BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 11 2012, pages 1446–1454
doi:10.1093/bioinformatics/bts155

Genome analysis

Advance Access publication April 5, 2012

LOESS correction for length variation in gene set-based genomic sequence analysis

Anton Aboukhalil[1,2] and Martha L. Bulyk[2,3,4,*]

[1]Department of Aeronautics and Astronautics, Massachusetts Institute of Technology, Cambridge, MA 02139, [2]Division of Genetics, Department of Medicine, [3]Department of Pathology, Brigham and Women's Hospital and Harvard Medical School and [4]Harvard-MIT Division of Health Sciences and Technology (HST), Harvard Medical School, Boston, MA 02115, USA

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Sequence analysis algorithms are often applied to sets of DNA, RNA or protein sequences to identify common or distinguishing features. Controlling for sequence length variation is critical to properly score sequence features and identify true biological signals rather than length-dependent artifacts.

Results: Several cis-regulatory module discovery algorithms exhibit a substantial dependence between DNA sequence score and sequence length. Our newly developed LOESS method is flexible in capturing diverse score-length relationships and is more effective in correcting DNA sequence scores for length-dependent artifacts, compared with four other approaches. Application of this method to genes co-expressed during *Drosophila melanogaster* embryonic mesoderm development or neural development scored by the Lever motif analysis algorithm resulted in successful recovery of their biologically validated cis-regulatory codes. The LOESS length-correction method is broadly applicable, and may be useful not only for more accurate inference of cis-regulatory codes, but also for detection of other types of patterns in biological sequences.

Availability: Source code and compiled code are available from http://thebrain.bwh.harvard.edu/LM_LOESS/

Contact: mlbulyk@receptor.med.harvard.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on September 22, 2011; revised on March 5, 2012; accepted on March 24, 2012

1 INTRODUCTION

An important task in computational biology is to score DNA, RNA or protein sequences for a variety of features. These features may range from relatively basic sequence properties such as GC content or evolutionary conservation, to more complex measures of putative biological function, such as the presence of transcription factor binding sites (TFBSs) or of TFBS clusters organized into cis-regulatory modules (CRMs), e.g. transcriptional enhancers (Frith *et al.*, 2003; Hallikas *et al.*, 2006; Johansson *et al.*, 2003; Kielbasa *et al.*, 2010; Sinha *et al.*, 2003; Warner *et al.*, 2008; Zhou and Wong, 2004). Such general computational endeavors often face the central challenge of properly accounting for input sequences of varying lengths.

*To whom correspondence should be addressed.

Accounting properly for dependence between a sequence score and its length is critical to identify a true biological signal, rather than a correlation artifact. A greater sequence length offers greater opportunities for finding good local matches by chance alone. For instance, artifacts associated with length dependence were noted in the early days of searching sequence databases. A nearly linear correlation was often observed between the best local similarity score and the length of the sequence match in the queried database, even when those sequences were unrelated (Durbin, 1998; Pearson, 1995). More recently, adjusting for sequence length dependence was also found necessary in analysis of next-generation RNA sequencing (RNA-Seq) data (Mortazavi *et al.*, 2008). Since the number of reads mapped to a gene depends on its transcript abundance and length (Cloonan *et al.*, 2008; Gao *et al.*, 2011; Lee *et al.*, 2011; Mortazavi *et al.*, 2008), transcript length may confound gene expression analysis (Oshlack and Wakefield, 2009), with a bias for calling a higher proportion of long transcripts as expressed more highly.

In the *Drosophila melanogaster* genome, for instance, genes have varying non-coding sequence lengths, ranging from 10^2 to 10^6 bp (Fig. 1). A prior study found that genes with more complex expression patterns tend to have longer non-coding sequences (Nelson *et al.*, 2004). In prior investigations of CRMs and cis-regulatory codes (Philippakis *et al.*, 2006; Warner *et al.*, 2008), it was observed that for various input TFBS motifs there was a notable correlation between a gene's non-coding sequence length (i.e. total

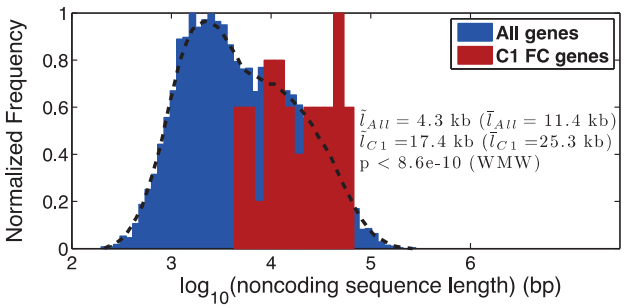

Fig. 1. In *D. melanogaster* the length of a gene's non-coding sequence varies across four orders of magnitude, from $\sim 10^2$ to 10^6 bp. For instance, the *C*1 somatic mesoderm FC gene set has on average lengthier non-coding sequences, with a median of 17.4 kb (25.3 kb mean) as compared with 4.3 kb (11.4 kb mean) for all genes in the genome, a difference that must be accounted for to avoid computational artifacts.

upstream, downstream and intronic sequence) and its CRM score. We here find that this length-dependence artifact is more general than previously expected. It can affect several other CRM discovery algorithms (Fig. 2) and to our knowledge is not accounted for by currently existing algorithms for CRM prediction and inference of cis-regulatory codes.

In fields other than computational prediction of CRMs, a few approaches have been proposed to correct for dependence between two variables (Cox and Hinkley, 1974; Pearson, 1901; Seok *et al.*, 2002). Here, we focus on approaches that can be applied in our context. Given a sequence score of interest, S, and a sequence length, L, one may adjust S for potential length dependence by subtracting $\ln(L)$ as in Equation (1) (Durbin, 1998), shown below. Alternatively, one may regress a linear function, with slope m and intercept b, and adjust the score as shown in Equation (2) (Durbin, 1998; Pearson, 1995).

$$S^{\text{adj}} = S - \ln(L) \quad (1)$$

$$S^{\text{adj}} = S - m\ln(L) - b \quad (2)$$

In scoring of protein sequence similarity, several variations on Equations (1) and (2) yielded different performances depending on the scoring matrices and gap penalty combinations used (Shpaer *et al.*, 1996). However, the general consensus was that even simple log-length normalization improved performance and increased the selectivity of the queries (Shpaer *et al.*, 1996). In RNA-Seq analysis, Gao *et al.* (2011) either subtract or divide their score by a length-dependent correction factor $c\sqrt{L-d}$, where L is the gene length, d is the sequenced tag length (~25–32 bp), and c is an empirically derived weight (a constant). There also exists a diversity of mathematical procedures based on principal component analysis (PCA) (Cox and Hinkley, 1974; Nam, 2010; Pearson, 1901; Salomon, 2007) to convert potentially correlated datasets into uncorrelated ones. However, available approaches either assume an a priori model of the correlation relation (e.g. linear or of order $^1/_2$), or might overlook potential underlying structure of the correlation relation, thereby being inflexible or inapplicable to problems with varying correlation instances. Applying normalization methods that are blind to the correlation structure in any particular context may remove the overall correlation but might still yield undesired artifacts due to data interpretation (described further below in Section 3.3).

We present a general and flexible approach to adjust for different dependence relations between two variables of interest (e.g. a sequence score and its length). Our strategy derives the local correlation structure of the dataset under consideration by fitting a LOESS curve (Cleveland, 1979; Cleveland and Devlin, 1988) to the score versus length (S–L) scatter plot, and dividing the scores by the values of this fitted curve at the corresponding length value. We compare the performance of our LOESS approach to that of four other normalization methods and benchmark them in the context of cis-regulatory code discovery in embryonic mesoderm development and nervous system development in the fruit fly *D. melanogaster*.

2 METHODS

We first present the cis-regulatory code and CRM discovery frameworks, followed by the datasets and normalization methods used in this study.

2.1 Cis-regulatory code discovery

To computationally predict cis-regulatory codes, we use a gene set (GS) enrichment analysis framework (Subramanian *et al.*, 2005) implemented in the Lever algorithm (Warner *et al.*, 2008). Lever is a general method that infers cis-regulatory codes for predefined sets of putatively co-regulated genes (e.g. genes co-expressed under different conditions, with similar biological function, within the same pathway, etc.), by considering over-represented TFBS motif combinations within putative CRMs, typically predicted by the CRM prediction algorithm PhylCRM (Warner *et al.*, 2008). A main advantage of using the algorithms Lever and PhylCRM (detailed in Sections 2.2.1 and 2.3, respectively) is that they provide the capability of exploring several tens of kb of non-coding sequences, as opposed to other algorithms (Ho Sui *et al.*, 2005; Kreiman, 2004) that might intrinsically consider only proximal promoter regions of a fixed length (e.g. ±1 kb around transcription start sites). Although Lever was developed in conjunction with PhylCRM, Lever can be used with any CRM prediction algorithm.

2.2 CRM discovery

2.2.1 PhylCRM CRM prediction algorithm: Briefly, the PhylCRM score $\text{PS}(\text{MC}_i, w)$ is a statistical measure of TF binding site density and evolutionary conservation for a given TFBS motif combination MC_i over a sequence window of width w.

Specifically, $\text{PS}(\text{MC}_i, w)$ represents the –log(P-value) of observing another window of size w with a higher density of conserved motif matches for MC_i in the entire non-coding genome. Thus, the PhylCRM score by itself accounts for the window length over which it is computed. To assess conservation of a TF binding site within a set of aligned genomic sequences, PhylCRM uses the probabilistic evolutionary model MONKEY (Moses *et al.*, 2004). This model computes the extent to which each position along a putative binding site is both conserved and a close match to the TFBS motif. The PhylCRM score $\text{PS}(\text{MC}_i, w)$ is then based on a Fuzzy Boolean logic integration of the MONKEY score across individual motif occurrences, for a particular combination MC_i and a window size w (Warner *et al.*, 2008).

To identify the most likely CRM for a given motif combination MC_i and gene g_l, we define the 'gene CRM score' $S(\text{MC}_i; g_l)$ as the highest (maximum) PhylCRM score. Note that $S(\text{MC}_i; g_l)$ is the maximum PhylCRM score evaluated over a continuous range of user-specified window sizes (e.g. 50–1000 bp) and over the entire non-coding sequence of a gene g_l.

Thus, a gene CRM score of zero, $S(\text{MC}_i; g_l) = 0$, indicates that the gene has no sequence windows that contain TFBS motif matches for the motif combination MC_i. In a biological sense, a gene CRM score of 0 indicates that MC_i is unlikely to directly regulate the gene g_l through a cis-acting DNA regulatory element.

2.2.2 Other CRM prediction algorithms: To assess whether the length variation problem affects other CRM prediction algorithms, we focused on the following algorithms that were reviewed and performed well in Klepper *et al.* (2008) and Su *et al.* (2010): MSCAN (Alkema *et al.*, 2004), Cluster-Buster (Frith *et al.*, 2003) and STUBBMS (i.e. Stubb with conservation) (Sinha *et al.*, 2006). These algorithms use different computational strategies and types of inputs, as described in Section 3.1. Default parameter settings were used. For each algorithm and for each motif combination, the 'gene CRM score' represents the score of a gene's highest scoring, predicted CRM.

2.3 Lever analyses

Lever is a cis-regulatory code finder (Warner *et al.*, 2008) based on GS enrichment analysis (Subramanian *et al.*, 2005). Lever identifies significant over-representation of particular combinations of evolutionarily conserved TFBS motif occurrences within putative CRMs in the non-coding regions of predefined foreground gene sets. Formally, given an input collection of N foreground gene sets $F_1, \ldots, F_N$, a corresponding collection of N disjoint background gene sets $B_1, \ldots, B_N$ is constructed to form the compound foreground–background gene sets $GS_m = \{F_m, B_m\}, m \in \{1, \ldots, N\}$. For each

given motif combination MC_i, the gene CRM score $S(MC_i; g_l)$ is used to rank every gene g_l in each GS_m; the ranks are subsequently used to compute the Wilcoxon–Mann–Whitney (WMW) statistic [equivalent to the area under the receiver operating characteristic curve (AUC)]. The AUC reflects the likelihood of observing a higher-scoring CRM in a randomly chosen foreground gene $g_l \in F_m$ with respect to a randomly chosen background gene $g_l \in B_m$. Thus, for a given motif combination, a high AUC value indicates that the foreground genes tend to have higher-scoring CRMs than their background gene counterparts; a motif combination with a high AUC value is thus hypothesized to be a cis-regulatory code of the foreground GS under investigation. To assess the statistical significance of observing such an AUC value (or higher) by chance, Lever performs many permutations (default 1000) of the labels between the foreground and background genes in each compound GS_m, re-computes the AUC statistic on each permuted dataset, and estimates the resulting false discovery rate (FDR; Storey, 2002).

2.3.1 Parameter settings: For all Lever and PhylCRM analyses, window widths ranged between 50 and 1000 bp. To assess evolutionary conservation, the genomes of all 12 sequenced fly species (Adams *et al.*, 2000; Celniker *et al.*, 2002; Clark *et al.*, 2007; Richards *et al.*, 2005) were used. To identify putative TFBS motif matches, an information content cutoff of one SD below the motif's position weight matrix (PWM) average score was used (Stormo, 2000). The PWM average score represents the mean information score over all 4^k sequence variants, where k is the length of the PWM. We used Lever in two broadly defined biological contexts—embryonic mesoderm development and neural development—that encompassed seven different foreground GSs. For the mesodermal GSs, we used Lever to evaluate all Boolean AND motif combinations up to five-way combinations. For the neural GSs, we used Lever to evaluate Boolean AND as well as OR motif combinations up to seven-way combinations since the putative code could involve an OR logic. We used default settings for all other Lever parameters.

2.3.2 Genome pre-processing: The *D. melanogaster* genome (BDGP Release 5/dm3) and its 11 MultiZ *Drosophila* alignment genomes were pre-processed as in (Philippakis *et al.*, 2006). Both coding and repetitive sequences were masked. For each gene, the non-coding region that we examined included its introns and the entire upstream and downstream intergenic regions until the nearest flanking genes.

2.3.3 Permuted motif analysis: We conducted this analysis to further assess the specificity of a given motif combination for its target GS.

Initially, for each pairing of motif combination and GS, the FDR on the AUC Lever statistics was derived by randomly relabeling the foreground and background genes, yet using real TFBS motifs.

In this additional analysis, we generated column-permuted motifs (i.e. maintaining nucleotide composition, but shuffling the nucleotide positions) as negative controls for each input, real TFBS motif. We repeated the Lever analyses 100 times using different versions of these shuffled motifs. An empirical P-value equal to $N/100$ can be derived, where N is the number of shuffled (negative control) motif combinations that scored more significantly than their original real motif counterparts.

Therefore, for a given GS, a motif combination was considered a putative cis-regulatory code if it proved statistically significant with an AUC ≥ 0.6, FDR $q \leq 0.05$ (Warner *et al.*, 2008) and an empirically derived $p \leq 0.05$.

2.4 Biological datasets

We conducted our computational study on two biological systems—an embryonic somatic mesoderm founder cell (FC) GS and neural GSs—for which prior experimental support exists for their cis-regulatory codes (Castro *et al.*, 2005; Halfon *et al.*, 2000, 2002; Philippakis *et al.*, 2006; Reeves and Posakony, 2005). We also performed our analysis for additional systems—an additional embryonic somatic mesoderm GS and cardiac mesoderm GSs—where we sought to discover potentially novel codes (Zhu *et al.*, 2012). Our analyses center on *D. melanogaster* GSs particular to four different cell types in the embryonic mesoderm and to three cell types in the peripheral nervous system.

2.4.1 Gene sets: The *C*1 (cluster 1) GS (Philippakis *et al.*, 2006) is a collection of 37 genes that are co-expressed in a subset of somatic mesoderm FCs and that respond similarly across 12 genetic perturbations of myogenesis (Estrada *et al.*, 2006). The *FCM-all* GS comprises 104 genes expressed in fusion-competent myoblasts (FCMs) (Estrada *et al.*, 2006), cells which fuse with individual FCs to form somatic muscles. *PC-only* and *CC-only* are sets of 39 and 35 genes that are expressed only in pericardial cells (PCs) or cardial cells (CCs), respectively (Ahmad *et al.*, in press; Zhu *et al.*, 2012), the two main cell populations of the embryonic heart vessel in *Drosophila*. The *PNC* GS comprises 44 genes validated by in situ hybridization to be expressed in proneural cluster (PNC) cells, which have the potential to adopt neural cell fates in the larval wing (Reeves and Posakony, 2005). The *SOP* GS is a subset of 26 *PNC* genes expressed in sensory organ precursor (SOP) cells in the peripheral nervous system. The *non-SOP* GS is the remaining subset of 18 *PNC* genes expressed in the inhibited cells, which adopt an epidermal cell fate (Reeves and Posakony, 2005). The genes in each GS are provided in Supplementary Table S1.

2.4.2 Transcription factor binding site motifs: Our input collection of TFBS motifs that might regulate the four mesodermal GSs under consideration included motifs for the TFs activated by the Wg, Dpp and Ras pathways: T cell factor (dTCF), Pointed (Ets motif), Twist (Twi), Tinman (Tin) and Mothers against dpp (Mad). For the neural GSs, we additionally considered motifs for Achaete/Scute (Ac/Sc) and Suppressor of Hairless (Su(H)). All motifs (Supplementary Table S2) were obtained from (Philippakis *et al.*, 2006).

2.5 Normalization methods

We examined the performance of Lever in identifying putative cis-regulatory codes using the following five general normalization approaches. For each motif combination under consideration, normalization is applied accordingly to all genes under investigation in all compound foreground–background gene sets $g_l \in \bigcup_{m=1}^{N} GS_m$.

2.5.1 Length matching: Previously length matching (LM) was applied to genes analyzed by the CodeFinder algorithm to evaluate potential cis-regulatory codes (Philippakis *et al.*, 2006); CodeFinder was subsequently scaled-up and further developed in Lever (Philippakis *et al.*, 2006; Warner *et al.*, 2008). For every foreground GS F_m a corresponding background GS B_m was constructed. Each background set is chosen so as to contain at least 20 times as many genes as the foreground set, and so that the distribution of non-coding sequence lengths of the foreground and background sets are matched. We use LM here as the baseline method for length normalization, and the generated length-matched background sets are used in all other normalization methods examined in this article.

2.5.2 Linear log-regression normalization: This approach was initially applied in Lever because of its success in protein sequence homology searches (Durbin, 1998; Pearson, 1995). Considering only genes with non-zero CRM scores, a linear curve with slope m and intercept b is regressed to the scatter plot of 'gene CRM score' versus 'non-coding sequence length'. Then, for each gene g_l the CRM score is adjusted as in Equation (2). Linearity between sequence score and length is here assumed.

2.5.3 PCA-based normalization: Based on PCA for multivariate data, a linear transformation can be used to decorrelate two (or more) variables, despite a non-linear correlation between them (Cox and Hinkley, 1974; Nam, 2010; Pearson, 1901; Seok *et al.*, 2002). Thus, given a matrix X_{mxn} with n variables (here, $n=2$: 'gene CRM score' and 'non-coding sequence length') for m genes, and a matrix Z_{mxm} with all of its entries set equal to 1, we apply

the following two-step mapping:

$$D = X - \frac{1}{m} ZX \tag{3.a}$$

$$T = D\left(D^{\mathrm{T}} D\right)^{-1/2} \tag{3.b}$$

to transform the variables so that they have zero correlation with each other (and zero mean). We chose this method as a general representative for other more recent state-of-the-art variations of PCA-based decorrelation methods, such as DECO (Nam, 2010). DECO has recently been used to decorrelate expression datasets in GS analyses. This method conducts eigenvalue decomposition of a covariance matrix, followed by linear transformations and eigenvalue rescaling and truncation. DECO is similar, for instance, to the Mahalanobis-based transformation (Salomon, 2007) for pixel decorrelation in image compression, but is rather slow and may present numerical instabilities despite the sophisticated shrinkage covariance estimator used (Schafer and Strimmer, 2005). Therefore, because of the similarity in decorrelation performance between DECO and the general PCA-based approach described by Equations (3.a) and (3.b) (Supplementary Fig. S1), we chose to use the latter method.

2.5.4 Length division: This method is similar in spirit to the recent RNA-Seq correction approach from (Gao *et al.*, 2011). There, the score is divided by a correction factor $c\sqrt{L-d}$, with L being the effective gene length, d the sequenced tag length (∼25–32 bp), and c an empirically derived constant (0.0301 or 0.0436). However, since the score-length correlation relation is not generally of order $1/2$, normalizing by $\sqrt{L_l}$ or $\sqrt{\log_{10}(L_l)}$ is ineffective (Supplementary Fig. S2). Therefore, we instead use LD as a general approach where, for each gene g_l, the CRM score is divided by $\log_{10}(L_l)$, the logarithm of the associated non-coding sequence length.

2.5.5 LOESS-fit normalization: We implemented this method because of its flexibility in capturing the score-length relationship without any a priori assumptions. First, considering only genes with non-zero CRM scores, a LOESS curve is regressed to approximate the local dependence relationship between the gene CRM score and the non-coding sequence length. Then, for all genes, the CRM score is divided by the value of the LOESS curve at the corresponding non-coding sequence length. A second-order LOESS curve with a smoothing parameter of 0.5 was used. We found the normalization to be robust to the value of the smoothing parameter considering a range of values (Supplementary Fig. S3), and we chose this parameter to avoid overfitting the data (Supplementary Fig. S4).

2.6 Normalization performance evaluation

Normalization is applied using information from all available genes. However, using all genes to normalize and assess the post-normalization score-length correlation would not necessarily represent a stringent evaluation of performance. Therefore, we instead compute this correlation using different sampled foreground–background GSs. Because subsequent enrichment analyses are conducted on these same foreground–background GSs, we must ensure the successful removal of the correlation at the GS level as well, since any remnant correlation might bias the results.

3 RESULTS

Among other genes, mesodermal genes and *non-SOP* genes have complex gene expression patterns (Casal and Leptin, 1996; Philippakis *et al.*, 2006) and lengthy non-coding sequences (Fig. 1 and Supplementary Fig. S5). For example, as shown in Figure 1, the $C1$ GS has a median non-coding sequence length of 17.4 kb as compared with 4.3 kb for all *D. melanogaster* protein-coding genes (WMW P-value $< 8.6 \times 10^{-10}$). Thus, GS analyses that examine non-uniform sequence lengths must correct for length variability to avoid potential artifacts.

We first show that dependence between a gene's CRM score and its non-coding sequence length affects several CRM prediction algorithms. We then compare our LOESS method to other normalization methods in correcting for this dependence and in identifying validated cis-regulatory codes.

3.1 Length dependence of gene CRM score affects a variety of CRM discovery algorithms

As a result of length variability, longer sequences provide the opportunity for an algorithm to predict higher scoring CRMs by chance alone. Many well-known algorithms display to some extent a correlation between their gene CRM score and the gene's non-coding sequence length (Fig. 2).

Although results are depicted for only MC_i = 'Ets AND Twi AND Tin' (Fig. 2), the correlation effect is consistent across a wide variety of motif combinations, despite the different algorithmic strategies and types of inputs used (Supplementary Fig. S6). For instance, PhylCRM uses its own heuristic scoring scheme that exploits phylogenetic information across all 12 fly species, based on the MONKEY evolutionary model (Moses *et al.*, 2004). STUBBMS (Sinha *et al.*, 2006) employs a different phylogenetic strategy, uses only one other fly species (here, the distantly related *Drosophila grimshawi*), and is based on a hidden Markov model (HMM). Cluster-Buster (Frith *et al.*, 2003) is also an HMM scheme, but unlike STUBBMS does not use the Baum–Welch algorithm on the DNA sequence to derive the state transition probabilities. Cluster-Buster heuristically derives these probabilities from expected input motif cluster structure parameters, and from background distributions from sliding windows along the non-coding genome. Both Cluster-Buster and MSCAN (Alkema *et al.*, 2004) do not use phylogenetic information. MSCAN is not HMM-based and uses its own heuristic to evaluate the combined statistical significance of dense clusters of motif matches within a sliding window. Despite these considerable differences in the underlying algorithms and types of inputs used, their calculated gene CRM scores tend to correlate with non-coding sequence length.

3.2 Length matching

As an initial solution to this length variability problem, LM was devised to generate for each foreground GS a background set with a matched non-coding length distribution (Philippakis *et al.*, 2006). However, LM by itself does not solve the distinct problem of having a gene score depend on its non-coding sequence length. For example, considering the union of the $C1$ genes and their

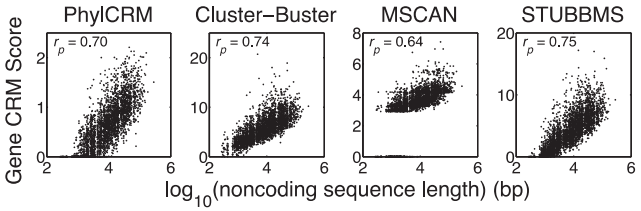

Fig. 2. Gene CRM scores generated by four different CRM prediction algorithms are highly correlated with genes' non-coding sequence length, across all *D. melanogaster* genes, considering the TFBS motif combination Ets AND Twi AND Tin as a representative example.

length-matched background genes for the three-way AND motif combination MC_i = 'Ets AND Twi AND Tin', there remains a considerable score-length correlation ($r_p \sim 0.62$), close to that of all genes under consideration ($r_p \sim 0.7$) (Supplementary Fig. S7). LM alone is insufficient to adjust for the score-length correlation artifact across different foreground and background gene sets GS_m for combinations of real (Fig. 3B, panel i; Supplementary Fig. S8, panels a and c) and shuffled motifs (Fig. 3B, panel ii; Supplementary Fig. S8, panels b and d). Highlighting the $C1$ genes emphasizes the length dependence of gene CRM scores (Supplementary Fig. S9). Since length dependence may confound subsequent motif analyses, further normalization is necessary in addition to LM.

3.3 Log-length regression

For different gene sets and motif combinations, applying log-length regression (LLR) reduces the median Pearson correlation to ~0.1–0.2 in the case of AND combinations (Fig. 3B, panels iii and iv; Supplementary Fig. S8, panels e and f) and to ~0 in the case of OR combinations (Supplementary Fig. S8, panels g and h). Despite this considerable reduction in correlation, LLR is incompatible with our cis-regulatory code discovery framework on two fronts, which may lead to serious artifacts.

First, in a gene set enrichment-based cis-regulatory code discovery framework, two genes with no candidate CRMs are interpreted as functionally equivalent 'non-target' genes, irrespective of their non-coding sequence length. That is, despite non-coding sequence length differences, both genes are unlikely to be targeted directly by a motif combination of interest. Therefore, it is desirable that 'non-target' genes receive an equal score (e.g. zero, both pre- and post-normalization) irrespective of non-coding sequence length. However, LLR adjusts zero-scoring 'non-target' genes to acquire unequal and length-dependent scores (Fig. 3A, panels iii and iv; encircled in blue in Supplementary Fig. S10, panels i and iii, and ii and iv), thereby breaking down the desired property that 'non-target' genes receive an equivalent score irrespective of non-coding sequence length.

Second, it is desirable for a 'potential target gene' to acquire an adjusted score greater than that of a 'non-target' gene. However, with LLR, some putative 'non-target' genes may rank higher than 'potential target genes' (Fig. 3A, panels iii and iv; Supplementary Fig. S10, panels iii and iv). The effect is stronger for the Ets AND Twi AND Tin motif combination (Supplementary Fig. S10, panel iv), for which very short 'non-target' genes (encircled in blue) score more highly than most of the 'potential target genes'.

As a result, in gene set analyses, artifacts may occur where strong enrichment is observed for a motif combination in a given GS, even though no or very few genes have motif occurrences in their non-coding sequence. This issue is more pronounced for AND combinations which, by the nature of the scoring scheme, tend to have a higher number of zero-scoring genes.

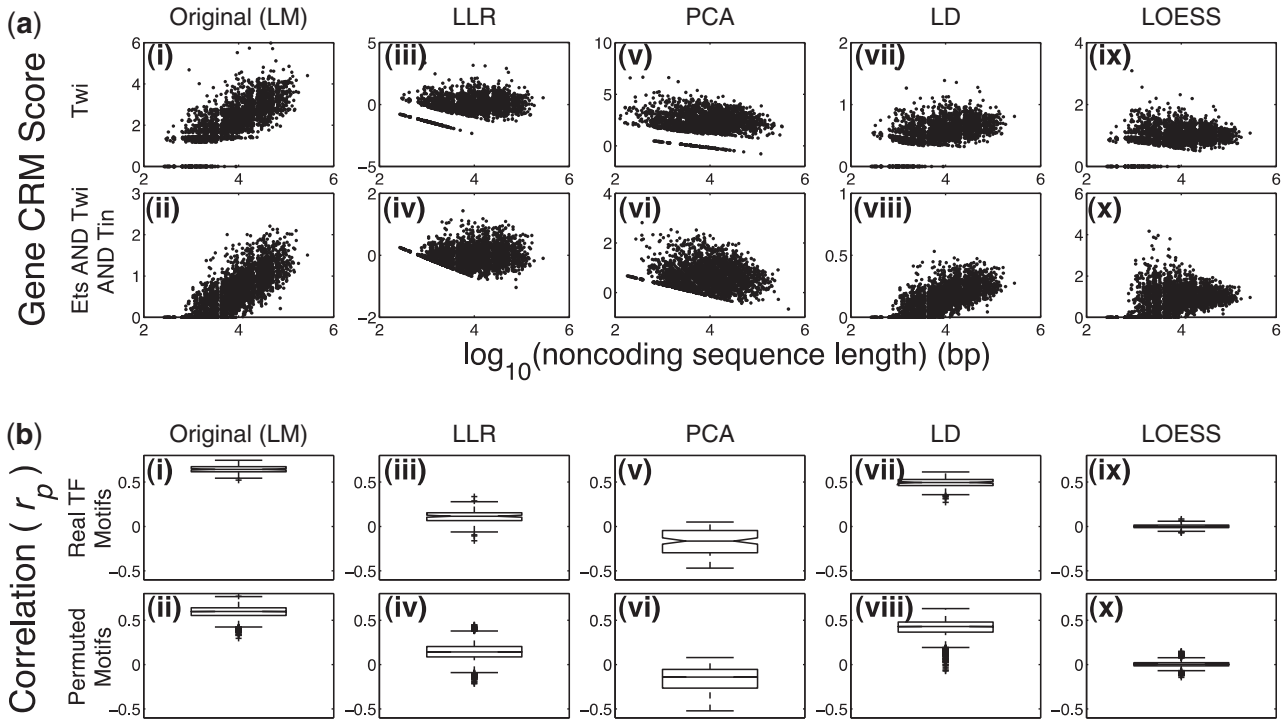

Fig. 3. Comparison of different length correction methods. (i, ii) LM (Original), (iii, iv) LLR, (v, vi) PCA-based normalization (PCA), (vii, viii) LD and (ix, x) division by the LOESS-fit curve (LOESS). (**a**) Overall effects of the different length correction methods on the Pearson correlation coefficients between the gene CRM score (for PhylCRM) and the noncoding sequence length when considering the Twist motif alone (*above*) and the motif combination Ets AND Twi AND Tin (*below*). (**b**) Box plots summarizing the Pearson correlation coefficients between the gene CRM score (for PhylCRM) and noncoding sequence length resulting from 31 different real TFBS motifs or motif combinations for each of four different GSs (i.e. $31 \times 4 = 124$ pairings of motifs and GSs) (*above*), and from 100 shuffled motifs or combinations thereof constructed and used in place of the real TFBS motifs (i.e. $31 \times 4 \times 100 = 12400$ pairings of motifs and GSs) (*below*).

3.4 PCA-based normalization

Although the PCA-based approach generates relatively similar *S–L* scatter plots as does LLR (Fig. 3A, panels iii and iv, and iv and vi), in general, it performs worse than LLR with a wider interquartile range of the Pearson correlations. For different GSs and motif combinations, applying PCA reduces the median Pearson correlation to −0.2–0 (range ∼−0.5–0.2) in the case of AND combinations (Fig. 3B, panels v and vi; Supplementary Fig. S8, panels i and j) and to a median of 0 (range −0.4–0.3) in the case of OR combinations (Supplementary Fig. S8, panels k and l). Nonetheless, PCA-based methods suffer from the same potential artifacts as LLR (Supplementary Fig. S10).

3.5 Length division

To avoid potential artifacts that may arise in LLR and PCA (due to an inherent subtraction operation [Equation (2) and Equation (3.a), respectively], we tested the LD normalization method, whereby the gene CRM score S_i is divided by $\log_{10}(L_i)$. However, since the relationship between S_i and $\log_{10}(L_i)$ is not linear, LD does not eliminate the correlation effectively (Fig. 3A, panels vii and viii; Fig. 3B, panels vi and vii; Supplementary Fig. S8, panels m to p), resulting in a median $r_p \sim 0.3$–0.5. Therefore, instead of using higher-order or exponential curves, for which one needs to select the function and its parameters a priori,we sought to find a fitting function that is general, simple and flexible.

3.6 LOESS-fit normalization

We pursued LOESS as an alternate approach to better characterize and reduce the length-dependence relationship in gene CRM scores. We use the LOESS-fit approach to derive empirically the relationship between the gene CRM score and $\log_{10}(L_i)$. This fit is then used to adjust the CRM score for each gene. After normalization by the LOESS curve, the Pearson correlation between the gene CRM score and non-coding length has a relatively tight distribution around zero across all seven GSs for the real TFBS motifs (i.e. 505 MC–GS pairings; Fig. 3B, panel ix; Supplementary Fig. S8, panels q and s; range ±0.1) and for the shuffled motifs (i.e. 50 500 MC–GS pairings; Fig. 3B, panel x; Supplementary Fig. S8, panels r and t; range ±0.2). LOESS-fit normalization was the only method for which the median of the distribution of Pearson correlations was consistently 0, as evidenced by a WMW AUC close to 0.5 (Supplementary Table S3). In sum, the LOESS-fit approach is the most consistent and most effective method for correcting the length-dependence correlation.

We also tested various exponential, logarithmic, trigonometric and hyperbolic functions based on the LOESS fit, and found that a simple LOESS fit reduced correlation the most (data not shown).

3.7 Cis-regulatory code motif analysis

To further assess the utility of the LOESS-fit correction of gene CRM scores, we considered the corrected gene CRM scores in cis-regulatory code motif analysis in real biological contexts.

3.7.1 LOESS-fit correction helps recover known cis-regulatory codes: First, we first focused on the GS *C*1 for which there is prior experimental support for the AND motif combination Ets AND Twi AND Tin acting as a cis-regulatory code (Philippakis *et al.*, 2006). The original version of Lever (i.e. length-correction by LM) (Warner *et al.*, 2008) does not yield any statistically significant codes for the GS *C*1 considering five TFBS motifs of relevance in mesoderm development [Fig. 4A (LM); Supplementary Table S4]. However, Lever with LOESS-fit normalization yields three statistically significant codes involving these five motifs (AUC ≥0.60, $q \leq 0.05$; Fig. 4A–C; Supplementary Tables S5–S7) and includes the Ets AND Twi AND Tin cis-regulatory code for the *C*1 GS. To further assess the robustness of our results, we ran the Lever analysis using 100 shuffled versions of the five real TFBS motifs as negative controls. No shuffled motifs or combinations thereof scored more significantly than the three putative codes ($p < 0.01$; Supplementary Tables S5–S7). Furthermore, comparing these significant codes to those identified by a prior cis-regulatory code evaluation framework (CodeFinder) that used a different CRM scoring scheme and phylogeny for only three fly species (ModuleFinder) (Philippakis *et al.*, 2005; Philippakis *et al.*, 2006), yielded generally consistent results (Supplementary Table S8).

To assess the generality of LOESS-fit correction of gene CRM scores, we applied it in a second biological context. In *D. melanogaster* neural development, there is experimental support for the motifs Ac/Sc and Su(H) participating in a cis-regulatory code for the *non-SOP* GS (Castro *et al.*, 2005; Reeves and Posakony, 2005). We ran Lever with LOESS-fit correction to investigate both AND and OR combinations of these motifs along with the five motifs analyzed for the mesodermal GSs. We repeated the analysis with 100 shuffled versions of each of the seven real TFBS motifs. We recovered Ac/Sc (alone), Su(H) (alone), as well as the Boolean AND and OR combinations of Ac/Sc and Su(H) among the top statistically significant codes, with five or fewer of 100 of their shuffled motif combinations scoring more significantly (i.e. $p \leq 0.05$) (Supplementary Table S7).

3.7.2 LOESS-fit normalization performs favorably against other competing methods in recovering known cis-regulatory codes: We repeated all our previous PhylCRM and Lever analyses using LLR, PCA and LD to correct for length dependence (Supplementary Tables S9–S11). In general, all these forms of normalization improved detection of putative codes as noted by an increase in the AUC and a decrease in the *q*-value with respect to the original Lever analyses performed using LM alone (Fig. 4). For the *non-SOP* GS, LOESS-fit compares favorably to the other normalization methods in recovering validated codes in a statistically significant fashion. For the validated Ets AND Twi AND Tin code for *C*1, the LOESS-fit approach performed the best (Fig. 4A). However, as shown above, both LLR and PCA normalization may yield undesired artifacts (Sections 3.3 and 3.4), and LD does not remove length dependence as effectively as does LOESS-fit (Sections 3.5 and 3.6).

3.7.3 LOESS-fit correction improves cis-regulatory code results obtained using MSCAN, Cluster-Buster and STUBBMS CRM prediction algorithms: We repeated our Lever analyses using three other well-known CRM prediction algorithms: MSCAN (Alkema *et al.*, 2004), Cluster-Buster (Frith *et al.*, 2003) and STUBBMS (Sinha *et al.*, 2006). We compared their performance in recovering validated cis-regulatory codes with versus without LOESS-fit correction. In general, LOESS-fit correction did not drastically change the AUC values, but tended to improve the statistical significance (reducing the *q*-value) of highly scoring code predictions (Supplementary Tables S12–S14; Fig. 5). We note that

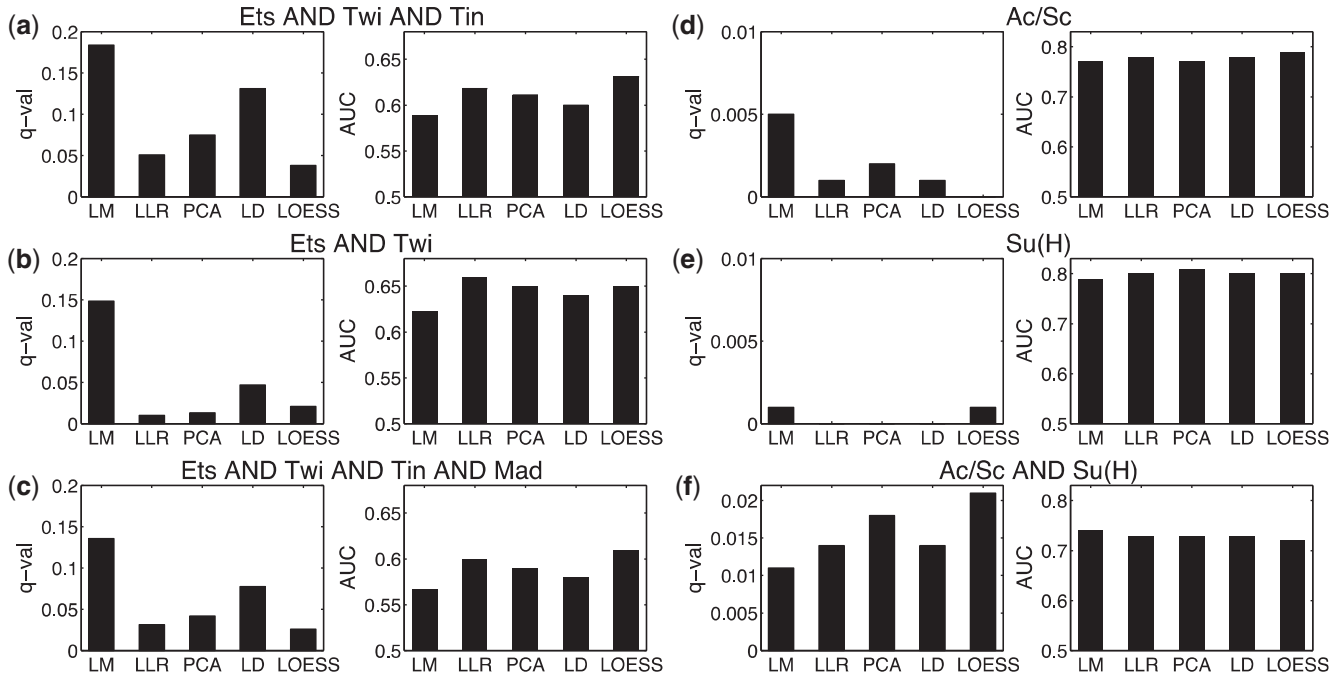

Fig. 4. Comparison of the different normalization methods in recovering known cis-regulatory codes in the context of *Drosophila* mesoderm development (**a**–**c**) and neural development (**d**–**f**). Good performance is characterized by both a high AUC value and a low *q*-value (Section 2.3).

these results must be interpreted with care, as LOESS-fit correction cannot improve the intrinsic ability of any CRM prediction algorithm to identify true biological signals; however, LOESS-fit correction does effectively reduce artifactual length dependency of the resulting CRM scores.

4 DISCUSSION

Dependence of the score of a biological sequence on the length of the sequence may confound or conceal true biological signals. Here, we have shown that this problem occurs for several well-known CRM prediction algorithms. We have presented a general method to adjust for length-dependence of a DNA sequence score. Our LOESS-fit approach is consistent in reducing the correlation between sequence length and score, not only in comparison to other methods (e.g. length division (LD), PCA-based methods and linear-log regression (LLR)), but also in absolute terms (Fig. 3B; Supplementary Fig. S8). In the context of cis-regulatory code discovery, LOESS-fit correction allowed the Lever algorithm (Warner *et al.*, 2008) to recover experimentally validated codes in two different biological contexts in *D. melanogaster* (Fig. 4).

Compared with other normalization approaches, LOESS-fit performed favorably in recovering the validated codes (Fig. 4), and does not suffer from the artifacts that LLR and PCA may yield (Supplementary Fig. S10). In addition, LOESS-fit correction is useful in application to a variety of CRM prediction algorithms in analyses aimed at identifying putative cis-regulatory codes (Fig. 5).

Our GSs were based on relatively stringent biological evidence. We focused on GSs comprising genes that are expressed in relatively homogeneous cell types, that were confirmed by in situ hybridization, and that have prior experimental evidence for a cis-regulatory code. GSs could be constructed alternatively to comprise genes with shared Gene Ontology (GO) annotation terms or pathway involvement.

Some major advantages of LOESS, as compared with other regression methods, are that it does not call for a priori specification of a fitting function and that it is a flexible and simple fitting approach used for a wide range of purposes (Gijbels and Prosdocimi, 2010). Our data points are generally dense (e.g. Fig. 2). However, potential limitations of the LOESS-fit method may arise in small or sparse datasets, where LOESS might overfit or inaccurately fit the data points. On another level, to avoid generating non-monotonic LOESS curves that might overfit the data, we have empirically chosen the value of the smoothing parameter to be robust across various motif combinations and GSs (Supplementary Figs S3 and S4). Potential artifacts might arise in a dataset with varying data density, such as around the extreme ends of a distribution, where LOESS interpolation might deviate from the sparse points that it tries to fit. We seldom encountered such deviations, which occurred mostly for genes with particularly short non-coding sequence lengths and with low gene CRM scores, where the LOESS curve could go below zero. To guard against such deviations, we implemented a small saturation threshold, $t \sim 0.01$, so that the LOESS curve fit_LOESS $= \max(t,$ fit_LOESS) always remains >0 (or a small value, t, above 0).

We considered the possibility of score inflation where the LOESS curve came close to zero. However, since the scores fit by the LOESS curve are also generally close to zero and of the same order of magnitude, their ratio will not be inflated. Of course, an outlier score that is much greater than the LOESS curve will be inflated; however;

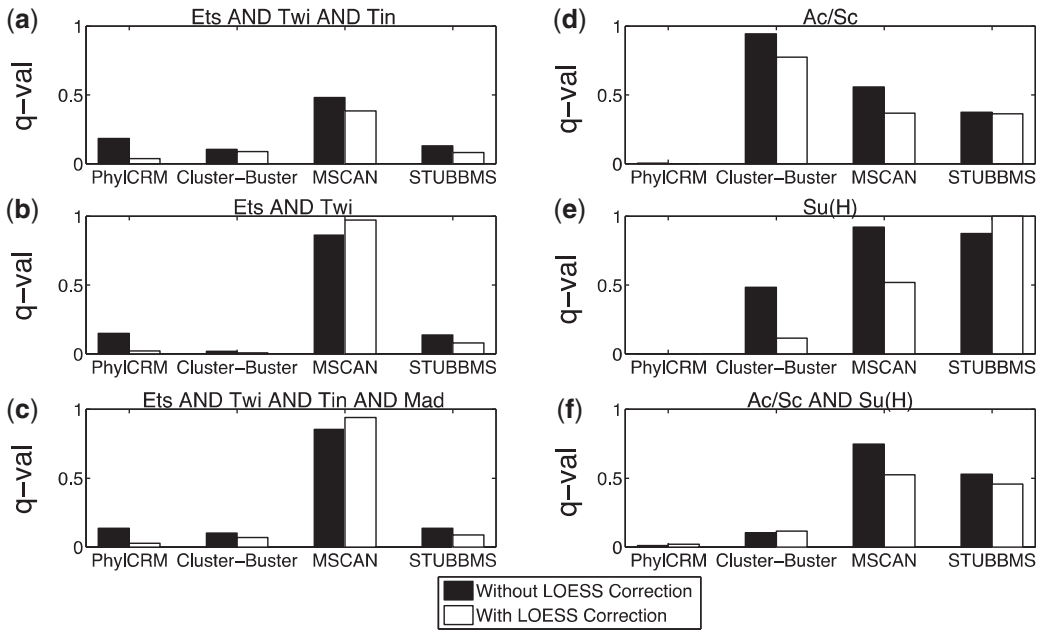

Fig. 5. Comparison of the performance of four different CRM prediction algorithms in identifying known cis-regulatory codes without LOESS correction (black bars), and with LOESS correction (white bars). In general, LOESS correction improves performance as noted by a decrease in the FDR *q*-values. Although LOESS correction is effective in reducing score-length correlation, it cannot improve the intrinsic ability of CRM finders in identifying true biological signals.

that could be the case of a short sequence containing a likely CRM. We control for potential low-score inflation in our Lever analyses, by first length-matching the foreground and background genes, and then transforming the scores into ranks to robustly assess enrichment of the foreground gene scores with respect to those in the background set, in the face of outliers.

As originally applied in the context of protein similarity scores (Durbin, 1998; Pearson, 1995), LLR did not encounter zero-scoring elements and thus did not lead to artifactual results analogous to those resulting from GS-based motif analysis of non-coding sequences. However, because of the structure of our data (i.e. the existence of zero-scoring genes and their biological interpretation as non-targets for a given code), PCA-based normalization methods (Cox and Hinkley, 1974; Nam, 2010; Pearson, 1901) can produce similar artifacts as LLR, making both of these methods inappropriate in the context of GS analyses (Supplementary Fig. S10).

Our use of gene CRM scores corrected by LOESS-fit normalization resulted in successful recovery in Lever analyses of both known and putative cis-regulatory codes for genes involved in embryonic somatic mesoderm development and in neural development in *D. melanogaster*. We have also used Lever with LOESS-fit correction to identify novel cis-regulatory codes involving other TFBS motifs for genes expressed in PCs and CCs, which were experimentally validated by Zhu *et al.* (2012); without LOESS-fit correction, no statistically significant codes could be obtained. Although here we focused on *Drosophila* GSs, we anticipate LOESS-fit correction of gene CRM scores to be useful for cis-regulatory code analyses in other organisms.

The LOESS-fit length normalization method is general and may be useful beyond cis-regulatory code analysis, for other DNA, RNA or protein sequence analyses where a score might depend on another variable such as the length of the sequence. We anticipate that such improvements in controlling for length dependence in scoring of biological sequence data will lead to improved discovery of important biological sequence patterns in gene regulation and function.

ACKNOWLEDGEMENTS

We thank Anthony Philippakis, Savina Jaeger, Steve Gisselbrecht, Jabier Gallego-Llamas and Trevor Siggers for helpful discussions, Ivan Adzhubey for technical assistance, and Steve Gisselbrecht and Luis Barrera for critical reading of the manuscript. We thank Alan Michelson, Shaad Ahmad, Xianmin Zhu and Brian Busser for sharing pre-publication data on *D. melanogaster* PC and CC gene sets.

Funding: This work was supported by the National Institutes of Health [NIH/NHGRI grant no. R01 HG005287 to M.L.B.]. A.A. was supported in part by an American Heart Association Predoctoral Fellowship.

Conflict of Interest: none declared.

REFERENCES

Adams,M.D. *et al.* (2000) The genome sequence of Drosophila melanogaster. *Science*, **287**, 2185–2195.

Ahmad,S.M. *et al.* (in press) Two Forkhead transcription factors regulate the division of cardiac progenitor cells by a Polo-dependent pathway. *Developmental Cell*.

Alkema,W.B. *et al.* (2004) MSCAN: identification of functional clusters of transcription factor binding sites. *Nucleic Acids Res.*, **32**, W195–W198.

Casal,J. and Leptin,M. (1996) Identification of novel genes in Drosophila reveals the complex regulation of early gene activity in the mesoderm. *Proc. Natl Acad. Sci. USA*, **93**, 10327–10332.

Castro,B. *et al.* (2005) Lateral inhibition in proneural clusters: cis-regulatory logic and default repression by Suppressor of Hairless. *Development*, **132**, 3333–3344.

Celniker,S.E. *et al.* (2002) Finishing a whole-genome shotgun: release 3 of the Drosophila melanogaster euchromatic genome sequence. *Genome Biol.*, **3**, RESEARCH0079.

Clark,A.G. *et al.* (2007) Evolution of genes and genomes on the Drosophila phylogeny. *Nature*, **450**, 203–218.

Cleveland,W.S. (1979) Robust locally weighted regression and smoothing scatterplots. *J. Am. Stat. Assoc.*, **74**, 829–836.

Cleveland,W.S. and Devlin,S.J. (1988) Locally weighted regression: an approach to regression-analysis by local fitting. *J. Am. Stat. Assoc.*, **83**, 596–610.

Cloonan,N. *et al.* (2008) Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat. Methods*, **5**, 613–619.

Cox,D.R. and Hinkley,D.V. (1974) *Theoretical Statistics*. Chapman and Hall, London.

Durbin,R. (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge.

Estrada,B. *et al.* (2006) An integrated strategy for analyzing the unique developmental programs of different myoblast subtypes. *PLoS Genet.*, **2**, 160–171.

Frith,M.C. *et al.* (2003) Cluster-Buster: finding dense clusters of motifs in DNA sequences. *Nucleic Acids Res.*, **31**, 3666–3668.

Gao,L. *et al.* (2011) Length bias correction for RNA-seq data in gene set analyses. *Bioinformatics*, **27**, 662–669.

Gijbels,I. and Prosdocimi,I. (2010) Loess. *Wiley Interdicipl. Rev. Comput. Stat.*, **2**, 590–599.

Halfon,M.S. *et al.* (2000) Ras pathway specificity is determined by the integration of multiple signal-activated and tissue-restricted transcription factors. *Cell*, **103**, 63–74.

Halfon,M.S. *et al.* (2002) Computation-based discovery of related transcriptional regulatory modules and motifs using an experimentally validated combinatorial model *Genome Res.*, **12**, 1019–1028.

Hallikas,O.K. *et al.* (2006) Identification of antibodies against HAI-1 and integrin alpha6beta4 as immunohistochemical markers of human villous cytotrophoblast. *J. Histochem. Cytochem.*, **54**, 745–752.

Ho Sui,S.J. *et al.* (2005) oPOSSUM: identification of over-represented transcription factor binding sites in co-expressed genes. *Nucleic Acids Res.*, **33**, 3154–3164.

Johansson,O. *et al.* (2003) Identification of functional clusters of transcription factor binding motifs in genome sequences: the MSCAN algorithm. *Bioinformatics*, **19** (Suppl. 1), i169–i176.

Kielbasa,S.M. *et al.* (2010) TransFind–predicting transcriptional regulators for gene sets. *Nucleic Acids Res.*, **38**, W275–W280.

Klepper,K. *et al.* (2008) Assessment of composite motif discovery methods. *BMC Bioinformatics*, **9**, 123.

Kreiman,G. (2004) Identification of sparsely distributed clusters of cis-regulatory elements in sets of co-expressed genes. *Nucleic Acids Res.*, **32**, 2889–2900.

Lee,S. *et al.* (2011) Accurate quantification of transcriptome from RNA-Seq data by effective length normalization. *Nucleic Acids Res.*, **39**, e9.

Mortazavi,A. *et al.* (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**, 621–628.

Moses,A.M. *et al.* (2004) MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. *Genome Biol.*, **5**, R98.

Nam,D. (2010) De-correlating expression in gene-set analysis. *Bioinformatics*, **26**, i511–i516.

Nelson,C.E. *et al.* (2004) The regulatory content of intergenic DNA shapes genome architecture. *Genome Biol.*, **5**, R25.

Oshlack,A. and Wakefield,M.J. (2009) Transcript length bias in RNA-seq data confounds systems biology. *Biol. Direct*, **4**, 14.

Pearson,K. (1901) On lines and planes of closest fit to systems of points in space. *Philos. Mag.*, **2**, 559–572.

Pearson,W.R. (1995) Comparison of methods for searching protein sequence databases. *Protein Sci.*, **4**, 1145–1160.

Philippakis,A.A. *et al.* (2006) Expression-guided in silico evaluation of candidate cis regulatory codes for Drosophila muscle founder cells. *PLoS Comput. Biol.*, **2**, 439–453.

Philippakis,A.A. *et al.* (2005) Modulefinder: a tool for computational discovery of cis regulatory modules. *Pac. Symp. Biocomput.*, 519–530.

Reeves,N. and Posakony,J.W. (2005) Genetic programs activated by proneural proteins in the developing Drosophila PNS. *Dev. Cell*, **8**, 413–425.

Richards,S. *et al.* (2005) Comparative genome sequencing of Drosophila pseudoobscura: chromosomal, gene, and cis-element evolution. *Genome Res.*, **15**, 1–18.

Salomon,D. (2007) *Data Compression: the Complete Reference*. Springer, London.

Schafer,J. and Strimmer,K. (2005) A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat. Appl. Genet. Mol. Biol.*, **4**, Article32.

Seok,J. *et al.* (2002) A novel audio watermarking algorithm for copyright protection of digital audio. *ETRI J.*, **24**, 181–189.

Shpaer,E.G. *et al.* (1996) Sensitivity and selectivity in protein similarity searches: a comparison of Smith-Waterman in hardware to BLAST and FASTA. *Genomics*, **38**, 179–191.

Sinha,S. *et al.* (2006) Stubb: a program for discovery and analysis of cis-regulatory modules. *Nucleic Acids Res.*, **34**, W555–W559.

Sinha,S. *et al.* (2003) A probabilistic method to detect regulatory modules. *Bioinformatics*, **19** (Suppl. 1), i292–i301.

Storey,J.D. (2002) A direct approach to false discovery rates. *J. Roy. Stat. Soc. B*, **64**, 479–498.

Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.

Su,J. *et al.* (2010) Assessing computational methods of cis-regulatory module prediction. *PLoS Comput. Biol.*, **6**, e1001020.

Subramanian,A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.

Warner,J.B. *et al.* (2008) Systematic identification of mammalian regulatory motifs' target genes and functions. *Nat. Methods*, **5**, 347–353.

Zhou,Q. and Wong,W.H. (2004) CisModule: de novo discovery of cis-regulatory modules by hierarchical mixture modeling. *Proc. Natl Acad. Sci. USA*, **101**, 12114–12119.

Zhu,X. *et al.* (2012) Differential regulation of mesodermal gene expression by Drosophila cell type-specific Forkhead transcription factors. *Development*, **139**, 1457–1466.