Whole-genome expression profiling through fragment display and combinatorial gene identification

Ats Metsis, Ulf Andersson, Göran Baurén, Patrik Ernfors¹, Peter Lönnerberg, Andreas Montelius¹, Mats Oldin, Arno Pihlak and Sten Linnarsson*

Global Genomics AB, Tomtebodavägen 21B, SE-171 77 Stockholm, Sweden and ¹Laboratory for Molecular Neurobiology, Department of Medical Biochemistry and Biophysics, Karolinska Institutet, Scheeles väg 1, SE-171 77 Stockholm, Sweden

Received February 25, 2004; Revised June 15, 2004; Accepted August 20, 2004

ABSTRACT

There is a growing demand for highly parallel gene expression analysis with whole genome coverage, high sensitivity and high accuracy. Open systems such as differential display are capable of analyzing most of the expressed genome but are not quantitative and generally require manual identification of differentially expressed genes by sequencing. Closed systems such as microarrays use gene-specific probes and are, therefore, limited to studying specific genes in well-characterized species. Here, we describe Tangerine, a PCR-based system that combines the scope and generality of open systems with a robust and immediate identification algorithm using publicly available sequence information. By combinatorial analysis of three independent and complete DNA indexing profiles, each displaying the complete set of expressed transcripts on capillary electrophoresis, the method allows transcripts to be simultaneously quantified and identified. The method is sensitive, accurate and reproducible, and is amenable to high-throughput automated operation.

INTRODUCTION

A number of methods for detection and quantification of gene expression are available. These include, for instance, northern blot analysis (1), real-time PCR (2), serial analysis of gene expression (SAGE) (3) and sequencing of cDNA libraries (4). Differential display (5) and related technologies which are based on generic primers require no prior sequence information to execute the experiment, and the identity of the genes which are under study can be determined by sequencing excised fragments. In contrast, microarrays (6,7) are based on specific hybridization on cDNAs or oligonucleotides arrayed at high density on a solid support, and, therefore, require no separate identification step.

In addition to these established technologies, a number of alternative approaches have been demonstrated [reviewed in (8)]. For example, adaptor-mediated PCR (9,10) has been used to improve the reproducibility of differential display, and

minisequencing (8) and universal microarray hybridization (11) have been used to automate identification of the displayed fragments. SAGE has also inspired alternatives such as massively parallel signature sequencing (MPSS) (12) and long-SAGE (13).

A recent, large-scale cDNA sequencing project (14) reported that the total number of transcriptional units in the mouse may be as high as 70 000 when non-coding RNAs are included. Although the true number of genes may never be known, this result also underscores the usefulness of open, unbiased methods of whole-genome expression analysis that avoid a limited set of gene-specific measurements.

For these purposes we have developed an improved gene expression analysis system. Our main objectives were as follows:

- (i) *Comprehensive, whole-genome coverage* avoiding the artificial bias introduced by using a set of gene-specific probes. To achieve this, we use an adapter-mediated fragment display method designed to permit systematic detection of >95% of all expressed transcripts.
- (ii) *Sensitive and robust quantitation*. To achieve this, we use competitive PCR, in which multiple transcripts can be co-amplified without introducing non-linear amplification artifacts (15).
- (iii) *Direct identification of known genes*. Although fragment display would yield an unbiased pattern of expressed transcripts, we also wanted to know the identities of genes wherever possible. To achieve this, we designed a combinatorial identification algorithm operating on three independent fragment display profiles each displaying a single (but different) fragment from each transcript. As a bonus, the built-in technical triplicate provides a more robust quantification.

A schematic illustration of the method is summarized in Figure 1a. cDNA was synthesized and then indexed (16) with one of three Type IIS restriction enzymes. Such enzymes cut outside their recognition motif, leaving in each case a 4 bp cohesive end specific to the upstream sequence. Because there are 256 possible 4 bp cohesive ends, each restriction product was split in 256 subreactions and ligated to one of 256 different adaptors, each carrying a different cohesive end but the same universal PCR primer template sequence. The ligated

*To whom correspondence should be addressed. Tel: +46 850 884 718; Fax: +46 830 1750; Email: slinnarsson@globalgenomics.com

Figure 1. Sensitive and reproducible fragment display. (a) Overview of the method. An aliquot of 300 ng of total RNA was cleaved separately with three restriction enzymes and the resulting fragments were indexed using 256 different adaptors, amplified in two rounds of PCR and separated by capillary electrophoresis. An example of the raw data generated by the method is shown in (b) for a single chromatogram across four samples, with the identified transcripts labeled. The samples were from the Myoblast experiment and the *Cdkn1* gene shows differential expression.

products were purified and amplified using nested PCR. The second PCR round used an anchored and labeled 3' primer that served to ensure that fragments were generated starting exactly adjacent to the polyadenylation sequence, and providing the identity of the terminal nucleotide. The products were then separated by capillary electrophoresis. Supplementary Figure S1 shows 256 out of 2304 chromatograms (a single enzyme and anchored primer combination) from a complete Tangerine expression profile.

As a result, any given transcript was potentially displayed three times, once for each of the three restriction enzymes. Each time, a detectable fragment should occur in only one of the 256 subreactions, namely the one that received an adapter complementary to the cohesive end generated by the enzyme. The indexing procedure thus revealed 10 bp of sequence information for each fragment (5 bp from the restriction enzyme recognition motif, 4 bp cohesive end and 1 bp anchor base), and the fragment length. However, 10 bp is not unique in a mammalian transcriptome. The problem was solved by combining information from the three different restriction enzymes. A combinatorial identification approach, based on three analyses of 10 bp per gene (Figure 4a; discussed below), was used to simultaneously quantify and identify virtually all genes.

MATERIALS AND METHODS

Tangerine experimental protocol

For each tangerine profile, 300 ng of mouse RNA was used. cDNA synthesis was performed with the SMART cDNA synthesis kit (Clontech) according to the manufacturer's instructions, except that the 3' primer was replaced by 5'-AAG CAG TGG TAT CAA CGC AGA GTT CAC GCT GGA CTG TTT CGG TTT TTT TTT TTT TTT TTT TTT TV-3' (note: we have preliminary data to suggest that the anchor base on this primer contributes to shadow peaks and would better be omitted in future experiments). The cDNA was then cleaved with 4 U of a Type IIS enzyme (FokI from Roche, BbvI and BsmAI from NEB) in 30 µl of the appropriate buffer for 2 h. After phenol-chloroform extraction and precipitation, 1.28 µg of each cleaved template was

resuspended in 280 µl 10 mM Tris, pH 8. An aliquot of 240 µl T4 DNA ligase buffer (Fermentas), 480 µl 0.25 M NaCl with 25% PEG-6000, 300 µl internal control fragment (adjusted by titration to yield an easily detectable peak after electrophoresis) and 500 µl water was added and the mix was distributed across each well of two 384-well microtiter plates containing predisposed adaptors (5'-biotin-AGG ACA TTT GTG AGT CAG GCG TGT CTT GGA TGC-3' and 5'-NNN NGC ATC CAA GAC ACG CCT GAC TCA CAA ATG TCC T-3', where N represents specific bases varied across the set of adaptors in all permutations). An aliquot of 600 µl ligase buffer, 600 µl T4 DNA Ligase (Fermentas), 1200 µl 0.25 M NaCl with 25% PEG-6000 and 3600 µl water was mixed and distributed across the two plates. After incubation at 37°C for 16 h, 20 µl paramagnetic beads (Dyna Dynabeads M280) were added to each well and the samples were washed three times in bind/wash buffer (10 mM Tris, pH 7.5, 1 mM EDTA, 2 M NaCl), ending up with dry beads retaining ligated fragments.

A first round of PCR (30 cycles of 94°C 30 s, 58°C 30 s, 72°C 60 s) was performed using 8 pmol each of an outer primer pair (5'-TTC ACG CTG GAC TGT TTC GG-3' and 5'-AGG ACA TTT GTG AGT CAG GC-3') and 48 nl TaqExpress (25 U/µl, GENETIX, UK). The samples were then diluted 1:100 and a second round of PCR (30 cycles of 94°C 30 s, 43°C 30 s, 72°C 60 s) was performed using 8 pmol each of an inner primer pair (5'-GTG TCT TGG ATG C-3' and labeled downstream primers 5'-Rox-TTT T5*T TTT TTT TTT TTT TTT TG-3', 5'-Hex-TTT TTT TTT TTT TTT TTT TTT TTT TTC-3' and 5'-Tama-TTT TT5*TTT TTT TTT TTT TTT TTT TTA-3' where 5* denotes dT-fluorescein) and 48 nl TaqExpress. An aliquot of 2.2 µl of the product was transferred into 20 µl formamide (HiDi, Applied Biosystems) containing size marker and separated on a MegaBACE 4000 capillary electrophoresis instrument (384 capillaries; Amersham).

Starting with adaptor ligation, the entire procedure was automated on a Biomek FX (Beckman Coulter) liquid handling robot in 384-well microtiter plate format.

Signal processing

First, raw fluorescence data from the electrophoresis was noise-reduced and baseline-corrected. Each fragment was

corrected for differential migration introduced by the fluorochrome and converted to base pair coordinates using the size marker. In control experiments, this procedure correctly sized fragments to within 1 bp across the range 50–1000 bp (data not shown).

Second, a corrected quantity for each fragment was calculated as follows. Peaks were located and a preliminary area was calculated by approximation to a triangular shape. Each quantity was then normalized to the nearest size marker fragment, to the experimentally determined DNA polymerase efficiency and to the internal control (in that order).

Signal processing thus generated a set of fragments, each associated with a length, an anchor base and a quantity, and with the enzyme and adaptor that was used to generate it.

Database construction

A sequence database was constructed from several sources, including annotated GenBank (17) sequences, Riken (14), Unigene and RefSeq (18) transcripts as well as internally generated 3' end cDNA sequences.

Sequences were aligned to each other, and when possible, ambiguities were corrected. Overlapping sequences were joined into longer transcripts. Branching alignments (due to e.g. differential splicing) were split into separate transcripts.

Poly(A) sites were marked based on annotation in the source database, poly(A) tails, alignment of multiple 3' ends of expressed sequence tags (ESTs) and the occurrence of poly(A) signals. Each putative poly(A) site was assigned a quality score according to the strength of supporting evidence, and the scores were used to adjust the stringency of gene assignment. In addition, internal poly(A) stretches were marked as potential internal priming sites and used downstream as if they were true polyadenylation sites.

Samples

The technical analyses in the present paper are based on several experiments, each of which are presented here as control versus treatment two-group experiments. First, the mouse myoblast cell line C2C12 (19) was grown to confluence and differentiated in 2% horse serum ('Myoblast' sample: control 0 h, treatment 24 h after serum withdrawal). Second, a mouse primary cortical neuron cell culture was depolarized with potassium ('Neuron': control low potassium, treatment high potassium). Third, mice fasted for 24 h were returned to feed ad libitum for 0, 2, 6 or 24 h and liver samples were analyzed ('Liver1' to 'Liver4': the control in each case was given food ad libitum, the treatment was fasted, then fed for 0, 2, 6 or 24 h). Detailed biological analysis of these experiments will be published elsewhere.

Gene assignment and comparative analysis

In a first pass, the database was scanned and candidate transcripts were assigned to each peak. In order to permit statistical evaluation, corresponding peaks in each comparative analysis were then aligned using dynamic programming, as follows.

A pairwise optimal global alignment similar to that used for pairwise alignment of DNA sequences and related to (20) was used. In such an algorithm, the optimal alignment is found by recursively applying a set of penalties: for match, for gap in the upper sequence and for a gap in the lower sequence. In our

case, the sequence was the ordered set of peaks in a chromatogram. Gap and match penalties were calculated as functions of the position of the peaks under consideration.

The match penalty $Z(i,j)$ for peaks i and j was taken as

$$Z(i,j) = |L_i - L_j|^{s_d},$$

and the gap penalty $Z(i)$ as

$$Z(i) = m_g + k_g L_i,$$

where L_i is position of peak i (in base pairs). Furthermore, there were two normalization constants k_g and k_r , and a scaling constant s_d . We used $m_g = 0$, $k_g = 0.005$ and $s_d = 2$ in all experiments. The fact that gaps are penalized more for longer fragments means that longer fragments will tend to be matched instead of gapped. Multiple alignments were obtained by piling up an arbitrary sequence of pairwise alignments.

Peak areas were then normalized using Tukey's one-step biweight method (21). A regularized t -test (22) was applied to each aligned set of peaks. Transcripts were reported as significantly regulated at $x\%$ only when assigned to at least two significantly ($P < x$) regulated peaks with consistent direction of change. The estimated false-positive rate of *change* was calculated as $3x^2/2$. The false-positive rate of *assignment* was determined empirically using 300 yeast genes included as negative controls in the database.

Real-time PCR

Transcript sequences were extracted from the database and primers were automatically designed (PrimerExpress, Applied Biosystems) to a target within 1 kb of the 3' end. Each sample was amplified in SYBR Green Master Mix (Applied Biosystems) in three 2-fold dilutions, each in triplicate, allowing the primer efficiency to be determined for each primer pair. A no-template control was included for each primer pair. Amplifications showing low efficiency, too high efficiency, extremely low RNA abundance (i.e. close to the no-template control) or where melting curves indicated more than one product were discarded. Quantitation was performed relative to two separate triplicate amplifications of *gapd* at each end of the 1 kb 3' end interval, to verify that cDNA synthesis had effectively spanned the interval.

Spiking experiment

Synthetic mRNA was generated by *in vitro* transcription from a cloned, polyadenylated cDNA as follows. pCR4-Topo plasmids (Invitrogen) containing *Saccharomyces cerevisiae* fragments E7 (accession no. NP_015262) and E11 (accession no. NP_012049) were linearized with Not I. Synthetic polyadenylated RNA was produced with the Promega RiboMax large-scale RNA production system. The concentration of the produced RNA was measured and dilutions with a 3.2 step were made from 2.8×10^9 to 8500 molecules per sample, corresponding to a range from 14 to 5×10^{-8} copies/cell (assuming cells contained 300 000 mRNA copies, of average length 2 kb).

The synthetic mRNA was spiked into a series of identical 2.5 μ g mouse normal liver mRNA samples (Clontech) and the resulting spiked samples were subjected to Tangerine as

described above. The synthetic mRNA peak was then identified in each sample and plotted against the copy number. Non-linear regression was used to fit a four-parameter sigmoidal equation:

$$Z = \frac{a}{1 + \left(\frac{k}{c}\right)^n} + m,$$

where Z is the measured peak area and c is the concentration of spiked mRNA (here expressed as 10^6 of molecules per sample). The results obtained with the two different mRNAs were similar ($a = 660\,000$, $k = 55$, $n = 1.25$ and $m = 0$ for E7 and $a = 805\,000$, $k = 33$, $n = 1.23$ and $m = 0$ with E11; E7 shown in the figure).

RESULTS

In order to determine the sensitivity of the method, we spiked mouse liver RNA samples with a dilution series of two synthetic polyadenylated mRNAs (*S.cerevisiae* fragments E7 and E11). Shown in Figure 2 is the expression level

Figure 2. The sensitivity of the procedure was assessed by spiking synthetic mRNA, showing accurate detection down to 0.04 copies/cell (slightly more than 1 molecule per ten million) and approximately three orders of magnitude of dynamic range. Each data point shows a single peak generated from a spiked mRNA, except the leftmost (lowest concentration) where the peak was indistinguishable from background noise. The curve shows a four-parameter sigmoidal fit (Materials and Methods).

(i.e. area of corresponding peak) versus the copy number per cell for the E7 fragment. We detected transcripts down to 7.5 parts per ten million, and results from the two different fragments were similar. Since a typical mammalian cell contains 300 000 RNA molecules (23), a sensitivity of 7.5×10^{-6} corresponds to 0.04 copies/cell.

The use of competitive PCR also resulted in highly reproducible quantitation. Figure 1b shows chromatograms from four samples belonging to two groups. The replicas showed high within-group overlap of genes expressed at both high- and low levels. To further examine the reproducibility of the method, we plotted the log-ratio for a replicated sample versus the log of the expression level (Figure 3a). The distribution was similar across several orders of magnitude with no tendency for skewness and no evidence of outliers. As a result there was no need to normalize the samples using a non-linear regression method [e.g. lowess (24)] as is often done with microarray data. The variance of the method was also significantly lower than the inherent biological variance, as shown in Figure 3b by the same plot of a biological replicate. While the technical replicate shows the variability of the complete Tangerine procedure operating on identical RNA samples, the biological replicate adds the biological variability inherent in any expression profiling experiment. Coefficients of variation were 0.24 for the technical and 0.33 for the biological replicate. The difference was evident across the whole range of expression levels, but was most pronounced at mid- to high levels. Note that these plots show the reproducibility of individual peaks, i.e. one of the three fragments generated for each gene. When performing downstream analyses, the built-in technical triplicate adds to the statistical power of the method.

The use of the PCR fragment display provided the basis for direct identification of differentially expressed transcripts. We combined three independent fragment display profiles to unambiguously identify fragments. As an example, Figure 4a shows how *glutathione peroxidase 1* (*gpx1*) gave rise to three fragments with characteristic length and inferred sequence information. The large number of subreactions (256 per enzyme) and the use of distinct fluorophores to distinguish anchor bases with each subreaction ensured that only a small number of fragments would be displayed in each individual chromatogram, which aided in separation and unambiguous

Figure 3. Reproducibility of the method. The figure is an R-I (ratio-intensity) plot of a technical (a) and a biological (b) replicate. Each data point shows the base two logarithm of the ratio of a peak identified in the two replicates plotted versus the base ten logarithm of the expression level. Both experiments were performed on mouse liver samples obtained from a single animal and split after RNA isolation (a) or from two independent animals (b).

Figure 4. Robust gene identification and comparative analysis. (a) Expressed gene transcripts were identified using the three predicted fragments obtained from the three independent enzyme profiles (FokI, BbvI and BsmAI). In the case of *glutathione peroxidase 1* (*gpx1*), the predicted fragments were 389 bp with cohesive end CACC (for BsmAI), 219 bp with cohesive end ACCA (for BbvI) and 118 bp with cohesive end CCCA (for FokI), in all cases with anchor base G. The figure shows the corresponding chromatograms, and the arrows point to the peaks used for assignment. (b) A scatterplot of transcripts from experiment 1 showed that 2-fold changes were significant for low-expressed genes, while as little as 30% change was detectable for highly expressed genes ($P < 0.005$ in both cases; the P -value was adjusted to compensate for multiple testing to give a false discovery rate of $\sim 10\%$). (c) Overview of a typical Tangerine experiment. As shown, we lost 15–20% of all chromatograms in quality controls prior to gene assignment. Peaks from the remaining chromatograms were assigned to a transcript database containing 30 919 genes, 94% of which were assignable [at least two predicted fragments and one poly(A) site], resulting in about 13 000 assigned genes with 98% concordance. After comparative analysis, 602 regulated genes were found. (d) Accuracy of change call was measured by real-time PCR, showing that 'up', 'down' and 'no change' calls were made with 78, 66 and 82% accuracy, respectively.

gene assignment. We assigned genes when there were peaks within a window of 2 bp plus 0.15% of the predicted fragment lengths in two out of three enzyme profiles (to allow for assignment of low-expressed genes whose peaks were close to the detection threshold). These criteria were chosen to be generous in order to avoid false negatives (see further below). The sample workflow shown in Figure 4c gives a sense of the scope of the method. Individual chromatograms for the full experiment (Myoblast) were automatically analyzed and approximately 100 000 peaks were extracted, giving an average of about 40 peaks per chromatogram. This set of peaks was then automatically compared with a transcript database containing slightly less than 30 000 assignable genes, resulting in approximately 13 000 assignments in each sample. Figure 4b shows a scatterplot of the experiment, with highly significant regulated transcripts indicated in red. In this experiment, changes in both low- and highly expressed genes could be detected at <2 -fold with $P < 0.005$.

In order to assess the performance of Tangerine for comparative analysis, we performed several experiments which are summarized in Table 1.

First, we addressed the specificity of gene assignment, using a set of yeast genes included in our gene database as negative controls. We wanted to measure the positive predictive value (PPV; true positives divided by all positives), as this measure takes into account the fact that a large number of genes are being simultaneously analyzed, and often a set of regulated genes are selected for further analysis. It is then important that this list contains as few false positives as possible. The parameters of assignment were optimized to provide a satisfactory compromise between sensitivity and specificity among regulated genes. We judged a transcript to be regulated if it was assigned to at least two peak positions that were significantly ($P < 0.025$) and consistently regulated. Based on the number of assigned negative controls, the false-positive rate was always $<2\%$ (corresponding to a P -value of <0.02).

Table 1. The performance of Tangerine for comparative analysis

Parameter	Myoblast	Liver1	Liver2	Liver3	Liver4	Neuron
Number of samples per group	4	3	3	3	3	4
Average number of assigned genes	12 931	13 550	13 020	13 381	13 141	16 475
Average % pos ctrl genes (60)	93	93	93	95	93	97
Average % neg ctrl genes (300)	22	21	20	20	21	30
<i>P</i> -value used	0.065	0.025	0.025	0.025	0.025	0.025
Number regulated	602	1517	713	767	1183	938
% Up-regulated	50	80	54	80	43	84
Estimated specificity of change call ^a	86	99	98	98	99	98
Specificity of assignment ^a	84	81	93	94	92	68
% PCR verified (number of assays) ^b	72 (71)	88 (8)	63 (8)	75 (8)	100 (8)	N/A

N/A not analyzed.

^aSee Materials and Methods, section 'Gene assignment and comparative analysis'.^bBoth regulated and unregulated genes were designated for verification by real-time PCR.

We then estimated the PPV for assignment by assuming that the whole gene database would have a false-positive rate equal to that of the negative control genes. For example, in the Myoblast experiment a single negative control was assigned, i.e. a rate of 0.3%. We would then expect 0.3% of all 29 185 assignable genes, or about 97 genes to be assigned by chance. The PPV of assignment was therefore 84% (1-97/602) and this rate was similar for most experiments.

Judging assignment parameters and setting the peak detection threshold is thus an important part of designing and analyzing a Tangerine experiment. The assignment criteria used here required consistent peak patterns for all three fragments analyzed for each gene. This caused an enrichment for correct assignments because false positives were more likely to be assigned to inconsistent peak patterns, and different settings could have been used if the goal had been to analyze *non-regulated* genes.

Second, we examined the PPV of change calls. The various experiments show different patterns both in terms of the number of genes which were regulated, and in the balance between up- and down-regulation. We estimated the false-positive rate simply by the *P*-value (i.e. $P < 0.05$ implies 5% false positives). We then calculated the PPV of change call as the number of true-regulated genes divided by the sum of true- and false-regulated genes. In each case, we found that the PPV exceeded 95%.

Third, we sought to experimentally verify these specificities. We performed real-time PCR on a selection of genes in each experiment. We were able to verify 72% of change calls tested in experiment 1 (based on the criterion that real-time PCR should show a fold change of >25% with $P < 0.05$ using Student's *t*-test, and in the right direction), which is consistent with the combined PPV of change and assignment (84 and 86%, respectively, giving an estimated overall PPV of 72%).

In conclusion, we have shown by a variety of approaches, both statistical and experimental, that Tangerine is capable of quantifying and correctly identifying regulated genes. Based on the summary of all our real-time PCR experiments in Figure 4d, we were in most cases able to confirm up-, down- and no regulation, averaging 75% success across all experiments.

DISCUSSION

We have described an open system of gene expression profiling for all transcripts with immediate and robust identification

of known genes. The method achieves high sensitivity by using PCR. The use of competitive PCR with a single primer pair (8), contributes to its low variability.

Tangerine provides a method for the discovery of new functions of known genes and the discovery of unknown differentially expressed genes. The quality of gene identification and quantification is not affected down to 100 ng of total RNA as starting material, and the method is among the most sensitive available (i.e. close to 1 part per ten million as shown by the spiking experiment). Such high sensitivity can be useful especially when studying low-abundant genes and heterogeneous cell populations where a gene of interest may be expressed in only a fraction of the cells represented in a sample.

Based on simple *in vitro* reactions, Tangerine is simple to scale up to high-throughput automated operation using a pipetting robot and capillary sequencer. The whole process of Tangerine including gene identification presently takes 48 h with a throughput of two samples per day. The maximal throughput is limited by the 2.5 h run time on the 384-capillary sequencer, and since each sample involves two 384-well plates the practical maximum is four samples in 24 h.

These qualities of Tangerine should make it an attractive alternative to microarrays for many important scientific problems. Recent estimates of the transcriptional potential of the genome hint at a very large unexplored transcriptome (14,25). The comprehensive coverage offered by Tangerine is rivaled only by SAGE and direct clone sequencing, both highly expensive and slow methods. This feature should be especially valuable when looking for diagnostic or prognostic biomarkers. The ability to directly identify novel transcripts by sequencing, combined with sensitivity to rare transcripts, provides assurance that no markers will be missed.

Because Tangerine displays actual gene fragments, newly discovered transcripts can easily be analyzed in old experiments simply by rerunning gene assignment. A further advantage of the Tangerine system is its broad applicability across species. Indeed any species with sequence data, no matter how little or how much, is immediately available for whole-genome expression profiling.

Recent work on expression profiling methods derived from differential display has focused on improving the gene identification step to avoid the laborious direct sequencing of excised fragments. GeneCalling (26) and TOGA (9) use database lookup to predict fragment identities; however, both suffer from too little band separation which leads to ambiguous

assignments and thus still requires confirmatory experiments for each gene of interest. For example, TOGA separates the whole transcriptome into 256 fractions, while Tangerine uses 768 fractions per profile, and, in addition, Tangerine fragments are twice as long on average. In spite of this, we find that a single profile is not enough to unambiguously identify a gene, and thus always perform three complete profiles (affording a built-in technical triplicate as an added benefit) and use a combinatorial identification scheme.

A different solution to the identification problem was recently proposed (11) using universal hexamer microarrays to read out each fraction. Conceptually, the approach is akin to replacing the electrophoretic separation of fragments in Tangerine with a microarray hybridization step; with the main drawback being that a very large number of arrays (e.g. 80 to 120) are needed to analyze a single sample.

In conclusion, Tangerine is the first open gene expression analysis system that combines high throughput and comprehensive scope with reasonable cost and a direct and robust identification algorithm.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

We thank Tobias Andersson, Oivvio Polite, Måns Sandström, Johanna Sagemark and Ylva Linderstam for the bioinformatics work, Staffan Ahlandsberg, Eva Büren, Zhila Hosseinejad, Christina Österlund and Anette Selander for running and optimizing Tangerine, Lena Nyberg for project management, Pekka Ojala and Paula Jalonen for setting up the sequence verification workflow, and Biovitrum for the mouse liver mRNA samples.

REFERENCES

- Alwine, J.C., Kemp, D.J. and Stark, G.R. (1977) Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. *Proc. Natl Acad. Sci. USA*, **74**, 5350–5354.
- Heid, C.A., Stevens, J., Livak, K.J. and Williams, P.M. (1996) Real time quantitative PCR. *Genome Res.*, **6**, 986–994.
- Velculescu, V.E., Zhang, L., Vogelstein, B. and Kinzler, K.W. (1995) Serial analysis of gene expression. *Science*, **270**, 484–487.
- Adams, M.D., Dubnick, M., Kerlavage, A.R., Moreno, R., Kelley, J.M., Utterback, T.R., Nagle, J.W., Fields, C. and Venter, J.C. (1992) Sequence identification of 2375 human brain genes. *Nature*, **355**, 632–634.
- Liang, P. and Pardee, A.B. (1992) Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science*, **257**, 967–971.
- Schena, M., Shalon, D., Davis, R.W. and Brown, P.O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467–470.
- Fodor, S.P., Read, J.L., Pirrung, M.C., Stryer, L., Lu, A.T. and Solas, D. (1991) Light-directed, spatially addressable parallel chemical synthesis. *Science*, **251**, 767–773.
- Scheel, J., Von Brevern, M.C., Horlein, A., Fischer, A., Schneider, A. and Bach, A. (2002) Yellow pages to the transcriptome. *Pharmacogenomics*, **3**, 791–807.
- Sutcliffe, J.G., Foye, P.E., Erlander, M.G., Hilbush, B.S., Bodzin, L.J., Durham, J.T. and Hasel, K.W. (2000) TOGA: an automated parsing technology for analyzing expression of nearly all genes. *Proc. Natl Acad. Sci. USA*, **97**, 1976–1981.
- Bachem, C.W., van der Hoeven, R.S., de Bruijn, S.M., Vreugdenhil, D., Zabeau, M. and Visser, R.G. (1996) Visualization of differential gene expression using a novel method of RNA fingerprinting based on AFLP: analysis of gene expression during potato tuber development. *Plant J.*, **9**, 745–753.
- Roth, M.E., Feng, L., McConnell, K.J., Schaffer, P.J., Guerra, C.E., Affourtit, J.P., Piper, K.R., Guccione, L., Hariharan, J., Ford, M.J. *et al.* (2004) Expression profiling using a hexamer-based universal microarray. *Nat. Biotechnol.*, **22**, 418–426.
- Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D.H., Johnson, D., Luo, S., McCurdy, S., Foy, M., Ewan, M. *et al.* (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotechnol.*, **18**, 630–634.
- Saha, S., Sparks, A.B., Rago, C., Akmaev, V., Wang, C.J., Vogelstein, B., Kinzler, K.W. and Velculescu, V.E. (2002) Using the transcriptome to annotate the genome. *Nat. Biotechnol.*, **20**, 508–512.
- Carninci, P., Waki, K., Shiraki, T., Konno, H., Shibata, K., Itoh, M., Aizawa, K., Arakawa, T., Ishii, Y., Sasaki, D. *et al.* (2003) Targeting a complex transcriptome: the construction of the mouse full-length cDNA encyclopedia. *Genome Res.*, **13**, 1273–1289.
- Siebert, P.D. and Larrick, J.W. (1992) Competitive PCR. *Nature*, **359**, 557–558.
- Unrau, P. and Deugau, K.V. (1994) Non-cloning amplification of specific DNA fragments from whole genomic DNA digests using DNA 'indexers'. *Gene*, **145**, 163–169.
- Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. and Wheeler, D.L. (2003) GenBank. *Nucleic Acids Res.*, **31**, 23–27.
- Wheeler, D.L., Church, D.M., Federhen, S., Lash, A.E., Madden, T.L., Pontius, J.U., Schuler, G.D., Schriml, L.M., Sequeira, E., Tatusova, T.A. *et al.* (2003) Database resources of the National Center for Biotechnology. *Nucleic Acids Res.*, **31**, 28–33.
- Goetsch, S.C., Hawke, T.J., Gallardo, T.D., Richardson, J.A. and Garry, D.J. (2003) Transcriptional profiling and regulation of the extracellular matrix during muscle regeneration. *Physiol. Genomics*, **14**, 261–271.
- Aittokallio, T., Ojala, P., Nevalainen, T.J. and Nevalainen, O. (2001) Automated detection of differently expressed fragments in mRNA differential display. *Electrophoresis*, **22**, 1935–1945.
- Hoaglin, D.C., Mosteller, F. and Tukey, J.W. (2000) Understanding robust and exploratory data analysis. Wiley classics library ed., Wiley, NY.
- Baldi, P. and Long, A.D. (2001) A Bayesian framework for the analysis of microarray expression data: regularized *t*-test and statistical inferences of gene changes. *Bioinformatics*, **17**, 509–519.
- Alberts, B., Bray, D., Louis, J., Raff, M., Roberts, K. and Watson, J. (1995) *Molecular Biology of the Cell*, 3rd edn. Garland, NY.
- Cleveland, W.S. and Devlin, S.J. (1988) Locally weighted regression: an approach to regression analysis by local fitting. *J. Am. Stat. Assoc.*, **83**, s596–610.
- Kapranov, P., Cawley, S.E., Drenkow, J., Bekiranov, S., Strausberg, R.L., Fodor, S.P. and Gingeras, T.R. (2002) Large-scale transcriptional activity in chromosomes 21 and 22. *Science*, **296**, 916–919.
- Shinkets, R.A., Lowe, D.G., Tai, J.T., Sehl, P., Jin, H., Yang, R., Predki, P.F., Rothberg, B.E., Murtha, M.T., Roth, M.E. *et al.* (1999) Gene expression analysis by transcript profiling coupled to a gene database query. *Nat. Biotechnol.*, **17**, 798–803.