Natural genetic variation affects transcription start site - positioning, shape and usage during *D. melanogaster* development

Jacob Degner*[1], Ignacio Schor*[2], Enrico Cannavo[2], Heejung Shim[1], Francesco Paolo Casale[3], Ewan Birney[3], Matthew Stephens[1], Oliver Stegle[3], Eileen E Furlong[2]

*These authors contributed equally
[1]Human Genetics, University of Chicago; Chicago USA
[2]Genome Biology Unit European Molecular Biology Laboratory; Heidelberg Germany
[3]European Bioinformatics Institute; Cambridge UK

Initiation of transcription is one of the most important steps in the regulation of gene expression across metazoans. This process is tightly controlled through the integration of transcription factors and co-factors with promoter-proximal and distal *cis*-regulatory elements, such as enhancers and silencers, and the participation of chromatin modifications and non-coding RNAs. Given that much of this regulation acuminates through the modulation of RNA Polymerase II activity at transcription start sites (TSSs), it is not surprising that sequence variants in the neighborhood of TSS are most likely to impact gene expression. Knowledge of TSS usage and variation is therefore an essential step toward understanding complex trait genetics and gene-regulatory evolution. While high-quality experimentally derived annotations of the locations of TSSs exist for several organisms, we do not fully understand the sequence features that define the position, number and usage of individual TSSs. To address these limitations, we performed a large-scale study of population variation in TSS usage during embryonic development using Cap-Analysis of Gene Expression (CAGE) and quantitative trait locus (QTL) mapping across a collection of *Drosophila melanogaster* wild isolates, bred to homozygous near clones.

CAGE experiments yield a uniquely high-resolution characterization of the pattern of transcription initiation and we have developed a novel computational approach to map QTLs that is capable of detecting both changes in magnitude and changes in spatial distribution of CAGE signal. Briefly, we treat the measures of CAGE signal at all base-pairs within a transcription initiation cluster as a set of correlated features that vary both across time points and across individuals. For QTL mapping, we treat the projections onto the top three principle components (PCs) of CAGE measurements within a TSS cluster as a reduced representation of this high-dimensional feature space. We perform joint association mapping on these PCs as if they were multiple correlated phenotypes. We account for multiple layers of correlation structure among the phenotypes and individual lines using a flexible linear mixed model framework. Overall, we find a marked improvement in power to detect QTL comparing single developmental timepoint analysis to simultaneous analysis of three developmental timepoints. Additionally, we find that we identify both more and fundamentally different types of QTL using the PC-based method as compared to association mapping against the mean CAGE level at TSS clusters.

While we find that our PC-based method for detecting QTL is powerful and efficient at detecting a variety of QTL types, it has several drawbacks for direct interpretation. In order to be scalable to genome-wide analysis, it was necessary to use a reduced representation of the full data. In order for tests to

adhere to model assumptions and for test statistics to remain well calibrated, it was necessary to do a quantile-based transformation of the data. For these reasons, we turned to a second method for characterizing the effects of each QTL across the base pairs of a TSS cluster. Here, we used a wavelet-based approach to estimate posterior distributions of effect size across bases in such a way as to keep effects in the original data space, use all data together in effect size estimation, and share information on effect size locally across bases.

Using these posterior estimates of effect size, we classify QTL into several distinct groups with differing biological interpretations. We find that only a subset of QTL affect the total level of initiation sites in a cluster, while the remaining QTL change only the spatial distribution of initiation. We refer to these as directional and redistribution type QTL, respectively, and find that these two major classes differ in cis-regulatory features and contributions from distal enhancer-based regulation or core-promoter elements.

The simplest interpretation of directional-type QTL is that, much like traditional expression quantitative trait loci (eQTL), they are act on the total level of RNA in the cell. In order to test this assumption, we performed RNA-seq experiments on a subset of the individual lines to measure whole-gene steady state expression levels. Indeed, we find a highly significant bias toward RNA-seq levels across individuals in the same direction as predicted by the directional CAGE QTL ($P < 1E-09$). While the comparison of concordance in direction of effect between CAGE and RNA-seq is highly significant, we were surprised to see a substantial fraction ($\sim$40%) of directional QTL are discordant in effect direction in the RNA data. We reasoned that the discordant directional QTL could represent inter individual differences in abortive transcription levels or inter-individual differences in proportion of capped transcripts perhaps as a result of partial degradation. To distinguish among these hypotheses, we performed poly-A capture followed by CAGE on a subset of individual lines giving us three data types to compare. Poly-A capture CAGE and CAGE experiments measure only capped transcripts while polyA CAGE and RNA-seq measure only polyadenylated transcripts. To identify QTL that were indeed unique to the polyadenylated or capped experimental data types, we reanalyzed our data in a multi-trait QTL analysis framework meant to estimate the probability of sharing of QTL across different data-types.

In addition to characterizing directional effects on TSS clusters, we describe a large class of QTL that are uniquely detected using CAGE and the analysis methods described here that specifically modify the spatial distribution of transcription initiation without affecting overall levels of initiation. We estimate that approximately 1/3 of the total QTL we detect are of the redistribution class. We find that redistribution-type QTL are more concentrated in the core promoter region compared to directional QTL and that redistribution type QTL are specifically enriched for known core promoter elements. Interestingly, redistribution of CAGE signal happens both among bases on the same strand, but also across strands leading to QTL affecting expression at two genes simultaneously. Taken together, this study reveals enormous diversity in 5' ends of transcripts between individuals and highlights transcription initiation as an important sequence-directed aspect of gene-regulation.