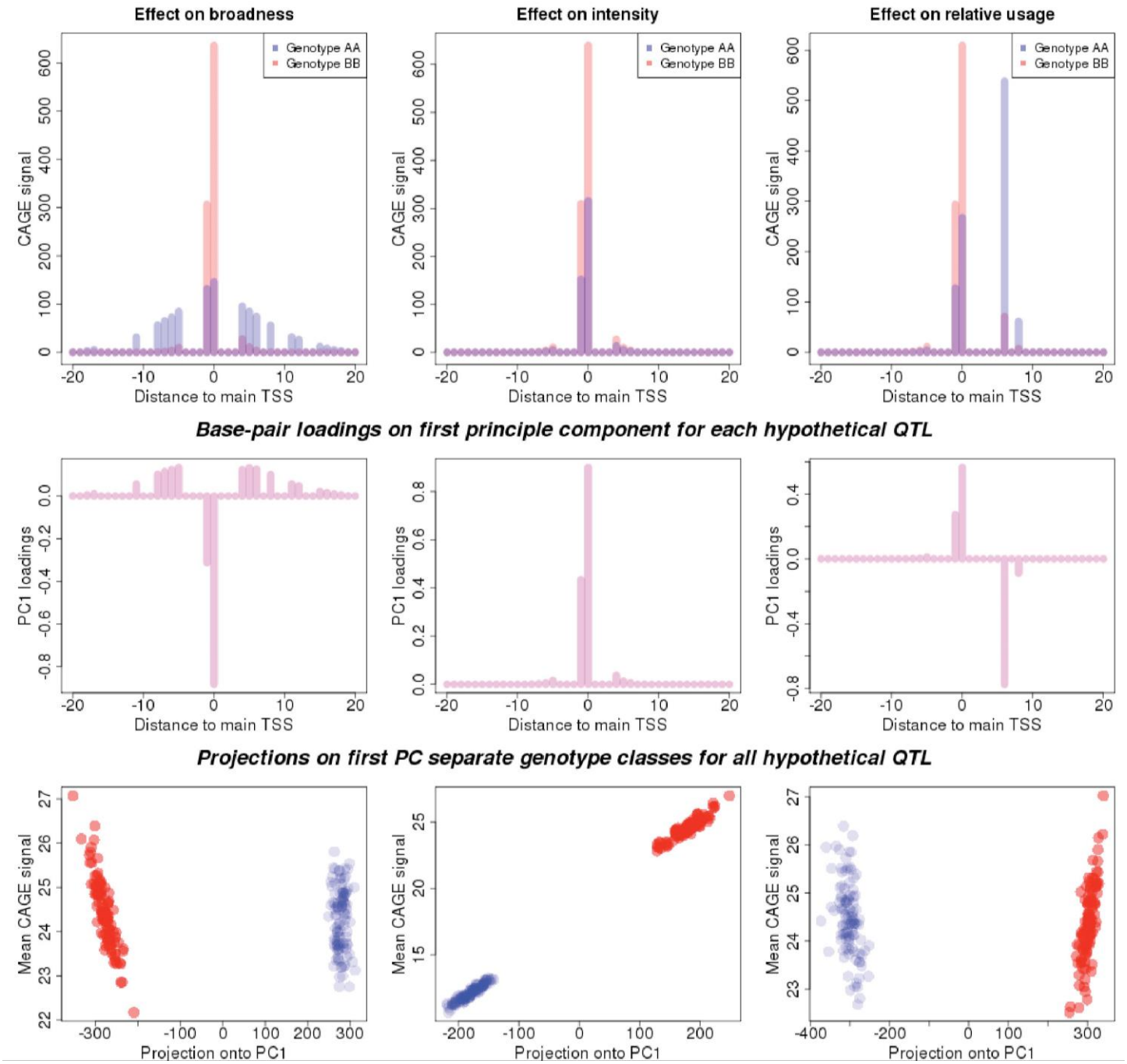
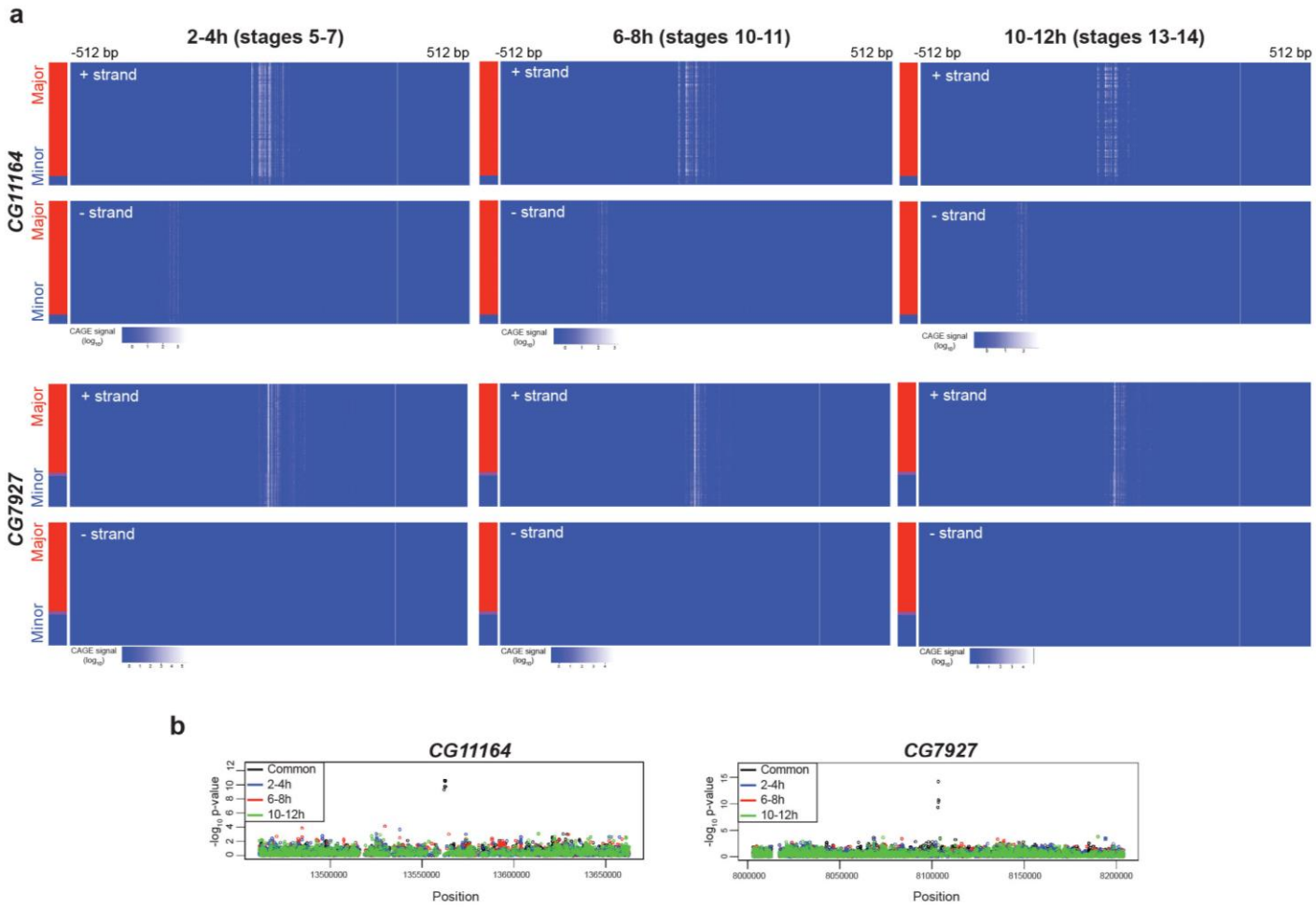


Supplementary Figure 1

Complex molecular phenotypes captured by the PCA-based QTL-calling approach.

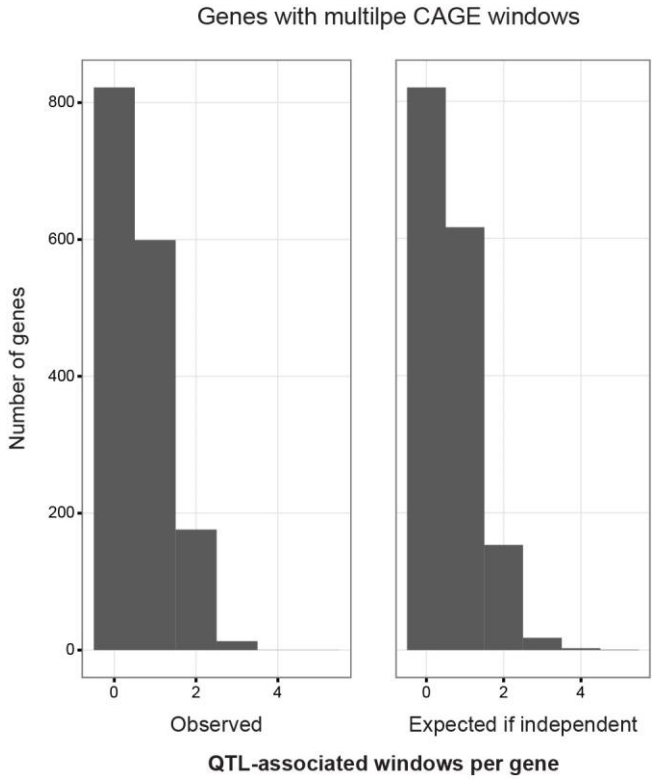
Top, three synthetic examples of QTL illustrating three aspects of CAGE signal - shape, level, and position of the highest peak within the cluster. Middle, loadings for the first principle component when genotypes are simulated from these hypothetical QTL. Bottom, separation of hypothetical genotype classes based on either projections onto the top principle component (y-axis) or average CAGE signal across entire window (x-axis). While genotype classes in the middle column can be distinguished by both the mean signal and the projections on the first PC, genotypes in the left and right columns are separated by the first PC but not by the mean signal.



Supplementary Figure 2

tssQTL generally have effects that are independent of developmental stage.

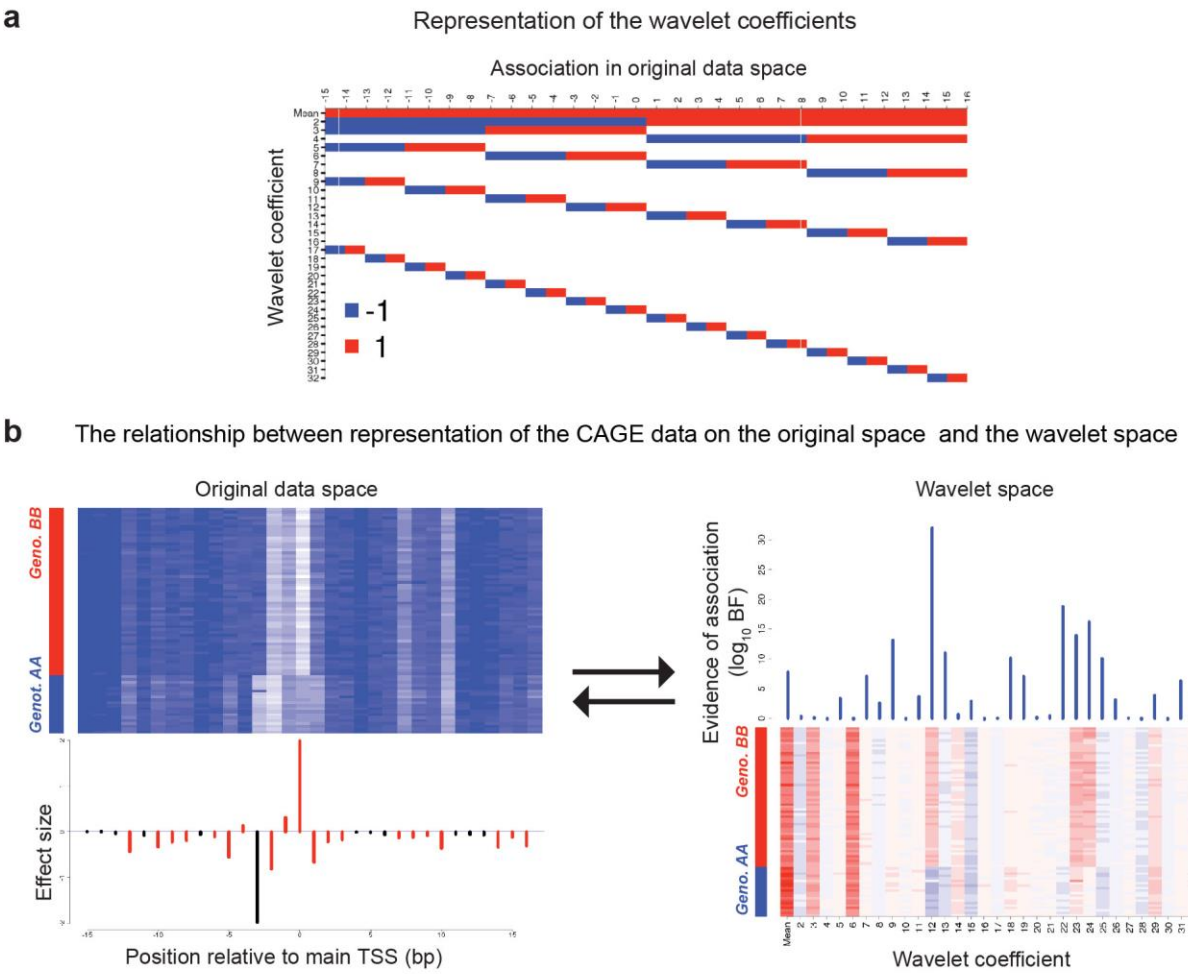
Shown are two representative tssQTL with effects at all three embryonic stages. **a,** Heatmaps of CAGE signal within a promoter window, with individual genotypes (rows) grouped into major (red, above) and minor (blue, below), for the same examples shown in Fig. 1D. Lead variants are chrX_13562404_SNP (*CG11164*) and chr3L_8103470_SNP (*CG7927*). **b,** Manhattan plots, showing the $\log_{10}$ (*p*-values) for all tested variants in each case. Significant associations tend to be common to all there developmental time-points and not stage-specific (results of time-specific effect tests are on Supplementary Tables 2 and 3).



Supplementary Figure 3

tssQTL in genes with multiple promoters.

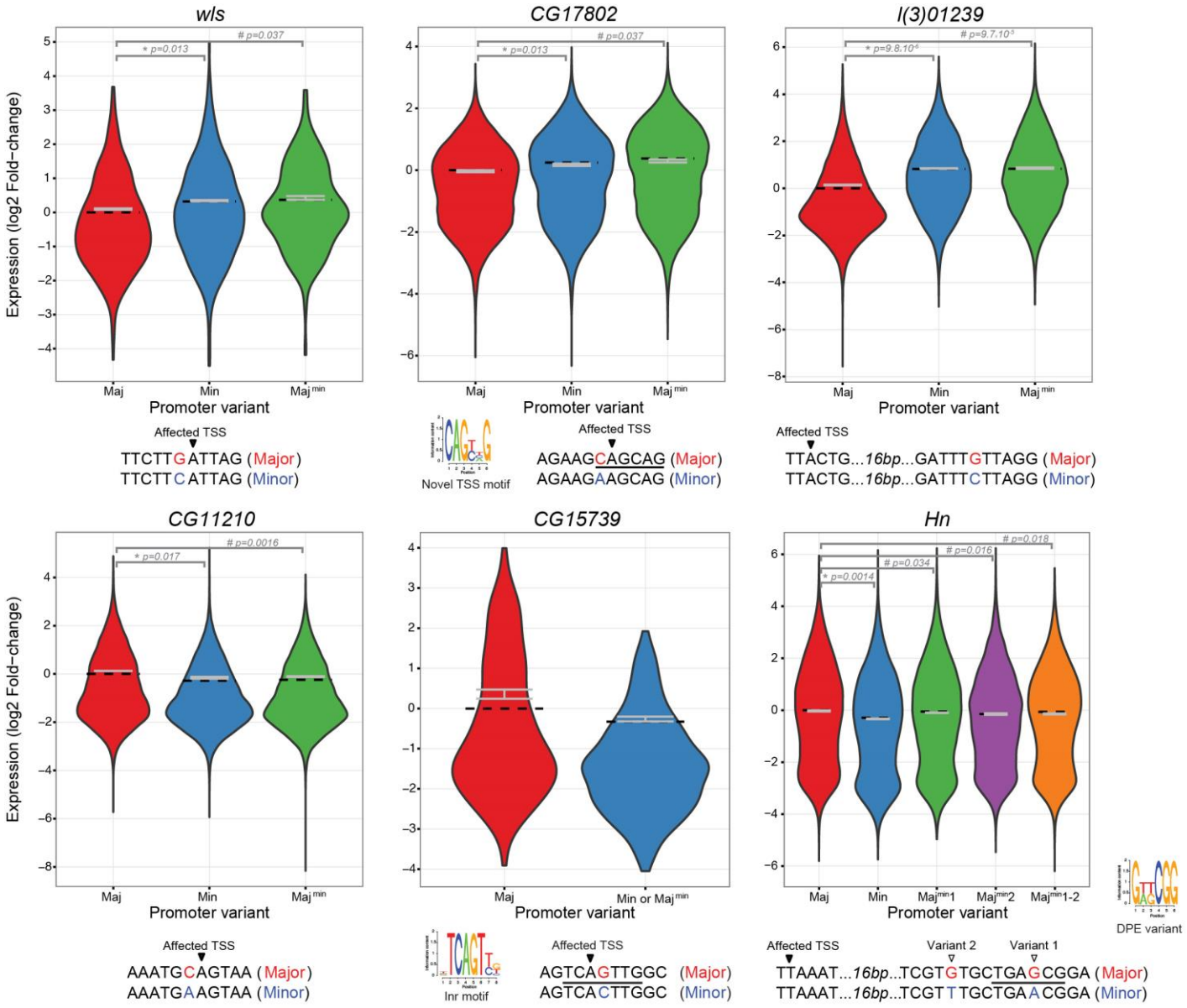
Left, histograms showing the number of QTL-associated promoters per gene, for 1,610 genes with more than one associated CAGE window (varying between 2 and 8 per gene). *Right*, expected number of QTL-associated promoters per gene, assuming that the probability of tssQTL across promoters is uniform (no dependence within a gene). The observed and predicted distributions are not significantly different (p = 0.9986, two-sided Kolmogorov-Smirnov test).



Supplementary Figure 4

Wavelet analysis.

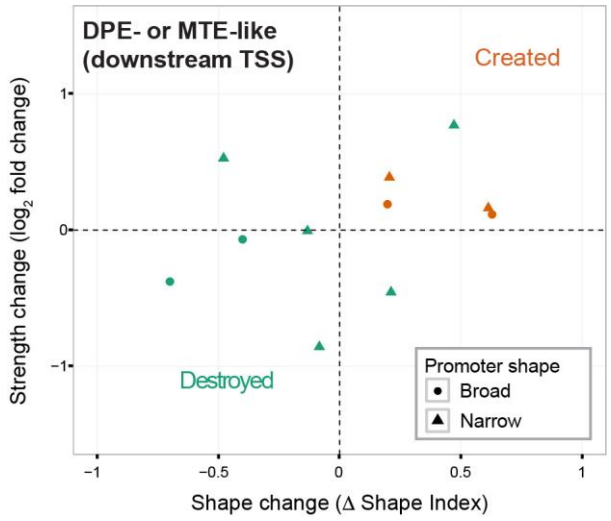
a, Representation of different wavelet coefficients in an example 30 bp window. **b,** Detailed relationship between CAGE signal (white in heatmap above) and single-base effect sizes (below) in the genomic space (*left*) and the evidence of effect on the different wavelet coefficients (*right)*, for the *CG7927* example (included in Fig. 2a). Significant effects are shown in red.



Supplementary Figure 5

Validation of lead tssQTL variants.

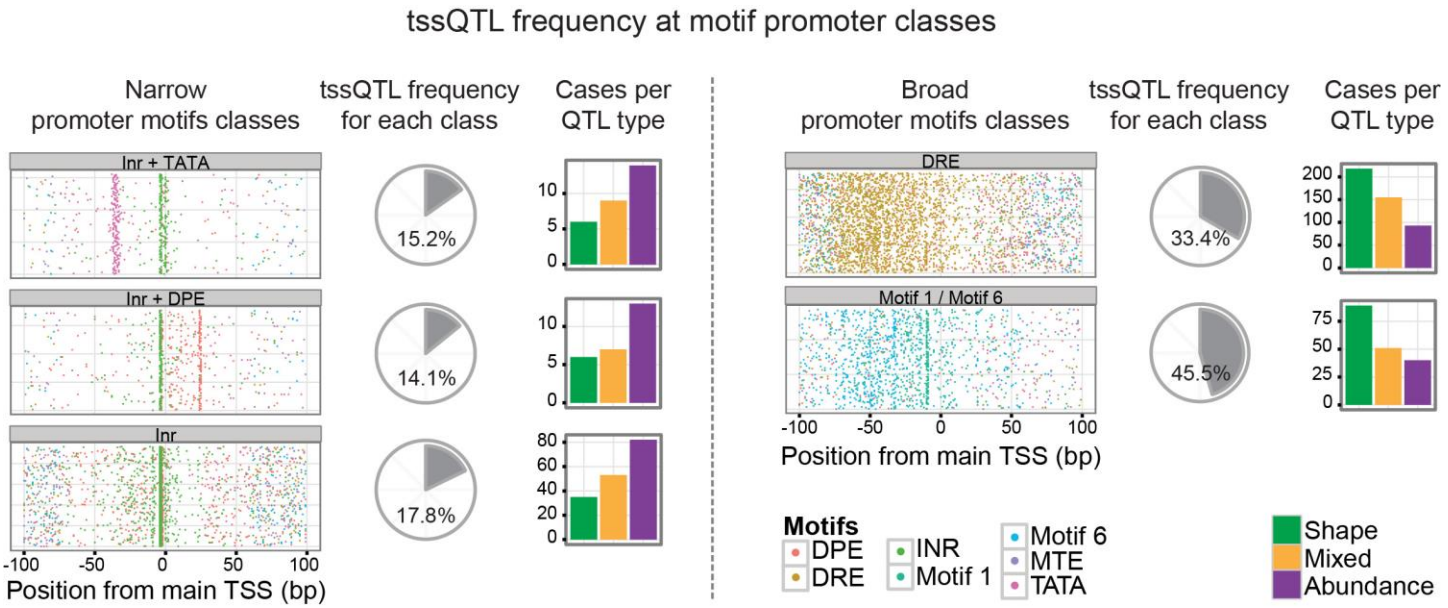
The distribution of single-cell expression values for tssQTL, showing promoter strength differences between Maj, Min and Maj^{min} promoter variants, extending those shown in Fig. 4b. Promoter single cell assay quantify promoter activation of sfGFP in individual measured by analytic FACS. We show the distribution of expression values (sfGFP/mCherry ratios) for both natural promoter variants (Maj and Min) and engineered promoters, placing the lead SNP into the Maj haplotype (Maj^{min}). Dashed black line: population median. Grey error bars: average expression value +/- SEM, calculated from 2 (*wls*, *CG15739* and *Hn* (Min, Maj^{min}1 and Maj^{min}12)) or 3 (*CG17802*, *l(3)01239*, *CG11210* and *Hn* (Maj and Maj^{min}2)) transfection replicates. * *p*-values for two-sided Welch t-test comparing population averages. # *p*-values from one-sided Welch t-test. Lead variants are: chr3L_ 11162703_SNP (*wls*), chr3R_13630942_SNP (*CG17802*), chr3L_11067477_SNP (*l(3)01239*), chr2R_3986212_SNP (*CG11210*), chrX_11767477_SNP (*CG15739*), chr3L_ 7753664_SNP (*Hn* variant 1) and chr3L_ 7753657_SNP (*Hn* variant 2).



Supplementary Figure 6

Effects of changes in positioned downstream motifs on promoter shape and strength.

Change in raw CAGE signal and shape index between the minor and the major genotypes for tssQTL affecting DPE- (downstream promoter element) and MTE- (motif ten element) like motifs located 10-40bp downstream of the highest effected TSS. *Created* or *destroyed* indicates a motif score that is higher or lower in the minor allele variant, respectively. Symbols indicate aggregate promoter shape (see Methods).



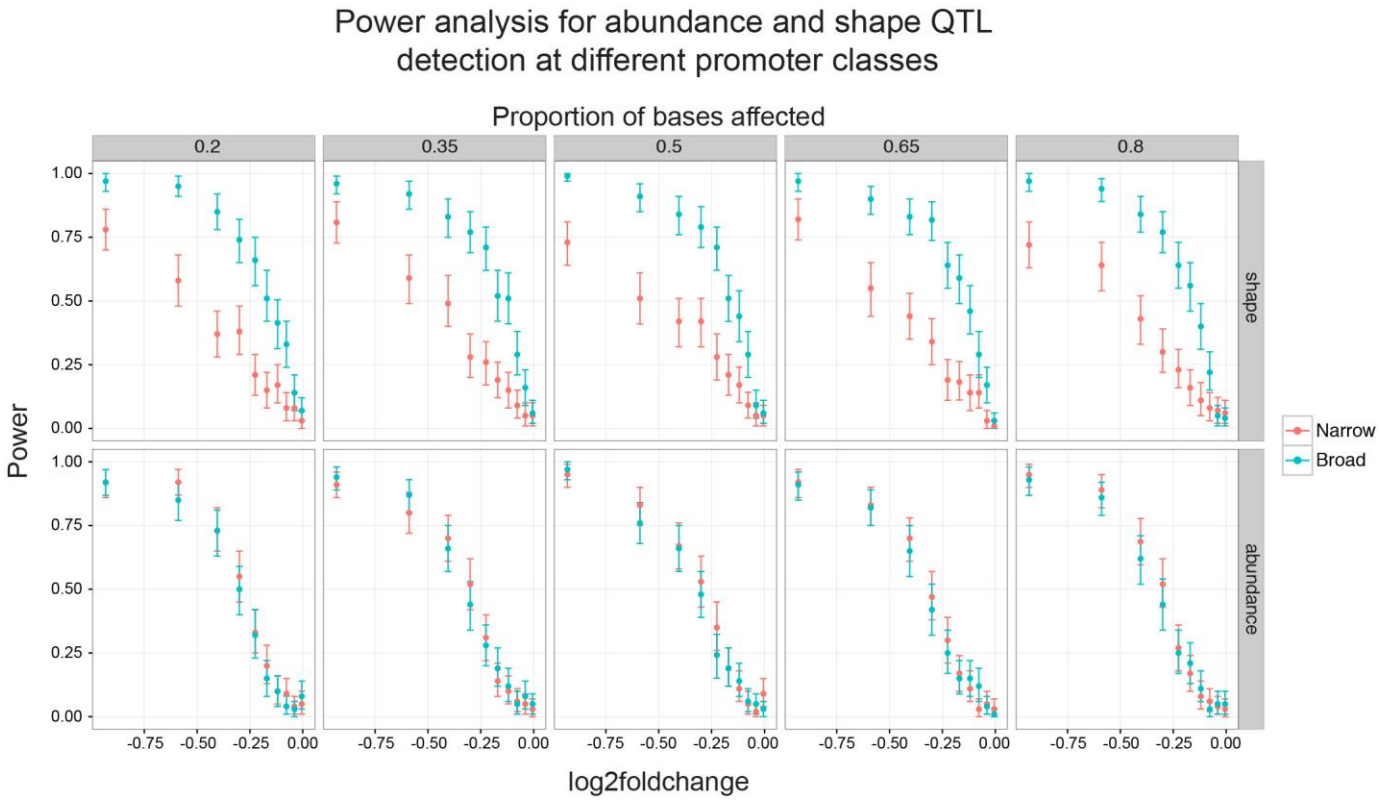
Supplementary Figure 7

Motif-based classification of promoters and their relationship to tssQTL.

We used Ohler *et al* proposed classification of *Drosophila* promoters based on five motif-based classes[1,2]: Inr+TATA, Inr+DPE, Inr-only, characteristic of narrow promoters and DRE and Motif1/6 motifs, characteristic of broad promoters. For each promoter class (defined by this motif classification[1]), the following is shown: *left,* positioning of the indicated motifs from the main TSS; *middle,* fraction of promoters with a tssQTL; *right,* the relative proportion of tssQTL classes. The classes corresponding to broad promoters (indicated at the bottom) show elevated tssQTL number (2.84-fold increase in the broad classes compared to narrow, $p = 2.2 \times 10^{-16}$, Fisher's exact test) and a higher proportion of shape QTL, recapitulating the differences observed between broad and narrow promoters classified based on their shape index (Fig. 5a).

1. Ohler, U. Identification of core promoter modules in Drosophila and their application in accurate transcription start site prediction. *Nucleic acids research* **34**, 5943-5950, doi:10.1093/nar/gkl608 (2006).

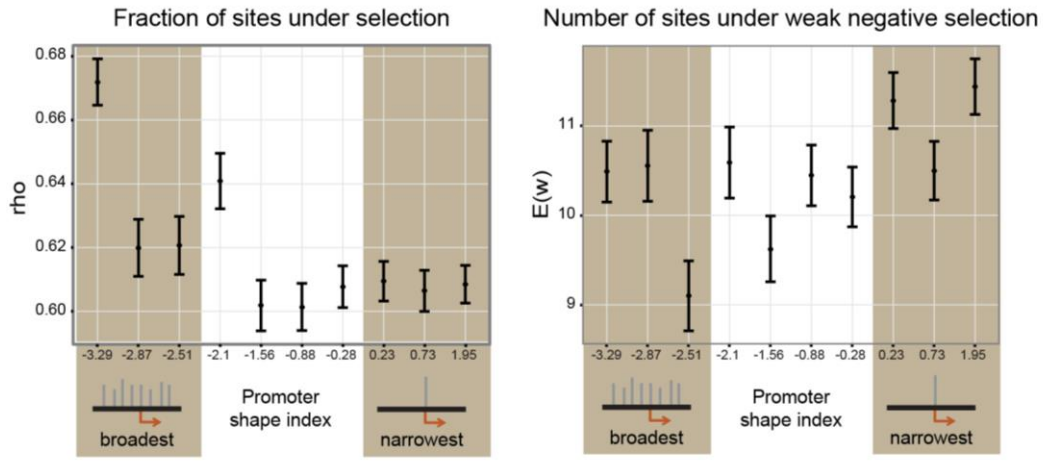
2. Ohler, U., Liao, G.-c. C., Niemann, H. & Rubin, G. M. Computational analysis of core promoters in the Drosophila genome. *Genome biology* **3**, Research0087 (2002).



Supplementary Figure 8

Estimated detection power of abundance and shape QTL in different promoter shape classes.

Simulation study to assess statistical power for detecting abundance or shape QTL in broad and narrow promoters and for different effect sizes. Given the small number of TSS within a narrow promoter, it is expected that detection of shape QTL within them may be inherently more difficult, potentially leading to their underestimation. Signal distribution, minor allele frequency and library sizes were simulated to match those observed in real data. *Top panel:* simulated shape QTL, where the increasing proportions (indicated on top) of the main peak were redistributed in the second genotype group. *Bottom panel:* Simulated abundance QTL, where increasing proportions of bases within promoter windows (indicated on top) are affected in the same direction with increasing fold-change effect (x axis). After generating the simulated counts (using a Poisson distribution), the simulated CAGE profiles were processed analogous to the approach taken for the real data, including normalization and principal component analysis to obtain the projections onto the first three PCs. Significance was assessed using a t-test for each PC to test differences between genotype classes. The proportion (error bars indicate 95% CI) of simulated windows for which significant differences could be detected (power at $p < 0.05$) is indicated in the y-axis.

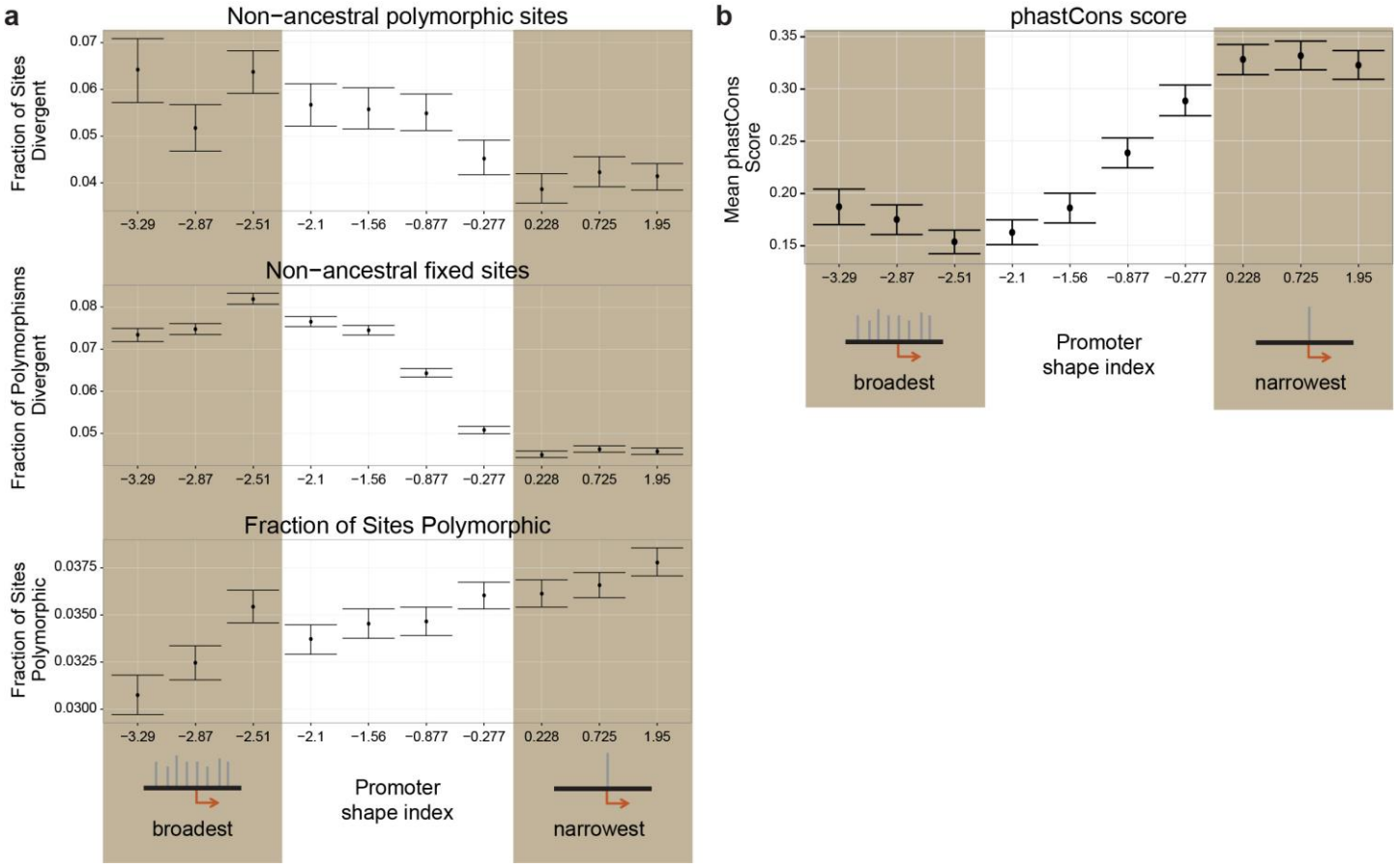


Supplementary Figure 9

Number of bases under negative selection at different promoter shapes.

INSIGHT analysis on bins of promoter windows according to their shape indexes. Brown shading indicates the 30% broadest or narrowest promoters. rho = fraction of sites under negative selection, E(w) = the number of sites under weak negative selection. All windows have the same size (1,024 bp). Error bars represent uncertainty estimates for INSIGHT parameter estimates (standard errors), which are derived from the curvature of INSIGHT's likelihood function[1].

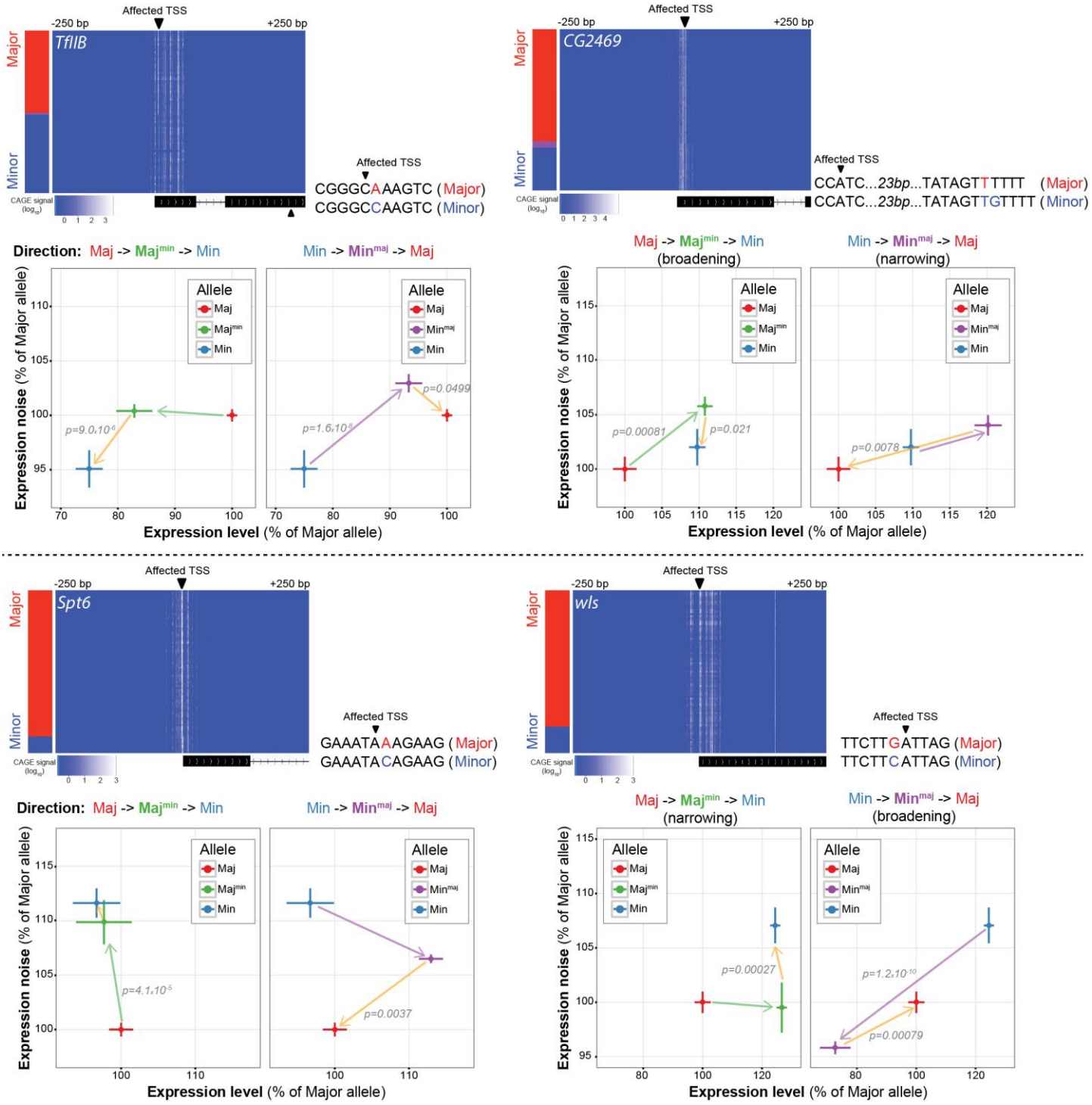
1. Gronau, I., Arbiza, L., Mohammed, J. & Siepel, A. Inference of natural selection from interspersed genomic elements based on polymorphism and divergence. *Molecular biology and evolution* **30**, 1159-1171, doi:10.1093/molbev/mst019 (2013).



Supplementary Figure 10

Patterns of genetic variation at different promoter shapes.

a, Broad promoters show an increase in the fraction of non-ancestral sites (y-axis) for promoters with different shape index (x-axis). Brown shading indicates the top 30% for each shape type. Both the number of polymorphic sites where none of the alleles is the inferred ancestral allele (top), and the number of fixed (monomorphic) sites diverging from the ancestral state (middle) show a clear dependence on promoter shape, with broadest promoters having elevated rates for the non-ancestral states. In contrast, polymorphism rates are equal or reduced in broader shape bins compared to narrow (bottom), arguing against elevated substitution rates being strictly a function of differences in the fraction of functional bases. **b,** Average phastCons scores for windows of the indicated shape indexes show evidence for increased conservation in narrow promoter categories. Scores where obtained from the UCSC Genome Browser. In all cases, error bars correspond to 95% CI.



Supplementary Figure 11

Examples of changes in expression level and noise due to genetic variants affecting promoter shape.

Promoter single cell assay using analytic FACS (as in Fig. 4a). Population median indicates expression level, while median absolute deviation from the median indicates cell-to-cell variation (expression noise). Four promoters harboring shape or mixed QTL with a predominant shape effect (extension of Fig. 6b,c). Lead variants are chr2L_10431345_SNP (*TfIIB*), chr3L_ 1307602_INS (*CG2469*), chrX_ 6161845_SNP (*Spt6*) and chr3L_11162703_SNP (*wls*). Expression level is the mean of the population medians, and noise is the mean of population MADs (3 transfection replicates, >15,000 cells per construct), with bars corresponding to SEM. Green and purple

arrows indicate the effect of the lead variant mutation alone, while orange arrows indicate the effect of changing all remaining variants in the transition to the opposite natural haplotype. Significance was assessed by a Levene test for homogeneity of variances, only significant *p*-values (< 0.05) are shown.

Promoter shape varies across populations and impacts promoter evolution and expression noise

Ignacio E. Schor, Jacob F. Degner, Dermot Harnett, Enrico Cannavò, Francesco P. Casale, Heejung Shim, David Garfield, Ewan Birney, Matthew Stephens, Oliver Stegle[†] & Eileen E. Furlong[†]

Nature Genetics: doi:10.1038/ng.3791

SUPPLEMENTARY NOTE

EXTENDED METHODS

Embryo collections and RNA extraction

Freshly eclosed adults were placed in embryo collection vials with standard apple cap plates. After at least three 1h pre-lays, the flies were allowed to lay for 2 h, after which the embryos were aged to the appropriate time-point. The embryos were then dechorionated using 50% bleach, washed alternately with water and PBS + 0.1% Triton X-100, and then snap-frozen in liquid nitrogen. For RNA isolation, a pool of ~100 embryos was homogenized in TRIzol®LS (Life Technologies) with Cordless Motor for Pellet Mix and pestels (VWR) in ice. RNA was extracted according to manufacturer instructions, digested with RNase-free DNase I (Roche) for 30 min and purified a second time using the RNeasy mini kit (QIAGEN).

5' CAGE library preparation

Samples were prepared in batches of 20-40 and therefore performed most steps in 96-well plates, cut into quarter or halves (twin.tec PCR plate 96, unskirted, Eppendorf). Reverse transcription, diol oxidation, biotinylation of the RNA diols and RNase treatment (and all the intermediate cDNA purification steps) were done as stated in the Carninci *et al*. protocol[1]. Biotinylation was performed overnight at 23°C.

For the first biotin capture (biotinylated 5' caps), 20 µl of Dynabeads® MyOne™ Streptavidin C1 (Life Technologies) was used per sample. Typically, we pooled batches of beads corresponding to 10 samples (200 µl of resuspended beads) in a LoBind 1.5 ml tube (Eppendorf) and washed them using 1 ml of 1x B&W buffer, 200 µl of solution A and 500 µl of solution B (all recipes provided by the Dynabeads manufacturer). For all dynabeads washing steps, the solution was added, the tube (or plate) was placed in a magnetic rack (Life Technologies) and then turned 4 times to force the beads to enter in solution. The beads were allowed to attach to the wall facing the magnet for 2-3 min and the solution removed. After initial washing of the beads, they were resuspended in 1 ml of solution B with 15 µl of 20 µg/µl tRNA (1.5 µl per sample) and incubated for 30-60 min at room temperature with agitation, in order for the tRNA to coat the beads, which reduces non-specific binding. The beads for 10 samples were then washed twice with 500 µl of wash buffer 1 (recipe from the Carninci protocol[1]), resuspended in 850 µl of the same buffer, and then 80 µl aliquots were mixed with the purified RNA/cDNA duplexes. The plate

Nature Genetics: doi:10.1038/ng.3791

containing the samples was incubated for 30 min at room temperature in a shaker at 200 rpm. After the first 15 min, samples were re-mixed by pipetting. Then we continued with the original protocol for the washes and elution.

After ligation of barcoded biotinylated 5' linkers, 10 samples were pooled together during the two-step purification, as suggested in the original protocol. Once digested, the resulting pools of fragments were ligated with the 3' linker, and purified using the biotin on the 5' linker. For this second biotin capture step, 10 μl of resuspended Dynabeads MyOne C1 were used for each pool of 10 samples.

After washing, beads were resuspended in 20 μl of water and the PCR was performed on beads.

Processing of 5' CAGE reads

Reads were demultiplexed and the 27nt-long tags were extracted between nucleotides 13 and 39 (first 6 bases correspond to the barcode and next 6 bases to the EcoP15I restriction site). The Carninci's lab reported that a G (rather than the expected A) is often observed as the first nucleotide in CAGE data, and reasoned that this is most likely due the addition of a non-template G by the reverse transcriptase[1]. We observed a similar finding in our data. We therefore trimmed the first base (G at position 13), resulting in 26nt-long CAGE tags, which also slightly increased the percentage of successfully mapped reads. The 26nt tags from the 81 different DGRP lines were mapped to the *Drosophila melanogaster* reference genome (dm3 assembly), using BWA version 0.6.1-r104, allowing for up to 1 mismatch (bwa aln parameter n=1). Mapped reads were filtered to only retain reads that uniquely map to a single location in the genome and have an overall mapping quality score > 10 (samtools view parameter –q 10). Finally, bam files were sorted and indexed for rapid access.

Verifying genotype identity from raw CAGE reads

To verify that the identity of each sample was correct (prior to genetic analyses), the sequence reads for each independent library was compared to the published Freeze 2.0 genotypes generated by the DGRP consortium[2], as follows: For each CAGE library, we extracted the base-call at every position within all reads overlapping SNP positions. For each of the 81 DGRP lines, we calculated a genotype concordance statistics, based on the fraction of CAGE reads that were consistent with the genotype of the line. The distribution of this statistic was bimodal with one mode representing matched individual to library comparisons and the second mode representing miss-matched samples. The 6 miss-matched samples (out of 282) were discarded and the sample prepared again.

Nature Genetics: doi:10.1038/ng.3791

Creating a Universal Mappability Map (UMM) for the DGRP lines

Sequence variation between individuals or lines can induce variation in the ability to correctly map short sequence reads back to their location in a reference genome. Such sequence-specific biases in mapping can lead to artificial associations between genotype and any functional genomics measurements made with short read sequences[3]. A very effective method for addressing these biases is to identify the genomic locations for which different genotypes have different mappabilities and remove these genomic locations from the analysis[3-5]. To address this potential bias, we created a genome-wide mappability map of each DGRP line for each position in the genome and used this to pre-filter regions that are likely to introduce artificial associations: Raw sequencing reads from the DGRP genome-sequencing project (75 bp paired end Illumina reads) were used to generate each unique 26 bp short read (matching the size of the CAGE tags) that could be obtained from each line (i.e., each 75 bp read was chopped into 49 overlapping 26 bp reads). These shortened reads were mapped to the dm3 reference genome and the number of reads starting at each position was counted for each line. From this alignment map, a Universal Mappability Map (UMM) was created by considering genomic positions for which each DGRP line contained at least one read that mapped to that location. Finally, a summary of this map was created such that there were mapped reads represented at that position for all DGRP lines with modest sequencing depth (zero read counts were ignored if they only occurred for a single DGRP line and that DGRP line had an average sequencing depth < 10X). This thereby removes all genomic positions that failed to uniquely map to DNA-seq reads in even one of the 81 lines, which should provide a conservative method to reduce mapping biases and trivial associations with indels segregating within the population. On average, each 1kb window has 77 bases filtered out by the UMM, corresponding to 3.7% of all sites. The efficiency of the UMM filter also manifests in the tssQTL set, where the percentage of lead variants being an indel (11%) is lower compared to the tested variant set as a whole (18%). This is expected given the impact of indels on mappability.

The use of such a stringent method for eliminating mapping biases allowed us to systematically sample the functional affect of all variants around the TSS, even considering variants that fall into the CAGE read itself: While the lead variants of the vast majority of tssQTL lie outside the CAGE tag, 785 variants (19% of all tssQTL) fall within the 26nt region (length of CAGE read) downstream the most affected TSS, and yet can be mapped successfully and also, very importantly, their functional differences in promoter activity can be experimentally validated (both examples shown in Fig. 4 and two examples in supplementary Fig. 5 (*l(3)0139* and *CG15739*) have variants in the CAGE read).

Nature Genetics: doi:10.1038/ng.3791

Application of WASP for additional mappability filtering

While the UMM based approach has the advantage that it does not require accurate variant calls to identify potentially biased regions of the genome, it does rely on the DGRP lines being fully homozygous. When lines have residual heterozygosity at a variant location, the UMM may miss some read mapping bias. For example, when two alleles are segregating in a line and only one of these alleles is mappable, the UMM might consider the location mappable when it is only mappable for a subset of the reads arising from that location. To address this potential second source of bias, we additionally filtered all bam files for mappability using the WASP pipeline[5] using all DGRP variants for variant mapping simulations.

Molecular phenotypes for CAGE tssQTL analysis

For each of the 13,508 active promoter windows, described above, we aggregated the CAGE signal in two different ways (PC-based and mean based).

Mean based phenotypes: For every base contained in the active promoter-phenotype window (excluding those determined to be unmappable as described above), we derived the single-base pair read coverage by counting the number of CAGE reads that start at a given position. This was applied separately for each CAGE library. When multiple CAGE libraries exist for the same individual line at the same time-point (Supplementary Table 6), the resulting coverage matrices were added together. These single-base pair count matrices were then normalized by the total sequencing depth for that individual/time-point combination. To estimate total promoter abundance levels, the normalized read counts across single bases within a given promoter window were aggregated, and normalized by the length of the window (average coverage per base). These promoter-level abundance estimates for each promoter window and time-point combinations were quantile normalized to a standard normal distribution (across lines, to strictly adhere to the assumptions of our statistical model). This procedure resulted in a single measurement for each promoter window and time-point/individual combination, which were used as phenotypic traits to test for associations in a multi-phenotype linear mixed model framework, as described below.

***PC-based phenotypes*:** Analogous to the approach above, the starting point was a coverage matrix for each promoter window, derived from the number of CAGE reads that start at a given position. This was applied separately for each CAGE library, again summing across multiple CAGE libraries exist for the same individual line and timepoint, where multiple libraries exist. Analogous to the mean approach, the resulting coverage matrices were normalized by the total sequencing depth for that individual/timpoint

Nature Genetics: doi:10.1038/ng.3791

combination. To limit the impact of extreme values and to approximately variance stabilize these coverage counts, the normalized base-level read counts were then adjusted using the Anscombe transformation (essentially a square root function). For each promoter window, we constructed a local coverage matrix where each row corresponds to an individual/timepoint combination and each column is a single base-pair in the 1kb window. This matrix was then decomposed using principal component analysis (PCA), thereby reducing the number of bases (of 1,024 bp windows) to a smaller number of components. Intuitively, the top PCs derived using this approach capture, in a ~1,000 dimensional space, the direction in which there is the most variation in CAGE signal across lines and developmental stages. A change in the direction of one PC then corresponds to a combination of changes in the raw CAGE signal across the 1,024 bp window. For example, an increase in CAGE signal in one base may correspond to the same movement along PC1 as a decrease in CAGE signal at an adjacent base. The first PCs thereby represent the directions in which the 1,024 bases making up the window tend to co-vary across lines. Importantly, projections onto these PCs can detect both changes in total TSS usage (transcript abundance) as well as changes in promoter shape (Fig 1b, and Supplementary Fig. 1). For the genetic analyses, we considered the leading 3 principal components as a phenotypic trait, which were again quantile normalized to a standard normal distribution prior to genetic analysis.

This procedure results in 3 measurements for each individual/timepoint combination (or 9 total phenotypes per *Drosophila* line). These were tested for association with genotypes using a multi-phenotype modelling framework, as described below.

Adjusting for hidden covariates using PEER

Before QTL calling, we used PEER[6] to adjust for hidden structure in the expression data. PEER was applied separately to the data from each of the 3 developmental stages (mean based phenotypes) or the 9 phenotypes derived from PCs (PC-based phenotypes, 3 PCs x 3 timepoints). The PEER model adjusts for both for measured and unlabelled (hidden) confounding covariates. We fit PEER providing relevant measured covariates (collection number, library construction personnel, library prep date, barcode sequence, embryo collection date, PCR cycle number, library purification method, and sequencing batch) and 10 additional hidden (unmeasured) factors. Residuals from this model fit were taken forward as phenotypes for QTL mapping.

Nature Genetics: doi:10.1038/ng.3791

Models for testing genetic associations with CAGE signal

QTL calling was carried out using three different approaches, either considering individual time points or all three time points jointly, either using the mean CAGE signal or projections into the first 3 principal components (PCs) as phenotype.

Single-trait analysis of mean expression phenotpyes:

Briefly, denoting with N the number of lines, the N-sized phenotype vector $\boldsymbol{y}$ is modeled as sum of an intercept term, a fixed effect for the variant to be tested, a random effect accounting for genetic relatedness between lines and residual noise:

$$\boldsymbol{y} = \underbrace{\mathbf{1}_N \mu}_{\text{intercept}} + \underbrace{\boldsymbol{x}\beta}_{\text{variant effect}} + \underbrace{\boldsymbol{u}}_{\text{relatedness}} + \underbrace{\boldsymbol{\psi}}_{\text{noise}} \quad (1)$$

$$\text{where } \boldsymbol{u} \sim N\left(\mathbf{0}, \sigma_{\mathrm{g}}^2 \boldsymbol{K}\right), \boldsymbol{\psi} \sim N(\mathbf{0}, \sigma_{\mathrm{n}}^2 \boldsymbol{I}_{N\times N})$$

Here $\mathbf{1}_N$ denotes an N-sized vector of ones, μ is the mean, $\boldsymbol{x}$ is the genotype vector of the variant being tested, β its effect size, $\boldsymbol{K}$ is the realized relatedness matrix (RRM), $\boldsymbol{I}_{N\times N}$ denotes the $N\times N$ identity matrix while σ_{g}^2 and σ_{n}^2 are the genetic and the noise variance component respectively. We test for the effect of the considered genotype being different from zero ($\beta \neq 0$) using a log likelihood ratio test (LRT). This single-trait analysis was applied to each promoter window and developmental stage separately.

Multi-trait analysis of mean expression phenotypes:

Indicating with D the number of developmental stages (here $D = 3$) let $\boldsymbol{Y}$ denote the $N\times D$ phenotype matrix of normalized expression values with columns corresponding to developmental stages and rows to different lines.

As previously derived in[7,8], we extend the model in Eq. (1) to jointly analyze multiple traits (here corresponding to the different developmental stages) using a liner model of the form

$$\boldsymbol{Y} = \underbrace{\mathbf{1}_N \boldsymbol{\mu}^T}_{\text{intercept}} + \underbrace{\boldsymbol{x}\boldsymbol{\beta}^T}_{\text{variant effect}} + \underbrace{\boldsymbol{U}}_{\text{relatedness}} + \underbrace{\boldsymbol{\Psi}}_{\text{noise}}, \quad (2)$$

where $\boldsymbol{\mu}$ ($D\times 1$) and $\boldsymbol{\beta}$ ($D\times 1$) denote stage-specific mean- and genotype effects respectively and $\boldsymbol{U}$ denotes the random effect relatedness component and $\boldsymbol{\Psi}$ correlated residual noise:

$$\boldsymbol{U} \sim \mathrm{MVN}\left(\mathbf{0}, \boldsymbol{C}_{\mathrm{g}}, \boldsymbol{K}\right), \boldsymbol{\Psi} \sim \mathrm{MVN}(\mathbf{0}, \boldsymbol{C}_{\mathrm{n}}, \boldsymbol{I}_{N\times N}).$$

Nature Genetics: doi:10.1038/ng.3791

The matrix-variate normal distributions permit modeling both trait-to-trait and individual-to-individual covariances, where the trait-to-trait covariances $\boldsymbol{C}_{\mathrm{g}}$ and $\boldsymbol{C}_{\mathrm{n}}$ correspond to the genetic and residual covariance of an individual promoter window across developmental stages.

We considered the following models to test for genetic effects:

- no-effect model ($\mathcal{H}_0$: $\boldsymbol{\beta} = \mathbf{0}$): the genotype has no effect across the three developmental stages;
- common-effect model ($\mathcal{H}_c$: $\boldsymbol{\beta} = \mathbf{1}_D\beta_0$): the genotype has same effect size across the three developmental stages;
- specific-effect model in stage s_i ($\mathcal{H}_{s_i}$: $\boldsymbol{\beta} = \mathbf{1}_{\sim s_i}\beta_0 + \mathbf{1}_{s_i}\beta_i$): modeling different effect size (β_i) in stage s_i with respect to the common effect across other stages (β_0).

By comparing these different models of $\boldsymbol{\beta}$ using LRT, it is possible to test for common effects ($\mathcal{H}_c$ vs $\mathcal{H}_0$) and stage-specific genetic effects (differential effect size) for each developmental stage ($\mathcal{H}_{s_i}$ vs $\mathcal{H}_c$). The common effect test and the three specific effect tests were applied to each variant in *cis* candidate regions, resulting in distinct p-values for each of the four tests.

Multi-trait analysis using PC-based phenotypes:

When considering PC-based phenotypes we model total of 9 traits across the three developmental stages (3 PCs by 3 developmental stages).

Indicating with $\boldsymbol{Y} = [\boldsymbol{Y}_1\ \boldsymbol{Y}_2\ \boldsymbol{Y}_3]$ the $N\times P$ phenotype matrix where P ($P = 9$) denotes the number of traits and $\boldsymbol{Y}_i$ is the $N\times D$ phenotype matrix for PC i across the developmental stages, we consider the model as in Eq. (2) across all traits. We remark that $\boldsymbol{\mu}$ and $\boldsymbol{\beta}$ are $P\times 1$ vectors of trait-specific means and while $\boldsymbol{C}_{\mathrm{g}}$ and $\boldsymbol{C}_{\mathrm{n}}$ denote $P\times P$ covariance matrices that explain genetic and residual covariances across both PCs and developmental stages respectively.

Similarly to the multi-stage analysis of mean expression levels we considered different models for the genotype effect:

- no-effect model ($\mathcal{H}_0$: $\boldsymbol{\beta} = \mathbf{0}$): the genotype has no effect across all PCs and developmental stages;
- common-effect model ($\mathcal{H}_c$: $\boldsymbol{\beta} = \mathbf{1}_D \otimes \boldsymbol{\gamma}$, where $\boldsymbol{\gamma}$ is the vector of the effect sizes on the three PCs): the genotype has same effect size across the three developmental stages on each of the three PCs (effect sizes across different PCs can be different);
- specific-effect model in stage s_i ($\mathcal{H}_{s_i}$: $\boldsymbol{\beta} = \mathbf{1}_{\sim s_i} \otimes \boldsymbol{\gamma}_0 + \mathbf{1}_{s_i} \otimes \boldsymbol{\gamma}_i$, where $\boldsymbol{\gamma}_i$ and $\boldsymbol{\gamma}_0$ are the vectors of the effect sizes on the three PCs respectively in state s_i and in any other state): the genotype can have different effect size on the three PCs in stage s_i ($\boldsymbol{\gamma}_i$) with respect to the other stages ($\boldsymbol{\gamma}_0$).

Nature Genetics: doi:10.1038/ng.3791

Again, we tested for common and specific effects comparing models $\mathcal{H}_c$ vs $\mathcal{H}_0$ and $\mathcal{H}_{s_i}$ vs $\mathcal{H}_c$ by considering LRT and obtaining p-values for each of the four tests. While the multi-stage mean expression test has 1 degree of freedom (dof), the test in the PC analysis has 3 degrees of freedom (#dof = #PCs).

Assessment of developmental-stage specific effects

To explore to what extent genetic effects on transcriptional initiation are stage specific, we additionally considered stage-specific tests applied at lead variants identified using the common effect model. Here, as only one test was performed per promoter window, we used direct P-values rather than the permutation based "window-wise" phenotypes described in the previous section. Global FDRs were estimated from the distribution of P-values obtained for each time point individually. Using this targeted approach, we identified 103, 23 and 34 main effect QTL with a time point specific term in the 2-4h, 6-8h and 10-12h time points, respectively (FDR < 5%). These results strongly suggest that the tssQTL identified are markedly persistent across developmental stages.

Cage Tag Clustering Methods

To define high confidence TSS clusters for *de novo* promoter-associated motif discovery we used a method based on local smoothing, while integrating a background filter similar to that used by Hoskins *et al*[9]. This is a gene-centric approach where we assumed that reads are distributed over genes as a mixture of the signal distribution of interest, plus some flat background spread over the gene body, with mixture parameter lambda. To estimate lambda for each gene, we first chose the value minimizing the total distance between predicted noise tag values and actual values over all base pairs. We then calculated the probability of each base pair having it's current tag values under the null, and excluded all values with $p > 1x10^{-5}$. Lambda was then re estimated for each gene, after excluding sites with the new set of putative noise sites, and the procedure repeated. This process was iterated 3 times for each time point, and further iterations did not alter the estimated value of lambda.

With background noise due to highly expressed genes removed, we then derived peaks by first smoothing the data (taking the local mean over 150 bp) and then selecting regions between local minima of length greater than 150 bp. The window size 150bp was empirically selected and found to yield good separation between closely spaced cage clusters. The resulting peaks were then filtered to remove those having a read count of less than 50 across all libraries (Supplementary Table 7). The remaining 60,700

Nature Genetics: doi:10.1038/ng.3791

CAGE peaks were associated to annotated transcript models as follows: We reasoned cage peaks upstream of known genes were likely to be un-annotated peaks, while in the body of known genes were more likely to be due to re-capping or experimental artifacts. We therefore defined a zone of interest for each gene extending 250 bp downstream of annotated TSS (UCSC dm3 annotation), and up to 2 kb upstream (while not including other genes on the same strand). Peaks overlapping these regions and on the correct strand were included in our 'main' set. Also included were extragenic peaks overlapping a DNaseI hypersensitive region (using Hotspot peaks called with Hotspot from available data[10]), which may correspond to promoters of annotated genes or eRNAs. Internal peaks, overlapping gene bodies in the same strand, were excluded from the 'main' peak set.

Construction of the promoter reporter system

We designed an experimental system that can allow for measurements of expression levels in single cells, in a way that allows us to extract population averages ('promoter strength') as well as cell-to-cell variation ('promoter noise'). The method uses GFP expression as a readout, and analytical flow cytometry to accurately quantify expression levels in thousands of single cells per construct. Constructs can be transfected in cultured *Drosophila* S2 cells, allowing for rapid testing of many core promoter variants in a few days. We used the fast-folding version of the GFP, sfGFP[16] to minimize the time between transcription and fluorescence production.

Since we were working with transient transfection, we included an internal control of expression for each transfected cell. The 'transfection indicator' (mCherry) was placed on the same plasmid to avoid problems associated with differences between plasmid proportions in individual cells: the reporter vector carried the coding sequence of mCherry under the control of a constitutive *Drosophila* promoter (*actin5C* gene).

The vector of origin is the pTetOne vector included in the 3rd generation Tet-One™ Inducible Expression System (Clontech). sfGFP was amplified with primers CGGTGGATCCAGATCATGAGCAAGGGCGAGGAG and TATGCTGCAGAGATCTTACTTGTACAGCTCGTCC, and cloned into the BglII-digested pTetOne vector using the In-Fusion® cloning system (Clontech), generating pTetOne-sfGFP vetor. mCherry harboring a nuclear localization signal (ATGGCTCCAAAGAAGAAGCGTAAGGTG) was cloned into the KpnI and EcoRI sites of the pAc5.1/V5-His C vector (Life Technologies), a BamHI restriction site present in this vector was eliminated using the DpnI method for site-directed mutagenesis (primers GGGGTACCGAGCTCGAATCCACTAGTAACG and

Nature Genetics: doi:10.1038/ng.3791

CGTTACTAGTGGATTCGAGCTCGGTACCCC). A fragment containing the NLS-mCherry under control of *actin5c* promoter was then amplified using GCCTTTTCCAAGGCATAAAAAAAATCATGAATGGCATC and CTCTTGTCCAGTCTAGACTTGTACAGCTCGTCCAT primers, and subcloned into XcmI-digested pTetOne-sfGFP vector using the In-Fusion system. This generates an in-frame fusion of NLS-mCherry with the Tet-On 3G activator, and at the same time replaces the mammalian hPGK constitutive promoter with the fly *actin5c* promoter. We called this basic vector TIPR-cherry (Tet-inducible promoter reporter).

In this vector, the sfGFP is under the control of the TRE3Gs promoter, which includes the TetO binding sites and the MMTV LTR as core promoter. This promoter is almost inactive in S2 cells, so it was typically used as a negative control. This MMTV fragment can be precisely excised by cutting with BamHI + HindIII, so it can be replaced by the desired *Drosophila* core promoter variant. The promoters tested, and lines (genotypes) from which they were amplified from, are listed below:

Gene	Original CAGE Window	Cloned region	Lead SNP position	Ref allele line	Alt allele line	Ref allele is
CG2469	chr3L:1307085-1308108	chr3L:1,307,438-1,307,699	1307602	324	820	Major
TfIIB	chr2L:10430873-10431896	chr2L:10,431,106-10,431,402	10431345	786	362	Major
CG17802	chr3R:13630404-13631427	chr3R:13,630,891-13,631,300	13630942	362	313	Major
l(3)01239	chr3L:11066937-11067960	chr3L:11,067,347-11,067,740	11067477	707	362	Major
CG12576	chrX:21883296-21884319	chrX:21882430-21884113	21883861	313	714	Major
Spt6	chrX:6161332-6162355	chrX:6,161,361-6,161,940	6161845	786	313	Major
CG11210	chr2R:3985701-3986724	chr2R:3,986,030-3,987,069	3986212	362	786	Minor
CG10347*	chrX:11748554-11749577	chrX:11,748,692-11,748,901	11748798	313	517	Minor
wls	chr3L:11162186-11163209	chr3L:11,162,560-11,162,856	11162703	313	399	Major
CG31436	chr3R:21129353-21130376	chr3R:21,129,620-21,129,934	21129867	313	786	Major
CG15739	chrX:11766964-11767987	chrX:11,767,356-11,767,533	11767477	313	786	Major
Hn	chr3L:7753127-7754150	chr3L:7,753,560-7,753,726	7753664	313	362	Minor

Promoter regions cloned and assessed with the promoter reporter. We indicate the DGRP RAL numbers of the fly lines used to amplify the Reference and Alternative genotypes, as well as the correspondence between Ref/Alt and Major/Minor for each promoter.
* The window encompassing the CG10347 promoter is centered of the ATP7 promoter on the opposite strand.

For generating the point mutants (Maj^{min} and Min^{maj}), we used overlap-extension PCR[17] with the same primers used for initial amplification as outer primers and specially-designed ~20nt-long primers overlapping the mutation position as inner primers. All primer sequences are available under request.

Flow cytometry and single-cell quantification of gene expression

S2 cells were grown in Schneider's medium (Gibco, Life Technologies) supplemented with 10% fetal bovine serum (Sigma). For transfection, approximately 1.10^6 S2 cells per well were plated into 12-well plates (Falcon) and the next day transfected using Cellfectin II reagent (Life Technologies) according to

Nature Genetics: doi:10.1038/ng.3791

the manufacturer's instructions, with 46 μl of reagent and 500 ng of total DNA per well (250 ng of the TIPR-cherry vector + 250 ng of an unrelated plasmid). After transfection (in serum-free medium), they were switched again to 10% FBS-containing medium, but Tet-free certified medium (Clontech) was used. The next day, the cells were directly harvested by pipetting and measured with an LSR-Fortessa Analyzer (BD) at the EMBL Flow Cytometry Core Facility, using the FITC and PE-Texas Red channels for sfGFP and mCherry fluorescence respectively.

Results are shown by comparing expression of samples (for example all four variants of the same promoter) measured in the same experiment, to minimize experiment-to-experiment variability. A mock-transfection (no TIPR-cherry plasmid) and a positive control (HSP core promoter from pUAST) were always included in each experiment. Flow cytometry raw data was processed using FlowJo v.10.0.8 software. Typically FSC-A and SSC-A parameters were used to filter out dead cells, and then FSC-A and SSC-W were used to select single cells. Afterwards, we apply a first threshold on mCherry value to exclude untransfected cells. This threshold was selected to exclude all cells in the mock-transfected sample. The cells that passed this threshold were selected as transfected cells. Values for all positive cells were then exported in csv files, and further analysis was performed in R. Importantly, transfected cells have a log-linear relation between the sfGFP and mCherry vectors, confirming that the mCherry signal is useful to correct for typical causes of variation, such as plasmid copy number per cell or differences in the expression potential. In contrast, cells with low mCherry values (cells expressing at the lowest levels, although clearly separated from untransfected cells) do not have a log-linear relationship between the presence of mCherry and GFP. To remove any bias from these cells, we therefore discarded data from cells corresponding to the bottom 10% of mCherry raw signal after gating.

For statistical testing of differences in promoter strength to verify causality of the lead variant, the following procedure was applied. First, a two-tail Welch's t-test (which accounts for differences in variance by adjusting the degrees of freedom) is performed to compare Maj and Min variants. For significant cases on this test ($p<0.05$), we then tested if the mutant variant (Maj^{min}) was able to mimic this effect. Therefore, we requested an effect of the Maj vs. Maj^{min} on the same direction as Maj vs. Min, and performed a one-tail Welch's t-test.

Nature Genetics: doi:10.1038/ng.3791

References

1. Takahashi, H., Lassmann, T., Murata, M. & Carninci, P. 5' end-centered expression profiling using cap-analysis gene expression and next-generation sequencing. *Nature protocols* **7**, 542-561 (2012).
2. Huang, W. *et al.* Natural variation in genome architecture among 205 Drosophila melanogaster Genetic Reference Panel lines. *Genome Res* **24**, 1193-208 (2014).
3. Degner, J.F. *et al.* Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics* **25**, 3207-12 (2009).
4. Degner, J.F. *et al.* DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482**, 390-4 (2012).
5. van de Geijn, B., McVicker, G., Gilad, Y. & Pritchard, J.K. WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nature Methods* **12**, 1061-1063 (2015).
6. Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat Protoc* **7**, 500-7 (2012).
7. Lippert, C., Casale, F.P., Rakitsch, B. & Stegle, O. LIMIX: genetic analysis of multiple traits. *bioRxiv* (2014).
8. Zhou, X. & Stephens, M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nat Methods* **11**, 407-9 (2014).
9. Hoskins, R.A. *et al.* Genome-wide analysis of promoter architecture in Drosophila melanogaster. *Genome research* **21**, 182-192 (2011).
10. Thomas, S. *et al.* Dynamic reprogramming of chromatin accessibility during Drosophila embryo development. *Genome biology* **12**, R43 (2011).
11. Gronau, I., Arbiza, L., Mohammed, J. & Siepel, A. Inference of natural selection from interspersed genomic elements based on polymorphism and divergence. *Mol Biol Evol* **30**, 1159-71 (2013).
12. Mackay, T.F. *et al.* The Drosophila melanogaster Genetic Reference Panel. *Nature* **482**, 173-178 (2012).
13. Drosophila 12 Genomes, C. *et al.* Evolution of genes and genomes on the Drosophila phylogeny. *Nature* **450**, 203-18 (2007).
14. Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* **20**, 110-21 (2010).
15. Cannavo, E. *et al.* Shadow Enhancers Are Pervasive Features of Developmental Regulatory Networks. *Curr Biol* **26**, 38-51 (2016).
16. Pedelacq, J.D., Cabantous, S., Tran, T., Terwilliger, T.C. & Waldo, G.S. Engineering and characterization of a superfolder green fluorescent protein. *Nat Biotechnol* **24**, 79-88 (2006).
17. Heckman, K.L. & Pease, L.R. Gene splicing and mutagenesis by PCR-driven overlap extension. *Nat Protoc* **2**, 924-32 (2007).

Nature Genetics: doi:10.1038/ng.3791

SUPPLEMENTARY TABLES

LEGENDS FOR SUPPLEMENTARY TABLES

Supplementary Tables 1, 2, 3, 4, 6 and 7 are provided as separate files. They are also available on the Furlong lab web page, http://furlonglab.embl.de/data/download.

Supplementary Table 1. Promoter regions used for tssQTL calling.

This is a gff3 file containing the 1024bp windows used to call the tssQTL.
These columns contain no information but are gff3 mandatory: source,type,score,phase
Extra fields:
gene_name: name of associated gene
gene_id: ID of associated gene
Window: the ID of the window
total_reads: total number of CAGE reads in the window
shape.index: shape index for the window calculated from the aggregate CAGE data

Supplementary Table 2. Summary of all significant associations found with the joint model.

This is a tab delimited text file.
Columns:
Window: Window ID of the CAGE window
chromosome: chromosomal location of Window and variant
topSNP - the position of the genetic variant with the highest P-value (lead variant)
GeneWise.Q.1012h.Mean.oneTP: permutation based significance for a single-time-point model (10-12h) using mean as phenotype
GeneWise.Q.68h.Mean.oneTP: permutation based significance for a single-time-point model (6-8h) using mean as phenotype
GeneWise.Q.24h.Mean.oneTP: permutation based significance for a single-time-point model (2-4h) using mean as phenotype
GeneWise.Q.Common.Effect.Mean: permutation based significance for a joint model (3 time points together) using mean as phenotype
GeneWise.Q.Common.Effect.3PC: permutation based significance for a joint model (3 time points together) using first 3 PCs as phenotype
GeneWise.Q.1012h.3PC: permutation based significance for a single-time-point model (10-12h) using first 3 PCs as phenotype
GeneWise.Q.68h.3PC: permutation based significance for a single-time-point model (6-8h) using first 3 PCs as phenotype
GeneWise.Q.24h.3PC: permutation based significance for a single-time-point model (2-4h) using first 3 PCs as phenotype
Timespecific_1012h: Logical denoting significant stage-specific effects at this timepoint
Timespecific _68h: Logical denoting significant stage-specific effects at this timepoint
Timespecific _24h: Logical denoting significant stage-specific effects at this timepoint

Nature Genetics: doi:10.1038/ng.3791

QTLwise_1012h_P_3PC_FF: P-value for stage-specific effects at this timepoint
QTLwise_68h_P_3PC_FF: P-value for stage-specific effects at this timepoint
QTLwise_24h_P_3PC_FF: P-value stage-specific effects at this timepoint
GeneName: name of associated gene
GeneID: ID of associated gene
NumTopSnps: number of extra variants with a P-value within one order of magnitude of that of the lead variant
RelativeTopQTLPosition: position of the lead variant with respect to the center of the window
RelativeStrongestSignificantPositions: the location of the site with highest single-base-pair effect, relative to the center of the Window
strand: the strand where the window center was defined
set: whether the QTL is shape, mixed or abundance
peakid: the ID of the TSS cluster associated with the Window
Internal: logical, whether the variant is associated only with an internal cage peak (excluded from the high confidence set)
is_enzyme_artifact: logical, whether the tssQTL is associated with a variant which plausibly causes variation in artifactual signal due to affecting EcoP15I restriction site (excluded from the high confidence set)
minfreq: the minor allele frequency (MAF) of the variant
l2fc: log2 fold change in expression due to tssQTL, calculated using library size normalized data
shapechange: change in shape index due to tssQTL, calculated using the raw reads and relevant time points

Supplementary Table 3. Summary of high-confidence associations found with the joint model.

This is a tab delimited text file.
This table contains the same windows as Table S2, with exception of those QTL labeled as internal or enzyme artifact (see Table S2).
All columns in Table S2 are present.
In addition, it contains the following columns derived from the waveQTL analysis:
NumberMainEff: number of positions with significant effects in the primary direction
NumberOppEff: number of positions with significant effects in the secondary direction
SumMainEff: sum of significant effects in the primary direction
SumOppEff: sum of significant effects in the secondary direction
BayesFactorWave1: log10 Bayes factor for wave 1 (evidence of effect on the mean)
MaxBayesFactor: maximum log10 Bayes factor for any of the wave coefficients

Supplementary Table 4. Core promoter motifs obtained *de novo*.

This is a tab delimited text file, containing the motifs found, grouped in similarity clusters and with those having IC < 8 filtered out.
Columns:
Name: unique identifier for each Motif
Cluster: the cluster of similar motifs which includes this motif
Motif_word: consensus sequence for the motif
EnrichedInSet: the set(s) of TSS which were used as test and background sets to discover the motif
E_value: the motif's associated E-value
Algorithm: whether the motif was found using MEME or DREME
IC: Information Content of motif

Nature Genetics: doi:10.1038/ng.3791

Prevalence: number of TSS clusters with this motif
Strand_bias: the degree to which the motif is found on a particular strand
is_positioned: logical, whether the motif shows significant positional bias relative to TSS as per centrimo
Pos_E_value : the E value of the motifs positional enrichment
Position: the location of the motifs enriched zone
Bin_width: the width of the motifs positional enrichment zone
Shape_shift: the difference between the global mean of shape index, and the mean shape index for promoters with this motif
Novel: it states 'No' if the motif has a match with those appearing in Ohler et al 2002, FitzGerald et al 2006, or Down et al 2007; it states 'TF' if the motif only has a match to a TF motif according to TomTom; if none of the previous cases, it states 'Yes'
Ohler: matching motif from Ohler et al 2002, if any
FitzGerald: matching motifs from FitzGerald et al 2006, if any
Tiffin: matching motifs from the Tiffin database (Down et al 2007), if any
tomtomMatch: matching TF motifs according to Tom Tom, if any
locationset: whether the motif is positioned upstream the TSS, overlapping the TSS, or downstream the TSS
Abundance: number of abundance QTL whose lead variant alter this motif
Mixed: number of mixed QTL whose lead variant alter this motif
Shape: number of shape QTL whose lead variant alter this motif

Supplementary Table 5. Results from the INSIGHT test for promoter windows.

SI bins: the range of shape indexes on each bin of the CAGE windows
transcr overlap: average overlap with transcripts for the windows tested
rho: INSIGHT's best estimate of the proportion of base pairs under selection
rho stderr: INSIGHT 's standard error for its estimate of rho
E[A]: INSIGHT's best estimate of the number of base pairs under adaptive evolution, per kb
E[A] stderr: standard error for the above
E[W]: INSIGHT's best estimate of the number of base pairs under weak negative selection, per kb
E[W] stderr: standard error for the above
alpha: INSIGHT's best estimate of the proportion of substitutions driven by positive selection
alpha stderr: standard error for the above

Supplementary Table 6. Libraries information.

This is a tab delimited text file.
Columns:
Library: identifier of the individual library
Line: RAL number identifying the *D. melanogaster* line from the DGRP.
TP: timepoint of embryo collection
Barcode: sequence of the barcode added during library preparation
Total.Reads: total number of sequenced reads
Reads.uniquely.mapped: reads successfully and uniquely mapping to the reference genome
Reads.after.WASP: total uniquely mapped reads remaining after WASP filtering
Total.Reads.Windows: total number of mapped reads contained in all the considered promoter windows
Reads.Windows.after.WASP: reads in windows remaining after WASP filtering
Reads.Windows.after.WASP.UMM: reads in windows remaining after WASP and UMM filtering steps

Nature Genetics: doi:10.1038/ng.3791

Supplementary Table 7. TSS clusters.

This is a gff3 file containing the cage peaks or TSS clusters.
These columns contain no information but are gff3 mandatory: source,type,score,phase
Extra fields:
peakid: unique ID associated with the TSS cluster
gene_name: name of associated gene
gene_id: ID of associated gene
isinternal: whether the peak is in the internal set of peaks
shape.ind: the peak's shape index
shape: whether the peak is broad or narrow (threshold is -1)
totalexpr: total reads associated with the peak
The following columns are only present if the TSS cluster overlaps a QTL window:
Window: the associated CAGE window, where it has an associated tssQTL
set: the type of associated tssQTL (shape, mixed or abundance)
majorshape: the peak's shape index in the major genotype
minorshape: the peak's shape index in the minor genotype
shapechange: the shape index difference between minor and major

Nature Genetics: doi:10.1038/ng.3791

Supplementary Table 5. Results from the INSIGHT test for promoter windows.

Shape Index (SI) bins	transcr. overlap	rho	rho stderr	E[A]	E[A] stderr	E[W]	E[W] stderr	alpha	alpha stderr
0.725 to 1.95	39.4%	0.5231	0.00778	4.7848	0.4226	10.8223	0.3923	0.1962	0.01588
0.228 to 0.725	40.6%	0.5378	0.00795	6.3567	0.4390	10.8999	0.3944	0.2492	0.01539
-0.277 to 0.228	44.5%	0.5559	0.00807	5.6895	0.4442	10.7355	0.4090	0.2373	0.01664
-0.877 to -0.277	44.7%	0.5449	0.00858	8.3143	0.4849	10.3059	0.4243	0.3052	0.01544
-1.56 to -0.877	43.0%	0.4932	0.00977	14.7267	0.5650	8.0687	0.4338	0.4101	0.01291
-2.1 to -1.56	42.9%	0.5069	0.01023	20.9753	0.6251	7.4767	0.4709	0.5079	0.01153
-2.51 to -2.1	40.3%	0.4869	0.01080	20.8812	0.6557	7.0354	0.4785	0.4945	0.01201
-2.87 to -2.51	40.5%	0.4370	0.01190	22.8856	0.7221	6.5510	0.5102	0.4921	0.01210
-3.29 to -2.87	43.0%	0.5328	0.01180	22.4057	0.7269	8.3317	0.5239	0.5354	0.01308
-4 to -3.29	50.4%	0.5388	0.01384	21.7851	0.8428	7.9616	0.5926	0.5256	0.01559

Nature Genetics: doi:10.1038/ng.3791