national cohort of individuals with SCA, this study exemplifies how biological sample repositories linked to clinical databases may be efficiently used to successfully perform large disease association studies. However, because cerebrovascular disease in SCA is heterogeneous, manifested as ischemic stroke, intracranial hemorrhage and silent infarction, rigorous phenotypic characterization of cases and controls is imperative.

The lack of available clinical and neuroimaging data needed for optimal phenotypic classification limited this study. Despite these limitations, Sebastiani and co-workers have powerfully demonstrated that multiple SNP sites from different genes over distant parts of the genome are better at identifying overt stroke in SCA than any single SNP or previously identified clinical variable alone. Their results highlight the combined influence of several candidate susceptibility genes on stroke and suggest biological pathways to be explored in future mechanistic studies.

The potential utility of the Bayesian network algorithm is illustrated by the model's ability to determine accurately the relative genetic and clinical effects on stroke risk, find the most probable combination of genetic variants leading to stroke and predict an individual's odds for developing stroke given his/her genotypic profile. As more candidate SNPs and clinical markers are identified, this predictive algorithm will undoubtedly become an invaluable tool in genetic association studies aimed at identifying disease susceptibility genes. The complex interactions modeled through this approach might ultimately translate into clinical benefit through early identification and targeted intervention in those individuals at greatest risk for a particular disease phenotype such as stroke. The computer may well replace the clinician in determining stroke risk, but it will be left to the clinician to apply this information in caring for the patient■

Carolyn Hoppe is at the Department of Hematology/Oncology, Children's Hospital Oakland, Oakland, CA, USA.
E-mail: choppe@mail.cho.org

References

1 Ohene-Frempong K, Weiner SJ, Sleeper LA *et al*: Cerebrovascular accidents in sickle cell disease: rates and risk factors. *Blood* 1998; **91**: 288–294.
2 Carroll JE, McKie V, Kutlar A: Are sickle cell disease patients with stroke genetically predisposed to the event by inheriting a tendency to high tumor necrosis factor levels? *Am J Hematol* 1998; **58**: 250.
3 Taylor JG, Tang D, Foster CB, Serjeant GR, Rodgers GP, Chanock SJ: Patterns of low-affinity immunoglobulin receptor polymorphisms in stroke and homozygous sickle cell disease. *Am J Hematol* 2002; **69**: 109–114.
4 Kahn MJ, Scher C, Rozans M, Michaels RK, Leissinger C, Krause J: Factor V Leiden is not responsible for stroke in patients with sickling disorders and is uncommon in African Americans with sickle cell disease. *Am J Hematol* 1997; **54**: 12–15.
5 Cumming AM, Olujohungbe A, Keeney S, Singh H, Hay CR, Serjeant GR: The methylenetetrahydrofolate reductase gene C677T polymorphism in patients with homozygous sickle cell disease and stroke. *Br J Haematol* 1999; **107**: 569–571.
6 Andrade F, Annichino-Bizzacchi J, Saad S, Costa F, Arruda V: Prothrombin mutant, factor V Leiden, and thermolabile variant of methylenetetrahydrofolate reductase among patients with sickle cell disease in Brazil. *Am J Hematol* 1998; **59**: 46–50.
7 Zimmerman SA, Ware RE: Inherited DNA mutations contributing to thrombotic complications in patients with sickle cell disease. *Am J Hematol* 1998; **59**: 267–272.
8 Sebastiani P, Yu YH, Ramoni MF: Bayesian machine learning and its potential applications to the genomic study of oral oncology. *Adv Dent Res* 2003; **17**: 104–108.
9 Risch N, Merikangas K: The future of genetic studies of complex human diseases. *Science* 1996; **273**: 1516–1517.

Genomics

The amazing complexity of the human transcriptome

Martin C Frith, Michael Pheasant and John S Mattick

European Journal of Human Genetics (2005) **13,** 894–897.
doi:10.1038/sj.ejhg.5201459; published online 22 June 2005

New work from Tom Gingeras and colleagues extends the findings of a series of recent global analyses of transcription[1–7] by revealing a much larger number of non-polyadenylated (polyA−) transcripts than expected and an extraordinary level of organizational complexity in the human transcriptome.

A variety of recent evidence indicates that the majority of sequences in eukaryotic genomes are transcribed and that the proportion of transcribed nonprotein-coding sequences increases with developmental complexity (Table 1). However, it is the novel approaches that Gingeras and colleagues employed that allowed them to add spectacularly to the findings of these previous studies. In particular, the use of tiling arrays to identify transcribed fragments ('transfrags') of the human genome gives more complete and global coverage of the transcriptome than standard cDNA cloning and sequencing approaches, although the relationship between adjacent transfrags that derive from the nearby genomic region is initially uncertain (see below).

Cheng *et al*[6] isolated mature (ie post-spliced) cytoplasmic polyA+ RNA from eight human cell lines and interrogated tiling chips covering 10 human chromosomes in triplicate. They found that the detectable transfrags in each cell line covered on average 5% of the genomic sequences on the arrays. Cumulatively, 10% of the genomic sequences were represented in the polyA+ RNA fraction of one or more cell lines, indicating that many of the observed RNAs were cell type

Table 1 Increase in nonprotein-coding transcription in metazoa

Organism	No. of protein-coding genes	Genome size (Mb)	Coding sequences		UTR sequences		Total transcribed noncoding sequences		Ratio of noncoding to coding sequences
			Mb	%	Mb	%	Mb	%	
Whole genome									
Human	~20–25 000	2851	34	1.2	32	1.1	1619	57	47:1
Mouse	~20–25 000	2490	31	1.3	26	1.1	1339	54	43:1
Fruit fly	~13 500	120	22	18	6.4	5.3	53	44	2.4:1
Nematode	~19 000	100	26	26	0.4	0.4	33	33	1.3:1
Nonrepetitive portion of genome only									
Human		1455	33	2.3	26	1.8	867	60	27:1
Mouse		1422	29	2.0	22	1.6	811	57	28:1
Fruit fly		109	21	20	6.2	5.7	48	44	2.2:1
Nematode		86	25	29	0.3	0.4	26	31	1.1:1

The data were taken from the UCSC human (NCBI build 35), mouse (NCBI build 33), *Drosophila melanogaster* (Flybase release 3.2) and *Caenorhabditis elegans* (Wormbase WS120) databases. For each species, we merged the annotated 'Known Genes' (human, mouse) or equivalents thereof (FlyBase and WormBase genes) with RefSeq genes to create a superset of known protein-coding genes. 'Genome size' shows either the whole genome size or the size of the genome that is not masked out by either RepeatMasker or Tandem Repeat Finder. The protein-coding sequences were calculated as the sum of all annotated coding sequences (CDS); untranslated regions (UTRs) as the difference between annotated exons and the CDS. 'Total transcribed noncoding sequences' consist of UTR and introns in annotated protein coding genes and other sequences (intronic and exonic) in genomic regions covered by spliced cDNAs/ESTs that are not annotated as protein coding. We excluded all spliced cDNAs, which were above the 99th percentile of length of the known genes for the species, the purpose being to remove extremely large cDNAs that may be due to chimeric clones or other artifacts (see, for example, http://www.fruitfly.org/EST/EST.shtml). Gaps were excluded from all calculations as the size of these gaps are not always reliably known. These estimates are conservative and do not include many of the new transcripts observed in global tiling array studies[3–7] or unspliced noncoding RNAs observed in cDNA studies.[2] 'Ratio of noncoding to coding sequences' is the ratio of the total transcribed noncoding sequences to coding sequences. The total amount of UTR sequence increases dramatically from nematodes to mammals, in contrast to the total amount of coding sequence that remains between 21 and 34 Mb. In addition, the ratio of noncoding to coding transcribed sequence increases dramatically from nematodes to mammals, even when adjusted for the much larger amount of repetitive sequence found in mammalian genomes.

specific. The average length of the transfrags (exons) was approximately 120 bp, but PCR cloning and sequencing showed that the detected transfrags are derived from much longer primary transcripts covering extended genomic regions.

The current annotation of the human genome indicates that less than 2% of the genome is present in known or predicted mRNAs. So, most of the RNAs that the authors observed are not derived from known or predicted transcripts: over 56% of the transfrags do not overlap with any well-characterized exon, mRNA or EST annotation; 30% map to 'intergenic' regions and 26% to introns of known genes.

Transcriptome analyses have traditionally focused on cytoplasmic polyA+ RNA. This strategy was used partly to exclude infrastructural RNAs (rRNAs and tRNAs) and incompletely processed primary transcripts, and partly because it was assumed that most transcripts are derived from protein-coding genes and so are processed to polyadenylated mRNAs that are exported to the cytoplasm for translation.

In a radical departure from this tradition, Cheng *et al* extended their study to examine polyA+ and polyA− RNAs fractionated from the nucleus and the cytoplasm of the cell line HepG2. In both fractions, they found more nonpolyadenylated than polyadenylated RNA, a pattern that is consistent with some early but largely forgotten studies 30 years ago.[8–10] Over half of the detected transfrags are unique to the largely unstudied polyA− and the nuclear polyA+ fractions of the transcriptome. Kiyosawa *et al*[11] recently reported similar observations in mouse. A very big and almost completely unexplored area of the expressed RNA repertoire in mammalian cells has just been reopened.

The tiling array technique has some limitations: it does not reveal which strand of the chromosome is transcribed (because the RNA sample is converted into double-stranded cDNA before hybridization), and it does not indicate which transfrags are connected in different transcripts. To address these limitations, the authors studied several hundred randomly selected, nonannotated transfrags in more detail. They used rapid amplification of cDNA ends (RACE) to generate the extended sequences that are linked upstream and downstream of the transfrags *in vivo*. To map and characterize the transcripts that contain the transfrags, these PCR-amplified products were used to reinterrogate the tiling arrays, as well as cloned and sequenced to confirm their structure.

Over half of the studied transfrags show evidence of transcription from both strands. In a number of cases, the authors found exact reverse complement transcripts, so that one transcript has the standard GT–AG sequence at its intron boundaries and its partner has the complementary but nonstandard sequence CT–AC. The most plausible explanation for this pattern is that an RNA-dependent RNA polymerase uses the partner transcript as a template, although no such enzyme has yet been identified in mammals.[12] In this context, however, it is

worth noting that a reservoir of replicable RNA molecules has been proposed to be responsible for the non-Mendelian inheritance of ancestral alleles not present in parental chromosomes in Arabidopsis.[13]

Perhaps, the most basic question about these mysterious unannotated transcripts, termed TUFs (transcripts of unknown function), is whether or not they encode proteins. The cloning and sequencing of 178 TUFs reveals that most do not possess open-reading frames greater than 100 amino acids. This pattern is consistent with these TUFs being noncoding, a conclusion also reached by others.[2] However, some might encode short proteins, and more powerful techniques such as synonymous versus nonsynonymous substitution analysis will need to be employed to provide tighter bounds on this question. In any case, potential short protein coding sequences do not explain the vast extent of the hidden transcriptome that is being brought to light.

These studies also demonstrate the interlaced nature of transcription, so that rather than neatly separated genes, the genome harbors a network of nested and overlapping transcripts on both strands, where introns of one harbor exons of another. Large-scale cDNA sequencing projects such as FANTOM have also revealed such complex patterns, at least in part[2] (Carnici *et al*., submitted for publication). Transcript overlap occurs not only on opposite strands but also on the same strand, so that there is often no clear distinction between splice variants and overlapping neighboring genes.

Kapranov *et al*[7] explore these complex patterns further in a subsequent paper. They examined the structures of transcripts from 14 transcribed loci, representing both known genes and unannotated transcripts taken from those described in Cheng *et al*.[6] They show that there is an amazing world of previously unknown and again barely explored transcripts. Even loci that encode well-known proteins, such as sonic hedgehog, are shown to have previously unknown exons and novel isoforms that are likely to have important functions. They also report that it is not uncommon that a single base pair is part of an intricate network of multiple isoforms of overlapping sense and antisense transcripts, the majority of which are unannotated.

The picture that emerges is that the human genome, far from being a desert with islands of protein-coding sequences, is a nest of interwoven transcriptional units that cover a large fraction of the genome, including many 'intergenic' regions previously considered to be inert. This complexity will undoubtedly have consequences for our understanding of genetic information and pleiotropy, since a mutation may affect multiple overlapping 'genes'. Indeed, the utility of the gene concept itself is no longer clear, both in terms of its discreteness and in terms of the usual presumption that proteins express and transmit most genetic information.

The cDNA cloning and the tiling array approaches give complementary but incomplete views of the transcriptome. The depth of interrogation of the range of expressed transcripts, particularly rare transcripts, by whole tissue cDNA approaches has obvious limitations and is subject to diminishing returns, even using aggressive normalization techniques to remove common transcripts. Tiling arrays are more global, but the data are inherently more noisy and disconnected. Not only are the strand and exon linkages uncertain but also the exact exon boundaries are not revealed with confidence. Even the RACE/array technique does not provide exact exon boundaries and transcript sequences. Cloning and sequencing of these RACE/PCR products is required to reveal the actual transcripts in the detail required for analyzing their characteristics and function. This in turn means that all transfrags might have to be examined in this way to provide a comprehensive view of the human transcriptome. Even if these procedures were refined to be high throughput, this task would be a huge undertaking, although not beyond the scale of past genome projects.

Finally, it is unlikely that tiling arrays and other techniques will have detected all the stable processed transcripts from the human genome. The cDNA approaches used to interrogate tiling arrays will easily not detect short RNAs, for example microRNAs, of which around 1000 have been thus far been identified in human.[14,15] These miRNAs regulate a wide variety of important developmental processes, and are probably just the tip of a very big iceberg of small regulatory RNAs, most of which remain to be discovered.[15] Potentially important[16] regulatory RNAs expressed below the detection limit of such assays are also likely to have been missed. Moreover, since a high proportion of the transcripts show cell-type-specific expression, and only eight cell lines were analyzed, almost certainly many new transcripts will be found in different cell types.

It is now beyond question that the majority of the human genome is transcribed, and that the vast majority of the transcribed sequences are nonprotein coding. There are only two choices – either this transcription is largely meaningless or it is fulfilling some unexpected function. The former explanation is becoming more difficult to sustain, as many of these transcripts and their splicing patterns show cell specificity, although very few have yet been experimentally studied.[17] Both logic and a wide variety of molecular genetic evidence now suggest that there are two inter-related levels of genetic information expressed in complex organisms – that specifying the analog components of cells (mainly proteins, including their many isoforms) and an extensive regulatory RNA network (including microRNAs) transacted by sequence-specific recognition to form various RNA:RNA and RNA:DNA complexes that are in turn recognized and acted upon by different types of nucleic acid-binding proteins.[15,18,19]

We predict that the coming years will see an avalanche of studies demonstrating function for noncoding RNAs, including the many intronic and 'antisense' RNAs that are transcribed. If current indications hold, we may have to reassess many, if not most, of our conceptions of how genetic information is encoded and transacted in our genome■

MC Frith, M Pheasant and JS Mattick are at the Institute for Molecular Bioscience at the University of Queensland, Brisbane, Qld 4072, Australia.
E-mail: j.mattick@imb.uq.edu.au
Martin Frith has a joint appointment at RIKEN Genomic Sciences Centre, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045 Japan.

References

1 Kapranov P, Cawley SE, Drenkow J *et al*: Large-scale transcriptional activity in chromosomes 21 and 22. *Science* 2002; **296**: 916–919.
2 Okazaki Y, Furuno M, Kasukawa T *et al*: Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* 2002; **420**: 563–573.
3 Cawley S, Bekiranov S, Ng HH *et al*: Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* 2004; **116**: 499–509.
4 Kampa D, Cheng J, Kapranov P *et al*: Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Res* 2004; **14**: 331–342.
5 Stolc V, Gauhar Z, Mason C *et al*: A gene expression map for the euchromatic genome of *Drosophila melanogaster*. *Science* 2004; **306**: 655–660.
6 Cheng J, Kapranov P, Drenkow J *et al*: Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* 2005; **308**: 1149–1154.
7 Kapranov P, Drenkow J, Cheng J *et al*: Examples of the complex architecture of the human transcriptome revealed by RACE and high density tiling arrays. *Genome Res* 2005, (in press).
8 Milcarek C, Price R, Penman S: The metabolism of a poly(A) minus mRNA fraction in HeLa cells. *Cell* 1974; **3**: 1–10.
9 Katinakis PK, Slater A, Burdon RH: Non-polyadenylated mRNAs from eukaryotes. *FEBS Lett* 1980; **116**: 1–7.
10 Salditt-Georgieff M, Harpold MM, Wilson MC, Darnell Jr JE: Large heterogeneous nuclear ribonucleic acid has three times as many 5′ caps as polyadenylic acid segments, and most caps do not enter polyribosomes. *Mol Cell Biol* 1981; **1**: 179–187.
11 Kiyosawa H, Mise N, Iwase S *et al*: Disclosing hidden transcripts: mouse natural sense-antisense transcripts tend to be poly(A) negative and nuclear localized. *Genome Res* 2005; **15**: 463–474.
12 Stein P, Svoboda P, Anger M, Schultz RM: RNAi: mammalian oocytes do it without RNA-dependent RNA polymerase. *RNA* 2003; **9**: 187–192.
13 Lolle SJ, Victor JL, Young JM, Pruitt RE: Genome-wide non-Mendelian inheritance of extra-genomic information in *Arabidopsis*. *Nature* 2005; **434**: 505–509.
14 Berezikov E, Guryev V, van de Belt J, Wienholds E, Plasterk RH, Cuppen E: Phylogenetic shadowing and computational identification of human microRNA genes. *Cell* 2005; **120**: 21–24.
15 Mattick JS, Makunin IV: Small regulatory RNAs in mammals. *Hum Mol Genet* 2005; **14**: R121–R132.
16 Johnston RJ, Hobert O: A microRNA controlling left/right neuronal asymmetry in *Caenorhabditis elegans*. *Nature* 2003; **426**: 845–849.
17 Pang KC, Stephen S, Engström PG *et al*: RNAdb – a comprehensive mammalian noncoding RNA database. *Nucleic Acids Res* 2005; **33**: D125–D130.
18 Mattick JS: Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. *Bioessays* 2003; **25**: 930–939.
19 Mattick JS: RNA regulation: a new genetics? *Nat Rev Genet* 2004; **5**: 316–323.