nature genetics

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

Ignacio E Schor[1,5], Jacob F Degner[1], Dermot Harnett[1], Enrico Cannavò[1], Francesco P Casale[2], Heejung Shim[3], David A Garfield[1], Ewan Birney[2], Matthew Stephens[4], Oliver Stegle[2] & Eileen E M Furlong[1]

Animal promoters initiate transcription either at precise positions (narrow promoters) or dispersed regions (broad promoters), a distinction referred to as promoter shape. Although highly conserved, the functional properties of promoters with different shapes and the genetic basis of their evolution remain unclear. Here we used natural genetic variation across a panel of 81 *Drosophila* lines to measure changes in transcriptional start site (TSS) usage, identifying thousands of genetic variants affecting transcript levels (strength) or the distribution of TSSs within a promoter (shape). Our results identify promoter shape as a molecular trait that can evolve independently of promoter strength. Broad promoters typically harbor shape-associated variants, with signatures of adaptive selection. Single-cell measurements demonstrate that variants modulating promoter shape often increase expression noise, whereas heteroallelic interactions with other promoter variants alleviate these effects. These results uncover new functional properties of natural promoters and suggest the minimization of expression noise as an important factor in promoter evolution.

TSSs are defined by the core promoter, a complex DNA element that facilitates recruitment of the basal transcriptional machinery and RNA polymerase II (RNA Pol II)[1–3] in addition to providing specificity for interactions with particular enhancers[4–10]. Methods such as cap analysis of gene expression (CAGE)[11,12] map TSS position at single-nucleotide resolution and provide a global view of TSS distribution within a promoter[11,13,14]. This has led to the classification of animal promoters into two major types[3,15]. Narrow promoters, typical of genes with restricted tissue-specific expression, are often associated with positioned motifs such as the TATA box or initiator (Inr), and have a single predominant TSS. Broad promoters, typical of ubiquitously expressed genes, have more dispersed patterns of transcriptional initiation and do not contain a TATA box[16–19]. Broad promoters were initially associated with mammalian CpG island promoters[11] but later found to be common in *Drosophila*[20,21], indicating that they function independently of the CpG island itself. Broad and narrow promoters also have differences in the positioning and histone modifications of the first nucleosome downstream of the promoter[22–24], suggesting different regulatory mechanisms.

Although both promoter classes are found in organisms ranging from flies to humans[11,14,23], their inherent functional differences are not understood. Cross-species studies indicate that promoter shape is generally conserved[11,25], suggesting functional importance, yet the natural variability of promoter shape within populations and its genetic determinants and evolutionary constraints are unknown.

Studies of expression quantitative trait loci (eQTLs) in human[26–28] and model systems[29–32] have identified extensive functional genetic variants in the vicinity of promoters that affect transcript abundance. This suggests that natural sequence variation within a highly polymorphic species such as *Drosophila melanogaster* could also be used as a perturbation tool to gain mechanistic insights into promoter function by testing for associations with complex molecular readouts. Following this rationale, we used CAGE[12,33] sequencing across a panel of 81 *Drosophila melanogaster* inbred lines to investigate genetic effects on 5′ transcriptional initiation at single-base-pair resolution during embryonic development. Mapping TSS-associated QTLs (tssQTLs) using CAGE-derived phenotypes highlighted three types of associations: variants affecting transcript abundance (either promoter strength or transcript turnover), variants affecting promoter shape (a new genetic trait), or variants affecting both. Genetic variants primarily associated with promoter shape are located within the core promoter region itself. Using a single-cell quantitative assay combined with promoter engineering, we found that tssQTLs affecting promoter shape typically increase expression noise. In their natural sequence context, this noise effect was often buffered by other variants within the promoter, suggesting that promoter shape–associated variants, although frequent, are found only in specific allelic combinations. Taken together, our results identify promoter shape as a genetically controlled molecular trait with important implications for both promoter function and evolution.

[1]European Molecular Biology Laboratory (EMBL) Genome Biology Unit, Heidelberg, Germany. [2]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Hinxton, UK. [3]Department of Statistics, Purdue University, West Lafayette, Indiana, USA. [4]Department of Human Genetics, University of Chicago, Chicago, Illinois, USA. [5]Present address: Instituto de Fisiología, Biología Molecular y Neurociencias (IFIBYNE-CONICET), Buenos Aires, Argentina. Correspondence should be addressed to E.E.M.F. (furlong@embl.de) or O.S. (oliver.stegle@ebi.ac.uk).

Received 14 September 2016; accepted 20 January 2017; published online 13 February 2017; doi:10.1038/ng.3791

RESULTS

A strategy to map genetic determinants of TSS usage

As a source of genetic perturbations, we made use of the extensive natural sequence variation within *Drosophila*, which is characterized by inherently small (<5 kb) blocks of linkage disequilibrium (LD) and therefore well suited for high-resolution QTL mapping within endogenous core promoter elements. As part of the *Drosophila* Genetic Reference Panel (DGRP)[34], wild *D. melanogaster* isolates were inbred to near homozygosity, representing near genetic clones. We selected 81 unrelated lines and used CAGE to measure transcription initiation at base-pair resolution for each. Given the importance of specific promoter types in the regulation of developmental and tissue-specific gene expression[4,8–10,14,24,35], we studied promoter function during embryonic development. CAGE libraries were prepared from tightly staged embryos at three important transitions during embryogenesis (**Fig. 1a**): 2–4 h after egg laying (AEL) (stages 5–8, early in development, including blastula), 6–8 h AEL (stages 10–11, during major linage specification in the mesoderm and ectoderm) and 10–12 h AEL (stages 13–14, terminal tissue differentiation). This yielded a total of 243 samples.

Prior to the genetic analysis of CAGE-derived phenotypes, we applied conservative filters to minimize potential biases due to differential mappability (Online Methods). First, to avoid line-specific mapping effects, we created a 'universal mappability map' (UMM) of the reference genome that identifies all sites that can be accurately mapped across all lines. All genomic positions that failed to uniquely map DNA reads (using the same read length as CAGE) in even 1 of the 81 lines were excluded from the analysis (Online Methods), providing a conservative method to eliminate mapping biases and trivial associations with insertions or deletions (indels) segregating within the population. Second, to remove residual variation in mappability that may have escaped the UMM, we applied WASP, a computational method to identify potential biases in allele-specific read mapping[36]. To identify active promoters, both annotated and unannotated, we iteratively selected 1-kb windows centered on the highest CAGE peaks across the entire genome (**Fig. 1a**) until 99% of genomic regions with mapped CAGE signal (with >10 reads) were covered. This identified 13,249 transcriptionally active regions (windows) at one or more stages of embryogenesis (**Supplementary Table 1**).

We developed an analysis strategy to detect genetic effects on different traits derived from the CAGE signal. Specifically, we used principal component (PC) analysis to project the full CAGE output within 1-kb windows onto the first three PCs, which capture the major axes of covariation of individual bases in a promoter among *Drosophila* lines. These PC-based traits capture changes in total TSS usage (promoter strength) and in spatial distribution of CAGE signal (promoter shape) (**Fig. 1b** and **Supplementary Fig. 1**).

We then applied a multi-trait linear mixed model[37–39] (LMM) to test for associations between genetic variants in 200-kb *cis*-regions centered on each active promoter window and the PC-based traits (**Fig. 1a** and Online Methods). This model extends existing LMMs to enable testing for genetic effects that are common or time-specific across developmental stages, and it accounts for genetic structure between lines as well as nongenetic correlations between traits (Online Methods). Using this approach, we tested ~2.2 million common variants (minor allele frequency > 5%) extracted from the Freeze 2 genotype catalog for *cis*-associations[34,40]. After filtering associations influencing intragenic CAGE clusters and variants disrupting the restriction site used for CAGE (Online Methods), we identified a high-confidence set of 4,075 promoters with a tssQTL (at genome-wide false discovery rate (FDR) < 0.01) (**Supplementary Tables 2** and **3**).

Almost 30% (1,183/4,075) of tssQTLs had a distinct lead variant (with *P* values more than an order of magnitude different from those of other variants), and more than 70% contained five or fewer equivalent lead variants. This suggests that our tssQTL set is highly enriched for causal variants affecting transcriptional initiation, which our experimental validation (described below) strongly supports.

The joint model using PCs as traits yielded greater power to detect tssQTLs, almost doubling the number of associations achieved with a conventional approach using mean CAGE signal as a trait (**Fig. 1c**), and detected a large set of additional regulatory associations (2,363) (**Fig. 1c** and **Supplementary Table 3**). These PC-based tssQTLs included both 'simple' associations, in which the genetic variant influences transcript abundance, affecting all TSSs within a promoter in the same direction (abundance QTLs) (**Fig. 1d**), and more complex QTLs, where transcription is affected in opposite directions at different TSSs in the promoter (promoter shape QTLs) (**Fig. 1d**). Notably, the majority of tssQTLs consistently affected transcriptional initiation at all three time points tested (**Supplementary Fig. 2**), with only 5.2% (210/4,075) showing evidence of stage-specific effects (**Supplementary Table 3** and Online Methods). Moreover, tssQTLs for alternative promoters for the same gene appeared independent, showing no evidence for coordinated changes that would suggest compensation between TSSs (**Supplementary Fig. 3**).

Single-bp effect sizes distinguish three classes of tssQTL

Whereas PC-based LMM provides an excellent method to discover genetic variants associated with abundance and shape changes, it does not show the extent to which a tssQTL affects individual bases, information that is important to dissect their mechanisms. To obtain such spatially resolved effect sizes, we applied a wavelet-based decomposition model[41] to the raw CAGE signal using the set of tssQTLs (Online Methods and **Supplementary Fig. 4**). The wavelet analysis assesses genetic effects on the CAGE signal in the entire window, on halves of the window, on halves of these halves, and so on until reaching effect size estimates at base-pair resolution. By specifically targeting the tssQTLs discovered in the PC-based approach, we limited the computational burden of the wavelet analysis while gaining effect size estimates at high spatial resolution. This identified the most affected TSSs within promoters that have multiple initiation sites (**Fig. 2a**). This modeling approach also yielded a Bayes factor (BF) for evidence of associations at different spatial scales (wavelet coefficients (WCs)), which estimates the extent to which any tssQTL affects transcript abundance or promoter shape (**Fig. 2a**, **Supplementary Fig. 4** and **Supplementary Table 3**). tssQTLs that affect the first WC capture effects on transcript abundance, as shown for *CG11164* (**Fig. 2a**). Alternatively, tssQTLs with little (or no) evidence of effects on the first WC but genetic effects on higher-order WCs capture nonhomogeneous effects, which affect specific TSSs, resulting in changes of promoter shape, as illustrated by *CG7927* (**Fig. 2a**).

Different spatial architectures of tssQTLs are also evident in the single-base effect size estimates (**Fig. 2a**), which we used to classify individual tssQTLs more formally (Online Methods). Plotting the sum of effect sizes for positions with an effect in the same direction as the most affected TSS (primary direction) (**Fig. 2a**) versus the sum of effect sizes of positions with opposite effects (secondary direction) highlighted three types of tssQTL (**Fig. 2b** and Online Methods). Abundance QTLs (28.7% of tssQTLs, 1,171/4,075) are primarily associated with the mean CAGE signal (first WC) (**Fig. 2b**) and therefore are likely to affect promoter strength and transcriptional and/or post-transcriptional regulation. Shape QTLs (40.5% of tssQTLs, 1,649/4,075) are associated with changes in the distribution

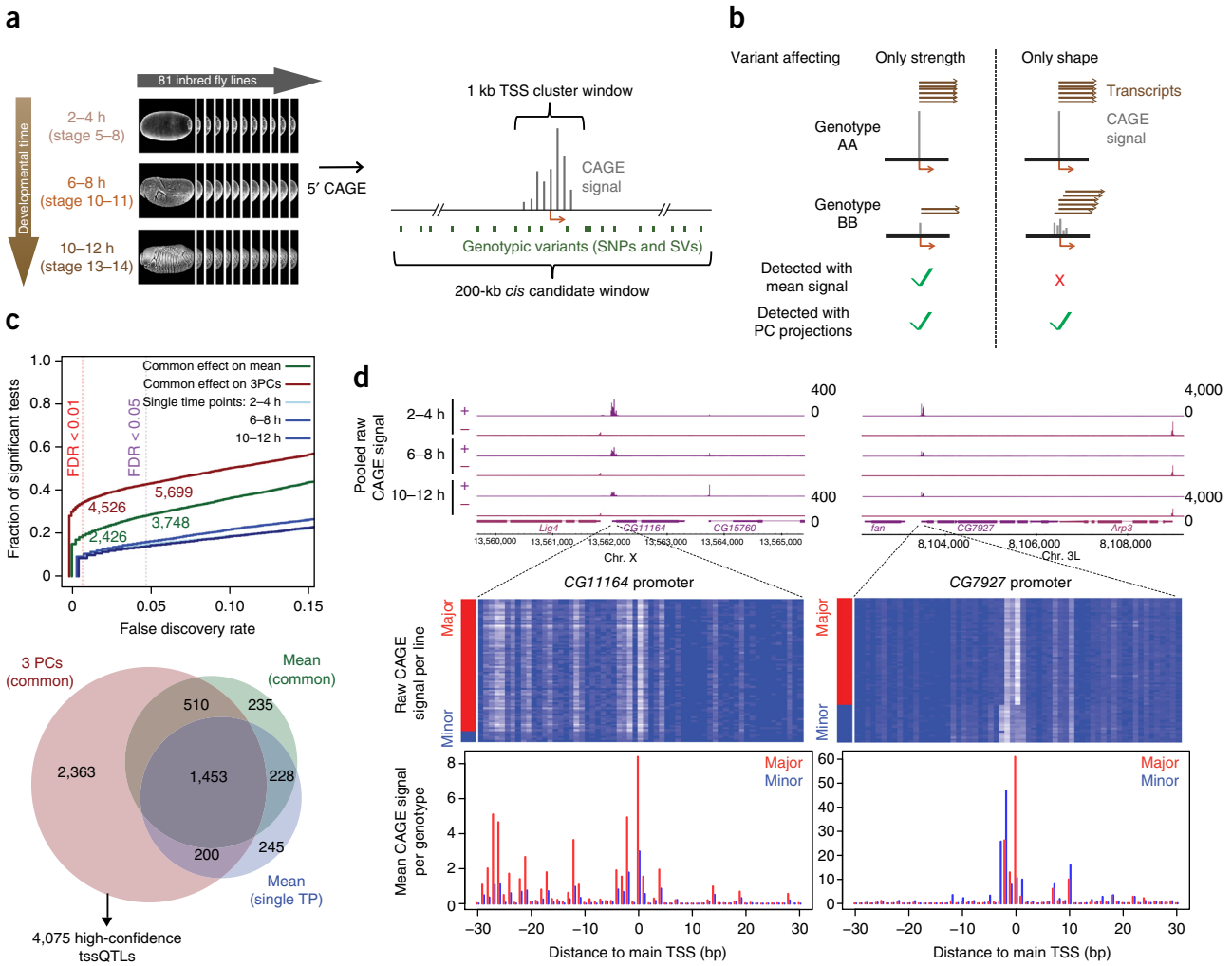

Figure 1 Identification of tssQTL using 5′ CAGE and PC analysis. (**a**) Experimental design for QTL calling. CAGE reads from embryonic samples (81 lines, 3 developmental stages) clustered into 13,249 1-kb regions (TSS cluster window) and tested for association with variants ±100 kb. (**b**) PC-based versus mean-based QTL mapping. (**c**) Joint analysis using the first three PCs (3PCs). Top, fraction of TSS clusters with a significant QTL at variable FDR for alternative models. Numbers of significant tssQTLs at FDR values <1% and <5% are indicated. Bottom, overlap of tssQTL genes at FDR < 1%. (**d**) Examples of simple (left) and complex (right) tssQTLs detected by PC method. Top, genomic locus showing aggregate CAGE signal around affected promoters at three different time points and for + and – strands. Middle, raw CAGE signal at 6–8 h for promoter window, with individual *Drosophila* lines (rows) grouped by the major (red) and minor (blue) allele at the lead variant (*CG11164*, chrX_13562404_SNP; *CG7927*, chr3L_8103470_SNP). Bottom, single-base average CAGE signal per genotype.

of TSSs (promoter shape), with little or no evidence of a change in mean signal, indicating that transcriptional changes at one position (primary direction) are balanced by transcriptional changes at another (secondary direction). Consequently, these are located close to the diagonal (**Fig. 2b**). Mixed QTLs (30.8%, 1,255/4,075) affect both transcript abundance and promoter shape and populate regions above and below the diagonal (**Fig. 2b**).

Shape and abundance tssQTLs have different properties

Genetic variants associated with shape and abundance tssQTLs preferentially occurred in different genomic contexts: variants in active promoter regions, for example, were enriched for shape QTLs and low in abundance QTLs (**Fig. 3a** and Online Methods). Similarly, genetic variants in promoter-proximal DNase I–hypersensitive sites (DHS) were 3–4 times more likely to harbor shape tssQTLs than abundance tssQTLs (**Fig. 3a**). In contrast, coding regions and the 3′ UTR were enriched for abundance QTL (odds ratio (OR) > 2 and > 4, respectively). Consistent with this, lead variants of shape QTLs were concentrated in core promoter regions (±100 bp around the TSS; **Fig. 3b**), whereas abundance QTLs were more dispersed and biased toward locations downstream of the TSS (**Fig. 3b**).

The single-base-pair effect size estimates derived from the wavelet analysis enabled us to more accurately survey the positions of lead variants relative to the most affected TSSs within promoters. This enhanced resolution showed that shape QTLs tend to be located in bases immediately adjacent to the affected TSS (**Fig. 3b**). This position strongly suggests that shape tssQTLs have local effects that are likely to involve sequence changes that strengthen or weaken motifs around the TSS itself (a hypothesis we confirm below).

Taken together, these results provide the first insights into how common genetic variants affect endogenous transcription initiation sites, highlighting promoter shape as a genetically regulated trait. The prevalence of shape tssQTLs at core promoter regions indicates extensive functional genetic variation within the population, specifically

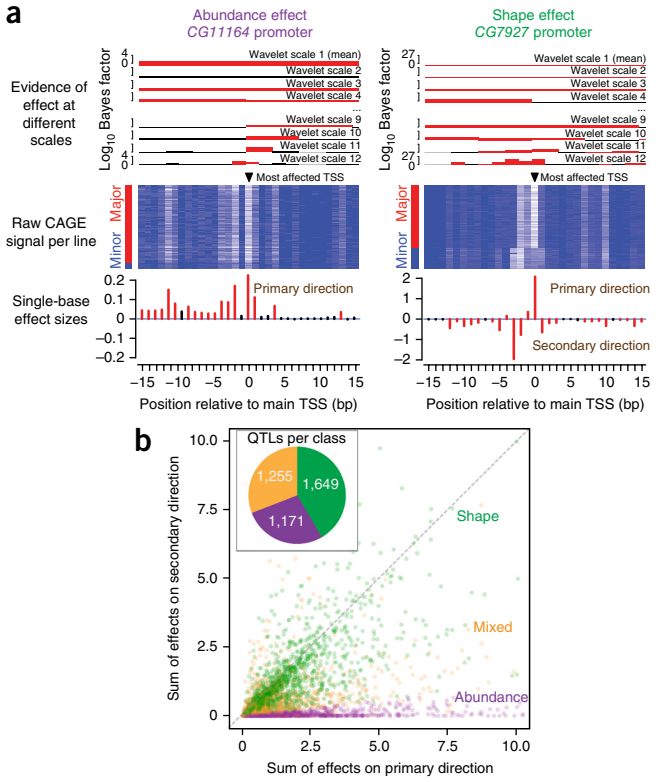

Figure 2 Classification of tssQTLs. (**a**) Wavelet analysis of tssQTL effects. Top, evidence for association ($\log_{10}$ Bayes factor) at different spatial scales (WCs) for tssQTL lead variants. Middle, raw CAGE signal for individual lines. Bottom, effect size estimates for each base pair. Arrowhead indicates the most affected TSS. Significant effects (0 is outside posterior mean for effect size ± 2 posterior s.d.) on the primary and secondary direction are shown in red. (**b**) Classification of tssQTLs. Scatterplot shows sum of significant effects on the primary versus secondary directions for 4,075 tssQTLs. Inset, tssQTL number per class.

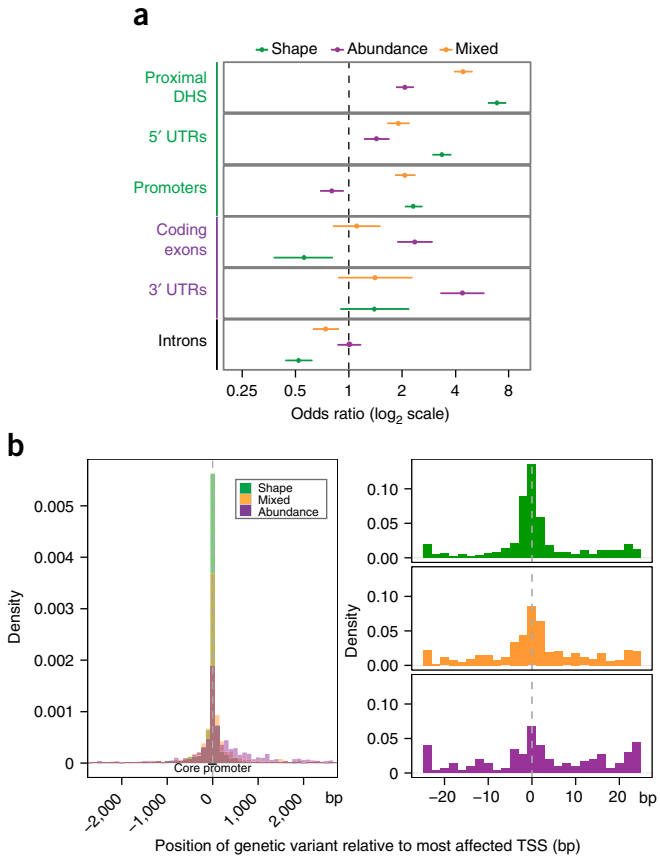

Figure 3 Genomic properties of variants affecting promoter shape or transcript abundance. (**a**) Enrichment of lead variants of shape, abundance or mixed QTLs in different genomic contexts. ORs estimated using multivariate logistic regression of TSS occurrence considering minimum allele frequency, distance to the TSS, and CAGE peak expression levels as covariates[45]. Error bars, 95% confidence interval (CI). (**b**) Lead variant positions relative to the most affected TSS. Left, densities of the relative lead variant positions for tssQTL classes (100-bp bins). Right, tssQTL density for TSS-proximal regions at higher resolution (2-bp bins).

affecting the distribution of TSS usage within a promoter without necessarily affecting transcript abundance. Consequently, promoter shape may evolve freely as an independent trait from overall transcript levels.

Lead variants are causal, affecting core promoter motifs

To assess the functional consequence of genetic variants on promoter activity, we designed a single-cell-based assay to quantify expression of a promoter across thousands of individual cells using analytical flow cytometry (**Fig. 4a**). We selected eight representative promoters, across all three types of tssQTL, that (i) had a distinct lead variant in the promoter region and (ii) induced a measurable change in expression in the cell culture assay. We then compared the promoter activity for a natural major and minor allele haplotype, as well as an engineered promoter with the minor allele lead variant placed into the major allele background (Maj^{min}). The tssQTL for *CG31436*, for example, is due to a minor allele variant that creates an Inr motif, resulting in increased promoter strength (**Fig. 4b**). Placing the minor allele into the major haplotype reproduced this effect, thus confirming that the tssQTL lead variant is the causal SNP. Similarly, the QTL for *CG12576* is due to a SNP that destroys an instance of motif 1 (ref. 42) in the minor allele, thereby reducing promoter activity (**Fig. 4b**). Again, allele replacement of the minor allele into the major haplotype phenocopied the promoter activity of the natural minor haplotype. We confirmed the function of the remaining six tssQTLs in an analogous manner (**Supplementary Fig. 5**), demonstrating that the lead variants of all eight tested tssQTLs were causal, as measured by their effect on expression in the reporter assay (**Fig. 4b** and **Supplementary Fig. 5**). Taken together, the results of this validation strongly suggest that the 4,075 tssQTLs are highly enriched for causal variants, with sufficient resolution to pinpoint functional motifs.

To more globally dissect the underlying mechanism of tssQTLs, we used *de novo* motif discovery on all active promoters to identify potentially novel core promoter motifs (Online Methods). By combining reads from all lines and time points, we obtained sequence information of TSSs at depths (~2.6×10^9 uniquely mapped reads) unprecedented for these stages of embryogenesis. This enabled us to recover all eight common *Drosophila* core promoter motifs[20] and discover 22 new motif clusters, 15 of which were enriched at specific positions with respect to the main TSS, suggesting a specific function in transcription initiation (**Supplementary Table 4** and Online Methods).

Promoter motifs with specific positioning relative to the TSS (upstream, at the TSS, or downstream) were globally enriched for tssQTLs, even when compared to promoter motifs that lack

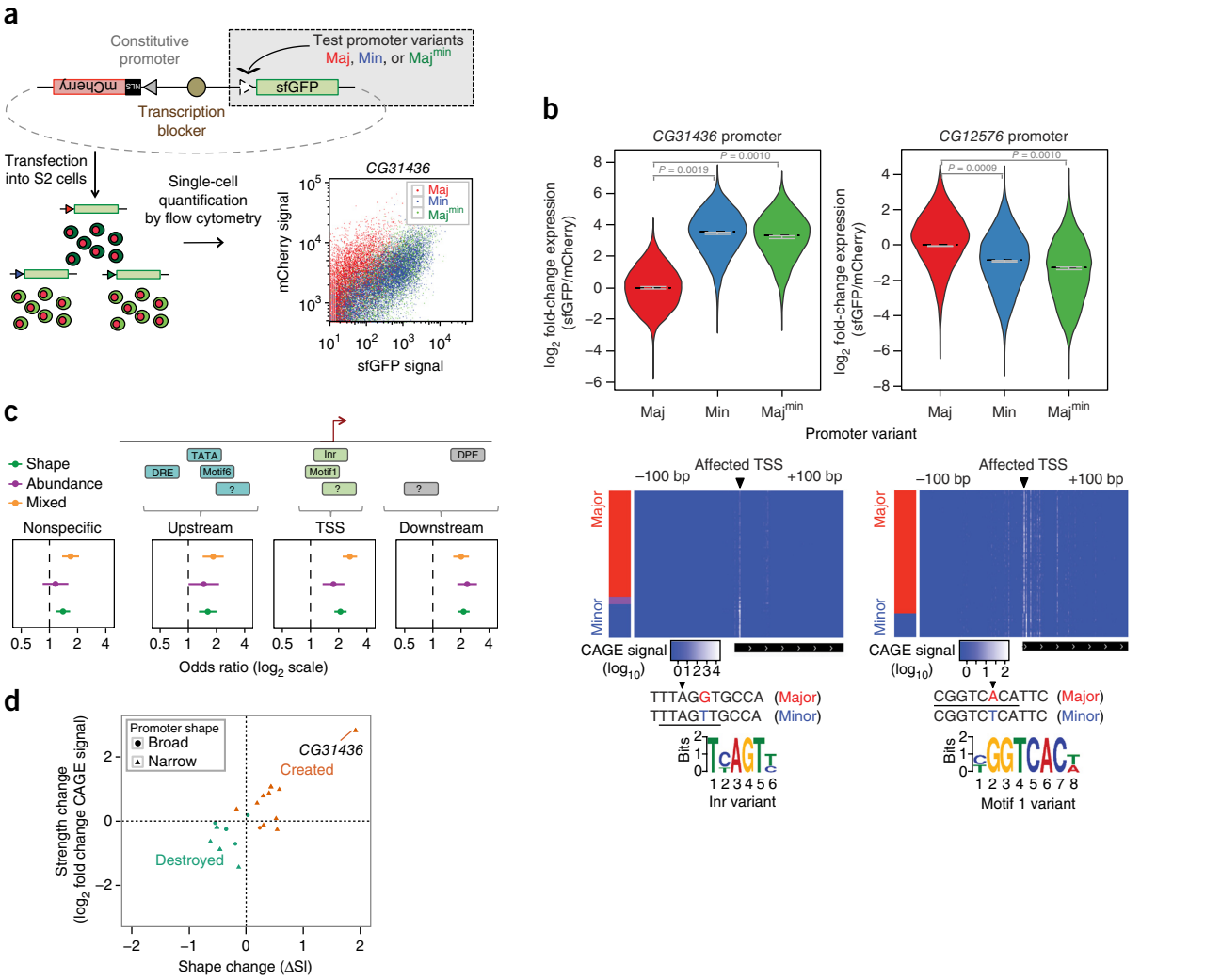

Figure 4 Core promoter motifs affected in tssQTLs. (**a**) Promoter activation of sfGFP measured in individual transiently transfected cells by analytic FACS (mCherry, transfection control) for natural promoter variants (major (Maj) and minor (Min)) and engineered promoters (Maj^{min}). (**b**) Inr motif creation (left, variant chr3R_21129867_SNP) and motif 1 disruption (right, variant chrX_21883861_SNP). Top, distribution of single-cell expression levels for the different promoter alleles. Dashed black lines, population median; error bars, mean ± s.e.m. of 2 (*CG31436*) or 3 (*CG12576*) replicates. *P* values calculated by two-sided (Maj vs. Min) or one-sided Welch's *t*-test (Maj vs. Maj^{min}). Both experiments were repeated 3 times. Bottom, heat maps for the corresponding promoter CAGE signals across all lines for the 10–12 h time point. Sequence logos show the affected motif consensus (underlined). (**c**) Enrichment of tssQTLs in *de novo*–discovered motifs, including known and novel core promoter motifs. Nonspecific motifs are not positioned, directional, or discriminative between promoter types. Error bars, 95% CI. (**d**) Change in raw CAGE signal and SI between minor and major genotypes for 19 tssQTLs affecting Inr-like motifs (Online Methods) overlapping the highest-effect TSS. Created or destroyed is determined relative to the minor allele variant.

preference in their localization, strand, or promoter type (nonspecific motifs), with no significant differences between tssQTL types (**Fig. 4c**). Moreover, the direction of motif changes caused by tssQTLs was largely concordant with the observed direction of effect on transcript abundance or promoter shape—for example, for Inr-like motifs located at the TSS (**Fig. 4d**)—indicating their functional role. We observed that when an Inr-like motif is created (in the minor genotype), for example, the promoter becomes stronger and narrower, and the opposite happens when it is destroyed (**Fig. 4d**). This population resource complements cross-species data in mice and humans, showing that substitutions in Inr-like PyPu motifs disrupt TSS usage[11]. We found that 55.7% (167/300) of shape QTLs located within ±1 bp of a TSS modify a PyPu motif, providing further support for the functional importance of this dinucleotide in defining the precise position of transcriptional initiation. However, disruption of Inr-like PyPu motifs only accounted for 10% (167/1649) of all shape QTLs, which probably reflects the higher information content of *Drosophila* TSS-defining motifs[1,18,42] and the importance of other core promoter motifs. For example, changes in downstream-positioned DPE- or MTE-like motifs were also generally concordant with the direction of change in promoter output (**Supplementary Fig. 6**). Transcription initiation patterns could also be affected by tssQTL disruption of nucleosome positioning, as recently suggested[43]. Although 228 of our tssQTLs (including 64 shape QTLs) were located in the general region of the first nucleosome (+50 to +250 bp from the affected TSS), we did not find an association between changes in predicted nucleosome positioning sequences[44] and the effects of these tssQTLs.

Broad initiation confers distinct evolutionary properties

The extensive genetic variation affecting TSSs highlights the potential evolvability of features such as promoter shape and raises questions about the selective pressures acting on this variation. To address these

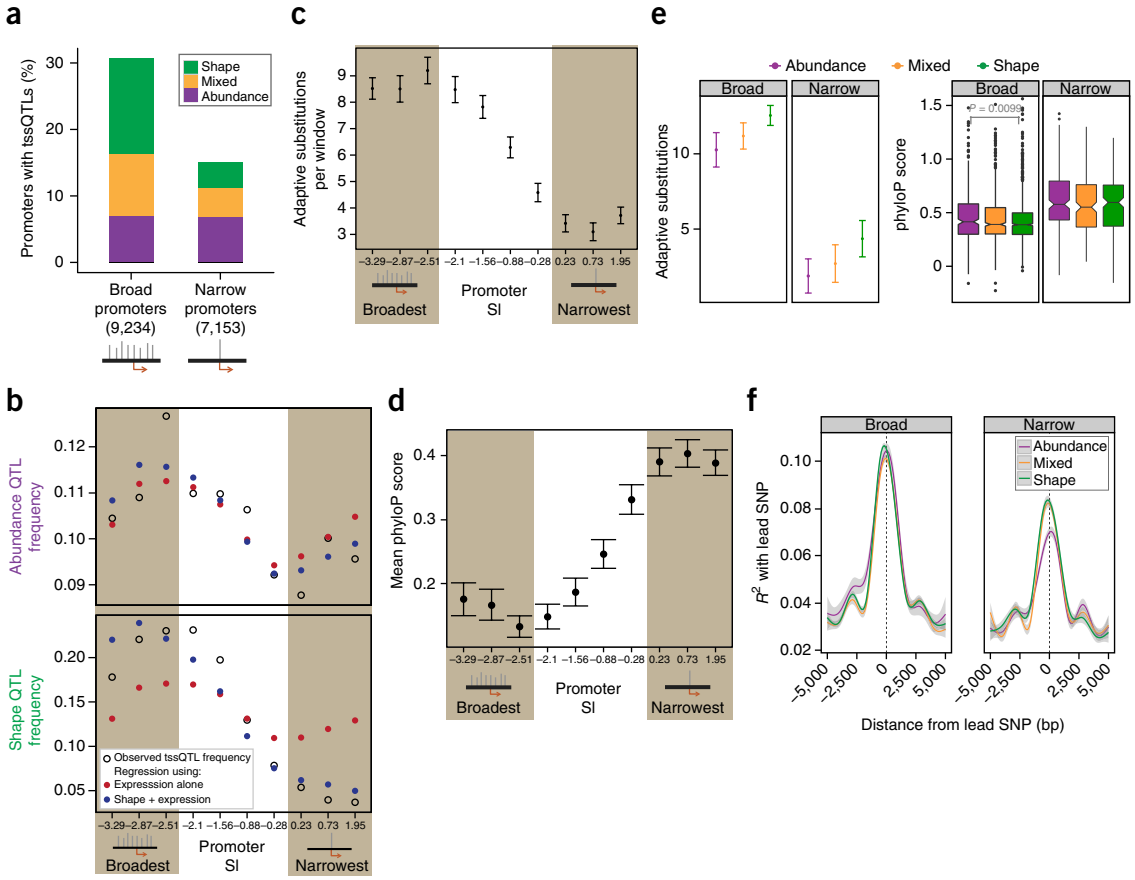

Figure 5 Promoter shape influences promoter evolution. (**a**) Percentages of broad or narrow promoters with tssQTLs, stratified by class. (**b**) Predicted (filled circles) and observed (open circles) tssQTL frequencies for promoters binned according to SI (Online Methods). Predictions obtained from logistic regressions considering gene expression level (red dots) or gene expression plus SI (blue dots). Brown shading indicates first three deciles (bottom 30%) and last three deciles (top 30%). (**c**) INSIGHT[46] analysis of promoters of different shapes. E(*A*), estimated number of adaptive substitutions. Error bars, standard error, estimated on the basis of the curvature of INSIGHT's likelihood function[47]. (**d**) PhyloP[48] score (between-species substitutions rates) calculated for different SI bins. Error bars, 95% CI (bootstrap). (**e**) Regions ±250 bp around tssQTLs located in promoter windows stratified by QTL type and promoter shape class. Left, E(*A*) parameter from INSIGHT; error bars as in **c**. Right, phyloP score distribution; notches, 1.58× interquartile range/$\sqrt{n}$; boundaries, first and third quartiles; whiskers, ±1.5× IQR. *P* values determined by Wilcoxon rank sum test. (**f**) Mean correlation coefficients between lead tssQTLs and proximal variants (±100 bp from most affected TSS) in broad or narrow promoters, calculated across 200 DGRP lines in a region ±5 kb from lead variant. Shading, 95% CI.

questions, we first classified promoters genome-wide as broad or narrow using a computational estimate of promoter shape (shape index (SI))[14] derived from the aggregated CAGE data from all lines and time points. The defined promoter shape classes showed a significantly higher proportion of broad promoters affected by tssQTLs (**Fig. 5a**, 2.2-fold difference, $P = 3.08 \times 10^{-91}$, Fisher's exact test), where the majority are shape QTLs, with smaller proportion of mixed QTLs. This result held true in an orthogonal, CAGE-independent classification of promoters into broad and narrow based only on their motif content[19,42] (**Supplementary Fig. 7**). Notably, the difference in the frequency of abundance tssQTLs between promoters of different shapes could be explained by differences in expression level between genes with broad (mainly housekeeping) and narrow (mainly tissue-specific) promoters (**Fig. 5b**), a property commonly observed in conventional eQTL studies[45]. The frequency of shape-changing tssQTLs, however, remained higher within broad promoters, even after accounting for differences in expression level (**Fig. 5b**).

This effect may reflect a biological property of broad promoters, such as increased robustness to genetic variation, but may also be explained by differences in power to detect shape tssQTLs in broad versus narrow promoters (**Supplementary Fig. 8**). To further explore this, we examined patterns of segregating variation and substitution rates, which can provide information on how selection has historically acted on promoters of different shapes. To examine positive and negative selection in broad versus narrow promoters, we applied a probabilistic extension of the McDonald–Kreitman test designed for noncoding sequence (INSIGHT)[46,47]. This showed only a weak relationship between promoter shape and the fraction of sites under selection and no relationship with the number of segregating sites under weak negative selection (**Supplementary Fig. 9** and **Supplementary Table 5**). Promoter shape classes differed markedly in the number of adaptive events inferred by INSIGHT: broad promoters had consistently higher substitutions rates than narrow promoters (**Fig. 5c** and **Supplementary Table 5**), a feature also observed in mice and humans[11] and consistent with more frequent positive selection acting on broader promoters. This conclusion was further supported by other metrics of selection (**Supplementary Fig. 10**), including interspecies conservation, which showed that narrow promoters are under stronger constraint than broad promoters[48] (**Fig. 5d**).

Genomic regions associated with abundance and shape tssQTLs also have distinct signatures of selection. Within each promoter class, regions flanking abundance QTLs were more constrained, whereas

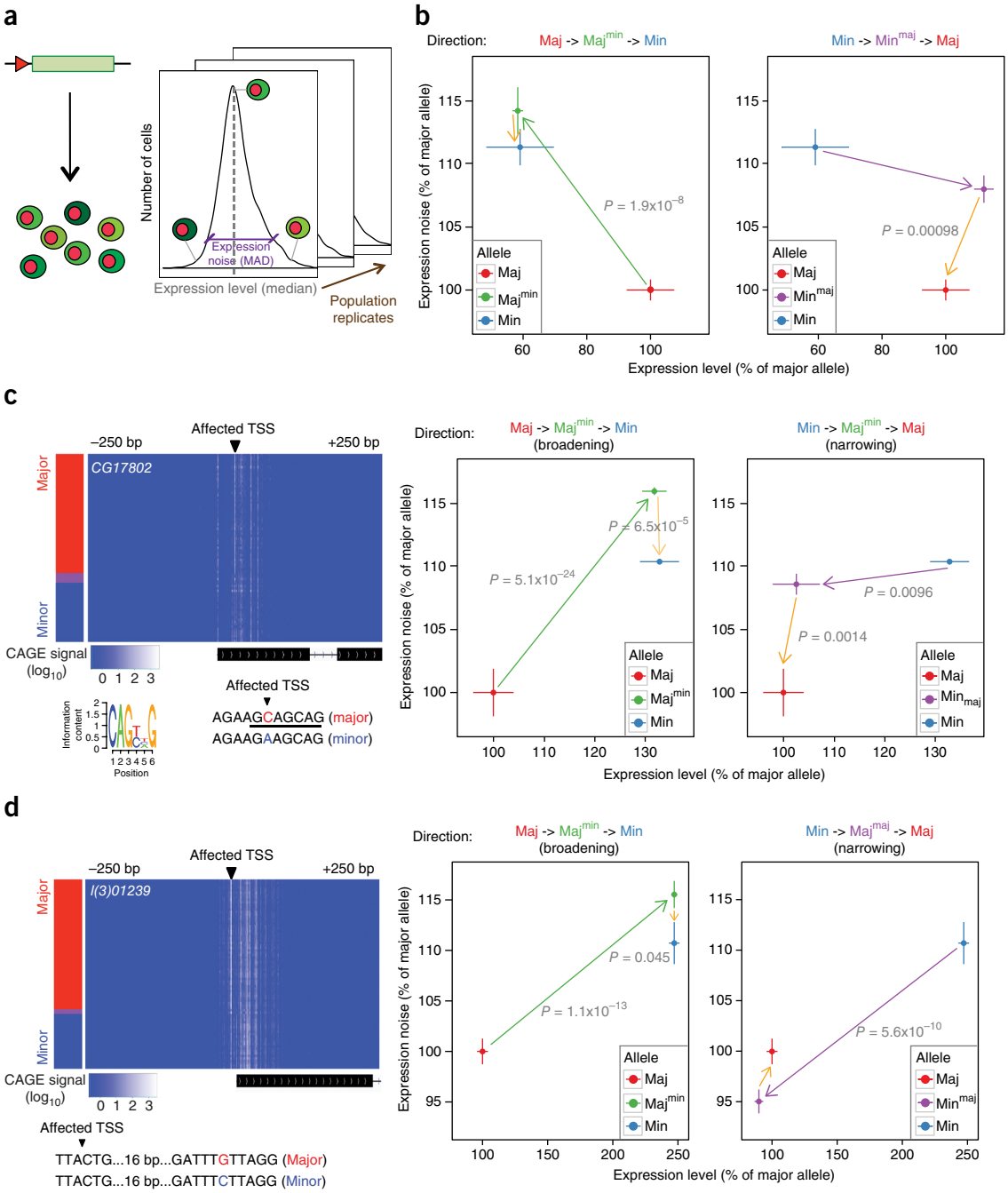

Figure 6 Changes in promoter shape are associated with increased expression noise. (**a**) Measurement of expression noise based on the assay in **Figure 4a**. MAD, median absolute deviation from the median. Each group consisted of 3 samples. Experiments were performed at least twice. (**b**) Transitions between shape haplotypes for the *CG12576* promoter. (**c**,**d**) Examples of promoters showing epistatic (**c**) and additive (**d**) effects of lead QTLs and remaining variants. Left, heat maps showing CAGE signal in a promoter window at the 10–12 h time point. For *CG17802* promoter (**c**), a novel TSS-positioned motif (underlined) is interrupted by the chr3R_13630942_SNP variant. For the *I(3)01239* promoter (**d**), mutation occurs downstream of the affected TSS (chr3L_11067477_SNP), although no motif could be identified. Right, expression level and noise for shape transitions. *P* values were assessed by Levene test for homogeneity of variances.

those around shape QTLs showed increased evidence of positive selection (**Fig. 5e** and Online Methods). This is supported by an increased number of adaptive substitutions in shape QTL promoter regions of both classes (**Fig. 5e**) and increased interspecies conservation around abundance QTLs, particularly within broad promoters, suggestive of purifying selection (**Fig. 5e**). The extent of LD between tssQTLs and flanking variants also differed according to tssQTL type and promoter shape class; genetic variants in the vicinity of broad promoters showed more pronounced LD with the lead variant than narrow promoters (**Fig. 5f**), consistent with longer haplotypes and again supporting more frequent, recent positive selection in broad promoters. When we stratified these LD patterns by tssQTL types, shape QTLs showed stronger linkage than abundance QTLs in narrow promoters (**Fig. 5f**).

Taken together, these data further support promoter shape as a functional property that leads to distinct selective pressures, with

broad promoters having longer LD blocks and more frequent substitutions, consistent with the actions of (recent) positive selection. Within each promoter class, genetic variants that change promoter shape are less constrained than those affecting transcript abundance, suggesting that shape as a trait may evolve more rapidly.

Variants changing promoter shape increase expression noise

The identified set of genetic variants affecting promoter shape provides a unique opportunity to examine the functional consequences of variation in shape for a given promoter, as opposed to comparing the general properties of genes with broad versus narrow promoters[1–3]. To explore this, we measured expression levels of promoter variants in thousands of individual cells, including population replicates, to estimate both expression level and expression noise, defined as the dispersion of single-cell expression values for a given promoter (**Fig. 6a**).

We focused on shape or mixed tssQTLs in which genetic variation alters promoter shape, studying seven broad promoters in four haplotypes. For example, the functional impact of the tssQTL disrupting motif 1 in the *CG12576* promoter (**Fig. 4b**) was examined in the two 'natural' haplotypes as well as in engineered promoters with the minor allele placed into the Major haplotype (Maj^{min}) and the major allele placed into the minor haplotype (Min^{maj}) (**Fig. 6b**). Using this reciprocal design, we assessed the effect of both alleles of tssQTL lead variants in either genetic background, thereby assessing the effect of changes in genetic context for both alleles (**Fig. 6b**).

We drew three conclusions from this experiment. First, the tssQTL disrupting motif 1 caused not only a difference in *CG12576* expression (**Fig. 4b**) but also a ~15% increase in expression noise (**Fig. 6b**). Second, when considering expression level and noise, different variants within the *CG12576* promoter showed distinct effects: the lead variant had an effect on both parameters (Maj versus Maj^{min} and Min versus Min^{maj}), but there are clearly additional variants within the promoter that have a specific effect to reduce noise as, for example, noise in the natural major haplotype was lower than the engineered promoter with the major lead variant in the context of the minor haplotype (**Fig. 6b**). Third, the effects of each transition were not additive but rather point to an epistatic interaction that compensates for the increase in noise of variants in their natural context.

Examining the remaining six promoters led to similar results. In five cases, we observed a significant increase in expression noise in Maj^{min} versus major (**Fig. 6b–d** and **Supplementary Fig. 11**). This increase was not always accompanied by an increase in expression, suggesting that elevated noise levels are not merely a consequence of higher expression. Notably, for all seven promoters, we observed interaction effects, within the promoter's natural heteroallelic context, on expression noise (**Fig. 6b–d** and **Supplementary Fig. 11**). In six out of seven cases, the natural promoter context significantly reduced expression noise for either the major or minor allele; the only exception was the *wls* gene promoter. In the promoter of *l(3)01239*, the effects of both transitions appeared to be additive; the variants surrounding the lead SNP decreased expression noise in the direction of major to minor while increasing it in the opposite direction (**Fig. 6d**). In contrast, four promoters (of *CG12576*, *CG17802*, *TfIIB* and *CG2469*) (**Fig. 6b**,**c** and **Supplementary Fig. 11**) showed patterns consistent with epistatic interactions among the lead variant and the remaining polymorphisms; for a given allele of the lead variant (both major and minor), the natural promoter context led to a decrease in noise with respect to the mutant construct.

This suggests that although genetic variants associated with a change in promoter shape often increase transcriptional noise, in natural promoter haplotypes they occur in the context of other *cis*-associated variants that reduce noise levels. Considering that polymorphisms in broad promoters showed strong LD (**Fig. 5f**), these results suggest that only particular combinations of alleles that maintain a minimum level of expression noise can reach high frequency in the population.

DISCUSSION

The regulatory regions controlling embryonic gene expression must impart robustness to developmental programs while leaving flexibility for evolutionary innovations. Broad and narrow core promoters have different signatures of evolution that reflect their regulatory needs. Narrow promoters, which are associated with tissue- or stage-specific genes, are evolutionary constrained. These genes have complex regulatory landscapes in which many developmental enhancers at diverse distances give input to the core promoter[49]. Moreover, narrow promoters typically have motifs working in close cooperation at fixed distances from each other, such that weakening of any one motif or changing the distance between them could have a major impact on transcriptional initiation and overall transcript levels. Conversely, ubiquitously expressed genes generally have broad promoters, with many regulatory elements located close to the promoter itself[10], implying a need to achieve robustness and evolvability within the promoter region. Our results suggest that this is achieved by the distributed initiation architecture of broad promoters: the effect of a genetic variant on one TSS is often buffered by other initiation sites within the same broad promoter, generating a shape effect with little or no impact on promoter strength (transcript abundance).

Promoter evolution appears to be influenced by a need to maintain low levels of expression noise, as recently shown for the *TDH3* promoter in yeast[50]. In that case, purifying selection acts to eliminate variants leading to an increase in noise. Here we showed that the situation is more complex in animal promoters; 'noisy' alleles associated with changes in promoter shape are common within natural populations. However, the impact of these mutations is partially buffered by the 'noise-reducing haplotypes' in which they reside, owing to allelic interactions with other variants that attenuate noise (**Fig. 6**). We propose that the dispersed architecture (which probably reduces the impact of mutations) and high frequency of adaptive substitutions in broad promoters may provide the substrate for the formation of such haplotypes, effectively allowing the presence of promoter shape variants in spite of the constraints associated with promoter function. We recently found a different instance of heteroallelic interactions within enhancer elements, which act to partially buffer deleterious effects on enhancer activity[32]. Taken together, these results suggest that allelic combinations with balancing effects are commonplace within haplotypes of diverse *cis*-regulatory elements, at least within drosophilids.

Overall, this study demonstrates how high-resolution measurements such as CAGE coupled with multivariate QTL mapping strategies enable probing of genetic effects on a new dimension of molecular variability, namely promoter shape. Many of the shape and mixed QTLs we identified are missed by conventional expression QTL mapping (**Fig. 1c**) yet highlight an important link between promoter shape and expression noise.

URLs. UCSC Genome Browser, https://genome.ucsc.edu/.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank all members of E.E.M.F.'s laboratory for discussions and comments. We are very grateful to P. Carninci and A.M. Suzuki for support regarding the CAGE protocol. This work was technically supported by the EMBL Genomics and Flow Cytometry Core Facilities. This work was financially supported by the European Research Council (FP/2007-2013)/ERC grant agreement 322851 (ERC advanced grant CisRegVar) to E.E.M.F. and post-doctoral fellowships from the Human Frontiers in Science Program (HFSP) and EMBO to J.F.D.

AUTHOR CONTRIBUTIONS

E.E.M.F., I.E.S., and O.S. designed the study, explored the results and prepared and edited the manuscript, together with contributions from all authors, including E.B. I.E.S. performed all CAGE and FACS experiments and led data analysis associated with QTL effects and frequency. J.F.D. developed the QTL calling and classification procedure with contributions from I.E.S., F.P.C., H.S., M.S., and O.S. H.S. and M.S. developed the wavelet analysis. D.H. performed all data analysis downstream of QTL calling. E.C. provided staged embryos and RNA. D.A.G. performed natural selection analysis.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Kadonaga, J.T. Perspectives on the RNA polymerase II core promoter. *Wiley Interdiscip. Rev. Dev. Biol.* **1**, 40–51 (2012).
2. Sandelin, A. *et al.* Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nat. Rev. Genet.* **8**, 424–436 (2007).
3. Lenhard, B., Sandelin, A. & Carninci, P. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nat. Rev. Genet.* **13**, 233–245 (2012).
4. Li, X. & Noll, M. Compatibility between enhancers and promoters determines the transcriptional specificity of gooseberry and gooseberry neuro in the *Drosophila* embryo. *EMBO J.* **13**, 400–406 (1994).
5. Hansen, S.K. & Tjian, R. TAFs and TFIIA mediate differential utilization of the tandem Adh promoters. *Cell* **82**, 565–575 (1995).
6. Butler, J.E. & Kadonaga, J.T. Enhancer-promoter specificity mediated by DPE or TATA core promoter motifs. *Genes Dev.* **15**, 2515–2519 (2001).
7. Zehavi, Y., Kuznetsov, O., Ovadia-Shochat, A. & Juven-Gershon, T. Core promoter functions in the regulation of gene expression of *Drosophila* dorsal target genes. *J. Biol. Chem.* **289**, 11993–12004 (2014).
8. Merli, C., Bergstrom, D.E., Cygan, J.A. & Blackman, R.K. Promoter specificity mediates the independent regulation of neighboring genes. *Genes Dev.* **10**, 1260–1270 (1996).
9. Juven-Gershon, T., Hsu, J.Y. & Kadonaga, J.T. Caudal, a key developmental regulator, is a DPE-specific transcriptional factor. *Genes Dev.* **22**, 2823–2830 (2008).
10. Zabidi, M.A. *et al.* Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. *Nature* **518**, 556–559 (2015).
11. Carninci, P. *et al.* Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.* **38**, 626–635 (2006).
12. Shiraki, T. *et al.* Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl. Acad. Sci. USA* **100**, 15776–15781 (2003).
13. Forrest, A.R. *et al.* A promoter-level mammalian expression atlas. *Nature* **507**, 462–470 (2014).
14. Hoskins, R.A. *et al.* Genome-wide analysis of promoter architecture in *Drosophila melanogaster*. *Genome Res.* **21**, 182–192 (2011).
15. Akalin, A. *et al.* Transcriptional features of genomic regulatory blocks. *Genome Biol.* **10**, R38 (2009).
16. Suzuki, Y. *et al.* Diverse transcriptional initiation revealed by fine, large-scale mapping of mRNA start sites. *EMBO Rep.* **2**, 388–393 (2001).
17. Cooper, S.J., Trinklein, N.D., Anton, E.D., Nguyen, L. & Myers, R.M. Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome. *Genome Res.* **16**, 1–10 (2006).
18. FitzGerald, P.C., Sturgill, D., Shyakhtenko, A., Oliver, B. & Vinson, C. Comparative genomics of *Drosophila* and human core promoters. *Genome Biol.* **7**, R53 (2006).
19. Ohler, U. Identification of core promoter modules in *Drosophila* and their application in accurate transcription start site prediction. *Nucleic Acids Res.* **34**, 5943–5950 (2006).
20. Ni, T. *et al.* A paired-end sequencing strategy to map the complex landscape of transcription initiation. *Nat. Methods* **7**, 521–527 (2010).
21. Rach, E.A., Yuan, H.-Y.Y., Majoros, W.H., Tomancak, P. & Ohler, U. Motif composition, conservation and condition-specificity of single and alternative transcription start sites in the *Drosophila* genome. *Genome Biol.* **10**, R73 (2009).
22. Nozaki, T. *et al.* Tight associations between transcription promoter type and epigenetic variation in histone positioning and modification. *BMC Genomics* **12**, 416 (2011).
23. Rach, E.A. *et al.* Transcription initiation patterns indicate divergent strategies for gene regulation at the chromatin level. *PLoS Genet.* **7**, e1001274 (2011).
24. Haberle, V. *et al.* Two independent transcription initiation codes overlap on vertebrate core promoters. *Nature* **507**, 381–385 (2014).
25. Main, B.J., Smith, A.D., Jang, H. & Nuzhdin, S.V. Transcription start site evolution in *Drosophila*. *Mol. Biol. Evol.* **30**, 1966–1974 (2013).
26. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).
27. Montgomery, S.B. *et al.* Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464**, 773–777 (2010).
28. Pickrell, J.K. *et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**, 768–772 (2010).
29. Francesconi, M. & Lehner, B. The effects of genetic variation on gene expression dynamics during development. *Nature* **505**, 208–211 (2014).
30. Huang, W. *et al.* Genetic basis of transcriptome diversity in *Drosophila melanogaster*. *Proc. Natl. Acad. Sci. USA* **112**, E6010–E6019 (2015).
31. Massouras, A. *et al.* Genomic variation and its impact on gene expression in *Drosophila melanogaster*. *PLoS Genet.* **8**, e1003055 (2012).
32. Cannavo, E. *et al.* Genetic variants regulating expression levels and isoform diversity during embryogenesis. *Nature* **541**, 402–406 (2017).
33. Takahashi, H., Lassmann, T., Murata, M. & Carninci, P. 5′ end-centered expression profiling using cap-analysis gene expression and next-generation sequencing. *Nat. Protoc.* **7**, 542–561 (2012).
34. Mackay, T.F. *et al.* The *Drosophila melanogaster* Genetic Reference Panel. *Nature* **482**, 173–178 (2012).
35. Engström, P.G., Ho Sui, S.J., Drivenes, O., Becker, T.S. & Lenhard, B. Genomic regulatory blocks underlie extensive microsynteny conservation in insects. *Genome Res.* **17**, 1898–1908 (2007).
36. van de Geijn, B., McVicker, G., Gilad, Y. & Pritchard, J.K. WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nat. Methods* **12**, 1061–1063 (2015).
37. Korte, A. *et al.* A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nat. Genet.* **44**, 1066–1071 (2012).
38. Casale, F.P., Rakitsch, B., Lippert, C. & Stegle, O. Efficient set tests for the genetic analysis of correlated traits. *Nat. Methods* **12**, 755–758 (2015).
39. Lippert, C., Casale, F.P., Rakitsch, B. & Stegle, O. LIMIX: genetic analysis of multiple traits. Preprint at http://biorxiv.org/content/early/2014/05/21/003905 (2014).
40. Huang, W. *et al.* Natural variation in genome architecture among 205 *Drosophila melanogaster* Genetic Reference Panel lines. *Genome Res.* **24**, 1193–1208 (2014).
41. Shim, H. & Stephens, M. Wavelet-based genetic association analysis of functional phenotypes arising from high-throughput sequencing assays. *Ann. Appl. Stat.* **9**, 665–686 (2015).
42. Ohler, U., Liao, G.C., Niemann, H. & Rubin, G.M. Computational analysis of core promoters in the *Drosophila* genome. *Genome Biol.* **3**, RESEARCH0087 (2002).
43. Dreos, R., Ambrosini, G. & Bucher, P. Influence of Rotational Nucleosome Positioning on Transcription Start Site Selection in Animal Promoters. *PLoS Comput. Biol.* **12**, e1005144 (2016).
44. Xi, L. *et al.* Predicting nucleosome positioning using a duration Hidden Markov Model. *BMC Bioinformatics* **11**, 346 (2010).
45. Brown, C.D., Mangravite, L.M. & Engelhardt, B.E. Integrative modeling of eQTLs and cis-regulatory elements suggests mechanisms underlying cell type specificity of eQTLs. *PLoS Genet.* **9**, e1003649 (2013).
46. Arbiza, L. *et al.* Genome-wide inference of natural selection on human transcription factor binding sites. *Nat. Genet.* **45**, 723–729 (2013).
47. 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**, 1159–1171 (2013).
48. Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20**, 110–121 (2010).
49. Spitz, F. & Furlong, E.E. Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.* **13**, 613–626 (2012).
50. Metzger, B.P.H., Yuan, D.C., Gruber, J.D., Duveau, F. & Wittkopp, P.J. Selection on noise constrains variation in a eukaryotic promoter. *Nature* **521**, 344–347 (2015).

ONLINE METHODS

5′ CAGE library preparation from DGRP embryos. We selected 81 genotyped fly lines from the *Drosophila* Genetic Reference Panel (DGRP)[40], including lines with the highest quality and depth of genome sequencing, avoiding pairs of highly-related lines, and avoiding lines with unusual levels of residual heterozygosity. Tightly staged embryos were collected at the specified time points, and RNA was extracted as described previously[32] and in the **Supplementary Note**.

5′ CAGE produces libraries containing strand-specific 27-nt-long tags, with their 5′ ends indicating the position of the 5′ cap-bound nucleotide and therefore the TSS[33]. We applied the protocol developed by Takahashi *et al.*[33], starting with 2.5 µg total RNA for each sample, with some minor modifications (**Supplementary Note**). A set of test PCRs was used to assess whether adequate numbers of cycles were performed for the preparative reaction. The final number of cycles was 11 for most of the samples, although it ranged from 9 to 12. To remove large fragments from the CAGE tags, we routinely included a second purification and size-selection step by adding 0.75 volumes of Agencourt AMPure XP beads (Beckman Coulter) for 5 min, recovering the supernatant and precipitating the CAGE tags by addition of extra 1.25 volumes of beads. This also removed minor contamination of adapters, leaving a unique peak of ~100 bp.

The final libraries were quality checked using a 2100 Bioanalyzer system with a HS DNA kit (Agilent) and quantified using a Qubit fluorometer with dsDNA HS reagent (Life Technologies). Libraries were sequenced (50-bp single end) by the EMBL Genomics Core Facility using a HiSeq 2000 Illumina Sequencer, adding 15% PhiX genomic DNA (molar concentration) to increase sequence complexity.

All oligos follow the sequences suggested previously[33], except that we extended the length of the barcodes from 3 to 6 nt. The upper oligos for the 5′ linkers were ordered in two variants (N_6 and GN_5), and mixed in 1:4 proportions.

Processing and mapping of CAGE reads is described in the **Supplementary Note**. The numbers of total and mapped reads, including barcode information for each sequence, are provided in **Supplementary Table 6**.

Defining active promoter windows. All individual CAGE libraries, from all time points and genotypes, were merged into a single BAM file. On the basis the combined sequencing data, we used a greedy search algorithm to select promoter windows (1,024-bp windows) for analysis, which together contained more than 99% of the total CAGE reads. The algorithm consist of the following steps:

1. Tally number of reads originating from each base in the genome.
2. Select the base with the highest read count.
3. Identify the 1,024-bp region centered on this base as the next phenotype window and replace count at all bases contained in this window with 0 in the tally made in step 1. Record the total fraction of all reads contained in phenotype windows.
4. Repeat steps 2–3 until total fraction of reads contained in all chosen phenotype windows is >99%.

This identified 13,508 active promoter windows (which generally represent clusters of TSSs (CAGE tags)) at one or more stages of embryogenesis (**Supplementary Table 1**). By definition, these contain 99% of all mapped CAGE signal (with >10 reads).

Testing for tssQTLs within a 200-kb *cis*-candidate window. Before the QTL analysis, we sequentially applied two approaches to reduce mappability differences between lines: UMM (developed here) and WASP[36]. A detailed description of both approaches is provided in the **Supplementary Note**.

For each promoter window, we tested common (minor allele frequency >5%) biallelic variants (Freeze 2.0 from the DGRP consortium) in 200-kb windows centered on the TSS with maximum CAGE signal within the active promoter window for association with the observed expression levels. Multiallelic variants were reduced to biallelic variants, considering the reference allele (Ref) versus all alternative (Alt) alleles. Association tests were performed using a multi-trait linear mixed model[35,37], jointly testing for genetic effects across developmental stages while accounting for population structure[37,39]. We considered two alternative phenotypes derived from the CAGE signal: the mean CAGE signal of the promoter window or the projections of the raw CAGE signal onto the first three PCs. All traits were adjusted for observed and hidden covariates using PEER[51].

Using these traits, we carried out three different analyses: (i) single-trait analysis of the mean CAGE signal for individual developmental stages, (ii) a joint analysis across the three time points using the mean CAGE signal, or (iii) the full model that performs joint genetic analysis across nine phenotypes derived from the PC-based phenotypes (three time points and three PCs). For full details of all models as well as the processing steps for the CAGE-derived phenotypes, see **Supplementary Note**.

Multiple testing, model comparison, and downstream evaluation of tssQTLs. We adjusted for multiple testing using a two-stage approach. First, for each promoter window, we adjusted for multiple testing in *cis*-regions using 10,000 permutations as described[26]. Specifically, we estimated *cis*-region adjusted *P* values by comparing the *P* value of the lead variant for a given test to the empirical distribution of null lead variants from 10,000 permutations. These *cis*-region adjusted *P* values were stored for each test and promoter window. Next, we applied Benjamini and Hochberg's method[52] to adjust for tests across promoter windows, thereby controlling for the global, genome-wide FDR. For single-trait methods, we adjusted for all tests and promoter windows.

We compared the performance of mean-based and PC-based phenotypes using two criteria. First, we compared the overall number of promoter windows with a significant association, thereby assessing the power to detect genetic associations. Second, we evaluated the ability of one method to recover the QTLs identified by other methods. The PC-based approach identified the majority (85%) of QTLs identified by the mean-based method and yielded increased power, but the converse was not true (**Fig. 1c**). Thus, for subsequent analysis, we considered lead QTLs obtained from the PC-based multi-trait analysis (at global FDR < 1%), yielding 4,526 promoter windows with a tssQTL.

We applied two additional filters to obtain a high-confidence set. First, tssQTLs at internal intragenic peaks were discarded, as they are probably the result of re-capping events in highly expressed genes (flagged as internal = TRUE in **Supplementary Table 2**). Although these associations may represent interesting genetic regulation of exonic promoter activity (as described by Carninci *et al.*[11]), these are outside the scope of our study focusing on canonical promoter function. Second, we identified variants that disrupt an EcoP15I restriction site in the same orientation as the one provided by the CAGE RT primer, which can cause technical associations with variation in CAGE signal. The creation of such sites in a genomic region proximal to the start site results in a significant enhancement of the CAGE signal, probably because they are the preferred matching sites owing to their proximity to the site on the 5′ adapter. These variants were therefore also filtered out (flagged as is_enzyme_artifact = TRUE in **Supplementary Table 2**). This resulted in 4,075 high-confidence tssQTLs, which are those included in **Supplementary Table 3** and used for all analysis in this study. Confidence intervals of lead QTL variants were derived by examining the sizes of regions with loci with association *P* values within one order of magnitude of the top *P* value for that region. Stage-specific effects were tested for these tssQTLs as described in the **Supplementary Note**.

Estimating single-base-pair effect sizes of significant QTLs using wavelets. Although the PC-based approach proved to be powerful and scalable for the large number of individual SNP × phenotype tests required here, this model has limitations. First, only the first three PCs were considered to test for associations, which may discard subtle changes in promoter shape. Second, although we expect the biology of TSS choice to be at least somewhat spatially dependent, this spatial structure is not explicitly accounted for by the PC-based method. Both limitations are effectively addressed in a wavelet-based method[41], which models CAGE data using a spatial model based on wavelets. By specifically targeting the tssQTL lead variants discovered by the PC-based approach, we limit the computational burden of the wavelet analysis.

The matrix of CAGE signal is decomposed into projections onto wavelet space, using the full representation of signal in this space, which allows a posterior of significant genetic effects to be estimated at single-base-pair

resolution. Moreover, as wavelets capture successively larger chunks of the base-pair space, estimation of effect sizes in this space enables analyzing spatially restricted sections of the phenotype window (in our case the 1-kb promoter region) to share the same effect size when effects are distributed on the lower-order (larger segment) WCs. This procedure results in estimations of effect size both in base-pair space and in the space of wavelets, both of which were used to characterize the architecture of tssQTLs (estimation is done in wavelet space, which can be subsequently transformed to base-pair space) (**Supplementary Fig. 4**).

For each significant tssQTL (FDR < 1%), we calculated effect sizes and Bayes factors for the association between the top associated SNP and the 1,024-bp region surrounding the primary TSS location. Forward and reverse strands were concatenated for each 1,024-bp region so that the total phenotype input matrix had 2,048 columns (1,024 bp for forward strand and 1,024 bp for reverse strand). As the majority of QTLs were consistent across stages, however, we focused our analysis on the middle (6–8 h) developmental stage. Other WaveQTL parameters were -f 2048 -numPerm 0 -fph 1 –gmode 1.

Criteria for classification of tssQTLs on the basis of pattern of wavelet and single-base effect sizes. To classify tssQTLs, we used estimates of effect size for individual WCs as well as single-base-pair effects. tssQTLs were classified as shape effects if the overall evidence was in favor of no association with the mean CAGE level ($\log_{10}$ Bayes factor (BF) < 0 for evidence of association with the lowest WC), or when only weak statistical evidence for an association with mean ($\log_{10}$ BF < 2.5) was observed, but much higher evidence was observed for effects at higher WCs (difference between maximum $\log_{10}$ BF and lowest WC $\log_{10}$ BF at least 10). At the other extreme, we use the directionality of the effect as the main determinant of abundance tssQTLs: when all the bases with a significant effect size were in the same direction (significance at the base-pair level corresponds to an approximate credible interval—calculated as posterior mean ±2 posterior s.d.—on base-level effect size that does not overlap 0) or the sum of effects on the primary direction was at least 10 times higher than in the second direction. Finally, the mixed tssQTL class included instances where the criteria for either both or none of the previous types were met.

***De novo* motif discovery.** The Meme Suite[53] was used for all *de novo* motif discoveries. We first defined high-confidence TSS clusters (**Supplementary Note** and **Supplementary Table 7**) and subsequently ran MEME-ChIP on these regions, centering on the most highly tagged site within each and extending outwards ±250 bp, using a maximum motif size of 15 bp and an E-value cutoff of 5×10^{-5}. A separate MEME-ChIP analysis was carried out for broad CAGE peaks, narrow CAGE peaks, and their union, in addition to contrasting broad against narrow, and vice versa. We ran CentriMo on all discovered promoter motifs.

To assess motifs for similarity, a column-wise correlation score was used as described previously[54]. To this we added 0.1, where motifs had overlapping positional enrichments, to yield a similarity score. From this score we constructed a similarity matrix and performed hierarchical clustering, cutting the tree at a distance of 0.6 to yield clusters of similar motifs. This procedure resulted in 59 motif clusters. Motifs with information content lower than 8 were discarded, giving a final result of 183 motifs grouped in 58 similarity clusters (**Supplementary Table 4**). After analyzing the properties of the different motifs (mainly the positioning with respect to main TSSs and the shape bias of promoters containing them) we split clusters 15, 24, 46, 47, and 56 in two (naming them 15 and 15a, 24 and 24a, and so on). Motifs were considered novel if they did not match promoter motifs discovered by Ohler *et al.*[42], FitzGerald *et al.*[18], or Down *et al.*[55].

Feature enrichments. Previous studies in human populations have shown that eQTLs are enriched in specific biochemical, annotational, and sequence features within the genome[45,56]. We applied a multivariate logistic regression framework similar to the method used by Brown *et al.*[45] to estimate the enrichments of tssQTLs in specific annotations (**Figs. 3a** and **4c**)

$$P(\text{Eqtl}) \sim \text{logit}^{-1}(\text{B}_0 + \text{Feature} + \text{AF} + \log_{10}(\text{Expr})) \quad (1)$$

Where Eqt1 is a binary variable denoting whether a given variant-window pair is a significant tssQTL, B_0 is the intercept, Feature is a binary variable denoting the variant's overlap with the feature, and Expr denotes to the total number of CAGE reads of the corresponding promoter window.

This model was fit to all variants tested for each CAGE window and models the probability of a given variant being a tssQTL lead variant on the basis of a given feature of interest and additional covariates. The variant's frequency (AF) and the total expression level were included as covariates in the regression framework, as both affect the power to detect tssQTLs. For promoters, introns, exons, and UTRs (**Fig. 3a**), we included ($\log_{10}$) distance to the CAGE window as an additional covariate. For chromatin immunoprecipitation (ChIP) and DNase I peaks (data from ref. 57), we did not include distance as a continuous covariate but instead tested proximal (within 1 kb of a CAGE window) and distal variants using separate models, because distal peaks are likely to have distinct functions and more often represent enhancers than promoters. For promoter-associated motifs (**Fig. 4c**), we reasoned that motif functionality should be limited to promoter-proximal regions so tested for enrichment relative to other variants within the CAGE clusters used to discover the motifs.

tssQTLs affecting core promoter motifs. We assessed changes to transcription factor binding sites by constructing local haplotype sequences, scanning them for the presence of motif matches, and assigning the best score in each haplotype to the corresponding variants. To select variants strongly affecting position weight matrix (PWM) scores, therefore, we counted only differences of 3 or more points in PWM scores between haplotypes. After obtaining all lead tssQTL variants affecting any of the core promoter motifs, we removed those located outside a ±100-bp window centered in the most affected TSS to enrich in the actual core promoter motifs instances. Only cases where the SI and CAGE signal were measurable for both genotypes were considered. For assessing turnover of Inr-like motifs, we took tssQTsL affecting motifs from clusters 10, 15, and 31 (**Supplementary Table 4**), located ±4 bp from the most affected TSS. For downstream promoter element (DPE)- and motif ten element (MTE)-like motifs (downstream-positioned motifs) we took tssQTLs affecting motifs from clusters 12, 24a, 26, 46a, 47a, 56a, 57, and 59 (**Supplementary Table 4**), located between 10 and 40 bp downstream of the most affected TSS. In both cases, for inclusion in this analysis, we considered only those cases where the most affected TSS was located at the promoter window center to avoid considering the effect of minor secondary peaks on the overall promoter region signal.

SI calculation and promoter classification. Shape index (SI) was calculated as previously described[14] using the aggregated CAGE signal for all time points and all lines (or all lines for major and minor allele separately, when indicated). Promoter regions were classified as broad if SI ≤ −1.5; otherwise they were scored as narrow (unique SI per window is shown in **Supplementary Table 1**). For binning promoters according to shape, we took 10% quantiles of the SI distribution.

Scans for selection. For analysis on promoters, scans were carried out using the initial 1-kb promoter windows. For tssQTL-centric analyses, we considered lead variants located inside the promoter window and took a region of ±250 bp centered on the SNP position. In all cases, exonic sequences were excluded.

INSIGHT analysis was carried out using scripts from Gronau *et al.*[47]. The scripts require input on both intraspecies variation, for which we used the complete genotypes of the DGRP lines[34], and interspecies variation, for which we used the 12 sequenced *Drosophila* genomes[56]. We modified the scripts to use nearby fourfold-degenerate sites rather than flanking intergenic regions as neutral proxies, as the high density of the *D. melanogaster* genome means that most of these regions are not evolving neutrally.

As an alternative method to look for positive and negative selective forces, we used the phyloP tool[48], as previously described[58]. In addition, we downloaded phastCons scores from the UCSC Genome Browser.

Flow cytometry and single-cell quantification of gene expression. We designed a reporter system for single-cell measurements in transient transfection, consisting of a single plasmid (TIPR-cherry) harboring both an sfGFP coding sequence under the control of the test promoter and an mCherry

doi:10.1038/ng.3791

coding sequence under the control of a constitutive promoter. Expression values per transfected cell were measured by analytical FACS for thousands of individual cells and calculated as $\log_{10}$ sfGFP/mCherry. For testing statistically significant differences between constructs, we used the average expression value from the population as a magnitude. All promoters were tested in at least 2 independent (different day) experiments; in each experiment we typically transfected each construct in duplicate or triplicate, and each transfection was considered an independent sample. Details of the TIPR-Cherry plasmid construction and promoter cloning, as well as the measurement protocol, are provided in the **Supplementary Note**.

Measuring expression noise. Single-cell expression levels from clonal cells can be used to determine expression noise, defined as the dispersion of single-cell expression values for a given promoter over thousands of cells. All promoters were tested in at least 2 independent (different day) experiments. In each experiment, to accurately determine the dispersion within the population, we required at least 3 independent transfections with a minimum of 15,000 cells per variant. Because the distribution of expression values for some promoters or promoter variants was not entirely normal (for example owing to long tails), we applied a conservative measure of dispersion, namely the median absolute deviation from the population median (MAD). Accordingly, when comparing changes in expression noise, we used the population median to indicate changes in expression levels. The qualitative results did not differ if we used population mean and s.d. as measures of expression level and noise, respectively.

To determine statistically significant differences in variance, we used the missMethyl package[58], which implements a Levene test for homogeneity of variances between groups, using replicate information to distinguish technical effects on variability from those due to construct genotype. We contrasted each mutant haplotype (Maj^{min} and Min^{maj}) against both natural haplotypes (major and minor).

Data availability. Raw data, comprising 317 demultiplexed files of CAGE-seq, have been submitted to EBI's ArrayExpress under accession number E-MTAB-4787. The mapped CAGE clusters and other processed data, including raw *P* values for all variants as well as HTML table for convenient visualization of tssQTL plots and information, are available at http://furlonglab.embl.de/data.

51. 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**, 500–507 (2012).
52. Storey, J.D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA* **100**, 9440–9445 (2003).
53. Bailey, T.L., Johnson, J., Grant, C.E. & Noble, W.S. The MEME Suite. *Nucleic Acids Res.* **43**, W39–W49 (2015).
54. Pietrokovski, S. Searching databases of conserved sequence regions by aligning protein multiple-alignments. *Nucleic Acids Res.* **24**, 3836–3845 (1996).
55. Down, T.A., Bergman, C.M., Su, J. & Hubbard, T.J. Large-scale discovery of promoter motifs in Drosophila melanogaster. *PLoS Comput. Biol.* **3**, e7 (2007).
56. Gaffney, D.J. *et al.* Dissecting the regulatory architecture of gene expression QTLs. *Genome Biol.* **13**, R7 (2012).
57. Thomas, S. *et al.* Dynamic reprogramming of chromatin accessibility during Drosophila embryo development. *Genome Biol.* **12**, R43 (2011).
58. Phipson, B., Maksimovic, J. & Oshlack, A. missMethyl: an R package for analyzing data from Illumina's HumanMethylation450 platform. *Bioinformatics* **32**, 286–288 (2016).