LETTER

doi:10.1038/nature25966

Dysregulation of expression correlates with rare-allele burden and fitness loss in maize

Karl A. G. Kremling[1], Shu-Yun Chen[2,3], Mei-Hsiu Su[2], Nicholas K. Lepak[4], M. Cinta Romay[2], Kelly L. Swarts[1,5], Fei Lu[2,6], Anne Lorant[7], Peter J. Bradbury[4] & Edward S. Buckler[1,2,4]

Here we report a multi-tissue gene expression resource that represents the genotypic and phenotypic diversity of modern inbred maize, and includes transcriptomes in an average of 255 lines in seven tissues. We mapped expression quantitative trait loci and characterized the contribution of rare genetic variants to extremes in gene expression. Some of the new mutations that arise in the maize genome can be deleterious; although selection acts to keep deleterious variants rare, their complete removal is impeded by genetic linkage to favourable loci and by finite population size[1–4]. Modern maize breeders have systematically reduced the effects of this constant mutational pressure through artificial selection and self-fertilization, which have exposed rare recessive variants in elite inbred lines[5]. However, the ongoing effect of these rare alleles on modern inbred maize is unknown. By analysing this gene expression resource and exploiting the extreme diversity and rapid linkage disequilibrium decay of maize[6], we characterize the effect of rare alleles and evolutionary history on the regulation of expression. Rare alleles are associated with the dysregulation of expression, and we correlate this dysregulation to seed-weight fitness. We find enrichment of ancestral rare variants among expression quantitative trait loci mapped in modern inbred lines, which suggests that historic bottlenecks have shaped regulation. Our results suggest that one path for further genetic improvement in agricultural species lies in purging the rare deleterious variants that have been associated with crop fitness.

The phenotypic consequences of rare deleterious alleles are of interest for their role in disease and fitness, but their effects are difficult to detect without prohibitively large sample sizes[4,7,8]. Gene expression, a phenotype for which millions of observations are obtainable, has previously been associated with rare alleles through pedigrees[9] and correlated with the burden of rare alleles in putative *cis*-regulatory regions[10]. However, links between regulation of expression, fitness and altered allele frequencies resulting from population bottlenecks are not well established.

Maize provides a powerful system with which to evaluate these questions. Although the mutation rate of maize ($9–20 \times 10^{-9}$ mutations per base pair per generation) is similar to that of humans[11], the rapid decay of linkage disequilibrium in much of the genome is an order of magnitude faster than in other large eukaryotic genomes, which improves the resolution of associating phenotypes with both rare and common genotypes (average $r^2 < 0.1$ in 2 kb)[6,12]. Maize is also very diverse (nucleotide diversity, $\pi \approx 1.4\%$[13]; approximately 14× more than humans) having preserved much of the diversity from its wild tropical relative teosinte[12]. However, population bottlenecks during temperate adaptation and modern breeding have severely reduced the diversity of maize and severely reduced its effective population size (N_e) from more than one million in pre-bottleneck tropical landraces[5,14,15].

Because it has undergone intensive selection and systematic inbreeding since approximately 1900, which has contributed to an eightfold increase in productivity[16,17], maize also provides an extreme case in which to examine the capacity of selection and inbreeding to purge deleterious variants. Additionally, the rapid linkage disequilibrium decay in maize enables the testing of whether rare alleles or novel combinations of common alleles drive extreme expression.

Here we quantified mRNA expression from 299 maize lines that represent the genotypic and phenotypic diversity of modern inbred maize[18]. We automated a 3′ mRNA sequencing method (QuantSeq, Lexogen GmBH), which is more efficient and accurate than mRNA sequencing and deals well with paralogues[19,20]. We used this method to profile seven diverse tissues (Extended Data Fig. 1 and Supplementary Table 1). All lines had previously been genotyped by whole genome sequencing to produce a set of 61,430,377 segregating variants[21], which we used to map expression quantitative trait loci (eQTL). We also calculated the fraction of variance in expression explained by individual single nucleotide polymorphisms (SNPs). We further used the variance explained to quantify whether alleles that were rare in pre-bottleneck tropical lines[22] have a disproportionately large role in the regulation of expression in modern temperate bottlenecked germplasm (see Methods).

To determine whether rare alleles contribute to extreme gene expression, we investigated whether rare-allele abundance is greatest in individuals with extreme gene expression. We calculated an expression rank and upstream (5 kb) rare-allele count for each combination of gene and individual maize line[10] (Fig. 1a). In each tissue, this was done separately for each of the 5,000 most highly expressed genes. At each expression rank we plotted the mean number of rare alleles across the 5,000 gene-line combinations, which gave a mean number of rare *cis* variants at each rank. If rare alleles drive extreme expression, then individuals at the lowest and highest expression ranks should have the highest number of rare *cis* variants. Significance was tested using a quadratic regression to compare the expression rank with the *cis* rare-allele count as previously described[10]. Using allele frequencies from previous whole genome sequencing of these lines[21], we find that the abundance of rare SNPs within 5-kb upstream of the nearest gene (minor allele frequency (MAF) ≤ 0.05) is significantly correlated with extreme over- and underexpression in all tissues, relative to the population mean per gene ($P < 1.90 \times 10^{-80}$ and $R^2 = 0.72$, Fig. 1b and Extended Data Fig. 2). Although the quadratic regression is highly significant, even more-extreme departures exist at the tails than predicted by the regression (Fig. 1b, c); we therefore focus on the rare-allele count deviations of the five highest- and lowest-expressing individuals across the 5,000 most-expressed genes. Across the tissues, individuals that express at the lowest five expression ranks are enriched 1.68–1.86-fold for local rare alleles compared to the middle two quartiles, whereas the top five

[1]Section of Plant Breeding and Genetics, 175 Biotechnology Building, Cornell University, Ithaca, New York 14853, USA. [2]Institute for Genomic Diversity, 175 Biotechnology Building, Cornell University, Ithaca, New York 14853, USA. [3]Institute of Plant and Microbial Biology, Academia Sinica 128, Sec 2nd, Academia road, Taipei, 11529, Taiwan. [4]USDA-ARS, R. W. Holley Center, Cornell University, Ithaca, New York 14853, USA. [5]Research Group for Ancient Genomics and Evolution, Department of Molecular Biology, Max Planck Institute for Developmental Biology, Spemannstr. 35, 72076 Tübingen, Germany. [6]The State Key Laboratory of Plant Cell and Chromosome Engineering, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing 100101, China. [7]Department of Plant Sciences, University of California Davis, Davis, California 95616, USA.

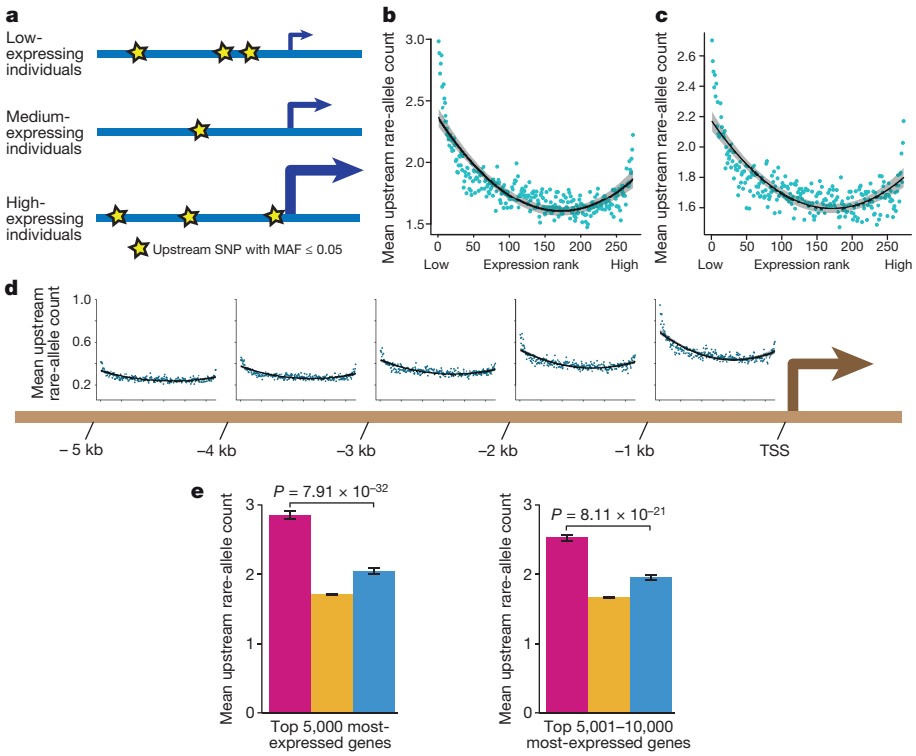


Figure 1 | The abundance of local rare alleles correlates with extremes in expression. a, A greater number of rare alleles is expected upstream of a gene in individuals that over- or underexpress a given gene relative to the mean of the population. **b**, Significant quadratic relationship between the expression rank of each line, in each of the top 5,000 most-expressed genes and the average local (5-kb upstream) rare-allele count. Quadratic regression[10] of root tissue is shown here. ($n = 273$ unique inbred samples, $P = 1.95 \times 10^{-75}$, $R^2 = 0.72$). Each point is the mean number of rare alleles within 5-kb upstream of genes for lines that share an expression rank for one of the 5,000 genes. **c**, Quadratic relationship between expression rank for medium-expressed genes (5,001–10,000) and local (5-kb upstream) rare-allele count ($n = 273$ unique inbred samples, $P = 2.16 \times 10^{-66}$, $R^2 = 0.67$). **d**, Expression ranks versus 5-kb upstream rare-allele counts divided into 1-kb windows ($n = 273$ unique inbred samples). TSS, transcription start site. **e**, Comparison of the number of rare *cis* alleles near genes for individuals in the bottom five expression ranks (fuchsia, $n = 5$ unique inbred samples measured for each of 5,000 genes) versus the middle two quartiles (yellow, $n = 137$ unique inbred samples measured for each of 5,000 genes) versus the top five expression ranks (blue, $n = 5$ unique inbred samples measured for each of 5,000 genes) (mean ± s.e.m.) within the top 5,000 and next 5,000 most-expressed genes. *P* values from two-sided *t*-tests, comparing the top and bottom five expression ranks in the top 5,000 and next 5,000 most-expressed gene sets, are shown. This is consistent across tissues and gene sets (Extended Data Fig. 2, 3).

ranks are enriched by 1.22–1.46-fold over the middle two quartiles (Fig. 1b, e and Extended Data Fig. 2).

When broken down by proximity to the transcription start site, we observe that rare variants nearest to the transcription start site have the greatest effect (Fig. 1d). It is also notable that the ratio of underexpressing to overexpressing individuals possessing rare *cis* variants is greater nearest to the transcription start site (Fig. 1d). This is consistent with the proposition that disruption to gene-proximal promoters lowers expression, whereas distal disruptions have a relatively greater chance of increasing expression.

Although both over- and underexpression are potentially deleterious consequences of dysregulation, we note that there are significantly more *cis* rare alleles in underexpressing individuals than there are near the same genes in overexpressing individuals ($P = 7.91 \times 10^{-32}$, Fig. 1e). This observation is consistent with stronger selection against underexpression, and with the expression dosage model of fitness and heterosis[23]. Additionally, it may be the case that rare variants are more likely to disrupt a promoter element than a repressor element, which would lead to a loss rather than a gain of expression. However, the greater apparent abundance of rare *cis* variants in underexpressing individuals may also partly be the result of under-calling expression, owing to mapping bias for genes that are located on rare haplotypes.

To evaluate how regulatory load affects genes at different expression levels, we compared how this load affects more- and less-expressed genes in each tissue. Purifying selection acts on regulatory variants[24] and is expected to be strongest near the most highly expressed genes[25]. This can result in the purging of dysregulatory variants or—if selection is not strong enough—an increase in rare variants because deleterious variants are reduced in frequency but not purged. By comparing rare-allele burdens 5-kb upstream of the most highly expressed genes (top 5,000) with those upstream of the medium-expressed genes (next 5,000 most-expressed genes), we note that across all tissues the 5,000 most highly expressed genes contain more *cis* rare alleles (Fig. 1e and Extended Data Figs 2, 3). However, consistent with an increased effect of purifying selection on the most highly expressed genes, we also note that the deleterious *cis* variants appear to be purged from the top 1,000 most highly expressed genes, which in some tissues have fewer rare *cis* variants than the next 1,000 most-expressed genes (Extended Data Fig. 4).

These analyses reveal the cumulative effects of rare alleles; we can also directly estimate the effect of ancestrally rare alleles on expression by exploiting the evolutionary history and bottlenecks in maize. This is possible because we can sample allele frequencies in extant tropical maize populations that have not experienced the same bottlenecks, which enables us to infer the ancestral allele frequency for each SNP. Approximately half a million SNPs in the maize HapMap3 (about 1% of total SNPs) that are rare (≤0.05 MAF) in 335 genotyped tropical lines[22] are common (>0.2 MAF) across our predominantly temperate expression-profiled lines (the base of the third leaf is shown in Fig. 2; remaining tissues shown in Extended Data Figs 5, 6), illustrating the effects of the tropical-to-temperate bottleneck. This class of formerly rare tropical SNPs—which is now of sufficient frequency to be easily detected in our eQTL association study using modern inbred lines—enables powerful tests of ancestrally rare-allele effects. By comparing

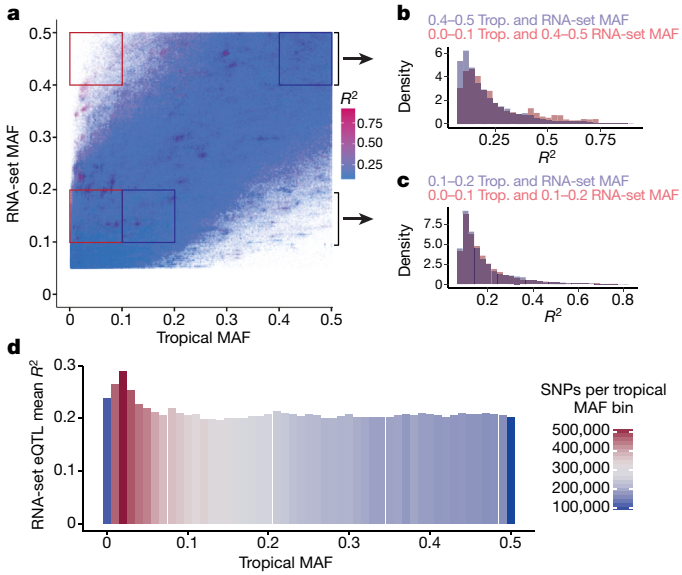


Figure 2 | Ancestral rare alleles are significantly enriched for highly explanatory *cis* eQTL in modern germplasm. a, Two-dimensional MAF plot coloured by R^2 from *cis* eQTL for SNPs associated below $P = 0.00001$. Using $n = 263$ unique inbred lines and conducting an eQTL association study for each expression trait, we quantified the fraction of variance in expression explained (R^2) by each *cis* SNP for its most strongly associated gene. eQTL R^2 was determined by linear regression (matrix eQTL). **b**, **c**, Across temperate MAF levels, significantly more variance in expression is explained by SNPs that are rare in the tropical (Trop.) maize germplasm that has not undergone a bottleneck (two-sided Wilcoxon signed-rank and Kolmogorov–Smirnov tests, $P < 1.12 \times 10^{-240}$). Results of tissue from base of leaf three shown here ($n = 263$ unique inbred samples). **d**, Histogram comparing RNA-set *cis* eQTL R^2 across tropical MAF bins for SNPs with MAF over 0.2 in the RNA set. R^2 determined by eQTL regression in matrix eQTL. Bars are coloured by the number of SNPs in each tropical MAF bin to illustrate that rarer alleles, which explain the most variance in expression, constitute the most abundant class of SNPs. A minimum RNA-set MAF of 0.2 is used, given the limited statistical power to map eQTL beneath this threshold with our RNA-set sample sizes. $n = 263$ unique inbred samples used in eQTL mapping.

the variance explained by alleles that are common in tropical maize lines versus those that are rare in these lines, while controlling for equal allele frequency in the RNA set, it is clear that tropical rare alleles explain significantly more variance in expression than do tropical common alleles across the whole genome in all tissues (Fig. 2b, c and Extended Data Figs 5, 6, Kolmogorov–Smirnov and Wilcoxon signed-rank tests, $P < 9.06 \times 10^{-33}$ for all tissues). This is clear from the high-R^2 SNPs, which can be seen to the left of the diagonal in the 2D MAF plot in Fig. 2a. Although some of these differences are adaptive, there are only a few hundred loci in the maize genome with a highly significant F_{st} (fixation index) that are thought to be adaptive[12,26]. Therefore, many of the highly explanatory eQTL SNPs are likely to be deleterious. This highlights the fact that formerly rare alleles have significantly larger effects on quantitative expression phenotypes than do common variants, and suggests that the temperate bottleneck imposed substantial changes on gene regulation.

Finally, we investigated the association between dysregulation in gene expression and a fitness trait. To quantify fitness, we used previously published seed yields of the expression-profiled lines (multi-environment best linear unbiased predictions[27]). First, we tested whether overall expression of the top 5,000 most-expressed genes could predict seed yields using ridge regression, which works across tissues (Extended Data Fig. 7). However, if dysregulation has a major role in altering phenotype, then the deviation in expression should also be predictive. We find support for this in two analyses: the cumulative deviation in expression for the 5,000 most-expressed genes in adult leaves and kernels significantly correlates with seed-weight fitness (Extended Data Fig. 8, Extended Data Table 1 and Supplementary Methods). Furthermore, by using the matrix of absolute deviation for the top 5,000 most-expressed genes in each tissue we were able to predict seed-weight fitness using ridge regression in six of seven tissues (Fig. 3). This predictive ability is significant in all tissues (Fig. 3, $r = 0.19$–0.38, $P = 0.01$ to $P < 2.9 \times 10^{-7}$) except for immature roots ($P = 0.68$). Consistent with expectations, an increased rare-allele burden does correlate with decreased fitness (Extended Data Fig. 9), but this result is likely to be partially confounded with the recent breeding history of maize and pedigree relationships. Large bi-parental populations without population structure would be better suited to addressing this question.

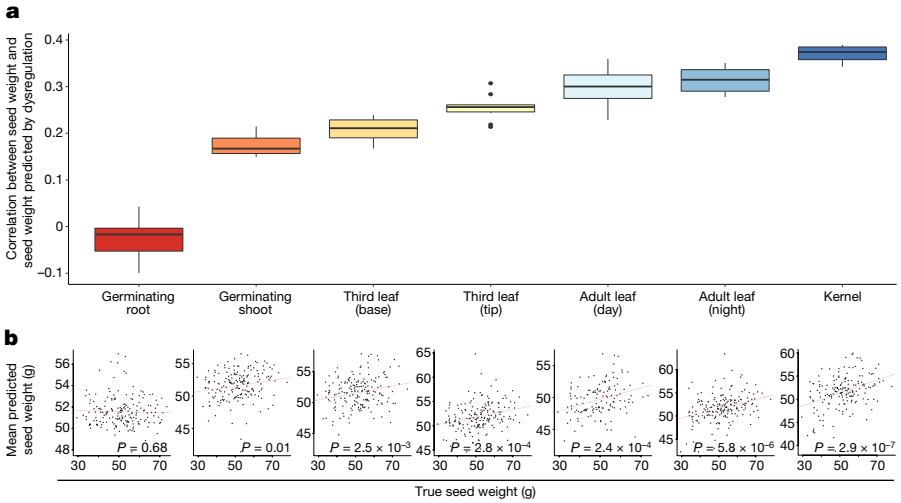


Figure 3 | Dysregulation of expression can predict fitness. Dysregulation of expression in the top 5,000 most-expressed genes from six of seven tissues significantly predicts seed-weight fitness. **a**, Range of correlations between predicted and true seed weight from ten repetitions of nested tenfold cross validation (ten inner and ten outer) using ridge regression. In the box plots, the middle horizontal lines represent the median, hinges represent the 25th and 75th percentiles (the interquartile range), the upper and lower whiskers extend to maximum and minimum points no more than 1.5× the interquartile range beyond the hinges, and individual dots are outliers beyond the whiskers. **b**, True seed weight versus mean of predicted seed weight in grams. *P* values obtained from linear regression between true seed weight and mean of predicted seed weight. For both **a** and **b**, sample sizes were as follows: 2-cm root tips (unique $n = 181$) and shoots (unique $n = 183$) of germinating seedlings; 2-cm base (unique $n = 181$) and tip (unique $n = 182$) of leaf 3, leaves collected in the field during the day (unique $n = 135$) and night (unique $n = 187$); and 350-growing-degree-day kernels (unique $n = 171$), post sexual maturity (anthesis).

The ability of dysregulation to predict fitness is notable, given that the expression profiles were not collected in the same environments or years in which the seed-weight fitness phenotypes were determined. In brief, environmental variance cannot be controlled between the individuals in which expression and fitness were measured, which adds noise to these regressions. For this reason, the effect of expression dysregulation on fitness calculated here may be underestimated.

Our results demonstrate the influence of rare genetic variation on gene expression, and provide an example that connects the dysregulation of expression and a fitness trait. Consistent with population genetic expectations and evidence that recombination is insufficient to purge deleterious variants in modern maize[28], these results illustrate the disproportionate effect of rare alleles on thousands of expression phenotypes as well as the consequences of newly common alleles in modern low-N_e populations. Despite intensive selection and millions of yearly field trials by breeders[5], maize provides evidence for the persistence of rare deleterious alleles in modern agricultural species after a strong bottleneck. This suggests that even intensive artificial selection is insufficient to purge genetic load. Although genomic selection has successfully combined favourable sets of common variants to improve yields, additional targeted breeding approaches and genetic manipulation would enable further removal of deleterious mutations and their phenotypic consequences.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 7 October 2016; accepted 1 February 2018.
Published online 14 March 2018.

1. Kimura, M., Maruyama, T. & Crow, J. F. The mutation load in small populations. *Genetics* **48,** 1303–1312 (1963).
2. Marth, G. T. *et al.* The functional spectrum of low-frequency coding variation. *Genome Biol.* **12,** R84 (2011).
3. Henn, B. M., Botigué, L. R., Bustamante, C. D., Clark, A. G. & Gravel, S. Estimating the mutation load in human genomes. *Nat. Rev. Genet.* **16,** 333–343 (2015).
4. Gibson, G. Rare and common variants: twenty arguments. *Nat. Rev. Genet.* **13,** 135–145 (2012).
5. Troyer, A. F. A retrospective view of corn genetic resources. *J. Hered.* **81,** 17–24 (1990).
6. Remington, D. L. *et al.* Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proc. Natl Acad. Sci. USA* **98,** 11479–11484 (2001).
7. Kono, T. J. Y. *et al.* The role of deleterious substitutions in crop genomes. *Mol. Biol. Evol.* **33,** 2307–2317 (2016).
8. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461,** 747–753 (2009).
9. Li, X. *et al.* Transcriptome sequencing of a large human family identifies the impact of rare noncoding variants. *Am. J. Hum. Genet.* **95,** 245–256 (2014).
10. Zhao, J. *et al.* A burden of rare variants associated with extremes of gene expression in human peripheral blood. *Am. J. Hum. Genet.* **98,** 299–309 (2016).
11. Jiao, Y. *et al.* Genome-wide genetic changes during modern breeding of maize. *Nat. Genet.* **44,** 812–815 (2012).
12. Gore, M. A. *et al.* A first-generation haplotype map of maize. *Science* **326,** 1115–1117 (2009).
13. Tenaillon, M. I. *et al.* Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. *mays* L.). *Proc. Natl Acad. Sci. USA* **98,** 9161–9166 (2001).
14. Vigouroux, Y. *et al.* Rate and pattern of mutation at microsatellite loci in maize. *Mol. Biol. Evol.* **19,** 1251–1260 (2002).
15. Beissinger, T. M. *et al.* Recent demography drives changes in linked selection across the maize genome. *Nat. Plants* **2,** 16084 (2016).
16. Duvick, D. N. The contribution of breeding to yield advances in maize (*Zea mays* L.). *Adv. Agron.* **86,** 83–145 (2005).
17. Troyer, A. F. & Wellin, E. J. Heterosis decreasing in hybrids: yield test inbreds. *Crop Sci.* **49,** 1969–1976 (2009).
18. Flint-Garcia, S. A. *et al.* Maize association population: a high-resolution platform for quantitative trait locus dissection. *Plant J.* **44,** 1054–1064 (2005).
19. Eveland, A. L., McCarty, D. R. & Koch, K. E. Transcript profiling by 3′-untranslated region sequencing resolves expression of gene families. *Plant Physiol.* **146,** 32–44 (2008).
20. Lohman, B. K., Weber, J. N. & Bolnick, D. I. Evaluation of TagSeq, a reliable low-cost alternative for RNAseq. *Mol. Ecol. Resour.* **16,** 1315–1321 (2016).
21. Bukowski, R. *et al.* Construction of the third generation *Zea mays* haplotype map. *Gigascience* http://doi.org/10.1093/gigascience/gix134 (2017).
22. Romay, M. C. *et al.* Comprehensive genotyping of the USA national maize inbred seed bank. *Genome Biol.* **14,** R55 (2013).
23. Yao, H., Dogra Gray, A., Auger, D. L. & Birchler, J. A. Genomic dosage effects on heterosis in triploid maize. *Proc. Natl Acad. Sci. USA* **110,** 2665–2669 (2013).
24. Josephs, E. B., Lee, Y. W., Stinchcombe, J. R. & Wright, S. I. Association mapping reveals the role of purifying selection in the maintenance of genomic variation in gene expression. *Proc. Natl Acad. Sci. USA* **112,** 15390–15395 (2015).
25. Gout, J.-F., Kahn, D., Duret, L. & Paramecium Post-Genomics Consortium. The relationship among gene expression, the evolution of gene dosage, and the rate of protein evolution. *PLoS Genet.* **6,** e1000944 (2010).
26. Hufford, M. B. *et al.* Comparative population genomics of maize domestication and improvement. *Nat. Genet.* **44,** 808–811 (2012).
27. Hung, H.-Y. *et al.* The relationship between parental genetic or phenotypic divergence and progeny variation in the maize nested association mapping population. *Heredity* **108,** 490–499 (2012).
28. Rodgers-Melnick, E. *et al.* Recombination in diverse maize is stable, predictable, and associated with genetic load. *Proc. Natl Acad. Sci. USA* **112,** 3823–3828 (2015).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank J. Pardo, J. Wallace, R. Punna, K. Shirasawa and S. Miller for assistance with tissue collection; J. Budka and G. Inzinna for field and greenhouse assistance; R. Bukowski for running the maize HapMap genotyping pipeline; L. Johnson and Z. Miller for database curation; G. Gibson, M. Wolfe, J.-L. Jannink, M. Hufford and J. Ross-Ibarra for discussions; P. Schweitzer, J. Mosher, A. Tate, J. Mattison, M. Magallanes-Lundback, I. Holländer and D. Daujotyte for guidance on RNA extraction, library preparation automation and sequencing; and S. Miller for copy-editing. This work was supported by the US Department of Agriculture–Agricultural Research Service and the National Science Foundation grants IOS-0922493 and IOS-1238014 to E.S.B. The National Science Foundation Graduate Research Fellowship Program grant DGE-1650441 and the Section of Plant Breeding and Genetics at Cornell University provided support to K.A.G.K. The Taiwanese Ministry of Science and Technology Overseas Project for Post Graduate Research grant 104-2917-I-564-015 supported S.-Y.C.

Author Contributions K.A.G.K. and E.S.B. designed the experiments and wrote the manuscript. K.A.G.K performed the analyses and made the RNA-seq libraries. K.A.G.K., S.-Y.C., and M.-H.S. extracted RNA. N.K.L. managed germplasm and plants with K.A.G.K., M.C.R., K.L.S. and A.L. produced and imputed HapMap genotypic data. P.J.B. implemented matrixEQTL in Java/TASSEL. F.L. implemented SNP calling from RNA-seq data.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to E.S.B. (esb33@cornell.edu) and K.A.G.K. (kak268@cornell.edu).

Reviewer Information *Nature* thanks N. Springer and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

No statistical methods were used to predetermine sample size. The investigators were not blinded to allocation during experiments and outcome assessment. Planting order was randomized for chamber and greenhouse experiments, as well as within maturity groups for field-grown plants. Flat order was also randomized.

Tissue collection and RNA extraction. Two centimetres of the base of leaf three from three uniform plants per genotype at third leaf stage were collected from 10:30 to 12:00. Tissue was collected 16–20 June 2015 and immediately frozen in liquid nitrogen. Seedlings were grown at one per cell in 96-cell flats in LM-111 mix (Lambert) at Guterman Greenhouse. Plants were watered daily and grown with 40–60% humidity under supplementary high-pressure 600-W sodium lighting from 8:30 to 20:30. Day temperature was 30 °C and night temperature was 24 °C. Flat positions were randomized daily.

Two centimetres of germinating seedling roots and whole germinating seedling shoots were collected from three plants per genotype from 11:00 to 13:00 on the day of germination and immediately frozen in liquid nitrogen. Seeds were germinated at eight per cell in 24-cell flats in medium grain vermiculite (Lambert) to facilitate collection of roots without soil contamination. A walk-in growth chamber was used with two 12-inch 24-W 6400K T5 fluorescent lights per shelf. Lights were on from 8:00 to 24:00. The lights were 36.5 cm from the shelves. Humidity ranged between 40 and 60%. Seeds were watered daily and the temperature was maintained between 22 and 24 °C. Flat positioning was randomized daily.

Seven developing kernels were collected from three plants at 350-growing-degree days after self-pollination and immediately frozen in liquid nitrogen. Based on pollination date, kernels were harvested between 11:00 and 13:00 from 25 August 2014 to 19 September 2014, from plants grown in field M1 (Lima silt loam soil) at the Cornell Musgrave Research Farm. Seven immature kernels were collected and combined from three plants per genotype. End plants from each plot were avoided whenever possible.

Using these same plots, mature leaves were collected during the day (11:00–13:00) and during the night (23:00–01:00) in two batches, on 8 August 2014 and 26 August 2014. Leaves were collected in two batches with the collection date assigned depending on what had already flowered by the dates above (see the column titled 'Tissue' in Supplementary Table 1 for collection date of adult leaves). Leaf sections from three plants per plot were collected into liquid nitrogen from a 1-cm section to one side of the midrib from the second leaf below the tassel. The proximal–distal middle of the leaf blade was determined by folding the leaf in half such that the leaf tip touched the ligule.

Non-kernel tissues from three plants per genotype × tissue combination were homogenized with two 3-mm steel beads using a Genogrinder in 30-s increments to ensure samples remained frozen (Spex Sample Prep). Between each grinding increment, samples were placed in liquid nitrogen. RNA was extracted using TRIzol (Invitrogen) with Direct-zol columns (Zymo Research). Twenty-one frozen kernels (seven kernels from three individuals) per genotype were ground in an IKA seed mill (IKA). Hot borate and lithium chloride were used to extract RNA from kernels[29]. For kernels, RNA integrity was assessed using gel electrophoresis and extractions were repeated for degraded samples. For all tissues, RNA was quantified using RNA QuantiFluor (Promega) and diluted on a Beckman Biomek NXp to a concentration of 100 ng/μl for library preparation. 3′ RNA-seq libraries were prepared robotically from 500 ng total RNA in 96-well plates on an NXp liquid handler (Beckman Coulter) using QuantSeq FWD kits (Lexogen) according to the manufacturer's instructions. Post-PCR cleanup was performed manually according to the QuantSeq protocol. Libraries were pooled to 96-plex, based on concentration as measured by DNA QuantiFluor (Promega). Molar concentrations for each pool were calculated from Bioanalyzer (Agilent) fragment lengths and digital PCR quantifications. Pools were sequenced with 90 nucleotide single-end reads using Illumina TruSeq primers on an Illumina NextSeq 500 with v2 chemistry at the Cornell University Sequencing facility.

Trimmomatic[30] version 0.32 was used to remove the first 12 bp and Illumina Truseq adaptor remnants from each read. The first 12 bp were removed based on kit maker instructions, and the propensity for errors to occur in sequencing after random priming. The splice-aware STAR[31] aligner v.2.4.2a was used to align reads against the maize genome annotation version AGPv.3.29, allowing a read to map in at most 10 locations (–outFilterMultimapNmax 10) with at most 4% mismatches (–outFilterMismatchNoverLmax 0.04), while filtering out all non-canonical intron motifs (–outFilterIntronMotifs RemoveNoncanonicalUnannotated). Default settings from HTSeq[32] v.0.6.1 were used to obtain gene-level counts from the resulting BAM files. Counts were normalized by library sequencing depth using the -estimateSizeFactors method in DESeq2[33] in R. Expression counts were transformed using Box–Cox transformation after adding a small random value beneath the minimum detection threshold in order to enable transformation of zeroes.

Sample sizes. After positive sample identification by SNP-calling from RNA-seq data (as described below), sample sizes for each tissue were as follows: 2-cm root tips ($n = 291$, unique $n = 273$) and shoots ($n = 295$, unique $n = 278$) of germinating seedlings; 2-cm base ($n = 302$, unique $n = 263$) and tip ($n = 295$, unique $n = 265$) of leaf 3; 350-growing-degree-day kernels ($n = 254$, unique $n = 229$); and post-sexual-maturity (anthesis) leaves during the day ($n = 210$, unique $n = 204$) and night ($n = 276$, unique $n = 260$). Each sample was pooled from three individuals per genotype.

Genotyping. Genotypes from maize HapMap 3.2.1 were called as previously described[21]. Hapmap 3.2.1 was called on 1,268 inbred genotypes from across the world, with highly variable depth of coverage, and paralogous sites were retained, as they provide signal in genome-wide association studies. Because paralogous sites were retained, we used *k*-nearest neighbour imputation[34] (KNNi) as implemented in TASSEL[35] to impute; this was robust to high error rates in genotype calling, but KNNi over-imputes missing data to the major allele. Despite this bias, KNNi imputation for HapMap 3.2.1 had an overall accuracy of 0.988 and a minor allele imputation accuracy of 0.94 for imputed genotypes. Imputed HapMap 3.2.1 genotypes were projected onto the genotyping by sequencing (GBS)-genotyped inbred diversity panel (including tropical lines[22]) using FILLIN[36]. Projection was anchored by 465,085 consensus sites between HapMap3 and GBS, in which the physical positions match and the major/minor alleles are shared; the projection accuracy was $r = 0.99$ between masked and subsequently imputed genotypes (0.96 for minor alleles).

To rapidly confirm the identity of the RNA-seq samples, SNPs were called from RNA-seq reads using FastCall[37]. Using SNPs called from the first 20 Mb on chromosome 10 (AGPv.3), all samples that did not match the reference genotype were discarded, leaving 1,960 samples. These SNPs called from RNA-seq were not used for eQTL association studies or other downstream analyses.

Gene set filtering. After library size normalization with DESeq2 (described above), genes were ranked by expression in each tissue and the top 5,000 were placed in the 'top 5,000' set; the next 5,000 were placed in the '5,001–10,000' set for each tissue. The blocks were further broken down into groups of 1,000 for the plots included in the Extended Data.

eQTL mapping and 2D MAF comparisons between tropical and RNA sets. Before mapping eQTL, 25 hidden factors were calculated using PEER[38] for each tissue individually and these were used as covariates together with 5 multidimensional scaling (MDS) coordinates calculated from HapMap 3.2.1[21]. Using the PEER factors and MDS values, eQTL were mapped using a parallelized version of MatrixEQTL[39] implemented in TASSEL using Java. *P* values and R^2 from eQTL hits with significance below $P = 0.00001$ were recorded after performing association tests individually for each tissue with SNPs that had a MAF ≥ 0.05 in the RNA-set samples for a specific tissue.

Each SNP present in the temperate-biased RNA set and previously published tropical lines[22] was then plotted in a 2D minor-allele frequency plot, with its position on the plot determined by its frequency in the tropical and predominately temperate RNA set. Recorded R^2 from *cis* eQTL (within ± 1 Mb) were then used to colour each dot, revealing the effects of ancestral allele frequencies on modern eQTL in Fig. 2. Within each tissue, the most strongly associated SNP–gene pair was kept for each SNP for plotting in the 2D MAF plots.

The distribution of variance explained by SNPs that were formerly rare (tropical MAF under 0.1), but are now common (RNA-set MAF over 0.4) was compared to SNPs that were common in the tropical populations and remain common in the RNA set, which revealed that formerly rare SNPs explain more variance.

Calculation of seed-weight fitness. Multi-location multi-year seed weight best linear unbiased predictions[27] were used as a proxy for fitness. Seed weights from sweet corn and popcorn lines were excluded from Extended Data Fig. 8 and Extended Data Table 1, given that dry seed weight is not a rational proxy for fitness or target of selection in these subpopulations. To be conservative, tropical lines were also excluded from Fig. 3 and Extended Data Fig. 7 because they are likely to be substantially dysregulated relative to the temperate material and have lower seed weights, thus inflating prediction accuracy of fitness from dysregulation. However, significant prediction accuracy is still achieved when tropical lines are included (data not shown).

Prediction of fitness from dysregulation of expression. To quantify the deviation of expression, we subtracted the mean expression level for a single gene from the expression of that gene in each line. This was done for the 5,000 highest expressed genes in each tissue. These 5,000 deviations were then used as *X* values in a ridge regression (alpha = 0) implemented using the glmnet package[40] in R to predict the seed weight best linear unbiased predictions, discussed above. Ten repetitions of nested tenfold (tenfold inner, tenfold outer) cross-validation were carried out using glmnet. In Extended Data Fig. 8, the cumulative absolute deviation in expression for the top 5,000 most-expressed genes (see formula below) was also correlated against the seed weights.

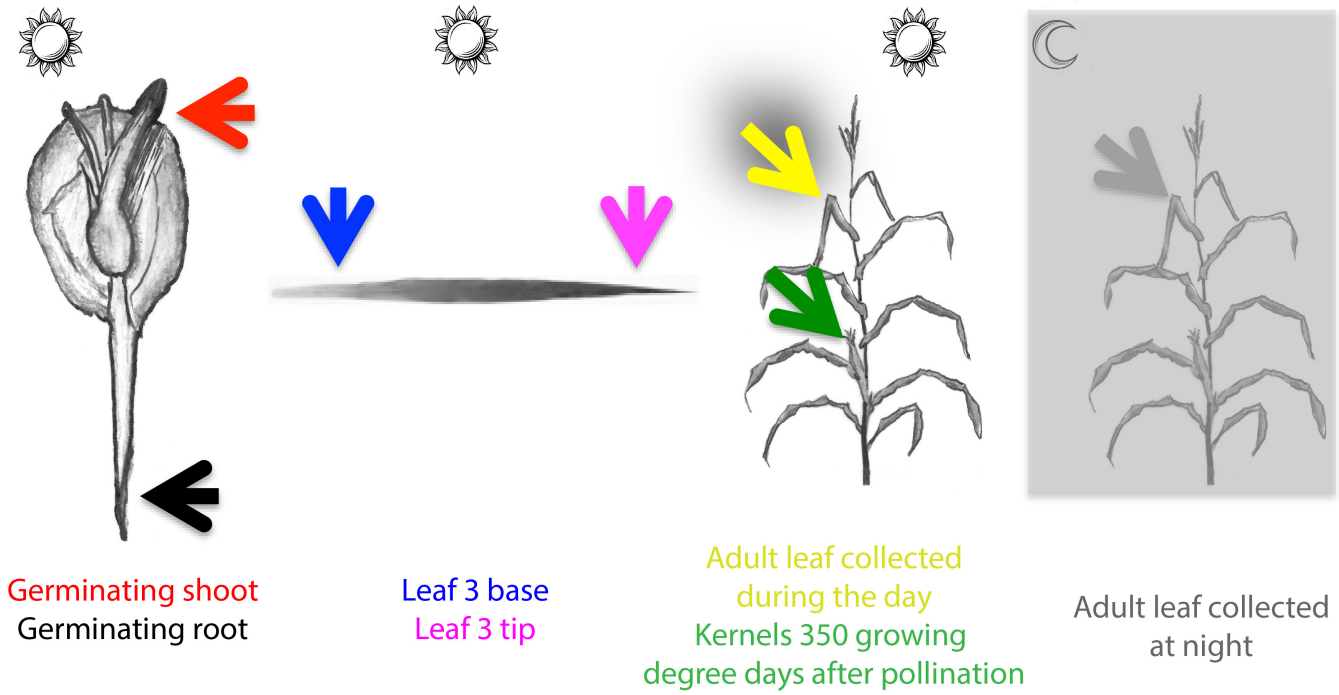
$$D_i = \log_{10}\frac{\sum_{j=1}^{n}(e_{ij} - \mu_j)^2}{n}$$

in which D_i is the deviation of expression of individual i, e_{ij} is the expression of gene j in individual i, n is the number of genes and μ_j is the mean expression of gene j across all samples profiled in the tissue.

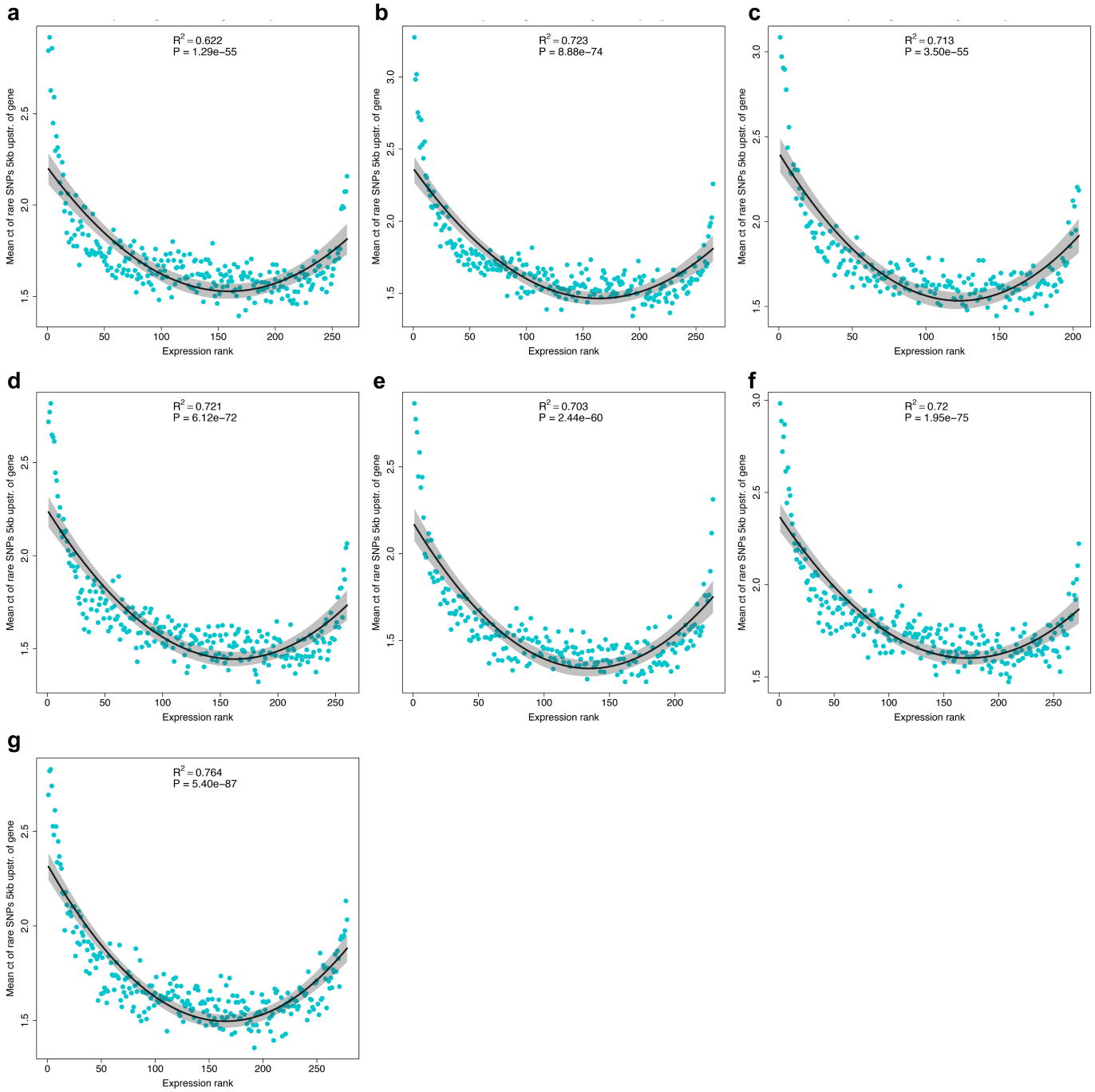
Code availability. Robotic code for Biomek NXp RNA-seq library production can be found at http://www.maizegenetics.net/robotic-code; parallelized implementation of matrix eQTL in TASSEL can be found at http://www.maizegenetics.net/tassel; and FastCall for calling SNPs from whole genome sequencing data can be found at https://github.com/Fei-Lu/FastCall. Further scripts for making plots are available from the corresponding authors upon reasonable request.

Data availability. Sequence data that support the findings of this study have been deposited in the Sequence Read Archive under accession number SRP115041, and in BioProject under accession number PRJNA383416. Processed expression counts are available at the Cyverse Discovery Environment (de.cyverse.org/de/) under the directory: /iplant/home/shared/panzea/dataFromPubs/. All other data are available from the corresponding author(s) upon reasonable request.

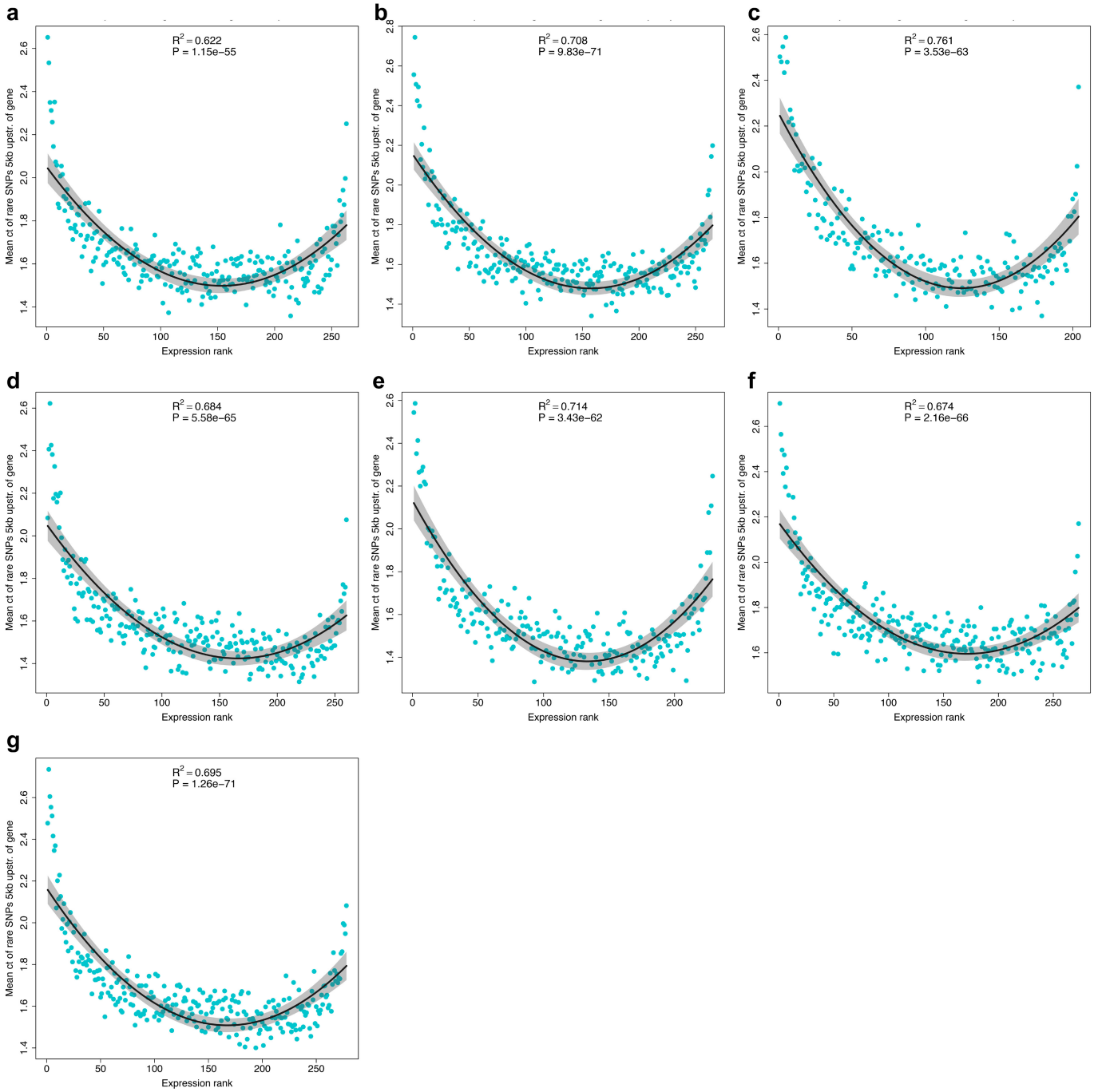
29. Wan, C. Y. & Wilkins, T. A. A modified hot borate method significantly enhances the yield of high-quality RNA from cotton (*Gossypium hirsutum* L.). *Anal. Biochem.* **223,** 7–12 (1994).
30. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30,** 2114–2120 (2014).
31. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013).
32. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31,** 166–169 (2015).
33. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11,** R106 (2010).
34. Money, D. *et al.* LinkImpute: fast and accurate genotype imputation for nonmodel organisms. *G3* **5,** 2383–2390 (2015).
35. Bradbury, P. J. *et al.* TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* **23,** 2633–2635 (2007).
36. Swarts, K. *et al.* Novel methods to optimize genotypic imputation for low-coverage, next-generation sequence data in crop plants. *Plant Genome* **7,** http://doi.org/10.3835/plantgenome2014.05.0023 (2014).
37. Ramu, P. *et al.* Cassava haplotype map highlights fixation of deleterious mutations during clonal propagation. *Nat. Genet.* **49,** 959–963 (2017).
38. Stegle, O., Parts, L., Durbin, R. & Winn, J. A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLOS Comput. Biol.* **6,** e1000770 (2010).
39. Shabalin, A. A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28,** 1353–1358 (2012).
40. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33,** 1–22 (2010).
41. Kisselbach, T. A. *The Structure and Reproduction of Corn* (Cold Spring Harbor Laboratory, 1999).



Extended Data Figure 1 | Tissues that were expression profiled by 3′ RNA-seq. See additional details regarding tissue collection in Methods. Illustrations inspired by ref. 41.

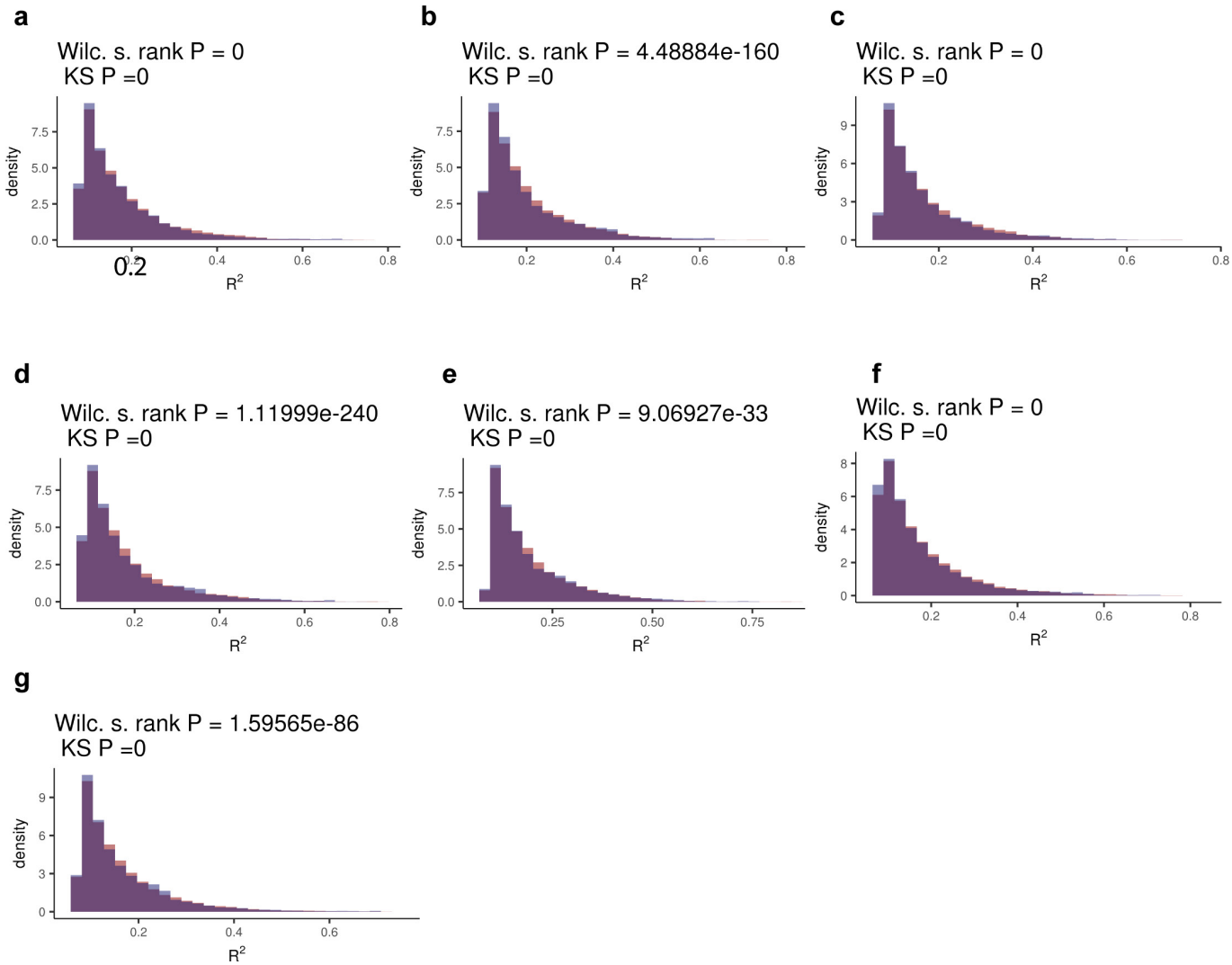


Extended Data Figure 2 | Higher numbers of rare alleles are upstream of genes in extreme-expressing individuals, for the most highly expressed genes. Quadratic regression of the expression rank of each line, for each of the top 5,000 most-expressed genes versus the average local (5-kb upstream) rare-allele count. **a**, Base of leaf three ($n = 263$ unique inbred samples). **b**, Tip of leaf three ($n = 265$ unique inbred samples). **c**, Adult leaves collected during the day ($n = 204$ unique inbred samples). **d**, Adult leaves collected at night ($n = 260$ unique inbred samples). **e**, Kernels at 350-growing-degree days ($n = 229$ unique inbred samples). **f**, Roots of germinating seedling ($n = 273$ unique inbred samples). **g**, Shoots of germinating seedling ($n = 278$ unique inbred samples).

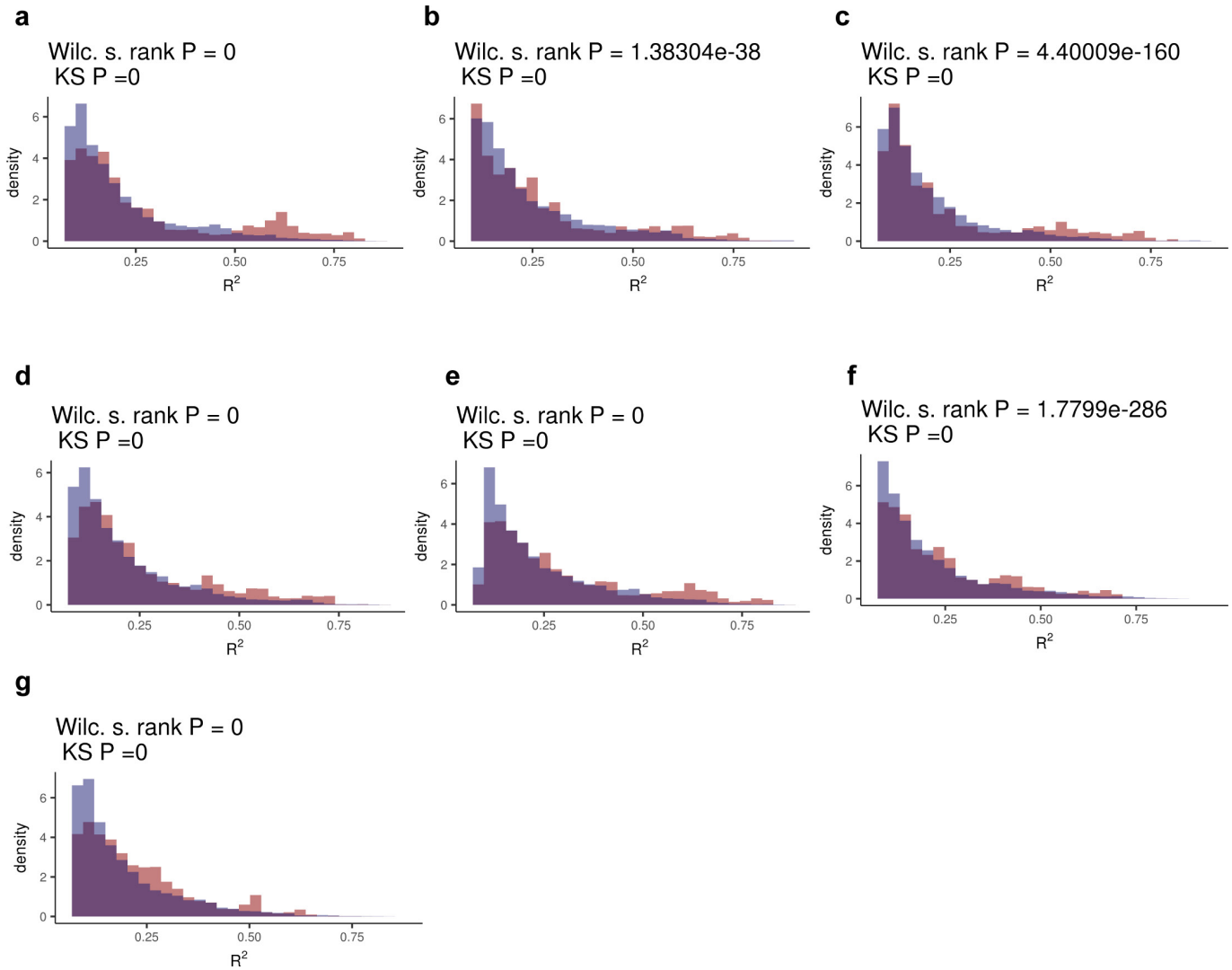


Extended Data Figure 3 | Higher numbers of rare alleles are upstream of genes in extreme-expressing individuals, for the medium-expressed genes. Quadratic regression of the expression rank of each line, for each of the top 5,001–10,000 most-expressed genes versus the average local (5-kb upstream) rare-allele count. **a**, Base of leaf three ($n = 263$ unique inbred samples). **b**, Tip of leaf three ($n = 265$ unique inbred samples). **c**, Adult leaves collected during the day ($n = 204$ unique inbred samples). **d**, Adult leaves collected at night ($n = 260$ unique inbred samples). **e**, Kernels at 350-growing-degree days ($n = 229$ unique inbred samples). **f**, Roots of germinating seedling ($n = 273$ unique inbred samples). **g**, Shoots of germinating seedling ($n = 278$ unique inbred samples).

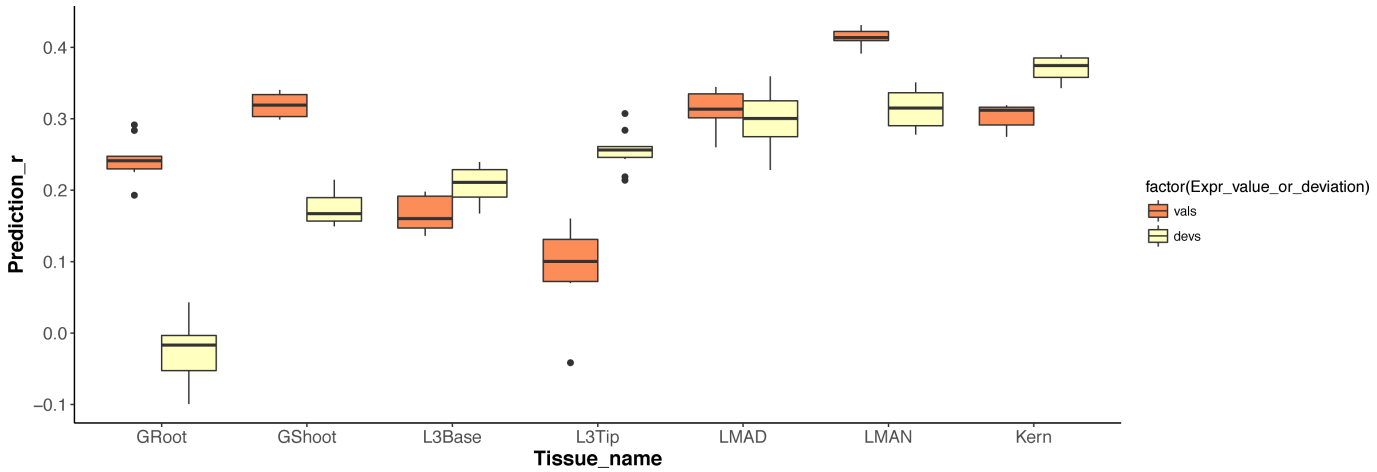
Extended Data Figure 4 | Comparison of the number of rare *cis* alleles near genes with differing expression levels. The 10,000 most-expressed genes in each tissue are divided into groups of 1,000 on the basis of expression level. Plots in each panel show genes ranked 1–1,000, 1,001–2,000, …, 9,001–10,000 from left to right. Each of the individuals represented in each tissue is ranked for expression for each of the 1,000 genes in each group. Individuals in the bottom five expression ranks (fuchsia) versus the middle two quartiles (yellow) versus the top five expression ranks (blue) (mean ± s.e.m.). *Y* axes refer to mean upstream (within 5 kb) rare-allele count. **a**, Roots of germinating seedling ($n = 273$ unique inbred samples). **b**, Shoots of germinating seedling ($n = 278$ unique inbred samples). **c**, Kernels at 350-growing-degree days ($n = 229$ unique inbred samples). **d**, Base of leaf three ($n = 263$ unique inbred samples). **e**, Tip of leaf three ($n = 265$ unique inbred samples). **f**, Adult leaves collected during the day ($n = 204$ unique inbred samples). **g**, Adult leaves collected at night ($n = 260$ unique inbred samples).



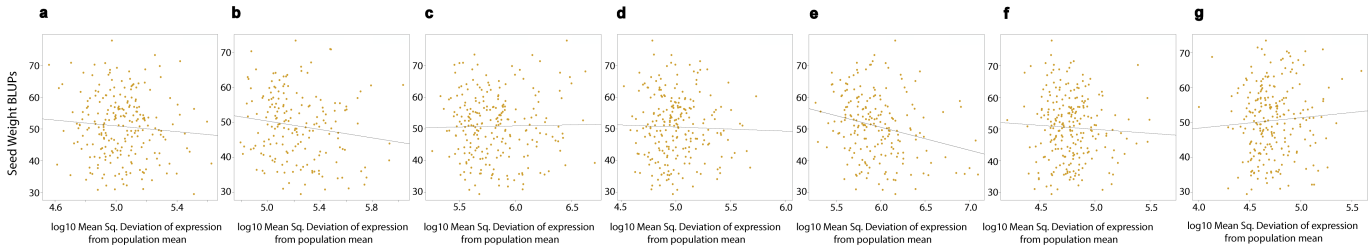
Extended Data Figure 5 | eQTL R^2 distribution comparisons between SNPs in 0.0–0.1 (tropical MAF) and 0.1–0.2 (RNA-set MAF) versus 0.1–0.2 (RNA-set and tropical MAF). a, Adult leaves collected at night ($n = 260$ unique inbred samples). **b**, Adult leaves collected during the day ($n = 204$ unique inbred samples). **c**, Tip of leaf three ($n = 265$ unique inbred samples). **d**, Base of leaf three ($n = 263$ unique inbred samples). **e**, Kernels at 350-growing-degree days ($n = 229$ unique inbred samples). **f**, Shoots of germinating seedling ($n = 278$ unique inbred samples). **g**, Roots of germinating seedling ($n = 273$ unique inbred samples). All pairs of distributions within each tissue are significantly different. $P < 2.2 \times 10^{-16}$ two-sided Wilcoxon signed-rank test and Kolmogorov–Smirnov test.



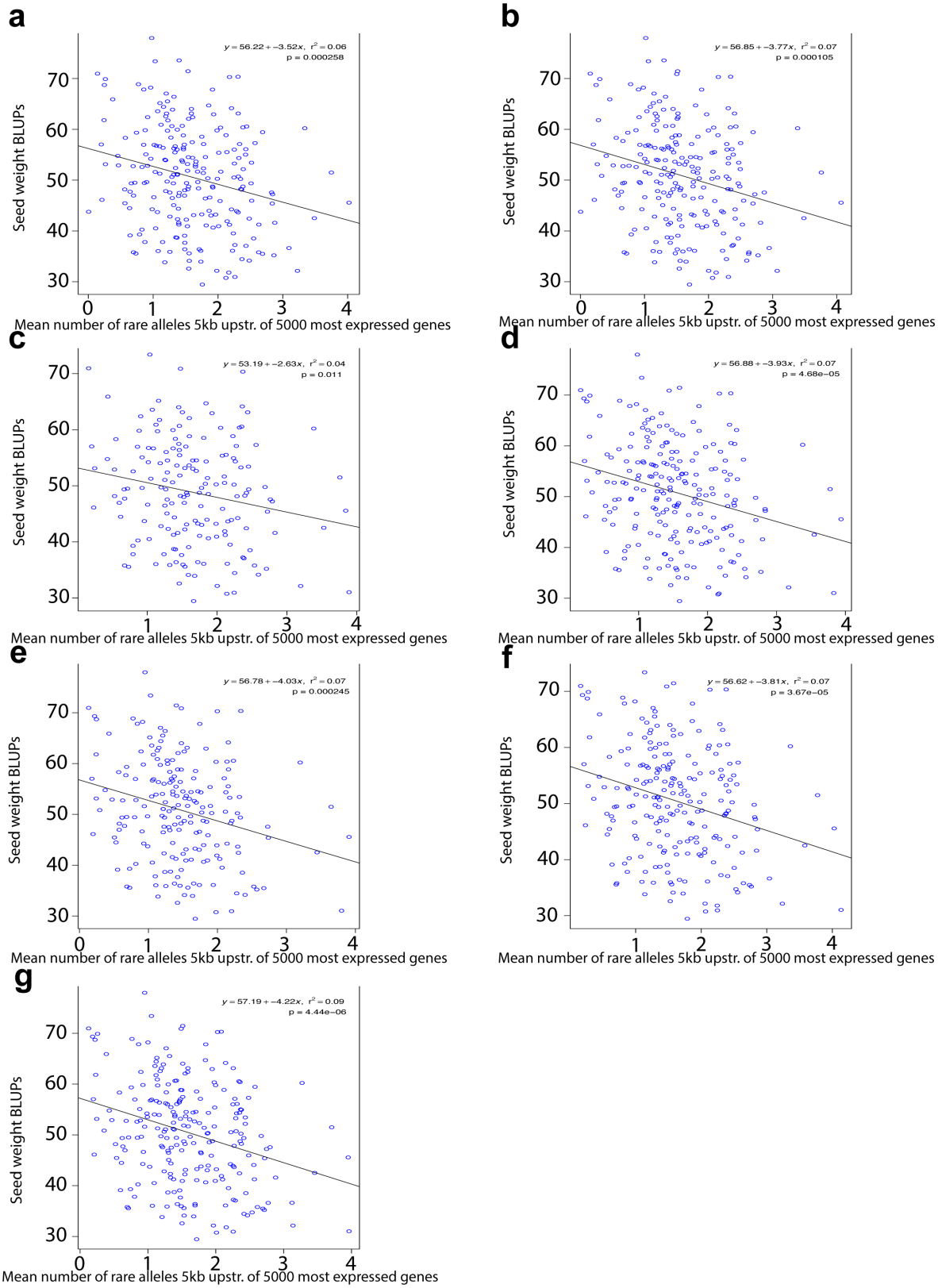
Extended Data Figure 6 | eQTL R^2 distribution comparisons between SNPs in 0.0–0.1 (tropical MAF) and 0.4–0.5 (RNA-set MAF) versus 0.4–0.5 (RNA-set and tropical MAF). a, Adult leaves collected at night ($n = 260$ unique inbred samples). **b**, Adult leaves collected during the day ($n = 204$ unique inbred samples). **c**, Tip of leaf three ($n = 265$ unique inbred samples). **d**, Base of leaf three ($n = 263$ unique inbred samples). **e**, Kernels at 350-growing-degree days ($n = 229$ unique inbred samples). **f**, Shoots of germinating seedling ($n = 278$ unique inbred samples). **g**, Roots of germinating seedling ($n = 273$ unique inbred samples). All pairs of distributions within each tissue are significantly different. $P < 2.2 \times 10^{-16}$ two-sided Wilcoxon signed-rank test and Kolmogorov–Smirnov test.



Extended Data Figure 7 | Expression value and dysregulation of 5,000 most-expressed genes are both predictive of fitness. Orange boxes represent correlations between predicted and true seed weight when using expression values. Yellow boxes represent correlations between predicted and true seed weight when using absolute deviation in expression from the population mean. Range of correlations between predicted and true seed weight is displayed from ten repetitions of nested tenfold cross validation (ten inner and ten outer) using ridge regression. In the box plots, the middle horizontal lines represent the median, hinges represent the 25th and 75th percentiles (the interquartile range), the upper and lower whiskers extend to maximum and minimum points no more than 1.5× interquartile range beyond the hinges, and individual dots are outliers beyond the whiskers. Sample sizes: 2-cm root tips of germinating seedlings (unique $n = 181$) and whole shoots of germinating seedlings (unique $n = 183$); the 2-cm base (unique $n = 181$) and tip (unique $n = 182$) of leaf 3; leaves collected in the field during the day (unique $n = 135$) and night (unique $n = 187$); and 350-growing-degree-day kernels (unique $n = 171$), post sexual maturity (anthesis).



Extended Data Figure 8 | Cumulative expression dysregulation of the 5,000 most-expressed genes in each tissue versus seed weight. a, Adult leaves collected at night ($n = 221$ unique inbred samples). **b**, Adult leaves collected during the day ($n = 171$ unique inbred samples). **c**, Tip of leaf three ($n = 226$ unique inbred samples). **d**, Base of leaf three ($n = 224$ unique inbred samples). **e**, Kernels at 350-growing-degree days ($n = 195$ unique inbred samples). **f**, Shoots of germinating seedling ($n = 235$ unique inbred samples). **g**, Roots of germinating seedling ($n = 226$ unique inbred samples). Regression statistics in Extended Data Table 1. Sweet corn and popcorn lines were excluded from these regressions.



Extended Data Figure 9 | Mean upstream rare-allele count from the 5,000 most highly expressed genes versus seed weight. a, Adult leaves collected at night ($n = 221$ unique inbred samples). **b**, Adult leaves collected during the day ($n = 171$ unique inbred samples). **c**, Tip of leaf three ($n = 226$ unique inbred samples). **d**, Base of leaf three ($n = 224$ unique inbred samples). **e**, Kernels at 350-growing-degree days ($n = 195$ unique inbred samples). **f**, Shoots of germinating seedling ($n = 235$ unique inbred samples). **g**, Roots of germinating seedling ($n = 226$ unique inbred samples).

Extended Data Table 1 | Regression statistics for cumulative expression dysregulation in each tissue against seed-weight fitness

Tissue	P-value	r	n
Kernel	0.00046	-0.242	195
Adult Leaves - day	0.03886	-0.154	171
Germinating Shoot	0.42226	-0.052	235
Adult Leaves - night	0.20301	-0.084	221
Germinating root	0.29569	0.069	226
Leaf 3 base	0.66049	-0.029	224
Leaf 3 tip	0.77367	0.019	226

Sample size *n* refers to genetically unique inbred samples after excluding sweet corn and popcorn lines.

nature research

Corresponding author(s): Karl Kremling & Edward Buckler

☐ Initial submission ☐ Revised version ☒ Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

Describe how sample size was determined.

Sample size (mean 255 individuals per tissue) was chosen based on the availability of lines with WGS data. Para. 3-4. Additionally, previous studies demonstrated the ability to map eQTL with < 300 individuals (see GTEx).

2. Data exclusions

Describe any data exclusions.

To eliminate samples which were potentially collected into the wrong tube(errors), RNAseq data were used to call SNPs from the first 20 mb of chr 10 and these SNPs were matched to existing WGS DNAseq variant calls in the Maize HapMap. Expression profiles from the middle of leaf three were also excluded because the sample size for that tissue was below 50.

3. Replication

Describe whether the experimental findings were reliably reproduced.

The results were replicated across 7 tissues which were collected and sequenced separately. Additionally, during revisions independent expression data from a previous publication (Hirsch et al PlantCell 2014) was also used to replicate the results linking rare alleles to dysregulation by quadratic regression.

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

Field grown plants were planted in 4 randomized blocks based on maturity dates so that collection would coincide. Growth chamber and greenhouse grown plants were completely randomized and rotated each day.

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

The investigators were not blinded to the maize lines being used nor the tissues being collected.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒ A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ A statement indicating how many times each experiment was replicated
☐	☒ The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
☐	☒ A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒ The test results (e.g. *P* values) given as exact values whenever possible and with confidence intervals noted
☐	☒ A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒ Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.	Robotic code, Java code, and custom R and Unix scripts are released through the following websites as described at the end of the paper and from the authors: http://www.maizegenetics.net/robotic-code http://www.maizegenetics.net/tassel https://github.com/Fei-Lu/FastCall

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.	All lab materials are available from Lexogen GMBH and Beckman Coulter. Germplasm for the maize lines in the study is available from the North Central Regional Plant Introduction Station.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).	NA

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.	NA
b. Describe the method of cell line authentication used.	NA
c. Report whether the cell lines were tested for mycoplasma contamination.	NA
d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.	NA

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.	NA

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.	NA