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

Schor IE, Degner JF, Harnett D, Cannavo E, Casale FP, Shim H, Garfield D, Birney E, Stephens M, Stegle O‡, Furlong EE‡ (2017) Nature Genetics, In press [‡ Co-corresponding authors]

Supplementary Tables Description

Supplementary Table 1: Promoter regions used for tssQTL calling

This is a GFF3 file containing the 1024bp windows used to call the tssQTL.

Extra columns:

- gene_name: Name of associated gene
- gene_id: ID of associated gene
- Window: 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)

NB: These columns contain no information but are GFF3 mandatory: source,type,score,phase

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
- 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: Location of the site with highest single-base-pair effect, relative to the center of the Window
- Strand: Strand where the window center was defined
- Set: Whether the QTL is shape, mixed or abundance
- Peakid: 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: 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: Cluster of similar motifs which includes this motif
- Motif_word: Consensus sequence for the motif
- EnrichedInSet: Set(s) of TSS which were used as test and background sets to discover the motif
- E_value: Motif’s associated E-value
- Algorithm: Whether the motif was found using MEME or DREME
- IC: Information Content of motif
- Prevalence: Number of TSS clusters with this motif
- Strand_bias: 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: E value of the motifs positional enrichment

- Position: Location of the motifs enriched zone
- Bin_width: Width of the motifs positional enrichment zone
- Shape_shift: Difference between the global mean of shape index, and the mean shape index for promoters with this motif
- Novel: 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; States ‘TF’ if the motif only has a match to a TF motif according to TomTom; States ‘Yes’ if none of the previous cases
- 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 7: CAGE TSS clusters

This is a GFF3 file containing the cage peaks or TSS clusters.

Extra columns:

- 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: 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

NB: These columns contain no information but are GFF3 mandatory: source,type,score,phase