nature genetics

Genome-wide analysis of mouse transcripts using exon microarrays and factor graphs

Brendan J Frey[1,2,6], Naveed Mohammad[2,6], Quaid D Morris[1,2,6], Wen Zhang[2,3,6], Mark D Robinson[1,2], Sanie Mnaimneh[2], Richard Chang[2], Qun Pan[2], Eric Sat[4], Janet Rossant[3,4], Benoit G Bruneau[3,5], Jane E Aubin[3], Benjamin J Blencowe[2,3] & Timothy R Hughes[2,3]

Recent mammalian microarray experiments detected widespread transcription and indicated that there may be many undiscovered multiple-exon protein-coding genes. To explore this possibility, we labeled cDNA from unamplified, polyadenylation-selected RNA samples from 37 mouse tissues to microarrays encompassing 1.14 million exon probes. We analyzed these data using GenRate, a Bayesian algorithm that uses a genome-wide scoring function in a factor graph to infer genes. At a stringent exon false detection rate of 2.7%, GenRate detected 12,145 gene-length transcripts and confirmed 81% of the 10,000 most highly expressed known genes. Notably, our analysis showed that most of the 155,839 exons detected by GenRate were associated with known genes, providing microarray-based evidence that most multiple-exon genes have already been identified. GenRate also detected tens of thousands of potential new exons and reconciled discrepancies in current cDNA databases by 'stitching' new transcribed regions into previously annotated genes.

Mammalian genome and transcript sequencing efforts indicate that most protein-coding genes have already been identified[1]. But microarray-based analyses suggest that polyadenylated transcripts are produced from a considerably larger proportion of the genome, including regions that are conserved and seem to be noncoding, as well as regions that contain potential coding exons[2].

To reconcile this discrepancy, we reasoned that much of the functional mammalian transcriptome could be rapidly identified and characterized by surveying exon expression across multiple normal tissues, because most known genes consist of exons and are expressed at different levels across tissues and developmental states. We designed microarrays[3] containing 1,140,421 sequences selected from the combined outputs of five exon-finding and gene-like sequence detection algorithms (GenScan[4], HMMGene[5], GrailEXP[6], BlastX and BlastN) applied to the mouse genome[7]. We expected the resulting set of putative exons to have broad coverage because we used low stringency settings for the search algorithms (**Supplementary Methods** online). The resulting set of putative exons was more than five times larger than the set of exons in known genes. We analyzed data from a previous study[8] to select twelve tissue pools, encompassing 37 different tissue samples (**Table 1**), in a way that maximizes both differential expression between pools and global activity in every pool (**Supplementary Methods** online). We analyzed wild-type mouse tissues, rather than cell lines, to ensure that genes contained in the pools were expressed under normal physiological conditions. To achieve high fidelity, we hybridized unamplified first-strand fluor-labeled cDNA obtained from polyadenylation-purified mRNA primed with oligo-dT and random nonamers (**Supplementary Methods**

Table 1 Compositions of the 12 mRNA pools analyzed

Pool	Composition (mRNA per array hybridization)
1	Heart (2 μg), skeletal muscle (2 μg)
2	Liver (2 μg)
3	Whole brain (1.5 μg), cerebellum (0.48 μg), olfactory bulb (0.15 μg)
4	Colon (0.96 μg), intestine (1.04 μg)
5	Testis (3 μg), epididymis (0.4 μg)
6	Femur (0.9 μg), knee (0.4 μg), calvaria (0.06 μg), teeth and mandible (1.3 μg), teeth (0.4 μg)
7	15-d embryo (1.3 μg), 12.5-d embryo (12.5 μg), 9.5-d embryo (0.3 μg), 14.5-d embryo head (0.25 μg), embryonic stem cells (0.24 μg)
8	Digit (1.3 μg), tongue (0.6 μg), trachea (0.15 μg)
9	Pancreas (1 μg), mammary gland (0.9 μg), adrenal gland (0.25 μg), prostate gland (0.25 μg)
10	Salivary gland (1.26 μg), lymph node (0.74 μg)
11	12.5-d placenta (1.15 μg), 9.5-d placenta (0.5 μg), 15-d placenta (0.35 μg)
12	Lung (1 μg), kidney (1 μg), adipose tissue (1 μg), bladder (0.05 μg)

[1]Electrical and Computer Engineering, University of Toronto, 10 King's College Rd., Toronto, Ontario M5S 3G4, Canada. [2]Banting and Best Department of Medical Research, University of Toronto, 112 College St., Toronto, Ontario M5G 1L6, Canada. [3]Medical Genetics and Microbiology, University of Toronto, 1 King's College Ct., Toronto, Ontario M5S 3G4, Canada. [4]Samuel Lunenfeld Research Institute, Mount Sinai Hospital, 600 University Avenue, Toronto, Ontario M5G 1X5, Canada. [5]The Hospital for Sick Children, 555 University Ave., Toronto, Ontario M5G 1X8, Canada. [6]These authors contributed equally to this work. Correspondence should be addressed to T.R.H. (t.hughes@utoronto.ca) or B.J.F. (frey@psi.toronto.edu).

Published online 28 August 2005; doi:10.1038/ng1630

online). This technique generated a data matrix of 1,140,421 exon expression profiles across the 12 tissue pools. **Figure 1a** shows a subset of the expression data. The data are available from our project website, along with an interface (linked to the University of California Santa Cruz genome browser[9]) that enables browsing of microarray data, *ab initio* exon predictions, mappings of known genes and genes predicted by our analysis.

Detection of transcripts in exon and genome tiling data is influenced by cross-hybridization, probe sensitivity, probe noise and experimental procedures, among other variables. Previous analyses detected transcripts by applying thresholds to individual signal intensities, correlations of coregulation patterns in multiple samples, the number of consecutive probes that constitute a 'hit' and genomic distances between probes[10–15]. Substantial increases in detection sensitivity can be gained by analyzing multiple samples jointly. In particular, because most multiple-exon protein-coding mammalian genes vary in expression to some extent across tissues[8], a subset of similar expression profiles from probes derived from putative exons that are close to each other in the genome can be taken as evidence of a functional transcript[10]. A disadvantage of previous applications of this approach is that decisions to link putative exons are irreversible. In particular, a decision to assign a probe to a particular transcript removes the probe from further consideration, even if another transcript that is better suited to the probe emerges later in the analysis.

To carry out a global analysis of our microarray data, we derived a genome-wide scoring function that describes relationships between hidden variables and expression profiles. **Figure 1b** provides a graphical depiction of the relationships between n microarray probe signals, each containing 12 expression levels, and hidden

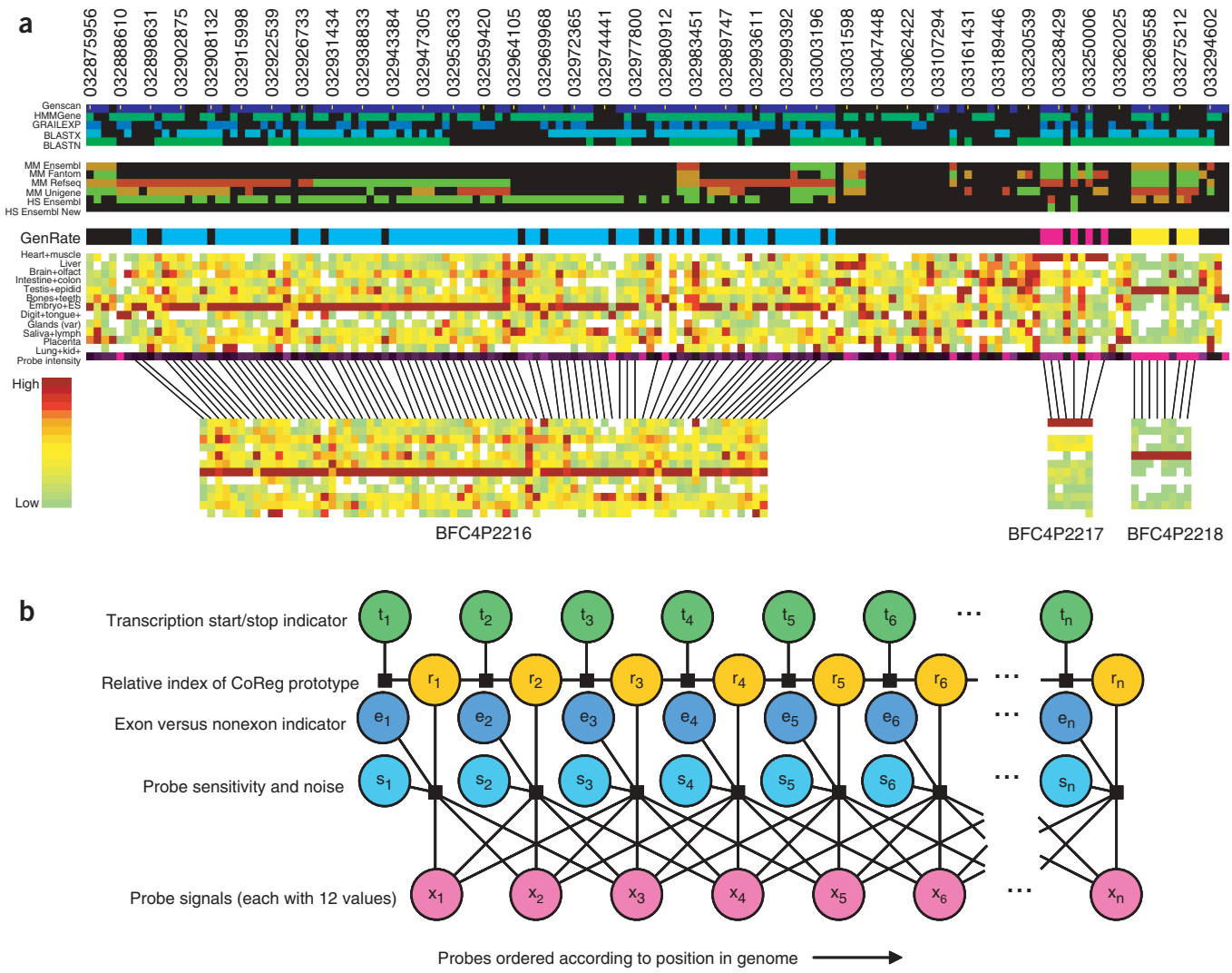

Figure 1 Example of results and illustration of analysis method. (**a**) Sample of exon-resolution data and GenRate output from the positive strand of chromosome 4, map positions 32833512–33300999 (build 33) from left to right. Colored rows indicate the origin of the exon prediction, sequence matches to cDNAs in six databases, the expression data (scaled from minimum to maximum), the maximum log-probe intensity and the GenRate-predicted genes at 2.7% exon FDER. A change in color of a cDNA database match indicates the beginning of a different transcript. Similar views for the entire data set are available at our project website. This example shows that GenRate can correctly connect together erroneously disjoined sequences in gene databases and that coregulation across tissues can be more useful than probe intensity (purple or black track) for detecting genes. HS, human; MM, mouse. (**b**) The factor graph that GenRate uses to find CoRegs in microarray data. Each black box corresponds to a local scoring function that depends on nearby hidden and observed variables, as indicated.

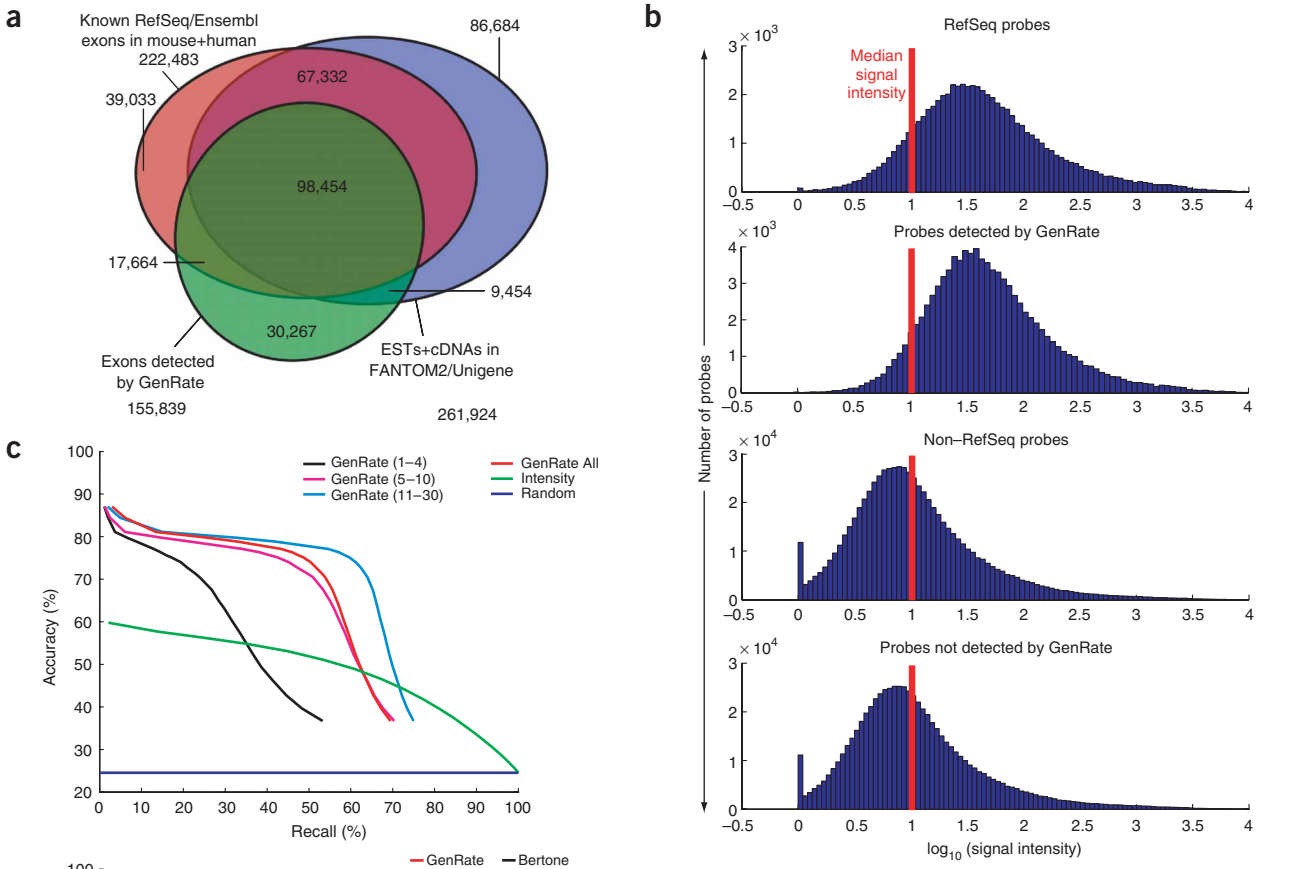

Figure 2 GenRate detects exons with high sensitivity and high specificity. (**a**) A comparison of exons detected by GenRate with mouse and human exons in RefSeq and Ensembl, in addition to exons in the FANTOM2 and Unigene databases. (**b**) Distributions of maximum probe signal intensities for RefSeq exons, probes detected by GenRate, non-RefSeq exons and probes not detected by GenRate. GenRate detects many probes that have low intensity but correspond to known exons and rejects many probes that have high intensity but do not correspond to known exons. (**c**) Accuracy versus recall of GenRate for various gene size categories (number of exons), and the method of thresholding the probe intensity. A comparison with a previously reported system[15] (Bertone) using closely corresponding regions of the mouse and human genomes (Chromosome X) shows that GenRate achieves higher accuracy and recall. The portion of each recall level expected to be due to false detections is indicated for several points on the plots.

variables, including transcription start and stop sites; relative transcript length; a true or false flag for each putative exon; and the sensitivity, cross-hybridization level and additive noise level for each probe. This network is formally called a factor graph[16], and the global score (probability distribution) is equal to the sum (product) of a large number of local scores (probability functions). Each local score reflects how well neighboring observations and hidden variables match. Our technique, called GenRate (generative model for finding and rating transcripts), uses the max-product algorithm to efficiently find the globally optimum score (B.J.F., Q.D.M. & T.R.H., unpublished results) and identifies sets of probe signals, called CoRegs, that represent coregulated transcription. By maximizing a global scoring function, GenRate achieves higher sensitivities than standard clustering techniques (**Supplementary Methods** online).

To validate the reliability of the predictions made by GenRate, we used a permutation test (*i.e.*, randomly reordering the probes) to estimate exon and CoReg false detection rates (FDERs), the fraction of detections that are expected to be false. To limit effects of cross-hybridization noise, we applied GenRate to the 837,251 probes that map uniquely to build 33 of the mouse genome. By varying GenRate's sensitivity, we obtained exon and CoReg FDERs varying from 0.13% to 32% and from 0.2% to 37%, respectively.

To test GenRate's ability to recover previously annotated exons, we compared our predictions with exons mapped from human and mouse genes in six cDNA databases, as well as transcripts detected in a recent human liver microarray analysis[15]. At a stringent exon FDER of 2.7%, GenRate detected 155,839 exons (4,186 expected false detections) comprising 12,145 CoRegs. GenRate detected 64% of the exons in the 17,577 RefSeq Golden Path mouse genes and identified

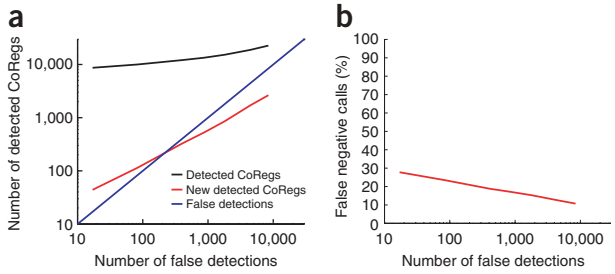

Figure 3 Performance of GenRate on detecting genes. (**a**) False detection analysis. The number of CoRegs identified in the randomized probe data (horizontal axis; average over ten repetitions) and the original probe order (vertical axis) is plotted. (**b**) False negative analysis. For each false detection level (horizontal axis), the fraction of false negative calls among RefSeq genes (vertical axis) is plotted.

70,913 putative new exons. We next expanded our comparison set to include all mouse and human genes in the RefSeq and Ensembl cDNA databases (**Fig. 2a**). GenRate detected 116,118 (52%) of the exons in these databases and identified 39,721 putative new exons. We also expanded our comparison set to include all previous annotations, including poorly characterized expressed-sequence tags (ESTs) and cDNAs in the FANTOM2 and Unigene databases (**Fig. 2**).

Notably, GenRate detected known exons whose probe signal intensities were below the median intensity and rejected putative exons whose probe signals had high intensity but were not coregulated (**Fig. 2b**). We compared the sensitivity and specificity of GenRate with results from a recent study of expression in human liver[15]. At an FDER of 2.9%, their system identified 13,889 exon-size transcripts, 4,931 of which corresponded to previously annotated exons. At an FDER of 2.7%, GenRate detected ~11 times more exons (155,839) and confirmed ~24 times more previously annotated exons (116,118).

We also estimated accuracy using the fraction of detected exons that map to the reference set of RefSeq genes. **Figure 2c** shows the accuracy of GenRate as a function of the fraction of RefSeq exons that are detected (recall), for various sizes of genes. As expected, the recall of GenRate for exons in short genes (<5 exons) was low, because there is less evidence of coregulation. **Figure 2c** also shows the accuracy versus recall obtained by applying a threshold to the maximum intensity for each probe. For all but high levels of recall, where false detections are expected to dominate predictions for all methods, GenRate had substantially higher accuracy than intensity thresholding. We compared the accuracy versus recall of our system with the previously reported system[15] on the X chromosome. GenRate achieved higher accuracy over a much wider range of recall levels (**Fig. 2c**) and achieved higher recall levels with a much lower fraction of expected false positives. Intensity thresholding in our data also achieved substantially higher accuracies than the previously reported system[15], partly because we used a wider selection of tissues.

We next studied all CoRegs detected by GenRate and how they compared with RefSeq genes. At an exon FDER of 2.7%, GenRate detected 12,145 CoRegs, of which 412 (3.4%) were expected to be false detections. **Figure 1a** shows a sample of the output at this FDER and shows two general trends: (i) long transcripts, which are the most difficult to clone, could be identified by this approach; and (ii) coregulation of expression among adjacent probes yielded substantially different predicted transcripts than would be identified by probe intensity level alone. The mean and median number of exons per CoReg were 12.8 and 10, respectively, and the mean and median genomic lengths were 67,592 bp and 29,483 bp, respectively. GenRate detected 11,395 (51%) RefSeq genes, including 8,121 (81%) of the 10,000 RefSeq genes most highly expressed in our data.

Despite the high sensitivity of our system, we did not detect a substantial number of CoRegs consisting entirely of exons not included in any of the databases (**Fig. 3a**). At an exon FDER of 2.7%, only 332 of the CoRegs were entirely new and only 96 of these did not overlap substantially with RepeatMasker sequence (*i.e.*, these CoRegs contain less than 10% of exons that map to RepeatMasker sequence). On average, 83 CoRegs detected in the randomly permuted data consisted entirely of new exons, suggesting that most, if not all, of the 96 new CoRegs not found in RepeatMasker are false detections. To confirm this prediction, we tested 35 of them by RT-PCR, using primers that bridge putative exons and distinguish spliced transcripts (**Supplementary Table 1** online). For 18 of these, we obtained products of some form after repeated attempts, but sequencing confirmed in all cases that the product was aberrant amplification of either genomic sequence or nontargeted highly expressed

Figure 4 New exons detected by GenRate and associated with RefSeq Golden Path genes are categorized by 3′ or 5′ extensions of known genes, bridges that join together known genes, new exons that map to an EST or cDNA in the FANTOM2 or Unigene database, new exons that can be stitched together with the known gene by a previously detected EST or cDNA, and new exons that do not map to any previously detected sequences. The expected number of false detections is 4,186. This analysis was repeated for new exons that were detected among the probes with maximum signal intensity above the 90th percentile. Among these exons, the fraction of completely new exons decreased and the fraction of new exons that are confirmed by ESTs or cDNAs that overlap with known genes increased.

mRNAs. In contrast, we obtained correct RT-PCR products for >75% of known genes from the same samples in the first attempt (ref. 8 and data not shown), indicating that our RT-PCR technique was reliable.

How comprehensive is the set of CoRegs predicted by GenRate? **Figure 3b** shows the fraction of RefSeq genes not detected by GenRate (*i.e.*, the rate of false negative calls) for the same numbers of false detections shown in **Figure 3a**, among the 10,000 RefSeq genes that were most highly expressed in our data. The low false negative rates in combination with the lack of a significant number of new CoRegs detected by GenRate provides compelling evidence that almost all the multiple-exon genes with expression in the 37 tissue pools we studied are already known.

We next examined the relationship between exons detected by GenRate and 17,577 well-characterized RefSeq Golden Path genes. **Figure 4** shows the number of detected exons in each of several categories, including extensions of RefSeq genes, bridges joining RefSeq genes, new exons in RefSeq genes that map to cDNA or EST databases (FANTOM2, Unigene and Ensembl mouse and human) and completely new exons in RefSeq genes. To be especially stringent in making predictions, we repeated this analysis for exons detected by GenRate whose maximum probe intensities were above the 90th percentile.

Two lines of evidence indicate that most of the new exons that we identified are valid. First, there is a bias against new exons internal to RefSeq genes (**Fig. 4**), where errors or omissions are least likely, and a corresponding bias toward new exons flanking cDNAs in Unigene or FANTOM2, which are most likely to be incomplete. Second, we verified new exons by RT-PCR experiments. For example, CoReg BF_C4_2262 (**Fig. 1a**) is fragmented in current mouse cDNA and EST databases; virtually all the exons in this CoReg are contained in a single transcript of >11 kb (**Supplementary Table 2** and **Supplementary Fig. 1** online), which seems to be the mouse homolog of Midasin, the largest gene in yeast[17], and which has not been completely identified by comparison to its human counterpart (**Fig. 1a**). More than half of the CoRegs included at least one exon that did not map to the best-matching cDNA, and so our analysis provides a revised view of the potential exon composition of mammalian genes.

Many of the new exons detected by GenRate reconcile discrepancies in current gene databases. For example, GenRate detects 3,266 non-RefSeq exons that are internal to RefSeq genes and are confirmed by sequences in the FANTOM2 or Unigene databases. There are also examples where GenRate detects a CoReg that bridges together distinct, neighboring cDNAs. Because such a CoReg may correspond to two separate but coregulated genes, we used RT-PCR to confirm that the longest such example in our data is expressed as a single transcript (**Supplementary Table 2** online).

The International Human Genome Sequencing Consortium recently estimated that the human genome contains ~25,000 protein-coding genes (presumably, this is similar for mouse), of which most have already been identified by transcript sequencing[1]. In contrast, previous tiling microarray analyses[10–15] focused on the discovery of thousands of new transcripts in intergenic regions, in introns and antisense to known genes. Although some of these have been confirmed by RT-PCR[13–15] and, in some cases, distinct molecular species have been identified by northern blotting, rapid amplification of cDNA ends or cDNA cloning[14,18,19], the function and origin of these transcripts is largely unknown. Thousands of putative new transcripts probably evolve at a neutral rate[20], suggesting that their function (if any) is independent of sequence. These transcripts might be 'cryptic', potentially resulting from incomplete quality control in Pol II transcription[21], or simply undegraded fragments of heterogeneous nuclear RNA, as more than half of the genome seems to be transcribed as pre-mRNA[22]. Our primary data also include strong signals from many isolated probes (**Fig. 1**). Our results support the view that most multiple-exon genes expressed in diverse tissues are already identified, although there are probably thousands of additional exons that are not currently annotated. Our study therefore reconciles a discrepancy between previous approaches to gene identification and, furthermore, extensively revises our knowledge of the exon composition of the mammalian genome.

METHODS

Array design. To achieve broad coverage of putative exons, we used liberal detection criteria. The numbers of putative exons and of unique putative exons detected by each program were as follows: GenScan, 374,540 and 117,849; HMMGene, 385,759 and 159,523; GrailEXP, 307,911 and 139,906; BlastX, 327,746 and 32,869; and BlastN, 642,401 and 272,152. These yielded a total of 1,140,421 unique putative exons. Details of exon detection and probe selection are given in **Supplementary Methods** online. We selected a single T_m-balanced oligonucleotide to represent each exon on the basis of a scoring system that favors unique sequences without secondary structure and a minimum of simple repeats and homopolymeric runs. Six copies of each of 52 array designs, each containing 21,929 60-mer probes, were manufactured by Agilent Technologies. Sequences of probes are available from our project website, together with mappings to build 33 of the mouse genome.

Tissue pools. We combined the mRNA samples from tissues listed in **Table 1** and reverse-transcribed them for each of the 52 array designs. Typical cDNA yields were 25–50% of the amount of mRNA input. Full details of pool selection, tissue sources and RNA preparation are given in **Supplementary Methods** online.

Varying the sensitivity of GenRate and permutation-based estimates of the exon and CoReg FDERs. We estimated exon and CoReg FDERs as a function of the parameters used in the GenRate analysis. GenRate is a deterministic algorithm with three parameters: the probability θ that a probe is at the start of a CoReg; the probability that a probe in a CoReg corresponds to an exon; and the average number of probes encompassed by a CoReg. The analysis is most dependent on the first parameter, θ, which determines the number of CoRegs that are detected (*i.e.*, the sensitivity of the system) and the FDER. We report results as a function of FDER. The analysis is much less sensitive to the other two parameters, which we set to 0.7 and 20, respectively, using estimates obtained by mapping known human and mouse genes to our probe set. To estimate the FDER, we applied GenRate to a version of our data in which the probes were randomly reordered on a chip-by-chip basis (disrupting their order on the chromosome) and repeated this process ten times to obtain an accurate estimate. By varying θ, we obtained exon and CoReg FDERs varying from 0.13% to 32% and from 0.2% to 37%, respectively.

Mapping known human and mouse genes to our probe set. We compared the chromosomal locations of exons in mouse RefSeq Golden Path genes (build 33) directly with our probe locations, which were mapped to build 33. To include cDNA sequences not on the Golden Path, we used BLAT[23] to map cDNA sequences in RefSeq[24], Ensembl[25], FANTOM2 and Unigene[26] to the chromosomes of build 33 of the mouse genome. To minimize false discovery of genes by cross-hybridization[3], we allowed all probes with 19 in 20 contiguous nucleotide matches to a cDNA to be considered a match. With the exception of Unigene, more than 90% of genes in these five databases were represented by probes on our arrays. We also mapped all 33,930 (28,374 known and 5,556 'novel') of the Ensembl[25] human genes in a similar fashion, using $E < 10^{-4}$ (BLAST) as a cut-off for identity to the array probe.

Comparison with previous results[15] on the X chromosome. We mapped all mouse exons for which we have probes to the human-mouse homologous regions of the X chromosome, using the two-way human-mouse BlastZ alignments downloaded from the University of California Santa Cruz in March 2005. All probes that were previously designed for the human X chromosome[15]

were lined up with the corresponding matched human coordinates in the homologous regions. We constructed an evaluation set of 6,699 putative exons from aligned sequences that include at least one probe from our system and one probe from the previously reported system[15]. We found that 3,447 (52%) of these map to exons in the reference set of mouse RefSeq genes.

We normalized both microarray data sets on a chip by chip basis by applying an affine transformation to the probe signals on each chip, so that the median probe signal and the difference between the 75th percentile and 25th percentile were the same across all chips. To compare the accuracy and recall of GenRate against those of the previously reported system[15], we varied the probe intensity threshold in their method from the 20th percentile to the 90th percentile, obtaining multiple analyses with different sensitivities. The sensitivity of GenRate was varied as described previously.

Comparing GenRate CoRegs with RefSeq genes. A RefSeq gene was considered to be detected if at least one half or at least five of the exons in the gene were detected by GenRate. To determine the set of most highly expressed RefSeq genes, we computed a total expression level for every RefSeq gene. To limit the effects of probe sensitivity, we determined the total expression of a RefSeq gene by counting the number of exons with maximum probe signal (over the 12 tissue pools) in excess of 20. In a previous study[8], this threshold was useful in distinguishing positive signals from negative controls.

URL. Our project website is available at http://www.psi.toronto.edu/genrate/.

Accession codes. GEO, GSE3047.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank G.E. Hinton for conversations and C. Boone and B. Andrews for their support. This work was supported by grants from the Canadian Institutes of Health Research, the Natural Sciences and Engineering Research Council of Canada and the Canadian Foundation for Innovation (to T.R.H., B.J.F. and B.J.B.), by a PREA award (to B.J.F.) and by a Natural Sciences and Engineering Research Council of Canada postdoctoral fellowship (to Q.D.M.).

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Received 17 June; accepted 28 July 2005
Published online at http://www.nature.com/naturegenetics/

1. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).
2. Schadt, E.E. *et al.* A comprehensive transcript index of the human genome generated using microarrays and computational approaches. *Genome Biol.* **5**, R73 (2004).
3. Hughes, T.R. *et al.* Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat. Biotechnol.* **19**, 342–347 (2001).
4. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**, 78–94 (1997).
5. Krogh, A. Two methods for improving performance of an HMM and their applicatoin for gene finding. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **5**, 179–186 (1997).
6. Xu, Y., Mural, R.J. & Uberbacher, E.C. Inferring gene structures in genomic sequences using pattern recognition and expressed sequence tags. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **5**, 344–353 (1997).
7. Waterston, R.H. *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
8. Zhang, W. *et al.* The functional landscape of moues gene expression. *J. Biol.* **3**, 21 (2004).
9. Karolchik, D. *et al.* The UCSC genome browser database. *Nucleic Acids Res.* **31**, 51–54 (2003).
10. Shoemaker, D.D. *et al.* Experimental annotation of the human genome using microarray technology. *Nature* **409**, 922–927 (2001).
11. Stolc, V. *et al.* Global identification of human transcribed sequences with genome tiling arrays. *Science* **306**, 2242–2246 (2004).
12. Yamada, K. *et al.* Empricial analysis of transcriptional activity in the *Arabidopsis* genome. *Science* **302**, 842–846 (2003).
13. Kapranov, P. *et al.* Large-scale transcriptional activity in Chromosomes 21 and 22. *Science* **296**, 916–919 (2002).
14. Rinn, J.L. *et al.* The transcriptional activity of human Chromosome 22. *Genes Dev.* **17**, 529–540 (2003).
15. Bertone, P. *et al.* Global identification of human transcribed sequences with genome tiling arrays. *Science* **306**, 2242–2246 (2004).
16. Kschischang, F.R., Frey, B.J. & Loeliger, H.A. Factor graphs and the sum-product algorithm. *IEEE Trans. Inf. Theory* **47**, 498–519 (2001).
17. Garbarino, J.E. & Gibbons, I.R. Expression and genomic analysis of midasin, a novel and highly conserved AAA protein distantly related to dynein. *BMC Genomics* **3**, 18 (2002).
18. Okazaki, Y. *et al.* Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420**, 563–573 (2002).
19. Cheng, J. *et al.* Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* **308**, 1149–1154 (2005).
20. Wang, J. *et al.* Mouse transcriptome: Neutral evolution of 'non-coding' complementary DNAs (reply). *Nature* **431**, 757 (2004).
21. Wyers, F. *et al.* Cryptic Pol II transcripts are degraded by a nuclear quality control pathway involving a new poly(A) polymerase. *Cell* **121**, 725–737 (2005).
22. Wong, G.K., Passey, D.A. & Yu, J. Most of the human genome is transcribed. *Genome Res.* **11**, 1975–1977 (2001).
23. Kent, W.J. BLAT - The BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
24. Pruitt, K.D., Tatusova, T. & Maglott, D.R. NCBI Reference Sequence Project: update and current status. *Nucleic Acids Res.* **31**, 34–37 (2003).
25. Hubbard, T. *et al.* Ensembl 2005. *Nucleic Acids Res.* **33**, D447–D453 (2005).
26. Pontius, J.U., Wagner, L. & Schuler, G.D. Unigene: A unified view of the transcriptome. in *The NCBI Handbook* (National Center for Biotechnology Information, Bethesda, MD, 2003).