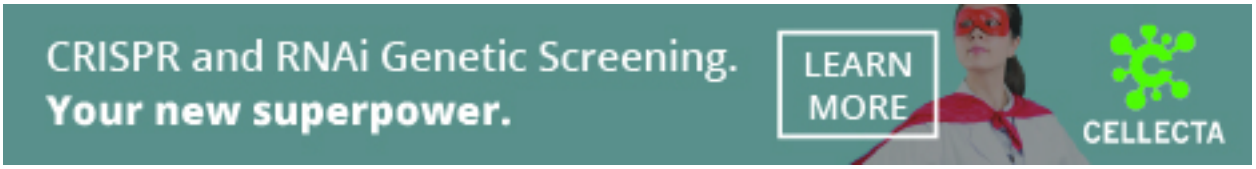
Promoting transcriptome diversity

Robert L. Strausberg and Samuel Levy

Genome Res. 2007 17: 965-968
Access the most recent version at doi:10.1101/gr.6499807



Perspective

Promoting transcriptome diversity

Robert L. Strausberg[1] and Samuel Levy

The J. Craig Venter Institute, Rockville, Maryland 20850, USA

Although the number of protein-encoding human genes is more limited than many had estimated, the human transcript repertoire is much more diverse than anticipated. In part, transcript diversity is generated through the use of alternative promoters and alternate splicing. In addition, based on discoveries using technologies such as full-length cDNA libraries and whole genome tiling microarrays, it is now likely that non-protein-encoding transcripts comprise a substantial fraction of the human RNA population. Much attention is currently focused on understanding the role of alternative promoters in generating transcript diversity, both for non-protein-encoding (ncRNAs) and protein-encoding RNAs.

Over the past decade, our concept of the human gene repertoire has changed dramatically such that the one-promoter–one-gene–one-transcript–one-protein concept no longer provides a realistic view of human (and other eukaryotic) genes (Carninci et al. 2005; Frith et al. 2006; Furuno et al. 2006; Mattick and Makunin 2006; Willingham and Gingeras 2006). It is not possible to fully appreciate the complexity of the human transcriptome without referring to an alternative universe including alternative promoters, alternatively spliced transcripts, and alternative polyadenylation. Perhaps most surprising has been the discovery of new families of RNA species that do not encode proteins (non-coding RNAs, ncRNAs). The interplay of these transcriptional elements has changed the concept of a "gene" such that it is now clear that one gene can encode a repertoire of transcripts engaged in diverse biological activities. Therefore, while there was much discussion of the limited coding content of the human genome based on analysis of the draft sequences, it almost immediately became evident that the transcriptome was very rich in complexity (Pennisi 2005). Currently, the substantial role of alternative promoter utilization in driving human transcript complexity is becoming more evident, especially based on the availability of full-length cDNA and genome sequences.

Characterizing alternative promoters

A major research area for modern genomics is the role of alternative promoters in driving transcript production. While it has been clear for many years that some eukaryotic genes use multiple promoters that can drive expression in specific tissues and/or developmental stages, evidence that alternative promoters play a major role in transcript diversity is expanding rapidly (Carninci 2006). The increased knowledge of the existence of alternative promoters creates new opportunities for discovering mechanisms for generating and regulating biological complexity and understanding features that make species unique. Recently a pioneering series of papers has appeared in *Genome Research* (Kimura et al. 2006; Baek et al. 2007), including the communication of Tsuritani et al. (2007) in the present issue. These studies present early analyses of the structure/function relationships of promoters and alternative promoters and provide context for pursuing key issues in eukaryotic gene expression. For example, are all alternative promoters functional? How do they participate in tissue-specific gene expression? How do they evolve? What is the relationship of promoters for protein-encoding RNAs and ncRNAs?

To study these and other questions in a comprehensive manner, current research has focused on the identification and characterization of the sequence structure of alternative promoters, and the delineation of transcription factors that interact with those promoters. Through these studies it is anticipated that a better understanding of the rules governing the structure of use of promoters in non-protein-encoding genes will emerge. Toward that end, full-length cDNA sequencing, especially from human and mouse, has provided data sets that give insight not only into alternative splice forms, but also the utilization of alternative promoters (Okazaki et al. 2002; Strausberg et al. 2002; Imanishi et al. 2004; Ota et al. 2004; Carninci et al. 2005; Takeda et al. 2006). Especially valuable for these studies have been cDNA libraries generated with technologies that select for the presence of the 5′-CAP. Use of these CAP selection techniques is especially critical to the analysis of transcription start sites, providing evidence that they are, indeed, transcripts and not genomic contaminants and that the cDNA sequence carries the in vivo initiation site (Kimura et al. 2006; Baek et al. 2007; Tsuritani et al. 2007). The study of Tsuritani et al. (2007) focuses on the characterization and utilization of alternative promoters in human and mouse, encompassing protein-encoding as well as non-protein-encoding genes. A particularly unique and bold aspect to this study is the analysis of promoters that are evolutionarily conserved, as well as those that seem to be species-unique. The results complement those present in an earlier publication of this group (Kimura et al. 2006) as well as the paper of Baek et al. (2007), and it is important to consider the current findings in relationship to these previous studies.

In their initial study of human promoters, Kimura et al. (2006) used nearly 1.8 million CAP-selected full-length cDNAs derived from >160 cDNA libraries. The 5′-end sequences were clustered into bins based on their physical separation (>500 bp) in the genome. These clustered transcriptional start sites (TSSs) were then considered in relation to the best characterized human genes, those in the RefSeq (http://www.ncbi.nlm.nih.gov/RefSeq/) set. Remarkably, this analysis revealed that one or more alternative promoters are used in more than half of the RefSeq genes. Given that the depth and breadth of the transcriptome based on cDNA sequences are still relatively light, this should be a minimum estimate. Analysis of the sequence content of these putative promoter regions revealed differences between genes

[1]Corresponding author.
E-mail rls@venterinstitute.org; fax (240) 268-4000.
Article is online at http://www.genome.org/cgi/doi/10.1101/gr.6499807.

that have one promoter compared with those with multiple alternative promoters. These differences were based on the content of TATA-boxes and CpG islands and revealed that the number of CpG island–containing promoters per gene did not increase commensurate with increased numbers of alternative promoters. Based on prior studies, it was hypothesized that the non-CpG island–containing alternative promoters might be associated with tissue or developmental stage-specific expression. Analysis of expression patterns based on the cDNA library tissue source confirmed this notion. For example, putative alternative promoters are preferentially used for expression in the testis and the brain, perhaps contributing to species-specific characteristics. Moreover, the pattern-specific expression extended to genes based on GO categories, with those genes associated with genes involved in signal-transduction-related activities, such as ATP binding, phosphorylation, and kinase, most highly associated with genes with putative alternative promoters.

Comparative genomic analysis of alternative promoters

With the increasing availability of genomic sequences across a diverse set of organisms, an attractive opportunity of assessing potential promoter sequence derives from comparative genomics. A very nice example of this approach is the assembly and evolutionary comparison of conserved alternative promoters based on the mouse and human full-length cDNAs (Baek et al. 2007). The structural discriminators used in this study were based on CpG richness. The results suggested that about half of all human genes have evolutionarily conserved alternative promoters. Interestingly, in comparison with promoter sequences that represent sole initiation sites within genes, the sequences of alternative promoters are more conserved between human and mouse. In keeping with the notion that alternative promoters enrich the biological-complexity-coding capacity of genes, differences in gene expression patterns were observed for genes with alternative promoters compared with those for single promoters. For example, genes with alternative promoters were observed to be more abundantly expressed in tissues such as the brain, heart, and liver, and in early development. Differences in expression patterns were also correlated with promoter sequence characteristics such as CpG island content and TATA-box frequencies.

For example, upstream promoters are more likely to be CpG-rich and associated with higher expression levels. Moreover, promoters that are in larger clusters (more alternative promoters) are more highly conserved, suggesting the need for conservation based on competitive signals. Interestingly, the presence of TATA-boxes is much reduced in alternative promoters and specific hexamers are more frequent, possibly representing, for example, the use of tissue and developmental stage-specific transcription activators. On the other hand, genes associated with widely expressed housekeeping functions generally have single promoters that are CpG-rich and more relaxed sequence conservation, as well as frequent use of TATA-boxes. Based on their studies, these investigators have developed discriminators that will be useful in predicting genes with alternative or single promoters, thereby enabling discovery of such putative elements lacking other experimental support.

The study of Tsuritani et al. (2007) in this issue of *Genome Research* expands the analysis of putative alternative promoters to those lacking evidence for evolutionary conservation. This type of promoter potentially includes those most associated with species-specific functions and that might be recently evolved or evolving. Similar to other studies, these investigators showed evidence to support the notion that at least half of the RefSeq genes use alternative promoters. Comparative sequence analysis revealed that most alternative promoters lack clear evidence of evolutionary conservation. The structural and functional characteristics of conserved and nonconserved alternative promoters are markedly different, and point to the challenges in characterizing the nonconserved elements. In general, the nonconserved promoters are of lower usage, have high AT content, and are rich in repetitive sequences. Moreover, they tend to be more associated with large changes in encoded protein, and with non-protein-encoding genes. These could include, for example, genes encoding microRNAs as well as longer ncRNAs. Typically, conserved promoters are associated with alternate usage of first exons, whereas the distribution of nonconserved promoter regions varies from first to distal exons, and about one-third are predicted to result in amino acid sequences of <100 amino acids, whereas the vast majority of conserved promoter regions drive expression of larger proteins. Therefore, a conclusion of this study is that a high proportion of the nonconserved promoters might be associated with non-protein-encoding genes, and cellular expression studies in which promoters were linked with a reporter were in support of this idea.

Overall, comparative sequence and functional analysis represents one of the exciting new opportunities to define the coding potential and regulation of mammalian genomes. Extending these types of analyses to an even wider diversity of organisms, including those that are very anciently related to humans (Venkatesh et al. 2006), will help to define features that are common among complex eukaryotes and those that may be lineage-specific.

Genome-wide association of transcription factors and epigenetic patterns with putative promoter regions

Complementary to the studies described above will be the integration of these findings with additional features that contribute to promoter function including epigenetic modification as well as the characterization of the regulatory proteins associated with putative promoter regions. Toward that end, several early studies have been reported that build on genome-wide approaches.

Using genomic tiling arrays comprising oligonucleotides spaced approximately every 35 nucleotides along the entire nonrepetitive region of chromosomes 21 and 22 (Kapranov et al. 2002), target regions were identified for three transcription factors, SP1, MYC, and TP53 (Cawley et al. 2004). Consistent with extensive genomic transcription, widespread (but specific) binding sites were identified, ranging from 1600 instances for TP53 to >12,000 instances for SP1. Several characteristics of the putative transcription factor binding sites were especially of interest, including the association of binding sites with regions producing ncRNAs and the coordinated regulation in response to retinoic acid of both protein-encoding RNAs and ncRNAs. Importantly, the numbers and positions of transcription factor binding sites were consistent with the large number of non-protein-encoding genes revealed by these microarrays.

Studies that extend these analyses of transcription-factor-binding sites to the entire human genome highlight the chal-

lenges and opportunities that exist (Yang et al. 2006). In this case, ~5800 binding sites were identified for the transcription factor TP73L (formerly known as p63). While a conserved sequence motif for TP73L binding was identified, many of those sites were not populated by TP73L in this assay, while other sequence motifs and nonconserved regions were also associated with TP73L binding. Therefore, it appears that transcription factor binding is complex, determined in part by other features in addition to direct DNA–protein interaction. Among those features might be epigenetic modifications of the genome and its regulatory regions.

Indeed, there already is evidence of differential methylation of promoters based on specificity of tissue expression as well as differences among promoters by type (Cheong et al. 2006; Hatada et al. 2006; Kawaji et al. 2006; Cooper et al. 2007; Heintzman et al. 2007). For example, in the study of Cheong and colleagues, the methylation status of tissue-biased promoters was compared with that of alternative promoters not exhibiting tissue bias (Cheong et al. 2006). Among the many interesting findings in this study was the observation that methylation status differed substantially among alternative promoters within a gene (in part related to the CpG content of the promoter) and that the methylation status correlated with tissue-specific expression. The study suggested that differential expression derived from alternative promoters might be related to differences in promoter sequence type, especially the presence of CpG islands.

Highly complementary to the studies described above are computational analyses that are already driving the discovery process for regulatory elements. For example, a comparative genomics approach involving a systematic analysis of the human and mouse genomes, together with those of the rat and dog, revealed 174 potential transcription factor binding sites, more than half of which are newly discovered (Xie et al. 2005). Moreover, examination of the 3′-untranslated regions revealed >100 motifs potentially associated with post-transcriptional regulation. Approximately half of these latter sites are associated with miRNAs, including many that were previously unknown. A remarkable result of this analysis is the notion that miRNAs are implicated in the regulation of at least 20% of human genes.

Keeping an open mind with open technology platforms

For many years, alternative transcript forms were very much below the radar screen of human genomics. Although there was awareness of interesting examples of alternative splicing events with functional consequences, much of the genomic era has still focused on the concept of one meaningful transcript per gene. For example, gene expression arrays based on oligonucleotides or cDNAs were designed to measure transcript production of a gene in a manner that captures the total signal of the transcript population of a gene based on the hybridization of 3′-end sequences. Therefore, the potential richness of those transcripts, including tissue-specific differences in expression that might encompass qualitative as well as quantitative differences, was undetectable. Moreover, the reliance on RefSeq genes and/or UniGene clusters (http://www.ncbi.nlm.nih.gov/RefSeq) meant that such platforms were essentially closed based on known genes and transcripts.

The underpinnings of transcript discovery have been open platforms that allow for observations that are not influenced by preconceived notions. This has been especially true for sequence-based approaches based on full-length cDNAs (Okazaki et al. 2002; Strausberg et al. 2002; Imanishi et al. 2004) and transcript tagging technologies (Iseli et al. 2002; Strausberg et al. 2003; Peters and Velculescu 2005; Carninci 2006; Gowda et al. 2006; Kodzius et al. 2006). It is now increasingly becoming the case for hybridization-based technologies such as oligonucleotide arrays, in large part based on the development of whole-genome tiling arrays.

As sequencing technology becomes higher throughput and less costly, the pace of discovery will increase dramatically. For example, pyrosequencing technology has recently been used to study the transcript population in human and plant cells (Bainbridge et al. 2006; Cheung et al. 2006). With read lengths increasing to ~250 nt and with ~400,000 reads per sequencing run, such technologies open new opportunities to look at transcript variation in a wide variety of biological states. In addition, Gowda et al. (2006) recently described the application of pyrosequencing technology to the analysis of 5′-transcript ends prepared by the robust analysis of 5′-transcript ends (5′-RATE), suggesting that a much higher coverage of 5′-end tags will become available. Therefore, it is likely that the discovery of both protein-encoding as well as non-protein-encoding transcripts is still at an early stage.

Key to overall progress will be the development of community tools including databases that will enable precise cataloging of transcripts, their promoters, and regulatory protein complexes. The integration of these data sets (The Encode Project Consortium 2004, 2007; Guigó et al. 2006; Thomas et al. 2007), hopefully including controlled experiments in which each of these features is assayed in the same biological material, will provide a basis for finding the most informative features from complexity and drive the development of the parts list of the transcriptome and how those parts work synergistically to drive our biology. A universe of opportunity awaits.

References

Baek, D., Davis, C., Ewing, B., Gordon, D., and Green, P. 2007. Characterization and predictive discovery of evolutionarily conserved mammalian alternative promoters. *Genome Res.* **17:** 145–155.

Bainbridge, M.N., Warren, R.L., Hirst, M., Romanuik, T., Zeng, T., Go, A., Delaney, A., Griffith, M., Hickenbotham, M., Magrini, V., et al. 2006. Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. *BMC Genomics* **7:** 246. doi: 10.1186/1471-2164-7-246.

Carninci, P. 2006. Tagging mammalian transcription complexity. *Trends Genet.* **22:** 501–510.

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M.C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C., et al. FANTOM Consortium, RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group). 2005. The transcriptional landscape of the mammalian genome. *Science* **309:** 1559–1563.

Cawley, S., Bekiranov, S., Ng, H.H., Kapranov, P., Sekinger, E.A., Kampa, D., Piccolboni, A., Sementchenko, V., Cheng, J., Williams, A.J., et al. 2004. Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* **116:** 499–509.

Cheong, J., Yamada, Y., Yamashita, R., Irie, T., Kanai, A., Wakaguri, H., Nakai, K., Ito, T., Saito, I., Sugano, S., et al. 2006. Diverse DNA methylation statuses at alternative promoters of human genes in various tissues. *DNA Res.* **13:** 155–167.

Cheung, F., Haas, B.J., Goldberg, S.M., May, G.D., Xiao, Y., and Town, C.D. 2006. Sequencing *Medicago truncatula* expressed sequenced tags using 454 Life Sciences technology. *BMC Genomics* **7:** 272. doi: 10.1186/1471-2164-7-272.

Cooper, S.J., Trinklein, N.D., Nguyen, L., and Myers, R.M. 2007. Serum

response factor binding sites differ in three human cell types. *Genome Res.* **17:** 136–144.
The Encode Project Consortium. 2004. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306:** 636–640.
The Encode Project Consortium. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447:** 799–816.
Frith, M.C., Bailey, T.L., Kasukawa, T., Mignone, F., Kummerfeld, S.K., Madera, M., Sunkara, S., Furuno, M., Bult, C.J., Quackenbush, J., et al. 2006. Discrimination of non-protein-coding transcripts from protein-coding mRNA. *RNA Biol.* **3:** 40–48.
Furuno, M., Pang, K.C., Ninomiya, N., Fukuda, S., Frith, M.C., Bult, C., Kai, C., Kawai, J., Carninci, P., Hayashizaki, Y., et al. 2006. Clusters of internally primed transcripts reveal novel long noncoding RNAs. *PLoS Genet.* **2:** e37.
Gowda, M., Li, H., Alessi, J., Chen, F., Pratt, R., and Wang, G.L. 2006. Robust analysis of 5′-transcript ends (5′-RATE): A novel technique for transcriptome analysis and genome annotation. *Nucleic Acids Res.* **34:** e126.
Guigó, R., Flicek, P., Abril, J.F., Reymond, A., Lagarde, J., Denoeud, F., Antonarakis, S., Ashburner, M., Bajic, V.B., Birney, E., et al. 2006. EGASP: The human ENCODE Genome Annotation Assessment Project. *Genome Biol.* **7:** S2.1–S2.31.
Hatada, I., Fukasawa, M., Kimura, M., Morita, S., Yamada, K., Yoshikawa, T., Yamanaka, S., Endo, C., Sakurada, A., Sato, M., et al. 2006. Genome-wide profiling of promoter methylation in human. *Oncogene* **25:** 3059–3064.
Heintzman, N.D., Stuart, R.K., Hon, G., Fu, Y., Ching, C.W., Hawkins, R.D., Barrera, L.O., Van Calcar, S., Qu, C., Ching, K.A., et al. 2007. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* **39:** 311–318.
Imanishi, T., Itoh, T., Suzuki, Y., O'Donovan, C., Fukuchi, S., Koyanagi, K.O., Barrero, R.A., Tamura, T., Yamaguchi-Kabata, Y., Tanino, M., et al. 2004. Integrative annotation of 21,037 human genes validated by full-length cDNA clones. *PLoS Biol.* **2:** e162.
Iseli, C., Stevenson, B.J., de Souza, S.J., Samaia, H.B., Camargo, A.A., Buetow, K.H., Strausberg, R.L., Simpson, A.J., Bucher, P., and Jongeneel, C.V. 2002. Long-range heterogeneity at the 3′ ends of human mRNAs. *Genome Res.* **12:** 1068–1074.
Kapranov, P., Cawley, S.E., Drenkow, J., Bekiranov, S., Strausberg, R.L., Fodor, S.P., and Gingeras, T.R. 2002. Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296:** 916–919.
Kawaji, H., Frith, M.C., Katayama, S., Sandelin, A., Kai, C., Kawai, J., Carninci, P., and Hayashizaki, Y. 2006. Dynamic usage of transcription start sites within core promoters. *Genome Biol.* **7:** R118.
Kimura, K., Wakamatsu, A., Suzuki, Y., Ota, T., Nishikawa, T., Yamashita, R., Yamamoto, J., Sekine, M., Tsuritani, K., Wakaguri, H., et al. 2006. Diversification of transcriptional modulation: Large-scale identification and characterization of putative alternative promoters of human genes. *Genome Res.* **16:** 55–65.
Kodzius, R., Kojima, M., Nishiyori, H., Nakamura, M., Fukuda, S., Tagami, M., Sasaki, D., Imamura, K., Kai, C., Harbers, M., et al. 2006. CAGE: Cap Analysis of Gene Expression. *Nat. Methods* **3:** 211–222.
Mattick, J.S. and Makunin, I.V. 2006. Non-coding RNA. *Hum. Mol. Genet.* **15:** 17–29.
Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H., Kondo, S., Nikaido, I., Osato, N., Saito, R., Suzuki, H., et al. FANTOM Consortium, RIKEN Genome Exploration Research Group Phase I & II Team. 2002. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420:** 563–573.
Ota, T., Suzuki, Y., Nishikawa, T., Otsuki, T., Sugiyama, T., Irie, R., Wakamatsu, A., Hayashi, K., Sato, H., Nagai, K., et al. 2004. Complete sequencing and characterization of 21,243 full-length human cDNAs. *Nat. Genet.* **36:** 40–45.
Pennisi, E. 2005. Why do humans have so few genes? *Science* **309:** 80.
Peters, B.A. and Velculescu, V.E. 2005. Transcriptome PETs: A genome's best friends. *Nat. Methods* **2:** 93–94.
Strausberg, R.L., Feingold, E.A., Grouse, L.H., Derge, J.G., Klausner, R.D., Collins, F.S., Wagner, L., Shenmen, C.M., Schuler, G.D., Altschul, S.F., et al. 2002. Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. *Proc. Natl. Acad. Sci.* **99:** 16899–16903.
Strausberg, R.L., Simpson, A.J., and Wooster, R. 2003. Sequence-based cancer genomics: Progress, lessons and opportunities. *Nat. Rev. Genet.* **4:** 409–418.
Takeda, J., Suzuki, Y., Nakao, M., Barrero, R.A., Koyanagi, K.O., Jin, L., Motono, C., Hata, H., Isogai, T., Nagai, K., et al. 2006. Large-scale identification and characterization of alternative splicing variants of human gene transcripts using 56,419 completely sequenced and manually annotated full-length cDNAs. *Nucleic Acids Res.* **34:** 3917–3928.
Thomas, D.J., Rosenbloom, K.R., Clawson, H., Hinrichs, A.S., Trumbower, H., Raney, B.J., Karolchik, D., Barber, G.P., Harte, R.A., Hillman-Jackson, J., et al. ENCODE Project Consortium. 2007. The ENCODE Project at UC Santa Cruz. *Nucleic Acids Res.* **35:** D663–D667.
Tsuritani, K., Irie, T., Yamashita, R., Sakakibara, Y., Wakaguri, H., Kanai, A., Mizushima-Sugano, J., Sugano, S., Nakai, K., and Suzuki, Y. 2007. Distinct class of putative human-specific promoters: Comparative studies of alternative promoters of human and mouse genes. *Genome Res.* (this issue) doi: 10.1101gr.6030107.
Venkatesh, B., Kirkness, E.F., Loh, Y.H., Halpern, A.L., Lee, A.P., Johnson, J., Dandona, N., Viswanathan, L.D., Tay, A., Venter, J.C., et al. 2006. Ancient noncoding elements conserved in the human genome. *Science* **314:** 1892.
Willingham, A.T. and Gingeras, T.R. 2006. TUF love for "junk" DNA. *Cell* **125:** 1215–1220.
Xie, X., Lu, J., Kulbokas, E.J., Golub, T.R., Mootha, V., Lindblad-Toh, K., Lander, E.S., and Kellis, M. 2005. Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature* **434:** 338–345.
Yang, A., Zhu, Z., Kapranov, P., McKeon, F., Church, G.M., Gingeras, T.R., and Struhl, K. 2006. Relationships between p63 binding, DNA sequence, transcription activity, and biological function in human cells. *Mol. Cell* **24:** 593–602.