than 3 epochs. The network was trained for 16 epochs. The trained network was validated by visually evaluating it on unseen images, and then applied to all flattened leaf images.

Leaf 8 length and width at 81 DAS were measured on the flattened leaf images using ImageJ v:1.50 (97). For measuring leaf 8 length, the midvein was traced from the leaf tip to the cutting point (i.e. where the petiole first lacked conspicuous laminar tissue) using the ImageJ segmented line tool. Leaf 8 width was measured at the widest point.

For measuring the projected area of the rosettes photographed at 42 DAS (i.e. the rosette imaging date closest to leaf sampling), a dedicated script was developed using the image analysis software HALCON (version 13.0.1.1, MVTec Software GmbH, Germany). First, the images were cropped to remove parts of adjacent plants visible on the pictures. To remove noise, both a gentle Gaussian filter and a median filter were applied. Each RGB image was

then converted to the HSV color space, where the Hue channel was used to select the green plant parts using a threshold range for the green pixels (34-80) defined based on trial and error. Care was taken to also include the petioles. After this, a 'closing_circle' operator was used and remaining small lesions (due to insect damage) were filled up using the 'fill_up' operator. Only the largest segmented area was taken into account, to differentiate between the plant of interest and small weeds nearby.

The HALCON segmentation strategy worked well for the rosette images taken at 42 DAS, but regularly produced segmentation errors for images of smaller rosettes taken closer to the sowing date. An alternative segmentation approach was therefore used on rosette images taken at 14, 18, 21, 25, 28, 32, 35 and 39 DAS (and 42 DAS as control). The main difficulty for the earlier time points is distinguishing small rosettes from weeds and other distracting objects occurring on the field. This requires an algorithm with a larger field of view than what a HALCON script or standard U-net (see above) can provide. Instead, a standard pre-trained DenseNet M161 (98) was taken and augmented with additional bilinear upsampling layers after each 'dense' layer of the original algorithm. That is, the last feature layer of DenseNet was upsampled with bilinear interpolation and a weighted sum was made with the higher resolution 'dense' features. This was repeated for each dense layer up to the original input resolution. The network was trained for 175 epochs (final mean epoch loss = 0.01) on 54 hand-labeled images (6 images per time point) using stochastic gradient descent (SGD) with momentum (learning rate = 0.001 and momentum = 0.99). The learning rate was divided by 10 each time the train loss plateaued for more than 4 epochs. Image rotations, mirroring and HSV augmentations were used to augment the training data. The trained model was used to segment all rosette images. After segmentation, a post-processing step was performed to

remove segmented parts of *B. napus* plants adjacent to the plant of interest and remaining weeds, using scikit-image v: 0.19.2 (99). Only the connected component closest to the centroid of the image and other components within a 25-pixel distance of this central component (e.g. leaves of which the stalk was segmented incorrectly because of a lower chlorophyll content) were associated with the plant of interest. Connected components with an area less than 10,000 pixels were filtered out to eliminate small weeds. This approach was evaluated visually for all segmentations and proved to work well for most plants. Segmentations with missing plant parts or weeds that weren't filtered out by this post-processing step were manually corrected. A grey calibration card (Novoflex grey card 15 x 20 cm, NOVOFLEX Präzisionstechnik GmbH, Germany) was used as a reference to convert pixels to areas in cm^2. The projected rosette areas at 42 DAS estimated by this segmentation approach exhibit a Pearson correlation of 0.997 with the areas estimated by the aforementioned HALCON script.

RNA sequencing

The frozen leaf samples for the 62 harvested non-border plants were grinded, and total RNA was extracted using the guanidinium thiocyanate-phenol-chloroform extraction method using TRI-reagent (Thermo Fisher Scientific) followed by DNA digestion using the RQ1 RNase-free DNase kit (Promega). ds cDNA was prepared using the Maxima H Minus Double-Stranded cDNA Synthesis Kit (#K2561, Thermo Fisher Scientific) to a concentration of ~17-38 ng/ul in 10mM Tris-Cl buffer (pH 8.5) at a minimum volume of 30ul. (~0.6 - 1.1 ug total). ds cDNA samples were sent to the University of Missouri Genomics Technology Core, where library preparation was performed (average insert size of 500 bp) using the Illumina TruSeq DNA PCR-Free Library Prep Kit according to the protocol described in (100). 250 bp paired-end

sequencing was performed at the Tufts University Genomics Core on an Illumina HiSeq 2500 machine in Rapid Run mode. The samples were sequenced in 3 batches (**Additional File 1: Table S1**).

The raw RNA-seq data was processed using a custom Galaxy pipeline (101) implementing the following steps. First, the fastq files were quality-checked using FastQC (v:0.5.1) (102). Next, Trimmomatic (v:0.32.1) (103) was used to remove adapters, read fragments with average quality below 20 and trimmed reads shorter than 125 base pairs. The trimmed and filtered reads were mapped to the *Brassica napus* Darmor-bzh reference genome v:5 (https://www.genoscope.cns.fr/brassicanapus/data/) (90) using HISAT2 v:2.0.5 (104) with default values for all parameters. Only the uniquely mapping reads or (in the case of multiple mappings) the best secondary alignment were kept for the following analyses. The mapping files were quantified using HTSeq v:0.6.1p1 (105) with the option 'Intersection-union', using the genome annotation of the *Brassica napus* Darmor-bzh reference genome v:5 (https://www.genoscope.cns.fr/brassicanapus/data/). No filtering steps were performed during preprocessing except for removing genes that were not expressed in any samples. Counts were normalized across samples and batches using a modified regularized log (*rlog*) model of the DESeq2 (106) package in R. Counts are still modeled in the same way as in the original *rlog* implementation, that is :

$$k_{ij} \sim NB\left(\mu_{ij}, \alpha_i\right)$$

$$\mu_{ij} = s_j \times q_{ij} \qquad \textbf{(Eq. 2)}$$

$$\log_2\left(q_{ij}\right) = \mathbf{x}_j \cdot \boldsymbol{\beta}_i$$

Where $k_{ij} \in \mathbb{N}^+$ is the count of gene i in sample j, which is assumed to be sampled from a negative binomial distribution (*NB*) with estimated mean $\mu_{ij} \in \mathbb{R}^+$ and estimated dispersion of the ith gene α_i. μ_{ij} is taken as the expected count q_{ij} for a 'typical' library size (i.e. with a size factor s_j = 1), scaled by a library size normalization factor s_j for sample j. Note that q_{ij} still contains batch effects : $\mathbf{x}_j \in \mathbb{R}^p$ is a vector of $p = 65$ predictors for sample j, including an intercept, 2 dummy variables for the smallest sequencing batches (1 and 3) that capture batch effects relative to the largest sequencing batch (2, the effects of which are absorbed in the intercept) and dummy variables for each of the 62 plants that were sampled. $\boldsymbol{\beta}_i \in \mathbb{R}^p$ contains the estimated coefficients for those predictors for gene i. As in (106), an empirical Bayes shrinkage procedure is used to estimate $\boldsymbol{\beta}_i$, using a flat prior for the intercept β_{i0} and the sequencing batch coefficients, and a zero-centered normal prior for each plant coefficient β_{ip_j} (with p_j the index of the plant corresponding to sample j), with prior variance estimated using quantile matching as described in Love et al. (106). There are only two differences compared to Love et al. (106): the first is the addition of two batch coefficients as fixed effects in the design matrix, and the second is that log-fold changes used in the prior random effect variance computation are estimated relative to the mean of each batch instead of to the mean of all samples. Once the model is estimated, *rlog* counts are computed as in Love et al. (106), that is:

$$rlog_{ij} \equiv \beta_{i0} + \beta_{ip_j} \qquad \textbf{(Eq. 3)}$$

Note that all samples j belonging to the same plant (technical repeats) have the same value for β_{ip_j}. The modified *rlog* transformation removes library size effects and batch effects,

unites technical repeats into one estimate and log-transforms the data (reducing heteroscedasticity) in a single step. In addition, using random effects for each plant allows pooling information from technical repeats while simultaneously basing variance estimates on all samples (including samples without technical repeats). This method therefore makes maximal use of the available data. The resulting data is show in **Additional File 2: Fig. S10**.

SNP detection and population structure analysis

Trimmed and filtered RNA-seq reads were aligned to the *Brassica napus* Darmor-bzh reference genome v:5 (https://www.genoscope.cns.fr/brassicanapus/data/) (90) using HISAT2 v:2.0.5 (104) with default values for all parameters. Genomic variants were detected for each plant using NGSEP v:3.3.2 (107) on the aligned reads. For downstream analyses, we focused on biallelic SNPs with a minimum genotype quality of 40 and called in at least 49 samples (80% of the population). Missing calls were imputed using Beagle v:5.1 (108) using default parameters, and only SNPs with minor allele frequency (MAF) ≥ 0.05 after imputation were kept, resulting in a dataset of 23,188 SNPs.

A neighbor-joining tree was made based on the SNP dataset with TASSEL v:5.2.60 (109), using 1-IBS (identity by state) as the distance measure while setting the distance from an individual to itself to zero. The tree was rendered using the polar tree layout in FigTree v:1.4.3 (110).

Spatial autocorrelation analysis

Moran's I was calculated for each gene (phenotype) as $I = \frac{n}{w} \frac{(\mathbf{x}-\bar{\mathbf{x}})^T \mathbf{C} (\mathbf{x}-\bar{\mathbf{x}})}{\|(\mathbf{x}-\bar{\mathbf{x}})\|^2}$. Where $\mathbf{x}$ is a column vector of *rlog* gene expression (phenotype) values, n is the number of samples and w

is the sum of elements of the connectivity matrix $\mathbf{C}$. For $\mathbf{C}$ a binary $n \times n$ 'queens' connectivity matrix was chosen, meaning that neighboring horizontal, vertical and diagonal plants are seen as connected. Note that $\mathbf{C}$ can differ from one phenotype to the next since not all phenotypes were available for all samples. For each gene (phenotype), the Moran's I was recalculated on 10^5 random permutations of $\mathbf{x}$ to obtain an empirical null distribution, which was then compared to the real Moran's I to obtain a *p*-value. Finally, *p*-values were corrected across all genes (phenotypes) using the Benjamini-Hochberg (BH) procedure (111). All calculations were done using the PySAL python library (112).

Variance analysis

Principal component analysis (PCA) was done on various normalized versions of the gene expression count matrix and on the phenotype dataset (including qualitative phenotypes such as leaf 6 lesion severity (74 DAS) but excluding the plant height and rosette area time series except for the final time points, i.e. plant height (278 DAS) and rosette area (42 DAS)), using the 'prcomp' function in the R stats package on the centered gene expression datasets and the 'ppca' method in pcaMethods v:1.88.0 (113) on the z-scored phenotype dataset. Phenotype distributions were plotted using the 'histogram' function in Matlab R2018b with probability normalization option. Shapiro-Wilk and Anderson-Darling tests were performed using the 'normalitytest' script (114) and 'adtest' functions in Matlab R2018b, respectively. Outliers were defined as values more than three scaled median absolute deviations (MAD) away from the median, as is default in the Matlab R2018b 'isOutlier' function. Outliers were only removed for the purpose of calculating their effect on the phenotypes' normality and coefficient of variation (CV), all other analyses used the complete phenotype dataset.

Normalized coefficients of variation (*normCV*s) for gene expression profiles were computed on batch and library size corrected data (without *rlog* transform). Normalized counts were obtained as $x_{ij} = k_{ij}/(\beta_{ib_j} \times s_j)$ where β_{ib_j} is the batch effect for gene i in sample j as estimated in the *rlog* calculation (see above). Since batch 2 is absorbed in the intercept, $\beta_{ib_j} = 1$ for samples of batch 2. Contrary to the *rlog* transform, this method does not collapse technical repeats, and they were instead collapsed by averaging (as in **Additional File 2: Fig. S10**, panel B, but without the $\log_2$-transform). From here on, variance analysis followed the same procedure as described in Cruz, De Meyer et al. (28). Briefly, a trendline was fitted to the CV^2 versus mean expression relationship (omitting genes expressed in <10 samples) using a generalized linear model (GLM) of the gamma family with identity link of the form $CV^2(\mathbf{x}) = {a}/{\bar{\mathbf{x}}} + b$, with fitting parameters a and b (115) (**Additional File 2: Fig. S11**). Code from the M3Drop R package (116) was used for this purpose. A normalized CV accounting for the observed mean-variance relationship was then calculated as $normCV(\mathbf{x}) = \log_2(CV^2(\mathbf{x})/trend(\bar{\mathbf{x}}))$ where $trend(\bar{\mathbf{x}})$ is the fitted value at the mean of $\mathbf{x}$.

GO enrichment analysis

A Gene Ontology (GO) annotation for *Brassica napus* was generated using the TRAPID v.2.0 platform (117) with default parameters on April 16, 2020. Transcript sequences parsed from the *B. napus* Darmor-bzh reference genome annotation v:5 (90) using the gffread v.0.9.6 utility (118) were used as input for TRAPID, and PLAZA 4.5 dicots (119) was used as the reference database. GO enrichment *p*-values were calculated with hypergeometric tests and adjusted for multiple testing (*q*-values) using the BH procedure (111), either using custom R scripts or using BiNGO v:3.0.3 (120). GO categories gravitating toward the top or bottom of

gene lists ranked in order of decreasing Moran's I or normalized CV were detected using two-sided Mann-Whitney U tests (with genes belonging to the GO category of interest classified as group 1 and other genes as group 2), as implemented in the 'wilcox.test' function in the R stats package v:4.0.5, followed by BH *p*-value adjustment.

Ortholog inference

Putative *A. thaliana* orthologs of *B. napus* genes were identified in two steps. First, putative orthologs of *B. napus* genes were identified in *B. rapa* and *B. oleracea* (source of the A and C subgenomes of *B. napus*, respectively), based on best similarity hits returned by TRAPID v.2.0 (117) and on the syntenic relationships reported in Chalhoub et al. (90) and Sun et al. (121). Second, putative *A. thaliana* orthologs of the identified *B. rapa* and *B. oleracea* genes were retrieved from PLAZA 4.5 dicots (119), which provides orthology inferences integrating four different lines of evidence : orthogroup inference within gene families using OrthoFinder (122), orthology inference using gene tree-species tree reconciliation, orthology inference from best DIAMOND (123) hits and their inparalogs, and positional orthology inference through collinearity analysis (124). The most likely *A. thaliana* orthologs of a given *B. napus* gene were taken to be the putative orthologs that are most strongly supported across both inference steps.

Single-feature phenotype prediction models

Single-gene models. Linear mixed-effects (lme) models (125) were fitted to predict a phenotype given the expression profile of a single gene. That is, given a phenotype vector $\mathbf{y}$

and a vector $\mathbf{x}$ of a given gene's z-scored expression values across the field, we fit the following model:

$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{x} + \boldsymbol{\varepsilon}$$
$$\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Sigma) \quad \textbf{(Eq. 4)}$$

where β_0 is the intercept (average phenotype value), β_1 the gene effect coefficient, and $\boldsymbol{\varepsilon}$ the residual error which is assumed to follow a multivariate normal distribution with a Gaussian covariance structure Σ given by:

$$\Sigma_{ij} = \sigma_\varepsilon^2 \times \left(\nu \times I_{ij} + (1 - \nu) \times exp\left[-\left({}^{d_{ij}}/_{r} \right)^2 \right] \right) \quad \textbf{(Eq. 5)}$$

where d_{ij} is the physical distance between plant i and j on the field, σ_ε^2 is the overall residual phenotype variance, the nugget ν (between 0 and 1) determines the proportion of the residual variance that is independently and identically distributed (*iid*) as opposed to governed by spatial autocorrelation, the range r determines how fast the residual phenotype correlation between plants drops when the distance between them increases, and I is an identity matrix. The same model form was used to predict final yield phenotypes, e.g. total seed weight, as a function of one of the phenotypes measured early in the growing season, e.g. leaf 8 area (81 DAS). All parameters $(\beta_0, \beta_1, \sigma_\varepsilon, n, r)$ are estimated from the data by Restricted Maximum Likelihood (ReML) estimation, implemented in the nlme package (126) in R. In some cases the lme model didn't converge and a regular linear model (lm) was used instead. *p*-values for the β_1 coefficients were determined using Wald tests and adjusted for multiple testing using the BH procedure (111).

For each of the 100 genes with the lowest BH-adjusted β_1 *p*-value for a given phenotype, a 9-times repeated 10-fold cross-validation scheme was used to assess the gene's predictive power (see section on multi-gene models for details). The median test R^2score across all 90 splits was used as a measure of prediction performance.

Single-phenotype models. The same linear mixed-effects (lme) modeling and cross-validation strategy as used for the single-gene models was also used also to model spring phenotypes as a function of autumnal leaf or rosette phenotypes. Leaf 6 and leaf 8 phenotypes and the rosette area at 42 DAS were used as features for predicting all spring phenotypes. In a separate analysis, also earlier rosette areas (14-42 DAS) were used as features, in order to assess how the predictive power of the projected rosette area for yield phenotypes evolves over time.

Alternative single-gene models for ratio phenotypes. For seeds per silique (on stem 1 or the entire plant), the following alternative log-link model was fitted using the nlme package (126) in R :

$$\ln(E(\mathbf{n} \oslash \mathbf{d})) = \beta_0 + \beta_1 \mathbf{x} \qquad \textbf{(Eq. 6)}$$

where $\oslash$ stands for the element-wise division of the numerator **n**, a vector containing the seed count stem 1 for all plants, by the denominator **d**, a vector containing the silique count stem 1 for all plants. **x** is the expression profile of a given gene across plants. The numerator is assumed to follow a normal distribution given the denominator **d** and the gene expression profile $\mathbf{x}$:

$$\mathbf{n} \sim \mathcal{N}(\mathbf{d} \cdot \exp(\beta_0 + \beta_1 \mathbf{x}), \Sigma) \quad \textbf{(Eq. 7)}$$

Various error models Σ were tried out. For each gene, Σ is either a constant σ^2 across all plants estimated from the data, a spatially covarying error structure (using a Gaussian covariance structure as for the other single-gene models, see above), a heteroscedastic error structure with the error variance increasing linearly with the estimate, or a both spatially covarying and heteroscedastic error structure. The parameters β_0, β_1, σ^2 (and optionally the nugget and range for spatial models) were estimated using the 'ngls' function in nlme. *p*-values for the gene expression coefficients β_1 were determined using Wald tests and adjusted for multiple testing using the BH procedure (111).

A similar model was used for the branches per stem phenotype :

$$\ln(E((\mathrm{c} + \mathbf{n}) \oslash \mathbf{d})) = \beta_0 + \beta_1 \mathbf{x} \quad \textbf{(Eq. 8)}$$

where **n** is a vector containing the total branch count for all plants, **d** is a vector containing the stem count for all plants, and c is an extra offset introduced to account for the amount of branches per stem decreasing with increasing numbers of stems on a plant.

Multi-feature phenotype prediction models

Multi-gene models. Predictive models were made for each phenotype based on z-scored *rlog* gene expression data, using either all genes or only transcription factors as potential features. Random forest (127) and elastic net (128) models were constructed with scikit-learn v:0.23.2 (129) using a 10-fold cross-validation scheme. Model learning on the training data in each

cross-validation split was done in two steps. First a feature selection model was used to select promising features, and then a random forest or elastic net model was built on the selected features. Three methods were used as alternatives for feature selection. The first feature selection technique used was HSIC lasso (130) as implemented in the pyHSICLasso package (131), which generally selected at most 200 genes. The second feature selection technique was a filter selecting gene expression profiles exhibiting a significant Spearman correlation with the phenotype of interest (BH-adjusted $q \leq 0.01$; if no features survived this filter, the threshold was set at $p \leq 0.001$). The third feature selection technique was a filter selecting genes with *rlog* gene expression > 0 in at least half of the samples (median *rlog* gene expression > 0). Elastic net models were built using a fourfold inner cross-validation loop to estimate the model hyperparameters. For random forest models, 1000 trees were estimated (n_estimators = 1000) using bootstrapping (bootstrap=True), and $\sqrt{n}$ features (with n the total number of features) were considered when looking for the best split (max_features = "auto"). The hyperparameters 'max_depth' (the maximum number of nodes) and 'min_samples_leaf' (the minimal number of samples at each leaf node) were optimized using a grid search with possible values (1, 2, 5, 10, 20, 50) and (1, 2, 5) for 'max_depth' and 'min_samples_leaf', respectively. Optimal hyperparameters were selected based on generalization scores on out-of-bag (oob) samples (oob_score=True).

For each combination of phenotype, machine learning method and feature selection technique, 9 repeats of the aforementioned 10-fold cross-validation scheme were performed, giving rise to 90 train-test data splits in total. For each split, an out-of-bag (oob) R^2 score was computed from the predicted and observed phenotype values in the test set, and the median oob R^2 across all 90 splits (= median test R^2) is reported as a measure of model prediction

performance. Alternative R^2 values and Pearson correlation (PCC) values were computed based on the combined set of test predictions across all 10 splits of a cross-validation repeat. The medians of those R^2 and PCC values across the 9 cross-validation repeats for a given model are reported as the median pooled R^2 and median pooled PCC score of the model, respectively.

For both elastic net and random forest models, genes of potential interest for a given phenotype were ranked based on their median importance across the 90 cross-validation splits of the model version with the highest median test R^2score (the difference between model versions being the use of different feature selection techniques). For random forest models, the gini importance of a gene was used as its importance score. For elastic net models, the absolute value of a gene's estimated model coefficient was used.

Models on permuted datasets. For all continuous and high-count phenotypes and for both the 'all genes' and 'transcription factors' feature sets, models were trained and tested on 90 datasets in which the phenotype values were permuted, using the same machine learning method and feature selection technique as for the model with the best median test R^2 score on real data for the given phenotype and feature set. For each phenotype and feature set, one model was trained per permuted dataset, using a single 90-10 train-test split mimicking one fold of the cross validation setup used on real data.

Multi-phenotype models. For all phenotypes measured in spring, additional predictive models were made based on z-scored data for 14 leaf and rosette phenotypes measured in the preceding autumn. We used the same modeling approach as for the expression-based

models (random forest and elastic net, 9 repeats of 10-fold nested cross-validation), except that the feature selection step of the expression-based modeling protocol was skipped given the low number of potential model features. In this respect, using elastic net models instead of a simple linear regression framework is technically also unnecessary, but elastic nets were used nevertheless to maximize comparability of the early phenotype-based and expression-based modeling results.

DECLARATIONS

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The raw RNA-seq data generated in this study is available at ArrayExpress, experiment E-MTAB-11904 (https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-11904). Data analysis scripts are available from Zenodo (https://zenodo.org/record/7072001) and GitHub (https://github.com/MMichaelVdV/Brassica_segmentation).

Competing interests

The authors declare that they have no competing interests.

Funding

SDM is a fellow of the Research Foundation-Flanders (FWO, grant 1146319N). FWO had no role in the design of the study, the collection, analysis, and interpretation of data or the writing of the manuscript.

Author contributions

SM conceived and supervised the study; TDS, PL, IR-R and SM designed the field trial ; HN, DI and IR-R provided resources ; TDS and SM set up the field trial ; SDM, DC, TDS, KB and SM phenotyped plants during the growing season ; SDM, DC, TDS, PL, JDB, KB, TVH, HN, IR-R and SM sampled leaves for RNA-seq ; JDB prepared samples for RNA-seq ; SDM, DC, KB, HS and SM harvested and phenotyped plants at the end of the field trial ; SDM, DC, TDS, PL, MVdV, SH and SM analyzed data ; SDM, DC and SM wrote the manuscript with input from the other authors. All authors read and approved the final manuscript.

Acknowledgements

The authors thank Benjamin Wittkop and Rod Snowdon for providing *B. napus* Darmor seeds, Luc van Gyseghem, Thomas Vanderstocken and Katleen Sucaet for their help with setting up and maintenance of the field trial, Dorota Herman and Kirin Demuynck for assistance with leaf sampling, and Chris Pires for advice on sample prep for RNA sequencing.

REFERENCES

1. Stewart-Ornstein J, Weissman JS, El-Samad H. Cellular noise regulons underlie fluctuations in *Saccharomyces cerevisiae*. Mol Cell. 2012;45(4):483-93.
2. Dunlop MJ, Cox RS, 3rd, Levine JH, Murray RM, Elowitz MB. Regulatory activity revealed by dynamic correlations in gene expression noise. Nat Genet. 2008;40(12):1493-8.
3. Munsky B, Neuert G, van Oudenaarden A. Using gene expression noise to understand gene regulation. Science. 2012;336(6078):183-7.
4. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol. 2014;32(4):381-6.
5. Moignard V, Woodhouse S, Haghverdi L, Lilly AJ, Tanaka Y, Wilkinson AC, et al. Decoding the regulatory network of early blood development from single-cell gene expression measurements. Nat Biotechnol. 2015;33(3):269-76.
6. Moignard V, Macaulay IC, Swiers G, Buettner F, Schutte J, Calero-Nieto FJ, et al. Characterization of transcriptional networks in blood stem and progenitor cells using high-throughput single-cell gene expression analysis. Nat Cell Biol. 2013;15(4):363-72.
7. Aibar S, González-Blas CB, Moerman T, Huynh-Thu VA, Imrichova H, Hulselmans G, et al. SCENIC: single-cell regulatory network inference and clustering. Nat Methods. 2017;14(11):1083-6.
8. van Dijk D, Sharma R, Nainys J, Yim K, Kathail P, Carr AJ, et al. Recovering gene interactions from single-cell data using data diffusion. Cell. 2018;174(3):716-29 e27.
9. Wagner A, Regev A, Yosef N. Revealing the vectors of cellular identity with single-cell genomics. Nat Biotechnol. 2016;34(11):1145-60.
10. Pratapa A, Jalihal AP, Law JN, Bharadwaj A, Murali TM. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. Nat Methods. 2020;17(2):147-54.
11. Bhosale R, Jewell JB, Hollunder J, Koo AJK, Vuylsteke M, Michoel T, et al. Predicting gene function from uncontrolled expression variation among individual wild-type *Arabidopsis* plants. Plant Cell. 2013;25(8):2865-77.
12. Massonnet C, Vile D, Fabre J, Hannah MA, Caldana C, Lisec J, et al. Probing the reproducibility of leaf growth and molecular phenotypes: a comparison of three Arabidopsis accessions cultivated in ten laboratories. Plant Physiol. 2010;152(4):2142-57.
13. Cortijo S, Bhattarai M, Locke JCW, Ahnert SE. Co-expression networks from gene expression variability between genetically identical seedlings can reveal novel regulatory relationships. Front Plant Sci. 2020;11:599464.
14. Atkinson NJ, Urwin PE. The interaction of plant biotic and abiotic stresses: from genes to the field. J Exp Bot. 2012;63(10):3523-43.
15. Barah P, Naika MBN, Jayavelu ND, Sowdhamini R, Shameer K, Bones AM. Transcriptional regulatory networks in *Arabidopsis thaliana* during single and combined stresses. Nucleic Acids Res. 2016;44(7):3147-64.
16. Cabello JV, Lodeyro AF, Zurbriggen MD. Novel perspectives for the engineering of abiotic stress tolerance in plants. Curr Opin Biotech. 2014;26:62-70.

17. Davila Olivas NH, Kruijer W, Gort G, Wijnen CL, van Loon JJ, Dicke M. Genome-wide association analysis reveals distinct genetic architectures for single and combined stress responses in *Arabidopsis thaliana*. New Phytol. 2017;213(2):838-51.
18. Johnson SM, Lim FL, Finkler A, Fromm H, Slabas AR, Knight MR. Transcriptomic analysis of *Sorghum bicolor* responding to combined heat and drought stress. BMC Genom. 2014;15:456.
19. Rasmussen S, Barah P, Suarez-Rodriguez MC, Bressendorff S, Friis P, Costantino P, et al. Transcriptome responses to combinations of stresses in Arabidopsis. Plant Physiol. 2013;161(4):1783-94.
20. Suzuki N, Rivero RM, Shulaev V, Blumwald E, Mittler R. Abiotic and biotic stress combinations. New Phytol. 2014;203(1):32-43.
21. Thoen MP, Davila Olivas NH, Kloth KJ, Coolen S, Huang PP, Aarts MG, et al. Genetic architecture of plant stress resistance: multi-trait genome-wide association mapping. New Phytol. 2017;213(3):1346-62.
22. Poorter H, Fiorani F, Pieruschka R, Wojciechowski T, van der Putten WH, Kleyer M, et al. Pampered inside, pestered outside? Differences and similarities between plants growing in controlled conditions and in the field. New Phytol. 2016;212(4):838-55.
23. Mittler R. Abiotic stress, the field environment and stress combination. Trends Plant Sci. 2006;11(1):15-9.
24. Nelissen H, Moloney M, Inzé D. Translational research: from pot to plot. Plant Biotechnol J. 2014;12(3):277-85.
25. Nelissen H, Sprenger H, Demuynck K, De Block J, Van Hautegem T, De Vliegher A, et al. From laboratory to field: yield stability and shade avoidance genes are massively differentially expressed in the field. Plant Biotechnol J. 2020;18(5):1112-4.
26. Oh SJ, Kim YS, Kwon CW, Park HK, Jeong JS, Kim JK. Overexpression of the transcription factor *AP37* in rice improves grain yield under drought conditions. Plant Physiol. 2009;150(3):1368-79.
27. Maathuis MH, Colombo D, Kalisch M, Bühlmann P. Predicting causal effects in large-scale systems from observational data. Nat Methods. 2010;7(4):247-8.
28. Cruz DF, De Meyer S, Ampe J, Sprenger H, Herman D, Van Hautegem T, et al. Using single-plant-omics in the field to link maize genes to functions and phenotypes. Mol Syst Biol. 2020;16(12):e9667.
29. FAO. World Food and Agriculture - Statistical Yearbook 2021. Rome; 2021.
30. Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. Ann Math Statist. 1947;18(1):50-60.
31. Cortijo S, Aydin Z, Ahnert S, Locke JC. Widespread inter-individual gene expression variability in *Arabidopsis thaliana*. Mol Syst Biol. 2019;15(1):e8591.
32. Efron B. Tweedie's formula and selection bias. J Am Stat Assoc. 2011;106(496):1602-14.
33. Aoyama T, Dong CH, Wu Y, Carabelli M, Sessa G, Ruberti I, et al. Ectopic expression of the Arabidopsis transcriptional activator Athb-1 alters leaf cell fate in tobacco. Plant Cell. 1995;7(11):1773-85.
34. Ribone PA, Capella M, Arce AL, Chan RL. A uORF represses the transcription factor AtHB1 in aerial tissues to avoid a deleterious phenotype. Plant Physiol. 2017;175(3):1238-53.

35. Miguel VN, Manavella PA, Chan RL, Capella MA. The AtHB1 transcription factor controls the miR164-*CUC2* regulatory node to modulate leaf development. Plant Cell Physiol. 2020;61(3):659-70.
36. Wang Y, Henriksson E, Söderman E, Henriksson KN, Sundberg E, Engström P. The Arabidopsis homeobox gene, *ATHB16*, regulates leaf development and the sensitivity to photoperiod in Arabidopsis. Dev Biol. 2003;264(1):228-39.
37. Otsuga D, DeGuzman B, Prigge MJ, Drews GN, Clark SE. *REVOLUTA* regulates meristem initiation at lateral positions. Plant J. 2001;25(2):223-36.
38. Byrne ME. Shoot meristem function and leaf polarity: the role of class III HD-ZIP genes. Plos Genet. 2006;2(6):e89.
39. Talbert PB, Adler HT, Parks DW, Comai L. The *REVOLUTA* gene is necessary for apical meristem development and for limiting cell divisions in the leaves and stems of *Arabidopsis thaliana*. Development. 1995;121(9):2723-35.
40. Prigge MJ, Otsuga D, Alonso JM, Ecker JR, Drews GN, Clark SE. Class III homeodomain-leucine zipper gene family members have overlapping, antagonistic, and distinct roles in Arabidopsis development. Plant Cell. 2005;17(1):61-76.
41. Ochando I, Jover-Gil S, Ripoll JJ, Candela H, Vera A, Ponce MR, et al. Mutations in the microRNA complementarity site of the *INCURVATA4* gene perturb meristem function and adaxialize lateral organs in Arabidopsis. Plant Physiol. 2006;141(2):607-19.
42. Baima S, Possenti M, Matteucci A, Wisman E, Altamura MM, Ruberti I, et al. The Arabidopsis ATHB-8 HD-zip protein acts as a differentiation-promoting transcription factor of the vascular meristems. Plant Physiol. 2001;126(2):643-55.
43. Gardiner J, Donner TJ, Scarpella E. Simultaneous activation of *SHR* and *ATHB8* expression defines switch to preprocambial cell state in Arabidopsis leaf development. Dev Dyn. 2011;240(1):261-70.
44. Kim J, Jung JH, Reyes JL, Kim YS, Kim SY, Chung KS, et al. microRNA-directed cleavage of *ATHB15* mRNA regulates vascular development in Arabidopsis inflorescence stems. Plant J. 2005;42(1):84-94.
45. Bou-Torrent J, Salla-Martret M, Brandt R, Musielak T, Palauqui JC, Martínez-García JF, et al. ATHB4 and HAT3, two class II HD-ZIP transcription factors, control leaf development in *Arabidopsis*. Plant Signal Behav. 2012;7(11):1382-7.
46. Brandt R, Salla-Martret M, Bou-Torrent J, Musielak T, Stahl M, Lanz C, et al. Genome-wide binding-site analysis of REVOLUTA reveals a link between leaf patterning and light-mediated growth responses. Plant J. 2012;72(1):31-42.
47. Zhang Z, Runions A, Mentink RA, Kierzkowski D, Karady M, Hashemi B, et al. A WOX/auxin biosynthesis module controls growth to shape leaf form. Curr Biol. 2020;30(24):4857-68 e6.
48. Zhang Z, Tucker E, Hermann M, Laux T. A molecular framework for the embryonic initiation of shoot meristem stem cells. Dev Cell. 2017;40(3):264-77 e4.
49. Nakata M, Matsumoto N, Tsugeki R, Rikirsch E, Laux T, Okada K. Roles of the middle domain-specific *WUSCHEL-RELATED HOMEOBOX* genes in early development of leaves in *Arabidopsis*. Plant Cell. 2012;24(2):519-35.
50. Cominelli E, Galbiati M, Vavasseur A, Conti L, Sala T, Vuylsteke M, et al. A guard-cell-specific MYB transcription factor regulates stomatal movements and plant drought tolerance. Curr Biol. 2005;15(13):1196-200.

51. Wang L, Hua D, He J, Duan Y, Chen Z, Hong X, et al. *Auxin Response Factor2* (*ARF2*) and its regulated homeodomain gene *HB33* mediate abscisic acid response in *Arabidopsis*. Plos Genet. 2011;7(7):e1002172.
52. Hong SY, Kim OK, Kim SG, Yang MS, Park CM. Nuclear import and DNA binding of the ZHD5 transcription factor is modulated by a competitive peptide inhibitor in *Arabidopsis*. J Biol Chem. 2011;286(2):1659-68.
53. Aharoni A, Dixit S, Jetter R, Thoenes E, van Arkel G, Pereira A. The SHINE clade of AP2 domain transcription factors activates wax biosynthesis, alters cuticle properties, and confers drought tolerance when overexpressed in Arabidopsis. Plant Cell. 2004;16(9):2463-80.
54. Zhao H, Wu D, Kong F, Lin K, Zhang H, Li G. The *Arabidopsis thaliana* Nuclear Factor Y transcription factors. Front Plant Sci. 2016;7:2045.
55. Zhang M, Hu X, Zhu M, Xu M, Wang L. Transcription factors NF-YA2 and NF-YA10 regulate leaf growth via auxin signaling in *Arabidopsis*. Sci Rep. 2017;7(1):1395.
56. Li WX, Oono Y, Zhu J, He XJ, Wu JM, Iida K, et al. The *Arabidopsis* NFYA5 transcription factor is regulated transcriptionally and posttranscriptionally to promote drought resistance. Plant Cell. 2008;20(8):2238-51.
57. Zhao H, Lin K, Ma L, Chen Q, Gan S, Li G. Arabidopsis NUCLEAR FACTOR Y A8 inhibits the juvenile-to-adult transition by activating transcription of *MIR156s*. J Exp Bot. 2020;71(16):4890-902.
58. Wenkel S, Turck F, Singer K, Gissot L, Le Gourrierec J, Samach A, et al. CONSTANS and the CCAAT box binding complex share a functionally important domain and interact to regulate flowering of *Arabidopsis*. Plant Cell. 2006;18(11):2971-84.
59. Leyva-González MA, Ibarra-Laclette E, Cruz-Ramírez A, Herrera-Estrella L. Functional and transcriptome analysis reveals an acclimatization strategy for abiotic stress tolerance mediated by *Arabidopsis* NF-YA family members. Plos One. 2012;7(10):e48138.
60. Xu MY, Zhang L, Li WW, Hu XL, Wang MB, Fan YL, et al. Stress-induced early flowering is mediated by miR169 in *Arabidopsis thaliana*. J Exp Bot. 2014;65(1):89-101.
61. Siriwardana CL, Gnesutta N, Kumimoto RW, Jones DS, Myers ZA, Mantovani R, et al. NUCLEAR FACTOR Y, subunit A (NF-YA) proteins positively regulate flowering and act through *FLOWERING LOCUS T*. Plos Genet. 2016;12(12):e1006496.
62. Yan Y, Shen L, Chen Y, Bao S, Thong Z, Yu H. A MYB-domain protein EFM mediates flowering responses to environmental cues in *Arabidopsis*. Dev Cell. 2014;30(4):437-48.
63. Jung JH, Lee S, Yun J, Lee M, Park CM. The miR172 target TOE3 represses *AGAMOUS* expression during *Arabidopsis* floral patterning. Plant Sci. 2014;215-216:29-38.
64. Schiessl S. Regulation and subfunctionalization of flowering time genes in the allotetraploid oil crop *Brassica napus*. Front Plant Sci. 2020;11:605155.
65. Tudor EH, Jones DM, He Z, Bancroft I, Trick M, Wells R, et al. QTL-seq identifies *BnaFT.A02* and *BnaFLC.A02* as candidates for variation in vernalization requirement and response in winter oilseed rape (*Brassica napus*). Plant Biotechnol J. 2020;18(12):2466-81.
66. Torti S, Fornara F, Vincent C, Andrés F, Nordström K, Göbel U, et al. Analysis of the *Arabidopsis* shoot meristem transcriptome during floral transition identifies distinct regulatory patterns and a leucine-rich repeat protein that promotes flowering. Plant Cell. 2012;24(2):444-62.

67. Melzer S, Lens F, Gennen J, Vanneste S, Rohde A, Beeckman T. Flowering-time genes modulate meristem determinacy and growth form in *Arabidopsis thaliana*. Nat Genet. 2008;40(12):1489-92.
68. O'Maoiléidigh DS, van Driel AD, Singh A, Sang Q, Le Bec N, Vincent C, et al. Systematic analyses of the *MIR172* family members of *Arabidopsis* define their distinct roles in regulation of *APETALA2* during floral transition. PLoS Biol. 2021;19(2):e3001043.
69. Balanzà V, Martinez-Fernández I, Sato S, Yanofsky MF, Kaufmann K, Angenent GC, et al. Genetic control of meristem arrest and life span in *Arabidopsis* by a *FRUITFULL-APETALA2* pathway. Nat Commun. 2018;9(1):565.
70. Zúñiga-Mayo VM, Marsch-Martínez N, de Folter S. JAIBA, a class-II HD-ZIP transcription factor involved in the regulation of meristematic activity, and important for correct gynoecium and fruit development in Arabidopsis. Plant J. 2012;71(2):314-26.
71. Kamata N, Okada H, Komeda Y, Takahashi T. Mutations in epidermis-specific HD-ZIP IV genes affect floral organ identity in *Arabidopsis thaliana*. Plant J. 2013;75(3):430-40.
72. Wang W, Sijacic P, Xu P, Lian H, Liu Z. *Arabidopsis* TSO1 and MYB3R1 form a regulatory module to coordinate cell proliferation with differentiation in shoot and root. Proc Natl Acad Sci U S A. 2018;115(13):E3045-E54.
73. Ranftl QL, Bastakis E, Klermund C, Schwechheimer C. LLM-domain containing B-GATA factors control different aspects of cytokinin-regulated development in *Arabidopsis thaliana*. Plant Physiol. 2016;170(4):2295-311.
74. Lozano-Sotomayor P, Chávez Montes RA, Silvestre-Vañó M, Herrera-Ubaldo H, Greco R, Pablo-Villa J, et al. Altered expression of the bZIP transcription factor DRINK ME affects growth and reproductive development in *Arabidopsis thaliana*. Plant J. 2016;88(3):437-51.
75. Gaillochet C, Jamge S, van der Wal F, Angenent G, Immink R, Lohmann JU. A molecular network for functional versatility of HECATE transcription factors. Plant J. 2018;95(1):57-70.
76. Meng WJ, Cheng ZJ, Sang YL, Zhang MM, Rong XF, Wang ZW, et al. Type-B ARABIDOPSIS RESPONSE REGULATORs specify the shoot stem cell niche by dual regulation of *WUSCHEL*. Plant Cell. 2017;29(6):1357-72.
77. Azodi CB, Pardo J, VanBuren R, de Los Campos G, Shiu SH. Transcriptome-based prediction of complex traits in maize. Plant Cell. 2020;32(1):139-51.
78. O'Neill CM, Lu X, Calderwood A, Tudor EH, Robinson P, Wells R, et al. Vernalization and floral transition in autumn drive winter annual life history in oilseed rape. Curr Biol. 2019;29(24):4300-6 e2.
79. Diepenbrock W. Yield analysis of winter oilseed rape (*Brassica napus* L.): a review. Field Crop Res. 2000;67:35-49.
80. Searle I, He Y, Turck F, Vincent C, Fornara F, Kröber S, et al. The transcription factor FLC confers a flowering response to vernalization by repressing meristem competence and systemic signaling in Arabidopsis. Genes & development. 2006;20(7):898-912.
81. Turnbull C. Long-distance regulation of flowering time. J Exp Bot. 2011;62(13):4399-413.
82. Lang A. Physiology of flower initiation. Differentiation and Development Encyclopedia of Plant Physiology. 15. Berlin, Heidelberg: Springer; 1965. p. 1380–536.
83. Klepikova AV, Kasianov AS, Gerasimov ES, Logacheva MD, Penin AA. A high resolution map of the *Arabidopsis thaliana* developmental transcriptome based on RNA-seq profiling. Plant J. 2016;88(6):1058-70.

84. Mendham NJ, Scott RK. Limiting effect of plant size at inflorescence initiation on subsequent growth and yield of oilseed rape (*Brassica napus*). J Agr Sci. 1975;84(Jun):487-502.
85. Gan YT, Stobbe EH, Moes J. Relative date of wheat seedling emergence and its impact on grain yield. Crop Sci. 1992;32(5):1275-81.
86. Liu T, Chen W, Li F, Wu W, Sun C, Ding J, et al. Characterization of the 3D structure of a cultivated land surface and its influence on wheat seedlings growth using Kinect. Sci Rep. 2017;7(1):3927.
87. Soltani A, Robertson MJ, Torabi B, Yousefi-Daz M, Sarparast R. Modelling seedling emergence in chickpea as influenced by temperature and sowing depth. Agr Forest Meteorol. 2006;138(1-4):156-67.
88. Forcella F, Benech Arnold, R.L., Sanchez, R., Ghersa, C.M. Modeling seedling emergence. Field Crop Res. 2000;67(2):123-39.
89. Varma V, Iyengar SB, Sankaran M. Effects of nutrient addition and soil drainage on germination of N-fixing and non-N-fixing tropical dry forest tree species. Plant Ecol. 2016;217(8):1043-54.
90. Chalhoub B, Denoeud F, Liu SY, Parkin IAP, Tang HB, Wang XY, et al. Early allopolyploid evolution in the post-Neolithic *Brassica napus* oilseed genome. Science. 2014;345(6199):950-3.
91. Yin X, Goudriaan J, Lantinga EA, Vos J, Spiertz HJ. A flexible sigmoid function of determinate growth. Ann Bot. 2003;91(3):361-71.
92. Habekotté B. A model of the phenological development of winter oilseed rape (*Brassica napus* L.). Field Crop Res. 1997;54(2-3):127-36.
93. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. arXiv. 2015;doi:10.48550/arXiv.1505.04597.
94. Dutta A, Zisserman A. The VIA annotation software for images, audio and video. Proceedings of the 27th ACM International Conference on Multimedia. 2019;doi:10.1145/3343031.3350535.
95. Kingma DP, Ba J. Adam: a method for stochastic optimization. arXiv. 2014;doi:10.48550/arXiv.1412.6980.
96. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: an imperative style, high-performance deep learning library. In: H. Wallach HL, A. Beygelzimer, F. d'Alché-Buc, E. Fox and R. Garnett, editor. Advances in Neural Information Processing Systems 32: Curran Associates, Inc.; 2019. p. 8024-35.
97. Schneider CA, Rasband WS, Eliceiri KW. NIH Image to ImageJ: 25 years of image analysis. Nat Methods. 2012;9(7):671-5.
98. Huang G, Liu Z, Weinberger KQ. Densely connected convolutional networks. arXiv. 2016;doi:10.48550/arXiv.1608.06993.
99. van der Walt S, Schönberger JL, Nunez-Iglesias J, Boulogne F, Warner JD, Yager N, et al. scikit-image: image processing in Python. PeerJ. 2014;2:e453.
100. Bivens NJ, Zhou M. RNA-Seq library construction methods for transcriptome analysis. Curr Protoc Plant Biol. 2016;1(1):197-215.
101. Goecks J, Nekrutenko A, Taylor J. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol. 2010;11(8):R86.
102. Andrews S. FastQC: a quality control tool for high throughput sequence data. http://wwwbioinformaticsbabrahamacuk/projects/fastqc. 2010.

103. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114-20.
104. Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nat Protoc. 2016;11(9):1650-67.
105. Anders S, Pyl PT, Huber W. HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015;31(2):166-9.
106. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550.
107. Tello D, Gil J, Loaiza CD, Riascos JJ, Cardozo N, Duitama J. NGSEP3: accurate variant calling across species and sequencing protocols. Bioinformatics. 2019;35(22):4716-23.
108. Browning BL, Zhou Y, Browning SR. A one-penny imputed genome from next-generation reference panels. Am J Hum Genet. 2018;103(3):338-48.
109. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics. 2007;23(19):2633-5.
110. Rambaut A. FigTree. http://treebioedacuk/software/figtree/. 2016.
111. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Methodol. 1995;57(1):289-300.
112. Rey SJ, Anselin L. PySAL: A Python library of spatial analytical methods. Handbook of applied spatial analysis: Springer; 2010. p. 175-93.
113. Stacklies W, Redestig H, Scholz M, Walther D, Selbig J. pcaMethods--a bioconductor package providing PCA methods for incomplete data. Bioinformatics. 2007;23(9):1164-7.
114. Öner M, Kocakoç ID. JMASM 49: A compilation of some popular goodness of fit tests for normal distribution: their algorithms and MATLAB codes (MATLAB). J Mod Appl Stat Meth. 2017;16(2):547-75.
115. Brennecke P, Anders S, Kim JK, Kolodziejczyk AA, Zhang X, Proserpio V, et al. Accounting for technical noise in single-cell RNA-seq experiments. Nat Methods. 2013;10(11):1093-5.
116. Andrews TS, Hemberg M. M3Drop: dropout-based feature selection for scRNASeq. Bioinformatics. 2019;35(16):2865-7.
117. Bucchini F, Del Cortona A, Kreft Ł, Botzki A, Van Bel M, Vandepoele K. TRAPID 2.0: a web application for taxonomic and functional analysis of de novo transcriptomes. Nucleic Acids Res. 2021;49(17):e101.
118. Pertea G, Pertea M. GFF Utilities: GffRead and GffCompare [version 2; peer review: 3 approved]. F1000Research. 2020;9:304.
119. Van Bel M, Diels T, Vancaester E, Kreft Ł, Botzki A, Van de Peer Y, et al. PLAZA 4.0: an integrative resource for functional, evolutionary and comparative plant genomics. Nucleic Acids Res. 2018;46(D1):D1190-D6.
120. Maere S, Heymans K, Kuiper M. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics. 2005;21(16):3448-9.
121. Sun F, Fan G, Hu Q, Zhou Y, Guan M, Tong C, et al. The high-quality genome of *Brassica napus* cultivar 'ZS11' reveals the introgression history in semi-winter morphotype. Plant J. 2017;92(3):452-68.

122. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 2019;20(1):238.
123. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015;12(1):59-60.
124. Proost S, Fostier J, De Witte D, Dhoedt B, Demeester P, Van de Peer Y, et al. i-ADHoRe 3.0--fast and sensitive detection of genomic homology in extremely large data sets. Nucleic Acids Res. 2012;40(2):e11.
125. Pinheiro JC, Bates DM. Linear mixed-effects models: basic concepts and examples. Mixed-effects models in S and S-Plus. New York, NY: Springer; 2000. p. 3-56.
126. Pinheiro J, Bates D, DebRoy S, Sarkar D, R Core Team. nlme: Linear and nonlinear mixed effects models. 2019.
127. Breiman L. Random forests. Mach Learn. 2001;45(1):5-32.
128. Zou H, Hastie T. Regularization and variable selection via the elastic net. J Roy Stat Soc B. 2005;67:301-20.
129. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825-30.
130. Yamada M, Jitkrittum W, Sigal L, Xing EP, Sugiyama M. High-dimensional feature selection by feature-wise kernelized Lasso. Neural Comput. 2014;26(1):185-207.
131. Climente-González H, Azencott CA, Kaski S, Yamada M. Block HSIC Lasso: model-free biomarker detection for ultra-high dimensional data. Bioinformatics. 2019;35(14):i427-i35.