

# Experimental annotation of the human genome using microarray technology

D. D. Shoemaker\*, E. E. Schadt\*, C. D. Armour, Y. D. He, P. Garrett-Engele, P. D. McDonagh, P. M. Loerch, A. Leonardson, P. Y. Lum, G. Cavet, L. F. Wu, S. J. Altschuler, S. Edwards, J. King, J. S. Tsang, G. Schimmack, J. M. Schelter, J. Koch, M. Ziman, M. J. Marton, B. Li, P. Cundiff, T. Ward, J. Castle, M. Krolewski, M. R. Meyer, M. Mao, J. Burchard, M. J. Kidd, H. Dai, J. W. Phillips, P. S. Linsley, R. Stoughton, S. Scherer & M. S. Boguski

Rosetta Inpharmatics, Inc., 12040 115th Avenue N.E., Kirkland, Washington 98034, USA

\* These authors contributed equally to this work

**The most important product of the sequencing of a genome is a complete, accurate catalogue of genes and their products, primarily messenger RNA transcripts and their cognate proteins. Such a catalogue cannot be constructed by computational annotation alone; it requires experimental validation on a genome scale. Using ‘exon’ and ‘tiling’ arrays fabricated by ink-jet oligonucleotide synthesis, we devised an experimental approach to validate and refine computational gene predictions and define full-length transcripts on the basis of co-regulated expression of their exons. These methods can provide more accurate gene numbers and allow the detection of mRNA splice variants and identification of the tissue- and disease-specific conditions under which genes are expressed. We apply our technique to chromosome 22q under 69 experimental condition pairs, and to the entire human genome under two experimental conditions. We discuss implications for more comprehensive, consistent and reliable genome annotation, more efficient, full-length complementary DNA cloning strategies and application to complex diseases.**

The initial interpretation of a genome sequence rests upon conclusions derived solely from bioinformatics approaches—*ab initio* gene predictions, homology studies, motif analysis and other non-experimental methods<sup>1–3</sup>. The limitations and fallibility of this process have been discussed<sup>4,5</sup> and one group has concluded<sup>6</sup> that, despite more than 17 years of research effort<sup>7</sup>, precise annotation of every gene in the human genome by computational methods alone is still a distant goal. Bioinformatics analyses of fragmentary experimental data have led to widely varying estimates of the number of human genes<sup>8–10</sup>. Comparative genomics approaches, particularly between human and mouse<sup>11–13</sup>, will help to identify candidate genes and refine their structures, but cannot alone show that a gene is active. Consequently, projects to clone and catalogue ‘full-length’ cDNA clones from human<sup>14</sup> and mouse<sup>15</sup> have been undertaken. Although these projects may capture the complete coding sequences of many genes in time, cDNA cloning fixes a gene product at a particular time and under particular conditions, and thus cannot efficiently reveal the multifunctional nature of a metazoan transcriptome.

Recent work indicates that the human genome may contain fewer genes than anticipated<sup>8,9</sup>, and that frequent alternative splicing might account for much physiological complexity<sup>16–19</sup>. This situation makes it essential to pursue a course that efficiently yields empirical validation of the structures of genes and simultaneously provides an accurate and complete catalogue of their expressed products (mRNA and cognate protein sequences).

We describe a high-throughput, microarray-based experimental method to validate predicted exons, group the exons into genes by co-regulated expression and define full-length mRNA transcripts. The method involves the design and fabrication of ‘exon arrays’ consisting of long (50–60 bases) oligonucleotide probes derived from predicted exons, followed by hybridization with fluorescently labelled cDNAs derived from specific cell lines or normal or diseased tissues. Absolute intensities (measuring cellular abundances) or intensity ratios (measuring differential expression regulation) from hybridized cDNAs are used to identify those probes that represent authentic exons under the conditions tested. In addition, the expression data can define gene boundaries, because adjacent exons that are co-regulated across many conditions are likely to be from the same transcript. For a higher-resolution view of gene

structure, we use ‘tiling arrays’ in which overlapping oligonucleotides are designed to blanket an entire genomic region of interest. This approach can potentially reveal exons not identified by current gene prediction algorithms and provide information about alternative splicing.

We applied the exon array approach to a detailed analysis of human chromosome 22 under 69 pairs of experimental conditions. Tiling arrays were used to refine the structure of new genes discovered by exon analysis. Finally, a preliminary analysis of the entire human genome using exon arrays under two experimental conditions demonstrated the power of being able experimentally to validate hundreds of thousands of exon predictions, anticipating the prospect of analysing the entire human genome to a depth similar to that achieved on chromosome 22.

## Analysis of chromosome 22q using exon arrays

Chromosome 22 was the first human chromosome to be completely sequenced and subjected to exhaustive computational annotation<sup>2</sup>. It has thus served as a benchmark for new computational and experimental methods of analysis<sup>20,21</sup>. We designed a single ink-jet array to monitor the 8,183 exons annotated<sup>2</sup> on chromosome 22q under diverse experimental conditions. Specifically, mRNAs from human cell lines and normal and diseased tissues (Fig. 1) were fluorescently labelled with two colours and hybridized in pairs to 69 individual chromosome 22 exon arrays (see Methods). Figure 2a shows a graphical display of error-weighted log expression ratios<sup>22</sup> for all 8,183 exons across 69 condition pairs. We developed a gene identification algorithm that uses intensity and ratio information to identify exons in a local neighbourhood that are strongly correlated across condition pairs, and then to extend such regions by incorporating other local exons with similar expression behaviour. The resulting 572 groups of co-regulated exons are referred to as expression-verified genes (EVGs). Figure 2b–e shows expanded views of specific regions of chromosome 22. Expression data can be used to confirm the exons and structure of a known gene (Fig. 2b), to identify potential false positive exon predictions (Fig. 2c), to merge UniGene clusters into a single gene (Fig. 2d) and to verify *ab initio* gene predictions experimentally (Fig. 2e).

For a chromosome-wide performance summary, we compared our experimentally derived EVGs to the list of 545 genes annotated

by Dunham *et al.*<sup>2</sup> (Table 1). These annotated genes were divided into four categories (known, related, predicted and *ab initio*) on the basis of the level of experimental support for the predictions. We identified 210 (85%) of the 247 known genes by analysing the expression data from the 69 condition pairs with our gene detection algorithm. The remaining 15% of known genes did not exhibit sufficient differential expression regulation among the conditions tested to enable ratio-based algorithms to verify them. We detected 66% of the related genes and 53% of the predicted genes using our expression regulation criteria. The most interesting result comes from the 325 *ab initio* genes that represented pure Genscan predictions. Dunham *et al.*<sup>2</sup> speculated that only 100 of these predicted transcripts would represent portions of 'real' genes, but we found experimental support for 185 (57%) of the genes in this category.

A few of the EVGs that we detected contained more than one gene. This occurred when adjacent genes were co-regulated across the 69 experimental conditions tested. In most cases, this situation can be addressed by testing additional conditions or by using additional bioinformatics techniques (for example, open reading frame (ORF) analysis, identification of internal polyadenylation sites, and supporting expressed sequence tag (EST) and protein sequence data). In a few cases, a single gene was represented by more than one EVG, indicating possible alternative splicing. Other algorithms are being developed to address this issue.

### Applications of tiling arrays

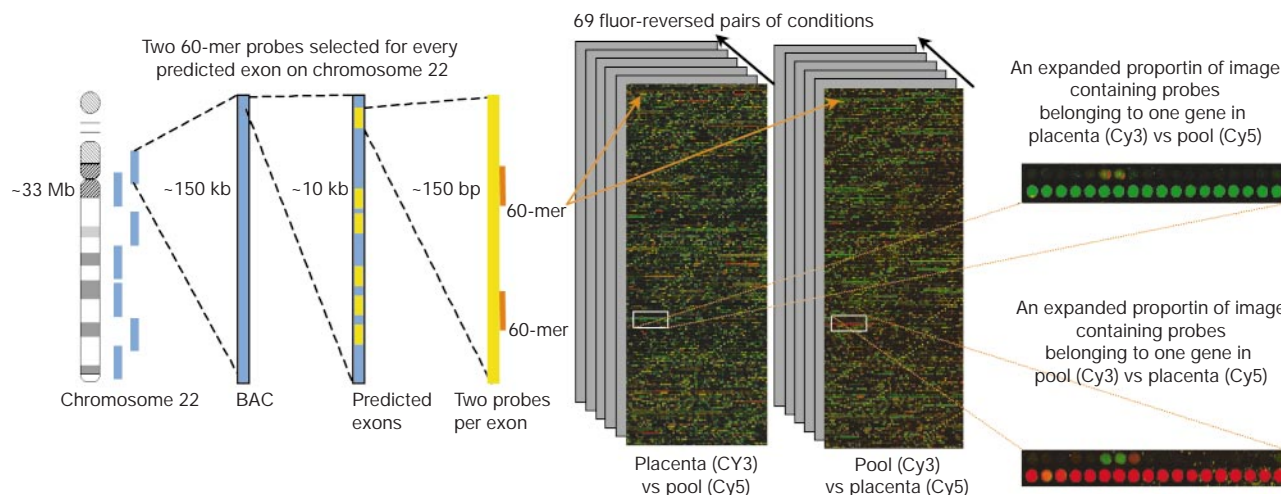
Exon-based gene validation arrays can be limited by the fact that gene prediction programs perform best on 'internal' exons and not very well on initial and terminal exons, or exons that correspond to the 5' and 3' untranslated regions (UTRs) of mRNAs<sup>6</sup>. Oligonucleotide tiling arrays of overlapping probes (Fig. 3) can effectively address this challenge because they are constructed without any *a priori* knowledge of the possible exon content of a genomic sequence. We designed tiling arrays covering both strands of various genomic regions on chromosome 22 defined by EVGs where the underlying gene structure was thought to be incomplete.

Figure 3 shows how the tiling approach was used to refine the structure of the novel testis transcript described in Fig. 2e. We fabricated an ink-jet array that contained 60-mer probes spaced in

10-base-pair (bp) intervals across both strands of the 113-kilobase (kb) bacterial artificial chromosome (BAC) clone containing the EVG of interest. The array was hybridized with fluorescently labelled testis mRNA and the resulting probe intensities were analysed to determine the approximate locations of the exons within this region. For each exon, the hybridization data effectively reduced the search for the intron-exon boundaries to regions of around 20–30 bp. The exact splice junctions can generally be identified within these narrow windows by using common rules (for example, GT-AG consensus sequence and ORF analysis). For the gene shown in Fig. 3, only four of the six exons were correctly predicted by Genscan. Our results extend the 3' UTR by 450 bp and one of the internal coding exons by 102 bases (34 amino acids). These results were confirmed by polymerase chain reaction with reverse transcriptase (RT-PCR) and sequencing (data not shown). The mRNA (GenBank accession no. AF324466) derived from this validated and corrected gene is 1,312 nucleotides long, including a 649-base 3' UTR with a polyadenylation signal at base 1,293. It encodes a 217-residue protein and a BLASTP search revealed only one significant match (*E*-value  $\sim 10^{-15}$ ) to a predicted gene product, CG5280 from the *Drosophila* genome project<sup>23</sup>.

### Human genome scan using exon arrays

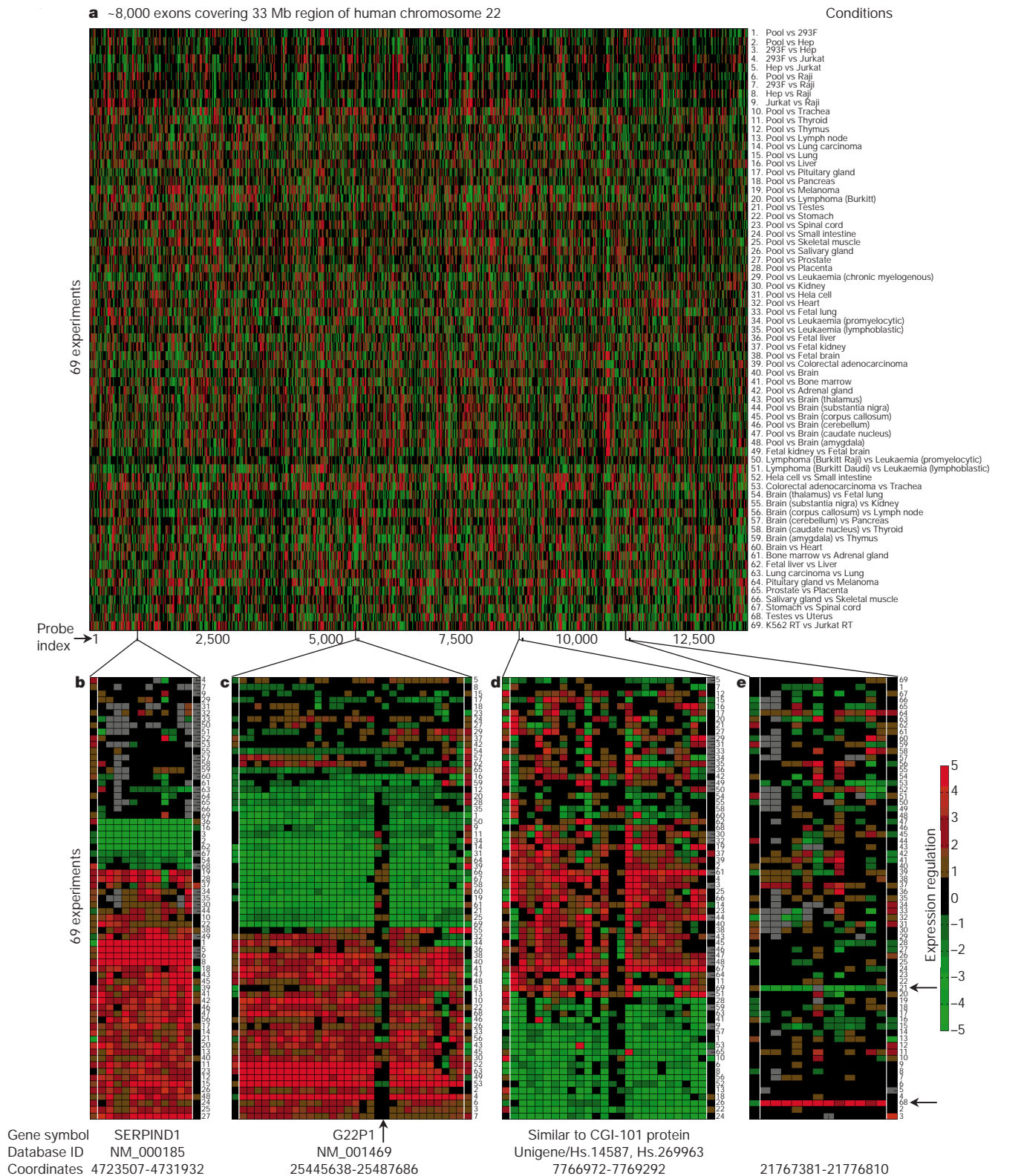
To show that the approach described above can scale to survey the entire human genome, we used the 15 June 2000 version of the Ensembl human genome annotation data set (<http://www.ensembl.org/>)<sup>24</sup> to make 50 arrays containing 1,090,408 oligonucleotide probes representing 442,785 exons predicted by Genscan<sup>25</sup>. Fluorescently labelled cDNAs from a human lymphoma cell line and a colorectal carcinoma cell line were hybridized to the arrays. Analysis of fluorescence intensities from this single pair of experimental conditions provided experimental evidence for 58% of the 78,486 Ensembl confirmed exons. We detected 34% of the 364,299 predicted exons that did not meet the Ensembl 'confirmed' criteria. The false positive rate for this analysis was estimated to be around 5%, from an analysis of a set of negative control probes included on the arrays. A summary of the exons validated by this genome survey is given (see Methods) in Fig. 4 and a full listing is available as Supplementary Information or from the Rosetta website at [www.rii.com](http://www.rii.com).



**Figure 1** Design and fabrication of exon arrays for the predicted exons on human chromosome 22. Two 60-mers were selected from each of 8,183 predicted exons on human chromosome 22q and printed on a single 1 x 3 inch array (~25,000 60-mers). This array was hybridized with 69 pairs of RNA samples using a two-colour hybridization technique. Each experiment was performed in duplicate with a fluor reversal to minimize

possible bias caused by the molecular structure of the Cy3 and Cy5 dyes (138 arrays in total). Red and green spots, as shown in the expanded panels on the right, are probes representing experimentally verified genes (groups of differentially expressed exons that are located next to each other in the genome).

**a** ~8,000 exons covering 33 Mb region of human chromosome 22



**Figure 2** Using expression data from multiple conditions to validate exons and define gene boundaries on chromosome 22. **a**, Pseudocolour image showing error-weighted  $\log_{10}$  expression ratios (red/green) for each of the ~8,000 exons ( $x$ -axis) across the 69 fluor-reversed experiments ( $y$ -axis). A brief description of each experiment is listed on the right side of the image; the numbers (1–69) are reference points for the Table in the Supplementary Information. The 15,511 probes representing the 8,183 predicted exons are arranged linearly across the 33 Mb of chromosome 22. **b**, Expanded region showing a known gene (SERPIND1, NM\_000185). The experiments on the  $y$ -axis have been clustered to emphasize how co-regulation across diverse experiments can be used

group exons into genes. The vertical white lines indicate the boundaries predicted by our gene finding algorithm; numbers on  $y$ -axis indicate experimental conditions. **c**, Expanded region showing a set of co-regulated exons from another known gene (G22P1, NM\_001469), illustrating the detection of potential false positives (arrow) made by the Genscan prediction program. **d**, Expanded region representing an EVG that collapses two Unigene EST clusters (HS.269963 and HS.14587) into a single transcript. **e**, Expanded region showing an EVG containing six exons that are part of a novel testis-expressed transcript (arrows, two experiments involving testis RNA samples).

**Discussion**

Post-genome biology and medicine will increasingly rely on complete and accurate catalogues of human genes, mRNAs and proteins. This ‘parts list’ is currently a patchwork of mostly hypothetical entities with varying degrees of supporting evidence. Computational techniques for sequence annotation provide invaluable clues to gene structure and function but experimental data will be required to provide a full and satisfying picture. Our microarray-based technology represents a comprehensive and consistent

approach to the simultaneous validation of gene predictions and study of the transcriptome under any number of biologically or medically interesting conditions. Our approach is applicable on a genome scale and also on the scale of defining the structure of a single, novel cDNA.

The exon-based approach is well suited to high-throughput screening of diverse cell types, growth conditions and disease states. Differential expression is an important tool for assembling exons into genes. We could detect differential expression for only 15% of the confirmed exons across the human genome with a single condition pair. Clearly, larger data sets will be essential for defining the structures of genes, detecting rarely expressed genes and addressing more complex issues such as alternative splicing. In addition, information from the exon analysis can be used to select genomic regions and samples for comprehensive tiling arrays.

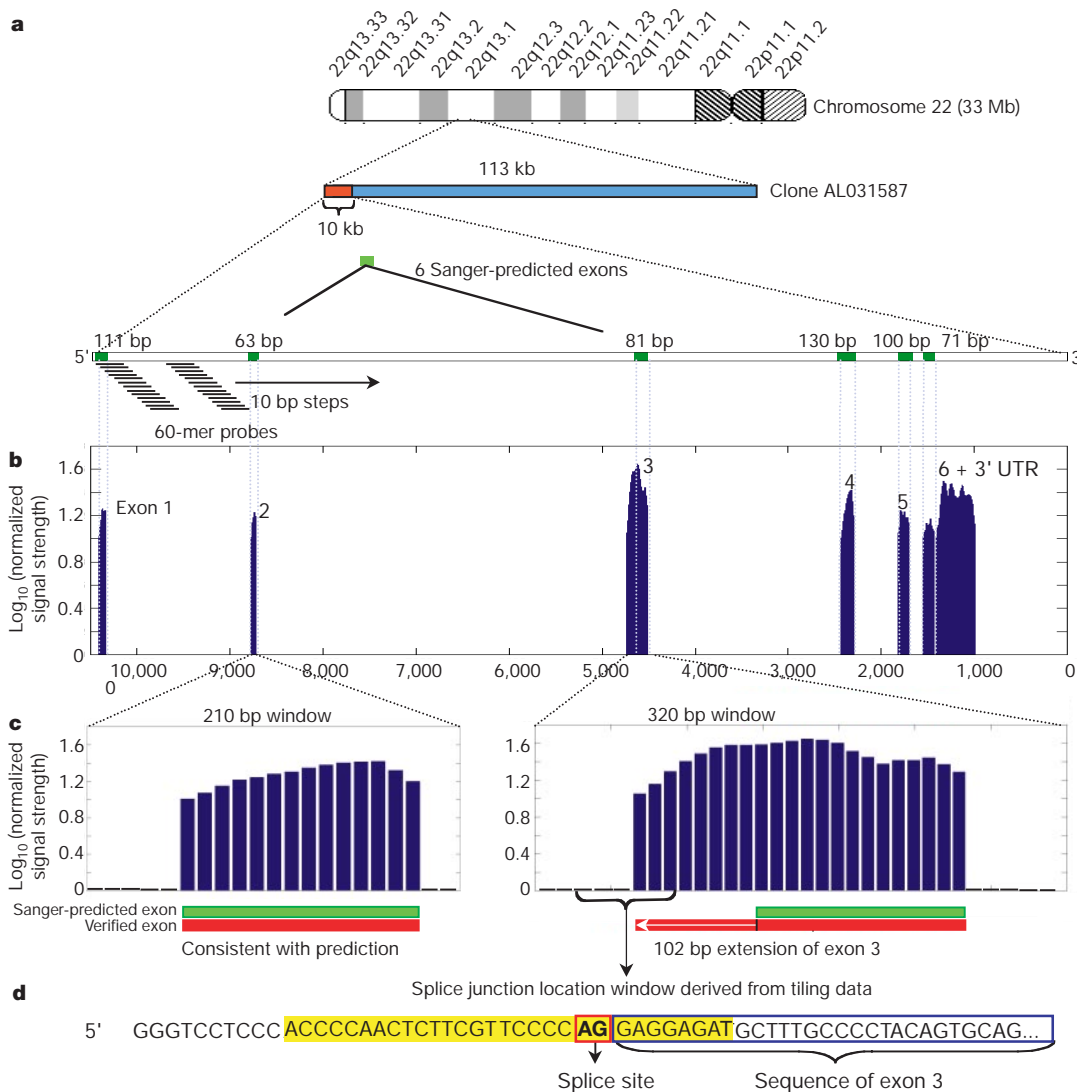
Ambitious efforts to clone and sequence ‘full-length’ cDNAs for the human<sup>14</sup> and mouse<sup>15</sup> genomes have begun with the purpose not only of helping to validate computational gene predictions but also of providing physical reagents for functional and structural geno-

**Table 1 Gene validation summary of human chromosome 22q**

	Annotation from ref. 2	Expression-verified genes (EVGs)	Validation fraction
Known genes*	247	210	85%
Related genes*	150	99	66%
Predicted genes*	148	78	53%
<i>Ab initio</i> genes*	325	185	57%

EVG sequences were searched against current versions of dbEST and nr (www.ncbi.nlm.nih.gov) and significant matches were defined as those having an *E*-value < 10<sup>-20</sup>.

\*Category definitions according to Dunham *et al.*<sup>2</sup>.



**Figure 3** Characterization of a novel testis transcript using tiling arrays. **a**, An EVG discovered in the analysis of chromosome 22 (Fig. 2e) was localized to a 10-kb region at one end of the insert of BAC clone AL031587. Both strands of this 113-kb genomic interval were tiled with 60-mer probes at 10-bp steps. The tiling array was hybridized with RNA from human testis. **b**, Hybridization signals corresponding to tiling probes from this region were filtered and plotted as log<sub>10</sub> values of the normalized signal strengths. Of the

six Genscan predicted exons in this region, two (exons 3 and 6) were at variance with the hybridization data. **c**, Detailed views of tiling data showing one correctly predicted exon and one incorrectly predicted exon. **d**, Typically, tiling data narrow the search window for an intron/exon boundary to 20–30-bp. The exact splice junction is then identified using consensus sequences (GT-AG rule) and ORF information. The exact splice junction can also be determined by sequencing RT-PCR products.



mics. The comprehensive set of EVGs generated by our approach will accelerate these efforts by allowing a more directed cloning strategy. We also expect that hybridization data defining EVGs will be useful in ‘training’ the next generation of gene prediction algorithms, in much the same manner that sequence similarity data enhances *ab initio* predictions in the current state-of-the-art programs. In this way, the maximum value can be realized from the intersection of computational and high-throughput experimental biology.

Our experimental method of annotating the human genome could be rapidly reiterated for updated sequence information from the Human Genome Project, and could easily be extended to the genomes of other organisms. Generating exon and tiling arrays requires only the availability of genomic sequence and exon predictions, from which probes can be rapidly and efficiently synthesized onto an array. The flexibility and short time scale for designing

and fabricating exon and tiling arrays using the ink-jet platform could substantially accelerate gene discovery.

Finally, our approach could be useful in the identification and analysis of genes underlying complex diseases. Genetic linkage studies of polygenic traits typically yield a dozen loci, each up to 20–30 megabases long. It is feasible to construct tiling arrays across all loci and probe them with mRNA samples from relevant normal and diseased tissues to ascertain both gene content and activity. Such analyses may provide not only more direct routes to the culpable genes, but also have the potential to uncover regulatory mutations by observed alterations in gene activity. □

Methods

Sources of predicted exons

To analyse chromosome 22q, we designed a single ink-jet oligonucleotide microarray to represent 8,183 sequences that had been identified or confirmed as having coding potential (Sanger Centre). We used two sources of information: 6,650 Genscan-predicted exon sequences, and 3,381 validated exon sequences identified by aligning the first complete version of the human chromosome 22 sequence with sequences from experimentally validated transcripts<sup>2</sup>. Of this set of 10,031 exons, 1,847 had coordinates identical to those of other exons and were removed from the sequence pool. The remaining 8,183 exon sequences were subjected to an oligonucleotide design process to identify the two best candidate probes for a given exon sequence (see below). For the whole-genome exon scan, we designed ink-jet oligonucleotide microarrays to 442,785 predicted exons selected from the publicly available assembled sequence in the Ensembl database as of 15 June 2000. Specifically, we selected 554,202 non-redundant sequences from an initial set of 628,635 Genscan predicted exons<sup>24</sup>. We removed 111,417 more sequences from the list after they were flagged by the RepeatMasker algorithm (<http://ftp.genome.washington.edu/cgi-bin/RepeatMasker>).

Probe selection for the exon-scanning arrays

For each of the predicted exons, we selected the top two 60-mers using an algorithm that takes into account binding energies, base composition, sequence complexity, cross-hybridization binding energies and secondary structure. For exon sequences of 60 nucleotides or less, we designed a single probe consisting of the entire exon sequence. For the 8,183 predicted exons on chromosome 22, 15,511 60-mers were selected and printed on a single array. For the whole-genome exon arrays, we selected 1,090,408 60mers to represent the 442,785 GenScan predicted exons from the Ensembl database. For 78,486 of the exons annotated as ‘confirmed’, the reverse-complement probes were also selected and placed next to the regular probes on the array as negative controls.

Probe selection for tiling arrays

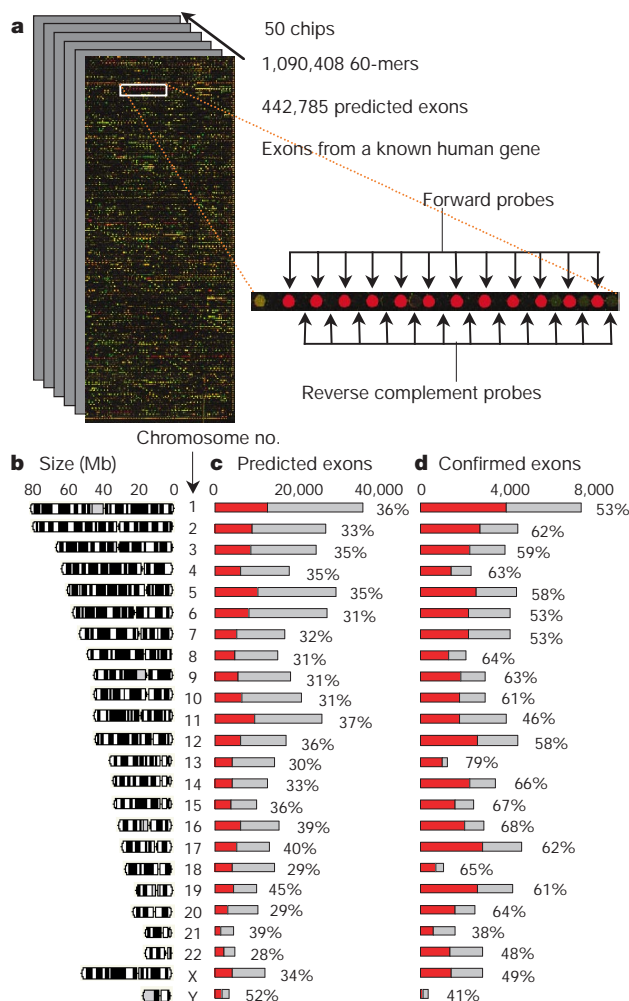
In the tiling experiment described in Fig. 3, 60-mer probes were placed in 10-bp intervals across a 113.8-kb region of chromosome 22 containing the novel testis transcript described in Fig. 2e (BAC clone AL031587). The reverse complements for each of the tiling probes were also included on the array to allow transcripts on either strand to be detected. The genomic sequences used in the tiling experiments were repeat-masked before probe selection but no other exclusionary filters were applied.

Array synthesis

We synthesized the oligonucleotide arrays on 1 × 3-inch glass slides using ink-jet technology<sup>26</sup>. The phosphoramidite monomers were delivered by a standard ink-jet printer head to defined positions on a glass surface containing exposed hydroxyl groups. The remaining synthesis steps are similar to traditional oligonucleotide synthesis. Using this approach, up to 25,000 different 60-mers can be synthesized on a single slide. Around 1,000 ‘gridline’ probes (5’ CCTATGTGACTGGTCGATGCTACTA 3’) are placed around the perimeter of each array. Fluorescently labelled synthetic oligonucleotides complementary to the control probes are included in all hybridizations. Arrays based on Rosetta designs were purchased from Agilent Technologies.

Preparation of labelled cDNA

We used the following human cell lines: Jurkat (T lymphocyte, ATCC no. TIB-152), K562 (chronic myelogenous leukaemia, ATCC no. CCL-243), Raji (Burkitt’s lymphoma, ATCC no. CCL-86), Colo (colorectal adenocarcinoma, ATCC no. CCL-220), 293 (embryonic kidney, ATCC no. CRL-1573.1) and HepG2 (hepatocellular carcinoma, ATCC no. CRL-11997). Poly-A<sup>+</sup> RNA (mRNA) was isolated from each of the cytoplasmic RNA samples as described<sup>27</sup>. The ‘pool’ RNA sample described in Fig. 2 contains an equal mixture of four human cell lines (Jurkat, K562, Raji and Colo). The 41 mRNA samples from the human tissues described in Fig. 2 were purchased from commercial sources and are described at [www.rii.com/Publications](http://www.rii.com/Publications). For a single hybridization, we combined 1.5 µg of mRNA with 1.0 µg of random 9-mers and incubated the mixture for 10 min at 70 °C, 5 min at 4 °C and 10 min at 22 °C. To this mixture we added 0.5 mM amino-allyl dUTP (Sigma A-0410), 0.5 mM dNTP, 1 × RT buffer, 5 mM MgCl<sub>2</sub>, 10 mM DTT and 200 units of Superscript (GibcoBRL), bringing the final reverse transcription reaction volume to 40 µl. This reverse transcription reaction was incubated for 20 min at 42 °C and the RNA was hydrolysed by adding 20 µl EDTA + NaOH and incubating at 65 °C for 20 min. The reaction was



**Figure 4** Whole-genome scan for validating predicted exons. **a**, Fifty 1 × 3-inch ink-jet arrays were used to test 442,785 Genscan predicted exons under two conditions. For each predicted exon, the best one or two 60-mer probes were selected, resulting in the set of 1,090,408 probes which were distributed over 50 arrays (~25,000 60-mers per array). The arrays also included 110,000 reverse complement probes and 48,500 control probes. The arrays were hybridized with Cy-3 or Cy-5 labelled mRNA from two human cell lines (Raji and Colo). Enlarged image, probes representing exons from a known gene with alternating forward and reverse complement probes. All experiments were performed in duplicate with a fluor reversal (100 arrays total). **b**, The sizes of the 24 human chromosomes (left). **c**, The number of predicted exons that were experimentally verified (red bars) for each of the chromosomes. Grey bars, number of predicted exons on each chromosome. **d**, A similar analysis for the confirmed exons across the human genome.

neutralized by adding 20  $\mu$ l of 1M Tris-HCl pH 7.6. We concentrated the resulting amino-allyl labelled single-stranded cDNA using a Microcon-30 (Millipore), and coupled it to Cy3 or Cy5 dye using a CyDye kit (Amersham Pharmacia Q15108). The per cent dye incorporation and total cDNA yield were determined spectrophotometrically. Pairs of Cy5/Cy3-labelled cDNA samples were combined and hybridized as described<sup>22</sup>.

### Analysis and visual display of expression data

Array images were processed as described<sup>22</sup> to obtain background noise, single channel intensity and associated measurement error estimates. Expression changes between two samples were quantified as  $\log_{10}$  (expression ratio) where the 'expression ratio' was taken to be the ratio between normalized, background-corrected intensity values for the two channels (red and green) for each spot on the array. An error model for the log ratio was applied<sup>22</sup> to quantify the significance of expression changes between two samples. The colour displays in Fig. 2 show  $\log_{10}$ (expression ratio) as red when the red channel is upregulated relative to the green channel, green when the red channel is downregulated relative to the green channel, black when  $\log_{10}$ (expression ratio) is close to zero, and grey when data from one or both of the channels for a given probe are unreliable.

### Identifying EVGs by co-regulation

Exons were grouped into EVGs by a two-step gene identification algorithm. First, each probe was assigned a similarity measure, based on the moving average (using a window size equal to six probes) of pair-wise Pearson correlation coefficients between the log ratios of probe intensities in neighbouring exons. Probes with correlation coefficients above 0.5 in a given window were selected as seeds for EVGs. The 0.5 threshold and window size were determined empirically by training the model on a subset of the known chromosome 22 genes. Second, probes neighbouring a seed region were merged into the region if the pair-wise correlation coefficients between the neighbouring probe and the average in the seed region exceeded 0.5. This process continued, allowing for gaps between probe pairs to account for failed probes and/or false exon predictions (gaps were not allowed to exceed five probes), until no probes flanking the candidate region met the significance threshold of correlation with the exon cluster. The final exon clusters resulting from the gene detection algorithm were identified as an EVG. Not all condition pairs (rows) were considered in forming EVGs. Elements in a given row had to have significant  $P$  values ( $\leq 0.01$ ) to be included in the analysis. Once an EVG was formed, the colour display (as in Fig. 2) was updated by reordering the condition pairs according to a hierarchical clustering algorithm, as described<sup>28</sup>.

### Annotation of EVGs

Predicted transcripts for all EVGs identified from the chromosome 22 exon data across the 69 condition pairs were formed by combining the individual exons into a single sequence. Each of these sequences was searched against dbEST and the NR protein databases using gapped BLASTN and BLASTX (www.ncbi.nlm.nih.gov), respectively, to determine the extent to which the EVG sequences were similar to other sequence data. We declared sequences similar if the corresponding  $E$ -value for the alignment was less than  $10^{-9}$ , using default parameters for gapped BLAST. BLAST results were used to determine the degree of sequence support defining a predicted transcript. These results were also used to determine the degree of existing sequence support for each of the EVGs detected from the chromosome 22 exon arrays.

### Quantitative analysis of whole-genome exon data

We used an intensity-based algorithm to verify predicted exons experimentally across the entire human genome. Specifically, we used raw intensity measurements for the forward-strand (FS) probes and the corresponding raw intensity measurements for the reverse-complement (RC) probes in conjunction with the respective standard deviations of those measurements to determine the significance of the FS probe intensities. We controlled for nonspecific cross-hybridization using RC probes, given that the reverse complement of a DNA sequence has equivalent sequence complexity to the forward strand sequence with respect to a variety of measures (such as GC content and GC trend). An exon was called 'present' if the intensity difference between an FS probe and the RC probe had  $P < 0.01$  in either the red or green channel, and if the FS probe intensity had a  $P < 0.01$  for being above background in the channel where the difference was considered most significant.  $P$  values were calculated using a  $t$ -test applied to the difference of the mean pixel intensities and to the difference of the mean FS/background intensities.

These single channel exon detection methods were applied only to those exons in which reverse-complement probes were designed. In the remaining cases, the significance of the single channel intensities was determined using the above-background criterion described above. We applied a correction to the detection percentages given for the predicted exons listed in Fig. 4, based on false positive estimates for above-background calls that were determined using the FS/RC probe intensity difference calls for the confirmed exons. Error

models used in this analysis to assess ratio significance were as described<sup>28</sup>. Of the 88,374 confirmed exons represented on the genome-wide exon arrays, 78,486 had corresponding RC probes. To assess the rate of false positives expected in the single-channel assessments, we used a similar detection procedure to determine the number of RC probe intensity measurements that were significantly greater than the corresponding FS probe intensity. Our results indicate that the false positive rate of detection using the single channel method was  $\sim 5\%$ .

Received 28 November 2000; accepted 9 January 2001.

1. The *C. elegans* Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282**, 2012–2018 (1998).
2. Dunham, I. *et al.* The DNA sequence of human chromosome 22. *Nature* **402**, 489–495 (1999).
3. Rubin, G. M. *et al.* Comparative genomics of the eukaryotes. *Science* **287**, 2204–2215 (2000).
4. Wheelan, S. J. & Boguski, M. S. Late-night thoughts on the sequence annotation problem. *Genome Res.* **8**, 168–169 (1998).
5. Boguski, M. S. Biosequence exegesis. *Science* **286**, 453–455 (1999).
6. Guigo, R., Agarwal, P., Abril, J. F., Burset, M. & Fickett, J. W. An assessment of gene prediction accuracy in large DNA sequences. *Genome Res.* **10**, 1631–1642 (2000).
7. Claverie, J. M. Computational methods for the identification of genes in vertebrate genomic sequences. *Hum. Mol. Genet.* **6**, 1735–1744 (1997).
8. Ewing, B. & Green, P. Analysis of expressed sequence tags indicates 35,000 human genes. *Nature Genet.* **25**, 232–234 (2000).
9. Roest Croliius, H. *et al.* Estimate of human gene number provided by genome-wide analysis using *Tetraodon nigroviridis* DNA sequence. *Nature Genet.* **25**, 235–238 (2000).
10. Liang, F. *et al.* Gene index analysis of the human genome estimates approximately 120,000 genes. *Nature Genet.* **25**, 239–240 (2000).
11. Makalowski, W. & Boguski, M. S. Evolutionary parameters of the transcribed mammalian genome: an analysis of 2,820 orthologous rodent and human sequences. *Proc. Natl Acad. Sci. USA* **95**, 9407–9412 (1998).
12. Batzoglu, S., Pachter, L., Mesirov, J. P., Berger, B. & Lander, E. S. Human and mouse gene structure: comparative analysis and application to exon prediction. *Genome Res.* **10**, 950–958 (2000).
13. Marshall, E. Public-private project to deliver mouse genome in 6 months. *Science* **290**, 242–243 (2000).
14. Strausberg, R. L., Feingold, E. A., Klausner, R. D. & Collins, F. S. The mammalian gene collection. *Science* **286**, 455–457 (1999).
15. The RIKEN Genome Exploration Research Group Phase II Team and the FANTOM Consortium. Functional annotation of a full-length mouse cDNAs collection. *Nature* **409**, 685–690 (2001).
16. Hanke, J. *et al.* Alternative splicing of human genes: more the rule than the exception? *Trends Genet.* **15**, 389–390 (1999).
17. Mironov, A. A., Fickett, J. W. & Gelfand, M. S. Frequent alternative splicing of human genes. *Genome Res.* **9**, 1288–1293 (1999).
18. Black, D. L. Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology. *Cell* **103**, 367–370 (2000).
19. Brett, D. *et al.* EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. *FEBS Lett.* **474**, 83–86 (2000).
20. de Souza, S. J. *et al.* Identification of human chromosome 22 transcribed sequences with ORF expressed sequence tags. *Proc. Natl Acad. Sci. USA* **97**, 12690–12693 (2000).
21. Penn, S. G., Rank, D. R., Hanzel, D. K. & Barker, D. L. Mining the human genome using microarrays of open reading frames. *Nature Genet.* **26**, 315–318 (2000).
22. Roberts, C. J. *et al.* Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. *Science* **287**, 873–880 (2000).
23. Adams, M. D. *et al.* The genome sequence of *Drosophila melanogaster*. *Science* **287**, 2185–2195 (2000).
24. Hubbard, T. & Birney, E. Open annotation offers a democratic solution to genome sequencing. *Nature* **403**, 825 (2000).
25. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**, 78–94 (1997).
26. Blanchard, A. P., Kaiser, R. J. & Hood, L. E. High-density oligonucleotide arrays. *Biosens. Bioelectron.* **6/7**, 687–690 (1996).
27. Marton, M. J. *et al.* Drug target validation and identification of secondary drug target effects using DNA microarrays. *Nature Med.* **4**, 1293–1301 (1998).
28. Hughes, T. R. *et al.* Functional discovery via a compendium of expression profiles. *Cell* **102**, 109–126 (2000).

Supplementary Information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

### Acknowledgements

We thank S. H. Friend for encouragement and support and J. Rine, M. V. Olson, C. Roberts and T. Hughes for critical readings of the manuscript.

Correspondence and requests for materials should be addressed to M.S.B. (e-mail: [msb@rii.com](mailto:msb@rii.com)).