Review TRENDS in Genetics Vol.21 No.2 February 2005

Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments

Jason M. Johnson[1], Stephen Edwards[1], Daniel Shoemaker[2] and Eric E. Schadt[1]

[1]Rosetta Inpharmatics LLC*, 401 Terry Avenue North, Seattle, WA 98109, USA
[2]GHC Technologies, 505 Coast Boulevard South, Suite 309, La Jolla, CA 92037, USA

Microarrays provide the opportunity to measure transcription from regions of the genome without bias towards the location of known genes. This technology thus offers an important source of genomic sequence annotation that is complementary to cDNA sequencing and computational gene-finding methods. Recent 'tiling' microarray experiments that assay transcription at regular intervals throughout the genome have shown evidence of large amounts of transcription outside the boundaries of known genes. This transcription is observed in polyadenylated RNA samples and appears to be derived from intergenic regions, from introns of known genes and from sequences antisense to known transcripts. In this article, we discuss different explanations for this phenomenon.

Introduction

With the completion of the human genome sequence, attention has shifted to determining the complete set of genes and other functional elements. Estimates for the number of protein-coding genes in the human genome have converged towards 20 000–25 000 [1]. There are also many known non-coding genes including rRNAs, tRNAs, small nucleolar RNAs (snoRNAs) and microRNAs. Full-length cDNA sequencing projects provide high-quality transcripts that define the current set of known genes [2–4]. However, most cloning strategies have been biased towards genes that are abundantly expressed in readily accessible tissues. As a result, the discovery rate of new genes via cDNA sequencing has decreased. Although new techniques and new tissue types have improved the novelty rate [5], new high-throughput approaches will be needed to accelerate and complete the characterization of the transcriptome of each organism. Genomic tiling arrays (Box 1) offer three advantages for discovery of new genes: (i) the sensitivity of microarrays enables rare transcripts to be detected; (ii) the parallel nature of the arrays enables numerous samples and genomic sequences to be analyzed; and (iii) the experimental design is not dependent on current genome annotations.

Corresponding author: Johnson, J.M. (jason_johnson@merck.com).
* Rosetta Inpharmatics LLC is a wholly owned subsidiary of Merck & Co., Inc.
Available online 21 December 2004

Multiple lines of evidence suggest that more of the genome is transcribed than represented by current annotations. After 5′ end-sequencing >1 000 000 cDNA clones and removing those that matched known genes, Ota and colleagues found 7000 novel, full-length human transcripts from a set of human tissues and cell lines [6]. Only about half of these appeared to be protein coding, suggesting that both coding and non-coding genes remain to be discovered. Estimates from long serial analysis of gene expression ('LongSAGE') experiments suggest there are at least 15 000 more exons than currently described, at least half of which will be from novel genes [7]. In addition, a recent analysis of the distribution of transposable elements excluded from transcribed regions suggests that 50% of the human genome might be transcribed [8]. Genome-scale transposon studies in yeast have also provided evidence of much broader transcription of the eukaryotic genome [9]. These results complement recent observations from tiling microarray experiments that suggest there is almost ten times as much transcription on human chromosomes 21 and 22 than accounted for by transcripts in public databases [10]. Widespread transcription has also been detected in different species using different microarray tiling strategies [11–16]. These studies call into question the completeness of our current view of transcription and suggest there could be a large amount of genomic 'dark matter' awaiting explanation [15], analogous to the unexplained dark matter comprising much of the mass in the universe. In this article, we briefly review recent microarray tiling publications and discuss potential explanations for the transcription that is observed.

Genomic tiling arrays

The first genomic-scale study of transcription using microarrays was performed in *Escherichia coli* using probes 'tiled' at regular intervals (Box 1) to provide an unbiased view of transcription with respect to the locations of known and predicted genes [12]. Both strands of the intergenic regions and the open reading frames (ORFs) of *E. coli* were tiled using 25mer oligonucleotides, with 6-bp and 30-bp steps, respectively. Expression was detected for most of the ~4000 ORFs in *E. coli* in the growth conditions surveyed. Surprisingly, when a reverse-complement array was used, transcription was also

 0168-9525/$ - see front matter doi:10.1016/j.tig.2004.12.009

Box 1. Tiling microarray experiments

Tiling microarrays are designed to assay transcription at regular intervals of the genome using regularly spaced probes (horizontal red lines) that can be overlapping (Figure I) or separated. The distance between the centers of successive probes is the 'step' size and probes can be selected to be complementary to one strand (as shown) or both strands. Probes can be synthesized directly onto or spotted onto glass slides, and can be synthesized oligonucleotides or PCR products. They are hybridized with fluorescently labeled cRNA or cDNA prepared from cell samples. Regions of greater fluorescent intensity (green peaks in lower panel) can reveal transcription within a large genomic region. In addition, the correlation of probe intensities in several different tissues (co-expression analysis) can be used to identify probes that are detecting exons of the same transcript. The lower panel shows the extent of a hypothetical transcript within the genome. The middle panel is a schematic, magnified view of the hybridization of a genomic region containing an exon.

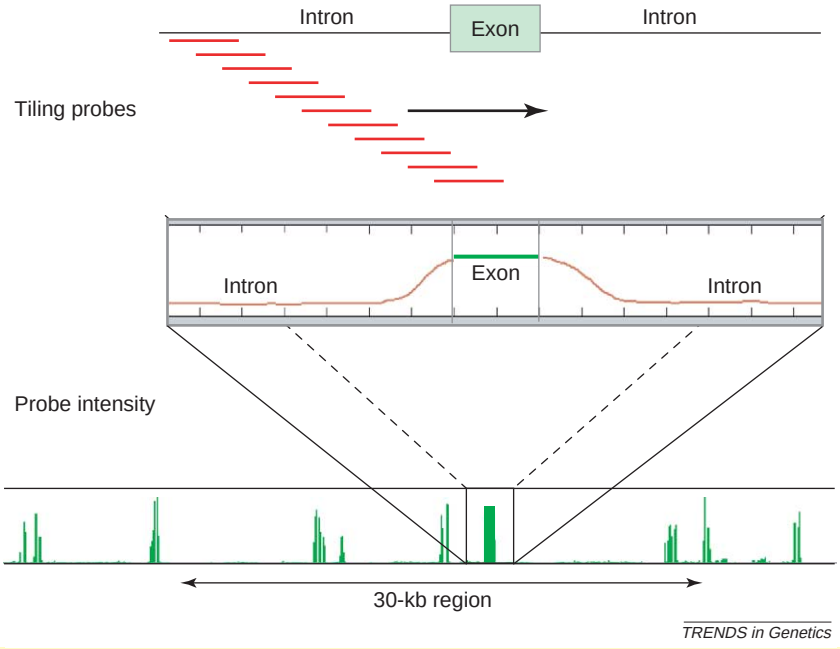

Figure I. Genomic tiling microarray experiments.

detected in >3000 of these ORFs from the antisense strand, suggesting that most of the *E. coli* genome is transcribed at a low level [12]. The first reported human microarray tiling experiment involved tiling through the predicted exons of human chromosome 22q [13]. The results showed that microarray tiling data were useful for the refinement and validation of computational gene predictions, and that correlated behavior of tiling probes across experimental conditions was a helpful indication of exons derived from the same transcript. Microarray experiments have also been used to test gene predictions in the *Drosophila melanogaster* [17], *Saccharomyces cerevisiae* [18,19] and human genomes [11,20]. Recent tiling experiments that are independent of computational gene predictions have been performed for human chromosomes 21 and 22 using 25mer oligonucleotide arrays [10,21], for human chromosomes 20 and 22 using ink-jet arrays with 60mer oligonucleotides [11], for human chromosome 22 using spotted PCR products [14], for the *Arabidopsis* genome using 25mers [15], and for the *Drosophila* genome using 36mers [16]. Table 1 summarizes these genomic tiling experiments.

For the *Arabidopsis* genome, tiling experiments at 8-bp step resolution by Yamada *et al.* identified many novel regions of transcription from poly(A)$^+$ RNA samples including 2000 intergenic regions and regions antisense to some portion of 30% of the annotated genes in *Arabidopsis*. The authors also observed expression of 5817 computationally predicted genes with no previous evidence of expression [15]. The *Drosophila* tiling

Table 1. Recent microarray-based genomic tiling experiments

Study	Microarray type	Samples	Species (chromosomes)
Kapranov *et al.* [10], Kampa *et al.* [21]	Synthesized 25mers	ds-cDNA from poly(A)$^+$ RNA of 11 cell lines, poly(A)$^+$ cRNA from two cell lines	*Homo sapiens* (Chr. 21–22)
Rinn *et al.* [14]	Spotted PCR products	Placental poly(A)$^+$ RNA	*Homo sapiens* (Chr. 22)
Selinger and Church [12]	Synthesized 25mers	Stationary and growth phase total RNA	*Escherichia coli*
Yamada *et al.* [15]	Synthesized 25mers	Poly(A)$^+$RNA from four to six samples	*Arabidopsis thaliana*
Schadt *et al.* [11]	Synthesized 60mers	Poly(A)$^+$ RNA from nine cell lines	*Homo sapiens* (Chr. 20 and 22)
Stolc *et al.* [16]	Synthesized 36mers	Poly(A)$^+$RNA from six developmental stages	*Drosophila melanogaster*

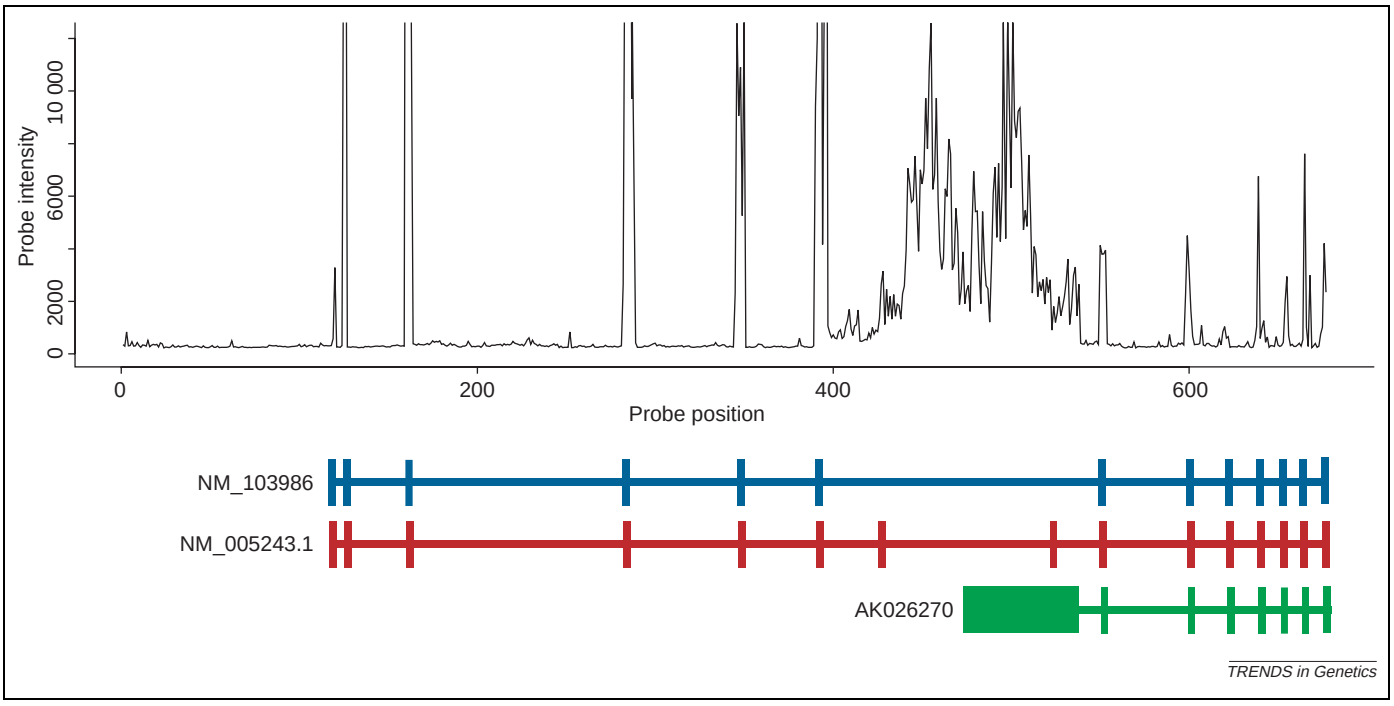

Figure 1. Microarray intensity profile for human thymus poly(A)$^+$ cDNA profiled on a 60mer ink-jet tiling array representing the genomic locus of the Ewing sarcoma breakpoint region 1 gene (*EWSR1*) in 30-nt steps, with probe index as the *x*-axis. No probes are shown for repeat-masked regions. The tiling data for this locus are shown in relation to the exon positions (below the plot) of three EWSR1 cDNAs (Genbank accession numbers: NM_013986, NM_005243.1 and AK026270). The 5′-most exon lies in a repeat-masked region and is not shown. A few peaks with the highest intensity have been truncated in Figures 1–3.

experiments assayed six developmental stages and detected RNA expression for 93% of annotated genes and 41% of the probes in intronic and intergenic regions [16]. For the human genome, Kapranov and colleagues used tiling probes at 35-bp resolution through chromosomes 21 and 22 and assayed transcription using labeled double-stranded cDNA samples prepared from cytoplasmic, polyadenylated RNA from 11 cell lines [10]. Because the samples were double stranded, the strand specificity was not determined. As expected, the authors noted a significant correlation between the density of known or predicted exons and the density of hybridizing probes from tiling data; in other words, regions containing known exons tended to be transcribed. However, most of the hybridizing probes lay outside the positions of known exons. Of probes that were positive when hybridized to samples isolated from at least one of 11 cell lines, 94% lay outside known exon positions [10]. Additional analysis of the same data set by Kampa *et al.* showed that about half of the regions of transcription that can be grouped into contiguous blocks, or 'transfrags' (which are presumably part of the same transcript), lay outside known mRNAs or ESTs [21].

Finally, microarray experiments tiling through human chromosomes 20 and 22 were performed on ink-jet synthesized arrays using 60mer probes in 30-bp steps with samples derived from six different conditions. These experiments found that 47% of positive probes were outside currently annotated exon positions, 22% were within introns and 25% were within intergenic regions [11]. Although estimates from tiling arrays of the amount of dark matter transcription vary, three of the four human tiling studies conclude that there are approximately twice as many nucleotides in poly(A)$^+$ transcripts (i.e. in exons) than are currently annotated in transcript databases [11,14,21]. However, this does not imply that twice as much of the genome is transcribed because the contribution to transcription from dark matter introns is unknown.

Examples of microarray tiling data

The use of tiling microarrays for identifying regions of transcription is illustrated here in three separate examples from our own work [11]. Similar data have been obtained by others [10,21]. Transcription is detected in regions of the human genome that are supported to varying degrees by independent computational and experimental data. The first example shows expression activity for a well-characterized gene, Ewing sarcoma breakpoint region 1 (*EWSR1*; Figure 1). The tiling data clearly detect exons corresponding to the RefSeq transcript of this gene but there also appears to be additional signal in at least one intron. The transcript represented by the cDNA sequence AK026270, shown in Figure 1, might correspond to a portion of this activity (as might others such as AL833489 and BX648769, not shown). In this case, the unexplained transcription suggests there are alternatively spliced isoforms or other uncharacterized RNAs transcribed from this locus.

The second example shows how tiling data can support low-confidence gene predictions (Figure 2). This chromosome 20 region contains an *ab initio* gene prediction, and a few of the predicted exons are also supported by ESTs. Here the exons predicted from microarray tiling data show good concordance with the computationally predicted exons, adding independent support for the existence of this novel gene.

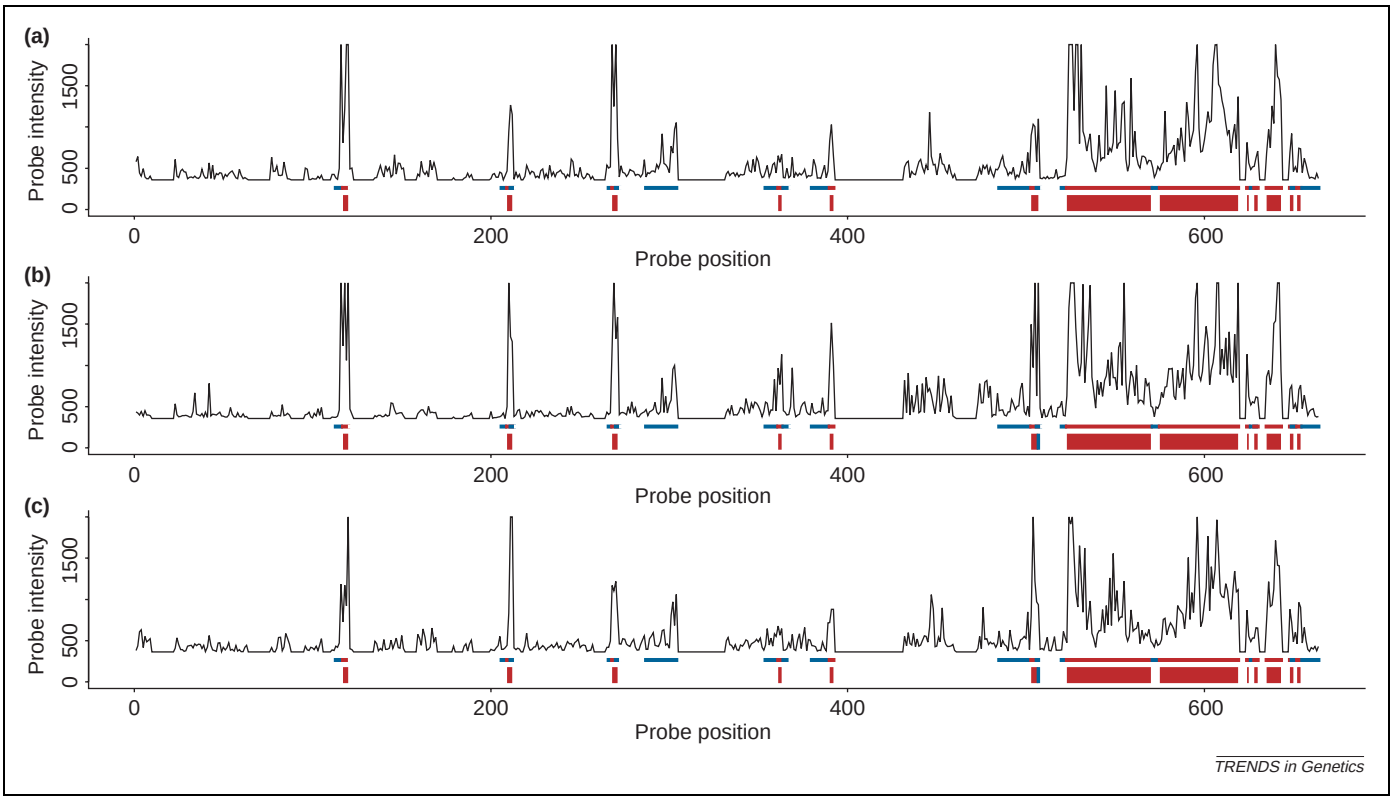

Figure 2. Microarray tiling confirmation of a predicted gene. Microarray tiling probe intensities for a region of human chromosome 20 that contains an *ab initio* gene prediction (made by the program GENSCAN [48]). Predicted exons for this gene are shown with blue lines. The transcription detected by microarrays has been grouped into a single transcriptional unit (dark red) by the correlated behavior of these probes across different human samples [11]. The conditions displayed are **(a)** thalamus, **(b)** testes and **(c)** uterus.

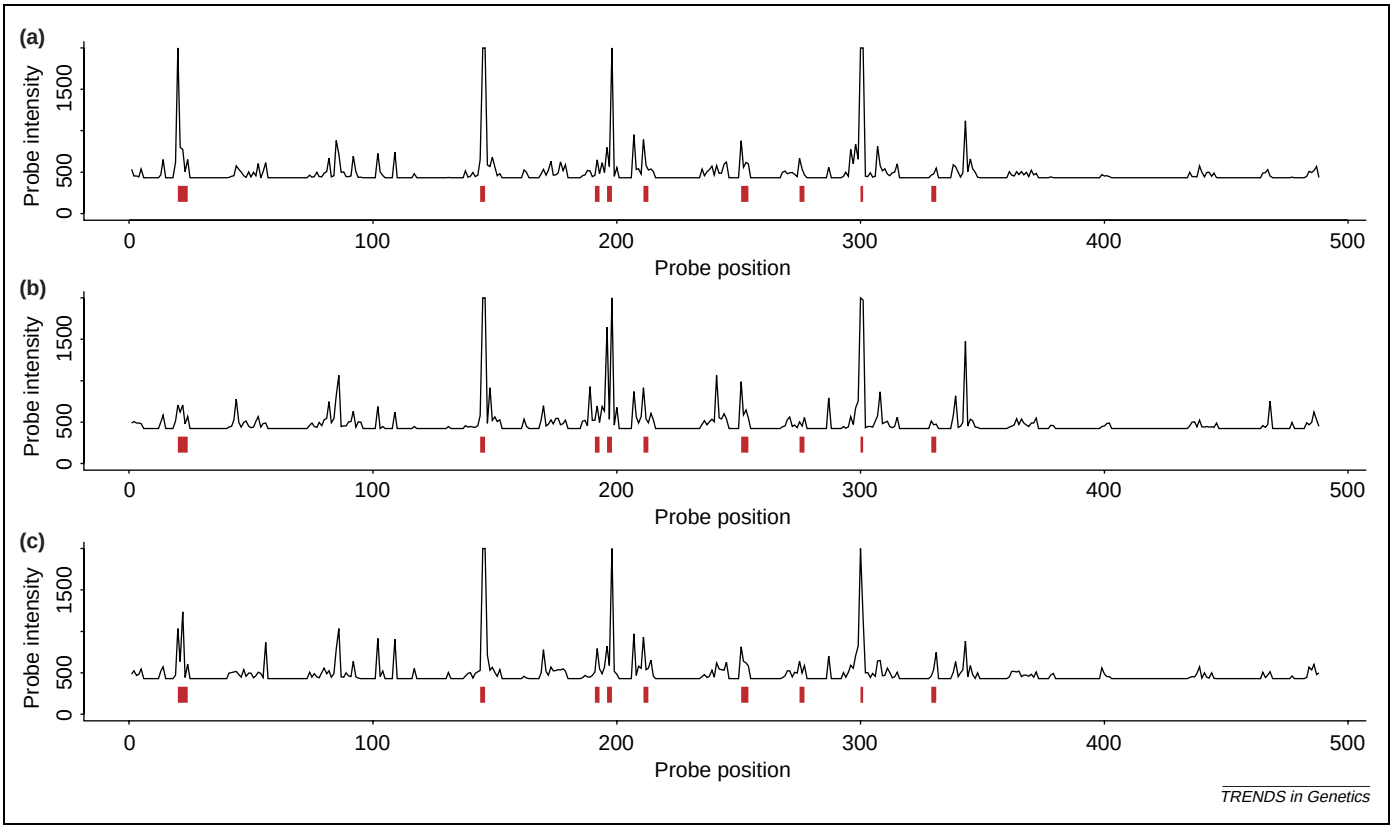

Figure 3. Tiling intensity profile for poly(A)$^+$ RNA samples from **(a)** thalamus, **(b)** testes and **(c)** uterus from an intergenic region of human chromosome 20 that contains no mapped ESTs or computational gene predictions. Dark red bars indicate transcription activity that has been grouped into a single transcriptional unit given the relative activity of these probes with respect to one another across different samples as described in Schadt *et al.* [11].

The third example illustrates transcriptional activity in an intergenic region of human chromosome 20 devoid of other experimental or computational support (Figure 3). This type of transcription comprises approximately half of the total detected in chromosomes 20 and 22 [11]. Here the transcriptional activity (highlighted by the dark red bars) has been grouped into a single transcript by analysis of the co-expression of these probes over six tissues [11].

Possible explanations for dark matter transcription

There are many potential explanations for the transcription activity that appears outside of known and predicted exons, such as biological 'noise,' experimental artifacts or previously undetected protein-coding or non-coding genes. These possibilities are discussed in more detail below.

Novel protein-coding genes

One explanation for dark matter transcription is that it arises from many uncharacterized protein-coding genes. Although a core set of well-characterized (RefSeq) genes is common to most estimates of human gene number, a comparison of predicted genes from the two initial analyses of the human genome from different assemblies [22,23] indicated little overlap among non-RefSeq genes [24], and suggested that there could be thousands more protein-coding genes than originally estimated. Further evidence arose from a variety of different sources, including SAGE, full-length cDNA and EST mapping, sequence conservation across species, analysis of ORFs and statistical analysis of sequence features [25]. The strongest evidence that protein-coding genes remain to be discovered comes from full-length cDNA sequencing projects. A recent study found 2000 novel, non-redundant cDNAs with ORFs >300nt that do not overlap any Ensembl gene models (derived from computational methods, mRNAs and ESTs) [6]. LongSAGE experiments have also detected hundreds of new protein-coding genes [7].

Ab initio computational gene finding programs predict numerous (>10 000) protein-coding genes in the genome not supported by other lines of evidence that could potentially explain much of the dark matter transcription [11]. However, when the expression of these sequences was tested in 60 human samples using 'predicted transcript arrays', <4% of the *ab initio* predictions could be verified using co-expression analysis (Box 1) compared with 75% of the genes with RefSeq transcripts [11]. This does not rule out the possibility that the *ab initio* gene predictions are real, but are expressed in rare tissues or have low expression.

ESTs are also a rich source of predicted genes that do not overlap with other lines of evidence [26]. Of 7170 EST-predicted genes in the human genome without other support, 20% (1428) could be verified by co-expression analysis using tiling arrays in 60 tissues [11]. Overall, however, co-expression analysis of probes representing a genome-wide set of ~50 000 putative transcripts derived from ESTs, *ab initio* algorithms and alignment of protein sequences to the genome (excluding predictions made from 'tiling' microarrays) does not lead to a significantly higher count of protein-coding genes [11].

Although cDNA sequencing indicates that more human protein-coding genes remain to be discovered – most likely those with low expression levels, expression that is specific to particular developmental stages or cell types, or with atypical sequence properties (e.g. unusual GC-richness) [6] – the discovery rate is decreasing. For example, <9% of the 21 243 non-redundant transcript clusters sequenced by Ota *et al.* are novel and have ORFs that are >300nt [6]. Furthermore, the number of well-annotated protein-coding loci in the human genome has actually decreased during the past few years as the quality of the genome has improved [1]. These trends suggest that new protein-coding genes are unlikely to account for a large proportion of dark matter transcription.

Expressed pseudogenes are another potential source of dark matter transcription. Yamada *et al.* detected expression of ~20% of the *Arabidopsis* loci currently annotated as pseudogenes [15]. There are at least 20 000 pseudogenes in the human genome, with some known to be transcribed (or even functional) [1,27–29]. However, because these pseudogenes have sequence properties of protein-coding genes, they are relatively straightforward to predict with computational gene-finding programs, and do not overlap a significant fraction of tiling-detected transcription [11].

Novel non-coding genes

There are a variety of types of non-coding RNAs (ncRNAs) with known functions (reviewed by Storz [30]) and these functions might represent only a fraction of the regulatory roles for RNAs in higher organisms [31]. Each of the tiling articles reviewed here concludes that ncRNAs are likely to explain much of the observed transcription. Although many known ncRNAs are localized to the nucleus and not polyadenylated, the evidence for the existence of many poly(A)$^+$ ncRNAs is increasing. More than 40% of a non-redundant set of 33 000 mouse cDNAs are apparently non-coding, in that they do not have ORFs >300nt [4]. Approximately 11 000 of these appear to be novel and many appear to have multiple exons. Similar proportions are observed for novel human cDNA sequences; almost 5500 of the ~7500 novel clusters of full-length cDNA sequences show no obvious reading frames, indicating that the potential for new discovery is probably greater for non-coding transcripts than for coding transcripts [6]. From these putative ncRNAs clones, 74% of those tested (32) were validated by RT–PCR [6].

Additional evidence that ncRNAs account for some of the dark matter transcription comes from validation of microarray-detected transcription. For example, to validate predictions from human 25mer microarrays, 193 regions of dark matter transcription detected by the arrays were tested by RT–PCR, and 65% of these could be cloned or sequence verified [21]. Although these and other experiments show that many ncRNAs exist, it is not certain that these are functional genes. Some evidence of biological function comes from the fact that a subset are conserved in mouse, and that ~20% are regulated in response to retinoic acid [32,33]. In addition, in a recent chromatin immunoprecipitation study of three transcription factors, comparable numbers of binding sites occur

near known non-coding and coding cDNAs, suggesting that many ncRNAs are biologically regulated and functional [33]. For the *Drosophila* genome, 15% of the intronic and intergenic probes with significant RNA expression were also differentially expressed across tissues, although it should be noted this is substantially less than the 57% of exon probes for which differential expression was detected. In addition, there was a statistically significant increase in sequence identity between *D. melanogaster* and *D. pseudoobscura* genomic sequences containing expressed intronic and intergenic probes [16].

Antisense transcription

Some of the dark matter transcription is antisense to known genes. Antisense RNA is likely to have a regulatory role in bacteria [34], and most genes in *E. coli* appear to have antisense transcription [12]. By analysis of cDNA sequence clusters, 7600 annotated genes in *Arabidopsis* (30%) had significant antisense expression [15]. For the rice genome, characterization of 32 000 full-length cDNAs led to the identification of 600 antisense transcript pairs, half of which have no ORF in one member of the pair [35].

Extensive antisense transcription has also been observed in other eukaryotic genomes, and much of the recent evidence has come from cDNA sequences deposited in the public domain. EST and cDNA databases were used to identify as many as 5880 human transcript clusters containing overlapping antisense transcripts and some of these have been validated by northern blot [36–38]. From 60 000 mouse cDNA clones, ~2400 sense–antisense pairs were detected [4,39]. In addition, although it has not yet been reported whether microarray-detected transcription has statistically significant proximity to transcription-factor-binding sites, antisense RNAs in public sequence databases do have a statistically significant proximity to transcription-factor-binding sites determined by chromatin immunoprecipitation experiments [33]. Finally, in tiling experiments designed to determine the strand specificity of dark matter transcription, 11% of probes that overlapped known exons showed transcription on the opposite strand [21].

Alternative isoforms

Approximately half of the positive tiling probes from array studies of human chromosomes 20–22 lie within introns or within a few hundred bases of exons of known genes, suggesting that much of the dark matter transcription could arise from extended or more complex transcription of known genes [10,11]. New isoforms created by alternative splicing, alternative transcript initiation sites and alternative polyadenylation are possible explanations for this novel transcription. Interestingly, only ~5% of the probes detected as expressed within the introns of annotated genes correspond to known alternative isoforms [10], suggesting that many alternative isoforms remain to be characterized in the human transcriptome. The existence of new exons within *Drosophila* introns was also suggested by the correlated activity of intronic tiling probes with neighboring exons and by overlap of these probes with Genscan predictions [16]. In *Arabidopsis*, 10% of the hybridizing fragments contained within known introns met the stringent criteria of expression in at least five of six experimental conditions [14]. However, half of human intronic transcription appears to be on the opposite strand, suggesting many of the intronic transcripts are novel RNAs rather than alternatively spliced isoforms [14].

Extensions of known genes

Because complete full-length transcripts are difficult to capture in cDNA cloning, and untranslated regions (UTRs) are difficult for gene-finding programs to predict, many known transcripts will undoubtedly turn out to be longer than currently annotated. This could explain some of the transcription adjacent to annotated genes in intergenic regions. It has been hypothesized that biases created by the overabundance of smaller genomic contigs and the limitations of gene-finding programs have obscured the reality that most intergenic regions are actually introns [40,41]. For example, the Usher Syndrome Type 2A gene (*USH2A*), originally annotated with 21 exons in 250kb of genomic sequence (GenBank accession number: NM_007123), was recently found by RT–PCR and 3′ rapid amplification of cDNA ends (RACE) to contain another 51 downstream exons, and to span a total of 790kb of genomic sequence [42]. The new 'long' isoform of this gene encodes a predicted protein of >5000 amino acids, and the nearest downstream gene (*KCTD3*) is now only ~4kb away from the *USH2A* 3′ UTR [42]. It seems likely that other known protein-coding genes will be found to occupy larger genomic regions. Thus, adding new exons to the existing protein-coding transcriptome might also explain a fraction of the dark matter transcription in intergenic regions.

Biological 'artifacts'

Transcription is certainly not free of noise. Given the high regulatory cost of perfection, some low-level, leaky transcription is expected [43], and if non-specific poly(A)$^+$transcripts are widespread they could explain a large proportion of dark matter transcription observed with tiling microarrays. The possibility of low-level transcriptional noise in the human genome is also consistent with the fact that many of the tiling-predicted transcripts appear to be expressed at such low levels that they are at the limits of detection of RT–PCR or northern blots [10,14,21]. Another argument against the majority of dark matter transcripts being functional is that only ~7–20% of the novel transcribed regions on human chromosomes 21 and 22 are conserved in the mouse genome [14,21]. This is in contrast to ~44% of the transcribed regions overlapping known genes. The large amount of intronic transcription signal suggests that retained introns could be another relevant source of biological noise. More than 40% of introns were detected as expressed in *Drosophila* [16]. It is clear that in-depth biochemical experiments will be required to assess the biological relevance of dark matter transcripts. Validation experiments to date have demonstrated that many of these transcripts exist, some are tissue-specific and some are regulated in response to retinoic acid [33]; however, it is unknown how many of these transcripts are functional.

Experimental artifacts

The fact that dark matter transcription has been observed by different groups on different array platforms in different species with similar outcomes suggests artifacts are not the dominant cause. However, several different types of artifacts are possible contributors, including the following: genomic contamination of RNA samples, contamination from unspliced mRNAs, unintended double-stranded labeling of RNA and cross-hybridization. Most total RNA samples contain at least some residual genomic DNA as a contaminant, which will appear as dark matter transcription when labeled and hybridized to genomic tiling arrays. The level of this genomic contamination varies widely by tissue, vendor and purification method. A simple control experiment to check for genomic contamination is to omit the reverse transcriptase step in the RNA labeling protocol [21]. RNA samples can also be treated with DNaseI to remove genomic DNA before labeling and hybridization to genomic tiling arrays. However, this approach can still leave detectable amounts of genomic DNA and, therefore, it is essential to include appropriate controls. A second type of experimental artifact can be caused by incompletely processed mRNAs that are inadvertently isolated during the RNA extraction procedure and can result in false positive hybridization signals in the introns of known genes. This problem can be minimized by purifying mRNA from the cytosolic fraction of an RNA preparation [44]. However, significant dark matter transcription is observed even when cytosolic mRNA is used [10].

A third potential experimental artifact is inadvertent double-stranded labeling, which can be caused by spurious second-strand synthesis during labeling protocols that use reverse transcriptase for first-strand cDNA synthesis [45]. This artifact would manifest primarily as false antisense transcription. However, it is unlikely to be the major explanation for dark matter transcription because similar amounts of antisense transcription have been observed using a direct RNA end-labeling strategy (using T4 RNA ligase) that avoids this double-stranded labeling artifact [21], and most of the dark matter transcription observed is not antisense to known transcripts [10,11,14].

Cross-hybridization is another potential source of error, which can be of two types. Specific cross-hybridization is caused by sequence-similar regions of the genome (e.g. paralogous genes or recently duplicated sections of the genome), whereas non-specific cross-hybridization occurs between probes and samples that do not share significant sequence similarity. Some of the genomic tiling experiments using 25mer arrays employed mismatch probes as a control for both types of cross-hybridization, and this was shown to be useful in previous experiments [46]. Mismatch controls were particularly helpful for the *E. coli* tiling experiments, which used random-primed total RNA containing substantial amounts of rRNA and tRNA. In the *Arabidopsis* tiling experiments, only perfect-match probes were used, after a pilot experiment showed that mismatch probes were not necessary for identifying gene structures [15]. Specific cross-hybridization was determined not to be a major contributing factor to the human 25mer tiling results by investigation of gene families and pseudogenes (A. Piccolboni *et al.*, unpublished). For the human ink-jet, 60mer experiments (which did not use mismatch probes) it was estimated that 20–25% of intron and intergenic probes have sequence similarities that have some potential for specific cross-hybridization, and 15–20% of the intron probes contain sequence features (e.g. GC-richness) that might make them susceptible to non-specific cross-hybridization, with some overlap between the two categories [11]. The conclusion from this study was that $<35\%$ of unexplained transcript signals could be caused by cross-hybridization.

False positives

One related limitation of the microarray tiling experiments is the lack of good negative controls, that is, regions of the genome that are known not to be transcribed or present in a given poly(A)$^+$ RNA sample. This limitation prevents an accurate estimate of false positives. Although introns and promoters are two negative-control possibilities, they are not easy to define and some are transcribed and polyadenylated (e.g. transcription upstream of the SER3 promoter region in yeast [47]). Validation efforts suggest that false positives might be a concern. Although thorough controls for genomic contamination are used in many cases, the majority of transcripts detected by tiling alone seem to have such low expression levels that they are not easily detected by RT–PCR or northern blot. Of the 118 transcriptionally active but unannotated regions tested from tiling experiments using placental RNA, only 30 were detected by RNA-blot analysis [14]. A similar rate was reported by Kapranov *et al.* [10]. By contrast, many known transcripts present at low copy numbers per cell reveal visible RT–PCR bands after a similar number of amplification cycles. To help rule out artifacts, it would also be helpful to know if there is a statistically significant correlation between the expression levels estimated by the tiling microarray data and expression levels assayed by quantitative PCR in the same tissue. Ultimately, however, the biochemical function of these transcripts will have to be demonstrated in the laboratory. Although each of the experimental artifacts discussed in this section could only account for a minor fraction of the dark matter transcription, the reported difficulty of experimental validation using techniques such as RT–PCR or northern blotting is strong evidence that a non-trivial fraction overall is due to artifacts.

Comparisons between independent tiling data sets

One potentially informative approach to evaluate possible explanations for dark matter transcription is to compare tiling data sets from different sample preparation methods or microarray platforms. A comparison of different sample preparation methods has been published: Kampa *et al.* [21] profiled cRNA from A375 and Jurkat cells using Affymetrix arrays (http://www.affymetrix.com/) representing chromosomes 21 and 22, and compared the results with tiling profiles from double-stranded cDNA samples from the same two cell lines on the same microarrays [10]. A different sample labeling method was used (direct

Box 2. Comparison of human chromosome 22 tiling data

The 60-nt data assayed transcription from both genomic DNA strands using reverse complement probes, and expressed regions were determined as described previously. For the purposes of this comparison, we counted only the positive strand probes and considered these probes to be expression-supported if expression was detected on either strand. For the 25-nt data set [10], the difference between the perfect match (PM) and corresponding mismatch probe, along with the estimates on chip-wide noise and background intensity computed by the authors, were used to determine the set of probes that were detected as two standard deviations above the background. The maximum expression statistic for each probe pair over the 11 conditions was used for these summaries (Table I). Only the chromosome 22 probes from this study are included here. The expressed regions from the PCR-fragment data set were chosen using the methods described by Rinn *et al.* [14].

Table I. A comparison of three microarray-based genomic tiling studies

Feature	Schadt *et al.* [11]	Kapranov *et al.* [10]	Rinn *et al.* [14]
Probe length (nt)	60	25	300–1400
Number of probes mapped	600 816	480 500	20 558
Number of nucleotides covered by all mapped probes	191 39 010	12 012 500	16 827 434
Number of probes overlapping probes in both other experiments	428 269	416 829	20 272
Number of mapped probes with expression support (eProbes)	22 866	85 560	2446
Number of eProbes overlapping eProbes in both other experiments	1826	1631	398

end-labeling of RNA with T4 ligase) and the hybridizations were performed with RNA instead of DNA, but otherwise the experiments were similar. However, only ~35% of the 'positive' probes were identical between the experiments [21]. The results were also analyzed by grouping the probes from both experiments into contiguous transcriptional units, or 'transfrags'. Of the 497kb of genomic sequence grouped into transfrags from the samples prepared as cDNA, 15% were confirmed by overlap with cRNA transfrags. The small fraction is explained at least in part by the lower sensitivity of the cRNA experiments. Of this overlapping sequence, 77% maps to known mRNAs and ESTs, a significant increase from 51% using cRNA data alone [21]. This enrichment suggests that tiling data are detecting biological transcription. It also suggests the intersection of different tiling data sets might be a valuable method for reducing experimental noise.

To compare tiling data from different microarray platforms, we performed a nucleotide-by-nucleotide comparison among the three human chromosome 22 tiling data sets discussed above. Although probe selection, probe length, probe density and repeat-masking methods differed, most of the probes used in each experiment overlapped with probes from the other two experiments by at least one base pair. This overlap varied from 99% for the PCR-fragment probes to 72% for the 60-bp probes, which cover additional genomic sequence regions. The overlap between the hybridizing, expression-supported probes of each experiment with those of the other experiments is shown in Box 2. The number of probes detecting expression in the three experiments varied widely, based on the stringency of the methods used to declare expression activity. Schadt *et al.* detected transcription with 22 866 of 600 816 probes (3.8%) across eight conditions using 60-bp probes, Rinn *et al.* (using longer PCR probes) detected 2446 of 20 558 (11.9%) in a single condition, and data from Kapranov *et al.* (from 25-bp oligos) show 85 560 of 480 500 (17.8%) detected across 11 conditions. Although different biological tissue samples and experimental methods were used, the overlap in observed expression activity for the three experiments is larger than what would be expected by chance, and each of the pair-wise comparisons is also statistically significant.

Examples of the three tiling data sets for common regions of human chromosome 22 are shown in Figure 4 for the locus of a known gene, for an intergenic region with transcriptional activity and for a second known gene with transcription in the upstream intergenic region. Because the PCR products are longer than the synthesized probes (300–1400bp), the intensity profiles have lower resolution. For many genomic regions, there is relatively good agreement between predictions from the three methods (e.g. Figure 4a,c). However, there are also many regions where the overlap is not convincing, particularly in intergenic regions (Figure 4b).

One conclusion that can be drawn from this comparison is that most of the observed dark matter transcription is either tissue-specific or platform-specific (i.e. detected with different specificities on different platforms) because a low percentage of positive probes overlap in these three experiments. Furthermore, most of the expression activity that is common among the different experiments is constrained to the loci of known genes and known exons. For example, 89% of the 25-bp positive probes overlapping the 60-bp positive probes were located within exons or introns of known transcripts, with 67% in exons.

Future steps

To resolve some of the issues about the significance of dark matter transcription, it could be helpful for new tiling studies to conduct replicate experiments or test the same biological samples on multiple platforms. It is also important for different data sets and approaches to be integrated; for example, by layering microarray evidence on top of cDNA, EST, *ab initio* gene predictions and cross-species conservation throughout the genome. This integration of techniques can also be performed in a directed fashion. For example, one could identify sets of probes from tiling-predicted loci that are conserved in rodents and humans, and then hybridize them across many more conditions to see if they are co-expressed with each other and with genes of known function. In general, more samples will enable better analysis of co-expression of neighboring probes, and further reduce the potential for false positives [11].

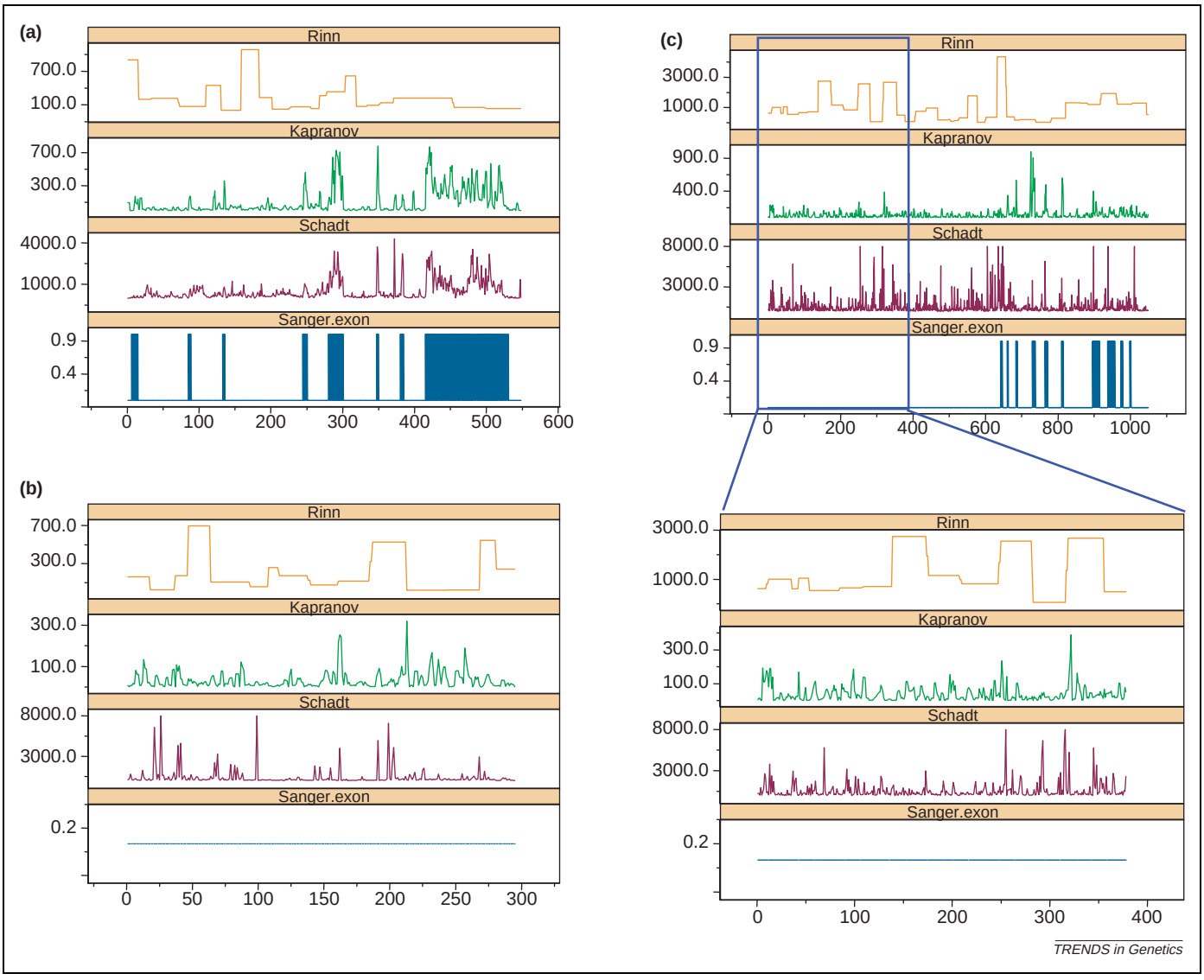

Figure 4. Comparison of three microarray-based annotation methods for regions of human chromosome 22 (Box 2). **(a)** Example of transcriptional activity in a known gene, nucleoporin (NUP50; SangerR3:bk217C2.C22.1) showing good agreement among the three methods and with known exon positions (Sanger.exon). The coordinates for the region shown are chromosome 22: 42 194 799 - 42 219 778). Synthesized oligonucleotide data are from Jurkat cells [10] and thymus [11]. **(b)** Example of an unannotated genomic region (chromosome 22: 42 658 868 – 42 673 838) showing transcriptional activity using all three methods, although exon boundaries are difficult to distinguish. Synthesized oligonucleotide data are from U-87mg cells [10] and chronic myelogenous leukemia (K562) cells [11]. **(c)** Upper panel: genomic region view (chromosome 22: 17 201 690 – 17 246 660) of another known gene (SangerR3:Em:AC007663.C22.2) with activity in the upstream intergenic region (blue box). Intensity data for Kapranov *et al.* and Schadt *et al.* are from NCCIT cells and testes, respectively. Lower panel: higher resolution view of the transcriptional activity in the highlighted intergenic region. Several candidate exons are evident from the Kapranov *et al.* and Schadt *et al.* data. The results of Rinn *et al.* [14] also support transcriptional activity in the region. All coordinates are relative to the UCSC (http://genome.ucsc.edu/) November 2002 human genome assembly.

Concluding remarks

Genomic tiling using microarrays is an important complement to other efforts to define the transcriptome because it provides an independent view of transcription unbiased to the positions of known and predicted genes. The fact that this transcription correlates with the positions of known genes, both statistically [11] and visually (Figure 1) is encouraging, but the scale of observed transcription is difficult to understand. It is clear that no single explanation can account for all, or even most, of the unexplained transcription but it is possible that several explanations together will account for much of it. For example, non-coding regulatory RNAs might explain much of the antisense transcription; new protein-coding genes and gene extensions will reduce the amount of unexplained intergenic transcription, alternative splicing and new RNA genes will explain a portion of intronic transcription, and cross-hybridization might account for the remainder. Whatever their sources are, the majority of these transcripts do not seem to be highly expressed in the tissues surveyed. This is attested to by the difficulty of validation and the fact that most are not represented in EST databases. Another puzzling fact is that little of the transcription lies in regions that are conserved between human and mouse, although statistically significant sequence homology between *Drosophila* species has been observed for dark matter transcription [16]. At this early stage in our understanding of array-detected dark matter transcription, it seems unwise to rule out major contributions from biological and experimental artifacts. However, the results from recent full-length cDNA sequencing projects make it clear that there will continue to be new

genes discovered – many of them non-coding genes, and many of them antisense – in addition to alternative isoforms of known genes, and it would be equally unwise to assume that any transcription we cannot explain is just 'noise'.

Acknowledgements

We thank J. Castle, G. Cavet and M. Parrish for helpful discussions and article review. We are also grateful for the thoughtful and helpful comments received from the journal referees.

References

1 International Human Genome Sequencing Consortium. (2004) Finishing the euchromatic sequencing of the human genome. *Nature* 431, 931–945
2 Strausberg, R.L. *et al*. (1999) The mammalian gene collection. *Science* 286, 455–457
3 Kawai, J. *et al*. (2001) Functional annotation of a full-length mouse cDNA collection. *Nature* 409, 685–690
4 Okazaki, Y. *et al*. (2002) Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* 420, 563–573
5 Carninci, P. *et al*. (2003) Targeting a complex transcriptome: the construction of the mouse full-length cDNA encyclopedia. *Genome Res.* 13, 1273–1289
6 Ota, T. *et al*. (2004) Complete sequencing and characterization of 21,243 full-length human cDNAs. *Nat. Genet.* 36, 40–45
7 Saha, S. *et al*. (2002) Using the transcriptome to annotate the genome. *Nat. Biotechnol.* 20, 508–512
8 Semon, M. and Duret, L. (2004) Evidence that functional transcription units cover at least half of the human genome. *Trends Genet.* 20, 239–232
9 Ross-Macdonald, P. *et al*. (1999) Large-scale analysis of the yeast genome by transposon tagging and gene disruption. *Nature* 402, 413–418
10 Kapranov, P. *et al*. (2002) Large-scale transcriptional activity in chromosomes 21 and 22. *Science* 296, 916–919
11 Schadt, E.E. *et al*. (2004) A comprehensive transcript index of the human genome generated using microarrays and computational approaches. *Genome Biol.* 5, R73
12 Selinger, D.W. *et al*. (2000) RNA expression analysis using a 30 base pair resolution *Escherichia coli* genome array. *Nat. Biotechnol.* 18, 1262–1268
13 Shoemaker, D.D. *et al*. (2001) Experimental annotation of the human genome using microarray technology. *Nature* 409, 922–927
14 Rinn, J.L. *et al*. (2003) The transcriptional activity of human Chromosome 22. *Genes Dev.* 17, 529–540
15 Yamada, K. *et al*. (2003) Empirical analysis of transcriptional activity in the *Arabidopsis* genome. *Science* 302, 842–846
16 Stolc, V. *et al*. (2004) A gene expression map for the euchromatic genome of *Drosophila melanogaster*. *Science* 302, 655–660
17 Hild, M. *et al*. (2003) An integrated gene annotation and transcriptional profiling approach towards the full gene content of the *Drosophila* genome. *Genome Biol.* 5, R3
18 Winzeler, E.A. *et al*. (1998) Direct allelic variation scanning of the yeast genome. *Science* 281, 1194–1197
19 Wodicka, L. *et al*. (1997) Genome-wide expression monitoring in *Saccharomyces cerevisiae*. *Nat. Biotechnol.* 15, 1359–1367
20 Penn, S.G. *et al*. (2000) Mining the human genome using microarrays of open reading frames. *Nat. Genet.* 26, 315–318
21 Kampa, D. *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
22 Venter, J.C. *et al*. (2001) The sequence of the human genome. *Science* 291, 1304–1351
23 Lander, E.S. *et al*. (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921
24 Hogenesch, J.B. *et al*. (2001) A comparison of the Celera and Ensembl predicted gene sets reveals little overlap in novel genes. *Cell* 106, 413–415
25 Snyder, M. and Gerstein, M. (2003) Genomics. Defining genes in the genomics era. *Science* 300, 258–260
26 Camargo, A.A. *et al*. (2001) The contribution of 700,000 ORF sequence tags to the definition of the human transcriptome. *Proc. Natl. Acad. Sci. U. S. A.* 98, 12103–12108
27 Hirotsune, S. *et al*. (2003) An expressed pseudogene regulates the messenger-RNA stability of its homologous coding gene. *Nature* 423, 91–96
28 Mighell, A.J. *et al*. (2000) Vertebrate pseudogenes. *FEBS Lett.* 468, 109–114
29 Harrison, P.M. *et al*. (2002) Molecular fossils in the human genome: identification and analysis of the pseudogenes in chromosomes 21 and 22. *Genome Res.* 12, 272–280
30 Storz, G. (2002) An expanding universe of noncoding RNAs. *Science* 296, 1260–1263
31 Mattick, J.S. (2003) Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. *BioEssays* 25, 930–939
32 Dermitzakis, E.T. *et al*. (2002) Numerous potentially functional but non-genic conserved sequences on human chromosome 21. *Nature* 420, 578–582
33 Cawley, S. *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
34 Wagner, E.G. and Simons, R.W. (1994) Antisense RNA control in bacteria, phages, and plasmids. *Annu. Rev. Microbiol.* 48, 713–742
35 Osato, N. *et al*. (2003) Antisense transcripts with rice full-length cDNAs. *Genome Biol.* 5, R5
36 Yelin, R. *et al*. (2003) Widespread occurrence of antisense transcription in the human genome. *Nat. Biotechnol.* 21, 379–386
37 Lehner, B. *et al*. (2002) Antisense transcripts in the human genome. *Trends Genet.* 18, 63–65
38 Chen *et al*. (2004) Over 20% of human transcripts might form sense-antisense pairs. *Nucleic Acids Res.* 32, 4812–4820
39 Kiyosawa, H. *et al*. (2003) Antisense transcripts with FANTOM2 clone set and their implications for gene regulation. *Genome Res.* 13, 1324–1334
40 Wong, G.K. *et al*. (2000) Is "junk" DNA mostly intron DNA? *Genome Res.* 10, 1672–1678
41 Wong, G.K. *et al*. (2001) Most of the human genome is transcribed. *Genome Res.* 11, 1975–1977
42 Van Wijk, E. *et al*. (2004) Identification of 51 novel exons of the syndrome type 2A (USH2A) gene that encode multiple conserved functional domains and that are mutated in patients with Usher syndrome type II. *Am. J. Hum. Genet.* 74, 738–744
43 Dennis, C. (2002) The brave new world of RNA. *Nature* 418, 122–124
44 Diehn, M. *et al*. (2000) Large-scale identification of secreted and membrane-associated gene products using DNA microarrays. *Nat. Genet.* 25, 58–62
45 Castle, J. *et al*. (2003) Optimization of oligonucleotide arrays and RNA amplification protocols for analysis of transcript structure and alternative splicing. *Genome Biol.* 4, R66
46 Lockhart, D.J. *et al*. (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.* 14, 1675–1680
47 Martens, J.A. *et al*. (2004) Intergenic transcription is required to repress the *Saccharomyces cerevisiae* SER3 gene. *Nature* 429, 571–574
48 Burge, C. and Karlin, S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* 268, 78–94