Origin of phenotypes: Genes and transcripts

Thomas R. Gingeras

Affymetrix, Inc., Santa Clara, California 95051, USA

While the concept of a gene has been helpful in defining the relationship of a portion of a genome to a phenotype, this traditional term may not be as useful as it once was. Currently, "gene" has come to refer principally to a genomic region producing a polyadenylated mRNA that encodes a protein. However, the recent emergence of a large collection of unannotated transcripts with apparently little protein coding capacity, collectively called transcripts of unknown function (TUFs), has begun to blur the physical boundaries and genomic organization of genic regions with noncoding transcripts often overlapping protein-coding genes on the same (sense) and opposite strand (antisense). Moreover, they are often located in intergenic regions, making the genic portions of the human genome an interleaved network of both annotated polyadenylated and nonpolyadenylated transcripts, including splice variants with novel 5′ ends extending hundreds of kilobases. This complex transcriptional organization and other recently observed features of genomes argue for the reconsideration of the term "gene" and suggests that transcripts may be used to define the operational unit of a genome.

New technical and conceptual insights have often prompted reconsiderations of what constitutes fundamental functional elements in a genome. In 1909, influenced by the writings of Hugo de Vries, Wilhelm Johannsen coined the term "gene" (Churchill 1974; Stamhuis et al. 1999). It was an attempt to provide a term that would represent an element that connected an inherited physical entity to an observable phenotype (Fig. 1A). Empirical findings and conceptual proposals made in the mid-20th century focused on the structural entities composing a gene. Notably, the elucidation of the structure of DNA (Watson and Crick 1953) and the subsequent unraveling of the processes of DNA replication and RNA transcription led to the identification of new elements in the genome, which, in turn, helped to sharpen an understanding of both the physical properties and definition of the term "gene" (Fig. 1B). Not long after the first description of the double helical structure of DNA, Francis Crick published a Central Dogma proposition as an operational framework describing how information stored in the sequence of DNA was transferred from the genome into functional protein products (Crick 1958). Two unstated implications from Crick's proposition emerged after publication. First, genes were viewed as discrete bounded elements, from which RNA was transcribed to carry stored information from the DNA to the cell for protein synthesis. Second, it was interpreted that the flow of information from DNA was unidirectional, with genes having the limited role of encoding protein synthesis information. Twelve years later, Crick responded to criticisms that the Central Dogma proposition was an oversimplification and further clarified his intended meaning. He restated that there was a versatile role for RNA that could allow for information to flow back to genome. Although the details were understandably sparse, Crick noted that RNA should not be considered as single purpose functional elements (Crick 1970). However, as the field of molecular genetics matured, with few notable exceptions, the functional roles for RNAs as products of genes remained focused on the production of proteins.

E-mail tom_gingeras@affymetrix.com; fax (408) 481-0422.
Article is online at http://www.genome.org/cgi/doi/10.1101/gr.6525007. Freely available online through the *Genome Research* Open Access option.

Are genes exclusively composed of protein-coding transcripts?

Efforts by subsequent generations of scientists have centered on adding greater molecular definition to the physical structure of genes and on achieving a greater understanding of how phenotypes are derived from genes. Studies aimed at defining the structure and organization of genes, characterizing the molecular structures of the RNA and protein products of genes, and determining the processes responsible for how gene expression at both the transcript and protein levels are regulated have led to many landmark discoveries. These include gene cloning, transcription factor-gene interactions, and RNA-splicing, RNA-editing, RNA-transport, and RNA-translation, to mention just a few (Fig. 1C). For the most part, these advances have taken place in the context of studying individual genes or restricted portions of genomes. Meanwhile, there was a growing realization that comprehensive answers to questions concerning the structure, function, and regulation of genes and their relationships to phenotypes required the analysis of large portions or entire genomes for most organisms.

The completion of a working draft of the human genome (Lander et al. 2001; Venter et al. 2001) provided one of the prerequisites for the development of the field of genomics. The number of genes that constitute the human genome was one of the first well-publicized genome-wide questions to be posed. While the ensuing debate may have been unnecessarily exaggerated, it seems clear that the proposed estimates were and still are based primarily on the default definition of a gene as a protein-coding functional element. As a reflection of this perception bias, protein-coding genes currently dominate the contents of most genome databases (Lander et al. 2001; Waterston et al. 2002; Zhang 2002; Parra et al. 2003).

The question of how many genes are present in the human genome led to a second query centered on the completeness of the cataloged collection of known protein-coding genes. The technical approaches used to answer this question, undertaken for not only the human genome but also for *Arabidopsis*, worm, fly, and mouse, have included in-depth full-length cDNA cloning, tiling microarrays, determination of the transcript 5′ ends

17:682–690 ©2007 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/07; www.genome.org

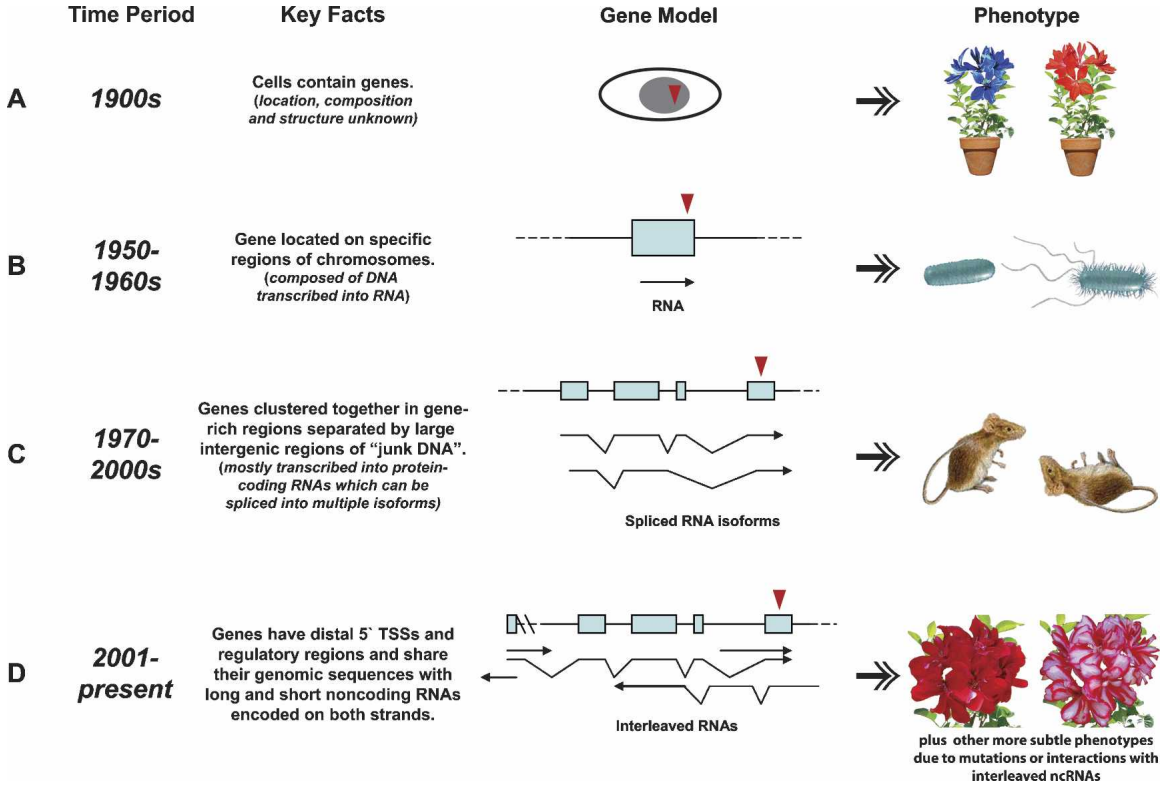

Figure 1. Evolution of the gene model and its relationship to wild-type and mutant phenotypes. Over the past century, the definition of a gene has been improved and refined from its conceptual origin in the early 1900's (*A*) with the discovery of RNA and DNA structures (*B*), splicing (*C*), and lastly, widespread unannotated transcription (*D*). Exonic regions are depicted as blue boxes with transcripts shown as arrows below (spliced and unspliced). A hypothetical mutation is shown as a red triangle. Note that as the definition of a gene grows to include multiple transcripts, a single mutation can now affect many different transcripts and thus potentially could have multiple and more subtle phenotypes.

using cap analysis of gene expression (CAGE), 3′ ends using serial analysis of gene expression (SAGE), and both 5′ and 3′ ends with gene-identification signature analysis using paired-end ditags (GIS-PET) (for review, see Johnson et al. 2005; Carninci 2006; Willingham and Gingeras 2006; Kapranov et al. 2007b). Each of these approaches has made an effort to interrogate genomes in an unbiased fashion (i.e., without regard to the knowledge of the location of previously identified protein-coding and noncoding genes). In this way, empirically based maps could be compared with the maps composed of annotated protein-coding genes. In turn, it would be possible to assess the completeness of catalogs of genes for each genome. In addition, estimates of the total number of genes for each genome could be measured, leading to the surprising observation that despite the 30-fold difference in genome size and vast differences in organismal complexity, humans have a comparable number of genes to the nematode worm (22,726 vs. 20,060 genes, respectively).

A by-product of these studies was the unanticipated, but unanimous conclusion that there was a significantly greater amount of transcriptional output from genomes than could be accounted for by our current collection of annotated protein-coding transcripts. Most of the newly identified unannotated transcripts were observed to have little protein-coding capacity (i.e., <100 amino acids) (Kapranov et al. 2002). These observations indicated that there exists a large collection of transcripts within cells that are not involved in directing protein synthesis (Figure 1D). This large collection of transcribed regions has been euphemistically been called the "dark matter" of the genome (Johnson et al. 2005) because until recently, these transcripts have escaped detection despite a considerable history of cDNA and EST cloning experiments. Although these transcripts appear to have reduced coding potential and have putatively been termed noncoding transcripts, there is no formal evidence that these transcripts do not encode short polypeptides. Thus, the term transcripts of unknown function (TUFs) has recently been suggested as their interim collective name (Cheng et al. 2005).

Prevalence of TUFs in nonprotein-coding regions of genomes

Additional confirmation of the prevalence of TUFs indicate a consistent picture of a large and until recently unannotated collection of stable cytosolic polyadenylated and nonpolyadenylated transcripts comprising approximately half of the human and mouse transcriptome. Initial analyses of the transcribed regions identified by independent technical approaches show more than half are observed by at least two different methods (Chen et al. 2002, 2004; Shiraki et al. 2003; Carninci et al. 2005, 2006; Ng et al. 2005; Ge et al. 2006).

The complexity and cellular localization of these unannotated transcripts has also proven to be unexpected. Transcriptional analysis of 10 human chromosomes demonstrates that unannotated nonpolyadenylated transcripts originating from intergenic regions of these chromosomes comprise the major proportion of the transcriptional output of the human genome (Cheng et al. 2005). In addition, nuclear and cytosolic compartmentalization of both polyadenylated and nonpolyadenylated

unannotated transcripts has been observed using tiling arrays and cDNA sequencing analyses (Cheng et al. 2005; Kiyosawa et al. 2005).

Several studies have estimated that ~10% of the nonrepeat sequences of the genome appear to be transcribed, polyadenylated, spliced in a high proportion of transcripts, and transported into the cytosol (Kapranov et al. 2002; Lian et al. 2003; Martone et al. 2003; Rinn et al. 2003; Yelin et al. 2003; Cheng et al. 2005). Considering the annotated transcripts present in RefSeq and GENCODE (Harrow et al. 2006) databases, as well as all ESTs recorded in dbEST, more than half of the detected transcribed sequences are not observed to align with these annotated transcripts (The ENCODE Project Consortium 2007; Kapranov et al. 2007a). These unannotated transcribed regions are approximately evenly distributed within and between gene boundaries.

These results were confirmed by several groups who participated in the National Human Genome Research Institute-sponsored Encyclopedia of DNA Elements (ENCODE) project, which focused its research efforts on 44 diverse regions of the human genome (~1%) to identify and characterize the functional elements present in these sequences (The ENCODE Project Consortium 2007). Analyses of the sites of transcription in these regions are presented in this special issue of *Genome Research*. Several striking observations consistent with the presence of a large representation of TUFs were made. First, it was estimated that for the nearly 400 annotated genes present in the ENCODE regions, the protein-coding loci averaged 5.4 transcripts per gene with only 1.7 potentially encoding proteins (Denoeud et al. 2007; The ENCODE Project Consortium 2007). Second, >65% of these genes possess 5′ distal (108,000 bp on average) previously unannotated, tissue-specific transcription start sites (TSS) and promoter regions, many of which are parts of TUFs (Denoeud et al. 2007). Third, large numbers of protein-coding genes in these regions have isoforms that are composed of exons located in genomic nonprotein-coding regions (introns and intergenic regions) (Rozowsky et al. 2007). Fourth, analysis of transcribed unannotated ENCODE regions reveal the potential to fold into stable RNA structures (Washietl et al. 2007). Fifth, a compilation of all previously annotated and empirically detected RNAs found in the ENCODE studies indicates that to produce these RNAs, >90% of genomic sequence appears to be transcribed as nuclear primary transcripts (The ENCODE Project Consortium 2007).

The existence of this additional layer of transcriptional complexity has prompted several questions concerning: (1) the likelihood of the functional significance of widespread transcription; (2) the relationship of TUFs to protein-coding transcripts; and (3) their regulation, structure, and genomic organization. While answers to some of these questions are emerging, studies focused on noncoding transcripts of known biological function have begun to reveal a complexity in genome organization not captured by the current collection of annotations, prompting a reconsideration of what constitutes the fundamental functional element of the genome and how it relates to phenotypic variation.

Well-characterized noncoding transcripts of known function

Well-characterized noncoding transcripts with known functions include ribosomal (r)RNAs, transfer (t)RNAs, small nuclear (sn)RNAs, small nucleolar (sno)RNAs, as well as small RNA components of RNase P and other protein complexes (for review, see Eddy 2001; Storz 2002; Prasanth and Spector 2007). Another class of noncoding RNAs includes microRNAs (miRNAs) and exogenous small interfering RNAs (siRNAs), both of which participate in the RNA interference pathway (RNAi) and have regulatory functions at transcriptional and post-transcriptional levels. Several of the structural and regulatory features of these known nonprotein-coding RNAs are notable and can be used as characteristics of functional transcripts.

snoRNAs

The first notable feature is the range of lengths of snoRNA transcripts, varying from 60 to 300 nucleotides (nt) in length. This variability in transcript size likely suggests either that the flexibility in the length of transcript sequences is required to carry out similar functions, or this class of noncoding RNAs may have multiple functions (see below). Second, these stable transcripts carry out their modifications of rRNAs in association with a set of proteins to form a collection of small nucleolar particles (Bachellerie et al. 2002). This association with specialized proteins to carry out their function is also shared by many other protein-coding and noncoding transcripts and the specificity conferred via this is instructive for how, throughout the cell, noncoding transcripts appear to provide a context-specific function to a common set of protein factors. Third, snoRNA transcripts in higher eukaryotes are processed from introns of mRNAs, thus serving as one of the first examples of the functional importance of intronic portions of preprocessed and blurring the boundaries of gene organization. Fourth, computational studies of the *Saccharomyces cerevisiae* genome have identified many novel methylation-guide snoRNAs that are involved in rRNA modification (Lowe and Eddy 1999; Schattner et al. 2004), indicating that although this is a well-established functional class of noncoding transcripts, the membership of this class is still growing. Finally, recent studies indicate that a number of snoRNA transcripts do not possess sequences that are fully complementary to rRNA targets (Jady and Kiss 2000; Li et al. 2005), which not only presents a challenge in identifying these targets, but also suggests that a larger network of cellular proteins and/or other transcripts outside of the rRNA complex may be required to assist snoRNAs in carrying out their functions. This later finding opens the possibility that snoRNAs may have functions other than modification of rRNAs and spliceosomal RNAs. One such function, regulation of alternative splicing of a transcript encoded in *trans,* has recently been demonstrated for one snoRNA, HBII-52 (Kishore and Stamm 2006).

RNA interference (RNAi): miRNAs and siRNAs

Both miRNAs and siRNAs have been shown to be sequence-specific transcriptional and post-transcriptional regulators of gene expression (Doench et al. 2003; Bartel 2004; Meister and Tuschl 2004; Zamore and Haley 2005; Kim and Nam 2006). These two classes of noncoding transcripts also possess many distinguishing characteristics that are essential for their biological functions, and as such, may exemplify common characteristics shared by newly identified noncoding transcripts.

RNAi noncoding transcripts operate as double-stranded RNA molecules, with each strand being ~21–23 nt in length in their ultimately functional forms. It is known that both types of RNAi molecules are produced from relatively long pri(mary)-transcripts by RNase III classes of endoribonucleases. The miRNAs are first processed in the nucleus by RNASEN (formerly

DROSHA). Following transport out of the nuclear compartment, DICER1, a dsRNA-specific endonuclease, processes the 70-mer pre-transcripts into the biologically active double stranded 21–23-mers. However, with the exception of a few cases, relatively little is known about the primary transcripts that give rise to the 70-mer precursors (pre-) of miRNAs or to siRNAs. The fully processed siRNAs and miRNAs are incorporated into the RNA-induced silencing complexes (RISC), which target specific mRNA transcripts to interfere with target RNA stability or translation (Nelson et al. 2003; Bartel 2004; Cullen 2004; Lee et al. 2004; Tijsterman and Plasterk 2004; Rivas et al. 2005; Zamore and Haley 2005; Kim and Nam 2006).

These two classes of nc-RNAi transcripts also possess several characteristics that are similar to those previously described for snoRNA transcripts, including: (1) both classes of RNAs are produced from much larger precursor RNA molecules; (2) the genomic location of pri-RNA transcripts often mapping to genomic sites previously considered less biologically relevant (i.e., intergenic and intronic regions); (3) the association of primary, precursor, and mature miRNA and siRNAs with specific protein complexes to achieve biological functionality; (4) a single RNAi or snoRNA has the ability to regulate multiple transcripts in *trans* using partial sequence complementarity; and (5) the likelihood that the current catalog of RNAi transcripts are significantly underestimated (Lewis et al. 2003, 2005; Krek et al. 2005; Kishore and Stamm 2006).

Other characterized functional noncoding RNAs

In addition to the short RNA species discussed above, there is a growing number of other noncoding RNAs with established or likely biological functions (for review, see Mattick and Makunin 2006; Willingham and Gingeras 2006; Prasanth and Spector 2007). These RNAs can range in length from 21 to 30 nt (e.g., 21U RNAs and piRNAs) through hundreds of nucleotides (e.g., 330 bp for 7SK snRNA) to the 100 kb (e.g., 108 kb for *Air* RNA) (Prasanth and Spector 2007). Furthermore, functional noncoding RNAs have been shown to act via protein (e.g., NRON) (Willingham et al. 2005), RNA (e.g., some natural antisense transcripts) (Wahlestedt 2006), DNA (e.g., *Xist*) (Avner and Heard 2001), or combinations of both types of interactions (e.g., the promoter-specific noncoding RNA of the *DHFR* gene that interacts with promoter DNA as well as components of the core transcriptional machinery (Martianov et al. 2007).

Thus, the characteristics observed to be part of the regulation, structure, and genomic organization of well-characterized noncoding transcripts of known function (e.g., snoRNAs, miRNAs, siRNAs, and others) represent potential hallmarks, several of which are shared by TUFs, which could be used to help to identify other classes of functional noncoding transcripts.

Transcripts of unknown function (TUFs)

TUFs identified from analysis of cytosolic polyadenylated RNAs appear to share at least four characteristics with RNAi and snoRNA transcripts. The first of these shared characteristics is that some of these unannotated transcripts appear to be part of a regulatory system for protein-coding gene expression. Several groups have shown that *cis*-encoded unannotated antisense transcripts on a wide genomic scale are found to be simultaneously expressed with their paired sense transcript (Cawley et al. 2004; Katayama et al. 2005; Kiyosawa et al. 2005). This expression is observed to be either coordinately or discordantly regulated with the sense transcript; therefore, antisense transcripts cannot be assumed to have a simple antagonistic RNAi-mediated influence on the complementary transcript. However, when compared with genes without antisense transcripts, antisense transcript pairs are considerably more likely to have this genomic organization evolutionarily preserved, suggesting that some functional relationship is being retained (Dahary et al. 2005). On an individual gene level, the antisense regulation of *MYCN* (Krystal et al. 1990), *HIF1A* (Thrash-Bingham and Tartof 1999), and *IME4* (Hongay et al. 2006) may point the way to how some of the antisense transcripts may carry out the regulation of their cognate sense genes. In yeast, entry into meiosis is controlled by *IME4* and its regulation by an antisense transcript through what appears to be a mechanism of transcription interference. Diploid cells with *IME4* antisense transcription have reduced sense transcripts and do not enter meiosis. Furthermore, human diseases ranging from breast cancer and lymphoma to thalassemia have been linked to naturally occurring antisense transcripts (for review, see Wahlestedt 2006).

The second shared characteristic is regulation of the TUFs by independent promoter elements not necessarily associated with the regulation of protein-coding genes. The majority of binding of MYC and SP1 to chromosomes 21 and 22 and of CREB1 to chromosome 21 were located in introns, exons, and intergenic regions (Cawley et al. 2004; Euskirchen et al. 2004). Many of these sites contained evidence of unannotated transcription in close proximity. The large-scale sequencing of more than 12 million CAGE tags from multiple mouse and human tissues permitted the genome-wide mapping of transcriptional start sites (TSSs) (Carninci et al. 2006). Widespread unannotated transcription was supported by an abundance of intergenic TSSs. Furthermore, the significant appearance of TSSs within internal exons and 3′ UTRs of annotated genes suggests multiple overlapping transcripts for many known genes.

The third shared characteristic is that the genomic locations encoding these TUFs correspond to regions thought to be biologically less important (introns and intergenic regions). Bertone et al. (2004) noted that 38% of their detected transcriptionally active regions (TARs) found while interrogating the entire human genome using tiling arrays were located more than 10 Kb from any previously annotated gene. Schadt et al. (2004) for human chromosomes 20 and 22 and Cheng et al. (2005) for 10 human chromosomes also reported that ~25% of the oligonucleotide probes on their respective microarrays detected evidence of transcription emanating from intergenic regions.

Of note, the maps of transcribed sequences created using microarrays are very conservative. The thresholds used to determine whether a hybridizing signal is background or real signal have been set to select for the highest 2%–10% of the possible probes. Since the estimated copy number of many of the detected TUFs is low (estimated to be between less than one and 10 copies per cell), most of these transcripts are not reported because of the possibility of increasing the amount of false positive calls. In addition, only a relatively small number of differentiated and undifferentiated mammalian cell types/tissues have been analyzed by each of the laboratories using the five methodological approaches mentioned previously (cDNA cloning, microarrays, CAGE, SAGE, PET ditags). In-depth analysis of the full range of cell types found in mammals is likely to reveal additional members for each of the TUF categories (see below). Therefore, a fourth characteristic shared with noncoding RNAs of known

function is that the transcript membership of each of the general categories of TUFs is undoubtedly underestimated.

Potential categories of TUFs

Three general organizational categories for the observed unannotated transcribed sequences can be identified. These categories are defined based on the relationship of TUFs to the structure and organization of the protein-coding transcripts. The first category consists of those TUFs that are complementary to sense transcripts. Relative to the sense transcript, these antisense transcripts can occur in *cis* (transcripts that overlap sense transcripts and for at least some portion of their length are completely complementary to exonic and/or intronic portions of sense transcripts) and *trans* (transcripts that are synthesized at a genomic site distal from the sense-transcribed region and may be only partially complementary to the sense transcript) (Kumar and Carmichael 1998; Vanhee-Brossollet and Vaquero 1998). The prominent presence of antisense transcripts in the genome has only recently been appreciated. Computational analyses of cDNA databases have estimated that from 8% (Shendure and Church 2002; Yelin et al. 2003) to 20% (Chen et al. 2004) of well-characterized coding genes have at least one overlapping antisense transcript. Empirical estimates have increased this estimate to >50% (Cheng et al. 2005), with the majority being unannotated transcripts. A comprehensive analysis of large cDNA, CAGE, and PET ditag libraries report similar occurrences of antisense transcription with as high as 72% of all annotated transcription units having an antisense transcript (Kiyosawa et al. 2003, 2005; Katayama et al. 2005).

The second category of unannotated transcribed sequences corresponds to isoforms of well-characterized protein-coding transcripts. Using a combination of techniques including microarray analysis, rapid amplification of cDNA ends (RACE), RT–PCR, and sequencing of isolated c-DNA clones, Kapranov et al. (2005) have noted that novel isoforms have been identified for almost every well-characterized protein-coding transcript examined. These experiments were later greatly expanded to include all annotated genes within the boundaries of the 1% of human genomes represented by the ENCODE regions (Denoeud et al. 2007; The ENCODE Project Consortium 2007). Strikingly, 90% of the 399 genes have either a previously unannotated exon or a new TSS (Denoeud et al. 2007; The ENCODE Project Consortium 2007). These novel isoforms include extended or shortened annotated exons as well as new exons. In fact, a combination of tiling arrays and RT–PCR/RACE experiments revealed that many human and *Drosophila* genes have extensive previously unannotated 5′ exons that are often noncoding UTRs. In *Drosophila*, the average size of newly predicted first introns was found to be >10-fold larger than estimated from RefSeq annotations (Manak et al. 2006), whereas in human ENCODE regions, new first introns averaged 108 kb with 23% of new introns >200 kb (The ENCODE Project Consortium 2007).

Expressed pseudogenes are a special version of this second category and may also contribute to the pool of unannotated transcribed sequences. While a pseudogene may have lost its ability to code for a functional protein, it may still be transcribed. An estimated 20,000 processed and unprocessed pseudogenes are present in the human genome (Torrents et al. 2003). However, this is likely to be an underestimate, since these analyses underrepresent evolutionary older and smaller pseudogenes. A recent revision of the state of the human genome sequence estimates that there will be more pseudogenes than functional protein-coding genes in the human genome (International Human Genome Sequencing Consortium 2004). Pseudogene transcripts have previously been shown to be functional by assisting to regulate the protein-coding mRNA stability and/or translation of their homologous coding genes (Hatfield et al. 2002; Zhang et al. 2002; Hirotsune et al. 2003; Yano et al. 2004). These findings demonstrate that expressed pseudogenes may be associated with specific regulatory role(s), and further highlight the potential functional significance of some of the unannotated transcripts. Approximately 10%–14% of the array-detected unannotated transcribed sequences found expressed in 10 human chromosomes may map to pseudogene loci (Cheng et al. 2005). Consistent with these results, the ENCODE Consortium, using a variety of experimental techniques, conservatively estimated that 19% of pseudogenes located within the ENCODE regions are transcribed (The ENCODE Project Consortium 2007; Zheng et al. 2007).

The third category consists of transcripts that either overlap intron regions of well-characterized annotated gene transcripts (on the same strand) or are entirely found within intergenic regions. Analysis of the structure and organization of TUFs using microarrays, RACE, and cloning/sequencing methods indicated that ~10% of the interrogated unannotated polyadenylated cytosolic TUFs were found to be located entirely in the intergenic regions, while another 10% of TUFs were found to be entirely included in the intronic regions of annotated protein-coding transcripts (Cheng et al. 2005). These transcripts often appear to be located near genomic regions that bind an assortment of transcription factors and contain localized histone modifications that alter the chromatin structure in a manner conducive for active transcription (The ENCODE Project Consortium 2004, 2007; Kapranov et al. 2007a).

Evolutionary conservation of TUFs

Overall, while ~5% of the human and mouse genomes appear to be under purifying evolutionary selection, and ~60% of these genomic regions occur outside the boundaries of the well-annotated exons, sequences detected as being part of unannotated transcribed sequences align to only a small percent of these conserved regions (The ENCODE Project Consortium 2007). Kampa et al. (2004) and Bertone et al. (2004) report that ~20%–24% of the unannotated transfrag and TAR sequences have substantial BLAST alignments with the mouse genome. Thus, the majority of detected unannotated transcribed sequence appears not to be strongly conserved relative to the mouse genome.

This characteristic of reduced evolutionary conservation makes TARs and TUFs unattractive in both being functionally important and being categorized as genes under traditional criteria (Snyder and Gerstein 2003). However, given the stated bias toward protein-coding transcripts in the formation of these criteria, it may prove premature to reach such conclusions. First, additional analyses are needed to address whether there is evolutionary conservation not detected using these traditional analysis approaches. One interesting possibility is that these unannotated transcribed sequences exhibit more recent evolutionary change, and thus may be more related to the primate limb of the mammalian lineages. Indeed, a search for sequences most rapidly evolving in the human lineage identified a noncoding RNA with brain-specific expression patterns (Pollard et al. 2006). Second, the types of sequence conservation observed for protein-

coding transcripts and for mature miRNA molecules may not be observed in either precursors to these short RNAs or other mature functional noncoding transcripts (e.g., NRSE dsRNA) (Kuwabara et al. 2004). Furthermore, the large noncoding RNA, *XIST*, is essential for sex chromosome dosage compensation in mammals, and yet exhibits rapid evolution of primary sequences despite an overall conservation of gene structure and organization (Nesterova et al. 2001). Third, noncoding transcripts may adopt secondary structures essential for their function, and these structures may permit certain latitude in primary sequence composition. Computational analysis of RNA structural conservation based on base pairing and thermodynamic stability identified more than 30,000 RNA elements across the human genome, with approximately half mapping outside of known genes (Washietl et al. 2005). Focusing on the approximate third of the human genome not alignable with mouse, a significant number of these nonconserved regions were found to have signatures of RNA structure and impressively were twice as likely to overlap a tiling array-detected transfrag (Torarinsson et al. 2006). Lastly, the general lack of evolutionary conservation for TUFs may be explained if the TUFs represent larger precursor transcripts that are post-transcriptionally processed to produce short RNAs, which themselves do have a higher degree of conservation, as noted by Ponjavic et al. (2007). For example, mature miRNA sequences can be quite conserved across the animal kingdom, and yet their longer precursor sequences often lack significant conservation. Indeed, this observation has recently been extended to a large number of entirely new classes of novel short RNAs that are overlapped by nuclear TUFs, raising the possibility that some distinct proportion of unannotated nuclear transcription could serve as precursors for short RNA species (Kapranov et al. 2007a).

A collective network of transcripts and other regulatory elements result in a phenotype

The finding that noncoding transcripts are an expanding class of biologically important molecules has been discussed by many authors (for reviews, see Eddy 2001, 2002; Mattick 2001, 2004, 2005; Mattick and Gagen 2001; Huttenhofer et al. 2002, 2005; Szymanski and Barciszewski 2002; Morey and Avner 2004). However, it is recognized that not all of the newly discovered transcripts are likely to be biologically important. Thus, additional independent empirical evidence is required to support their biological relevance. Traditionally, support for functionality is derived from genetic and biochemical experiments that demonstrate a measurable phenotype associated with the investigated RNAs. These experiments, however, require that a measurable phenotype be observable. This has not always been straightforward even for protein-coding genes. For 96% of the open reading frames in yeast mutated by gene deletions and assayed under six growth conditions, <7% were required for growth (Giaever et al. 2002). Similarly, 8.9% of the predicted genes in worm have a detectable phenotype after RNAi inhibition (Kamath et al. 2003). Thus, phenotypic responses for the newly identified TUFs will likely be challenging and certainly time consuming, as it has been for the vast majority of protein-coding transcripts.

It is likely that some of the TUFs and noncoding RNAs that have recently been identified will be members of the already identified classes of functional noncoding transcripts such as RNAi and snoRNAs. Yet, other TUFs and noncoding RNA transcripts will likely be involved in additional biological processes for which RNAs have been shown to be important components, such as genomic imprinting (Sleutels et al. 2002; Takada et al. 2002), regulation of transcription, DNA replication, RNA stability, processing, and translation (Storz 2002; Willingham and Gingeras 2006; Prasanth and Spector 2007). Some TUFs may simply be products of transcription and regulatory processes, and the RNAs themselves have little or no direct inherent functional value with the biological important function residing in the transcriptional process itself. Finally, since RNA is well suited to the recognition of other nucleic acids by base pairing and to interacting with cellular protein components by virtue of its folding capabilities, some of the identified TUFs and noncoding RNAs are likely to be involved in processes not currently associated with RNA transcripts.

It has been proposed that this additional layer of complexity embodied by the intricate network of noncoding transcripts within a cell provides two important higher order functional capabilities to genomes (Mattick 2001, 2004, 2005; Szymanski et al. 2003). The first functionality provides a means to increase the informational and operative capabilities of genomes, while the number of protein-coding genes remains relatively similar across evolutionary distances. Protein diversity can be substantially increased using multiple splice isoforms as well as using chimeric gene fusions (discussed in Kapranov et al. 2007b). Indeed, such "tandem-chimerism" and gene fusion has been proposed as a common cellular mechanism for increasing protein diversity (Akiva et al. 2006; Parra et al. 2006). The second functionality is to contribute to RNA-based mechanisms (discussed above) that carry out many of the regulatory processes required for the increased capabilities of higher organisms and to communicate the status of these regulated processes.

As described above, several classes of noncoding transcripts not only physically interact with protein-coding transcripts and their protein products, but are also organizationally embedded within or proximal to protein-coding transcripts (Fig. 2). This has served not only to blur the physical boundaries of genes, but also to increase the complexity of determining what sequences in a gene serve what functions. The abundant presence of *cis*-antisense transcripts, for example, allows for the same nucleotides present in a protein-coding transcript to be part of a noncoding transcript, which, in turn, may play a role in the regulation of the same (or another) protein-coding transcript. The recently reported whole-genome transcript mapping study of both long and short RNAs and their inter-relationship lends strong support to a model of gene organization that is decidedly not colinear (Kapranov et al. 2007a). Hundreds of thousands of new short RNA species were discovered and a significant class of promoter-associated short RNAs were found to correlate with expression of the associated long mRNAs (Kapranov et al. 2007a). Thus, in light of this overlapping interleaved network of protein-coding and noncoding transcripts, it seems appropriate to reconsider the concept of gene in describing the relationship of a portion of a genome to a phenotype.

What is a gene and are transcripts fundamental operational units?

The current definition of gene (as defined by HUGO's Human Genome Nomenclature Committee) is a DNA segment that contributes to phenotype/function, and in the absence of demonstrated phenotype/function, a gene may be characterized by se-

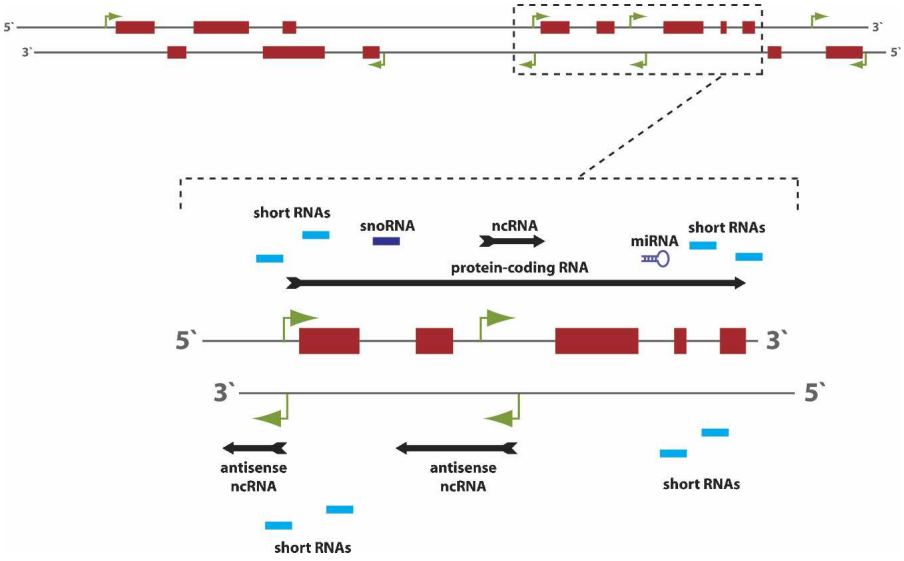

Figure 2. Transcriptional complexity of a gene. Hypothetical gene cluster with detailed zoom-in for highlighted gene demonstrates that a single gene can have multiple transcriptional start sites (TSSs) as well as many interleaved coding and noncoding transcripts. Exons are shown as red boxes and TSSs are green right-angled arrows. Known short RNAs such as snoRNAs and miRNAs can be processed from intronic sequences and novel species of short RNAs that cluster around the beginning and ends of genes have recently been discovered (see text).

quence, transcription, or homology (Wain et al. 2002). Accordingly, this definition would arguably include the DNA regions that regulate the "contribution" leading to the phenotype/function. Inclusion of regulatory regions along with the entire transcribed regions (intronic and exonic) is appropriate given that the levels of transcription and the efficiency of transcript processing (both examples of a contribution) often influence the phenotypes/functions. Proximal and distal regulatory elements such as promoters, enhancers, and insulators would therefore be considered parts of gene under such a definition. Thus, defining the functional components for any gene could include many clustered and dispersed portions of a genome. Additionally, multiple transcripts utilizing the same sequence space on the same and opposite strands often each controlled by their own distinct regulatory regions and that may extend the boundaries of protein-coding transcripts all together further complicates the concept of relating a DNA region with a corresponding phenotype/function (Fig. 2). If each of the transcripts sharing sequence space with a protein-coding gene are capable of effecting the same phenotype/function, then a gene can consist of multiple (coding and noncoding) transcripts and regulatory regions (Fig. 1D). This increased complexity of both the components of a gene and its boundaries begs for a simpler operational unit that can be used to link a specific DNA sequence to phenotype/function. Individual RNA transcripts provide these fundamental operational elements.

The consideration of the use of transcripts as a fundamental operational element in describing the linkage of discrete genomic sequences to specific phenotypes/function allows for the straightforward cataloging and the identification of singular or multiple RNAs that influence the same phenotype and the separation of the operational components that contribute to phenotypes/function from other genomic elements that directly contribute to phenotypes/functions, but whose influences may be subtle and/or whose location may be very distal from the site of transcription.

Clearly, our understanding of the complexity of how information in genomes is organized, regulated, and expressed has grown in recent years. The identification of an abundant collection of polyadenylated and nonpolyadenylated transcripts with highly reduced protein-coding potential, which are found in the many cell types from many organisms, together with the elucidation of the complex relationship of these transcripts to the protein-coding transcripts exemplifies this increased complexity. Correspondingly, if the biological relevance of the bulk of these novel transcripts continues to be confirmed by subsequent experiments, this increased complexity most certainly will necessitate a reconsideration of the definition of a gene and require the use of an alternative term to help to define the fundamental operational unit that relates genomic sequences to phenotypes/function.

Acknowledgments

A special thanks to A. Willingham who updated and assisted in restructuring this manuscript from an earlier written version and for the generation of the figures, as well as to K. Kong, P. Kapranov, and R. Duttagupta for helpful literature, organization editing suggestions, and helpful discussions. This work has been funded in part with Federal Funds from the National Cancer Institute, National Institutes of Health under Contract No. N01-CO-12400, the National Human Genome Research Institute under Grant number U01 HG003147, and from Affymetrix, Inc.

References

Akiva, P., Toporik, A., Edelheit, S., Peretz, Y., Diber, A., Shemesh, R., Novik, A., and Sorek, R. 2006. Transcription-mediated gene fusion in the human genome. *Genome Res.* **16:** 30–36.

Avner, P. and Heard, E. 2001. X-chromosome inactivation: Counting, choice and initiation. *Nat. Rev. Genet.* **2:** 59–67.

Bachellerie, J.P., Cavaille, J., and Huttenhofer, A. 2002. The expanding snoRNA world. *Biochimie* **84:** 775–790.

Bartel, D.P. 2004. MicroRNAs: Genomics, biogenesis, mechanism, and function. *Cell* **116:** 281–297.

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:** 2242–2246.

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. 2005. The transcriptional landscape of the mammalian genome. *Science* **309:** 1559–1563.

Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., Semple, C.A., Taylor, M.S., Engstrom, P.G., Frith, M.C., et al. 2006. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.* **38:** 626–635.

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.

Chen, J., Sun, M., Lee, S., Zhou, G., Rowley, J.D., and Wang, S.M. 2002.

Identifying novel transcripts and novel genes in the human genome by using novel SAGE tags. *Proc. Natl. Acad. Sci.* **99:** 12257–12262.
Chen, J., Sun, M., Kent, W.J., Huang, X., Xie, H., Wang, W., Zhou, G., Shi, R.Z., and Rowley, J.D. 2004. Over 20% of human transcripts might form sense-antisense pairs. *Nucleic Acids Res.* **32:** 4812–4820.
Cheng, J., Kapranov, P., Drenkow, J., Dike, S., Brubaker, S., Patel, S., Long, J., Stern, D., Tammana, H., Helt, G., et al. 2005. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* **308:** 1149–1154.
Churchill, F.B. 1974. William Johannsen and the genotype concept. *J. Hist. Biol.* **7:** 5–30.
Crick, F. 1970. Central dogma of molecular biology. *Nature* **227:** 561–563.
Crick, F.H. 1958. On protein synthesis. *Symp. Soc. Exp. Biol.* **12:** 138–163.
Cullen, B.R. 2004. Derivation and function of small interfering RNAs and microRNAs. *Virus Res.* **102:** 3–9.
Dahary, D., Elroy-Stein, O., and Sorek, R. 2005. Naturally occurring antisense: Transcriptional leakage or real overlap? *Genome Res.* **15:** 364–368.
Denoeud, F., Kapranov, P., Ucla, C., Frankish, A., Castelo, R., Drenkow, J., Lagarde, J., Alioto, T., Manzano, C., Chrast, J., et al. 2007. Prominent use of distal 5′ transcription start sites and discovery of a large number of additional exons in ENCODE regions. *Genome Res.* (this issue) doi: 10.1101/gr.5660607.
Doench, J.G., Petersen, C.P., and Sharp, P.A. 2003. siRNAs can function as miRNAs. *Genes & Dev.* **17:** 438–442.
Eddy, S.R. 2001. Non-coding RNA genes and the modern RNA world. *Nat. Rev. Genet.* **2:** 919–929.
Eddy, S.R. 2002. Computational genomics of noncoding RNA genes. *Cell* **109:** 137–140.
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* (in press).
Euskirchen, G., Royce, T.E., Bertone, P., Martone, R., Rinn, J.L., Nelson, F.K., Sayward, F., Luscombe, N.M., Miller, P., Gerstein, M., et al. 2004. CREB binds to multiple loci on human chromosome 22. *Mol. Cell. Biol.* **24:** 3804–3814.
Ge, X., Wu, Q., Jung, Y.C., Chen, J., and Wang, S.M. 2006. A large quantity of novel human antisense transcripts detected by LongSAGE. *Bioinformatics* **22:** 2475–2479.
Giaever, G., Chu, A.M., Ni, L., Connelly, C., Riles, L., Veronneau, S., Dow, S., Lucau-Danila, A., Anderson, K., Andre, B., et al. 2002. Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418:** 387–391.
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.* **7 (Suppl.):** S4.1–S4.9.
Hatfield, J.T., Rothnagel, J.A., and Smith, R. 2002. Characterization of the mouse hnRNP A2/B1/B0 gene and identification of processed pseudogenes. *Gene* **295:** 33–42.
Hirotsune, S., Yoshida, N., Chen, A., Garrett, L., Sugiyama, F., Takahashi, S., Yagami, K., Wynshaw-Boris, A., and Yoshiki, A. 2003. An expressed pseudogene regulates the messenger-RNA stability of its homologous coding gene. *Nature* **423:** 91–96.
Hongay, C.F., Grisafi, P.L., Galitski, T., and Fink, G.R. 2006. Antisense transcription controls cell fate in *Saccharomyces cerevisiae*. *Cell* **127:** 735–745.
Huttenhofer, A., Brosius, J., and Bachellerie, J.P. 2002. RNomics: Identification and function of small, non-messenger RNAs. *Curr. Opin. Chem. Biol.* **6:** 835–843.
Huttenhofer, A., Schattner, P., and Polacek, N. 2005. Non-coding RNAs: Hope or hype? *Trends Genet.* **21:** 289–297.
International-Human-Genome-Sequencing-Consortium. 2004. Finishing the euchromatic sequence of the human genome. *Nature* **431:** 931–945.
Jady, B.E. and Kiss, T. 2000. Characterisation of the U83 and U84 small nucleolar RNAs: Two novel 2′-O-ribose methylation guide RNAs that lack complementarities to ribosomal RNAs. *Nucleic Acids Res.* **28:** 1348–1354.
Johnson, J.M., Edwards, S., Shoemaker, D., and Schadt, E.E. 2005. Dark matter in the genome: Evidence of widespread transcription detected by microarray tiling experiments. *Trends Genet.* **21:** 93–102.
Kamath, R.S., Fraser, A.G., Dong, Y., Poulin, G., Durbin, R., Gotta, M., Kanapin, A., Le Bot, N., Moreno, S., Sohrmann, M., et al. 2003. Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. *Nature* **421:** 231–237.
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:** 331–342.
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.
Kapranov, P., Drenkow, J., Cheng, J., Long, J., Helt, G., Dike, S., and Gingeras, T.R. 2005. Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays. *Genome Res.* **15:** 987–997.
Kapranov, P., Cheng, J., Dike, S., Nix, D.A., Duttagupta, R., Willingham, A.T., Stadler, P.F., Hertel, J., Hackermueller, J., Hofacker, I.L., et al. 2007a. Genome-wide RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* (in press).
Kapranov, P., Willingham, A.T., and Gingeras, T.R. 2007b. Genome-wide transcription and the implications on genome organization. *Nat. Rev. Genet.* **8:** 1–11.
Katayama, S., Tomaru, Y., Kasukawa, T., Waki, K., Nakanishi, M., Nakamura, M., Nishida, H., Yap, C.C., Suzuki, M., Kawai, J., et al. 2005. Antisense transcription in the mammalian transcriptome. *Science* **309:** 1564–1566.
Kim, V.N. and Nam, J.W. 2006. Genomics of microRNA. *Trends Genet.* **22:** 165–173.
Kishore, S. and Stamm, S. 2006. The snoRNA HBII-52 regulates alternative splicing of the serotonin receptor 2C. *Science* **311:** 230–232.
Kiyosawa, H., Yamanaka, I., Osato, N., Kondo, S., and Hayashizaki, Y. 2003. Antisense transcripts with FANTOM2 clone set and their implications for gene regulation. *Genome Res.* **13:** 1324–1334.
Kiyosawa, H., Mise, N., Iwase, S., Hayashizaki, Y., and Abe, K. 2005. Disclosing hidden transcripts: Mouse natural sense-antisense transcripts tend to be poly(A) negative and nuclear localized. *Genome Res.* **15:** 463–474.
Krek, A., Grun, D., Poy, M.N., Wolf, R., Rosenberg, L., Epstein, E.J., MacMenamin, P., da Piedade, I., Gunsalus, K.C., Stoffel, M., et al. 2005. Combinatorial microRNA target predictions. *Nat. Genet.* **37:** 495–500.
Krystal, G.W., Armstrong, B.C., and Battey, J.F. 1990. N-myc mRNA forms an RNA-RNA duplex with endogenous antisense transcripts. *Mol. Cell. Biol.* **10:** 4180–4191.
Kumar, M. and Carmichael, G.G. 1998. Antisense RNA: Function and fate of duplex RNA in cells of higher eukaryotes. *Microbiol. Mol. Biol. Rev.* **62:** 1415–1434.
Kuwabara, T., Hsieh, J., Nakashima, K., Taira, K., and Gage, F.H. 2004. A small modulatory dsRNA specifies the fate of adult neural stem cells. *Cell* **116:** 779–793.
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. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.
Lee, Y.S., Nakahara, K., Pham, J.W., Kim, K., He, Z., Sontheimer, E.J., and Carthew, R.W. 2004. Distinct roles for *Drosophila* Dicer-1 and Dicer-2 in the siRNA/miRNA silencing pathways. *Cell* **117:** 69–81.
Lewis, B.P., Shih, I.H., Jones-Rhoades, M.W., Bartel, D.P., and Burge, C.B. 2003. Prediction of mammalian microRNA targets. *Cell* **115:** 787–798.
Lewis, B.P., Burge, C.B., and Bartel, D.P. 2005. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120:** 15–20.
Li, S.G., Zhou, H., Luo, Y.P., Zhang, P., and Qu, L.H. 2005. Identification and functional analysis of 20 Box H/ACA small nucleolar RNAs (snoRNAs) from *Schizosaccharomyces pombe*. *J. Biol. Chem.* **280:** 16446–16455.
Lian, Z., Euskirchen, G., Rinn, J., Martone, R., Bertone, P., Hartman, S., Royce, T., Nelson, K., Sayward, F., Luscombe, N., et al. 2003. Identification of novel functional elements in the human genome. *Cold Spring Harb. Symp. Quant. Biol.* **68:** 317–322.
Lowe, T.M. and Eddy, S.R. 1999. A computational screen for methylation guide snoRNAs in yeast. *Science* **283:** 1168–1171.
Manak, J.R., Dike, S., Sementchenko, V., Kapranov, P., Biemar, F., Long, J., Cheng, J., Bell, I., Ghosh, S., Piccolboni, A., et al. 2006. Biological function of unannotated transcription during the early development of *Drosophila melanogaster*. *Nat. Genet.* **38:** 1151–1158.
Martianov, I., Ramadass, A., Serra Barros, A., Chow, N., and Akoulitchev, A. 2007. Repression of the human dihydrofolate reductase gene by a non-coding interfering transcript. *Nature* **445:** 666–670.
Martone, R., Euskirchen, G., Bertone, P., Hartman, S., Royce, T.E., Luscombe, N.M., Rinn, J.L., Nelson, F.K., Miller, P., Gerstein, M., et al. 2003. Distribution of NF-κB-binding sites across human chromosome 22. *Proc. Natl. Acad. Sci.* **100:** 12247–12252.

Mattick, J.S. 2001. Non-coding RNAs: The architects of eukaryotic complexity. *EMBO Rep.* **2:** 986–991.
Mattick, J.S. 2004. RNA regulation: A new genetics? *Nat. Rev. Genet.* **5:** 316–323.
Mattick, J.S. 2005. The functional genomics of noncoding RNA. *Science* **309:** 1527–1528.
Mattick, J.S. and Gagen, M.J. 2001. The evolution of controlled multitasked gene networks: The role of introns and other noncoding RNAs in the development of complex organisms. *Mol. Biol. Evol.* **18:** 1611–1630.
Mattick, J.S. and Makunin, I.V. 2006. Non-coding RNA. *Hum. Mol. Genet.* (Suppl. 1) **15:** R17–R29.
Meister, G. and Tuschl, T. 2004. Mechanisms of gene silencing by double-stranded RNA. *Nature* **431:** 343–349.
Morey, C. and Avner, P. 2004. Employment opportunities for non-coding RNAs. *FEBS Lett.* **567:** 27–34.
Nelson, P., Kiriakidou, M., Sharma, A., Maniataki, E., and Mourelatos, Z. 2003. The microRNA world: Small is mighty. *Trends Biochem. Sci.* **28:** 534–540.
Nesterova, T.B., Slobodyanyuk, S.Y., Elisaphenko, E.A., Shevchenko, A.I., Johnston, C., Pavlova, M.E., Rogozin, I.B., Kolesnikov, N.N., Brockdorff, N., and Zakian, S.M. 2001. Characterization of the genomic Xist locus in rodents reveals conservation of overall gene structure and tandem repeats but rapid evolution of unique sequence. *Genome Res.* **11:** 833–849.
Ng, P., Wei, C.L., Sung, W.K., Chiu, K.P., Lipovich, L., Ang, C.C., Gupta, S., Shahab, A., Ridwan, A., Wong, C.H., et al. 2005. Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation. *Nat. Methods* **2:** 105–111.
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:** 108–117.
Parra, G., Reymond, A., Dabbouseh, N., Dermitzakis, E.T., Castelo, R., Thomson, T.M., Antonarakis, S.E., and Guigo, R. 2006. Tandem chimerism as a means to increase protein complexity in the human genome. *Genome Res.* **16:** 37–44.
Pollard, K.S., Salama, S.R., Lambert, N., Lambot, M.A., Coppens, S., Pedersen, J.S., Katzman, S., King, B., Onodera, C., Siepel, A., et al. 2006. An RNA gene expressed during cortical development evolved rapidly in humans. *Nature* **443:** 167–172.
Ponjavic, J., Ponting, C.P., and Lunter, G. 2007. Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Res.* **17:** 556–565.
Prasanth, K.V. and Spector, D.L. 2007. Eukaryotic regulatory RNAs: An answer to the 'genome complexity' conundrum. *Genes & Dev.* **21:** 11–42.
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:** 529–540.
Rivas, F.V., Tolia, N.H., Song, J.J., Aragon, J.P., Liu, J., Hannon, G.J., and Joshua-Tor, L. 2005. Purified Argonaute2 and an siRNA form recombinant human RISC. *Nat. Struct. Mol. Biol.* **12:** 340–349.
Rozowsky, J., Newburger, D., Sayward, F., Wu, J., Jordan, G., Korbel, J.O., Nagalakshmi, U., Yang, J., Zheng, D., Guigo, R., et al. 2007. The DART classification of unannotated transcription within the ENCODE regions: Associating transcription with known and novel loci. *Genome Res.* (this issue) doi: 10.1101/gr.5696007.
Schadt, E.E., Edwards, S.W., Guhathakurta, D., Holder, D., Ying, L., Svetnik, V., Leonardson, A., Hart, K.W., Russell, A., Li, G., et al. 2004. A comprehensive transcript index of the human genome generated using microarrays and computational approaches. *Genome Biol.* **5:** R73.
Schattner, P., Decatur, W.A., Davis, C.A., Ares Jr., M., Fournier, M.J., and Lowe, T.M. 2004. Genome-wide searching for pseudouridylation guide snoRNAs: Analysis of the *Saccharomyces cerevisiae* genome. *Nucleic Acids Res.* **32:** 4281–4296.
Shendure, J. and Church, G.M. 2002. Computational discovery of sense-antisense transcription in the human and mouse genomes. *Genome Biol.* doi: 10.1186/gb-2002-3-9-research0044.
Shiraki, T., Kondo, S., Katayama, S., Waki, K., Kasukawa, T., Kawaji, H., Kodzius, R., Watahiki, A., Nakamura, M., Arakawa, T., et al. 2003. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl. Acad. Sci.* **100:** 15776–15781.
Sleutels, F., Zwart, R., and Barlow, D.P. 2002. The non-coding Air RNA is required for silencing autosomal imprinted genes. *Nature* **415:** 810–813.
Snyder, M. and Gerstein, M. 2003. Genomics. Defining genes in the genomics era. *Science* **300:** 258–260.
Stamhuis, I.H., Meijer, O.G., and Zevenhuizen, E.J. 1999. Hugo de Vries on heredity, 1889-1903. Statistics, Mendelian laws, pangenes, mutations. *Isis* **90:** 238–267.
Storz, G. 2002. An expanding universe of noncoding RNAs. *Science* **296:** 1260–1263.
Szymanski, M. and Barciszewski, J. 2002. Beyond the proteome: Non-coding regulatory RNAs. *Genome Biol.* **3:** reviews0005.
Szymanski, M., Barciszewska, M.Z., Zywicki, M., and Barciszewski, J. 2003. Noncoding RNA transcripts. *J. Appl. Genet.* **44:** 1–19.
Takada, S., Paulsen, M., Tevendale, M., Tsai, C.E., Kelsey, G., Cattanach, B.M., and Ferguson-Smith, A.C. 2002. Epigenetic analysis of the Dlk1-Gtl2 imprinted domain on mouse chromosome 12: Implications for imprinting control from comparison with Igf2-H19. *Hum. Mol. Genet.* **11:** 77–86.
Thrash-Bingham, C.A. and Tartof, K.D. 1999. aHIF: A natural antisense transcript overexpressed in human renal cancer and during hypoxia. *J. Natl. Cancer Inst.* **91:** 143–151.
Tijsterman, M. and Plasterk, R.H. 2004. Dicers at RISC; the mechanism of RNAi. *Cell* **117:** 1–3.
Torarinsson, E., Sawera, M., Havgaard, J.H., Fredholm, M., and Gorodkin, J. 2006. Thousands of corresponding human and mouse genomic regions unalignable in primary sequence contain common RNA structure. *Genome Res.* **16:** 885–889.
Torrents, D., Suyama, M., Zdobnov, E., and Bork, P. 2003. A genome-wide survey of human pseudogenes. *Genome Res.* **13:** 2559–2567.
Vanhee-Brossollet, C. and Vaquero, C. 1998. Do natural antisense transcripts make sense in eukaryotes? *Gene* **211:** 1–9.
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:** 1304–1351.
Wahlestedt, C. 2006. Natural antisense and noncoding RNA transcripts as potential drug targets. *Drug Discov. Today* **11:** 503–508.
Wain, H.M., Bruford, E.A., Lovering, R.C., Lush, M.J., Wright, M.W., and Povey, S. 2002. Guidelines for human gene nomenclature. *Genomics* **79:** 464–470.
Washietl, S., Hofacker, I.L., Lukasser, M., Huttenhofer, A., and Stadler, P.F. 2005. Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. *Nat. Biotechnol.* **23:** 1383–1390.
Washietl, S., Pedersen, J.S., Korbel, J.O., Stocsits, C., Gruber, A.R., Hackermüller, J., Hertel, J., Lindemeyer, M., Reiche, K., Tanzer, A., et al. 2007. Structured RNAs in the ENCODE selected regions of the human genome. *Genome Res.* (this issue) doi: 10.1101/gr.5650707.
Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520–562.
Watson, J.D. and Crick, F.H. 1953. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171:** 737–738.
Willingham, A.T. and Gingeras, T.R. 2006. TUF love for "junk" DNA. *Cell* **125:** 1215–1220.
Willingham, A.T., Orth, A.P., Batalov, S., Peters, E.C., Wen, B.G., Aza-Blanc, P., Hogenesch, J.B., and Schultz, P.G. 2005. A strategy for probing the function of noncoding RNAs finds a repressor of NFAT. *Science* **309:** 1570–1573.
Yano, Y., Saito, R., Yoshida, N., Yoshiki, A., Wynshaw-Boris, A., Tomita, M., and Hirotsune, S. 2004. A new role for expressed pseudogenes as ncRNA: Regulation of mRNA stability of its homologous coding gene. *J. Mol. Med.* **82:** 414–422.
Yelin, R., Dahary, D., Sorek, R., Levanon, E.Y., Goldstein, O., Shoshan, A., Diber, A., Biton, S., Tamir, Y., Khosravi, R., et al. 2003. Widespread occurrence of antisense transcription in the human genome. *Nat. Biotechnol.* **21:** 379–386.
Zamore, P.D. and Haley, B. 2005. Ribo-gnome: The big world of small RNAs. *Science* **309:** 1519–1524.
Zhang, M.Q. 2002. Computational prediction of eukaryotic protein-coding genes. *Nat. Rev. Genet.* **3:** 698–709.
Zhang, Z., Harrison, P., and Gerstein, M. 2002. Identification and analysis of over 2000 ribosomal protein pseudogenes in the human genome. *Genome Res.* **12:** 1466–1482.
Zheng, D., Frankish, A., Baertsch, R., Kapranov, P., Reymond, A., Choo, S.W., Lu, Y., Denoeud, F., Antonarakis, S.E., Snyder, M., et al. 2007. Pseudogenes in the ENCODE regions: Consensus annotation, analysis of transcription and evolution. *Genome Res.* (this issue) doi: 10.1101/gr.5586307.