Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation

Patrick Ng[1,3], Chia-Lin Wei[1,3], Wing-Kin Sung[1], Kuo Ping Chiu[1], Leonard Lipovich[1], Chin Chin Ang[1], Sanjay Gupta[1], Atif Shahab[2], Azmi Ridwan[2], Chee Hong Wong[2], Edison T Liu[1] & Yijun Ruan[1]

We have developed a DNA tag sequencing and mapping strategy called gene identification signature (GIS) analysis, in which 5′ and 3′ signatures of full-length cDNAs are accurately extracted into paired-end ditags (PETs) that are concatenated for efficient sequencing and mapped to genome sequences to demarcate the transcription boundaries of every gene. GIS analysis is potentially 30-fold more efficient than standard cDNA sequencing approaches for transcriptome characterization. We demonstrated this approach with 116,252 PET sequences derived from mouse embryonic stem cells. Initial analysis of this dataset identified hundreds of previously uncharacterized transcripts, including alternative transcripts of known genes. We also uncovered several intergenically spliced and unusual fusion transcripts, one of which was confirmed as a *trans*-splicing event and was differentially expressed. The concept of paired-end ditagging described here for transcriptome analysis can also be applied to whole-genome analysis of *cis*-regulatory and other DNA elements and represents an important technological advance for genome annotation.

With the completion of sequencing of the human[1–3] and other mammalian genomes[4,5], scientists have turned their attention to the annotation of genomes for functional content, including gene-coding transcription units and *cis*-acting regulatory and epigenetic elements that modulate gene expression[6]. Current approaches to genome annotation include the use of cDNA[7] and microarray data[8,9] as well as *ab initio* computer predictions[10,11] and comparison of different vertebrate genomes to identify evolutionarily conserved regions[12,13].

Despite considerable success, there are limitations to the current transcript-targeted approaches. Fundamentally, there is no method that can rapidly, efficiently and accurately characterize entire transcriptomes across a large number of cell samples and biological conditions (reviewed in ref. 14). The full-length cDNA (flcDNA) sequencing approach[15,16] provides substantial information, but it is labor-intensive and too costly for the in-depth analysis of multiple transcriptomes. cDNA short-tag strategies, such as serial analysis of gene expression (SAGE)[17,18] and massively parallel signature sequencing (MPSS)[19], can be used to efficiently quantify known transcripts but provide only limited information about transcript structure. To address these problems, we developed an approach that combines the efficiency of short-tag methods with the accuracy provided by flcDNA characterization, to exploit the information contained in assembled genome sequences. The core concept is to obtain only linked 5′ and 3′ short tag sequences for each transcript, map these terminal ‘signatures’ to the genome and thereby infer the complete transcription units by the genome sequence encompassed between these 5′ and 3′ signatures.

RESULTS

Construction of GIS paired-end ditags

As an interim procedure we developed the 5′ LongSAGE and 3′ LongSAGE protocols that extracted 20 base pair (bp) 5′ and 3′ terminal tags separately[20]. With this new capability, we proceeded to design a cloning strategy that would covalently link the 5′ and 3′ signatures of each full-length transcript into a ditag structure (**Fig. 1**). Such PETs representing individual transcripts would then be concatenated for cloning and high-throughput sequencing. A quality single-pass sequencing read (~ 700 bp) would, on average, enable the analysis of about 15 such PET sequences. The PET sequences were then mapped directly to the genome to define the transcription start sites and polyadenylation sites of individual transcripts. To demonstrate this strategy, we generated 116,252 PETs that represented 63,467 nonredundant PET sequences from the E14 mouse embryonic stem cell line.

Quality and mapping specificity of ditags

A typical PET structure should contain an 18-nucleotide (nt) 5′ signature (positions 1–18) and an 18-nt 3′ signature (position 19–36) including a residual AA dinucleotide derived from the mRNA poly(A) tail that indicates ditag orientation (**Supplementary Fig. 1** online). The PET sequences were mapped to the mouse genome assembly (mm3; http://hgdownload.cse.ucsc.edu/goldenPath/mmFeb2003/chromosomes) by a suffix tree–derived alignment algorithm (W.-K.S. *et al.*, unpublished data). When mapped correctly to the genome sequences, nucleotides 1–18 in a ditag sequence should be aligned with the 5′ boundary and nucleotides 19–34 with the 3′ boundary of the corresponding

[1]Genome Institute of Singapore, 60 Biopolis Street, Genome #02-01, Singapore 138672. [2]Bioinformatics Institute, 30 Biopolis Street, Matrix #08-01, Singapore 138671. [3]These authors contributed equally to this work. Correspondence should be addressed to Y.R. (ruanyj@gis.a-star.edu.sg) and E.T.L. (liue@gis.a-star.edu.sg).

PUBLISHED ONLINE 9 JANUARY 2005; DOI:10.1038/NMETH733

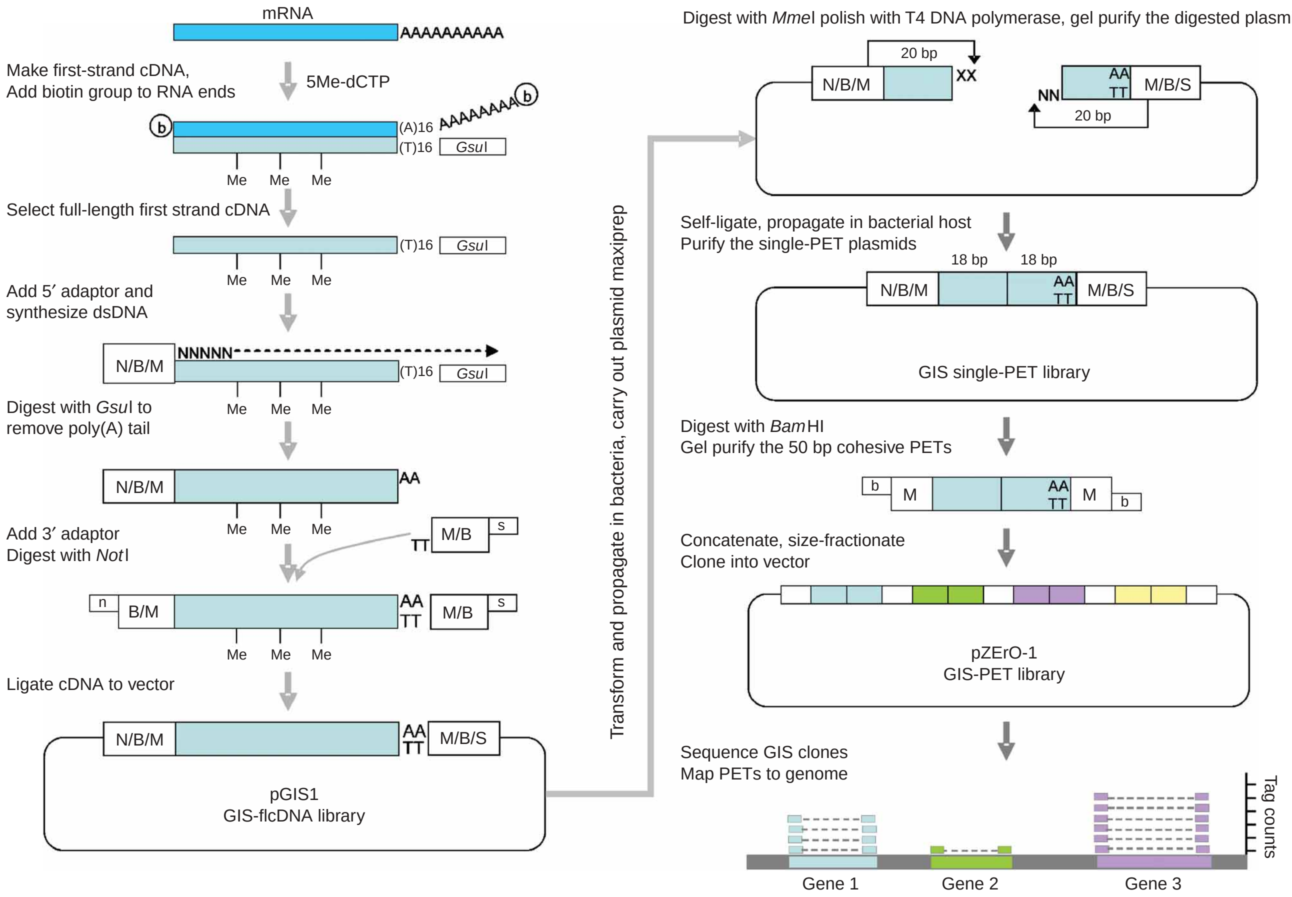

Figure 1 | Schematic view of the GIS analysis method. Starting from poly(A) RNA, single *Mme*I and *Bam*HI sites are introduced to both termini of flcDNA fragments. These are then cloned to create the GIS-flcDNA library. Plasmid DNA from this library is subjected to *Mme*I digestion, resulting in the retention of only a 5′ signature and a 3′ signature from the original cDNA clone. After end-polishing with T4 DNA polymerase, the 5′ and 3′ signatures (now 18 bp each) attached to the vector are self-ligated to form the GIS-PET structure, and this becomes the GIS–single PET library. Plasmid DNA from this intermediate library is digested with *Bam*HI to excise the ditags, which are then concatenated and cloned in pZErO-1 to form the 'GIS-PET library' for sequencing analysis. The PET sequences are subsequently mapped to the genome assembly to define the boundaries of transcription units.

transcript on a chromosomal locus. Of the 63,467 unique PETs obtained, 73.6% (46,714) were mapped to the genome through our automated pipeline, based on the criteria that the paired 5′ and 3′ signatures must be on the same chromosome, in the correct order and orientation (5′→3′) and within 1 million base pairs. In addition, based on our previous observations that both reverse transcriptase–derived nontemplated nucleotide incorporation and type II–restriction enzyme slippage could lead to ambiguities at the PET signature boundaries[20], we mandated a minimum 16-nt contiguous match for the 5′ signature and a 14-nt contiguous match for the 3′ signature. Further analysis showed that 34,815 (74.5%) of the 46,714 PETs mapped to a unique locus in the genome (**Supplementary Table 1** online). Based on their mapping coordinates, PETs with differing sequences but derived from transcripts of the same genes were grouped as PET clusters and assigned to the corresponding known genes in those particular loci. Using this approach, the complete ditag set of 63,467 PETs was grouped into 13,658 PET clusters representing the genes potentially expressed in E14 cells. The 34,815 single-locus PETs were themselves grouped into 7,035 clusters. Our subsequent analyses in this study were mainly focused on this set of single-locus PETs.

To assess overall PET quality and mapping specificity, we first examined the top ten most abundant PET clusters representing well-characterized housekeeping genes. Of the 1,173 unique PET sequences in this group, 98.9% represented full-length transcripts (**Supplementary Table 2** online), and the majority fell within 10 bp of the known 5′ and 3′ boundaries of these transcripts (**Supplementary Fig. 2** online). These results attest to the quality of the full-length cDNA clones and indicate that the PETs generated in GIS analysis were extracted accurately from intact transcripts and mapped correctly to their corresponding loci in the genome.

We then expanded the analysis and examined the entire dataset of 34,815 single-locus PETs and found that 95.2% of these single-locus PETs could be mapped to known transcripts based on the UCSC genome browser (http://genome.ucsc.edu/) annotation, which tracks the RefSeq database, genes defined as 'Known genes' (see **Supplementary Methods**), mammalian gene collection (MGC) and mouse mRNA from GenBank (**Table 1**). The remaining ditags mapped to genomic regions in which either only expressed sequence tags (ESTs; 1.0%) or gene predictions (0.6%) were sited, or no clear transcript information was available (3.1%).

Table 1 | Mapping characteristics of all PETs obtained by GIS analysis

Ditag mapping categories	Ditag counts	% total ditag counts	Unique ditags	% total unique ditags	Ditag clusters	% total ditag clusters
Total ditags generated	116,252	100.0	63,467	100.0		
Ditags mapped	92,213	79.4	46,714	73.6	13,658	
Ditags not mapped	24,039	20.7	16,753	26.4		
Total single-locus ditags	61,687	100.0	34,815	100.0	7,035	100.0
Ditags to transcripts						
Known transcripts	59,046	95.7	33,157	95.2	5,912	84.0
ESTs only	565	0.96	355	1.02	186	2.64
Gene prediction only:	523	0.89	224	0.64	139	1.98
Genscan	132	0.22	85	0.24	59	0.84
SGPGene	55	0.09	36	0.10	30	0.43
Twinscan	43	0.07	30	0.09	18	0.26
Ensembl	293	0.49	73	0.21	32	0.45
Novel 1	402	0.68	258	0.74	208	2.96
Novel 2 (ambiguous)	1,151	1.94	821	2.36	929	13.2

PETs mapped to the genome were assigned to various categories of transcripts if the PETs were located within ±10 kb of the transcript's boundaries. Known transcripts, PETs assigned to either known genes, genes in RefSeq, MGC sequences or mouse mRNAs from GenBank; Novel 1, PETs mapped to genome regions where no known transcripts, ESTs or predictions were found; Novel 2 (ambiguous), PETs that overlapped existing transcript sequences but were located > 10 kb from recorded transcription start site or polyadenylation site. 'Gene prediction only' indicates the sum of all ditags mapped to genes predicted by Genscan, SGPGene, Twinscan and Ensembl.

Of the PETs that mapped to known transcripts, the majority (90.7% for 5′ signatures and 86.9% for 3′ signatures) matched or were in close proximity to (within ±100 bp of the transcription start site or the polyadenylation site) the first and last exons of these known transcripts (**Table 2**). This again suggests that the PET sequences were of high quality, and that the tag-to-genome mapping algorithm was effective.

To examine whether GIS analysis reflects the quantitative nature of transcripts in cDNA libraries, a comparison of 32,540 PETs with EST data obtained from 5,671 random clones (mapping onto the same loci) from another E14 cDNA library was performed. This comparison revealed a correlation coefficient of 0.75 (**Supplementary Fig. 3** online). As a sampling control, an internal comparison between two pools of 10,000 PETs each from the same GIS library gave a correlation of 0.85. This indicated that although some sampling bias was present, the PET counts were quantitatively informative and provided an acceptable measure of relative gene expression.

We further verified the PET mapping by confirming the existence of physical cDNA clones represented by the ditags. We designed PCR primers based on the PET sequences, amplified the corresponding cDNA inserts from the parental GIS flcDNA library and then attempted to reamplify the primary amplicons with nested PCR primers that were designed based on genomic DNA encompassed by the tags. A set of 86 arbitrarily selected PETs representing a wide range of annotation categories, including known genes (38 PETs), predicted genes (2 PETs) and previously unidentified transcripts (46 PETs), was examined by nested PCR. Of these 86 PCRs, 84 (97.7%) generated specific nested PCR products, tentatively confirming the existence of *bona fide* transcript clones in the original GIS flcDNA library. Specific examples of these validated transcripts are shown (**Fig. 2**). To confirm this, we obtained quality bidirectional, single-pass sequences for 54 individual PCR products obtained above and verified that 51 (94.4%) of the 54 PCR-amplified sequences mapped to the expected genomic segments. This high PCR-verification success rate demonstrates an important additional benefit of GIS analysis, namely that the original flcDNA clones represented by the ditags can be efficiently retrieved from the parental GIS flcDNA libraries for further analysis by a simple PCR based on information provided by the ditags.

Table 2 | Mapping specificity of PETs matching Known genes, ESTs and gene predictions

	Known genes				ESTs				Predicted genes			
Ditag distance to transcript (bp)	5′	%	3′	%	5′	%	3′	%	5′	%	3′	%
> 10,000	102	0.31	140	0.42	14	3.94	14	3.94	33	14.7	20	8.93
10,000 to 1,000	336	1.01	783	2.36	29	8.17	93	26.2	23	10.3	27	12.1
1,000 to 100	1,935	5.84	2,054	6.19	105	29.6	116	32.7	40	17.9	76	33.9
100 to −100	30,082	90.7	28,804	86.9	177	49.9	106	29.9	83	37.1	47	20.9
−100 to −1,000	560	1.69	996	3.00	20	5.63	8	2.25	4	1.79	6	2.68
−1,000 to −10,000	114	0.34	325	0.98	8	2.25	16	4.51	25	11.2	36	16.1
< −10,000	28	0.08	55	0.17	2	0.56	2	0.56	16	7.14	12	5.36
Total	33,157	100.0	33,157	100.0	355	100.0	355	100.0	224	100.0	224	100.0

The number of ditags representing Known genes, ESTs and gene predictions is shown in relation to the distance between the ditag and the recorded 5′ or 3′ termini of the corresponding annotation category. Positive distances indicate PET-identified transcripts that are longer than their corresponding matches, whereas negative distances indicate PET-identified transcripts that are shorter than their corresponding matches.

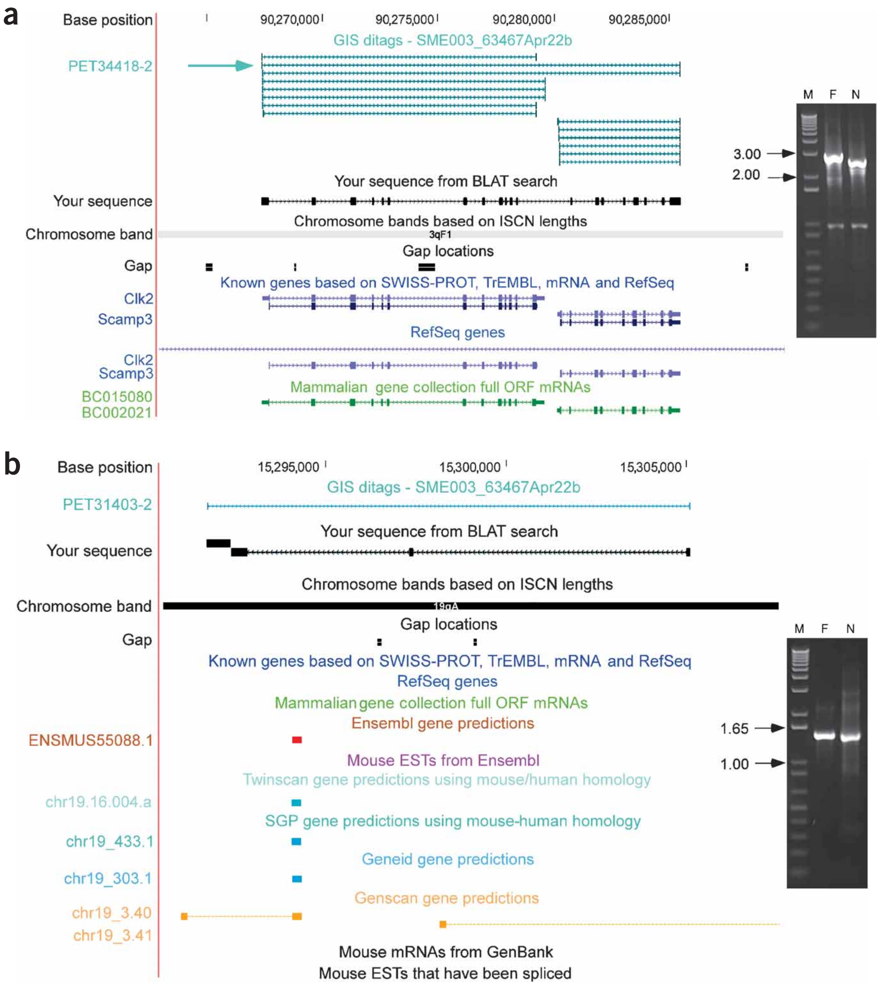
a
Base position
90,270,000
90,275,000
90,280,000
90,285,000
GIS ditags - SME003_63467Apr22b
PET34418-2
Your sequence from BLAT search
Your sequence
Chromosome bands based on ISCN lengths
Chromosome band
3qF1
Gap locations
Gap
Known genes based on SWISS-PROT, TrEMBL, mRNA and RefSeq
Clk2
Scamp3
RefSeq genes
Clk2
Scamp3
Mammalian gene collection full ORF mRNAs
BC015080
BC002021
M F N
3.00
2.00
b
Base position
15,295,000
15,300,000
15,305,000
GIS ditags - SME003_63467Apr22b
PET31403-2
Your sequence from BLAT search
Your sequence
Chromosome bands based on ISCN lengths
Chromosome band
19qA
Gap locations
Gap
Known genes based on SWISS-PROT, TrEMBL, mRNA and RefSeq
RefSeq genes
Mammalian gene collection full ORF mRNAs
Ensembl gene predictions
ENSMUS55088.1
Mouse ESTs from Ensembl
Twinscan gene predictions using mouse/human homology
chr19.16.004.a
SGP gene predictions using mouse-human homology
chr19_433.1
Geneid gene predictions
chr19_303.1
Genscan gene predictions
chr19_3.40
chr19_3.41
Mouse mRNAs from GenBank
Mouse ESTs that have been spliced
M F N
1.65
1.00

Figure 2 | Examples of previously uncharacterized transcripts identified by GIS analysis and verified by PCR and sequencing. The 5′ and 3′ signatures of ditag sequences were mapped to the mouse genome, and are shown as vertical bars linked by an arrowed line in teal indicating the ditag orientation from 5′→3′. Similarly, cDNA sequences of the amplification products derived from the ditags as PCR primers were also mapped to the genome. (**a**) A previously uncharacterized, intergenically spliced transcript represented by two ditag sequences (PET34418, two counts and PET56721, one count) was mapped to chromosome 3 (chr. 3: 90,267,377–90,285,454) and overlapped two adjacent known genes, *Clk2* and *Scamp3*. (**b**) A previously uncharacterized transcript identified by PET31403 (two counts) was mapped to chromosome 19 (chr. 19: 15,305,079–15,291,739). Insets: Agarose gels of PCRs. F, flcDNA fragment of the transcripts amplified by PCR using the ditag sequences as flanking primers; N, PCR with the same fragment, performed using nested primers. M, 1 kb DNA ladder; arrows indicate DNA sizes in kbp.

Unmapped PETs

Approximately 26% (16,753 of 63,467) of the PET sequences could not be mapped automatically to the mouse genome based on the predefined criteria. Possible reasons for this could include nucleotide mismatches including polymorphisms between the transcript sequences derived from the E14 cells (which originated from the 129/Ola strain) and the mouse genome assembly mm3 (which is derived from strain C57BL/6J), sequencing errors present in the ditags and the genome sequences, possible misassembly in the reference genome and potential genome rearrangements in the E14 cells.

To investigate *in silico* the possibility that the unmapped PETs resulted from polymorphisms between the tag sequences and the reference genome, we first aligned the 16,753 unmapped PETs against the mouse mRNA database in GenBank, applying the same parameters as for the genomic alignment and specifying that the 5′ and 3′ signatures of each ditag must match the same mRNA sequences. This resulted in 1,407 mRNA hits out of the 16,753 PETs. By subsequently allowing a single-nucleotide mismatch anywhere within the 5′ and 3′ signatures, we matched an additional 516 PETs to mRNA targets. In sum, we were able to map 1,923 of the originally unmappable PETs to mouse mRNA and subsequently to the mouse genome. In view of our estimated sequencing error rate of <0.01% (less than one error per 1,000 nt) in this study, it suggests that polymorphisms account for the majority of the mismatches. However, as the mRNA data in the public databases are still incomplete, not all PETs can be assigned to the genome by this indirect approach. Therefore, the percentage of PETs affected by polymorphism-dependent mismapping is inevitably underestimated by this analysis.

In an attempt to map the remaining 14,830 PETs that could not be mapped to the mRNA database, we again allowed a single-nucleotide mismatch in the predefined minimal 16-bp 5′ and 14-bp 3′ signatures and aligned the ditags, this time to the mm3 genome assembly. This resulted in specific hits to the genome for 2,479 of the 14,830 PETs. Collectively, mapping to both mRNA and genome sequence databases using relaxed criteria resulted in mapping of an additional 4,402 ditags, bringing the total number of mappable ditags from 46,714 to 51,116 (80.5% of the total 63,467 ditags, an increase of 6.9%). This process is outlined in **Supplementary Fig. 4** online.

To experimentally determine the underlying causes of the initial mapping failures, we randomly selected 99 originally unmapped PET sequences (each of which occurred at least twice, to increase confidence in their authenticity) for further testing. We designed PCR primers based on the PET sequences, amplified the corresponding cDNA inserts from the parental GIS flcDNA library and sequenced the PCR products bidirectionally. Using this approach, we succeeded in extending 94 PET sequences and aligned these longer sequences to the genome. The majority (91 pairs of extended PET sequences) could be mapped to the genomic loci of known transcripts, and individual examination of these sequences revealed that sequence differences were indeed responsible for mapping failures in 53 cases, with these differences ranging from single-base (17 PETs) to multiple-base (36 PETs) mismatches; in extreme cases, substantial stretches of one or both signatures could not be

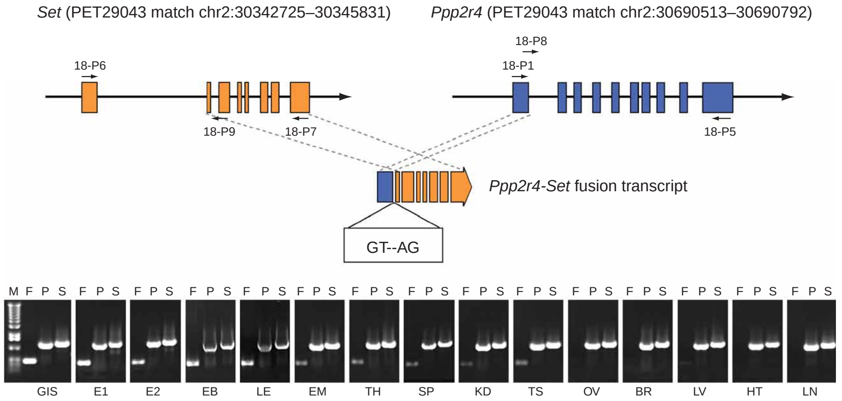
Set (PET29043 match chr2:30342725–30345831)
Ppp2r4 (PET29043 match chr2:30690513–30690792)
18-P8
18-P6
18-P1
18-P9
18-P7
18-P5
Ppp2r4-Set fusion transcript
GT--AG
M F P S F P S F P S F P S F P S F P S F P S F P S F P S F P S F P S F P S F P S F P S F P S
GIS
E1
E2
EB
LE
EM
TH
SP
KD
TS
OV
BR
LV
HT
LN

Figure 3 | Schematic view of the *Ppp2r4-Set* fusion transcript identified by GIS analysis in E14 cells and RT-PCR verification of its existence in various tissues. Exon arrangement of the *Ppp2r4-Set* fusion transcript identified by PET29043, which comprises the first exon of *Ppp2r4* and exons 2–8 of *Set*. The fusion point is a canonical (GT-AG) splice junction. The small labeled arrows indicate the positions of PCR primers used for validation. The existence of this fusion transcript in the E14 cells and various tissues was tested by semiquantitative RT-PCR. The fusion transcript PCR product (F) of 445 bp was generated by primers 18-P8 and 18-P9. The parental *Set* transcript RT-PCR product (S) of 1,146 bp was generated by primers 18-P6 and 18-P7, and the *Ppp2r4* transcript PCR product (P) of 1,299 bp was generated by 18-P1 and 18-P5. M, 1 kb DNA ladder. The source of templates for PCR and RT-PCR, products of which are shown in agarose gels at the bottom of the figure, were GIS, plasmid DNA of GIS-flcDNA library of E14 cells; E1, E14 cell RNA preparation 1; E2, E14 RNA preparation 2; EB, embryoid body; LE, late epiblast; EM, embryo; TH, thymus; SP, spleen; KD, kidney; TS, testes; OV, ovary; BR, brain; LV, liver; HT, heart; LN, lung.

seen on the reference genome. In four such instances, tag sequences that could not be found in the genome assembly were readily matched to mRNA sequences in GenBank. Some examples of these are given in the **Supplementary Data** online (Further GIS analysis). An additional 38 PETs were originally unmapped for reasons including lack of sufficient sequence information within the ditags (21 PETs), ambiguity caused by DNA repeats (12 PETs) and genome assembly errors (5 PETs). Three remaining PETs seemed to represent unusual fusion transcripts and are discussed later.

Our observation that 91 out of the original random selection of 99 (92%) previously unmapped PETs could in fact be successfully validated by PCR and sequencing once again strongly indicated the high quality of the PET sequences themselves. By extrapolation of these results, we might expect that most of the unmappable PETs (15, 413, that is, 92% of 16,753) actually represent authentic transcripts and should be verifiable by this experimental route and directly mappable if derived from the same genetic background as the genome sequences. This would imply that a combined total of 62,127 (46,714 + 15,413), or up to 98% of all PETs, should map to their corresponding transcription units in the mouse genome.

Novel transcripts identified by GIS analysis

The main feature of GIS analysis, simultaneous identification of the 5′ and 3′ ends of transcripts, coupled with enhanced sequencing efficiency, is uniquely useful for elucidating transcriptome complexity by studying individual transcripts. The definition of 'novel' in this context was taken to mean that there was no overlap of the PET-encompassed genomic segments with either *ab initio* gene predictions or previously published mRNA or EST data. In this survey of > 100,000 transcripts (equivalent to sequencing 100,000 cDNA clones) expressed in E14 cells, we were able to identify four major categories of novel transcripts: (i) 154 transcripts with novel alternative transcription start site and 221 transcripts with novel alternative polyadenylation site of known genes; (ii) 114 transcripts representing putative novel uncategorized transcription units; (iii) 14 putative intergenically spliced transcripts, and (iv) 3 unconventional fusion transcripts. These are described in the **Supplementary Data** online (Further GIS analysis) and in **Supplementary Table 3** online. Several examples of these previously uncharacterized transcripts are also illustrated in **Figure 2** and in **Supplementary Fig. 5** online, and these transcripts were validated by PCR and sequencing analysis. Most of these previously uncharacterized transcripts had distinct exon-intron structures, suggesting that they encode proteins, and a few of them were nonspliced, indicating that they might be noncoding RNA species.

The three unconventional fusion transcripts were particularly interesting. These transcripts were identified by three PETs (PET29043, four counts; PET24191, three counts; and PET14161, two counts), which were among the 99 initially unmapped PETs subjected to experimental verification by extended sequencing analysis. These three PET-extended cDNA sequences were unusual in that their 5′ and 3′ portions definitively mapped to different transcripts that were located either on different chromosomes or, if on the same chromosome, in the incorrect order (3′ → 5′) and at remote distances from each other. This raised the unusual possibilities of either chromosomal rearrangements or *trans*-splicing of primary transcripts.

The cDNA sequence derived from PET29043 was 1,475 bp long. Nucleotides 1–281 of this transcript mapped to the 5′ of first exon of the *Ppp2r4* gene on chromosome 2 (chr. 2: 30,690,513–30,690,792), whereas the remaining nucleotides (282–1,475) mapped to a region from the second-to-last exon of the *Set* gene (chr. 2: 30,342,725–30,345,831), also located on chromosome 2 but 344,682 kbp upstream (that is, in the wrong order). Sequence analysis showed that this PET-identified transcript was, in fact, a fusion between *Ppp2r4* and *Set*, in which the first exon of *Ppp2r4* was linked to the second exon of *Set* (**Fig. 3**; sequence details are shown in **Supplementary Fig. 6** online). We were subsequently able to RT-PCR–amplify the *Ppp2r4-Set* transcript directly from two preparations of E14 mRNA, which ruled out the possibility of chimeric clones arising as artifacts of the cDNA cloning procedure. Furthermore, the fusion transcript *Ppp2r4-Set* seemed to be preferentially expressed in embryonic cells (**Fig. 3**, lanes labeled 'F'), whereas the parental transcripts, *Ppp2r4* and *Set*, were expressed constitutively across all samples tested (**Fig. 3**, lanes labeled 'P' and 'S'). Therefore, these data also excluded the possibility of the fusion transcript arising as a genomic translocation event during cell culture adaptation. More intriguingly, the presence of a single continuous open reading frame (ORF) of 263 amino acids, including components of both *Ppp2r4* and *Set*, suggests that a new functional protein is produced, which may play a role in early development, given the preferential expression of *Ppp2r4-Set* in embryonic as compared to adult tissues. More details of the putative *trans*-spliced fusion transcripts are in the **Supplementary Data** online (*Trans*-spliced transcripts) and in **Supplementary Figure 7** online.

The phenomenon of *trans*-splicing has been observed in lower organisms including parasites and the fruit fly, but appears to be very rare in mammals (reviewed in ref. 21.) Our identification of fusion transcripts potentially generated by *trans*-splicing events suggests that *trans*-splicing may be an active, albeit not common, mechanism in mammalian systems for creating new proteins. To this end, the GIS analysis technology is uniquely suited to serve as a platform for the high-throughput and systematic identification of such recombinational events.

DISCUSSION

The GIS analysis described in this study is highly accurate for transcript demarcation. Over 80% of the PETs could be reliably mapped to genome sequences, and, as we estimated, up to 98.9% of the mapped PETs retain authentic terminal signatures of their corresponding transcripts. It is this level of accuracy that enabled us to identify, with a high degree of confidence, authentic alternative transcripts containing previously unidentified transcription start sites and polyadenylation sites, previously uncharacterized genes and transcripts derived from unconventional recombination events such as *trans*- or intergenic splicing. The paired-end nature of GIS analysis is demonstrably superior to the mono-tag based approaches, such as SAGE and MPSS, for elucidating transcriptome complexity.

An obvious advantage of GIS analysis is its unprecedented efficiency in delineating transcripts within the entire transcriptome. That two sequencing reads are required to define both ends of each standard flcDNA clone by conventional cDNA sequencing, whereas a single sequencing read of a GIS ditag clone reveals 15 PETs (corresponding to 15 individual transcripts), suggests that the GIS analysis is potentially 15-fold more efficient in clone and template preparation and 30-fold more efficient in sequencing for the demarcation of transcription boundaries than conventional cDNA sequencing approaches. To carry out a complete analysis of an entire transcriptome by surveying 1,000,000 cDNA clones, one has to generate at least 2,000,000 single-pass cDNA sequencing reads, which is impractical for most laboratories. However, it would require fewer than 100,000 sequencing reads of ditags to enumerate all cloned transcripts in a transcriptome. Nevertheless, one should keep in mind that because GIS analysis provides no information about the internal structure of transcripts, the PETs are unable to distinguish splicing variants formed by internal exon rearrangements.

The recent development of high-density microarrays that tile entire genomic regions demonstrated the possibility of using a whole-genome array approach to identify all possible exon units contributing to the transcriptome[22–24]. This approach, once validated and proven economical for whole-transcriptome analysis, would represent a substantial advance. However, constructing whole-genome arrays for large mammalian genomes currently remains impractical for reasons of cost and complexity. In addition, as a hybridization-based approach, it can only enable one to infer the existence of exons from signals that are averaged over all alternative transcripts, with no clear delineation of individual genes. In contrast, the PET sequencing approach we described can clearly demarcate individual transcript boundaries. A combined whole-genome and PET analysis would conceivably be fruitful, with PET data defining transcript boundaries and whole-genome analysis data helping to identify the internal exons of transcripts.

Like all technologies, GIS analysis has its limitations. First, the accuracy of PET sequences in mapping transcripts is dependent on the effective cloning of flcDNA, a process that is still technically challenging and is indeed the most demanding part of the GIS analysis procedure. Although we adopted the cap-trapper[25] approach, other established protocols for flcDNA selection (such as oligo-capping[26] or SMART cDNA synthesis[27]) may readily be substituted. Second, whereas a comparison of PET versus EST counts of separate E14 cDNA libraries showed a good correlation ($r = 0.75$), it is possible that the flcDNA cloning procedure may result in under-representation of certain transcripts, particularly long transcripts that are difficult to clone. One solution may be to complement the results of GIS analysis with a terminal-tag procedure, in which only short DNA fragments are cloned and sequenced. Such procedures have recently been developed by ourselves[20] and others[28,29].

Offsetting the limitations associated with flcDNA cloning, there are benefits to having flcDNA libraries that are constructed as part of the GIS analysis process. These serve as a sustainable source of cDNA templates and allow the assembly of an efficient pipeline for targeted full-length clone retrieval by simple PCR using information provided in PET sequences, thereby complementing large-scale full-length gene collection efforts such as MGC[15] that are now underway.

In conclusion, we believe that the combined advance in efficiency and accuracy conferred by GIS analysis makes this an ideal technology for comprehensive and systematic characterization of transcriptomes across large numbers of cell samples and biological conditions. In addition to helping define transcripts, the concept of paired-end ditagging for DNA analysis can also be applied, in conjunction with other methodologies, to whole-genome scans for the identification of *cis*-regulatory elements, epigenetic elements and chromosomal rearrangements. We are now starting pilot trials on these applications.

METHODS

Cell culture condition and RNA purification. The feeder-free E14 mouse embryonic stem cells[30] were cultured in the presence of leukemia inhibitory factor in DMEM. After about 15 passages, the cells were harvested for total RNA extraction using Trizol (Invitrogen). High-quality poly(A) RNA was purified from total RNA using the µMACS mRNA Isolation Kit (Miltenyi Biotec) for GIS library construction.

GIS analysis. The complete GIS analysis procedure (**Fig. 1**) comprises the construction of a GIS flcDNA library followed by the ditagging process that converts the GIS flcDNA library to a GIS PET library, which is subjected to computational analysis including PET sequence extraction and genome mapping. During flcDNA synthesis, *Mme*I sites are incorporated at each terminus of the cDNA insert via 5′ and 3′ adapters. These allow the subsequent excision of all cDNA other than the terminal tags, in the context of the unique vector pGIS1 (modified from pGEM3z (Promega)). Intramolecular recircularization results in the formation of a transient single-PET library, from which PETs are excised by

*Bam*HI digestion. Concatenated PETs are then cloned in pZErO-1 to form the GIS PET library. A detailed description of the GIS analysis process together with a step-by-step GIS library construction protocol is included in **Supplementary Protocols** online.

Validations of PET-identified transcripts. Experimental validations of PET-identified transcripts by PCR and sequencing were performed using standard molecular biology procedures. Detailed conditions are provided in the **Supplementary Methods** online, and the sequences of primers and adapters used are listed in **Supplementary Table 4** online.

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

ACKNOWLEDGMENTS

The authors acknowledge Q. Yu of the National University of Singapore and P. Kondu, W.Y. Au Yong, H.C. Yong, M. Hirwan and M. Rebhan of the Bioinformatics Institute of Singapore for information technology and bioinformatics support, as well as P. Li for providing the E14 cells. This work was funded by the Agency for Science, Technology and Research (A*STAR), Singapore.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Received 5 October; accepted 15 December 2004
Published online at http://www.nature.com/naturemethods/

1. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).
2. Lander, E.S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
3. Venter, J.C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
4. Gibbs, R.A. *et al.* Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**, 493–521 (2004).
5. Waterston, R.H. *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
6. The ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306**, 636–640 (2004).
7. Okazaki, Y. *et al.* Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420**, 563–573 (2002).
8. Rinn, J.L. *et al.* The transcriptional activity of human chromosome 22. *Genes Dev.* **17**, 529–540 (2003).
9. Kampa, D. *et al.* Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Res.* **14**, 331–342 (2004).
10. Brent, M.R. & Guigo, R. Recent advances in gene structure prediction. *Curr. Opin. Struct. Biol.* **14**, 264–272 (2004).
11. 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).
12. Guigo, R. *et al.* Comparison of mouse and human genomes followed by experimental verification yields an estimated 1,019 additional genes. *Proc. Natl. Acad. Sci. USA* **100**, 1140–1145 (2003).
13. Parra, G. *et al.* Comparative gene prediction in human and mouse. *Genome Res.* **13**, 108–117 (2003).
14. Ruan, Y., Le Ber, P., Ng, H.H. & Liu, E.T. Interrogating the transcriptome. *Trends Biotechnol.* **22**, 23–30 (2004).
15. Strausberg, R.L. *et al.* Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. *Proc. Natl. Acad. Sci. USA* **99**, 16899–16903 (2002).
16. Kawai, J. *et al.* Functional annotation of a full-length mouse cDNA collection. *Nature* **409**, 685–690 (2001).
17. Velculescu, V.E., Zhang, L., Vogelstein, B. & Kinzler, K.W. Serial analysis of gene expression. *Science* **270**, 484–487 (1995).
18. Saha, S. *et al.* Using the transcriptome to annotate the genome. *Nat. Biotechnol.* **20**, 508–512 (2002).
19. Brenner, S. *et al.* Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotechnol.* **18**, 630–634 (2000).
20. Wei, C-L. *et al.* 5′ Long serial analysis of gene expression (LongSAGE) and 3′ LongSAGE for transcriptome characterization and genome annotation. *Proc. Natl. Acad. Sci. USA* **101**, 11701–11706 (2004).
21. Maniatis, T. & Tasic, B. Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature* **418**, 236–243 (2002).
22. Kapranov, P. *et al.* Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296**, 916–919 (2002).
23. Shoemaker, D.D. *et al.* Experimental annotation of the human genome using microarray technology. *Nature* **409**, 922–927 (2001).
24. Yamada, K. *et al.* Empirical analysis of transcriptional activity in the *Arabidopsis* genome. *Science* **302**, 842–846 (2003).
25. Carninci, P. & Hayashizaki, Y. High-efficiency full-length cDNA cloning. *Methods Enzymol.* **303**, 19–44 (1999).
26. Maruyama, K. & Sugano, S. Oligo-capping: a simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides. *Gene* **138**, 171–174 (1994).
27. Zhu, Y.Y., Machleder, E.M., Chenchik, A., Li, R. & Siebert, P.D. Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction. *Biotechniques* **30**, 892–897 (2001).
28. Hashimoto, S. *et al.* 5′-end SAGE for the analysis of transcriptional start sites. *Nat. Biotechnol.* **22**, 1146–1149 (2004).
29. Shiraki, T. *et al.* Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl. Acad. Sci. USA* **100**, 15776–15781 (2003).
30. Hooper, M., Hardy, K., Handyside, A., Hunter, S. & Monk, M. HPRT-deficient (Lesch-Nyhan) mouse embryos derived from germline colonization by cultured cells. *Nature* **326**, 292–295 (1987).