Novel Transcribed Regions in the Human Genome

J. ROZOWSKY,* J. WU,† Z. LIAN,‡ U. NAGALAKSHMI,† J.O. KORBEL,* P. KAPRANOV,**
D. ZHENG,* S. DYKE,** P. NEWBURGER,†† P. MILLER,§,¶ T.R. GINGERAS,**
S. WEISSMAN,‡ M. GERSTEIN,*,¶ AND M. SNYDER*,†

*Molecular Biophysics & Biochemistry Department, †Molecular, Cellular & Developmental Biology Department, ‡Department of Genetics, §Center for Medical Informatics, and ¶Program in Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut 06520; **Affymetrix, Inc., Santa Clara, California 92024; ††University of Massachusetts Medical School, Children's Medical Center, North Worcester, Massachusetts 01605

We have used genomic tiling arrays to identify transcribed regions throughout the human genome. Analysis of the mapping results of RNA isolated from five cell/tissue types, NB4 cells, NB4 cells treated with retinoic acid (RA), NB4 cells treated with 12-O-tetradecanoylphorbol-13 acetate (TPA), neutrophils, and placenta, throughout the ENCODE region reveals a large number of novel transcribed regions. Interestingly, neutrophils exhibit a great deal of novel expression in several intronic regions. Comparison of the hybridization results of NB4 cells treated with different stimuli relative to untreated cells reveals that many new regions are expressed upon cell differentiation. One such region is the Hox locus, which contains a large number of novel regions expressed in a number of cell types. Analysis of the trinucleotide composition of the novel transcribed regions reveals that it is similar to that of known exons. These results suggest that many of the novel transcribed regions may have a functional role.

The human genome project has revealed the DNA sequence that governs all biological processes in humans (Lander et al. 2001; Venter et al. 2001). However, understanding how the sequence is interpreted to carry out cellular and developmental processes in humans is somewhat limited. In particular, we would like to identify the genes and the protein products they encode, the regulatory information that controls the level of expression of each RNA and protein, and how the different components function together to carry out complex molecular, cellular, developmental, and behavioral processes.

An important step in the process of characterizing the human genome is the identification of genes. This problem is particularly acute in mammals because the RNA coding segments of genes are split into relatively short exons of average length approximately 140 bp separated by often very large introns, which can sometimes be greater than 100,000 bp in length (Lander et al. 2001; Venter et al. 2001). In addition, the large number of alternatively spliced RNAs can make it difficult to determine which regions code for proteins. For these reasons, it can be extremely challenging to identify exons and genes, and computational approaches have generally been only partially successful (Burge and Karlin 1997; Guigo et al. 2006).

Other approaches for gene identification include (1) comparison of sequences among related species; in general, protein-coding genes are more conserved than noncoding regions (Cawley et al. 2003; Parra et al. 2003) and (2) mapping transcribed regions. For the latter case, sequences of cDNAs (ideally full length) and ESTs allow one to assign transcribed regions to segments of the genome. In general, ESTs have been less useful because of concerns that the sequences are often derived from unspliced RNAs and/or contaminating DNA in the starting RNA preparations. In contrast, full-length cDNAs have been quite successful for gene and exon identification (Carninci et al. 2005). However, many genes are expressed at a low level, and the large number of alternative spliced RNAs precludes a comprehensive analysis of transcribed regions using only cDNA information.

MAPPING TRANSCRIBED REGIONS USING TILING ARRAYS REVEALS EXTENSIVE TRANSCRIPTION IN THE HUMAN GENOME

We have been mapping transcribed regions using genomic tiling arrays. Tiling arrays enable the detection of transcribed regions unbiased by genome annotation. Initially, arrays containing 21,000 800-bp PCR products from human chromosome 22 were used to cover the nonrepetitive portions of the chromosome (Rinn et al. 2003). Subsequently, we built a 36-base oligonucleotide array that tiled the nonrepetitive sequence of both strands of the entire genome at a resolution of one oligonucleotide every 46 bp (Bertone et al. 2004). By probing the chromosome-22 array and whole-genome arrays using poly(A)$^+$ RNA isolated from placenta and liver, respectively, we found that approximately 60% of the transcribed regions lay outside of annotated exons. Validation by PCR confirmed that the majority of these regions are expressed in poly(A)$^+$ RNA. Thus, there is at least twice as much of the genome expressed as processed poly(A)$^+$ transcripts as previously identified.

More recently, we have probed an Affymetrix oligonucleotide array covering the nonrepetitive DNA of the ENCODE regions using RNAs isolated from five different cell types or tissues: NB4 cells (a lymphoid cell line); NB4 cells treated with retinoic acid (RA), NB4 cells treated with TPA, neutrophils (isolated from ten different patients), and placenta. NB4 cells treated with RA are thought to differentiate toward neutrophils; NB4 cells

treated with TPA are thought to differentiate toward monocytes. The placental sample was poly(A)$^+$ RNA, whereas the remaining samples used total RNA. The ENCODE regions comprise 44 genomic regions varying in size from 0.5 to 2 Mbp which collectively total 1% of the human genome (ENCODE Project Consortium 2004). The oligonucleotide probes are 25 nucleotides in length and overlap slightly such that they start on average every 20 bp. Hybridizing segments were scored using a sliding window approach described in Kampa et al. (2004) and Royce et al. (2005). Genomic sequences that are detected as transcribed are called either transcriptionally active regions, "TARs" (Rinn et al. 2003), or transcribed fragments, "transfrags" (Kapranov et al. 2002).

The results of these probings are revealed in Table 1. A large number (2,046) of novel transcribed regions were identified from all RNA samples. These new hybridizing regions lie in intergenic regions proximal (within 5 kb) or distal (greater than 5 kb) to annotated genes as well as in introns, proximal (within 5 kb) or distal (greater than 5 kb) to annotated exons or regions previously identified as ESTs which are not annotated as genes. In total, over twice as many transcribed regions were apparent as compared to previous annotation.

EXTENSIVE INTRONIC TRANSCRIPTON IN NEUTROPHILS

Although novel transcription was apparent in RNAs from each source, careful inspection of the results revealed that the intronic regions were often extensively expressed in neutrophils (see Table 1 and Fig. 1). Over 700 novel transcribed regions were expressed in neutrophil RNA; this pattern of expression was not generally observed for the other RNA samples. Most of the novel transcribed regions expressed in neutrophils were concentrated in a few ENCODE regions, in particular the HOXA locus.

COMPARISON OF EXPRESSED REGIONS REVEALS EXTENSIVE EXPRESSION CHANGES DURING CELL DIFFERENTIATION

Analysis of NB4 cells treated with different agents, RA and TPA, allowed us to compare transcriptional changes that occur upon cell differentiation. We found that NB4 cells treated with RA had transcriptional patterns similar to those of neutrophils, as expected. One such example in the TRIM22 region is shown in Figure 2. One region that exhibited particularly extensive transcription upon cell differentiation was the HOXA locus (Fig. 3). We observe that whereas a novel intergenic transcript is expressed in both the NB4 cells treated with RA and neutrophils, the HOXA1 gene is expressed in the RA-treated cells but not the neutrophils. Extensive intergenic transcription is apparent in this region in many cell types, such as NB4 cells treated with either RA or TPA. These results suggest that many novel transcribed regions are transcribed upon cell differentiation.

To gain further insight into the cell-type specificity of TARs, we determined the fraction of novel transcribed regions expressed in placental RNA that are expressed in other cell types and correspond to the same genomic coordinates. As shown in Figure 4, almost all of the novel transcribed regions expressed in placental RNA either entirely overlap with novel transcribed regions detected in other cell lines (ENCODE Project Consortium 2004) or do not overlap at all. These results demonstrate that novel transcribed regions are reproducible genomic regions which show varying expression profiles across different cell lines and tissues, similar to exons of known genes.

An initial attempt has been made to classify novel transcribed regions, using genomic location and expression profile across the mapped ENCODE RNAs, into those that are likely parts of alternative isoforms of known genes as well as those that correspond to novel transcribed loci (Rozowsky et al. 2006). Approximately 14% of the novel

Table 1. Distribution of Transcribed Regions (TARs) in the ENCODE Regions

	NB4 CTRL	NB4 RA	NB4 TPA	Neutrophil	Placenta
Count					
Gencode Exonic	727	552	728	880	2175
Intergenic Distal	52	29	69	61	147
Intergenic Proximal	55	35	70	81	99
Intronic Distal	26	15	36	211	80
Intronic Proximal	137	109	190	554	367
Other ESTs	60	46	79	299	168
Nucleotides					
Gencode Exonic	82,376	60,870	81,080	122,412	370,426
Intergenic Distal	3,564	2,009	4,544	4,353	10,668
Intergenic Proximal	3,499	2,157	4,468	7,185	6,638
Intronic Distal	1,991	1,283	2,598	35,631	5,171
Intronic Proximal	11,086	9,383	14,720	61,363	26,181
Other ESTs	3,958	4,247	7,050	38,209	16,723

Locations of transcribed sequences or TARs detected within the ENCODE regions for RNA from the following 5 cell lines/tissues: untreated NB4 cells, NB4 cells treated with RA, NB4 cells treated with TPA, neutrophils and placenta (poly(A)$^+$). Distributions are either the number of regions transcribed or the number of base pairs detected as being transcribed.

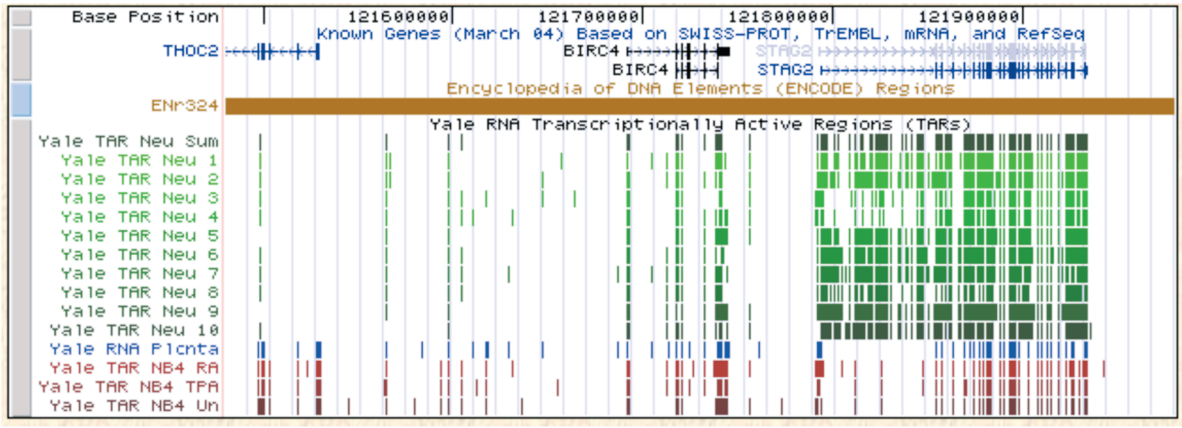

Figure 1. Plot of the genomic region surrounding the STAG2 locus on chromosome X. We observe that the entire STAG2 locus, including the intronic regions, is expressed in the RNA extracted from neutrophils.

TARs appear to connect with annotated loci, and approximately 21% of the remaining novel TARs are likely to be components of new annotated genes.

THE NOVEL TRANSCRIBED REGIONS HAVE A TRINUCLEOTIDE DISTRIBUTION SIMILAR TO EXONS

Our studies have revealed extensive transcription in the human genome. It is possible that this material represents random transcription throughout the human genome and that some fraction is maintained in stable poly(A)$^+$ RNA. To investigate this possibility, we analyzed the frequency of trinucleotide sequences of novel transcribed regions in intergenic and intronic regions and compared them with transcribed regions of exons, all GENCODE annotated exons (Harrow et al. 2006), and DNA from random selected nonrepetitive genomic intervals. As shown in Figure 5, transcribed regions and annotated exons show a nonrandom trinucleotide sequence composition. The composition of intergenic and intronic novel expressed regions is similar to that of transcribed annotated exons and all GENCODE exons, and very different from that of random DNA. Thus, novel transcribed regions have a sequence distribution similar to that of known genes, consistent with a functional role for these sequences.

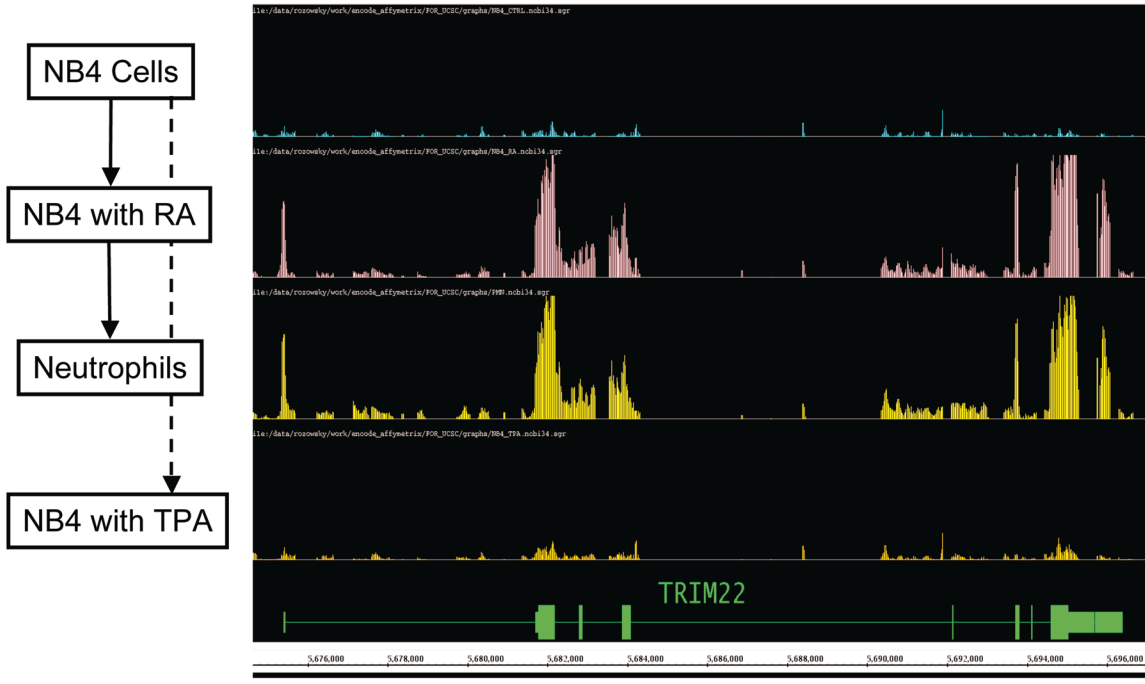

Figure 2. Plot of the expression profile of the TRIM22 locus on chromosome 11 for four of the RNAs mapped. This transcript is not expressed in the untreated NB4 cells; however, the transcript is present for both the NB4 cells treated with RA as well as the neutrophils, consistent with the hypothesis that RA differentiates NB4 cells toward neutrophils.

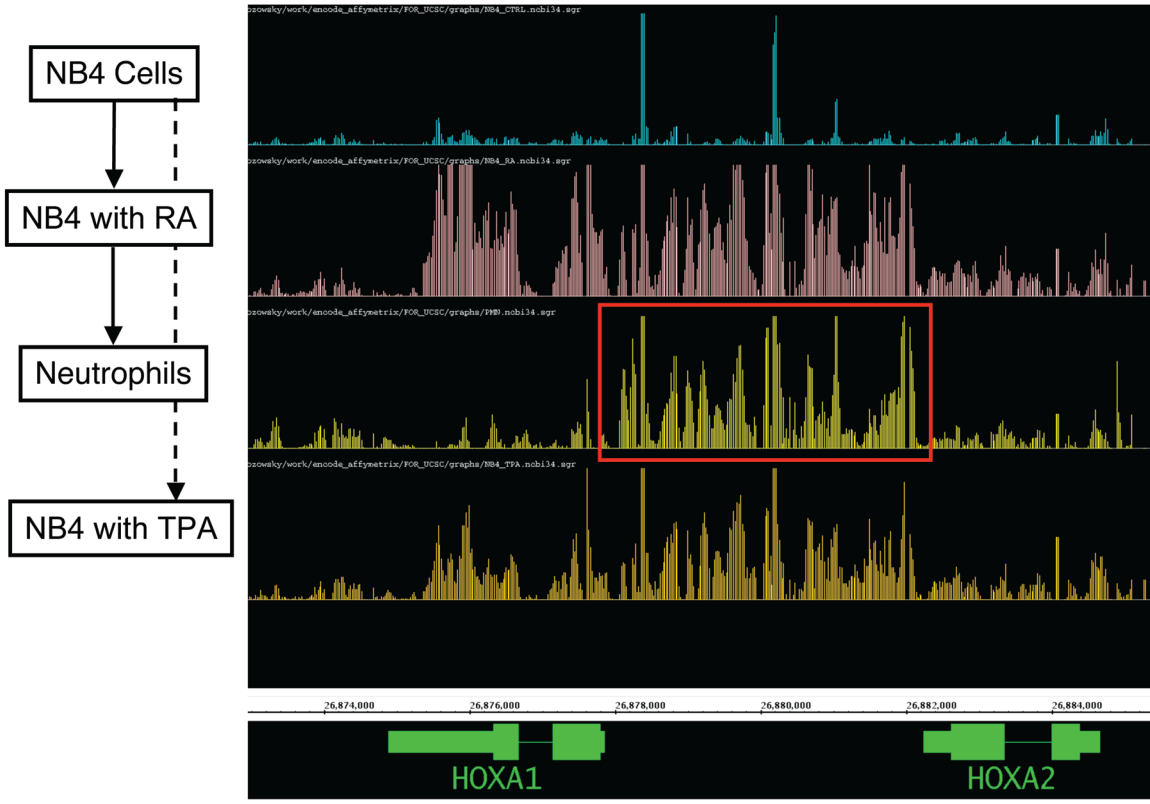

Figure 3. Plot of expression profiles of the genomic region from a portion of the HOXA locus on chromosome 7 for four of the RNAs mapped. The region displayed is not transcribed in the untreated NB4 cells; however, HOXA1 is expressed in the NB4 cells treated with RA or TPA. In addition, the intergenic region between HOXA1 and HOXA2 (shown with a *red box*) is transcribed in the NB4 cells treated with RA and TPA as well as the neutrophils. This transcript has been confirmed by RNA blot analysis.

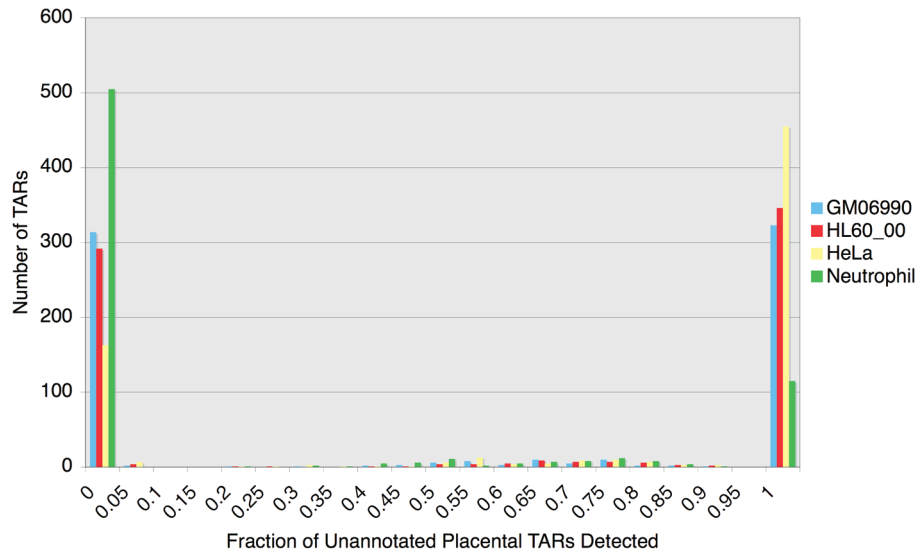

Figure 4. Plot of the number of novel TARs from four different cells lines mapped by the ENCODE project that overlap with the novel TARs detected in placental poly(A)$^+$ RNA. The horizontal axis shows the fraction of overlap between the novel TARs from each of the four cell lines vs the placental novel TARs. We observe that the novel TARs compared between the different cell/tissue types either overlap entirely or do not overlap at all. This is evidence that novel TARs are discrete transcribed regions similar to exons of known genes.

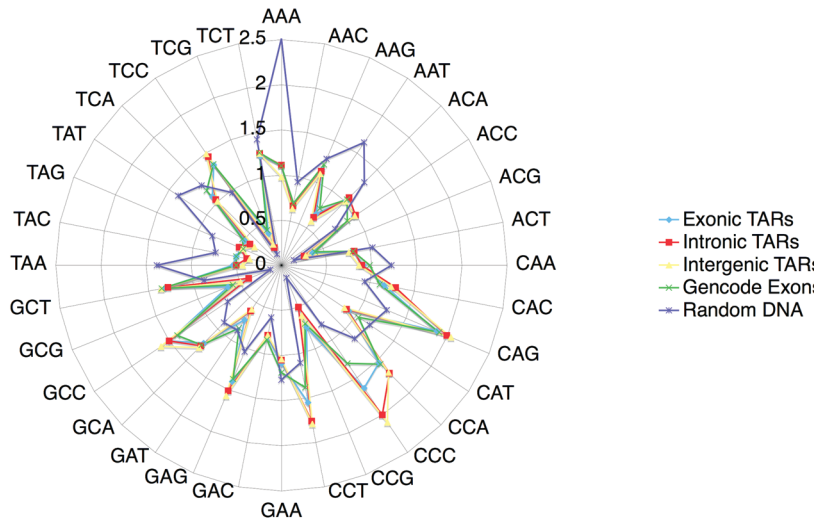

Figure 5. Radial plot of the relative trinucleotide frequency for the sequences of TARs that overlap exons of known genes, TARs that overlap introns of known genes, TARs that are in intergenic regions, exons of known genes, and DNA from randomly selected nonrepetitive genomic regions. We observe that all the TARs exhibit a trinucleotide frequency distribution similar to that for exons of known genes and which is unlike the distribution for randomly selected nonrepetitive genomic regions.

DISCUSSION

We demonstrate that there are many more expressed sequences in the human genome than previous appreciated. Neutrophils in particular have extensive intronic transcription compared to other cell types. The reasons for this are not clear. This may be because not all of the transcribed RNAs from neutrophils are fully processed RNAs. A fraction of RNAs from some loci may remain in the cell as unspliced primary transcripts.

By analyzing RNA samples derived from cells treated with different stimuli, we examined the pattern of transcribed regions upon cell differentiation. We found many new transcribed regions are expressed in cell types upon differentiation, suggesting that they may contribute to new cell identities. Of particular interest is the extensive transcription throughout the HOXA region. It is likely that there are many more transcribed in this region than previously known. Consistent with this, RNA blot analysis reveals at least one new transcript encoded in the intergenic region in this locus (data not shown).

Although it is possible that much of the new transcription is due to random transcription throughout the genome, analysis of the trinucleotide composition of the novel expressed regions reveals that it is similar to annotated regions and not random DNA, similar to that reported previously (Bertone et al. 2004). This suggests that the new transcribed regions are either likely to be derived from similar genomics regions and/or that they have functional roles. Further characterization of the novel transcribed RNAs and the regions that encode them is likely to determine the function of these transcripts.

ACKNOWLEDGMENTS

This work was supported by grants from the National Institutes of Health.

REFERENCES

Bertone P., Stolc V., Royce T.E., Rozowsky J.S., Urban A.E., Zhu X., Rinn J.L., Tongprasit W., Samanta M., Weissman S., et al. 2004. Global identification of human transcribed sequences with genome tiling arrays. *Science* **306:** 2242.

Burge C. and Karlin S. 1997. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268:** 78.

Carninci P., Kasukawa T., Katayama S., Gough J., Frith M.C., Maeda N., Oyama R., Ravasi T., Lenhard B., Wells C., et al. 2005. The transcriptional landscape of the mammalian genome. *Science* **309:** 1559.

Cawley S., Pachter L., and Alexandersson M. 2003. SLAM web server for comparative gene finding and alignment. *Nucleic Acids Res.* **31:** 3507.

ENCODE Project Consortium. 2004. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306:** 636.

Guigo 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.* (suppl. 1) **7:** S2.1.

Harrow J., Denoeud F., Frankish A., Reymond A., Chen C.K., Chrast J., Lagarde J., Gilbert J.G., Storey R., Swarbreck D., et al. 2006. GENCODE: Producing a reference annotation for ENCODE. *Genome Biol.* (suppl. 1) **7:** S4.1.

Kampa D., Cheng J., Kapranov P., Yamanaka M., Brubaker S., Cawley S., Drenkow J., Piccolboni A., Bekiranov S., Helt G., et al. 2004. Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Res.* **14:** 331.

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.

Lander E.S., Linton L.M., Birren B., Nusbaum C., Zody M.C., Baldwin J., Devon K., Dewar K., Doyle M., FitzHugh W., et al. (International Human Genome Sequencing Consortium). 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860.

Parra G., Agarwal P., Abril J.F., Wiehe T., Fickett J.W., and Guigo R. 2003. Comparative gene prediction in human and mouse. *Genome Res.* **13:** 108.

Rinn J.L., Euskirchen G., Bertone P., Martone R., Luscombe N.M., Hartman S., Harrison P.M., Nelson F.K., Miller P., Gerstein M., et al. 2003. The transcriptional activity of human Chromosome 22. *Genes Dev.* **17:** 529.

Royce T.E., Rozowsky J.S., Bertone P., Samanta M., Stolc V., Weissman S., Snyder M., and Gerstein M. 2005. Issues in the analysis of oligonucleotide tiling microarrays for transcript mapping. *Trends Genet.* **21:** 466.

Rozowsky J., Newburger D., Sayward F., Wu J., Jordan G., Korbel J.O., Nagalakshmi U., Yang J., Zheng D., Guigo R., et al. 2006. The DART classification of unannotated transcription within the ENCODE regions: Associating transcription with known and novel loci. *Genome Res.* (in press).

Venter J.C., Adams M.D., Myers E.W., Li P.W., Mural R.J., Sutton G.G., Smith H.O., Yandell M., Evans C.A., Holt R.A., et al. 2001. The sequence of the human genome. *Science* **291:** 1304.