Transcriptome-based prediction of complex traits in maize

Christina B. Azodi[1], Jeremy Pardo[1,3], Robert VanBuren[2,3], Gustavo de los Campos[4], and Shin-Han Shiu[1,5*]

[1] Department of Plant Biology, Michigan State University, East Lansing, MI, USA
[2] Department of Horticulture, Michigan State University, East Lansing, MI, 48824, USA
[3] Plant Resilience Institute, Michigan State University, East Lansing, MI, 48824, USA
[4] Epidemiology and Biostatistics and Statistics and Probability Departments, Michigan State University, East Lansing, MI 48824, USA
[5] Department of Computational, Mathematics, Science, and Engineering, Michigan State University, East Lansing, MI, USA

*Corresponding Author:
Shin-Han Shiu
Michigan State University
Plant Biology Laboratories
612 Wilson Road, Room 166
East Lansing, MI 48824-1312
517-353-7196
shius@msu.edu

Author Contributions: C.B.A., J.P., and S.-H.S. conceived and designed the study. J.P. assembled the data. C.B.A. wrote modeling code and the main manuscript. C.B.A. and J.P. ran the models. All authors assisted with interpretation of results and manuscript editing.

Abstract

The ability to predict traits from genome-wide sequence information (Genomic Prediction, GP), has improved our understanding of the genetic basis of complex traits and transformed breeding practices. Transcriptome data may also be useful for GP. However, it remains unclear how well transcript levels can predict traits, particularly when traits are scored at different development stages. Using maize genetic markers and transcript levels from seedlings to predict mature plant traits, we found transcript and genetic marker models have similar performance. Surprisingly, genetic markers important for predictions were not close to or identified as regulatory variants for important transcripts. Thus, transcript levels are predictive not simply due to genetic variation. Furthermore, genetic marker models identified only one of 14 benchmark flowering time genes, while transcript models identified five. Our findings highlight that transcriptome data is useful for GP and can provide a link between traits and variation that cannot be readily captured at the sequence level.

Introduction

The prediction of complex traits from genetic data is a grand challenge in biology and the outcome of such prediction has become increasingly useful for plant and animal breeding [1,2]. Among the different approaches for connecting genotypes to phenotypes, Genomic Prediction (GP, or Genomic Selection) using all available markers was developed to overcome the limitations of Marker-Assisted Selection, which uses only significant quantitative trait loci (QTLs), for breeding traits that are controlled by many small effect alleles [3,4]. Using GP, breeders are able to make data driven decision about what lines to include in their programs, speeding up and reducing the cost of developing the next generation of crops [5,6]. Furthermore, because GP models are associating genetic signatures with phenotypes, untangling GP models has the potential to improve our understanding of the genetic basis of complex traits. However, as with related approaches such as genome wide association studies and QTL mapping, it remains difficult to go from associated genetic markers to the molecular basis for a trait [7,8].

There are a number of factors contributing to this difficulty. The variation in markers associated with phenotypes may not be the causal variants but are linked to the genes that control the trait in question. Considering that linkage disequilibrium distance can range from 1 kilobase

58 (kb) in diverse maize populations [9] to ~250 kb in *Arabidopsis thaliana* [10], the linked candidate
59 genes can range from a few to a few hundreds. Even if the associated genetic variant is
60 controlling the underlying phenotype, most variants associated with complex traits have small
61 effect sizes and can be regulatory [11], which may not be linked to the genes they regulate.
62 Furthermore, multiple regulatory variants that have indiscernible effects on their own, could
63 interact epistatically to influence gene and ultimately trait expression. However, even with
64 sufficient statistical power to detect genetic variants with small effect sizes and interactions
65 between them, genetic information is connected to traits through multiple intermediate processes,
66 including, for example, transcription, translation, epigenetic modification, and metabolism. Each
67 of these intermediate processes represent an additional level of complexity that obscures the
68 association between genetic information and a trait.
69 One solution is to account for these intermediate processes by integrating relevant omics
70 data in addition to genetic variation. This approach has led to promising, but often mixed, results
71 in plants. Current efforts have focused primarily on predicting hybrid performance using
72 transcriptional information from the parental lines. For example, transcript level-based distance
73 measures generated from transcripts associated with the trait were better than genetic markers in
74 predicting hybrid performance in maize [12,13]. However, when all transcripts were used (instead of
75 a subset of pre-selected transcripts), model performance decreased [14]. The performance of
76 models based on transcript levels can be better or worse compared to those based on genetic
77 markers depending on the trait. For example, transcriptome data performed better for predicting
78 grain yield in hybrid maize populations, but genetic marker data performed better for predicting
79 grain dry matter content in the same population [15]. Similarly, in a maize diversity panel, GP
80 models that combined transcript and marker data only outperformed models using markers alone
81 for certain traits [16]. Finally, efforts to integrate additional omic information to predict various
82 traits in *Drosophila melanogaster* [17], and human diseases, such as breast cancer [18], and responses
83 to treatment interventions, including acute kidney rejection and response to infliximab in
84 ulcerative colitis [19,20], have demonstrated the potential usefulness of transcriptome data in the
85 field of precision medicine.
86 Overall, these efforts provide reasonable evidence that transcriptome data could be useful
87 for trait prediction. However, GP-based approaches that trained on the entire transcriptome data
88 have not been used to better understand the genetic mechanisms for a trait. In addition, it is not

89 known the degree to which transcriptomes obtained at a particular developmental stage can be
90 informative for predicting phenotypes scored at a different stage. To address these questions, we
91 used transcriptome data derived from maize whole seedling [22] to predict phenotypes (flowering
92 time, height, and grain yield) at much later developmental stages. In addition to comparing
93 prediction performance between genetic marker and transcriptome-based models, we also looked
94 at whether transcripts and genetic marker features important for the prediction models were
95 located in the same or adjacent regions. Finally, we determined how well our models were able
96 to identify a benchmark set of flowering time genes to explore the potential of using GP to better
97 understand the mechanistic basis of complex traits.
98

99 **Results and Discussion**

100 ***Relationships between transcript levels, kinship, and phenotypes among maize lines***

101 Before using the transcriptome data for GP, we first assessed properties of the
102 transcriptome data in three areas: (1) the quantity and distribution of transcript information
103 across the genome, (2) the amount of variation in transcript levels, and (3) the similarity in the
104 transcriptome profile between maize lines, with an emphasis on how these properties compared
105 to those based on the genotype data. After filtering out 16,898 transcripts that did not map to the
106 B73 reference genome or had zero or near zero variance across lines (see **Methods**), we had
107 31,238 transcripts. While the number of transcripts was <10% of the number of genetic markers
108 used in this study (332,178), the distribution of transcripts along the genome was similar to the
109 genetic marker distribution (**Fig. S1**). The $\log_2$-transformed median transcript level across lines
110 ranged from 0 to 12.4 (median=2.2) and the variance ranged from $3x10^{-30}$ to 14.5 (median=
111 0.13), highlighting that a subset of transcripts had relatively high variation in transcript levels
112 across maize lines at the seedling stage. To determine how similar transcript levels were between
113 lines, we calculated the expression Correlation (eCor) between all pairs of lines using Pearson's
114 Correlation Coefficient (PCC). The eCor values ranged from 0.84 to 0.99 (mean=0.93). As
115 expected, lines with similar transcriptome profiles were also genetically similar as there was a
116 significant correlation between eCor values with values in the kinship matrix generated from the
117 genetic marker data (Spearman's Rank $\rho = 0.27$, $p < 2.2x10^{-16}$; **Fig. 1A**). As a result, we were
118 able to find clusters of lines that had both high transcript and genetic similarities (e.g. cluster a,
119 b; **Fig. 1B, C**). However, most of the variation in eCor was not explained by kinship, which

120 explained why we identified other clusters that had similar transcriptome profiles, but were not
121 genetically similar (e.g. cluster c, **Fig. 1B, C**).
122 Because the basis of GP is to predict a phenotype from genetic data, we next asked if
123 kinship or eCor were anti-correlated with the phenotypic distances between lines (see **Methods**).
124 While both kinship (ρ = -0.03, $p < 2.2\text{x}10^{-16}$; **Fig. 1D**) and eCor (ρ = -0.08, $p < 2.2\text{x}10^{-16}$; **Fig.**
125 **1E**) were significantly, negatively correlated with the phenotype distance, the degree of
126 correlation was minor. Furthermore, the groups of lines that clustered together based on their
127 eCor (e.g. clusters a, b; **Fig. 1B,** 1C) did not have lower phenotypic distance (**Fig. 1F**). Taken
128 together, these findings suggest that transcriptome data may be similarly informative as genotype
129 data but capture difference aspect of phenotypic variation. We tested both of these interpretations
130 further in subsequent sections.
131
132 ***Predicting complex traits from transcript data***
133 To test how useful transcriptome data was for GP compared to genetic marker data, we
134 applied four approaches to predict three agronomically important traits in maize: flowering time,
135 height, and grain yield. Because no one GP algorithm always performs best [6,23], we tested two
136 linear algorithms (ridge regression Best Linear Unbiased Predictor (rrBLUP) and Bayesian-Least
137 Absolute Shrinkage and Selection Operator (BL)), one non-linear algorithm (random forest: RF),
138 and one ensemble approach (En; see **Methods**). To establish a baseline for our GP models, we
139 determined the amount of the phenotypic signal that could be predicted using population
140 structure alone, defined as the first five Principal Components from the genetic marker data.
141 Then we built models for each trait using genetic marker data (G), kinship (K) derived from G,
142 transcript levels (T), or expression correlation (eCor) derived from T (**Fig. 2**). Model
143 performance was measured using PCC between the actual and the predicted phenotypic values.
144 Across algorithms and traits, the K data resulted in models with the best predictive performance,
145 while models built using the eCor data performed the worst (**Fig. 2, Table S1**). Furthermore,
146 models built using G always outperformed models using T. Regardless, eCor and T-based
147 models were significantly better than the baseline predictions (dotted blue line, **Fig. 2**),
148 indicating transcriptome data can be informative in GP. Considering that the transcriptome is
149 from seedling, it is particularly surprising that mature plant phenotypes can be predicted.
150 Consistent with earlier findings [24,25], combining the predictions from multiple algorithms, known

151 as an ensemble approach, resulted in the best predictive models (**Fig. 2**), and is therefore used to
152 illustrate most of our findings.
153 Because the genetic marker and transcriptome data represented different types of
154 molecular information that could be associated with the traits of interest, we hypothesized that
155 their combination would be more informative and next built models that used combined data,
156 either K+T or G+T. For most combined models, adding the transcript data did not significantly
157 improve performance. The one exception was using RF to predict flowering time using G+T as
158 input (**Fig. 2**). To assess if G or T data features tend to be more informative in predicting traits,
159 we further quantified the importance score of each genetic marker and transcript feature for
160 models using G+T data. The importance score represents the impact that each feature had on
161 model performance defined according the algorithm used (see **Methods**). The importance scores
162 assigned to transcripts in the G+T models were correlated with the scores from the T-only
163 models (**Fig. S2A**), indicating that adding genetic marker features into the model did not impact
164 the relative importance of transcript features. Because RF importance measures tend to be biased
165 toward continuous features,[26] we focused on rrBLUP and BL importance scores. For all three
166 traits, the top 1,000 most important features were enriched for genetic markers relative to
167 transcript features (Odds Ratio = 0.17 ~ 0.44; all $p < 1x10^{-16}$; **Fig. S2B**; **Table S2**). However, the
168 top 20 most important features tended to be enriched for transcript relative to genetic marker
169 features (Odds Ratio = 2.66 ~ 13.0, $p = 0.087 \sim <1x10^{-16}$, **Table S3**), with transcript features
170 making up the top two most important feature in all cases (**Fig. S2B**). The consistency with
171 which transcript features were the most important for the models suggests that transcript
172 information is useful for GP. Further highlighting its usefulness, when either the 200 most
173 important transcripts or genetic markers were used to predict flowering time, models performed
174 equally well (r=0.70 ± 0.010; r=0.71 ± 0.009, respectively).
175
176 ***Comparison of the importance of transcripts verses genetic markers for model predictions***
177 Because models built using transcript features outperformed baseline models based solely
178 on population structure, we know transcriptome data contained information useful for explaining
179 phenotypic variation. However, combining both datasets does not improve the model (K+T and
180 G+T, **Fig. 2**), we hypothesized that this is because these two data types capture similar aspects of
181 phenotypic variation. To address this, we assessed the extent to which the important genetic

182 markers overlapped with or neighbored the genes where the important transcripts originated from
183 (top; **Fig. 3A**). The genic region and flanking sequences within a defined window of an
184 important transcript is referred to as the transcript regions (see **Methods**). For each trait and
185 algorithm, we compared the importance assigned to the transcript with that of the genetic marker
186 with the highest average importance in the transcript region (T:G pair).
187 Multiple window sizes were explored, and we used 2 kb (+/- 1kb from the center of a
188 gene) where the feature importance correlation between transcripts and genetic markers was
189 maximized (**Fig. S3**). Using this window size, 15,049 T:G pairs were identified. At the whole
190 genome level there appeared to be regions where both genetic markers and transcripts were
191 identified as important (**Fig. S4**). However, when we look closer, those regions mostly do not
192 overlap. In some cases, the important genetic markers and transcripts were in linkage
193 disequilibrium. Using the flowering time model as an example, we found the most important
194 genetic marker was located within a gene upstream the most important transcript
195 (GRMZM2G171650: *MADS69*; arrow a, **Fig. 3B**), but the two are in linkage disequilibrium [22]. In
196 most cases, there were no important genetic markers that were located nearby to important
197 transcripts. For example, the second most important flowering time genetic marker was not
198 located near important transcript regions (arrow b, **Fig. 3B**). Similarly, the second most
199 important flowering time transcript (GRMZM5G865543) was over 0.6 Mb from an important
200 genetic marker (arrow c, **Fig. 3B**). Across all traits and algorithms, T:G pairs were only
201 moderately correlated (ρ = 0.09-0.13; **Fig. 3C, Fig. S5A**).
202 This lack of correlation is notable for the most important genetic markers and transcripts.
203 For example, across the three traits, only 4-7 T:G pairs were both in the top 1% most important
204 features from the ensemble models, and those pairs were never the top ranked genetic markers or
205 transcripts from the model (**Fig. 3B)**. These findings argue against the notion that these two data
206 types capture similar aspects of phenotypic variation as we hypothesized earlier. In light of this,
207 we hypothesized that the lack of correlation was because important transcripts tend to be
208 regulated by important *trans* factors located far beyond the transcript region. To test this, we
209 assessed the degree to which important genetic markers identified as expression QTL (eQTLs)
210 were associated with important transcripts. We identified 58,361 *cis* (62) and *trans* (58,299)
211 eQTL associated with 7,052 transcripts and defined T:eQTL pairs for each of these transcripts by
212 selecting the genetic marker within +/- 1kb of an eQTL for that transcript (i.e. eQTL region) with

213 the highest average importance. Across all traits and algorithms, the importance of transcripts
214 and eQTL in T:eQTL pairs was actually negatively correlated (ρ = -0.15 ~ -0.06; **Fig. 3C, Fig.**
215 **S5B**).
216 The lack of correlation between importance scores for T:G and T:eQTL pairs was in
217 contrast to the relatively high correlation observed in feature importance between algorithms (ρ=
218 0.31-0.98), with rrBLUP and BL importance scores being the most correlated (ρ= 0.87-0.98) and
219 the average correlation between genetic markers (ρ = 0.75) being higher than for transcripts (ρ =
220 0.55) (**Fig. S6**). Together with the findings that important genetic markers were not co-located
221 and eQTL were not associated with genes that gave rise to the important transcripts for any of
222 the three traits, these findings may suggest that transcriptome data is capturing layers of
223 information, such as epigenetic signals, that are not captured by genome sequences. However,
224 we cannot rule of the possibility that the eQTL approach is not sufficiently sensitive in
225 identifying important *trans*-factors. Further study is needed to resolve these possibilities.
226
227 ***Assessment of benchmark flowering time genes***
228 Because the genetic basis for flowering time is well studied [27–30], we identified a set of 14
229 known flowering time genes (**Table S3**) and compared the ability of genetic marker and
230 transcript-based models to reveal them as important using the T:G and T:eQTL pairs described
231 earlier. Of the 14 benchmark genes, four had corresponding genetic markers in our T:G pair data.
232 When we increased the flanking regions threshold to 20kb from the center of the transcript for
233 defining T:G pairs, corresponding genetic markers were found for five additional benchmark
234 genes. Two benchmark genes, *CCT1* and *PEBP4*, neither of which were members of a T:G pair,
235 were associated with eQTLs. To account for differences in distribution and range of importance
236 scores generated by different algorithms and numbers of features, the importance scores were
237 converted to percentiles for comparison purposes.
238 Different benchmark genes were important (>95th percentile) for models using the two
239 different data types, with one and five benchmark gene considered important by the genetic
240 marker-based and the transcript-based models, respectively (**Figure 4A**; **Table S4**). For example,
241 the genetic marker located within the *RAP2* gene, which has been shown to be associated with
242 flowering time in multiple studies [22,31], was identified as important based on genetic marker
243 (99.7th-99.9th percentile), but not transcript (59th-79th percentile) data. In contrast, *MADS69*,

244 *MADS1*, *PEBP24*, and *PEBP8* were identified as important using transcript data (95th-100th
245 percentile), but not using genetic marker data (16th-93th percentile). Furthermore, with transcript
246 data we were able to assess the importance of three genes (*ZAG6, REPB5, and PEBP2*) that were
247 not located near genetic markers or associated with eQTL. For example, there were no eQTL
248 associated with or genetic markers within the 40bp window of *ZAG6*, but *ZAG6* was identified as
249 important (98th-99.9th percentile) in the transcript-based models (**Fig. 4A**). For some of these
250 benchmark genes, the region most closely linked to trait variation could be outside the +/- 20kb
251 window. For example, as described above, the important genetic marker for *MADS69*
252 (Chr3_160559109) is ~32 kb upstream, but the two are in linkage disequilibrium [22] (see arrows
253 in **a; Fig. 3B**). Taken together, these finding further highlight the usefulness of transcript data for
254 identifying the genetic basis for variation in a trait.
255
256 ***Improving our understanding of the genetic basis of flowering time using transcriptome data***
257 An open question was why transcript-based models were able to identify five benchmark
258 flowering time genes as important that were not identified by genetic marker-based models and if
259 transcriptome data could be used to better understand the genetic basis of flowering time. To
260 understand why benchmark genes were not uniformly identified as important for flowering time
261 when using both genetic marker and transcript data, we determined the extent to which transcript
262 levels and the genetic marker allele (i.e. major or minor) were related to flowering time. As
263 expected, we observed the most significant differences in flowering time for the transcripts (**Fig.
264 5A, Fig. S7A**) and genetic markers (**Fig. 5B, Fig. S7B**) that were identified as important by our
265 models. For example, *MADS1* was important only in the transcript-based models and transcript
266 level was significantly correlated with flowering time (p = 0.0001; **Fig. 5A**). In contrast, lines
267 with the major allele for the genetic marker that paired with the *MADS1* transcript (Chr9:
268 156980141) did not flower at a significantly different time than lines with the minor allele (p =
269 0.062; **Fig. 5B**). Another example was *RAP2*, which was important only in the genetic marker-
270 based models. Lines with the major allele in *RAP2* were more likely to flower late ($p < 1x10^{-4}$),
271 but *RAP2* transcript levels did not significantly correlate with changes in flowering time (p =
272 0.33). Overall, benchmark genes were more likely to have transcript levels associated with
273 flowering time (**Fig. 5C**) than genetic marker alleles associated with flowering time (**Fig. 5D**).

274 Importantly, using the transcriptome data we were also able to understand in more detail
275 the impact of the benchmark genes on flowering time. For example, variation in transcript levels
276 of *MADS69* accounted for 16.7% of the variation in flowering time, more than any other
277 transcript, where lines with lower levels of transcription flowered later. Modulation of *MADS69*
278 expression levels has recently been patented as an approach to controlling flowering time[32].
279 Similarly, *MADS1* transcript levels explained 3.7% of the variation in flowering time, with lines
280 with lower levels of transcription flowering later. This is consistent with what has been observed
281 experimentally, where down-regulation of *MADS1* results in delayed flowering time [33]. For
282 medium confidence benchmark genes (i.e. identified through association studies), the specific
283 roles of the genes on flowering time are not well understood, but by finding positive or negative
284 correlations between transcript levels and the underlying phenotypes, more mechanistic details
285 can be interred. For example, transcript levels of *ZAG6* had the second largest impact on
286 flowering time, accounting for 6% of variation, with increased transcript levels associated with
287 earlier flowering. Another example is *PEBP24*, with transcript levels of *PEBP24* accounting for
288 2.7% of the variation in flowering time. Unlike many of the other benchmark genes, increased
289 *PEBP24* transcript levels were associated with later flowering time. Overall, the identification of
290 these medium confidence benchmark genes as important transcript indicates the relevance of
291 transcriptional regulation in their flowering time functions.
292 While using the benchmark genes allowed us to assess the usefulness of transcript levels
293 compared to genetic marker information for identifying genes involved in flowering time, we
294 should note that many non-benchmark genes were also identified by our models as important.
295 For example, from the Ensemble model, there were 154 important, non-benchmark transcripts
296 with importance scores falling between the two most important benchmark genes (*MADS69*,
297 100th percentile; *ZAG6*, 99.5^{th} percentile; yellow, **Table S5**). While seven of those in between
298 transcripts were annotated with the Gene Ontology (GO) term "flower development"
299 (GO:0009908, green, **Table S5**), these 154 non-benchmark transcripts were not enriched for this
300 GO term (q = 1.0). In fact, neither these transcripts nor any other set of important transcripts
301 from models based on other algorithms (see **Methods**) were enriched for any GO terms.
302 Therefore, from our transcript-based GP models we have identified 147 high ranking transcripts,
303 many of which have unknown functions, that are among the most important in predicting
304 flowering time in maize but do not play known roles in this process. For example, both

305 GRMZM5G865543 and GRMZM2G023520, the second and third most important transcripts
306 respectively from the Ensemble model, are unknown genes. Note that the transcriptome data is
307 from the seedling stage. It is possible that genes of these important transcripts influence
308 biological processes in earlier stage of development that influence flowering time later. To
309 further our understanding of the genetic basis of flowering time control and the connections
310 between juvenile and adult phenotypes, these important transcripts are prime candidates for
311 future genetic studies.
312
313 **Conclusions**

314 We have generated predictive models that use genetic markers, transcripts, and their
315 combination to predict flowering time, height, and yield in a diverse maize population. While
316 models built using transcriptome data did not outperform models that used genotype data,
317 transcript-based models performed well above random expectation, and in many cases,
318 performance was similar to that of genotype-based models. We found that transcripts and genetic
319 markers from different genomic regions were identified as important for model predictions.
320 Furthermore, by assessing the relative importance of the features used to build the models, we
321 found that transcript-based models identified more known flowering time associated genes than
322 genetic marker-based models. These findings underscore the usefulness of transcript data for
323 improving our understanding of the genetic mechanisms responsible for complex traits.
324 There are four possible mechanistic explanations of why transcript levels could have a
325 similar predictive power as genetic markers. First, *cis*-regulatory variants that impact transcript
326 levels, are all more likely to be similar between closely related individuals. Therefore, the ability
327 of transcript data to predict phenotypes is simply a reflection of that dependency. However, we
328 demonstrated that the most informative transcript features for predicting maize phenotypes are
329 distinct from the most informative genetic marker features found in the transcript regions. While
330 for some important transcripts, the associated important genetic marker could be in linkage
331 disequilibrium but outside of the 2kb window used in our study (e.g. ~32 kb away in the case of
332 *MADS69*), overall as we increased the transcript region window size, the correlation between the
333 importance scores assigned to T:G pairs decreased, suggesting this is not generally the case.
334 Thus, the second explanation is that there are *trans*-regulatory variants, e.g. due to transposon
335 polymorphisms or transcriptional regulators, that play a major role. However, we found that the

336 importance of eQTLs (99.9% *trans*) and their associated transcripts were not positively
337 correlated, suggesting that the *trans*-regulatory variation we identified cannot explain why
338 transcript variation is predictive of phenotypic variation either. However, considering the
339 challenges in identifying eQTLs due to mixed tissues used [34] and in modeling epistatic
340 interactions [35], we cannot conclusively rule of this possibility. The third explanation is that
341 transcription is a molecular phenotype caused by the integration of multiple genetic marker
342 signals, both *cis* and *trans*, that may not have had strong signals individually. The fourth
343 explanation is that there are epigenetic variants contributing to expression variation. It remains to
344 be determined what the contribution of epigenetic variation is on our ability to use transcript data
345 to predict phenotypes.

346 One surprise is that the transcript data generated during the V1 seedling stage on whole
347 seedlings can predict adult plant phenotypes. We reason that complex traits, such as flowering
348 time, are influenced by more than just canonical genes that act immediately prior to the growth
349 and developmental sequences leading to flowering. For example, early developmental events
350 such as cotyledon damage [36], root restriction [37], and photoperiod and temperature changes [38] can
351 impact flowering time in mature plants. Therefore, early development transcript differences
352 could eventually result in different flowering time. However, we anticipate that if transcript data
353 collection occurred temporally and/or spatially closer to the phenotype data the predictive power
354 of transcript levels would increase, and likely perform better than genetic marker-based models.
355 Finally, an area of active research in GP is the incorporation of Genotype by Environment (GxE)
356 interactions into predictive models [39–41]. One potential benefit of using transcript information for
357 GP could be that GxE interactions would be picked up by transcript level signals. Because
358 transcriptome data used in our study was from whole seedlings (i.e. not the same individuals that
359 were phenotype), this could not be tested.

360 Our findings highlight an important benefit of using transcript data to better understand
361 the genetic basis of a trait. While it can be difficult to associate signals from a number of small
362 effect genetic markers or even a single large effect genetic marker back to a specific gene,
363 transcript level information is inherently associated with genes. Because of the importance of
364 regulatory variation on complex traits [11], the use of transcript information in GP could be crucial
365 for deciphering the contribution of regulatory variation to the genetic basis of traits. Therefore,
366 while we observed that in terms of predictive ability, genetic marker data outperformed transcript

367 data, expression differences are more straightforward to interpret than sequence polymorphisms.
368 In practice, this meant that transcript-based models identified five benchmark flowering time
369 genes, while genetic marker-based models only identified one and it highlighted our finding that
370 more insight into the genetic basis of complex traits can be gained when transcriptome data are
371 considered.
372

373 **Acknowledgements**
374 We thank Richard Amasino, Wolfgang Busch, and David Lowry for their help in interpreting our
375 findings. This work was partly supported by NSF Graduate Research Fellowship (Fellow ID:
376 2015196719), Graduate Research Opportunities Abroad (GROW) Fellowship to C.B.A.; the U.S.
377 Department of Energy Great Lakes Bioenergy Research Center (BER DE-SC0018409) and
378 National Science foundation (IOS-1546617, DEB-1655386) to S.-H.S.
379
380

381 **Methods**

382 ***Genotypic, transcriptomic, and phenotypic data processing***

383 The phenotypic [42], and genotypic and transcriptomic [22] data used in this study were
384 generated from the pan-genome population consisting of diverse inbred maize lines. Genotype,
385 transcriptome, flowering time, height, and yield data was all available for 388 lines out of the
386 503 maize pan-genome panel and were used for the study (**Table S6**). Genetic marker scores
387 derived from RNA-seq reads were converted to a [-1,0,1] format corresponding to [aa, Aa, AA]
388 with the more common allele (AA) designated as 1. The genetic marker positions were converted
389 from maize B73 reference genome A Golden Path v2 (AGPv2) to AGPv4.37. The AGPv2
390 genetic markers that did not map to AGPv4.37 and genetic markers with a minor allele frequency
391 less than 5% were removed, resulting in 332,178 genetic markers.
392 Transcriptomic data from whole-seedling tissue including root at the V1 stage from [22]
393 was processed to remove loci that did not map to AGPv4.37. The remaining maize B73 genes
394 were filtered with default settings of the nearZeroVar function from the R caret package to
395 remove genes with zero or near zero variance (> 95% of the lines sharing the same transcript
396 level) across lines. After the filtering steps, transcript counts for 31,238 genes were retained in
397 the final dataset. The raw transcripts per million count data were transformed with a $\log_e + 1$

398 transformation before the data were used in subsequent analyses. To assess if transcriptome data
399 had predictive power beyond random expectation, transcriptome data was permuted by gene, so
400 each gene had the same distribution of transcript values, but the values were randomly assigned
401 to different maize lines for building the transcriptome shuffled models. To compared important
402 transcripts and genetic markers from GP models, transcripts were converted from AGPv3 to v4,
403 only genes with one to one correspondence between AGPv3 and v4 were included in this
404 analysis.
405
406 ***Comparison of transcript and genetic marker data***
407 Three different approaches were used to determine the similarity between lines based on
408 the three different data types. For the genotype data, a kinship matrix was generated using the
409 centered Identity By State (IBS) method [43] implemented in TASSEL v5.20180517 [44]. For the
410 transcript data, we generated an expression Correlation (eCor) matrix by calculating the Pearson
411 Correlation Coefficients (PCCs) of transcript values between lines using the cor.test function in
412 R. The eCor matrix was normalized between 0 and 1 and the diagonal was set as 1. Finally, for
413 phenotype data, we calculated the Euclidean distance between lines using the distances package
414 in the R environment. The correlation between kinship, eCor, and Phenotype Distance between
415 pairs of lines was calculated using PCC.
416
417 ***Genomic prediction models and model performance***
418 Because part of the phenotypic signal observed in GP models may be due to population
419 structure within the breeding population, we established a baseline for our GP models by using
420 the first 5 principle components generated using the marker data alone, to predict phenotype
421 values. Four methods were used for each trait, two linear-parametric methods: ridge regression-
422 Best Linear Unbiased Predictor (rrBLUP)[45] and Bayesian Least absolute shrinkage and selection
423 operator (BL)[46], and one non-linear and non-parametric method: Random Forest (RF)[47], and one
424 ensemble based approach (En)[48]. Both rrBLUP and BL were implemented in R using the
425 "rrBLUP" and "BGLR" packages respectively. RF was implemented in python using Scikit-
426 Learn [49]. Ensemble predictions were generated by taking the mean of the predicted trait values
427 from rrBLUP, BL, and RF. A grid-search was performed on the first 10 of the 100 cross-
428 validation replicates to find the best combination of parameters for the RF model. Parameters

429 tested included max tree depth (3, 5, 10, and 50) and the max number of features included in
430 each tree (10%, 50%, 100%, square root, and $\log_2$).
431 The predictive performance of the models was compared using the PCC. The PCC
432 between the predicted ($\hat{Y}$) and the true trait value (Y) and was computed using the cor() function
433 in R for rrBLUP and BL or the NumPy corrcoef function in Python for RF. One hundred
434 replicates of a five-fold cross validation approach were applied to maximize the data available
435 for model training without resulting in overfitting. For each replicate, the lines were randomly
436 divided into 5 subsets, where each subset is used as the testing set once and the rest 4 subsets
437 combined to train the model, resulting in a total of 500 cross-validated runs. PCC was calculated
438 using only the predicted values from the testing set for each run.
439
440 ***Genetic marker/transcript importance analysis***
441 In order to identify features important for building the predictive models, feature
442 importance information was extracted from each model established with one of four methods:
443 rrBLUP, BL, RF, and Ensemble. For rrBLUP, the importance metric was the marker effect ($u)
444 calculated by mixed.solve in the R rrBLUP package. For BL, the importance metric was the
445 estimated posterior mean ($ETA) calculated using the R BGLR package. The absolute value of
446 marker effect and estimated posterior mean were used since the features are categorical with no
447 particular meaning for the sign of importance metrics. For RF, the importance metric was the
448 Gini importance, collected using the _importance_score function built into the Scikit-Learn
449 implementation of RF. The Gini importance is the total decrease in node impurity (i.e. the
450 homogeneity of classes in a node) after a particular feature is used to split a node. Node impurity
451 decreases as instances from one of the classes are removed from the node, leaving a greater
452 proportion of instances from the other class. Importance metrics from rrBLUP, BL, and RF were
453 averaged over the 100 cross-validation replicates. Ensemble importance scores were calculated
454 by normalizing the average importance scores from each model and each method between 0 and
455 1, then taking the mean of normalized importance scores across the three algorithms. Enrichment
456 for transcript compared to genetic marker features within the top 1000 or top 20 features was
457 done using Fisher's Exact Test, where the number of transcript features in and not in the top X
458 features was compared to the number of genetic marker features in and not in the top X features.

459 To determine the degree to which the importance of a transcript correlates with the
460 importance of nearby genetic markers, the genetic marker *G* with the greatest mean importance
461 score within a fixed window from the center of a genomic region *R* where a transcript *T* mapped
462 to was selected among genetic markers in region *R,* referred to as a T:G pair (**Fig. 3A**). To
463 identify the effect of window size, a series of window sizes ranging from 1-40kb were tested. For
464 each window size, the Spearman's Correlation (ρ) was calculated between the importance scores
465 of T:G pairs. The window size with the highest correlation (2kb) was chosen (**Figure S4**). For
466 this analysis, transcripts without location information or without one-to-one mapping between
467 AGP V3 to V4 were removed, leaving 24,412 transcripts. With a window size of 2kb, additional
468 transcripts were dropped because there was not a genetic marker within that window, resulting in
469 15,049 transcripts to be included in the downstream analysis.
470 To determine the degree to which the importance of a transcript correlated with the
471 importance of *trans*-regulatory variants, significant eQTLs (multiple testing corrected $p<0.05$)
472 were identified for each transcript using the linear regression (modelLINEAR) approach from
473 MatrixeQTL implemented in R. Benjamini-Hochberg false discovery rate correction was used to
474 adjust p for multiple testing and eQTLs were considered significant if adjusted $p < 0.05$. The
475 distance for considering eQTL as *cis* was 1 mega base [50], however, because <0.1% of eQTL
476 identified were *cis*, all eQTL were analyzed together. The importance of an eQTL or the
477 neighboring genetic marker located within a 2kb window of the eQTL with the greatest average
478 importance score was compared to the importance of the transcript with the eQTL in question
479 (T:eQTL pair).
480 Enrichment of Gene Ontology (GO) terms associated with important transcripts
481 compared to the reference genome was tested using agriGO v2 [51]. The enrichment *p*-values are
482 corrected for multiple testing by agriGOv2 using FDR. The top 10, 25, and 100 transcripts from
483 each algorithm, excluding the benchmark flowering time genes, were tested against the reference
484 genome. Additionally, the top 153 transcripts, excluding benchmark genes, from the ensemble
485 algorithm and the union of the top 10, 25, and 100 transcripts from all four algorithms were
486 tested.
487

488 ***Benchmark flowering time genes***

489 We compiled a list of genes known to be involved in flowering time based on evidence
490 from knockdown experiments [27–30,33] and/or association study [22,52]. Genes were assigned
491 confidence levels based on the type of evidence available, with experimental evidence
492 considered high confidence, association study evidence and significant similarity with known
493 flowering time genes from other species considered medium confidence (**Table S8**). Because
494 some of these genes did not have genetic markers located within the 2kb window of the center of
495 the transcript, progressively larger windows were used to identify the most important nearby
496 genetic marker up to 40kb. To compared importance scores across algorithms and between
497 models using G or T data as input, percentiles were used. To determine if transcripts or genetic
498 markers assigned to flowering time benchmark genes were associated with flowering time in this
499 study, linear models and t-tests, respectively, implemented in R were used.
500

501 ***Data Availability***

502 All data and code needed to reproduce the results from this study is available on GitHub
503 including genomic, transcriptomic, and phenotype data
504 (https://github.com/ShiuLab/Manuscript_Code/tree/master/2019_expression_GP/data), codes to
505 run rrBLUP and BL models (https://github.com/ShiuLab/GenomicSelection), codes to run RF
506 models (https://github.com/ShiuLab/ML-Pipeline), as well as R code used for preprocessing,
507 T:G/eQTL pairing, eQTL analysis, and additional statistical analyses
508 (https://github.com/ShiuLab/Manuscript_Code/tree/master/2019_expression_GP/scripts).
509

510 **Figures**

511 ***Figure 1. Relationship between lines from transcript and genetic marker data***

512 **(A)** Relationship between kinship based on genetic marker data (X-axis) and expression
513 correlation (eCor, in Pearson's Correlation Coefficient (PCC)) based on transcript data (Y-axis).
514 Boxplots show the median Y-axis value for each X-axis bin (bin size=0.15) with the 5th (blue)
515 and 95th (red) percentile range shown. The correlation between kinship and eCor was calculated
516 using Spearman's Rank Coefficient (ρ). **(B, C)** The relationships between lines based on eCor
517 **(B)** or kinship **(C)** for all pairs of maize lines. Lines are sorted based on hierarchical clustering
518 results using the eCor values. The blue, white, and red color scales indicate negative, no, or

519 positive correlations, respectively. Dotted rectangles: indicating cluster of lines discussed in the
520 main text. **(D, E)** The relationships between the Euclidian distance calculated with phenotype
521 values (Phenotype Distance: Y-axis) and kinship **(D)**, and eCor **(E)**. Colored line: follow those in
522 **(A). (F)** The relationships between lines based on Phenotype Distance, where the lines were
523 sorted as in **(B)**. Red: smaller distance (more similar). Blue: greater distances (less similar).
524
525 ***Figure 2. Genomic prediction model performance***
526 PCCs between predicted and true values for three traits and four algorithms using six different
527 input features. The darkest red indicate a normalized PCC of 1 (the algorithm/input feature
528 combination performed the best for the trait), while the darkest blue has a normalized PCC of 0
529 (performed the worst. Original PCC values were shown in the boxes with the top performing
530 model(s) in white. Right violin-plots show the PCC distributions among different input features
531 for each algorithm (right). The median PCCs are indicated with black bars. The model
532 performance PCCs based on only population structure are indicated with a blue dashed line.
533 Bottom violin-plots show the PCC distributions among different algorithms for each input
534 feature. rrB: ridge regression Best Linear Unbiased Predictor. BL: Bayesian Least Absolute
535 Shrinkage and Selection Operator. RF: Random Forest. En: Ensemble.
536
537 ***Figure 3. Correlation between genetic marker and transcript importance for flowering time***
538 **(A)** Illustration of how transcript (T):genetic marker (G) (top graph) and T:expression
539 Quantitative Trait Locus (eQTL) (bottom graph) pairs were determined. Genetic marker
540 importance percentiles are shown above the genetic markers (red triangle) and eQTL (yellow
541 triangle). A T:G pair was defined as the transcript and the most important genetic marker within
542 the transcript region (top graph). A T:eQTL pair was defined as the transcript and the most
543 important genetic marker within the eQTL region (bottom graph). **(B)** Manhattan plots of the
544 transcript (blue bar) and genetic marker (red dot) importance scores ($-\log_e$(1-importance
545 percentile)) in a 2Mb window surrounding top two genetic markers (top and middle plots) and
546 transcripts (top and bottom plots) based on the Ensemble models for predicting flowering time.
547 All genetic markers (i.e. not just the T:G pair) are shown. The threshold (gray dotted line) is set
548 at the 99th percentile importance. **(C)** Density scatter plot of the importance scores (see Methods)
549 of the genetic marker (Y-axis) and transcript (X-axis) for T:G pairs (top graphs) and of the eQTL

550 genetic marker (Y-axis) and transcript (X-axis) for the T:eQTL pairs (bottom graphs) for three
551 traits. The threshold (black dotted line) was set at the 99th percentile importance score for each
552 trait and input feature type. The correlation between importance scores between transcript and
553 genetic marker/eQTL pairs was calculated using Spearman's rank (ρ).
554
555 ***Figure 4. Comparison of transcript and genetic marker importance scores for benchmark***
556 ***flowering time genes***
557 Importance percentile of each transcript (left) and genetic marker (right) pair as determined by
558 each of the 4 algorithms (X-axis). Genes are sorted based on hierarchical clustering of their
559 importance percentiles. Gray boxes designate benchmark genes that did not have genetic markers
560 within a 40kb window. Confidence levels (high or medium) were assigned based on the type of
561 evidence available for the benchmark gene (see **Methods**). Algorithms were abbreviated as in
562 **Figure 2**.
563
564 ***Figure 5. Relationship between transcript level/allele type and flowering time for benchmark***
565 ***genes***
566 **(A)** Boxplots show the transcript levels ($\log_e$(Fold-Change)) over flowering time bin with the 5th
567 (blue) and 95th (red) percentile range shown. Flowering time was defined as the growing degree
568 days/100. Linear models were fit and adjusted r^2 and p-values are shown. Confidence levels of
569 benchmark genes were designated as in **(4)**. **(B)** Distributions of flowering time for lines with the
570 major (red) or minor (gray) alleles for the genetic marker paired with each benchmark gene as
571 indicated in **(A)**. Differences in flowering time by allele were tested using t-tests. **(C)** Number of
572 transcripts (Y-axis) for which transcript levels were associated with flowering time in linear
573 models within *p*-value bins (-$\log_{10}$(*p*-value); X-axis). Benchmark genes are labeled as in **(A)**. **(D)**
574 Number of genetic markers (Y-axis) for which differences in flowering time by allele from t-
575 tests were within p-value bins (-$\log_{10}$(*p*-value); X-axis). Benchmark genes are labeled as in **(A)**.
576

577 **Supplemental Figures**

578 ***Figure S1. Distribution of genetic marker and transcript data across maize chromosomes***

579 Number of genetic markers (top) and transcripts (bottom) included in this study in 1 Mb bins
580 across the maize chromosomes.
581

582 ***Figure S2. Feature importance analysis for G+T models***

583 **(A)** Relationships between importance scores for transcripts from the T (X-axis) and G+T (Y-
584 axis) flowering time prediction models established with rrBLUP (left column), BL (middle
585 column), and RF (right column). The Pearson's Correlation Coefficient (r) is shown in the top
586 left corner. **(B)** Distribution of importance scores for the top 1,000 (inset = top 20) features from
587 the G+T models for three traits using rrBLUP (top row) and BL (bottom row). Transcript
588 features are in purple and genetic marker features are in yellow.
589

590 ***Figure S3. Impact of transcript region sizes on importance correlation between***
591 ***transcript:genetic marker pairs***

592 The correlation (green) between importance scores for transcript:genetic marker pairs and the
593 number of pairs found (blue) as the transcript region size increases. Shown here are the
594 correlation scores when using top (solid) or the 95th percentile (dashed) mean importance score
595 of genetic markers in the transcript region.
596

597 ***Figure S4. Manhattan plot of importance scores from Genomic Prediction models***

598 Manhattan plots of genetic marker (top) and transcript (bottom) importance scores for predicting
599 **(A)** flowering time, **(B)** height, and **(C)** yield. Threshold importance scores (dotted blue) were set
600 at the 99th percentile importance score for each trait, algorithm, and input feature type (i.e.
601 genetic markers or transcripts). Genetic markers and transcripts falling above that threshold
602 colored in blue.
603

604 ***Figure S5. Correlation between genetic marker/eQTL and transcript importance***

605 Density plot of the importance scores of **(A)** genetic markers (G, Y-axis) and transcripts (T, X-
606 axis) from T:G pairs and **(B)** eQTL (eQTL, Y-axis) and transcripts (T, X-axis) from T:eQTL

pairs. The threshold was set (red dotted line) as the 99th percentile of the normalized importance score for each trait, algorithm, and input feature type. The correlation between transcript and genetic marker importance was calculated using Spearman's Rank (ρ).

Figure S6. Correlation between feature importance between algorithms

Density scatter plot of the importance scores of genetic markers (top) and transcripts (bottom) generated with rrBLUP and BL (left), rrBLUP and RF (middle), as well as BL and RF (right). The correlation between importance scores between algorithms was calculated using Spearman's Rank (ρ).

Figure S7. Relationship between transcript levels and alleles and flowering time for benchmark genes

(A) Boxplots show the median transcript level (log(Fold-Change)) for each flowering time (Growing Degree Days (GDD)/100) bin with the 95th (red) and 5th (blue) percentiles shown. Linear models were fit and adjusted r^2 and p-values are shown. (B) Violin-plots of the distribution of flowering time (GDD/100) for lines with the major (blue) or minor (gray) allele for the genetic marker paired with each benchmark gene. Significant differences in the GDD by allele were tested for using t-tests.

626 **Supplemental Tables**

627 ***Table S1. Model performance by feature input type and algorithm***

628 ***Table S2. Enrichment of transcript vs. genotype features among the top most important***
629 ***features from G+T models***

630 ***Table S3. Description of benchmark flowering time genes, including evidence for flowering***
631 ***time association and T:Gs and T:eQTL pair information***

632 ***Table S4. Importance scores and percentiles for benchmark gene transcripts, and genetic***
633 ***marker and eQTL pairs***

634 ***Table S5. Top 1000 most important transcripts for flowering time from the Ensemble models.***

635 ***Table S6. Account of data (Genetic Marker, Transcript, Phenotype) availability for maize lines***
636 ***and decision to include line in the study***

637

638 **References**

639 1. Heffner, E. L., Sorrells, M. E. & Jannink, J.-L. Genomic Selection for Crop Improvement.
640 *Crop Sci.* **49**, 1–12 (2009).
641 2. Jonas, E. & de Koning, D.-J. Does genomic selection have a future in plant breeding?
642 *Trends Biotechnol.* **31**, 497–504 (2013).
643 3. Meuwissen, T. H. E., Hayes, B. J. & Goddard, M. E. Prediction of Total Genetic Value
644 Using Genome-Wide Dense Marker Maps. *Genetics* 1–11 (2001).
645 4. Ribaut, J.-M. & Ragot, M. Marker-assisted selection to improve drought adaptation in
646 maize: the backcross approach, perspectives, limitations, and alternatives. *J. Exp. Bot.* **58**,
647 351–360 (2007).
648 5. Endelman, J. B. *et al.* Optimal Design of Preliminary Yield Trials with Genome-Wide
649 Markers. *Crop Sci.* **54**, 48–59 (2014).
650 6. Spindel, J. *et al.* Genomic Selection and Association Mapping in Rice (Oryza sativa): Effect
651 of Trait Genetic Architecture, Training Population Composition, Marker Number and
652 Statistical Model on Accuracy of Rice Genomic Selection in Elite, Tropical Rice Breeding
653 Lines. *PLoS Genet.* **11**, e1004982–25 (2015).
654 7. Drinkwater, N. R. & Gould, M. N. The long path from QTL to gene. *PLoS Genet.* **8**,
655 e1002975 (2012).
656 8. Solberg Woods, L. C. QTL mapping in outbred populations: successes and challenges.
657 *Physiol. Genomics* **46**, 81–90 (2014).
658 9. Tenaillon, M. I. *et al.* Patterns of DNA sequence polymorphism along chromosome 1 of
659 maize (Zea mays ssp. mays L.). *Proc. Natl. Acad. Sci. U. S. A.* **98**, 9161–9166 (2001).
660 10. Nordborg, M. *et al.* The extent of linkage disequilibrium in Arabidopsis thaliana. *Nature*

661 *Genetics* **30**, 190–193 (2002).
662 11. Albert, F. W. & Kruglyak, L. The role of regulatory variation in complex traits and disease.
663 *Nat. Rev. Genet.* **16**, 197–212 (2015).
664 12. Frisch, M. *et al.* Transcriptome-based distance measures for grouping of germplasm and
665 prediction of hybrid performance in maize. *Theor. Appl. Genet.* **120**, 441–450 (2010).
666 13. Fu, J. *et al.* Partial least squares regression, support vector machine regression, and
667 transcriptome-based distances for prediction of maize hybrid performance with gene
668 expression data. *Theor. Appl. Genet.* **124**, 825–833 (2012).
669 14. Zenke-Philippi, C. *et al.* Prediction of hybrid performance in maize with a ridge regression
670 model employed to DNA markers and mRNA transcription profiles. *BMC Genomics* **17**,
671 262 (2016).
672 15. Schrag, T. A. *et al.* Beyond Genomic Prediction: Combining Different Types of omics Data
673 Can Improve Prediction of Hybrid Performance in Maize. *Genetics* **208**, 1373–1385 (2018).
674 16. Guo, Z., Magwire, M. M., Basten, C. J., Xu, Z. & Wang, D. Evaluation of the utility of gene
675 expression and metabolic information for genomic prediction in maize. *Theor. Appl. Genet.*
676 **129**, 2413–2427 (2016).
677 17. Li, Z., Gao, N., Martini, J. W. R. & Simianer, H. Integrating Gene Expression Data Into
678 Genomic Prediction. *Front. Genet.* **10**, 126 (2019).
679 18. González-Reymúndez, A., de Los Campos, G., Gutiérrez, L., Lunt, S. Y. & Vazquez, A. I.
680 Prediction of years of life after diagnosis of breast cancer using omics and omic-by-
681 treatment interactions. *Eur. J. Hum. Genet.* **25**, 538–544 (2017).
682 19. Kang, T., Ding, W., Zhang, L., Ziemek, D. & Zarringhalam, K. A biological network-based
683 regularized artificial neural network model for robust phenotype prediction from gene
684 expression data. *BMC Bioinformatics* **18**, 565 (2017).
685 20. Zarringhalam, K., Degras, D., Brockel, C. & Ziemek, D. Robust phenotype prediction from
686 gene expression data using differential shrinkage of co-regulated genes. *Sci. Rep.* **8**, 1237
687 (2018).
688 21. Kremling, K., Diepenbrock, C., Gore, M. & Buckler, E. Transcriptome-wide association
689 supplements genome-wide association in Zea mays. *bioRxiv* (2018).
690 22. Hirsch, C. N. *et al.* Insights into the maize pan-genome and pan-transcriptome. *Plant Cell*
691 **26**, 121–135 (2014).
692 23. Heslot, N., Yang, H.-P., Sorrells, M. E. & Jannink, J.-L. Genomic Selection in Plant
693 Breeding: A Comparison of Models. *Crop Sci.* **52**, 146–115 (2012).
694 24. Shen, H.-B. & Chou, K.-C. Ensemble classifier for protein fold pattern recognition.
695 *Bioinformatics* **22**, 1717–1722 (2006).
696 25. Jia, J., Liu, Z., Xiao, X., Liu, B. & Chou, K.-C. iPPI-Esml: An ensemble classifier for
697 identifying the interactions of proteins by incorporating their physicochemical properties
698 and wavelet transforms into PseAAC. *J. Theor. Biol.* **377**, 47–56 (2015).
699 26. Strobl, C., Boulesteix, A.-L., Zeileis, A. & Hothorn, T. Bias in random forest variable
700 importance measures: illustrations, sources and a solution. *BMC Bioinformatics* **8**, 25

701 (2007).
702 27. Muszynski, M. G. *et al.* delayed flowering1 Encodes a basic leucine zipper protein that
703 mediates floral inductive signals at the shoot apex in maize. *Plant Physiol.* **142**, 1523–1536
704 (2006).
705 28. Lazakis, C. M., Coneva, V. & Colasanti, J. ZCN8 encodes a potential orthologue of
706 Arabidopsis FT florigen that integrates both endogenous and photoperiod flowering signals
707 in maize. *J. Exp. Bot.* **62**, 4833–4842 (2011).
708 29. Danilevskaya, O. N., Meng, X. & Ananiev, E. V. Concerted modification of flowering time
709 and inflorescence architecture by ectopic expression of TFL1-like genes in maize. *Plant*
710 *Physiol.* **153**, 238–251 (2010).
711 30. Meng, X., Muszynski, M. G. & Danilevskaya, O. N. The FT-like ZCN8 Gene Functions as
712 a Floral Activator and Is Involved in Photoperiod Sensitivity in Maize. *Plant Cell* **23**, 942–
713 960 (2011).
714 31. Buckler, E. S. *et al.* The genetic architecture of maize flowering time. *Science* **325**, 714–718
715 (2009).
716 32. US Patent Application for MODIFYING FLOWERING TIME IN MAIZE Patent
717 Application (Application #20140366213 issued December 11, 2014) - Justia Patents Search.
718 Available at: https://patents.justia.com/patent/20140366213. (Accessed: 10th October 2018)
719 33. Alter, P. *et al.* Flowering Time-Regulated Genes in Maize Include the Transcription Factor
720 ZmMADS1. *Plant Physiol.* **172**, 389–404 (2016).
721 34. Wills, Q. F. *et al.* Single-cell gene expression analysis reveals genetic associations masked
722 in whole-tissue experiments. *Nat. Biotechnol.* **31**, 748–752 (2013).
723 35. Becker, J., Wendland, J. R., Haenisch, B., Nöthen, M. M. & Schumacher, J. A systematic
724 eQTL study of cis-trans epistasis in 210 HapMap individuals. *Eur. J. Hum. Genet.* **20**, 97–
725 101 (2012).
726 36. Hanley, M. E. & May, O. C. Cotyledon damage at the seedling stage affects growth and
727 flowering potential in mature plants. *New Phytol.* **169**, 243–250 (2006).
728 37. Keever, G. J., Kessler, J. R., Jr, Fain, G. B. & Mitchell, D. C. Seedling Developmental
729 Stage at Transplanting Affects Growth and Flowering of Medallion Flower and Globe
730 Amaranth. *J. Environ. Hortic.* **33**, 53–57 (2015).
731 38. Song, Y. H., Ito, S. & Imaizumi, T. Flowering time regulation: photoperiod- and
732 temperature-sensing in leaves. *Trends in Plant Science* **18**, 575–583 (2013).
733 39. Burgueño, J., de los Campos, G., Weigel, K. & Crossa, J. Genomic prediction of breeding
734 values when modeling genotype× environment interaction using pedigree and dense
735 molecular markers. *Crop Sci.* **52**, 707–719 (2012).
736 40. Cuevas, J. *et al.* Bayesian Genomic Prediction with Genotype × Environment Interaction
737 Kernel Models. *G3* **7**, 41–53 (2017).
738 41. Granato, I. *et al.* BGGE: A New Package for Genomic-Enabled Prediction Incorporating
739 Genotype × Environment Interaction Models. *G3* **8**, 3039–3047 (2018).
740 42. Hansey, C. N., Johnson, J. M., Sekhon, R. S., Kaeppler, S. M. & Leon, N. de. Genetic

741 diversity of a maize association population with restricted phenology. *Crop Sci.* **51**, 704–
742 715 (2011).
743 43. Endelman, J. B. & Jannink, J.-L. Shrinkage estimation of the realized relationship matrix.
744 *G3* **2**, 1405–1413 (2012).
745 44. Bradbury, P. J. *et al.* TASSEL: software for association mapping of complex traits in
746 diverse samples. *Bioinformatics* **23**, 2633–2635 (2007).
747 45. Endelman, J. B. Ridge regression and other kernels for genomic selection with R package
748 rrBLUP. *Plant Genome* (2011).
749 46. Pérez, P. & de los Campos, G. Genome-wide regression and prediction with the BGLR
750 statistical package. *Genetics* **198**, 483–495 (2014).
751 47. Leo Breiman Statistics, L. B. Random Forests. in *Machine Learning* (2001).
752 48. Dietterich, T. G. Ensemble methods in machine learning. *International workshop on*
753 *multiple classifier systems* (2000).
754 49. Pedregosa, F. *et al.* Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**,
755 2825–2830 (2011).
756 50. Zan, Y., Shen, X., Forsberg, S. K. G. & Carlborg, Ö. Genetic Regulation of Transcriptional
757 Variation in Natural Arabidopsis thaliana Accessions. *G3* **6**, 2319–2328 (2016).
758 51. Tian, T. *et al.* agriGO v2.0: a GO analysis toolkit for the agricultural community, 2017
759 update. *Nucleic Acids Res.* **45**, W122–W129 (2017).
760 52. Salvi, S. *et al.* Conserved noncoding genomic sequences associated with a flowering-time
761 quantitative trait locus in maize. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 11376–11381 (2007).

762 50. Salvi, S. *et al.* Conserved noncoding genomic sequences associated with a flowering-time
763 quantitative trait locus in maize. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 11376–11381 (2007).

764

Figure 1

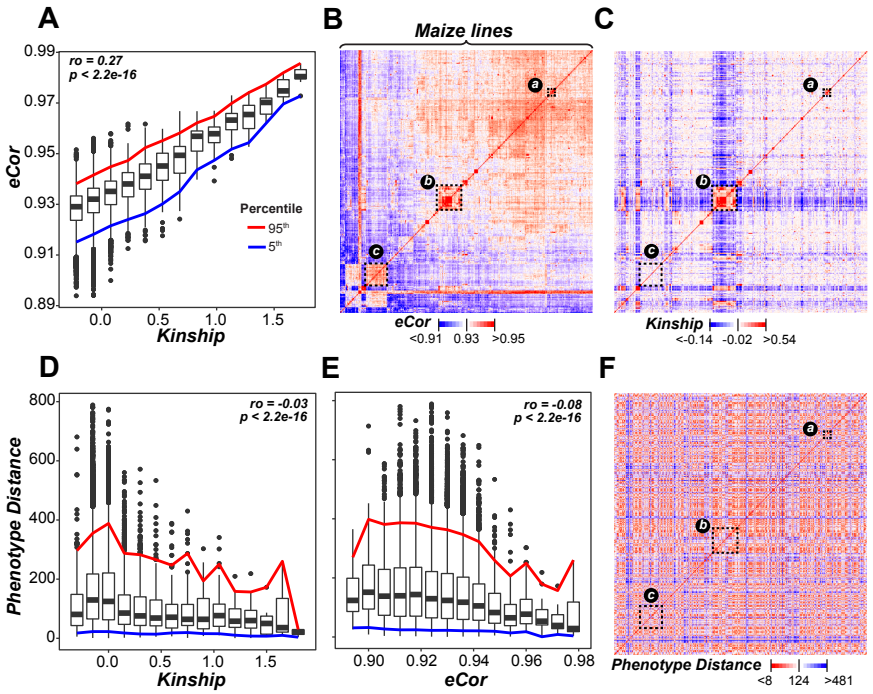

Figure 1. Relationship between lines from transcript and genetic marker data
(A) Relationship between kinship based on genetic marker data (X-axis) and expression correlation (eCor, in Pearson's Correlation Coefficient (PCC)) based on transcript data (Y-axis). Boxplots show the median Y-axis value for each X-axis bin (bin size=0.15) with the 5th (blue) and 95th (red) percentile range shown. The correlation between kinship and eCor was calculated using Spearman's Rank Coefficient (ρ). (B, C) The relationships between lines based on eCor (B) or kinship (C) for all pairs of maize lines. Lines are sorted based on hierarchical clustering results using the eCor values. The blue, white, and red color scales indicate negative, no, or positive correlations, respectively. Dotted rectangles: indicating cluster of lines discussed in the main text. (D, E) The relationships between the Euclidian distance calculated with phenotype values (Phenotype Distance: Y-axis) and kinship (D), and eCor (E). Colored line: follow those in (A). (F) The relationships between lines based on Phenotype Distance, where the lines were sorted as in (B). Red: smaller distance (more similar). Blue: greater distances (less similar).

Figure 2

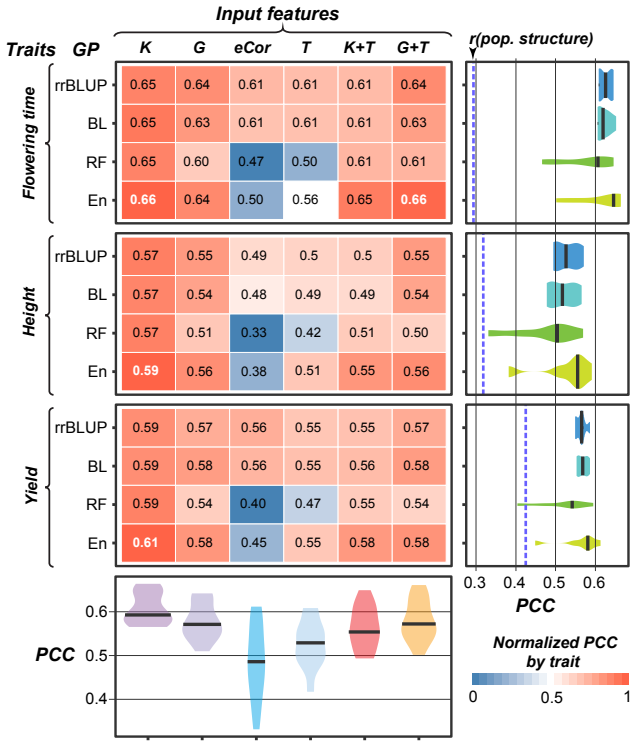

Figure 2. Genomic prediction model performance
PCCs between predicted and true values for three traits and four algorithms using six different input features. The darkest red indicate a normalized PCC of 1 (the algorithm/input feature combination performed the best for the trait), while the darkest blue has a normalized PCC of 0 (performed the worst. Original PCC values were shown in the boxes with the top performing model(s) in white. Right violin-plots show the PCC distributions among different input features for each algorithm (right). The median PCCs are indicated with black bars. The model performance PCCs based on only population structure are indicated with a blue dashed line. Bottom violin-plots show the PCC distributions among different algorithms for each input feature. rrB: ridge regression Best Linear Unbiased Predictor. BL: Bayesian Least Absolute Shrinkage and Selection Operator. RF: Random Forest. En: Ensemble.

Figure 3

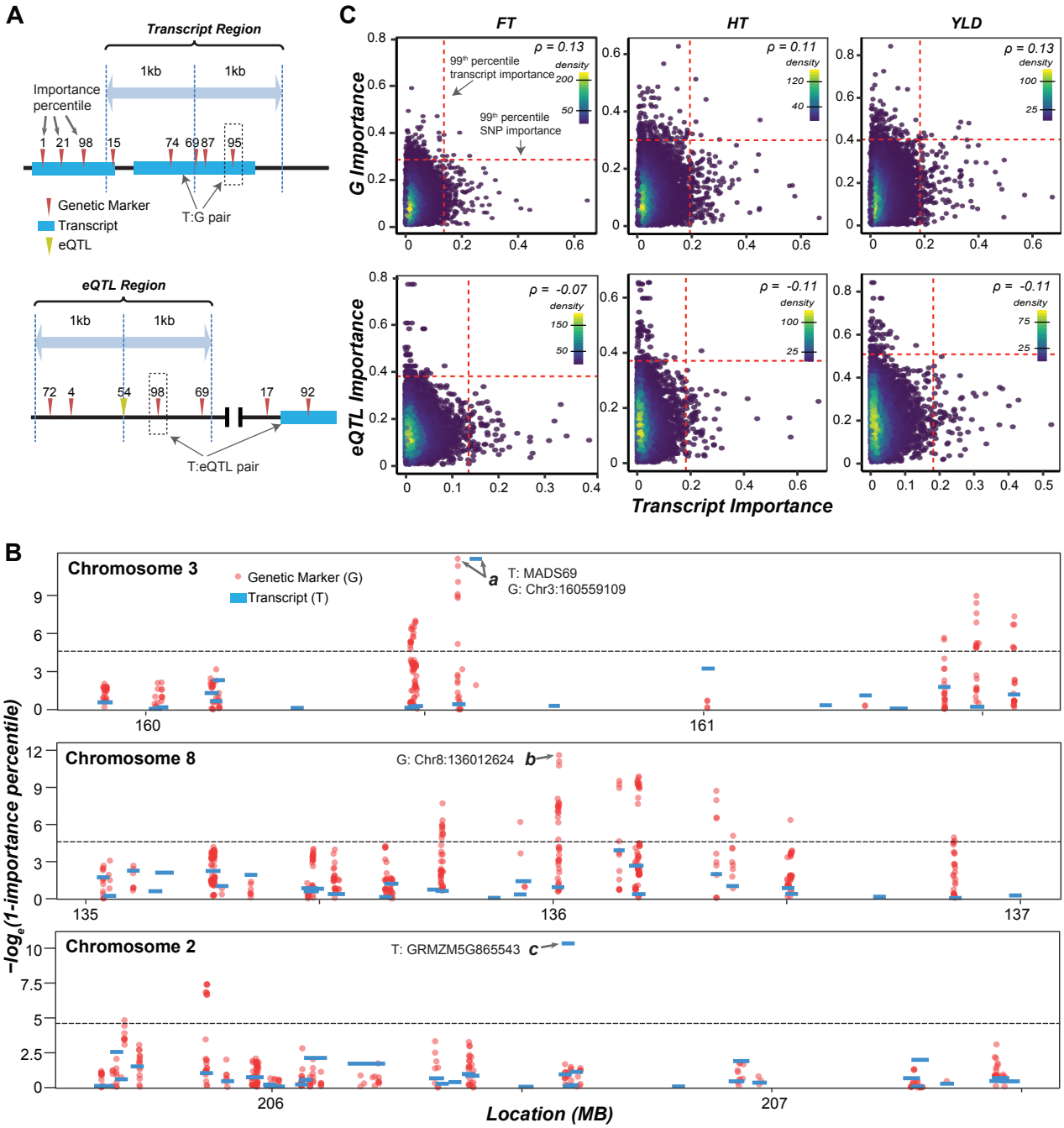

Figure 3. Correlation between genetic marker and transcript importance for flowering time

(A) Illustration of how transcript (T):genetic marker (G) (top graph) and T:expression Quantitative Trait Locus (eQTL) (bottom graph) pairs were determined. Genetic marker importance percentiles are shown above the genetic markers (red triangle) and eQTL (yellow triangle). A T:G pair was defined as the transcript and the most important genetic marker within the transcript region (top graph). A T:eQTL pair was defined as the transcript and the most important genetic marker within the eQTL region (bottom graph). (B) Manhattan plots of the transcript (blue bar) and genetic marker (red dot) importance scores (-loge(1-importance percentile)) in a 2Mb window surrounding top two genetic markers (top and middle plots) and transcripts (top and bottom plots) based on the Ensemble models for predicting flowering time. All genetic markers (i.e. not just the T:G pair) are shown. The threshold (gray dotted line) is set at the 99th percentile importance. (C) Density scatter plot of the importance scores (see Methods) of the genetic marker (Y-axis) and transcript (X-axis) for T:G pairs (top graphs) and of the eQTL genetic marker (Y-axis) and transcript (X-axis) for the T:eQTL pairs (bottom graphs) for three traits. The threshold (black dotted line) was set at the 99th percentile importance score for each trait and input feature type. The correlation between importance scores between transcript and genetic marker/eQTL pairs was calculated using Spearman's rank (ρ).

Figure 4

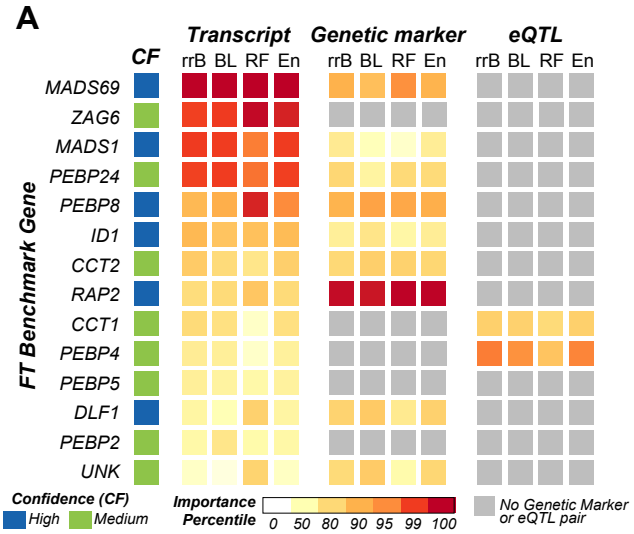

Figure 4. Comparison of transcript and genetic marker importance scores for benchmark flowering time genes
Importance percentile of each transcript (left) and genetic marker (right) pair as determined by each of the 4 algorithms (X-axis). Genes are sorted based on hierarchical clustering of their importance percentiles. Gray boxes designate benchmark genes that did not have genetic markers within a 40kb window. Confidence levels (high or medium) were assigned based on the type of evidence available for the benchmark gene (see Methods). Algorithms were abbreviated as in Figure 2.

Figure 5

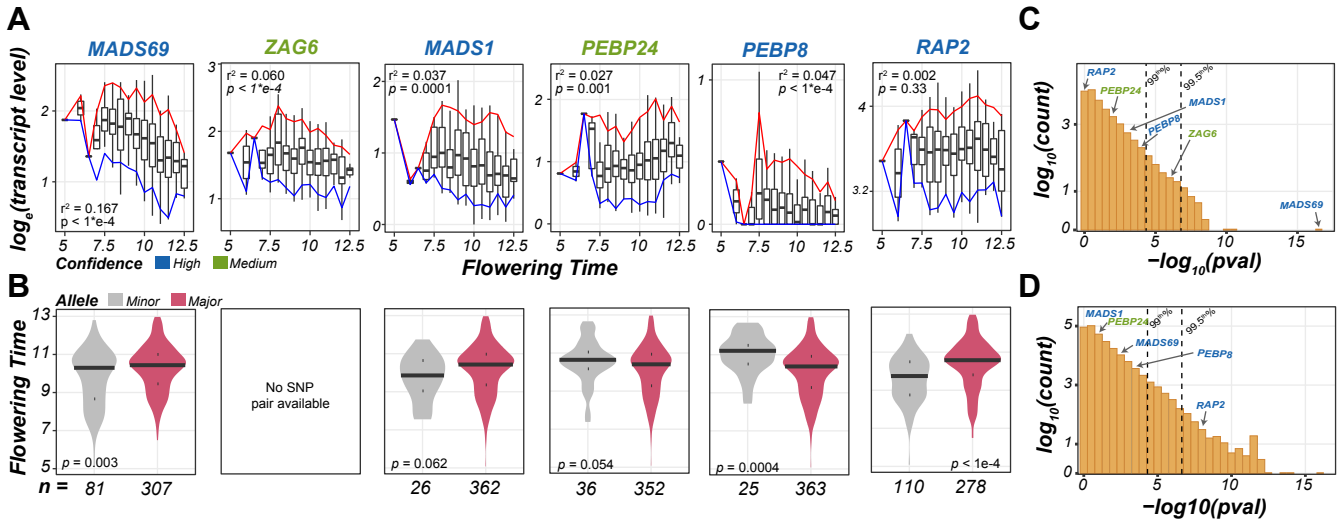

Figure 5. Relationship between transcript level/allele type and flowering time for benchmark genes

(A) Boxplots show the transcript levels (loge(Fold-Change)) over flowering time bin with the 5th (blue) and 95th (red) percentile range shown. Flowering time was defined as the growing degree days/100. Linear models were fit and adjusted r2 and p-values are shown. Confidence levels of benchmark genes were designated as in (4). (B) Distributions of flowering time for lines with the major (red) or minor (gray) alleles for the genetic marker paired with each benchmark gene as indicated in (A). Differences in flowering time by allele were tested using t-tests. (C) Number of transcripts (Y-axis) for which transcript levels were associated with flowering time in linear models within p-value bins (-log10(p-value); X-axis). Benchmark genes are labeled as in (A). (D) Number of genetic markers (Y-axis) for which differences in flowering time by allele from t-tests were within p-value bins (-log10(p-value); X-axis). Benchmark genes are labeled as in (A).