Identification of Novel Functional Elements in the Human Genome

Z. Lian,*[†] G. Euskirchen,*[‡] J. Rinn,*[¶] R. Martone,*[‡] P. Bertone,*[‡] S. Hartman,[‡]
T. Royce,[¶] K. Nelson,[‡] F. Sayward,[§] N. Luscombe,[¶] J. Yang,[§] J.-L. Li,[§] P. Miller,[§]
A.E. Urban,[‡] M. Gerstein,[¶] S. Weissman,[†] and M. Snyder[‡]

[†]Department of Genetics, [‡]Department of Molecular, Cellular and Developmental Biology,
[¶]Department of Molecular Biophysics and Biochemistry, and [§]Department of Anesthesiology,
Yale University, New Haven, Connecticut 06520

Recently, a nearly complete draft of the human genome has been determined, producing an enormous wealth of information (Olivier et al. 2001). However, the sequence by itself reveals little about the critical elements encoded in the DNA, and consequently, it is paramount to identify the functional elements encoded in the 3 billion base pairs and to determine how they work together to mediate complex processes such as development and responses to environmental alterations. Two essential tasks toward this goal are the identification of coding and transcriptionally active regions in the human genome and determining how they are regulated. The identification of these regions is an essential first step for the comprehensive and systematic analysis of gene and protein function. Thus far, a variety of different approaches have been used for identification of coding sequences and other functional elements in genomic DNA (Snyder and Gerstein 2003). Genes have been identified by generating and sequencing of cDNAs, expressed sequence tags (ESTs), and related approaches, and then mapping the mRNA coding sequences onto genomics DNA (Lander et al. 2001). Genes have also been identified by computational methods such as motif searches, identification of long open reading frames, and comparative genomic studies to identify conserved sequences, particularly those predicted to encode proteins (Lander et al. 2001; Venter et al. 2001; Waterston et al. 2002). The availability of the full genomic DNA sequence allows the direct identification of transcribed sequences by globally interrogating all regions of the genome using genomic DNA microarrays.

In addition to identification of genes, it is also of high interest to identify the elements that regulate their expression. Such information is crucial for understanding how the activity of genes is controlled and thereby what is essential for understanding cell proliferation and differentiation. Approaches to analyze gene regulation in the past have been hampered by the fact that the approaches are either not comprehensive or are indirect. For example, comparative analysis of gene expression using DNA microarrays in lines expressing or lacking a factor of interest is indirect—changes in gene expression may be due to downstream effects of the factor.

Recently, we have developed an approach for identifying the binding sites of transcription factors on a global scale (Iyer et al. 2001; Horak et al. 2002). This procedure involves immunoprecipitation of chromatin (ChIP) associated with a transcription factor of interest and using the associated DNA to probe a genomic DNA array containing the regulatory sequences or large segments of the genome. Thus, in one experiment, many binding sites for a transcription factor can be identified.

Below we describe the comprehensive analysis of transcribed regions and transcription factor-binding sites on a global scale. We have constructed an array containing most of the sequences of human Chromosome 22 and used it to identify novel transcribed regions and transcription factor-binding sites. These approaches are expected to be of broad utility for understanding the function and regulation of the human genome.

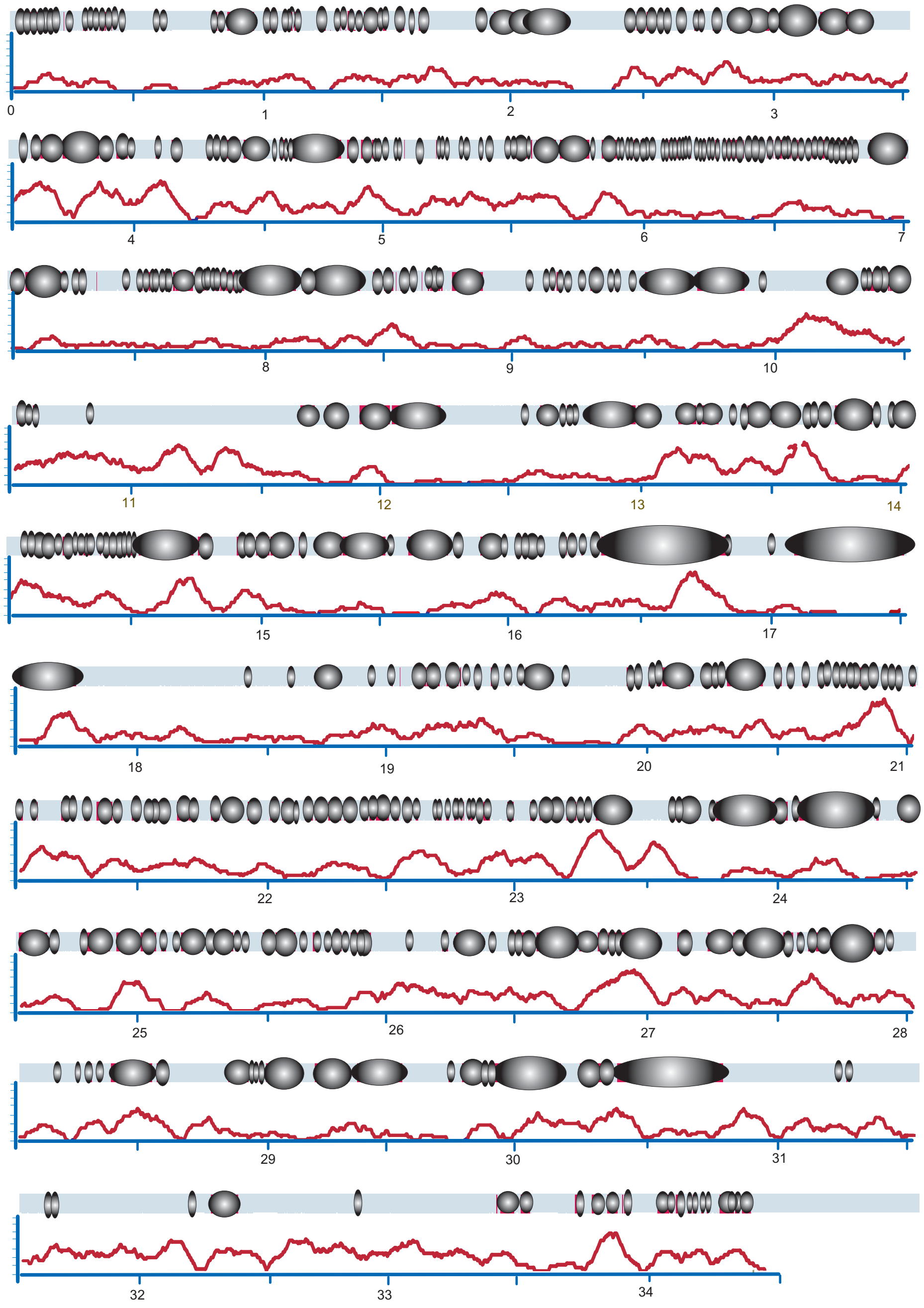
CONSTRUCTION OF A HUMAN CHROMOSOME 22 ARRAY

We have prepared an array containing nearly all of the unique sequences of human Chromosome 22 (Dunham et al. 1999; Rinn et al. 2003). Chromosome 22 contains 35.X Mbp of DNA; approximately one-half comprises repetitive DNA, and it contains 545 annotated genes (Sanger release 2.3). To prepare the array, the repetitive DNA of human chromosome was detected computationally and subtracted from the total sequence. The remaining single-copy DNA sequence was amplified in 0.3- to 1.4-kb segments (mean size 820 bp) using oligonucleotide primers and PCR; 21, 024 PCR products were attempted and 93% were successful. The DNA products were printed in duplicate onto 2.5 slides. Sequencing of several hundred products has revealed that 95% of the sequences are identical or close matches to the expected fragments.

IDENTIFICATION OF NOVEL TRANSCRIBED REGIONS

The chromosome 22 array was probed with a cDNA probe prepared from placental poly(A)$^+$ RNA (Rinn et al. 2003). The RNA had been purified three times using oligo(dT) cellulose. Labeled single-stranded cDNA was synthesized using a 50:50 mixture of oligo(dT) and ran-

*These authors contributed equally to this work.



The top gray bar represents the sequence on Chromosome 22, ordered from centromere to telomere (unit: Mega bp).

● Sanger Centre release 2.3 annotated genes in human Chromosome 22.

▬ the value of the number of positive hybridizing fragments divided by the total number of fragments in a 100-kb window.

Figure 1. The human Chromosome 22 placental transcriptome.

dom primers. Six sets of slides were probed, and 2470 fragments exhibited a significant signal in five or more replicas. The results are shown in Figure 1. Matching of the Sanger 2.3 annotation to the arrays revealed that 946 fragments corresponded to known exons; ~60% of known genes on Chromosome 22 were expressed in placental RNA. Importantly, more than half (1307 fragments) did not correspond to known exons or other annotations. These novel transcribed regions are designated TARs, for transcriptionally active regions.

To determine whether the novel TARs encoded discrete transcripts, 118 fragments were labeled and used to probe RNA blots of placental poly(A)$^+$ RNA (Rinn et al. 2003). Thirty (27%) reacted primarily with one RNA band, indi-

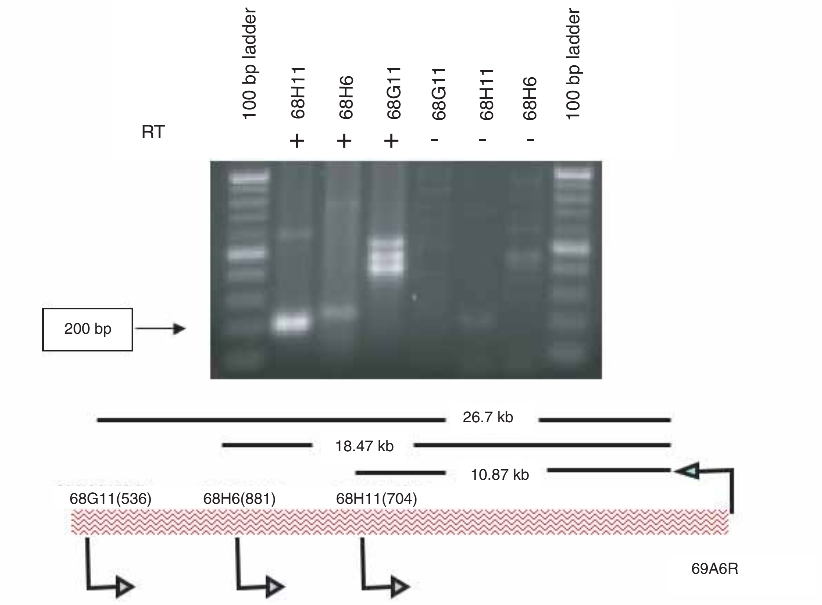


Figure 2. Characterization of a novel expressed coding fragment.

cating that they encode discrete transcripts. We further used BLAST to assure these probes could not be cross-hybridizing anywhere else in the genome. Indeed, 27 of the 30 matched only to Chromosome 22 sequence. The other 3 had weak homology with other regions in addition to a strong Chromosome 22 match. In one instance, two probes located 36 kb apart hybridized to the same 6-kb transcript, suggesting that they encoded different parts of the same message. We further investigated this region by a primer-walking experiment using a placenta cDNA library. Figure 2 demonstrates that sequential primers along this region produce increasingly larger amounts of transcript information, as evidenced by larger RT-PCR products. In each case, the increase in coding information is significantly smaller than the genomic sequence spanned by the primers, indicating the presence of a single mRNA coding region containing introns. Together the RNA blot analysis and RT-PCR primer-walking experiment indicate that many TARs encode discrete messages.

To further understand the nature of the TARs, 60-bp oligonucleotides were prepared to the regions of fragments from outside annotated genes or within introns. The sequences of the oligonucleotide were selected from the regions of the fragment that exhibited the highest score using gene prediction programs, e.g., GeneScan, Engrailed. Reverse complementary probes were also prepared. The oligonucleotides were spotted onto an array and probed with placental poly(A)$^+$ RNA. Fifty-three oligonucleotides showed a significant differential signal over the reverse complement oligo; interestingly, the proportion of signals from predicted coding versus reverse complementary oligonucleotides was identical, indicating that the gene prediction programs are not suitable for predicting novel TARs. We also found that, within introns, hybridization occurred nearly as often to the noncoding strand as to the coding strand. Thus, the transcription is unlikely to be residual unspliced messages, but rather most likely corresponds to novel transcribed regions.

We also explored the conservation of the TARs and their potential to encode protein. One third of the TARs are highly conserved with mouse sequences, indicating that they are functionally important. Approximately 8% of the hybridizing fragments with no prior annotation were found to have a homologous mouse protein. These are likely to represent novel exons associated with known annotated genes, as well as novel genes.

Recently, Karponov et al. (2002) mapped transcribed regions on human Chromosomes 21 and 22 using oligonucleotide arrays. A comparison of our data with theirs reveals that 90% of our hybridization results corresponded well with their results found for RNAs common to 6 of 11 cell lines. Thus, the independent methods each found common novel transcribed regions.

MAPPING OF NF-κB SITES ALONG CHROMOSOME 22

In addition to mapping transcribed regions, we also have mapped potential regulatory regions by identifying the binding sites of several human transcription factors along human Chromosome 22 (Martone et al. 2003; and our unpublished data). Our study was initiated by analyzing the binding-site distribution of the NF-κB family member, p65, which has been implicated in a variety of cellular responses, including inflammation and apoptosis. The number of binding sites for NF-κB along an entire chromosome was unknown and difficult to predict. Few known targets for this factor resided on this chromosome.

We analyzed the distribution of NF-κB in HeLa cells in response to TNF-α stimulation using chIP chip and our Chromosome 22 genomic DNA array (Martone et al. 2003). Briefly, HeLa cells were treated with TNF-α for

90 minutes and the cells were treated with 1% formaldehyde which crosslinks protein to DNA. The cells were then lysed and chromatin sheared to ~500–600 bp final DNA size by sonication. Anti-p65 antibodies were used to immunoprecipitate the p65-bound chromatin, the crosslinks were reversed by heating, and the DNA was purified and labeled with Cy5. As a control, cells that were not incubated with TNF-α (in which NF-κB remains in the cytoplasm) were treated in an identical fashion, and nonspecific DNA that was precipitated by the antibodies was labeled with Cy3. The labeled probes were mixed and hybridized to the Chromosome 22 array; three separate experiments were performed. Using the ExpressYourself program (Luscombe et al. 2003), we found 209 binding sites of p65 along Chromosome 22 (Fig. 3). Verification of 75 was confirmed using PCR with primers

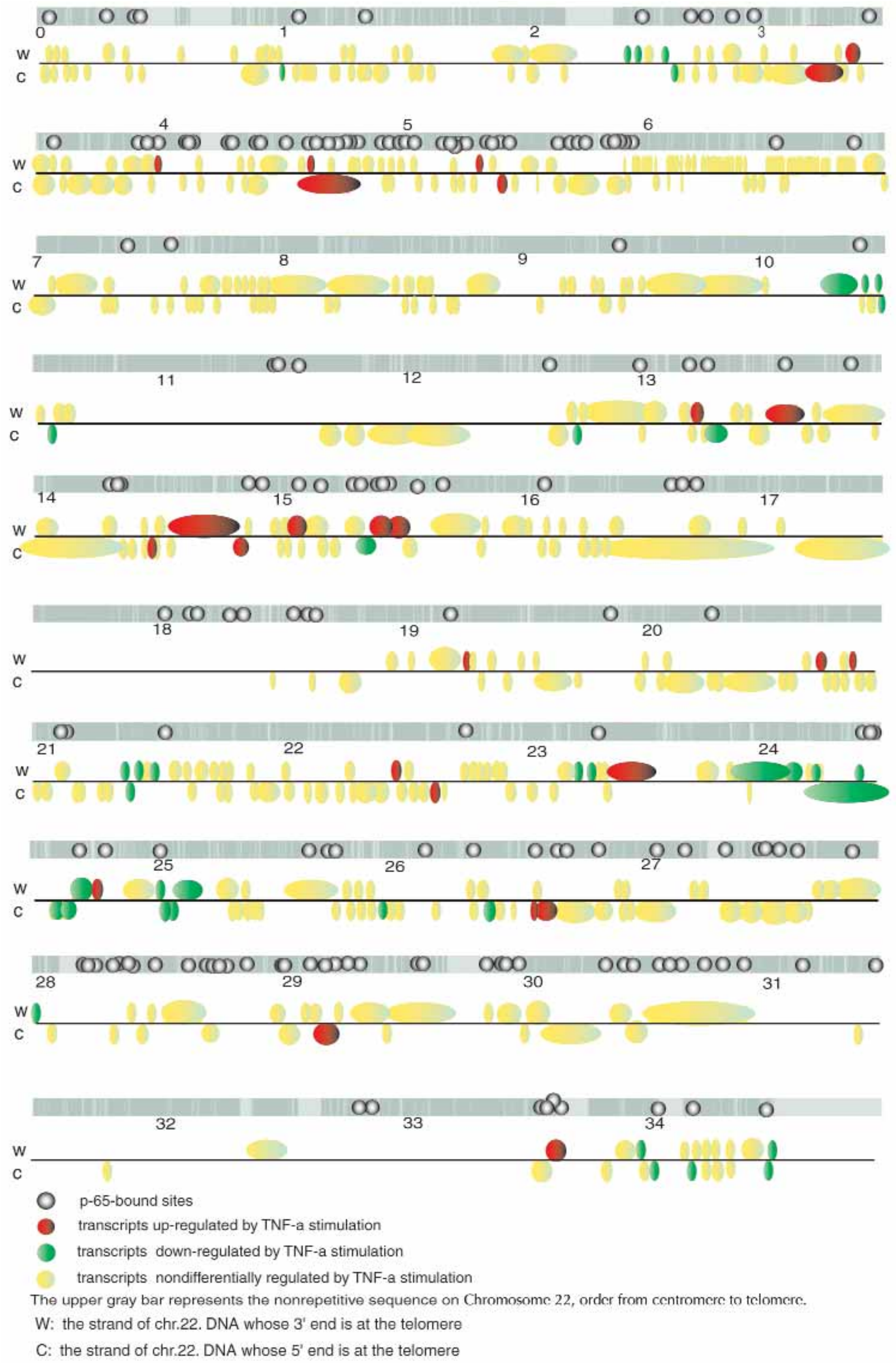


Figure 3. Chromosome 22q binding profile for p65.

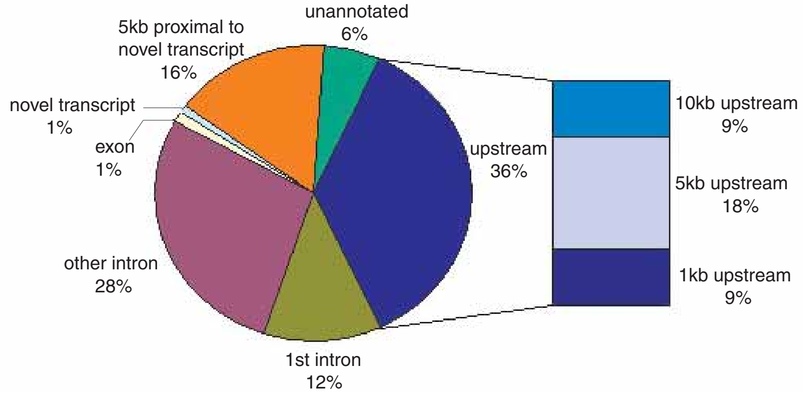


Figure 4. Distribution of p65-binding sites on Chromosome 22.

located within the hybridizing array fragment, and either gel electrophoresis and/or quantitative PCR. Approximately 80% of the sites confirm using these procedures.

The binding sites were mapped relative to known genes (Fig. 4). 77% of the sites are within 10 kb of annotated genes and 17% lie exclusively near a TAR. Only 6% do not lie near any annotated region or TAR. Interestingly, for annotated genes only a relatively small fraction (36%) of the total binding sites lie within 10 kb of the 5´ end of the gene. A large fraction lies in either the first intron (12%) or other introns (28%). Only 1% of p65-binding sites lie exclusively within an exon. The presence of NF-κB sites within introns is consistent with its initial discovery within introns (Baeuerle and Baltimore 1996). Nonetheless, these results suggest that many potential regulatory sites lie throughout a locus, not simply near the 5´ ends.

We also examined the binding distribution of p65 relative to its consensus binding site or that of a closely related factor, c-Rel, with which it is known to partner (Martone et al. 2003). 52% of the p65 binding fragments have identical matches to p65 or c-Rel consensus sites—the remainder have near matches. These results indicate that transcription factors are associated with both consensus and nonconsensus sites, indicating the importance of using experimental approaches for the detection of transcription factor-binding sites.

We also analyzed the types of genes that have p65-binding sites near or within them and found that the target genes often have biological functions that are consistent with those of p65. These include genes for PDGF, TIMP3, ATF4, EWSR1, IL2R-β, and PPAR. The binding of p65 suggests that NF-κB may be mediating some of its diverse effects through these gene targets.

Transcription factor binding does not always indicate regulation of gene expression. To correlate binding with gene expression, we examined the expression of HeLa cells upon treatment with TNF-α relative to untreated cells (Martone et al. 2003). By examining the median values of signals from exons, we found that 28 of human Chromosome 22 genes are up-regulated and 39 genes are down-regulated. Mapping of the p65-binding sites relative to expressed regions revealed that 12 lay near genes whose expression is induced by TNF-α and 6 lay near genes whose expression is repressed, indicating that TNF-α may have a previously unappreciated role as a transcriptional repressor. We also find that many p65-binding sites lie near genes whose expression is not affected by TNF-α. An example of the latter category is the λ-light chain genes; these genes may be regulated by p65 in B cells, but are not expected to be expressed in HeLa cells. These results indicate that both the gene and cellular context of binding are likely to be crucial for the regulation of gene expression; binding per se is not definitive for the regulation of gene expression. It is likely the p65 activity is modulated by functioning in concert with other transcription factors or modes of regulation (e.g., protein modification).

CONCLUSION

These results demonstrate that it is possible to map novel transcribed regions and the binding sites of transcription factors over an entire chromosome using genomic tiling arrays. Extension of these types of technologies would allow mapping over the entire human genome, allowing a comprehensive analysis of both transcribed regions and binding of regulatory factors. Given that there are ~1000–2000 transcription factors in humans (http://www.godatabase.org/dev/database/) and over 250 different cell types, it should be possible to deduce the binding sites for all factors in all possible cell types, thereby revealing the entirely transcriptional circuitry for a human being.

REFERENCES

Baeuerle P.A. and Baltimore D. 1996. NF-kappa B: Ten years after. *Cell* **87:** 13.

Dunham I., Shimizu N., Roe B.A., Chissoe S., Hunt A.R., Collins J.E., Bruskiewich R., Beare D.M., Clamp M., Smink L.J., Ainscough R., Almeida J.P., Babbage A., Bagguley C., Bailey J., Barlow K., Bates K.N., Beasley O., Bird C.P., Blakey S., Bridgeman A.M., Buck D., Burgess J., Burrill W.D., and O'Brien K.P., et al. 1999. The DNA sequence of human chromosome 22. *Nature* **402:** 489.

Horak C.E., Mahajan M.C., Luscombe N.M., Gerstein M., Weissman S.M., and Snyder M. 2002. GATA-1 binding sites mapped in the beta-globin locus by using mammalian chIp-chip analysis. *Proc. Natl. Acad. Sci.* **99:** 2924.

Iyer V.I., Horak C.A, Scafe C.S., Botstein D., Snyder M., and Brown P.O. 2001. Genomic binding distribution of the yeast cell-cycle transcription factors SBF and MBF. *Nature* **409:** 533.

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., Funke R., Gage D., Harris K., Heaford A., Howland J., Kann L., Lehoczky J., LeVine R., McEwan P., McKernan K., Meldrim J., Mesirov J.P., Miranda C., Morris W., and Naylor J., et al. (International Human Genome Sequencing Consortium). 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860.

Luscombe N.M., Royce T.E., Bertone P., Echols N., Horak C.E., Chang J.T., Snyder M., and Gerstein M. 2003. ExpressYourself: A modular platform for processing and visualizing microarray data. *Nucleic Acids Res.* **31:** 3477.

Martone R., Euskirchen G., Bertone P., Hartman S., Royce T.E., Luscombe N.M., Rinn J.L., Nelson F.K., Miller P., Gerstein M., Weissman S., and Snyder M. 2003. Distribution of NF-kappaB-binding sites across human chromosome 22. *Proc. Natl. Acad. Sci.* **100:** 12247.

Olivier M., Aggarwal A., Allen J., Almendras A.A., Bajorek E.S., Beasley E.M., Brady S.D., Bushard J.M., Bustos V.I., Chu A., Chung T.R., De Witte A., Denys M.E., Dominguez R., Fang N.Y., Foster B.D., Freudenberg R.W., Hadley D., Hamilton L.R., Jeffrey T.J., Kelly L., Lazzeroni L., Levy M.R., Lewis S.C., and Liu X., et al. 2001. A high-resolution radiation hybrid map of the human genome draft sequence. *Science* **291:** 1298.

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

Snyder M. and Gerstein M. 2003. Defining genes in the genomics era. *Science* **300:** 258.

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., Gocayne J.D., Amanatides P., Ballew R.M., Huson D.H., Wortman J.R., Zhang Q., Kodira C.D., Zheng X.H., Chen L., Skupski M., Subramanian G., Thomas P.D., Zhang J., Gabor Miklos G.L., and Nelson C., et al. 2001. The sequence of the human genome. *Science* **291:** 1304.

Waterston R.H., Lindblad-Toh K., Birney E., Rogers J., Abril J.F., Agarwal P., Agarwala R., Ainscough R., Alexandersson M., An P., Antonarakis S.E., Attwood J., Baertsch R., Bailey J., Barlow K., Beck S., Berry E., Birren B., Bloom T., Bork P., Botcherby M., Bray N., Brent M.R., Brown D.G., and Brown S.D., et al. (Mouse Genome Sequencing Consortium). 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520.