1 μg/ml). Cassettes[5] were cloned in pAV vectors, derived from pCWRSVN (ref. 11) by placing the promoter modules between *Bam*HI and *Hin*dIII sites, after modifying the vector. Modifications included destruction of the *Bam*HI site downstream of the Neo cassette, and removal of all sites between the original *Sal*I and *Xho*I sites, inclusive, by cleavage and religation. After inserting the cassettes, a new polylinker was created between the *Hin*dIII and *Sac*II sites. Sequences to be expressed were inserted as synthetic oligodeoxynucleotides precisely between the end of the unique *Sal*I site and the beginning of the unique *Xba*I site. Recombinant constructs were sequenced.

Transfections. Transient transfections were carried out on subconfluent HeLa cells. Synthetic RNA was transfected using Oligofectamine as described[4]. Recombinant DNA constructs were transfected using Lipofectin with Plus reagent according to the manufacturer's instructions. In transient transfections, cells were split after one day. Cells were fixed and examined for lamin protein after three days, and fixed and examined by *in situ* hybridization after two days.

Fluorescence microscopy. Transfected cells were fixed and subjected to previously described protocols for visualizing proteins[4] with antibodies (lamin A/C and β-Gal) or detecting small RNAs (http://singerlab.aecom.yu.edu/protocols) by hybridizing 5′-Cy3-labeled oligos (5′-Cy3-AAACUGGACUUCCAGAAGAACACGAA, 2′-*O*-methyl ribonucleotides) to the fixed preparations. Fluorescence was acquired with a Nikon Eclipse E800 (Tokyo, Japan) with a Hamamatsu Orca II camera (Hamamatsu-City, Japan). For each construct, hundreds of cells were examined to confirm that the selected images were representative. On multiple slides, lamin A/C fluorescence in transfected cells was deconvoluted and quantitated using Isee software (Inovision; Raleigh, NC) and is expressed in Table 1 as a percentage of lamin A/C signal from nontransfected cells on the same slides. Lamin signal was consistently higher in transfected cells than in untransfected cells on the same slide.

Acknowledgments

This work was supported by NIH grant AI40936 to D.R.E. and the Medical Scientist Training Program at the University of Michigan. We thank Gary R. Kunkel for the cloned, human U6 snRNA gene, and Michael Imperiale and Eric Fearon for other materials.

Competing interests statement

The authors declare competing financial interests: see the Nature Biotechnology *website (http://biotech.nature.com) for details.*

Received 28 December 2001; accepted 19 March 2002

1. Zamore, P.D. RNA interference: listening to the sound of silence. *Nat. Struct. Biol.* **8**, 746–750 (2001).
2. Bernstein, E., Denli, A.M. & Hannon, G.J. The rest is silence. *RNA* **7**, 1509–1521 (2001).
3. Gil, J. & Esteban, M. Induction of apoptosis by the dsRNA-dependent protein kinase (PKR): mechanism of action. *Apoptosis* **5**, 107–114 (2000).
4. Elbashir, S.M. *et al.* Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. *Nature* **411**, 494–498 (2001).
5. Good, P.D. *et al.* Expression of small, therapeutic RNAs in human cell nuclei. *Gene Ther.* **4**, 45–54 (1997).
6. Bertrand, E. *et al.* The expression cassette determines the functional activity of ribozymes in mammalian cells by controlling their intracellular localization. *RNA* **3**, 75–88 (1997).
7. Cheong, C., Varani, G. & Tinoco, I. Jr. Solution structure of an unusually stable RNA hairpin, 5′-GGAC(UUCG)GUCC. *Nature* **346**, 680–682 (1990).
8. Elbashir, S.M., Lendeckel, W. & Tuschl, T. RNA interference is mediated by 21- and 22-nucleotide RNAs. *Genes Dev.* **15**, 188–200 (2001).
9. Nykänen, A., Haley, B. & Zamore, P.D. ATP requirements and small interfering RNA structure in the RNA interference pathway. *Cell* **107**, 309–321 (2001).
10. Lipardi, C., Wei, Q. & Paterson, B.M. RNAi as random degradative PCR: siRNA primers convert mRNA into dsRNAs that are degraded to generate new siRNAs. *Cell* **107**, 297–307 (2001).
11. Chatterjee, S., Johnson, P.R. & Wong, K.K. Jr. Dual-target inhibition of HIV-1 *in vitro* by means of an adeno-associated virus antisense vector. *Science* **258**, 1485–1488 (1992).
12. Kunkel, G.R., Maser, R.L., Calvet, J.P. & Pederson, T. U6 small nuclear RNA is transcribed by RNA polymerase III. *Proc. Natl. Acad. Sci.* USA **83**, 8575–8579 (1986).
13. Kunkel, G.R. & Pederson, T. Upstream elements required for efficient transcription of a human U6 RNA gene resemble those of U1 and U2 genes even though a different polymerase is used. *Genes Dev.* **2**, 196–204 (1988).
14. Danzeiser, D.A., Urso, O. & Kunkel, G.R. Functional characterization of elements in a human U6 small nuclear RNA gene distal control region. *Mol. Cell. Biol.* **13**, 4670–4678 (1993).

Using the transcriptome to annotate the genome

Saurabh Saha[1,2†], Andrew B. Sparks[1,3†], Carlo Rago[1], Viatcheslav Akmaev[4], Clarence J. Wang[4], Bert Vogelstein[1], Kenneth W. Kinzler[1*], and Victor E. Velculescu[1*]

A remaining challenge for the human genome project involves the identification and annotation of expressed genes. The public and private sequencing efforts have identified ~15,000 sequences that meet stringent criteria for genes, such as correspondence with known genes from humans or other species, and have made another ~10,000–20,000 gene predictions of lower confidence, supported by various types of *in silico* evidence, including homology studies, domain searches, and *ab initio* gene predictions[1,2]. These computational methods have limitations, both because they are unable to identify a significant fraction of genes and exons and because they are unable to provide definitive evidence about whether a hypothetical gene is actually expressed[3,4]. As the *in silico* approaches identified a smaller number of genes than anticipated[5–9], we wondered whether high-throughput experimental analyses could be used to provide evidence for the expression of hypothetical genes and to reveal previously undiscovered genes. We describe here the development of such a method—called long serial analysis of gene expression (LongSAGE), an adaption of the original SAGE approach[10]—that can be used to rapidly identify novel genes and exons.

The LongSAGE method (Fig. 1) generates 21 bp tags derived from the 3′ ends of transcripts that can rapidly be analyzed and matched to genomic sequence data. The method is similar to the original SAGE approach[10], but uses a different type IIS restriction endonuclease (*Mme*I) and incorporates other modifications to produce longer transcript tags. The resulting 21 bp tag consists of a constant 4 bp sequence representing the restriction site at which the transcript was cleaved, followed by a unique 17 bp sequence derived from an adjacent sequence in each transcript. Theoretical calculations show that >99.8% of 21 bp tags are expected to occur only once in genomes the size of the human genome (Table 1A). Likewise, similar analyses based on actual sequence information from ~16,000 known genes suggest that >75% of 21 bp tags would be expected to occur only once in the human genome, with the remaining tags matching duplicated genes or repeated sequences (as discussed below). In contrast, conventional SAGE tags of 14 bp do not allow unique assignment of tags to genomic sequences, though they do allow such assignment to the much less complex compendium of expressed sequence tags (ESTs) and previously characterized mRNAs[10–12]. To optimize the quantification of transcripts, tags are ligated together to form "ditags," which are then concatenated and cloned. Sequencing tag concatemers in parallel allows the identification of up to ~30 tag sequences in each sequencing reaction. Matching tags to genome

[1]*Howard Hughes Medical Institute and the Sidney Kimmel Comprehensive Cancer Center, and* [2]*Program in Cellular and Molecular Medicine, Johns Hopkins Medical Institutions, Baltimore, MD 21231.* [3]*Current address: GMP Genetics, 200 Prospect Street, Waltham, MA 02451.* [4]*Genzyme Molecular Oncology, P.O. Box 9322, Framingham, MA 01701.* [†]*These authors contributed equally to this work.* **Corresponding authors (kinzlke@jhmi.edu and velculescu@jhmi.edu).*

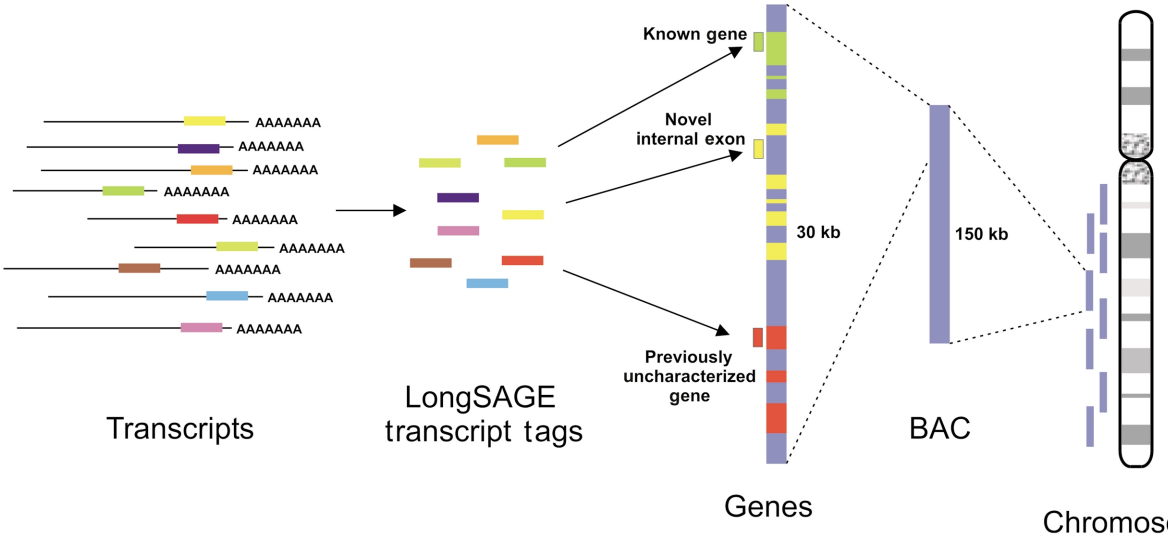

Figure 1. Schematic of LongSAGE method. Transcripts are isolated from cells of interest and tags are excised using specialized restriction endonucleases. Tag sequences are revealed by sequencing of tag concatemers. Matching of tags to genomic sequences allows precise localization of tags in the genome. Comparison of tag locations with positions of previously annotated genes can provide expression evidence for predicted genes, and identify novel internal exons and previously uncharacterized genes. See text and experimental protocol for details.

sequences identifies the gene corresponding to each tag, and the number of times a particular tag is observed provides a quantitative measure of transcript abundance in the RNA population.

To illustrate this approach, we characterized ~28,000 transcript tags expressed from the colorectal cancer cell line DLD-1 and compared these tags to the draft human genome sequence assembled by the Human Genome Project (HGP) and Celera Genomics (Rockville, MD) (Table 1B). The transcript tags matched slightly more sequences in the Celera database: 8,316 different tags matched genomic sequence, and 5,641 (68%) of these tags matched uniquely. The remaining tags corresponded to duplicated genes or gene domains (1,961; 24%) or to repetitive sequences (714; 9%). We found that 23 of the 25 most highly expressed tags were derived either from duplicated genes or from genes located in the mitochondrial genome. These included both functional and nonfunctional (pseudogene) copies of genes involved in protein synthesis, energy metabolism, and chromosomal structure. Working backwards from the genome to the transcriptome, our analyses showed that tags present at >10 copies per genome were on average fivefold more highly expressed than those that were present at only one copy per genome (Table 1B, $P < 10^{-10}$, χ^2 test). The association between elevated gene expression and gene copy number suggests that gene duplication may provide a mechanism for increasing gene expression or that highly expressed genes are more likely to undergo duplication through retrotransposition.

In order to identify potential undiscovered genes corresponding to LongSAGE tags, we first excluded tag loci that matched previously annotated genes, including both known genes and gene predictions identified through *in silico* analyses (Table 2). Of the 5,641 tags with single loci in the Celera genome, 3,419 precisely matched exonic sequences or 3′ untranslated regions (UTRs). Analysis of the HGP data revealed tag matches to a similar number of exonic and UTR sequences. Notably, our LongSAGE data provided direct experimental evidence for the expression of 245 predicted genes in the Celera genome, 111 of which had previously been predicted to be expressed solely on the basis of *in silico* evidence.

The remaining tags represent potential undiscovered genes or unrecognized exons of previously annotated genes (Table 2). To distinguish between these two possibilities, we analyzed the genomic regions surrounding tag loci. In the Celera database, 575 tags were found to match regions within introns of known genes. Such

Table 1A. Theoretical matching of tags to genome

Tag length (*n* base pairs)	Complexity[a]	Tag uniqueness probability[b]
14	1,048,576	0.00%
15	4,194,304	0.08%
16	16,777,216	16.73%
17	67,108,864	63.95%
18	268,435,456	89.43%
19	1,073,741,824	97.24%
20	4,294,967,296	99.30%
21	17,179,869,184	99.83%

[a]Complexity of tags ($C = 4^{(n-4)}$) is determined using a tag length comprising a constant 4 bp representing the restriction site at which the transcript was cleaved, followed by *n* bp derived from the adjacent sequence in each transcript. [b]The probability that a tag is unique in the genome ($P_u = [(C - 1)/C]^{30,000,000}$) is determined under the assumption that the genome contains ~30 × 10^6 *Nla*III-derived tags and is comprised of random sequence.

Table 1B. Experimental matching of LongSAGE tags to genome

Tag loci in genome[c]	Tags mapped to HGP database[d]	Tags mapped to Celera database[d]	Average expression level[e]
Nuclear genome			
1 copy/genome	5,605 (70%)	5,641 (68%)	1.73
2–10 copies/genome	1,638 (20%)	1,835 (22%)	3.38
>10 copies/genome	96 (<1%)	126 (<1%)	9.17
Repeats (e.g., *Alu*, LINE)	713 (9%)	714 (9%)	1.65
Total	8,052	8,316	2.25
Mitochondrial genome	54	54	38.1

[c]Number of nuclear genome locations matched by individual 21 bp tags. [d]Mapping was performed after removing tags matching repeated sequences using Celera CHGD Assembly Repeats Release 25g. [e]Average expression level was determined by dividing the total number of transcript tags observed in the library by the number of different tags (mapped to Celera database).

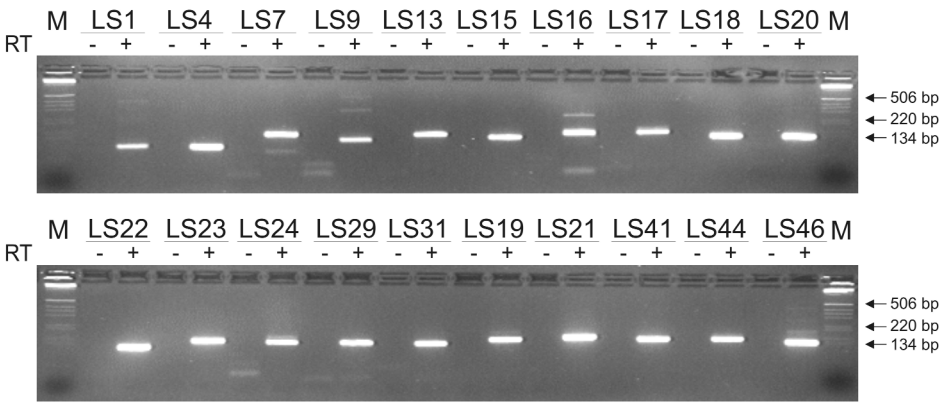

Figure 2. Expression analysis of candidate LS genes. RT-PCR was performed for each LS gene using cDNA from DLD-1 cells (RT^+) or mock templates (RT^-), and was analyzed by real-time PCR analysis and gel electrophoresis. In each case, the amplified product of the expected size was present only in the RT^+ lanes, and the signal from real-time PCR analysis was at least 60-fold greater for RT^+-containing templates than for mock templates.

sequences represent either unknown exons of annotated genes or novel genes embedded in the introns of known genes. A total of 803 tags matched regions at least 5 kb from the terminal exons of known or predicted genes. These sequences most probably represented novel genes. Examination of the tags mapped to the HGP database identified an even larger number of novel candidate internal exons and potential new genes (Table 2), the majority of which were not mapped to known or predicted genes in the Celera databases.

To independently confirm the expression of putative novel genes identified by LongSAGE (candidate LS genes), we arbitrarily selected 129 of these sequences for real-time polymerase chain reaction (PCR) analysis (Supplementary Table 1 online). The template for these analyses was cDNA produced through reverse transcription (RT) of mRNA from DLD-1 cells. Of the 129 candidate LS genes, 123 genes (95%) were shown to be expressed in DLD-1 cells, as the real-time PCR signals were over 60-fold higher in the presence of RT-derived templates than in mock templates produced from DLD-1 RNA without RT. Moreover, gel electrophoresis verified the presence of the expected PCR products in cDNA templates produced by RT (Fig. 2). Searches of the extant DNA databases for sequences of the predicted LS gene fragments revealed no similarity to characterized genes in DNA databases, although two predicted exons containing LongSAGE tags were identified by the *ab initio* program GENSCAN[13].

After the completion of this analysis, we re-analyzed the tags corresponding to candidate novel internal exons and genes to see if they matched newly annotated genes in more recent databases. Analysis of 100 of the 583 tags representing potentially novel internal exons against the current Celera database (release of 20 December 2001) revealed that 12 of 100 matched recently annotated exons. Likewise, examination of the 129 tags representing putatively novel genes in the same database identified 13 of these as corresponding to newly annotated genes (Supplementary Table 1 online). Furthermore, another 32 of these tags were shown to be represented in EST databases not included in the initial annotation of the genome. As these genes and exons were not present in the databases used in our initial analyses, these results further demonstrate that LongSAGE tags can identify previously unrecognized internal exons and uncharacterized genes.

Like other high-throughput approaches, LongSAGE has a number of limitations. First, a small number of genes would be expected to lack the appropriate 4 bp restriction site used in the first enzymatic step and would therefore be missed. This problem could be overcome by the generation of additional LongSAGE libraries using enzymes with different 4 bp recognition sites. Second, LongSAGE tags identify only a portion of each transcribed gene, and additional analyses are required to obtain full-length gene sequences. Different approaches could be employed for this purpose, including RT-PCR and sequencing across computationally predicted exons in nearby regions, rapid amplification of cDNA ends (RACE) using tags as PCR primers[14], and hybridization to cDNA libraries using tags as probes[10]. Finally, this study identified only a fraction of the genes expressed in colorectal cancer cells. A much larger number of tags from a variety of different cell types and environmental conditions would be required to thoroughly describe the compendium of expressed genes.

Nevertheless, our study has already shown that LongSAGE can provide experimental evidence for the expression of hypothetical genes previously predicted from *in silico* analysis of the human genome. Equally important, it is clear that a significant number of undiscovered genes and exons are present in the genome and can be readily detected by LongSAGE. Extrapolation from our current analysis would suggest that there are at least 15,000 uncharacterized exons in the genome, half of which are likely to be derived from previously unrecognized genes. This approximation is certainly an underestimate, as our study focused only on highly expressed genes in a well-characterized colorectal cancer cell line.

We envision that a systematic large-scale analysis of the genome using LongSAGE will be complementary to other approaches for gene identification. As LongSAGE represents one of the few high-throughput discovery approaches that does not depend on *a priori* knowledge of gene sequences, such data will immediately allow independent validation of *in silico* gene predictions and identification of regions not annotated by such methods. Additionally, although the approach is conceptually similar to EST sequencing[15,16], LongSAGE is at least an order of magnitude more efficient, allowing discovery of transcripts expressed at such low levels that they may not be represented in EST libraries. As any investigator with access to an automated sequencer can generate thousands of LongSAGE tags, this approach will allow genome annotation to be distributed across many different laborato-

Table 2. Identification of uncharacterized candidate genes and exons

Gene/exon category	Tags mapped to HGP database	Tags mapped to Celera database
Previously annotated[a]		
Known genes	2454	3174
Predicted genes	882	245
Total	3336	3419
Previously unannotated[b]		
Internal exons	583	575
Genes	920	803

[a]Tags matching exon sequences of annotated genes or regions within 5 kb of the terminal exon. Known genes refer to 17,969 Otto (Celera) or 9,550 RefSeq (HGP) annotations, whereas predicted genes refer to 21,350 *de novo* transcript predictions (Celera) or 11,597 Genie predictions (HGP), using Celera's or HGP's definitions[1,2]. [b]Tags matching regions between annotated exons of the same gene were considered novel internal exons, while tags matching regions >5 kb from the terminal exon were considered previously unannotated genes.

ries, permitting the discovery of genes expressed in unique tissues or experimental conditions. Finally, LongSAGE data could be used to focus array-based gene verification strategies[17] on specific genomic regions containing putatively novel genes. This would considerably expand the utility of expression arrays in general, as their content is usually limited to confirmed genes. The use of LongSAGE could considerably facilitate the annotation of genomes of other organisms whose genome sequences have been determined but whose transcript databases are less extensive than those for humans.

Experimental protocol

LongSAGE library construction. Total RNA was isolated from DLD-1 cells and mRNA was selected as previously described[14]. LongSAGE was performed with 2 μg mRNA using the standard SAGE protocol with the following modifications. Linkers containing the *Mme*I recognition site were ligated to 3′ cDNA ends after *Nla*III digestion (linker 1A (5′-TTTGGATTTGCTGGTGCAGTACAACTAGGCTTAATATCCGACATG-3′) and linker 1B (5′-TCGGATATTAAGCCTAGTTGTACTGCACCAGCAAATCC C7-amino-modified-3′) were annealed together and ligated to half the cDNA population, and linker 2A (5′-TTTCTGCTCGAATTCAAGCTTCTAACGATGTACGTCCGACATG-3′) and linker 2B (5′-TCGGACGTACATCGTTAGAAGCTTGAATTCGAGCAG C7-amino-modified-3′) were annealed together and ligated to the remaining half of the cDNA). Linker tag molecules were released from the cDNA using the *Mme*I type IIS restriction endonuclease[18,19] (University of Gdansk Center for Technology Transfer, Gdansk, Poland). Digestion was performed at 37°C for 2.5 h using 40 units *Mme*I in 300 μl of 10 mM HEPES, pH 8.0, 2.5 mM potassium acetate, 5 mM magnesium acetate, 2 mM DTT, and 40 μM *S*-adenosylmethionine. The linker 1 tag and linker 2 tag molecules were not polished and were directly ligated together in a 6 μl reaction containing 4 units T4 DNA ligase (Invitrogen, Carlsbad, CA) in the supplied buffer for 2.5 h at 16°C. The SAGE software was modified to allow extraction of 21 bp tags from sequences of concatemer clones. Detailed protocols for performing SAGE and LongSAGE and software for extraction of LongSAGE data are available at http://www.sagenet.org/sage_protocol.htm.

Quantitative PCR analysis of candidate LS genes. Single-stranded cDNA was synthesized from DLD-1 mRNA using Superscript II reverse transcriptase (Invitrogen) following the manufacturer's protocol, and mock template preparations were prepared in parallel without the addition of reverse transcriptase. We arbitrarily selected 129 LongSAGE tags matching regions >5 kb from the 3′ terminal exon of annotated genes in the HGP and Celera databases for RT-PCR analysis. As we expected that most tags would correspond to the last exon of the candidate LS genes, primers were designed using the Primer 3 interface (http://www-genome.wi.mit.edu/cgi-bin/primer/ primer3_www.cgi) to span a 100–200 bp region that included the tag, and were synthesized by GeneLink (Hawthorne, NY). These sequences were predicted to lie in the same exon as the tag because no splice sites intervened between the tag and primer sequences. Quantitative PCR was performed using an iCycler (Bio-Rad, Hercules, CA) using Pico Green dye (Molecular Probes, Eugene, OR), and threshold numbers were collected using iCycler software version 1.0. Quantitative PCR reactions were performed in triplicate and the threshold cycle numbers were averaged. Expression of LS genes was evaluated using the formula $2^{(Rt - Mt)}$ where Rt is the threshold cycle number observed for a LongSAGE gene in RT-derived templates from DLD-1 RNA, and Mt is the threshold cycle number observed in the mock template preparations from DLD-1 RNA. Expression for an LS gene was considered to be positive when the signal from RT-derived templates was at least 60-fold greater than that of mock templates and when gel electrophoresis showed products of the appropriate size in reactions using the RT-derived templates. The quality of RT reactions was assessed by quantitative PCR using primers for the amyloid-β (A4) precursor protein (*APP*), a gene known to be expressed in colorectal cancer cells[7]. Presence of contaminating genomic DNA and intronic RNA (heterogenous nuclear RNA) in RT reactions was determined by quantitative PCR using primers for an intronic region of the GATA-binding protein 4 gene. This analysis showed undetectable levels of product following 50 cycles of PCR from the synthesized cDNA, as opposed to robust amplification at 25 cycles using an equivalent amount of genomic DNA.

Matching LongSAGE tags to the genome. All 17 bp tags adjacent to the *Nla*III anchoring enzyme site (CATG) and corresponding position information were computationally extracted from the genome sequences assembled by HGP and Celera, using UCSC Golden Path Oct. 7, 2000 Assembly (http://genome.ucsc.edu/goldenPath/07oct2000/chromosomes/) and CHGD Assembly Release 25h (http://www.celera.com), respectively. The 27,737 experimentally derived LongSAGE tags comprising 14,020 different tags were electronically matched to ~3×10^7 extracted genomic tag sequences to identify the precise location of tag matches in the genome. A total of 8,570 different tags comprising 20,709 tags (75%) could be assigned to either the HGP or Celera draft of the genome, or to the mitochondrial genome. The remaining tags were likely due to sequencing errors in the tags or human genomic sequences, sequence polymorphisms in the tag region, or transcribed sequences not represented in the genome databases. Tags matching multiple locations in the genome were considered to be duplicated genes if they were not present in databases containing repeat sequences (Celera CHGD Assembly Repeats Release 25g) and if the majority of matching Celera hCG gene descriptions were identical. The 22 tags matching both the mitochondrial and nuclear genomes were considered to represent mitochondrial genes.

Comparison of LongSAGE tags to annotated genes. Annotated exon coordinates of known and predicted genes were obtained from the files Celera CHGD_transcripts_R25.26k and CHGD_transcripts_R25.12k (http://www. celera.com) and from the files HGP genieKnown.txt and genieAlt.txt (http://genome.ucsc.edu/goldenPath/07oct2000/database/). We considered the 17,969 Celera Otto and 9,550 HGP RefSeq annotations to represent known genes, and the 21,350 Celera *de novo* transcripts and 11,597 HGP Genie transcripts to represent predicted genes. Predicted genes without significant similarity to the Celera Human Gene Index[2] (match identity >90% and alignment >50%), were considered to lack human mRNA or EST expression data. As LongSAGE tags were derived from the *Nla*III site closest to the 3′ end of transcripts, tags were considered to match their corresponding genes when they identically matched annotated exonic sequences, or 3′ UTRs <5 kb from the terminal exon. Only tags matching in the sense orientation were considered for these analyses, as we could not distinguish whether antisense tags were the result of an undiscovered overlapping gene on the opposite strand or of internal oligo-dT priming during cDNA synthesis. Tags were considered to match novel internal exons of annotated genes when the tags matched intronic regions between two exons of the same gene in the appropriate orientation. Alternatively, tags were considered to match previously undiscovered genes when they matched regions >5 kb from the 3′ terminal exon of an annotated gene, or were present in contigs containing no annotated genes. Subsequent validation analyses were performed using manual BLAST analyses of 21 bp tags against dbEST (Release 011102, http://www.ncbi.nlm.nih.gov) and Celera human transcript database (Release of 20 December 2001). All tags corresponding to candidate novel internal exons and candidate novel genes are available in Supplementary Table 2 and Supplementary Table 3, respectively, online.

Note: Supplementary information is available on the Nature Biotechnology *website.*

Acknowledgments

We thank Kathy Romans for assistance with database searches, Jennifer Davis for statistical analyses, and Steve Madden, Kathy Klinger, Xiaohong Cao, and members of our laboratories for helpful discussions. This work was supported by NIH grant CA57345.

Competing interests statement

The authors declare competing financial interests: see the Nature Biotechnology *website (http://biotech.nature.com) for details.*

Received 2 October 2001; accepted 25 February 2002

1. Lander, E.S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
2. Venter, J.C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
3. Wheelan, S.J. & Boguski, M.S. Late-night thoughts on the sequence annotation problem. *Genome Res.* **8**, 168–169 (1998).
4. 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).
5. Fields, C., Adams, M.D., White, O. & Venter, J.C. How many genes in the human genome? *Nat. Genet.* **7**, 345–346 (1994).
6. Dunham, I. *et al.* The DNA sequence of human chromosome 22. *Nature* **402**,

489–495 (1999).
7. Velculescu, V.E. *et al.* Analysis of human transcriptomes. *Nat. Genet.* **23**, 387–388 (1999).
8. Liang, F. *et al.* Gene index analysis of the human genome estimates approximately 120,000 genes. *Nat. Genet.* **25**, 239–240 (2000).
9. 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).
10. Velculescu, V.E., Zhang, L., Vogelstein, B. & Kinzler, K.W. Serial analysis of gene expression. *Science* **270**, 484–487 (1995).
11. Lal, A. *et al.* A public database for gene expression in human cancers. *Cancer Res.* **59**, 5403–5407 (1999).
12. Caron, H. *et al.* The human transcriptome map: clustering of highly expressed genes in chromosomal domains. *Science* **291**, 1289–1292 (2001).
13. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**, 78–94 (1997).
14. Polyak, K., Xia, Y., Zweier, J.L., Kinzler, K.W. & Vogelstein, B. A model for p53-induced apoptosis. *Nature* **389**, 300–304 (1997).
15. Adams, M.D. *et al.* Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature* **377**, 3 ff. (1995).
16. Okubo, K., Yoshii, J., Yokouchi, H., Kameyama, M. & Matsubara, K. An expression profile of active genes in human colonic mucosa. *DNA Res.* **1**, 37–45 (1994).
17. Shoemaker, D.D. *et al.* Experimental annotation of the human genome using microarray technology. *Nature* **409**, 922–927 (2001).
18. Boyd, A.C., Charles, I.G., Keyte, J.W. & Brammar, W.J. Isolation and computer-aided characterization of *Mme*I, a type II restriction endonuclease from *Methylophilus methylotrophus*. *Nucleic Acids Res.* **14**, 5255–5274 (1986).
19. Tucholski, J., Skowron, P.M. & Podhajska, A.J. Mmel, a class-IIS restriction endonuclease: purification and characterization. *Gene* **157**, 87–92 (1995).

Quantitative proteome analysis by solid-phase isotope tagging and mass spectrometry

Huilin Zhou, Jeffrey A. Ranish, Julian D. Watts, and Ruedi Aebersold*

The adaptation of sequences of chemical reactions to a solid-phase format has been essential to the automation, reproducibility, and efficiency of a number of biotechnological processes including peptide and oligonucleotide synthesis and sequencing[1–4]. Here we describe a method for the site-specific, stable isotopic labeling of cysteinyl peptides in complex peptide mixtures through a solid-phase capture and release process, and the concomitant isolation of the labeled peptides. The recovered peptides were analyzed by microcapillary liquid chromatography and tandem mass spectrometry (μLC-MS/MS) to determine their sequences and relative quantities. The method was used to detect galactose-induced changes in protein abundance in the yeast *Saccharomyces cerevisiae*. A side-by-side comparison with the isotope-coded affinity tag (ICAT) method[5] demonstrated that the solid-phase method for stable isotope tagging of peptides is comparatively simpler, more efficient, and more sensitive.

We devised a method for site-specific, stable isotopic labeling of cysteinyl peptides using a solid-phase isotope tagging reagent (Fig. 1). The *o*-nitrobenzyl-based photocleavable linker was first attached to aminopropyl-coated glass beads by solid-phase peptide synthesis[6]. Next, the isotope tag, a leucine molecule containing either seven hydrogen (d0) or seven deuterium atoms (d7), was attached to the photocleavable linker, again by solid-phase peptide synthesis[6]. Finally, a sulfhydryl-specific iodoacetyl group was attached. Cysteinyl peptides from two samples to be compared were covalently captured on the solid phase containing isotopically heavy or normal tag. The beads were then combined, washed, and exposed to UV light (360 nm, chosen to minimize any possible photocatalyzed side reactions). This resulted in photocleavage of the linker and the transfer of isotope tags from the solid phase onto the side chain of cysteine residues. Finally, recovered tagged peptides were analyzed by μLC-MS/MS to determine the sequence and relative abundance of each peptide, essentially as described previously[5].

To illustrate the efficiency of the capture and release reactions, we used a mixture consisting of a cysteine-containing laminin B peptide and the non-cysteine-containing phosphoangiotensin (Fig. 2). Laminin B was quantitatively captured onto the solid phase (compare Fig. 2A, 2B). After 1 h of photocleavage, the tagged laminin B was recovered; it showed the expected mass

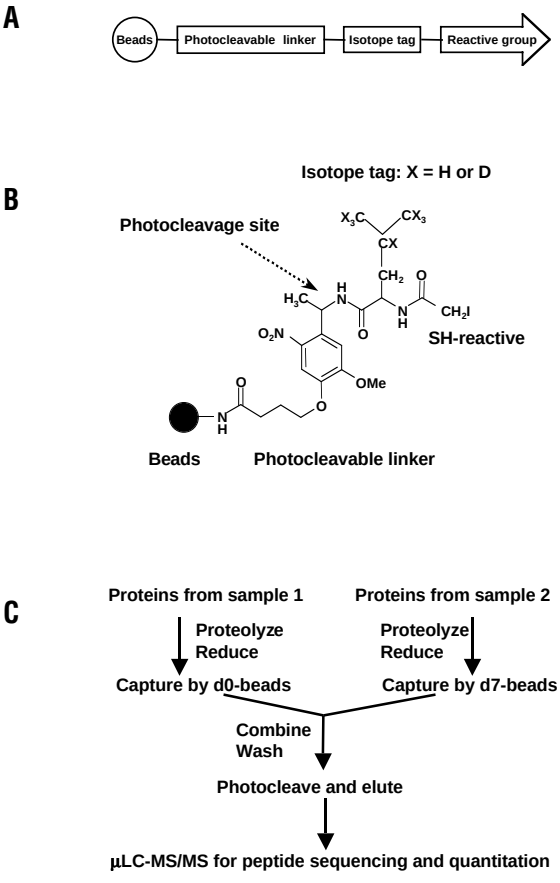

Figure 1. Schematic representation of the solid-phase isotope tagging method. (A) Modular composition of the solid-phase isotope tagging reagent, consisting of four elements: beads, photocleavable linker, stable isotope tag, and specific reactive group. (B) Chemical composition of the sulfhydryl (SH)–reactive solid-phase isotope tagging reagent. The *o*-nitrobenzyl-based photocleavable linker was coupled to aminopropyl glass beads. Peripheral to the photocleavable linker, a leucine molecule containing either 7 hydrogen (H) or 7 deuterium atoms (D), indicated by 'X', was attached as the isotope tag, followed by an iodoacetyl group as the SH-reactive group. (C) Strategy for quantitative protein analysis. Two protein samples to be compared were subjected to proteolysis. The Cys-containing peptides were reduced and captured by beads carrying either the d0-leucine or d7-leucine tag. The beads were then combined and, after stringent washing of the beads, the tagged peptides were released by photocleavage and analyzed by μLC-MS/MS.

Institute for Systems Biology, 1441 North 34th Street, Seattle, WA 98103-8904.
**Corresponding author (raebersold@systemsbiology.org).*