Review

Can We Predict Gene Expression by Understanding Proximal Promoter Architecture?

Łukasz Huminiecki^{1,*} and Jarosław Horbańczuk¹

We review computational predictions of expression from the promoter architecture – the set of transcription factors that can bind the proximal promoter. We focus on spatial expression patterns in animals with complex body plans and many distinct tissue types. This field is ripe for change as functional genomics datasets accumulate for both expression and protein–DNA interactions. While there has been some success in predicting the breadth of expression (i.e., the fraction of tissue types a gene is expressed in), predicting tissue specificity remains challenging. We discuss how progress can be achieved through either machine learning or complementary combinatorial data mining. The likely impact of single-cell expression data is considered. Finally, we discuss the design of artificial promoters as a practical application.

What Is Known About Proximal Promoters?

The study of gene expression and its regulation is a rapidly expanding field in which there is a growing interest in computer modeling. Progress in understanding the basic biology of gene regulation has been accompanied by technological progress in massively parallel measurements of gene expression and protein–DNA interactions. Here we focus on two very specific aspects of this large field: (i) defining and representing the **architectures of proximal promoters** (see [Glossary](#)); and (ii) using these representations to predict gene expression in animals with complex body plans and many cell and tissue types.

We strive to avoid overlaps with related reviews focusing on the regulation of gene expression in the context of yeasts or bacteria [1], sea urchins [2], the fruit fly [3], quantitative modeling [4], nucleosomes [5], enhancers [6], the divergence of *cis*-regulatory elements [7], computational methods for the identification of eukaryotic regulatory elements [8], or the 3D organization of the genome [9]. Instead, we are specifically interested in computational modeling of gene expression utilizing heuristic strategies that hold most promise for predicting the **breadth of expression** (i.e., the fraction of tissue/cell types a gene is expressed in) and the identity of expressing cell types and tissues. We mainly discuss two strategies, machine learning and combinatorics, but we also note the possibility of a hybrid strategy ([Figure 1](#), Key Figure). These strategies are closer in character to integrative data mining than to fully quantitative kinetic models, which are probably realistic in mammals only for studies that narrowly focus on individual loci.

What is the extent of the proximal promoter? In the literature the sizes vary ([Table 1](#)). Hurst *et al.* recently tried to locate the boundaries by integrating the fifth edition of the Functional Annotation of the Mammalian Genome Consortium (FANTOM5) [10] with the Encyclopedia of DNA

Trends

Integrative data mining of functional genomics datasets is an increasingly attractive research strategy that does not require investment in reagents or experimental facilities, but it does require qualified bioinformatics staff with expertise in applied statistics.

A wave of experimental data on promoter architectures is redefining how we model gene expression. Thus, we need new mathematical formalism to represent promoter architectures, implement computations on them, and facilitate combinatorial data mining.

Practical applications of the data mining of promoter architectures could include automatic genome annotation or the design of artificial promoters.

We can predict breadth of expression, but predicting tissue specificity remains a challenge. Moreover, the concepts of breadth of expression and tissue specificity will need to be redefined in view of single-cell expression data.

¹Institute of Genetics and Animal Breeding, Polish Academy of Sciences, ul. Postępu 36A, Jastrzębiec, 05-552 Magdalenka, Poland

*Correspondence: l.huminiecki@ighz.pl (Ł. Huminiecki).

Table 1. Estimates of the Boundaries of the Proximal Promoter

Species	Estimate	Motivation	Source	Refs
Mammals	First $\approx$ 200 bp upstream of the cap site ^a	None given	Classic textbook	[60]
	$\pm$ 500 bp from the TSS	Internal standard developed by the FANTOM5 consortium	Boundaries chosen for the annotation of CAGE peaks by the FANTOM5 consortium	[10]
	-2500 bp to $+500$ bp from the TSS	Authors saw these boundaries as arbitrary and noted that, in most cases, the signal relevant for the modeling of gene expression was detected within 500 bp of the TSS	Model aimed at integrated prediction of expression and enhancers based on Affymetrix microarray data from the Gene Expression Atlas	[39]
<i>Caenorhabditis elegans</i>	$\pm$ 2500 bp from the TSS	Cumulative plot of the spatial distribution of highly occupied target regions	Integrative analysis of modENCODE ChIP-seq data from <i>C. elegans</i>	[61]
	2000 bp upstream of the TSS	Not found	Probabilistic model of clustered expression patterns from promoter sequence	[32]
Yeasts	800 bp upstream of the TSS	Not found	Probabilistic model of clustered expression patterns from promoter sequence	[32]
	A few hundred base pairs from the cap site	None given	Classic textbook	[60]
Eukaryota	From -499 bp to $+100$ bp from the TSS	Focus on the regions flanking the putative TSS	Analysis focusing on nucleotide composition of promoters	[62]
<i>Escherichia coli</i>	-250 bp to $+100$ bp from the TSS	75% of all reported TF interactions in RegulonDB 8.5 occur within 100 bp of the TSS	Analyses of different releases of RegulonDB	[63–65]

^aThe cap site was defined by the location where transcription initiates (encoding the first nucleotide of the first exon of an mRNA).

Elements (ENCODE) [11]. (ENCODE was an international research consortium focusing on functional elements in the human genome.) The maximal effective size of the proximal promoter was estimated to be 6 kb [transcription start site (TSS) $\pm$ 3000 bp]. This size was approximated by estimating the distance from the TSS at which the rate of the increase in the number of mapped transcription factor (TF)-binding sites transformed to the background linear rate [12]. Interestingly, approximately the same estimate (TSS $\pm$ 4000 bp) was made by the modENCODE consortium in their **predictive model of expression** [13]. (ModENCODE was a project aiming to identify all functional elements in the genomes of *Caenorhabditis elegans* [13] and *Drosophila melanogaster* [14,15].) However, we note that these are genome-scale trends and that the sizes indicated are maximal rather than average.

Do proximal promoters determine expression patterns in animals? Currently there is mixed evidence (Box 1); whether we can or cannot make predictive models of gene expression from experimentally defined promoter architectures will weigh heavily on one or the other side of this debate.

In this context it is also important to mention enhancers (reviewed in [6]), which are *cis* acting and associate with proximal promoters but are more distally located. Most enhancers were

Glossary

Breadth of expression: the fraction of tissue or cell types in which a gene is expressed. A housekeeping gene is expressed in all tissue/cell types while a tissue-specific gene is expressed in few.

ChIP technologies: used to investigate interactions between proteins and DNA. In ChIP-on-chip crosslinked DNA–protein complexes are immunoprecipitated and the sequence of associated DNA is determined by hybridization to genomic microarrays. In ChIP-seq the identity of associated DNA is determined by sequencing.

Cis-plus-trans model: a model that includes as inputs the architecture of the promoter (*cis*-regulatory information) as well as the expression levels of TFs in a target tissue (*trans*-regulatory information). The output of the model is the expression output of the modeled promoter in a target tissue.

Fuzzy sets: extensions of classical sets that allow partial degrees of membership (between 0 and 1). Such partial membership can represent the measure of confidence one has that an element, such as a TF-binding site, belongs in the promoter architecture.

Housekeeping genes: genes that are transcribed in every cell and tissue type. These genes are involved in functions that are necessary for each cell, such as basic metabolism, transcription, and translation, or are the building blocks of fundamental cellular structures such as the cytoskeleton or organelles.

Multisets: extensions of classical sets that allow multiple occurrences of their elements.

Paralogs: homologous genes that are related to each other through a gene-duplication event.

Predictive models of gene expression: making a distinction between housekeeping and tissue-specific genes (i.e., predicting the breadth of expression of a gene) is one important aspect of the modeling of gene expression in mammals. Another important aspect is predicting precisely in which tissues a gene is expressed. Current approaches are quite good at predicting the breadth of expression of a gene in humans. However, limited success has been achieved at

Box 1. Mixed Experimental Evidence on the Key Role of the Proximal Promoter in Controlling Gene Expression Patterns

Which aspects of gene expression are determined by proximal promoters and to what extent? Published evidence is mixed. In the study of a single TF called STE12 [66], changes to the sequences of the binding motifs for STE12 were found to account for only approximately half of the expression differences observed between species [66]. The other half of the differences could not be easily attributed (but were most likely due to the divergence of sequences adjacent to the binding motif, or due to chromatin-level effects). In addition to data from yeasts, the authors also looked at human, mouse, and chimp expression compendia. When data from these species were analyzed, no correlation between the conservation of expression patterns and the conservation of *cis*-regulatory motifs could be observed.

Another interesting study focused on the mitogen-activated protein kinase pathway [67]. This study suggested that stimulation by growth factors, which led to the induction of immediate-early response genes, was accompanied by the upregulation of genes in the neighborhood of targets. The authors suggestively called this effect co-upregulation or the ripple effect. However, if the ripple effect were a rule rather than an exception throughout the genome, one would expect a very strong clustering of tissue-specific and environment-induced genes in the mammalian genome that extends beyond the clusters of tandem-duplicated genes.

On a more positive note, a simple metric of promoter architecture was found to predict the breadth of expression in human tissues [12]. In the same study, a correlation between the divergence of promoter architectures and the divergence of their expression patterns was found.

There have also been numerous focused studies of the architectures of individual mammalian promoters. For example, Okada *et al.* proved the importance of the 3-kb proximal promoter of an endothelium-specific **paralog** of roundabouts, ROBO4, for the control of its expression pattern [68]. Okada *et al.* detected several TF-binding sites, including SP1 and a site for GA-binding protein (GABP) [68]. Subsequently, Okada *et al.* showed using transgenic mice that the GABP site contributes to ROBO4 expression in the endothelium [69]. In another example a proximal promoter of a photoreceptor-specific tetraspanin called retinal degeneration slow (RDS) was examined and its 3.5-kb fragment was found to contain many TF-binding sites regulating cell-type-specific expression in photoreceptors (with maximal activity in just 350 bp proximal to the TSS) [70]. In yet further examples, the hepatocyte nuclear factor 4 (HNF4)-binding site in the proximal promoter was found to regulate the liver-specific expression of ABCC6 [71] while spatially and temporarily restricted TFs were found to regulate the proximal promoter of the Brambell receptor heavy chain, Fcgrt [72]. However, such locus-focused studies, while very numerous, are unlikely to look at a broad range of samples when verifying the tissue specificity of expression of different promoter constructs. (Thus, the measurement is of the level of expression, or at best of preferential expression, rather than strictly of tissue specificity.)

predicting exactly in which tissue a gene is 'on' or 'off'.

Promoter architecture: the set or multiset of all TFs that can potentially bind the proximal promoter.

Promoter architecture defined in this way is a global and static feature of promoter DNA in a given locus and species (it is constant across different cell/tissue types and conditions). Note that alternative definitions are possible: local (e.g., incorporating cell-specific epigenetic modifications) or dynamic (incorporating transient phenomena such as the temporal dynamics of post-translational modifications of DNA-bound TFs). Alternative mathematical representations may also allow for the order or locations of TF binding sites in relation to the transcription start site.

Proximal promoter: the DNA sequence, and the *cis*-regulatory elements it contains, directly adjacent to the TSS of a gene.

Tissue-specific genes: genes that are transcribed at a meaningful level only in a specific tissue. Examples include genes that enable a hepatocyte to function as the body's biochemical factory or the neuron to work in synaptic transmission or that make cellular progenitors of these cell types commit and differentiate into them.

found by high-throughput connectivity maps to interact with promoters beyond the nearest location on the linear map of the chromosome [16]. Enhancers can be identified by characteristic chromatin modifications (for example, the H3K4me1 histone mark reviewed in [17]) and bidirectional transcription into eRNAs (reviewed in [18]). However, in this review we focus on proximal promoters rather than enhancers.

What is meant here by the architecture of the proximal promoter? It is important to be precise because the phrase 'promoter architecture' has been frequently used in relation to proximal promoters but its use tends to be informal, sometimes referring to the type and arrangement of TF-binding sites, sometimes to the biased stretches of nucleotide sequences in core promoters, and sometimes to their epigenetic modifications. We must stress that we use this phrase in the first sense: as the type of functional TF-binding site within the proximal promoter. Specifically, we mean the set of all TFs that can bind the proximal promoter in any tissue or cell type. ('Set' is a term from the world of Cantor's classical set theory usually known from high-school mathematics; we consider alternatives in Box 2.) This set of potentially binding TFs might be derived from *in vivo* experimental data, such as ENCODE datasets merged across different cell lines and TFs. Thus, the architecture of a promoter defines the promoter's theoretical binding potential rather than specifying the TFs that are bound in a particular tissue or cell type, which it is usually not possible to know. (Indeed, the set of TFs bound in any cell type will be a subset of the architecture.) We propose that promoter architecture defined in this way should be regarded as a global and static feature of promoter DNA, a feature that is constant for a given locus and species.

Key Figure

A Conceptual Diagram Illustrating Two Strategies for Understanding Experimentally Defined Architectures of Proximal Promoters

Trends in Biotechnology

Figure 1. For a Figure360 author presentation of Figure 1, see the figure online at <http://dx.doi.org/10.1016/j.tibtech.2017.03.007#mmc1>.

The first strategy is to use machine learning. Inputs, machine-learning algorithms, and outputs should be interchangeable modules of the analytical pipeline. Model learning and the scoring of top-performing combinations of modules should be automatic. Combinatorial data mining is the second strategy. The aim is to identify the combinations of transcription factors (TFs) that support an expression characteristic of interest. This includes a computationally intensive step of calculating all TF combinations (of however many TFs are practical) and their frequencies. A hybrid strategy is also possible. Combinatorial data mining could be used to extract the most informative features of promoter architectures for subsequent machine-learning strategies. First, TF combinations that occur frequently in observed promoter architectures (versus randomized promoters) would be identified. Then, promoter architectures would be re-coded to be expressed as counts of these preselected combinations. This approach will reduce the dimensionality of the machine-learning problem.

A pause for thought and some careful consideration of wording is necessary. The word 'global', which we introduce in this context for the first time, signifies the fact that such a promoter architecture is constant across different cell and tissue types. The word 'static' signifies that it does not depend on environmental conditions or TF concentrations. This global and static

definition of promoter architecture has proved to be a useful research tool for predicting the breadth of expression in mammals [12,19].

We note that alternative definitions of promoter architecture are entirely possible and might be better suited for some applications. For example, a dynamic definition of promoter architecture could incorporate transient protein–DNA interactions or epigenetic modifications or cell-specific DNA–protein interactions. A local promoter architecture would be specific to a chosen cell type rather than generalized across the sample space. Additional comments are provided in Table 2, which compares the features of various definitions of promoter architecture accepted in several papers reviewed here.

Our goals are to: (i) ask to what extent it is possible using *in vivo* experimentally determined promoter architectures to predict animal expression patterns; (ii) argue for the usefulness of the proposed definition of promoter architecture (given a wave of ChIP data and especially for modeling gene expression in multicellular animals with complex body plans that have many tissue and cell types); (iii) propose a corresponding mathematical formalism and software implementations for computations on promoter architectures; (iv) review how published models of gene expression in animal model species and in mammals relate to the approach advocated here; and finally (v) consider how technological advances such as single-cell expression data are likely to impact the field. We stress that it is not our goal to review the literature on expression modeling in yeasts or bacteria (where data are easier to collect and consequently models tend to be more mathematical and more quantitative). We focus on animal model species and on mammals, where models of gene expression are less advanced, resembling the heuristics of data mining, but still have vast practical implications. For example, it would be extremely valuable if we could predict whether a proximal promoter leads to broad expression across tissue/cell types or whether it is likely to promote narrow expression restricted to a specific set of conditions. It would be even more valuable if we could predict in exactly which tissues or cell types, and under exactly what conditions, normal or pathological, the gene is expressed. The fields of application range from basic science through biomedical research to the pharmaceutical industry.

A Wave of New High-Throughput Experimental Data Defining Promoter Architectures

Our ability to determine the architectures of proximal promoters is rapidly being improved by high-throughput screens utilizing **ChIP technologies** ([20], see Box 1). For example, in the

Box 2. How Should the Architecture of the Promoter Be Represented for Computations?

Promoter architecture may be represented using the formalism of set theory [73] as a set of potential TFs that can bind a given proximal promoter. However, the dynamics of multiple occurrences of the binding sites of a TF are likely to differ from those of individual isolated sites [74]. Therefore, a more appropriate representation of the architecture of the promoter might be a **multiset**. (A multiset denotes a collection whose elements may occur multiple times.) These ideas for representing promoter architectures are graphically summarized in Figure 1.

Another consideration is the need to model ChIP peak intensities, which might be achieved using **fuzzy sets**. Fuzzy sets have already been used in the CisMiner itemset mining algorithm [75] and in an ANFIS-based fuzzy system for the detection of ChIP peaks [76]. The rationale for using fuzzy sets [77] for promoter architectures has been already been explained by Navarro *et al.* [75], and more generally in bioinformatics [78] and biomedicine [79]. Fuzzy technology is underused given that it outperforms classical technology when dealing with noisy datasets.

When the mathematical representation of the promoter architecture is chosen, it needs to be implemented in a programming language of choice. This point is technical but not trivial: performance is a concern when mining promoter architectures on the genome scale. Fortunately, popular compiled and interpreted languages have facilities for operations on sets that can be extended to implement multisets or fuzzy sets. For example, R has the package ‘sets’. Python has implementations of sets, lists, and dictionaries included in the core language definition. Scala has sets and maps in the package ‘collections’. An even faster alternative is to use C++ Standard Template Library containers: sets and multisets.

Figure 1. Representing the Architecture of the Proximal Promoter. Here we compare alternative mathematical representations of a single chosen proximal promoter architecture. The diagram in (A) (not to scale) shows a hypothetical proximal promoter of a three-exon gene with the following transcription factor (TF)-binding sites: TF1, TF2, TF3, TF4, TF4, TF5, TF6, TF6, and TF1. These TF-binding sites are listed according to the order of their 3' boundaries. In the figure we also illustrate the location of the transcription start site (TSS), which is located within the proximal promoter. Additionally, we show a hypothetical CTCF-binding site and an enhancer. The CTCF site may act as an insulator, which is illustrated with a STOP traffic sign. In (B) we show that our chosen architecture of the proximal promoter can be expressed as a set, $S = \{TF1, TF2, TF3, TF4, TF5, TF6\}$, or a multiset, $M = \{2/TF1, TF2, TF3, 2/TF4, TF5, 2/TF6\}$. In (C) we consider not one but three promoters with exactly the same architecture of the proximal promoter. We also consider their unions and multiset sums. In (D) we introduce fuzzy representations of our chosen promoter architecture. These representations comprise a fuzzy set, S_f , and a fuzzy multiset, M_f [77,92,93]. Here, Encyclopedia of DNA Elements (ENCODE) quality scores (which vary from 0 through 1000) are transformed to signify the degree of belonging in a fuzzy set representing the architecture of a promoter (which varies from 0 through 1). As mentioned in the main text, determining the functionality of low-quality-score but common ChIP peaks is a major challenge of the post-ENCODE era. For this reason promoter representations that can model the height of ChIP peaks will be at a premium. In (E) we still use fuzzy logic but now we consider three promoters instead of one.

Table 2. Definitions of Promoter Architectures

Type	Interpretation	Data	Application	Refs
Global ^a and dynamic	Promoter architecture depends on the concentrations of TFs and the strength of protein–DNA binding that can be modeled thermodynamically	Relative TF concentrations vary along the anteroposterior embryonic axis	Thermodynamic modeling: both the occupancy of TF-binding sites and target gene output are modeled	[29]
Global and static ^b	Although ChIP data came from a time course, peaks were merged so that the resulting promoter architecture does not vary	Expression and ChIP data are partially matched ^c in terms of tissue type	Heuristic data mining: classification of regulatory elements into one of the coexpression clusters	[31]
Global and static	Promoter architecture is constant across different cell and tissue types and under all conditions	Unmatched ENCODE ChIP data that was merged across different cell lines	Heuristic data mining: prediction of the breadth of expression	[12]
		<i>De novo</i> ^d computational predictions of TF-binding sites	Heuristic data mining: classification of a gene into one of the clusters of coexpressed genes	[32]
			Heuristic data mining: prediction of top and bottom expressed genes in a target tissue	[39]

^aGlobal signifies a promoter architecture that is cell-type independent. The opposite is local, which signifies a definition of promoter architecture that is cell-type specific; for example, because it was derived from ChIP-seq data from a single cell line.

^bStatic signifies the definition of promoter architecture that does not incorporate temporal effects that would depend on environmental conditions, TF concentrations, or post-translational modifications of TFs. (The opposite is dynamic.)

^cMatched expression and ChIP-seq datasets come from the same cell or tissue type.

^d*De novo* predictions of TF-binding sites are purely computational predictions calculated from the DNA sequence using specialized algorithms and databases of protein–DNA binding motifs.

past decade genome-wide ChIP screens of TF-binding sites became available for yeast [21], fly [15], worm [13], mouse, and human [11] genomes. These experimental screens augment the computational predictions of TF-binding sites available previously. Although ChIP-seq assays are run separately for individual TFs in separate cell lines, the same guidelines and protocols are followed [22]. Therefore, such datasets can be merged to infer *in vivo* experimentally defined global and static promoter architectures. (We note that some interpretations of ENCODE were criticized [23], but this criticism, although strongly worded, was focused on the claim that more than 80% of the human genome is functional rather than on the ENCODE datasets themselves.)

In addition to such *in vivo* experimental data on promoter architectures, there is also a large amount of new *in vitro* data on TF-binding profiles. Such profiles are derived from protein-binding microarrays (PBMs) or high-throughput SELEX [24]. Binding specificities can be deduced from the results of *in vitro* experiments where various double-stranded DNA oligonucleotides of 10–14 bp in length are tested for binding to the TF. While these *in vitro* experiments may excel at biophysically defining the affinity landscape of a TF, the inferred binding motifs are often not functionally occupied by a TF *in vivo*. Motif availability must be seen as the crucial issue for mammalian TFs. Motif availability can be modified through several mechanisms; for example, through: (i) the modification of a TF's affinity profile by protein cofactors or post-translational modifications; (ii) the impact of DNA shape on a TF's ability to bind its target motif [25,26]; (iii) competitive or cooperative effects of adjacent TF-binding sites [27]; or (iv) the impact of nucleosome formation on the availability of the binding motif. Therefore,

in vivo experimental data is hard to replace with *in vitro* data for the purposes of defining promoter architectures and the modeling of gene expression.

Still, the challenges of utilizing *in vivo* experimental data for modeling should not be underestimated. Levo and Segal [20] rightly note the conceptual difficulty in computing highly quantitative data on gene expression from semiquantitative data derived from ChIP-on-chip and ChIP-seq screens. ChIP technologies report many low-affinity/low-occupancy sites [22] whose functional significance is uncertain. Determining the significance of these low-intensity peaks is a priority for the post-ENCODE era, putting a premium on representations of promoter architecture that can model peak intensity instead of just binary presence or absence. Another obstacle is that data from different screens are difficult to integrate and meta-analyze because of the differing technological platforms used for sequencing and the differing reactivity profiles of antibodies. However, protein–DNA binding screens are likely to mature as technology and stricter standards are developed [22].

Nevertheless, for the foreseeable future mammalian expression data will be much more abundant than mammalian ChIP-on-chip or ChIP-seq data. This asymmetry poses a challenge for modeling and prediction: we will wish to predict gene expression in a broader range of cell types, tissue types, and physiological states than we have ChIP data for. For such practical reasons, the most universal strategies would work with promoter architectures that are akin to a menu of TFs from which given promoter can potentially ‘choose’ rather than a snapshot of TFs bound in a particular cell type (which would always have to be determined experimentally).

Lessons from Non-mammalian Animals: The Fruit Fly and the Worm

We start by mentioning two studies in the fruit fly, but note these papers have been already reviewed by Rister and Desplan [28]. Segal *et al.* [29] integrated expression levels of TFs (*trans*-regulatory elements) together with DNA-binding motifs (*cis*-regulatory elements) to predict gene expression during embryonic segmentation in the fruit fly. Protein–DNA interactions were modeled using thermodynamics, but such modeling required detailed knowledge of the biochemical properties of the modeled TFs. However, this thermodynamically aware ***cis-plus-trans* model** could predict expression with remarkable precision. Is this model applicable to mammals? Not directly: *Drosophila* embryogenesis (reviewed in [30]) is a rather specialized developmental model. Moreover, Segal’s modeling strategy demanded knowledge of the biochemical profiles of TFs whose concentrations had to be precisely quantified.

The second paper used ChIP-on-chip data for a few key TFs from a time series of mesoderm development [31]. Clusters of TF-binding sites were identified as *cis*-regulatory modules and used to train a machine-learning algorithm that was able to correctly predict *in vivo* expression activity for most modules. Importantly, the inputs for machine learning were not just binary (TF presence/absence) but directly reflected ChIP peak heights.

In *C. elegans* a probabilistic model developed originally for yeasts was successfully applied to microarray expression profiles from worm development [32]. However, the authors noted that the task was more difficult in *C. elegans* than in yeasts: the worm is multicellular, TF predictions are less established, genes are alternatively spliced, and regulatory regions are also present downstream of the TSS. Needless to say, these challenges are even more marked in mammals, where gene regulation is even more multilayered and there are hundreds of different cell and tissue types.

To what extent can models, such as the fly and worm models described above, be applied to mammals? As these models require copious amounts of experimental information, this would

probably be possible only under narrow and focused conditions such as gastrulation or the formation of individual organs. This would require considerable investment and the creation of consortia working on expression profiling, ChIP data, and TF proteomics for carefully selected biological models that are of vital interest to the scientific community. The investment might be worth the price because these models can tackle the intricacies of mammalian transcription; for example, the actions of regulatory noncoding RNAs. Perhaps, for well-established mammalian biological models, data-intensive quantitative approaches similar to those developed in the fruit fly should be prioritized. For more general applications, heuristic data-mining strategies that strive to make maximum use of the datasets that are already available might be prioritized.

Lessons from Mammals

The size of promoter architecture (or the cardinal number of the corresponding set) was found to be a good predictor of the breadth of expression in humans [12]. This insight was derived after the FANTOM5 database of expression profiles was integrated with ENCODE data on the binding profiles of 148 TFs in human cell lines. Therefore, this work might be of interest not only for its biological insights but also because it was a successful exercise in data integration. Gene expression (from FANTOM5) and *in vivo* promoter architectures (from ENCODE) were integrated to obtain insights that could not result from either dataset alone. After several control analyses, the merged ENCODE ChIP-seq datasets were found to be useful for defining global and static promoter architectures. Hurst *et al.* also considered and rejected [12] an alternative 'sticky' model, which was a trivial explanation of the correlation assuming random binding of TFs to transcriptionally active proximal promoters.

An exciting continuation of the work of Hurst *et al.* would be to quantify the contributions to **housekeeping** expression made by different TFs and TF combinations (Figure 2). Is there such a thing as a fixed housekeeping code of TFs – that is, a fixed set of TFs that are: (i) dominant; (ii) facultative; and (iii) excluded for the majority of housekeeping transcripts? Or are the data on the distribution of TFs overwhelmingly heterogeneous – that is, there are no dominant TFs, most TFs are facultative, and low-frequency TFs simply correspond to TFs with less numerous binding sites in the genome (without evidence for exclusion from housekeeping promoters)?

At the same time, we note that the strategy of Hurst *et al.* was a heuristic that tried to make the best use of the data already available and did not attempt to model many intricacies, including the cell specificity of ChIP-seq peaks [33], antisense transcription [34,35], temporal dynamics during development [36], long-range promoter–enhancer associations [16], and the existence of chromosomal domains and nuclear subcompartments [37,38]. To model these intricacies would be much more challenging and most likely not possible without new data being generated.

Some of these challenges inherent in predicting tissue-specific expression in mammals were previously underlined by Taher *et al.* [39]. These authors used computational predictions of TF-binding sites along with microarray expression data. A machine-learning algorithm was used to successfully discriminate between the promoters of the highest- and lowest-expressed genes in each of the tissues. However, the identified genes were not strictly tissue specific but rather highly expressed in a given tissue. These variables are hard to disentangle because the breadth of expression correlates with the average expression. Therefore, it is difficult to be certain what expression metric Taher *et al.* discriminated on: tissue specificity, average expression, or breadth of expression. Another methodological consideration is that microarray data do not discriminate between alternative TSSs. Therefore, the true proximal promoter might have been missed for some transcripts.

(A) Hypothetical TF code for housekeeping expression**Housekeeping genes** **(B) Overwhelming data heterogeneity****Housekeeping genes**

Trends in Biotechnology

Figure 2. Is There a Fixed Code of Transcription Factors (TFs) for Housekeeping Expression? The integration of the Functional Annotation of the Mammalian Genome Consortium (FANTOM5) and the Encyclopedia of DNA Elements (ENCODE) is making it possible to quantify the contributions to housekeeping expression made by different TFs. If there is a fixed housekeeping code (A), it should be possible to identify TFs that are: (i) dominant; (ii) facultative; and (iii) excluded for the majority of housekeeping transcripts. If the data are overwhelmingly heterogeneous (B), there should be no dominant TFs, most TFs will be facultative, and low-frequency TFs will correspond to TFs with less-numerous binding sites (without statistical evidence for exclusion).

In another study a ChIP dataset from mouse embryonic stem cells [40] was successfully used to predict gene expression in this cell type [41]. The practical limitation was that both the ChIP and the expression data had to come from exactly the same cell type (i.e., the datasets had to be matched). Such a strategy, requiring matching datasets, is hard to generalize.

Is there any alternative to settling for global promoter architectures or to expensive investment in generating ChIP data matching expression data in terms of cell type? Natarajan *et al.* [42] proposed using DNase I hypersensitive sites to predict cell-type-specific and housekeeping expression. Their success underlined the key role of TF motif accessibility, for which the DNase I signal can be regarded as a proxy. The strategy still relied on cell-type-matched data, but DNase I data are much easier to generate than ChIP data for many TFs. However, DNase I signal, like other chromatin features, is likely to be an effect of transcription rather than its cause. Only causative features could be used to construct synthetic promoters, which is one of most exciting practical applications in this area.

The Computation and Combinatorics of Promoter Architectures

Performance is likely to be a concern when working computationally with *in vivo* experimentally defined promoter architectures of whole genomes for the purpose of modeling gene

Trends in Biotechnology

Figure 3. Expression-to-Combination Data Mining of Promoter Architectures. In the expression-to-combination strategy outlined in this figure, all of the genes in the genome are divided into a target set of interest and its complement. For example, the target set could be all of the housekeeping genes. The complement will then comprise all genes that are not housekeeping genes. In the next step, combinations of transcription factors (TFs) with differential frequencies between the target and the complement are identified. Promoter architectures could be also pruned of TFs that are not coexpressed with their targets. *P* values should be calculated by a randomization-based test. This chart also illustrates three typical stages of data mining, highlighted using different colors: data representation in red, data remodeling in yellow, and the statistical testing stage in green.

expression. Therefore, the choice of a mathematical representation and computational implementations thereof is an important concern (Box 2).

Once represented computationally, how should promoter architectures be analyzed? One avenue would be to employ the mathematical toolkit of combinatorics, which works well with sets. Promoter architectures that are specific to a tissue type or condition could be identified by comparing the frequencies of occurrences of combinations of binding sites between groups of genes with and without the expression feature of interest. This approach is referred to here as an expression-to-combination workflow and is illustrated in Figure 3. For example, all genes

could be divided into two groups: (i) those with expression specific to a particular tissue type (e. g., liver); and (ii) the remaining genes. The next step would be to compare TF combinations in proximal promoters between the groups. Importantly, a randomized test would be required to account for the effect of differences in the distribution of the number of binding sites across genes in each group. (**Tissue-specific genes** tend to have fewer TF-binding sites, which would artifactually lower the frequencies of larger TF combinations.) This is just one example of many possible analytical workflows that are straightforward to implement once the architecture of the promoter is well defined, well represented mathematically, and well implemented computationally.

The Impact of Single-Cell Expression Data

There has recently been an increase in expression profiles derived from individual cells together with the development of statistical and bioinformatics tools to model [43–45] and interpret such

Trends in Biotechnology

Figure 4. Single-Cell Expression Data: Moving Beyond the Comparison of Means. Profiling the expression of single cells will necessitate a move beyond the simple comparison of means (A,C) and towards the analysis of signal distributions in cell populations (B,D). Many single-cell experiments result in a characteristic bimodal distribution. In such a distribution, the signal from non-expressing cells is mixed with the signal from actively transcribing cells (which are during the bursts of transcription from the locus of interest). This is true for both *in vitro* cell cultures (A,B) and bulk tissue samples (C,D). Thus, single-cell expression data will require rethinking and redefinition of housekeeping genes and of the tissue specificity of expression.

data spatially [46], in terms of stochasticity of expression [47,48], or in terms of allele-specific expression [49] (reviewed in [50–52]). These new technologies are an advance over older platforms for expression profiling that demanded an amount of RNA corresponding to approximately 100 000 cells (what amounted to bulk tissue samples or cells cultivated in a whole-cell culture dish) and precluded insights into the expression status of individual cells. Reviewing single-cell expression technologies fully is beyond the scope of this review, but in this section we note implications for the analysis of promoter architectures, the definition of the breadth of expression, and predictive expression modeling.

For example, promoter architectures in *Escherichia coli* were found to determine the amount of transcriptional noise that was observed [53], suggesting that transcriptional noise is a tunable and evolvable feature of living organisms. It will be interesting to see whether promoter architectures also affect transcriptional noise and heterogeneity in mammalian cells. From the modeling point of view, it would be valuable if gene expression in individual cells in different phases of the cell cycle could be predicted or how much transcriptional noise is likely to be associated with a target promoter architecture in different cell types. This area is of great interest but as yet largely uncharted. For the moment studies are limited to global promoter architectures because there are currently no technologies that can determine promoter architectures to match single-cell expression data.

Box 3. Complex Interactions between High- and Low-Affinity TF-Binding Sites

Explaining the function of common but weak ChIP peaks is a major challenge for the post-ENCODE era [80,81]. Published results suggest that the dominant mode of action for TFs is activation and that TFs have cumulatively positive effects on the breadth of expression [12]. However, more complex modes of interactions between TFs have been described. For example, some of the STAT1-binding sites in human glutaminase 1 were found to have inhibitory effects while others were excitatory [82]. In yeasts multiple binding sites for SBF in the promoter of the HO gene were found to create a complex spatiotemporal cascade that served to relay the signal in addition to activating transcription [83]. Other authors focused on the promoters of photoreceptors and the CRX TF [84] and suggested that high-affinity binding sites for CRX repress transcription while low-affinity sites activate transcription [85]. Interestingly, cooperative binding sites for other TFs might override the repressive logic of the strong binding sites for CRX.

Other interpretations of low-affinity binding sites include: (i) the hypothesis that such sites, although frequent, are not functionally relevant [81]; (ii) the hypothesis that sites with differential binding affinities may facilitate threshold-level activator responses to morphogen gradients, with low-affinity sites driving higher-threshold responses or responses that are spatially more restricted [86]; and (iii) the hypothesis that in some situations evolutionary selection may favor low-affinity sites [87]. This area has been excellently reviewed by Slattery *et al.*, who also proposed a general division of TFs into pioneers, settlers, and migrants and discussed the heterogeneity of ChIP samples as well as billboard and enhanceosome models of combinatorial TF interactions [80].

Furthermore, the need for a distinction between, and testing of, flexible versus constrained models of TF cooperativity has been pointed out by several other reviews focusing on enhancers [20,88]. A review of cooperative and non-cooperative models of interactions between homotypic clusters of TFs was also published [74] suggesting that such clusters can interact to modulate the temporal dynamics of TF binding. However, other scenarios are possible. For example, the TATA box was shown to be a highly modular component of synthetic promoter architectures in yeasts [56] (note that modularity is the opposite of interactivity).

Some practical implications for modeling are that good explanatory models of gene expression should help to quantify the frequencies of the different modes of interactions between TFs described above. For example, the hypothesis that high- and low-affinity binding sites have opposing effects can be tested by separating inputs for a machine-learning algorithm.

Assumptions about the mode of action of TFs should inform the choice of representation of promoter architecture. For example, allowing repeated TFs should allow the machine-learning algorithm to learn interactions between multiple binding sites, which can have complex dynamics (reviewed in [74]). Another modification would be to treat high- and low-affinity sites as separate inputs for machine learning because they could have opposing effects [85].

Single-cell expression data will also require radical rethinking of how to define the breadth of expression and tissue specificity. Already the trend is to move beyond the comparison of means and to model data heterogeneity [43]. There is usually a population of non-expressing cells mixed with a heavy right-tailed population of expressing cells (Figure 4). This characteristic bimodal distribution may be due to the cells being between bursts of transcription (e.g., due to the phase of the cell cycle) or stochastic effects such as promoter methylation. Thus, the expression signal for single-cell data will be a distribution rather than a point estimate (compare Figure 4B,D versus 4A,C). This phenomenon will additionally complicate the classification of a tissue as either expressing or not expressing a gene. Again, this is as-yet-uncharted territory: currently available single-cell datasets are from *in vitro* cultures of a single cell type (Figure 4B) rather than from *in vivo* tissue samples, which would be mixtures of cell subpopulations (Figure 4D). Only for mixed subpopulations would it make sense to define the breadth of expression. In conclusion, single-cell expression data are sure to have great impact but there are currently few results in mammals for prediction of single-cell spatiotemporal expression.

Concluding Remarks and Future Perspectives

Predicting spatial expression patterns from the architecture of proximal promoters in mammals remains challenging (see Outstanding Questions). We remain uncertain which features of expression patterns are controlled by proximal promoters and to what extent – and which are controlled by other *cis*-regulatory sequences, such as enhancers, or by chromatin modifications such as methylation. The literature includes many individual examples but few generalized studies performed on the scale of the genome.

Short-term challenges include: (i) determining the TF code for housekeeping expression; (ii) improving predictions of tissue-specific expression; (iii) determining the functionality of weak TF-binding sites (Box 3); (iv) quantifying the interactions of homotypic repeats of TF-binding sites; and (v) developing mathematical and computational tools to quantify functional trends in ChIP-derived promoter architectures on the scale of the genome.

In the longer term, computational predictions of promoter activity should aid in the design of synthetic promoters with custom expression characteristics (Box 4). Desirable expression

Outstanding Questions

To what extent is the spatial expression pattern in mammals controlled by the architecture of the proximal promoter? Alternative narratives stress the importance of genomic context and epigenetic modifications.

Can proximal promoters only drive basal levels of expression or do they control complex patterns of tissue-specific and/or inducible expression? There are many interesting individual examples on both sides of the argument. Quantification of global trends is highly desirable.

In general, do low-affinity and high-affinity TF-binding sites have the same or differing activities?

Do homotypic clusters of TFs generally have additive or cooperative dynamics?

Is there a fixed code of TFs for housekeeping expression in mammals?

Are there combinations of TFs that determine tissue-specific expression or are the data overwhelmingly heterogeneous?

Do promoter architectures determine transcriptional noise?

Box 4. A Practical Application: A Systematic Approach to Constructing Synthetic Promoters

When exogenous DNA is introduced into animal cells, the transgene should typically be expressed in a tissue-specific manner. For example, the success of so-called suicide gene therapy, where a cytotoxic gene is introduced into tumor cells, will rely on the gene being expressed only in transformed cancerous cells. Similarly, the success of a gene therapy strategy aimed at correcting a muscular dystrophy of monogenic origin by complementing with a wild-type version of the deficient gene will rely on the transgene being expressed in muscle cells alone. This logic applies not only to gene therapy but also to transgenic farming, where a transgene will be expressed in a tissue-specific manner, such as in the secretory epithelium of the mammary gland for the transgene to be secreted with milk. Clearly, uncontrolled or 'leaking' transgene expression is likely to be a problem both for gene therapy and for transgenic farming.

It is tempting to assume that a promoter of a well-known tissue-specific gene can drive tissue-specific expression when cloned upstream of a transgene [54]. For example, gene therapy strategies targeting vascular endothelial cells frequently utilize a promoter of vascular endothelial growth factor [89]. The problem is that the activity of the proximal promoter might be modified by the genomic context, nearby enhancers, epigenetic modifications, miRNA-binding sites, the 3' untranslated region (UTR), etc.

A more systematic approach was taken by Rincon *et al.*, who identified cardiomyocyte-specific transcriptional *cis*-regulatory motifs and used them for cardiac gene therapy [90]. Their bioinformatics strategy comprised of: (i) identifying highly and lowly expressed cardiac-specific genes; and (ii) subtracting their regulatory contents to identify cardiac-specific TF-binding sites associated with high expression in cardiomyocytes. We note that this procedure is similar to the expression-to-combination analytical workflow outlined in Figure 3 in main text.

Another systematic approach to the discovery of heart-specific *cis*-regulatory regions was explored by Narlikar *et al.* [91], who identified and tested *in vivo* human heart enhancers through a *de novo* classifier that combined linear regression with Gibbs sampling (achieving a validation rate of 62% in mouse and zebrafish).

characteristics might include tissue-specific or cancer-specific expression, induction of expression at certain stages of the cell cycle, or induction by a drug or a hormone (reviewed in [54,55]). Computational tools are needed to design such synthetic promoters. The current practice of cloning *cis*-regulatory cassettes from a handpicked tissue-specific promoter ([54], see Table 2) is inadequate. Such handpicked *cis*-regulatory cassettes usually do not robustly support analogous expression patterns when moved to a new genic and genomic context (as noted in [55]). In turn, the libraries of synthetic promoters can help to unravel the combinatorial rules that govern the activity of mammalian proximal promoters. Such an approach has already proved extremely fruitful in yeasts [56–59] and *E. coli* [53].

Supplemental Information

Supplemental Information associated with this article can be found, in the online version, at <http://dx.doi.org/10.1016/j.tibtech.2017.03.007>.

References

- Teif, V.B. (2010) Predicting gene-regulation functions: lessons from temperate bacteriophages. *Biophys. J.* 98, 1247–1256
- Bolouri, H. and Davidson, E.H. (2002) Modeling DNA sequence-based *cis*-regulatory gene networks. *Dev. Biol.* 246, 2–13
- Gaul, U. (2010) Decoding transcription and microRNA-mediated translation control in *Drosophila* development. *Biol. Chem.* 391, 767–770
- Segal, E. and Widom, J. (2009) From DNA sequence to transcriptional behaviour: a quantitative approach. *Nat. Rev. Genet.* 10, 443–456
- Teif, V.B. *et al.* (2013) Taking into account nucleosomes for predicting gene expression. *Methods* 62, 26–38
- Plank, J.L. and Dean, A. (2014) Enhancer function: mechanistic and genome-wide insights come together. *Mol. Cell* 55, 5–14
- Wittkopp, P.J. and Kalay, G. (2011) *Cis*-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nat. Rev. Genet.* 13, 59–69
- Narlikar, L. and Ovcharenko, I. (2009) Identifying regulatory elements in eukaryotic genomes. *Brief. Funct. Genomics Proteomics* 8, 215–230
- Dekker, J. and Mirny, L. (2016) The 3D genome as moderator of chromosomal communication. *Cell* 164, 1110–1121
- Forrest, A.R. *et al.* (2014) A promoter-level mammalian expression atlas. *Nature* 507, 462–470
- ENCODE Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74
- Hurst, L.D. *et al.* (2014) A simple metric of promoter architecture robustly predicts expression breadth of human genes suggesting that most transcription factors are positive regulators. *Genome Biol.* 15, 413
- Gerstein, M.B. *et al.* (2010) Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project. *Science* 330, 1775–1787
- Kharchenko, P.V. *et al.* (2011) Comprehensive analysis of the chromatin landscape in *Drosophila melanogaster*. *Nature* 471, 480–485
- Roy, S. *et al.* (2010) Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. *Science* 330, 1787–1797
- Zhang, Y. *et al.* (2013) Chromatin connectivity maps reveal dynamic promoter–enhancer long-range associations. *Nature* 504, 306–310
- Calo, E. and Wysocka, J. (2013) Modification of enhancer chromatin: what, how, and why? *Mol. Cell* 49, 825–837
- Natoli, G. and Andrau, J.C. (2012) Noncoding transcription at enhancers: general principles and functional models. *Annu. Rev. Genet.* 46, 1–19
- Hurst, L.D. *et al.* (2015) The constrained maximal expression level owing to haploidy shapes gene content on the mammalian X chromosome. *PLoS Biol.* 13, e1002315
- Levo, M. and Segal, E. (2014) In pursuit of design principles of regulatory sequences. *Nat. Rev. Genet.* 15, 453–468
- Abdulrehman, D. *et al.* (2011) YEASTRACT: providing a programmatic access to curated transcriptional regulatory associations in *Saccharomyces cerevisiae* through a web services interface. *Nucleic Acids Res.* 39, D136–D140
- Landt, S.G. *et al.* (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.* 22, 1813–1831
- Graur, D. *et al.* (2013) On the immortality of television sets: “function” in the human genome according to the evolution-free gospel of ENCODE. *Genome Biol. Evol.* 5, 578–590
- Orenstein, Y. and Shamir, R. (2014) A comparative analysis of transcription factor binding models learned from PBM, HT-SELEX and ChIP data. *Nucleic Acids Res.* 42, e63
- Mathelier, A. *et al.* (2016) DNA shape features improve transcription factor binding site predictions *in vivo*. *Cell Syst.* 3, 278–286
- Zhou, T. *et al.* (2015) Quantitative modeling of transcription factor binding specificities using DNA shape. *Proc. Natl Acad. Sci. U. S. A.* 112, 4654–4659
- Jolma, A. *et al.* (2015) DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature* 527, 384–388
- Rister, J. and Desplan, C. (2010) Deciphering the genome's regulatory code: the many languages of DNA. *Bioessays* 32, 381–384
- Segal, E. *et al.* (2008) Predicting expression patterns from regulatory sequence in *Drosophila* segmentation. *Nature* 451, 535–540
- Rivera-Pomar, R. and Jackle, H. (1996) From gradients to stripes in *Drosophila* embryogenesis: filling in the gaps. *Trends Genet.* 12, 478–483
- Zinzen, R.P. *et al.* (2009) Combinatorial binding predicts spatiotemporal *cis*-regulatory activity. *Nature* 462, 65–70
- Beer, M.A. and Tavazoie, S. (2004) Predicting gene expression from sequence. *Cell* 117, 185–198
- Rautio, S. and Lahdesmaki, H. (2015) MixChIP: a probabilistic method for cell type specific protein–DNA binding analysis. *BMC Bioinformatics* 16, 413
- Lavender, C.A. *et al.* (2016) Downstream antisense transcription predicts genomic features that define the specific chromatin environment at mammalian promoters. *PLoS Genet.* 12, e1006224
- Swaminathan, A. and Beilharz, T.H. (2016) Epitope-tagged yeast strains reveal promoter driven changes to 3'-end formation and convergent antisense-transcription from common 3' UTRs. *Nucleic Acids Res.* 44, 377–386
- Potier, D. *et al.* (2014) Identification of *cis*-regulatory modules encoding temporal dynamics during development. *BMC Genomics* 15, 534
- Rao, S.S. *et al.* (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 159, 1665–1680

38. Ulianov, S.V. *et al.* (2015) Nuclear compartments, genome folding, and enhancer-promoter communication. *Int. Rev. Cell Mol. Biol.* 315, 183–244
39. Taher, L. *et al.* (2013) Sequence signatures extracted from proximal promoters can be used to predict distal enhancers. *Genome Biol.* 14, R117
40. Chen, X. *et al.* (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* 133, 1106–1117
41. Ouyang, Z. *et al.* (2009) ChIP-seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. *Proc. Natl Acad. Sci. U. S. A.* 106, 21521–21526
42. Natarajan, A. *et al.* (2012) Predicting cell-type-specific gene expression from regions of open chromatin. *Genome Res.* 22, 1711–1722
43. Vallejos, C.A. *et al.* (2016) Beyond comparisons of means: understanding changes in gene expression at the single-cell level. *Genome Biol.* 17, 70
44. Katayama, S. *et al.* (2013) SAMstr: statistical test for differential expression in single-cell transcriptome with spike-in normalization. *Bioinformatics* 29, 2943–2945
45. Finak, G. *et al.* (2015) MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* 16, 278
46. Satija, R. *et al.* (2015) Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* 33, 495–502
47. Sherman, M.S. *et al.* (2015) Cell-to-cell variability in the propensity to transcribe explains correlated fluctuations in gene expression. *Cell Syst.* 1, 315–325
48. Daigle, B.J., Jr. *et al.* (2015) Inferring single-cell gene expression mechanisms using stochastic simulation. *Bioinformatics* 31, 1428–1435
49. Kim, J.K. *et al.* (2015) Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression. *Nat. Commun.* 6, 8687
50. Bacher, R. and Kendziora, C. (2016) Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biol.* 17, 63
51. Stegle, O. *et al.* (2015) Computational and analytical challenges in single-cell transcriptomics. *Nat. Rev. Genet.* 16, 133–145
52. Plessey, C. *et al.* (2013) Population transcriptomics with single-cell resolution: a new field made possible by microfluidics: a technology for high throughput transcript counting and data-driven definition of cell types. *Bioessays* 35, 131–140
53. Jones, D.L. *et al.* (2014) Promoter architecture dictates cell-to-cell variability in gene expression. *Science* 346, 1533–1536
54. Nettelbeck, D.M. *et al.* (2000) Gene therapy: designer promoters for tumour targeting. *Trends Genet.* 16, 174–181
55. Mehrotra, R. *et al.* (2011) Designer promoter: an artwork of cis engineering. *Plant Mol. Biol.* 75, 527–536
56. Mogno, I. *et al.* (2010) TATA is a modular component of synthetic promoters. *Genome Res.* 20, 1391–1397
57. Gertz, J. and Cohen, B.A. (2009) Environment-specific combinatorial cis-regulation in synthetic promoters. *Mol. Syst. Biol.* 5, 244
58. Gertz, J. *et al.* (2009) Analysis of combinatorial cis-regulation in synthetic and genomic promoters. *Nature* 457, 215–218
59. Mogno, I. *et al.* (2013) Massively parallel synthetic promoter assays reveal the *in vivo* effects of binding site variants. *Genome Res.* 23, 1908–1915
60. Lodish, H.F. *et al.* (2000) Regulatory sequences in eukaryotic protein-coding genes. In *Molecular Cell Biology*. (4th edn), W.H. Freeman pp 10.4
61. Van Nostrand, E.L. and Kim, S.K. (2013) Integrative analysis of *C. elegans* modENCODE ChIP-seq data sets to infer gene regulatory interactions. *Genome Res.* 23, 941–953
62. Gagniuc, P. and Ionescu-Tirgoviste, C. (2012) Eukaryotic genomes may exhibit up to 10 generic classes of gene promoters. *BMC Genomics* 13, 512
63. Rydenfelt, M. *et al.* (2014) The influence of promoter architectures and regulatory motifs on gene expression in *Escherichia coli*. *PLoS One* 9, e114347
64. Madan Babu, M. and Teichmann, S.A. (2003) Evolution of transcription factors and the gene regulatory network in *Escherichia coli*. *Nucleic Acids Res.* 31, 1234–1244
65. Cox, R.S., III *et al.* (2007) Programming gene expression with combinatorial promoters. *Mol. Syst. Biol.* 3, 145
66. Tirosh, I. *et al.* (2008) On the relation between promoter divergence and gene expression evolution. *Mol. Syst. Biol.* 4, 159
67. Ebisuya, M. *et al.* (2008) Ripples from neighbouring transcription. *Nat. Cell Biol.* 10, 1106–1113
68. Okada, Y. *et al.* (2007) A three-kilobase fragment of the human Robo4 promoter directs cell type-specific expression in endothelium. *Circ. Res.* 100, 1712–1722
69. Okada, Y. *et al.* (2008) A GABP-binding element in the Robo4 promoter is necessary for endothelial expression *in vivo*. *Blood* 112, 2336–2339
70. Cai, X. *et al.* (2010) A 350 bp region of the proximal promoter of *Rds* drives cell-type specific gene expression. *Exp. Eye Res.* 91, 186–194
71. de Bousnac, H. *et al.* (2010) The ERK1/2-hepatocyte nuclear factor 4 α axis regulates human *ABCC6* gene expression in hepatocytes. *J. Biol. Chem.* 285, 22800–22808
72. Tiwari, B. and Junghans, R.P. (2005) Functional analysis of the mouse *Fcgrt* 5' proximal promoter. *Biochim. Biophys. Acta* 1681, 88–98
73. Breuer, J. (2006) *Introduction to the Theory of Sets*, Dover
74. Ezer, D. *et al.* (2014) Homotypic clusters of transcription factor binding sites: a model system for understanding the physical mechanics of gene expression. *Comput. Struct. Biotechnol. J.* 10, 63–69
75. Navarro, C. *et al.* (2014) CisMiner: genome-wide *in-silico* cis-regulatory module prediction by fuzzy itemset mining. *PLoS One* 9, e108065
76. Wang, D. *et al.* (2016) ANFIS-based fuzzy systems for searching DNA-protein binding sites. *bioRxiv* Published online June 15, 2016. <http://dx.doi.org/10.1101/058800>
77. Zadeh, L. (1965) Fuzzy sets. *Inf. Control* 8, 338–353
78. Torres, A. and Nieto, J.J. (2006) Fuzzy logic in medicine and bioinformatics. *J. Biomed. Biotechnol.* 2006, 91908
79. Azar, A.T. and Hassani, A.E. (2015) Editorial on: fuzzy logic in biomedicine. *Comput. Biol. Med.* 64, 321–322
80. Slattey, M. *et al.* (2014) Absence of a simple code: how transcription factors read the genome. *Trends Biochem. Sci.* 39, 381–399
81. Fisher, W.W. *et al.* (2012) DNA regions bound at low occupancy by transcription factors do not drive patterned reporter gene expression in *Drosophila*. *Proc. Natl Acad. Sci. U. S. A.* 109, 21330–21335
82. Zhao, L. *et al.* (2013) STAT1 regulates human glutaminase 1 promoter activity through multiple binding sites in HIV-1 infected macrophages. *PLoS One* 8, e76581
83. Yarrington, R.M. *et al.* (2015) Spatiotemporal cascade of transcription factor binding required for promoter activation. *Mol. Cell Biol.* 35, 688–698
84. Corbo, J.C. *et al.* (2010) CRX ChIP-seq reveals the cis-regulatory architecture of mouse photoreceptors. *Genome Res.* 20, 1512–1525
85. White, M.A. *et al.* (2016) A simple grammar defines activating and repressing cis-regulatory elements in photoreceptors. *Cell Rep.* 17, 1247–1254
86. Peterson, K.A. *et al.* (2012) Neural-specific Sox2 input and differential Gli-binding affinity provide context and positional information in Shh-directed neural patterning. *Genes Dev.* 26, 2802–2816
87. Ramos, A.I. and Barolo, S. (2013) Low-affinity transcription factor binding sites shape morphogen responses and enhancer evolution. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 368, 20130018
88. Amosti, D.N. and Kulkarni, M.M. (2005) Transcriptional enhancers: intelligent enhanceosomes or flexible billboards? *J. Cell. Biochem.* 94, 890–898

89. Wu, K. *et al.* (2016) A double suicide gene system driven by vascular endothelial growth factor promoter selectively kills human hepatocellular carcinoma cells. *Oncol. Lett.* 11, 3152–3160
90. Rincon, M.Y. *et al.* (2015) Genome-wide computational analysis reveals cardiomyocyte-specific transcriptional *cis*-regulatory motifs that enable efficient cardiac gene therapy. *Mol. Ther.* 23, 43–52
91. Narlikar, L. *et al.* (2010) Genome-wide discovery of human heart enhancers. *Genome Res.* 20, 381–392
92. Zadeh, L.A. (1982) A note on prototype theory and fuzzy sets. *Cognition* 12, 291–297
93. Meyer, D. and Hornik, K. (2009) Generalized and customizable sets in R. *J. Stat. Softw.* 31 (1), 1–27

Correction

Can We Predict Gene Expression by
Understanding Proximal
Promoter Architecture?

Łukasz Huminiecki* and Jarosław Horbańczuk

*Correspondence: l.huminiecki@ighz.pl (Ł. Huminiecki).DOI of original article: <https://doi.org/10.1016/j.tibtech.2017.03.007>

(Trends in Biotechnology 35, 530–546, 2017)

The published article inadvertently omitted the Acknowledgments section, which should have acknowledged a funding source. The correct funding attribution is as follows:

“LH was funded to perform this work by the National Science Centre, Poland, grant POLONEZ 2 (grant number 2016/21/P/NZ2/03926). This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 665778. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.”