Published online 15 October 2008

Nucleic Acids Research, 2008, Vol. 36, No. 21 ***e141***
doi:10.1093/nar/gkn705

Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms

Peter A. C. 't Hoen[1,*], Yavuz Ariyurek[1], Helene H. Thygesen[1], Erno Vreugdenhil[2], Rolf H. A. M. Vossen[1], Renée X. de Menezes[1], Judith M. Boer[1], Gert-Jan B. van Ommen[1] and Johan T. den Dunnen[1]

[1]The Center for Human and Clinical Genetics and the Leiden Genome Technology Center, Leiden University Medical Center and [2]The Department of Medical Pharmacology from the Leiden/Amsterdam Center for Drug Research, Leiden, The Netherlands

Received August 12, 2008; Revised September 16, 2008; Accepted September 29, 2008

ABSTRACT

The hippocampal expression profiles of wild-type mice and mice transgenic for δC-doublecortin-like kinase were compared with Solexa/Illumina deep sequencing technology and five different microarray platforms. With Illumina's digital gene expression assay, we obtained ~2.4 million sequence tags per sample, their abundance spanning four orders of magnitude. Results were highly reproducible, even across laboratories. With a dedicated Bayesian model, we found differential expression of 3179 transcripts with an estimated false-discovery rate of 8.5%. This is a much higher figure than found for microarrays. The overlap in differentially expressed transcripts found with deep sequencing and microarrays was most significant for Affymetrix. The changes in expression observed by deep sequencing were larger than observed by microarrays or quantitative PCR. Relevant processes such as calmodulin-dependent protein kinase activity and vesicle transport along microtubules were found affected by deep sequencing but not by microarrays. While undetectable by microarrays, antisense transcription was found for 51% of all genes and alternative polyadenylation for 47%. We conclude that deep sequencing provides a major advance in robustness, comparability and richness of expression profiling data and is expected to boost collaborative, comparative and integrative genomics studies.

INTRODUCTION

Gene expression microarrays are at present the default technology for transcriptome analysis. Since they rely on sequence-specific probe hybridization, they suffer from background and cross-hybridization problems and measure only the relative abundances of transcripts (1). Moreover, only predefined sequences are detected. In contrast, tag-based sequencing methods like SAGE (Serial Analysis of Gene Expression) measure absolute abundance and are not limited by array content (2). However, laborious and costly cloning and sequencing steps have thus far greatly limited the use of SAGE. This has radically changed with the introduction of deep sequencing technology, enabling the simultaneous sequencing of up to millions of different DNA molecules. The shared idea behind the different deep sequencing approaches is the clonal detection of single DNA molecules at physically isolated locations(3–5). We used the Solexa/Illumina 1G Genome Analyzer, in which adapter sequences, ligated to both ends of the DNA molecule, are bound to a glass surface coated with complementary oligonucleotides. This is followed by solid-phase DNA amplification and sequencing-by-synthesis (6). The system yields millions of short reads (currently up to 36 bp), and is therefore very suitable for tag-based transcriptome sequencing. The technology is also referred to as Digital Gene Expression tag profiling (DGE), and is essentially an improved version of the earlier Massively Parallel Signature Sequencing (MPSS) technology(3,7).

The first steps of the procedure are similar to classical LONG-SAGE. Two restriction enzymes are used to generate tags, cutting at the most 3' CATG and 17 bp downstream of the first enzyme site. Unlike in classical SAGE,

*To whom correspondence should be addressed. Tel: +31 71 526 9421; Fax: +31 71 526 8285; Email: p.a.c.hoen@lumc.nl

tags are neither concatenated nor cloned, but sequenced immediately. The unprecedented sequencing depth now enables the analysis of individual biological samples, while pooling of samples was previously the only affordable option in SAGE. Our results include a striking example of the intrinsic hazards of pooling in expression profiling.

The biological question addressed in the current study was the identification of transcripts differentially expressed in the hippocampus between wild-type and transgenic mice overexpressing a splice variant of the doublecortin-like kinase-1 (*Dclk1*) gene. This splice variant, δC-doublecortin-like kinase (DCLK)-short, makes the kinase constitutively active (8), and causes subtle behavioral phenotypes (Schenk *et al.*, in preparation). The exact same RNA samples have been analyzed before on five different genome-wide microarray expression profiling platforms (9), which detected few differences in expression between the two groups. We report here that DGE detects a lot more small, yet significant differences between the two groups of mice, including those in antisense transcripts and transcripts with different 3′-untranslated regions (UTRs). Furthermore, we discuss the advantages of deep sequencing over microarray expression profiling.

MATERIALS AND METHODS

Samples

Wild-type male C57/BL6j and transgenic male mice overexpressing DCLK-short with a C57/BL6j background were individually housed 7 days prior to the start of the experiment. Animals were housed under standard conditions, 12 h/12 h light/dark cycle and had access to food and water *ad libitum*. Wild-type ($N = 4$) and transgenic ($N = 4$) tissue samples were collected by taking the brain from the skull and quickly dissecting out both hippocampi. Dissection was performed at 0° C to prevent degradation of RNA. Hippocampi were put directly in pre-chilled tubes containing Trizol reagent (Invitrogen Life Technologies, Carlsbad, CA, USA). All animal treatments were approved by the Leiden University Animal Care and Use Committee (UDEC# 01022).

RNA extraction

After transfer to ice-cold Trizol, hippocampi were homogenized using a tissue homogenizer (Salm&Kipp, Breukelen, The Netherlands) and total RNA was isolated according to the manufacturer's protocol. After precipitation, RNA was purified with Qiagen's RNeasy kit with on-column DNase digestion. The quality of the RNA was assessed with the RNA 6000 Labchip kit in combination with the Agilent 2100 Bioanalyzer (Agilent Technologies, Palo Alto, CA, USA), using the Eukaryote Total RNA Nano assay according to the manufacturer's instructions.

Sequence tag preparation

Sequence tag preparation was done with Illumina's Digital Gene Expression Tag Profiling Kit according to the manufacturer's protocol (version 2.1B). A schematic overview of the procedure is given in Supplementary Figure 1. One microgram of total RNA was incubated with oligo-dT beads to capture the polyadenlyated RNA fraction. First- and second-strand cDNA synthesis were performed while the RNA was bound to the beads. While on the beads, samples were digested with NlaIII to retain a cDNA fragment from the most 3′ CATG to the poly(A)-tail. Subsequently, the GEX adapter 1 was ligated to the free 5′ end of the RNA, and a digestion with MmeI was performed, which cuts 17 bp downstream of the CATG site. At this point, the fragments detach from the beads. After dephosphorylation and phenol extraction, the GEX adapter 2 was ligated to the 3′ end of the tag. A PCR amplifcation with 15 cycles using Phusion polymerase (Finnzymes) was performed with primers complementary to the adapter sequences to enrich the samples for the desired fragments. The resulting fragments of 85 bp were purified by excision from a 6% polyacrylamide TBE gel. The DNA was eluted from the gel debris with 1× NEBuffer 2 by gentle rotation for 2 h at room temperature. Gel debris were removed using Spin-X Cellulose Acetate Filter (2 ml, 0.45 μm) and the DNA was precipitated by adding 10 μl of 3 M sodium acetate (pH 5.2) and 325 μl of ethanol (–20°C), followed by centrifugation at 14 000 r.p.m. for 20 min. After washing the pellet with 70% ethanol, the DNA was resuspended in 10 μl of 10 mM Tris–HCl, pH8.5 and quantified the DNA with a Nanodrop 1000 spectrophotometer.

Sequencing using Solexa/Illumina Whole Genome Sequencer

Cluster generation was performed after applying 4 pM of each sample to the individual lanes of the Illumina 1G flowcell. After hybridization of the sequencing primer to the single-stranded products, 18 cycles of base incorporation were carried out on the 1G analyzer according to the manufacturer's instructions. Image analysis and basecalling were performed using the Illumina Pipeline, where sequence tags were obtained after purity filtering. This was followed by sorting and counting the unique tags. The raw data (tag sequences and counts) have been submitted to Gene Expression Omnibus (GEO) under series GSE10782.

Illumina DGE tag annotation

All tags were annotated using a database provided by Illumina. Briefly, a preprocessed database of all possible CATG + 17-nt tag sequences was created, using mouse genome (mm8 version from UCSC site) and mouse transcriptome (all refseq, mRNA and ESTs found in GenBank as of November 2006 and Unigene version Mm159). All tags were classified based on the location and orientation in the original sequence as outlined in Supplementary Table 1. The genome was used as a backbone for tag clustering, using tag per genome position as a unique key. Best possible 'local' annotation was chosen for each genome location. Finally, best annotation for each distinct tag sequence was chosen based on quality of local annotation and number of transcripts in that location. The total number of genome and transcriptome hits for each tag is

also recorded. This nonredundant set of all tags ('tophit') could be used as a lookup table for all experimental tags annotation. Only perfect matches were considered, and no mismatches were allowed.

The total set of all annotation tags could be separated into several groups: canonical transcriptomic tags—3′-most tags from known transcripts (the 52 281 tags most expected in a DGE tag profiling experiment); noncanonical transcriptomic tags–all tags in the mouse genome that map to any known exon (both strands) but not 3′-most or derived from few ESTs only (~1.6 million tags); tags derived from ribosomal (46 tags) and mitochondrial RNA (108 tags); REPEAT tags—tags that map to the genome more than 100 times (2900 tags); and tags that map to the genome but not to any known exon (~17 million 'just genome' tags).

Microarray analysis

The microarray analysis of the exact samples as used for DGE is described in our previous paper (9). Microarray data are available through Gene Expression Omnibus under series GSE8349.

Alignment to Ensembl transcripts

To enable comparison with microarray probes, all canonical sequence tags and microarray probe sequences were put in FASTA format and then aligned to the ENSEMBL mus_musculus_core_46_36g cDNA (transcript) database using the PERL API. The probe sequences on the Agilent (AGL: WMG G4122A), Illumina (ILL: Sentrix Mouse-6 Expression BeadChip) and home-spotted long oligonucleotide arrays (LGTC: 65-mer Sigma-Compugen mouse library, version 1), were provided by the manufacturer. For the Affymetrix chips (AFF: Mouse Genome 430 v2.0 Array), the sequence from the first probe in the probeset to the last probe in the probeset was taken. For the Applied Biosystems arrays (ABI: AB1700), only the surrounding 180 nt of the probe were given and these were taken into the alignment. Microarray and Ilumina DGE tag-profiling results were compared in pairs. Only ENSEMBL transcripts that were shared between the Illumina Genome Analyzer platform and a certain microarray platform were considered.

Statistical analysis of differential gene expression

Initially, a Student's *t*-test was performed to determine significant differences in gene expression between the group of wild-type and transgenic samples. Before performing the *t*-test, we corrected for differences in the total number of counts by multiplying with a linear scaling factor that is defined as the total number of tags obtained for a certain sample divided by the average number of obtained tags in all samples. In addition, we stabilized the variance by applying a square root transformation on the linearly scaled data. This square root transformation gives a better stabilization of the variance in the region of low abundance than a logarithmic transformation. In addition, the square root transformation can handle observations with zero counts.

As a better suited alternative for the *t*-test, we applied a Bayesian model developed by Vencio *et al.* (10). We considered only canonical tags which had at least one count in each group. It fits a probability density function per gene and per group, employing the Beta-Binomial distribution, and taking into account the number of observed tags in each sample and the library size (= total number of tags) for each sample. A Bayesian error rate is calculated that reflects the *posteriori* chance that the probability density function of the group of wild types is in reality not different from the one of the transgenic mice. To estimate the number of false positives in the list of differentially expressed genes obtained by setting a cutoff on the maximum Bayesian error rate, we calculated the number of genes below the same Bayesian error rate in all unique permutations for the comparison of two groups, where the first group contained two wild-type and two transgenic mice, and the second group contained the other two wild-type and transgenic mice.

Quantitative PCR analysis

The RNA samples used for the qPCR assays were the same as for the DGE experiments. cDNA was synthesized using the Transcriptor First Strand cDNA Synthesis Kit (Roche). Quantitative RT-PCRs (qPCRs) were done on the Lightcycler480 (Roche), with SYBR-Green detection or (when amplification efficiencies with SYBR-Green were below 90%) using the universal probe library (UPL, Roche). Each cDNA was analyzed in quadruplicate, after which the average threshold cycle (Ct) was calculated per sample. The relative expression levels were calculated with the $2^{-\Delta\Delta Ct}$ method, while using the average threshold cycles for all genes analyzed to correct for differences in cDNA input.

Biological pathway analysis

The global test (11) (available from Bioconductor: www.bioconductor.org) was used to test which Gene Ontology (GO)-defined pathways were significantly deregulated in DCLK compared to wild-type mice. After summarization of the tags for each Entrez Gene entry, the global test was run on the scaled and square root transformed data. The asymptotic method was used to calculate the *P*-values. Additional filtering of pathways was done on the median of the *z*-scores for each gene in the pathway (median should be >1.5), to retrieve only those pathways for which the majority of genes contribute to the significance of the pathways.

RESULTS

Sequencing statistics

We sequenced hippocampal DGE libraries from four individual wild-type and four individual DLCK transgenic mice. We obtained $2.4 \pm 1.2 \cdot 10^6$ sequence reads per sample with $\sim 2.0 \cdot 10^5$ unique tag sequences. Figure 1 shows the distribution of the tags over the different classes that we discriminate (see 'Materials and methods' section and Supplementary Table 1). Canonical tags, i.e. those

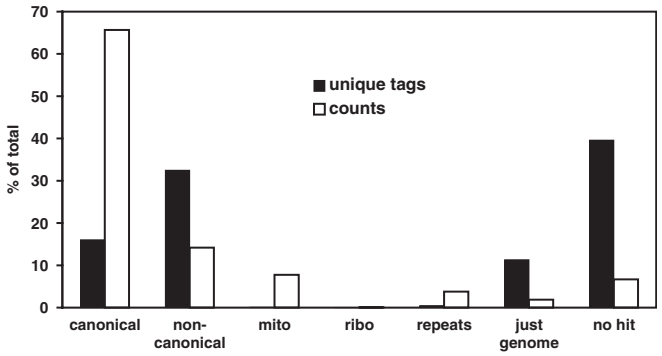

Figure 1. Categorization and abundance of tags. Distribution (in percentage of total) of unique tags (black bars) and individual reads (counts; open bars) over different categories (average from eight samples): high-confidence transcripts (canonical), low-confidence transcripts (noncanonical), mitochondrial RNA (mito), ribosomal RNA (ribo), genomic region with no evidence for transcription (just genome), repetitive genomic region (repeats) and tags with no hits in the genome.

which map to the most 3′ CATG site in high-confidence transcripts, account for 70% of the total number of reads. Since they account for only 20% of all unique tags, these appear to have an overall much higher abundance than tags corresponding to low-confidence transcripts (see also Supplementary Figure 2). Around 8% of the reads mapped to mitochondrial RNAs. The collective percentage of reads in repeat regions, regions with no evidence for transcription, and tags that could not be mapped to the genome was around 12%.

Reproducibility

To evaluate the reproducibility of DGE across different laboratories, the same RNAs were pooled and a wild-type and a transgenic pool were analyzed in triplicate at a different site (Illumina Inc., Hayward, CA) using the same protocol. The Pearson correlation coefficients for the number of counts and the normalized (scaled and square root-transformed) number of counts across technical replicates in the same laboratory were >0.99.The correlation between the normalized number of counts from the summed individual samples in our laboratory and the pool analyzed in the other laboratory were 0.98 and 0.96 for wild-type and transgenic samples, respectively (plots in Supplementary Figure 3). This is indicative of low technical variability, even across different laboratories.

Dynamic range

The dynamic range of DGE is three to four orders of magnitude. The most abundant transcript, arising from the *Ckb* gene (brain isoform of creatine kinase), comprises 0.55% of all canonical tags [$5.5{\cdot}10^3$ transcripts per million (t.p.m.)]. The lowest expressed transcripts which were still consistently detected in all samples had an abundance of 2 t.p.m., which corresponds with an average of ~0.3 copies per cell (12). The hippocampus is a rich source of unique transcripts: 28 341 different canonical tags were detected in both wild-type and transgenic groups; including noncanonical mappings increases the number even further. Within the noncanonical group alone, 45 550 tags were identified in both groups.

Alternative polyadenylation

DGE is able to discriminate between transcripts with different 3′-ends when they are separated by at least one restriction site. A remarkable 47% of the detected ENSEMBL transcripts were detected by more than one tag. This is unlikely to be caused by partial digestion of the NlaIII enzyme, in which case the more abundant and the less abundant tag for the same transcript would be found at an approximately fixed ratio. In addition, the majority of tags had been identified before in LONG-SAGE libraries. Most likely, it is due to the use of alternative polyadenylation signals in the 3′-UTR. In addition, a small fraction may be explained by alternative cleavage site selection from the same polyadenylation site (13). The observed 47% alternative polyadenylation is much higher than the 29% estimated previously based on EST sequences (14). We note that the actual incidence may yet be higher, because 3′-ends downstream of the annotated ENSEMBL transcripts are not mapped to the transcript, and alternative polyadenylation sites with no CATG sites in-between are missed. On the other hand, we have only investigated the hippocampus, while this incidence may well vary between tissues.

Antisense transcription

By considering canonical and noncanonical tags with an abundance of >2 t.p.m., and employing the strand-specific nature of the sequencing reads obtained, we find evidence for bidirectional transcription in 51% of all detectable Unigene clusters. While confirming earlier observations of bidirectional transcription in the majority of genes (15–19), our results show that the antisense transcripts are also expressed at substantial levels. Although in most cases the sense transcripts have higher abundance than the antisense transcripts, the opposite is true in 11% of the cases (Supplementary Figure 4). The on-the-bead cDNA synthesis, together with the absence of a correlation between the abundance of sense and antisense transcripts (i.e. antisense tags are generally not more prominent in highly abundant transcripts), almost excludes the possibility that the antisense tags are found due to reverse transcriptase artifacts, as suggested previously (20).

Differentially expressed genes

As a first indication for subtle, yet significant differential gene expression between the two groups of mice, the intragroup Pearson correlations (among wild-type or transgenic samples) were higher (0.96) than the correlations between samples from the different experimental groups (0.93) (*P*-value: 0.056, permutation test, Supplementary Table 2). A Fisher or similar 2×2 contingency table statistical test has previously been used to identify tags with significantly different abundance in two pooled SAGE libraries (21). In such experiments, biological variation between samples is not addressed. Our sequencing of the pooled samples clearly demonstrates the hazards of pooling. Table 1 shows tags that were highly significant in the

Table 1. Counts for blood-derived transcripts including *P*-values from Fisher test and Student's *t*-test

Gene	Name	Pool_WT	Pool_dC	Fisher	WT1	WT3	WT4	WT6	dC1	dC2	dC3	dC4	*t*-Test
Serpina3k	Serine (or cysteine) peptidase inhibitor, clade A, member 3K	87	0	1.22E-26	143	1	1	0	0	0	0	0	0.18
Gc	Group specific component	22	0	4.21E-19	41	0	1	0	0	0	0	0	0.36
Fgg	Fibrinogen, gamma polypeptide	60	0	1.69E-18	72	1	1	0	0	0	0	0	0.14
Serpina1a	Serine (or cysteine) peptidase inhibitor, clade A, member 1a	35	0	5.76E-11	71	0	0	0	0	0	0	0	0.36
Mug1	Murinoglobulin 1	20	0	2.96E-08	25	0	0	0	0	0	0	0	0.16
Itih4	Inter alpha-trypsin inhibitor, heavy chain 4	26	0	4.75E-07	51	1	1	0	0	1	0	0	0.28
Mup1	Major urinary protein 1	14	0	1.90E-06	4	0	0	0	0	0	0	0	0.36
Orm1	Orosomucoid 1	11	0	7.61E-06	22	1	0	0	0	0	0	0	0.36
Rdh7	Retinol dehydrogenase 7	17	0	1.52E-05	21	0	0	0	0	0	0	0	0.36
Exosc8	Exosome component 8	14	0	1.22E-04	28	2	0	0	0	0	1	0	0.17
Mup1	Major urinary protein 1	18	0	1.22E-04	8	0	0	0	0	0	0	0	0.36
Pnpo	Pyridoxine 5′-phosphate oxidase	12	0	9.76E-04	4	0	0	0	0	0	0	0	0.14

pooled experiment (based on Fisher's test), and not significant when analyzing the individual samples (Student's *t*-test). Clearly, these tags originate from wild-type sample 1 only. Significant expression of the *Mup1* transcript in wild-type sample 1 only was confirmed by qRT-PCR (Supplementary Table 3). Detailed study shows that all these transcripts are highly expressed in blood. Blood contamination of one of the samples, not noted during the tissue dissection procedure, thus leads to the false-positive identification of several differential transcripts in the pooled experiment. While sequencing of pooled SAGE libraries was previously the only option, it is now both advisable and affordable to sequence individual samples.

Since we sequenced multiple libraries from individual samples, we can estimate within- and between-group variation. Initially, we used a Student's *t*-test, which takes into account both types of variation, to identify genes differentially expressed between the two groups. In doing this, however, we found some important flaws of classic statistics: a *t*-test can only be applied in a meaningful sense after normalization for the total number of tags in the library and proper variance stabilization. We did this by linear scaling based on the total number of counts and subsequent square root transformation. The square root transformation (approximately) stabilizes the variance of raw counts but not of the scaled data. Hence, we cannot stabilize variance while normalizing for library size at the same time (22). This problem is particularly prominent in our experiment where one wild-type and one transgenic sample had, respectively, 3 and 10 times lower numbers of counts than the other samples. Vencio *et al.* (10) proposed a Bayesian method for the analysis of replicated SAGE data that takes into account stochastic effects for the low abundant genes as well as differences in library size. It reports the Bayesian error rate which can be interpreted as the chance that a gene is found differentially expressed under the null hypothesis. With a Bayesian error rate of 0.05, we detected 1559 up- and 1620 down-regulated canonical tags in the comparison of transgenic with wild-type mice. The distribution of the detected fold-changes can be inferred from the Volcano plot in Figure 2 and ranged between 71-fold (*2700089E24Rik*, found only once in all wild-type samples, but 19 times in transgenic samples) and 1.13-fold. Differentially expressed tags were found in the entire range of expression levels (Supplementary Figure 5). A list of the 20 most significant tags is given in Table 2. Vencio's test does not consider multiple testing. To estimate the number of false positives obtained, we calculated Bayesian error rates when permuting the samples (Supplementary Figure 6). The number of tags found differentially expressed with an error rate of 0.05 in the permutated situations was 270 ± 103. Thus, the false discovery rate in our list of 3179 differentially expressed genes is estimated to be 8.5%.

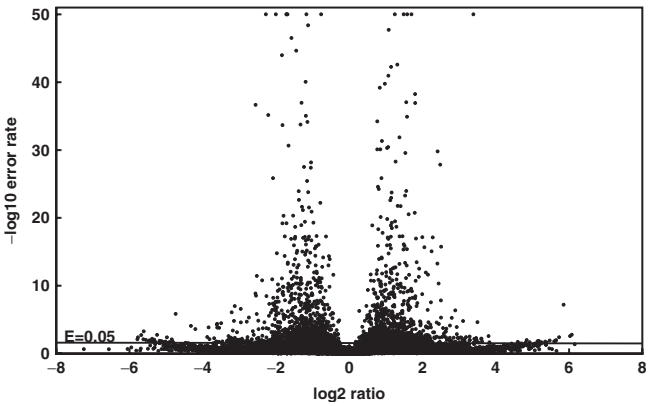

Figure 2. Volcano plot of canonical tags. For every tag, the ratio in expression levels of transgenic over wild-type mice (^{2}log scale, *x*-axis) is plotted against the Bayesian error rate (^{10}log scale, *y*-axis). The horizontal line indicates the significance threshold applied, the 3179 differentially expressed tags being above that line. The plot shows that the tags with highest average differences between trasngenic and wild-type mice (far left and right part of the plot) are not all significant (due to large intragroup variation). The most significant tags (top of the plot) generally display small differences in expression between transgenic and wild-type but are, due to relatively high expression levels, very accurately measured and therefore display low intragroup variation.

In addition to differentially expressed canonical tags, we detected differential expression of 2479 noncanonical and 15 mitochondrial tags.

Table 2. List of the 20 most significantly differentially expressed tags

Tag	Chr	Strand	Start	Unigene ID	Entrez ID	Gene symbol	Gene name	Ratio	Vencio's error rate
CATGCACTTAGAGTGTGAGAG	chr10	–	126575485	Mm.248373	216441	C78409	Expressed sequence C78409	2.48	<1E-50
CATGTCCACTACACAGAGCAT	chr6	+	55008968	Mm.250004	353172	Gars	Glycyl-tRNA synthetase	1.98	<1E-50
CATGGGGCAGGGAGCATTCAG	chr4	+	151150448	Mm.277464	57295	Icmt	Isoprenylcysteine carboxyl methyltransferase	2.75	<1E-50
CATGGTCAGAAGCAGAAGCTA	chr8	–	88150714	Mm.296520	65114	Vps35	Vacuolar protein sorting 35	3.83	<1E-50
CATGCTGCTAAGCAGAAGCAA	chr19	–	5274809	Mm.196532	319322	Sf3b2	Splicing factor 3b, subunit 2	18.36	<1E-50
CATGAAATTAATAAAAGTTAC	chr16	–	30232416	Mm.426334	106342	AU022875	Expressed sequence AU022875	0.34	<1E-50
CATGAAGGACTATGTCTAATC	chr19	–	60918807	Mm.29821	11757	Prdx3	Peroxiredoxin 3	0.31	<1E-50
CATGATGTCTAAGCTGAGAAA	chr12	–	80083926	Mm.265929	11847	Arg2	Arginasetype II	0.43	<1E-50
CATGTAGTCAGGGAGAAAACC	chr8	+	126289830	Mm.178818	66855	Tcf25	Transcription factor 25 (basic helix-loop-helix)	0.62	<1E-50
CATGGTGAACGTGCCTAAAAC	chrX	+	129932066	Mm.286408	19982	Rpl36a	Ribosomal protein L36a	0.30	<1E-50
CATGACAGACTTAAAACTGCT	chr9	+	54514230	Mm.52319	58233	Dnaja4	DnaJ (Hsp40) homolog, subfamily A, member 4	0.26	1.00E-50
CATGACAGCAGTATAAGGATC	chr10	+	83192493	Mm.271188	69784	1500009 L16Rik	RIKEN cDNA 1500009L16 gene	0.41	1.00E-50
CATGACTGACTCACACAGAGA	chr18	+	77175488	Mm.236127	76987	Hdhd2	Haloacid dehalogenase-like hydrolase domain containing 2	0.56	4.20E-49
CATGATGATAATGGACTGAGC	chr14	–	24757417	Mm.33344	211623	Plac9	Placenta specific 9	2.15	1.98E-48
CATGAAATAAATGTCAAGGGC	chr9	–	26724636	Mm.289244	66948	Acad8	Acyl-coenzyme A dehydrogenase family, member 8	0.43	3.12E-47
CATGTACAATGTGACAATAAA	chr18	+	33320540	Mm.391658	12326	Camk4	Calcium/calmodulin-dependent protein kinase IV	0.45	2.30E-45
CATGTTTCAAATAAAATTCTC	chr7	+	130555878	Mm.86322	57752	Tacc2	Transforming, acidic coiled-coil containing protein 2	0.26	1.09E-44
CATGGACCTGAAGCTCCTGGA	chr2	–	30782819	Mm.154994	30931	Tor1a	Torsin family 1, member A (torsin A)	2.08	2.57E-43
CATGCCAATTGTCCTGTGCAT	chr8	+	86886174	Mm.19111	18747	Prkaca	Protein kinase, cAMP dependent, catalytic, alpha	1.70	5.71E-43
CATGCTGTCTGGCCTTAGTGT	chr5	–	124379384	Mm.44261	19679	Pitpnm2	Phosphatidylinositol transfer protein, membrane-associated 2	1.74	1.13E-41

Displayed ratios are the ratios of the averaged normalized number of counts in transgenic over those in wild-type mice.

Biological implications

By alternative splicing the DCLK gene produces numerous proteins. Recent functional studies from a.o. knockout mice strongly indicate involvement of the DCLK gene in several molecular pathways. Some are microtubule-associated proteins (23) that may regulate microtubule-guided transport of SNARE-protein containing synaptic vesicles (24), while the DCLK-short variant has Ca^{++}/calmodulin-dependent protein kinase (CaMK) properties (8,25). In the current study, we evaluated which biological pathways were affected in the hippocampus by the expression of the DCLK-short isoform. The global test (11) was applied to the DGE data to identify the differential regulation of gene sets, as defined by the Gene Ontology consortium. Unlike commonly used overrepresentation tests or gene set enrichment analysis, this method uses the gene expression measurements of a particular set of genes, giving optimal power for small sample-size experiments and detection of gene sets where many genes display a small effect. The most significantly affected pathways are reported in Table 3. Strikingly, the CaMK pathway was the second most significant pathway. Disturbances in the expression of genes in the CaMK pathway are possibly a consequence of transcriptional feedback mechanisms. Also in line with the function of the DCLK gene, we find indications for disturbed synaptic vesicle transport along microtubules as a result of alterations in gene expression of vesicle SNARE proteins (rank 19) and microtubule plus-end binding proteins (rank 1), potentially affecting neurotransmitter release and axonal outgrowth.

The effect of sequencing depth on detection of differentially expressed genes

Before the development of deep sequencing technology, construction of a large-scale SAGE library containing up to 100 000 canonical tags would typically take 1 year and a considerable financial investment. The number of tags in such a library is 60 times smaller than the number of tags we obtained for each group of samples in a 3-day experiment. To illustrate the effect of the increased sequencing depth, we have compared our results to the results from a simulated SAGE experiment, which includes only 1/60 of the original number of DGE reads, randomly taken. The number of detected differentially expressed genes decreased 15-fold, from 3179 with the original number of reads to 200 in the simulated SAGE experiment (Bayesian error rate <0.05). The lowest abundance of a significantly detected differentially expressed transcript was 0.8 t.p.m. in our deep sequencing experiment versus 91 t.p.m. in the simulated SAGE experiment. As noted before (26), many of the genes with most significant changes in expression are low-abundant genes and would not have been identified in a typical SAGE experiment.

Table 3. Significantly deregulated pathways in DCLK transgenic mice

GOID	Term	Ontology	Genes tested	Statistic *Q*	Median *Z*	*P*-value
GO:0051010	Microtubule plus-end binding	MF	4	136	3.07	0.022
GO:0004683	Calmodulin regulated protein kinase activity	MF	8	161	2.79	0.011
GO:0005391	Sodium:potassium-exchanging ATPase activity	MF	6	416	2.71	0.013
GO:0016909	SAP kinase activity	MF	5	31	2.67	0.010
GO:0019238	Cyclohydrolase activity	MF	4	40	2.61	0.027
GO:0019209	Kinase activator activity	MF	9	70	2.31	0.014
GO:0043552	Positive regulation of phosphoinositide 3-kinase activity	BP	4	454	2.29	0.009
GO:0046339	Diacylglycerol metabolic process	BP	5	45	2.18	0.039
GO:0021782	Glial cell development	BP	7	118	2.07	0.015
GO:0048709	Oligodendrocyte differentiation	BP	5	143	2.07	0.017
GO:0014037	Schwann cell differentiation	BP	5	37	2.07	0.027
GO:0030325	Adrenal gland development	BP	5	23	2.07	0.031
GO:0001936	Regulation of endothelial cell proliferation	BP	5	27	2.07	0.035
GO:0009894	Regulation of catabolic process	BP	10	20	1.94	0.017
GO:0006970	Response to osmotic stress	BP	6	298	1.84	0.010
GO:0004602	Glutathione peroxidase activity	MF	6	44	1.80	0.012
GO:0042176	Regulation of protein catabolic process	BP	9	21	1.77	0.018
GO:0006265	DNA topological change	BP	8	38	1.75	0.027
GO:0015020	Glucuronosyltransferase activity	MF	9	34	1.66	0.016
GO:0000149	SNARE binding	MF	15	584	1.55	0.014
GO:0030295	Protein kinase activator activity	MF	7	75	1.51	0.016

The global test (11) was used to identify which pathways, as defined by the Gene Ontology consortium (BP = biological process; MF = molecular function), were significantly deregulated in DCLK mice. Only nonredundant pathways which contained at least four genes, had an asymptotic *P*-value <0.05, and for which the median of the *z*-scores of all genes in the pathway was at least 1.5, are shown.

Comparison with microarrays and qPCR

The exact same RNA samples had been analyzed previously by five different whole-genome expression microarray platforms: Applied Biosystems, Affymetrix, Agilent, Illumina and home-spotted oligonucleotide arrays (9). We compared the results from DGE and the microarray experiments after mapping all canonical tags and microarray probes to the ENSEMBL transcript database. With DGE, we detected 15 189 ENSEMBL transcripts with abundances >2 t.p.m. With most microarray platforms, a lower number of transcripts gave signal above background, except for Agilent, where there may have been considerable background signal caused by cross-hybridization (Table 4). Affymetrix had the highest percentage of transcripts in common with DGE. In general, less abundant transcripts were more difficult to detect with microarrays. The median expression of 538 transcripts detected by DGE but not by any of the microarray platforms had a median abundance of only 4 t.p.m., while the transcripts that were detected by all platforms had a median abundance of 106 t.p.m.

Figure 3 shows the correlation between absolute transcript abundance and microarray probe intensity. In line with other reports (27–29), we observed a reasonable correlation between the intensity of the microarray hybridization signal and the number of tags sequenced. The correlation was highest for Affymetrix chips (Pearson correlation: 0.63). For the Affymetrix data, intensities of the 11 perfectly matched probes were summarized into a single value. Indeed, the use of 11 different probes per transcript, in contrast to a single probe per transcript for the other platforms, should even out probe-specific hybridization characteristics. The correlation in detected transcripts was higher than previously found for SAGE or MPSS versus Affymetrix (30,31), mainly due to the higher number of tags sequenced with DGE.

Technical replicate measurements were used to compare the precision of DGE with that of microarrays. As a measure for precision we determined the distribution of the differences between independent replicate measurements of log ratios between wild-type and transgenic samples, as proposed by Irizarry *et al*. (1). Figure 4A shows the distribution of these differences for DGE and the two microarray platforms with highest and lowest precision (Agilent and home-spotted oligonucleotide arrays, respectively). The distribution of DGE is narrower (interquartile range (IQR): 0.51) than that of Agilent (IQR: 0.61) and home-spotted arrays (IQR: 0.75), indicating that DGE has a higher precision than microarrays.

With DGE we found a much wider distribution of fold-changes between the closely correlated groups of mice than for the microarray platforms, where the highest fold-change measured was 2. By DGE, we observed 1491 significantly differentially expressed tags (error rate <0.05) with an absolute fold change >2 (Figure 2). The only three genes that were significant on all microarray platforms and confirmed by qRT-PCR, *Plac9*, *D14Ertd449e* and *Gabra2*, were also significant in DGE (Bayesian error rates of $2.0{\cdot}10^{-48}$, $3.5{\cdot}10^{-47}$ and $3.9{\cdot}10^{-12}$, respectively). For the comparison between DGE and qPCR, we selected 29 significant genes from the DGE experiments (randomly chosen and covering the entire range of significance (Bayesian error rates between error rates between $1{\cdot}10^{-47}$ and 0.05) and fold-changes), and 33 genes significant genes from the microarray analyses (9). Results are given in Supplementary Table 3 and displayed in Figure 4B. The fold-changes obtained by DGE were generally also more extreme than those obtained by

quantitative PCR, as is evident from the slopes of the curve. Out of 62 genes assayed, 43 demonstrated a concordant direction of change for DGE and qPCR, but only five were significant according to both technologies.

We made a more general comparison of the lists of differentially expressed genes from the DGE and microarray experiments. Differential gene expression for DGE was established with Vencio's algorithm as described above (estimated FDR 8.5%) and for microarrays with the Empirical Bayes model LIMMA (32) (estimated FDR 10%). Complete results on correspondence between DGE tag counts and microarrays are reported in Table 5. The biggest overlap was found with the Affymetrix platform ($P = 1.2{\cdot}10^{-5}$; chi-square test): 31 transcripts were significant on both platforms with expression changes in the same direction. Also when assessing the correlation of the expression between transgenic and wild-type mice, Affymetrix chips were found to correlate better with DGE (Pearson correlation: 0.25) than the other microarray platforms (Supplementary Figure 7). The number of differentially expressed transcripts by DGE was closest to the number detected with the Agilent platform (2414 and 2710). However, the overlap between these transcripts was hardly greater than would be expected by chance and there was little correspondence in the direction of change.

Table 4. Overlap between DGE and microarrays in detectable transcripts

Platform	DGE	ABI	Affy	Agilent	Illumina	LGTC
Detectable	15 189	13 331	11 683	22 510	13 376	2017
Detected with DGE	100%	78%	89%	61%	82%	83%

For each platform we determined how many ENSEMBL transcripts could be reliably detected. For DGE, we put the threshold at 2 t.p.m., while for the microarray platforms the signal should be higher than the lowest 95% of all negative control spots. In the second row the number of transcripts detected by both—by a specific platform and by DGE—is expressed as a percentage of all transcripts detected by this specific platform.

Table 5. Overlap between DGE and microarrays in differentially expressed transcripts

	Differentially expressed			Statistics		Direction	
	MA	DGE	Overlap	Chi-square	P-value	Same	Opposite
ABI	8	2088	4	6.0	1.4E-02	4	0
AFF	153	2041	41	19.2	1.2E-05	31	10
ILL	52	2404	17	13.9	1.9E-04	14	3
AGL	2701	2414	400	1.9	1.7E-01	189	211
LGTC	33	1864	7	0.9	3.5E-01	6	1

For each subset of matching ENSEMBL transcripts between the DGE and one of the microarray platforms, we show the number of differentially expressed genes for DGE (Vencio's error rate < 0.05) and the microarray (MA; false discovery rate 10%), and the overlap. We calculate chi-square statistic and *P*-value, and indicate whether the overlapping genes are changed in the same or opposite direction.

DISCUSSION

Deep sequencing is a powerful technique for the identification of differentially expressed transcripts. The large sequencing depth clearly boosts the detection of differential expression of low-abundant transcripts that are well beyond the reach of classical SAGE. The sequencing depth of the Solexa/Illumina DGE technology compares favorably to the earlier MPSS system from Lynx Therapeutics [$7{\cdot}10^5$ sequences per run (7)] and Roche [454 sequencer, $3{\cdot}10^5$ sequences per run (33)], and is comparable to the polony multiplex analysis of gene expression (34).

Instead of sequencing SAGE tags, some recently published papers now describe the use of random shotgun RNA sequencing (RNASeq) (27–29,35–38). This overcomes the limitations of tag-based methods in the detection of transcripts alternative splicing in regions remote from the 3′-end, and enables detection of allele-specific transcription. With the continuously increasing number of reads at reduced costs, RNASeq will become affordable for standard differential gene expression analysis. However, at the present throughput it is favorable to use methods that provide a specific tag for each transcript, when the aim is to detect subtle expression differences in

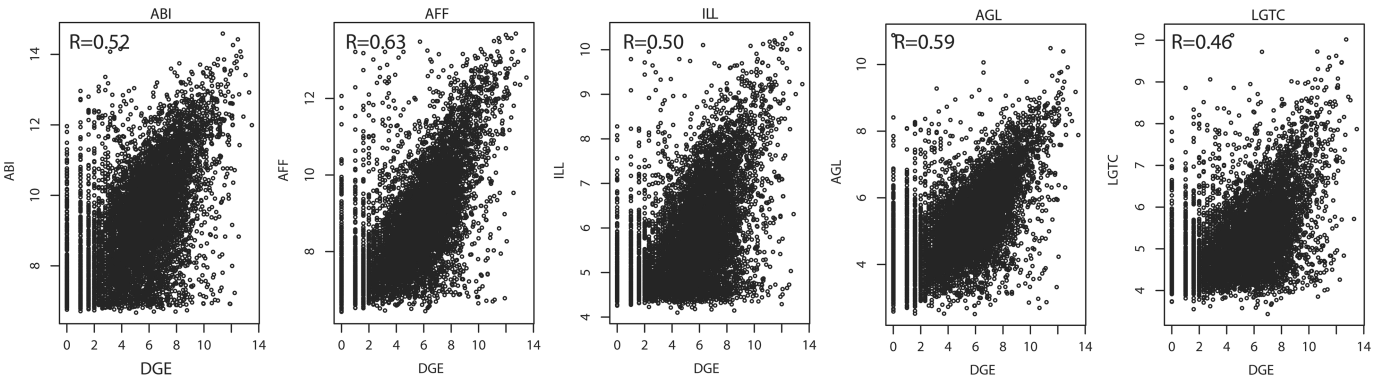

Figure 3. Correlation between absolute expression level (DGE) and microarrays signal intensity. Correlation of the tag abundance (square root transformed; *x*-axis) and intensities [normalized as described in (9)] on the five microarray platforms (*y*-axis) for matching ENSEMBL transcripts, for wild-type sample 1. Pearson correlations are indicated in the graphs. ABI: Applied Biosystems; AFF: Affymetrix; ILL: Illumina; AGL: Agilent; LGTC: home-spotted long oligonucleotide arrays.

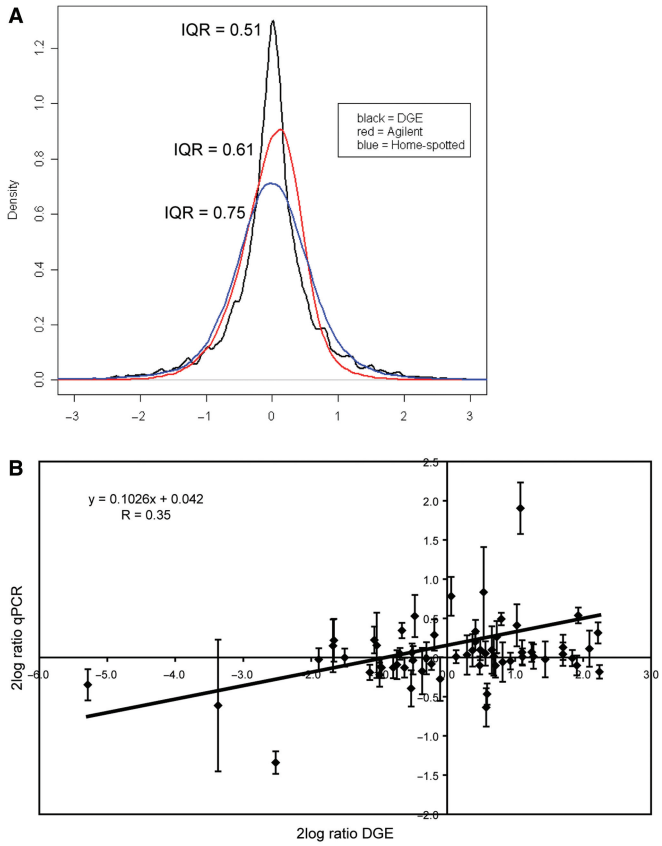

Figure 4. Assessment of precision and accuracy of DGE. **(A)** Samples from the wild-type and transgenic pools were sequenced in three different lanes. We calculated the three possible independent log ratios between transgenic and wild-type samples (technical replicates). As a measure of precision, we determined the pair-wise differences between these technical replicates. The distribution of these differences is plotted as a density function (black line). This is also done for three technical replicates of wild-type over transgenic ratios determined on Agilent (red) and home-spotted (blue) microarrays. We balanced the number of observations per platform through random selection of 21 886 features. **(B)** As a measure of accuracy, we correlated logged ratios of the expression in transgenic versus wild-type mice as obtained by DGE (*x*-axis) against those obtained by qPCR (*y*-axis). All data and primer sequences can be found in Supplementary Table 3.

larger group of samples: We demonstrate that ~2 million tags are required to reliably detect low abundant genes with DGE, whereas RNASeq requires at least 20 million tags per sample to obtain reasonable coverage of most transcripts (29,36).

We have implemented a dedicated Bayesian method to identify genes that are significantly differentially expressed between two groups of biological replicates. In most previously published reports analyzing differential gene expression in count-based data, the statistical tests applied did not account for within-group variation (28,34). We illustrated the importance of proper estimation of within- and between-group variation by showing that classical tests lead to the identification of false-positive genes due to the presence of a single blood contaminated sample. In an earlier deep sequencing report (27), in analogy to microarray data analysis, quantile normalization and a moderated *t*-statistic as implemented in the R package Limma were used to find differentially expressed genes. We believe that our method is better suited for the comparison of independent sequence libraries because one of the intrinsic properties of the test is that it puts more weight on samples which were sequenced at greater sequencing depth.

The availability of biological replicate measurements allowed us to use the global test (11), which takes into account the expression levels in individual samples, for the detection of disturbances in several biological pathways. Several of the identified pathways were highly relevant given the function of the DCLK1 protein (8,23–25). These pathways had not been identified by any of the microarrays using the same statistical test (9).

Our results demonstrate many advantages of DGE over expression microarray technology: (i) DGE gives an unbiased view of the transcriptome, not limited by predictions of expressed transcripts used to determine array content; (ii) DGE detects high levels of differential polyadenylation and antisense transcription, which are not detectable with standard microarrays; (iii) DGE data are more precise than microarray data; (iv) DGE data analysis requires a lower number of preprocessing steps (like background correction and normalization), which facilitates interlaboratory comparisons; (v) interlaboratory comparability of DGE data is high, probably due to the avoidance of hybridization processes, which are notoriously difficult to standardize (1); and (vi) DGE is more sensitive in the detection of low-abundant transcripts and of small changes in gene expression. This is probably due to the absence of background signal and saturation effects, which are major causes of ratio compression on microarrays (39). Some of these advantages have already been discussed in older literature comparing tag-based methods (SAGE, MPSS) and microarray data (2,26,30,31,40–45). The higher sequencing depth of DGE and the avoidance of laborious cloning steps add to the presumably superior precision and accuracy of DGE over these older methods, in particular when low-abundant transcripts are considered, and makes DGE a much more practical technique.

The correlation between DGE and microarrays and between DGE and qPCR assays was clear but modest. In accordance to what has been previously reported in comparisons between SAGE or MPSS and microarrays (31,40), the correlation between tag-based methods and microarrays was particularly poor for low-abundant transcripts. An important reason for the relatively low correlation across different technologies is the great similarity between our two sample sets. The resulting small differences in gene expression are difficult to pick up with microarrays, as also shown in the inter-microarray comparison of the same samples published recently (9), and also with qPCR assays. In samples with larger differences in gene expression, like the samples analyzed by the MAQC consortium (46), the correlation is likely higher. We believe that, apart from differences in sensitivity, an important reason is that the different platforms detect different transcripts. The microarray probes and qPCR assays detect, in many cases, a mix of different transcripts (1), where DGE can discriminate between transcripts with different 3′-ends; standard qPCR assays will detect cumulative presence of sense and antisense transcripts.

Indeed, when all DGE tags behave similarly, as with the *Gabra2* gene where we find 6 tags with an ~2.5-fold decrease in the DCLK mice (four from the sense and two from the antisense strand, see Supplementary Table 4), DGE results are consistent with all microarray platforms and qPCR (see Supplementary Figure 8). In many other cases, there will be no co-regulation between alternatively spliced transcripts or sense and antisense transcripts, which, especially in low-abundance situations, will result in poor correlation with microarrays and qPCR. In addition to the limited overlap in transcripts detected by both DGE and microarrays, many transcripts are detected only by one or a few of the platforms. For DGE, missing data for some transcripts are likely attributable to the absence of a CATG or a unique tag sequence (estimated frequency: 1% of murine RefSeq RNAs); for microarrays this is due to inadequate probe design. We also noted that there was a higher consistency between the fold-changes obtained by qPCR and microarrays when compared to those obtained by DGE. Apart from the explanations mentioned above, this is likely attributable to the fact that DGE measures absolute expression levels and DGE data are Poisson distributed (47), while qPCR and microarrays provide relative expression levels, which are log normal distributed.

Our finding that DGE results were more consistent with Affymetrix results than with other microarray platforms is consistent with an earlier study (31,42), in which MPSS results correlated better with Affymetrix than with other arrays. We think this lies in the use of multiple probes per gene, which should even out most probe-specific effects. Sequence biases in the different technologies have been described before. Comparative analysis of SAGE and microarrays shows that the GC content of microarray probes is important for detection sensitivity and for the correlation across technologies (26,30,41,43–45). We investigated GC bias in the DGE tags. The overall GC percentage observed in our tags is 42%. This is lower than for classical SAGE or MPSS (44) and better reflects the relatively low GC content of 3'-UTRs (48). By ranking the tags from high to low abundance, we find a higher percentage of Ts in the higher abundant tags (Supplementary Figure 9). This supports an earlier observation that highly expressed genes contain more T-rich 3′-UTRs than lowly expressed genes (48). Thus, the GC bias in DGE seems to be limited, but needs further investigation, also in the light of a recently published study where considerable overrepresentation of GC-rich sequences was observed in Solexa/Illumina-based resequencing experiments (49).

We foresee that further enhancements in sequencing depth will yet improve accuracy, in particular for low-abundant transcripts. Whole transcript sequencing (RNAseq) is another step forward. These advances, in combination with the currently achieved improvements in sensitivity, resolution and, notably, interlaboratory consistency, will tremendously boost the field of expression profiling. Multicenter biobanking and rare disease studies, where biological materials are scarce and widely spread and legal and logistical limitations may impede sharing of source materials, would gain enormously from better possibilities for robust post hoc integration of results. Also basic research and comparative genomics fields, which have been held back by extensive and lengthy standardization issues, will greatly benefit from the major improvement of data portability.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank Irina Khrebtukova and Gary Schroth (Illumina Inc., Hayward,CA) for sample preparation protocols, analysis of the pooled samples and assistance with data analysis. Michiel van Galen and Mattias den Hollander (LUMC and Hogeschool Leiden) are acknowledged for skillful assistance in bioinformatics. We would like to thank Jelle Goeman for helpful comments and Prof. Silvère van der Maarel and Prof. Rune Frants for critically reading the manuscript.

FUNDING

The Netherlands Genomics Initiative/Netherlands Organization for Scientific Research (NGI/NWO); a VENI-grant from the Dutch Organization for Scientific Research (NWO grant 2005/03808/ALW to P.A.C.'tH.).

Conflict of interest statement. None declared.

REFERENCES

1. Irizarry,R.A., Warren,D., Spencer,F., Kim,I.F., Biswal,S., Frank,B.C., Gabrielson,E., Garcia,J.G., Geoghegan,J., Germino,G. *et al*. (2005) Multiple-laboratory comparison of microarray platforms. *Nat. Methods*, **2**, 345–350.
2. Harbers,M. and Carninci,P. (2005) Tag-based approaches for transcriptome research and genome annotation. *Nat. Methods*, **2**, 495–502.
3. 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.
4. Margulies,M., Egholm,M., Altman,W.E., Attiya,S., Bader,J.S., Bemben,L.A., Berka,J., Braverman,M.S., Chen,Y.J., Chen,Z. *et al*. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.
5. Shendure,J., Porreca,G.J., Reppas,N.B., Lin,X., McCutcheon,J.P., Rosenbaum,A.M., Wang,M.D., Zhang,K., Mitra,R.D. and Church,G.M. (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, **309**, 1728–1732.
6. Bentley,D.R. (2006) Whole-genome re-sequencing. *Curr. Opin. Genet. Dev.*, **16**, 545–552.
7. Jongeneel,C.V., Iseli,C., Stevenson,B.J., Riggins,G.J., Lal,A., Mackay,A., Harris,R.A., O'Hare,M.J., Neville,A.M., Simpson,A.J. *et al*. (2003) Comprehensive sampling of gene expression in human cell lines with massively parallel signature sequencing. *Proc. Natl Acad. Sci. USA*, **100**, 4702–4705.
8. Engels,B.M., Schouten,T.G., van Dullemen,J., Gosens,I. and Vreugdenhil,E. (2004) Functional differences between two DCLK splice variants. *Brain Res. Mol. Brain Res.*, **120**, 103–114.
9. Pedotti,P., 't Hoen,P.A., Vreugdenhil,E., Schenk,G.J., Vossen,R.H., Ariyurek,Y., de Hollander,M., Kuiper,R., van Ommen,G.J., den Dunnen,J.T. *et al*. (2008) Can subtle changes in gene expression be consistently detected with different microarray platforms? *BMC. Genomics*, **9**, 124.

10. Vencio,R.Z., Brentani,H., Patrao,D.F. and Pereira,C.A. (2004) Bayesian model accounting for within-class biological variability in Serial Analysis of Gene Expression (SAGE). *BMC. Bioinformatics.*, **5**, 119.
11. Goeman,J.J., van de Geer,S.A., de Kort,F. and van Houwelingen,H.C. (2004) A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics.*, **20**, 93–99.
12. Velculescu,V.E., Madden,S.L., Zhang,L., Lash,A.E., Yu,J., Rago,C., Lal,A., Wang,C.J., Beaudry,G.A., Ciriello,K.M. *et al.* (1999) Analysis of human transcriptomes. *Nat. Genet.*, **23**, 387–388.
13. Pauws,E., van Kampen,A.H., van de Graaf,S.A., de Vijlder,J.J. and Ris-Stalpers,C. (2001) Heterogeneity in polyadenylation cleavage sites in mammalian mRNA sequences: implications for SAGE analysis. *Nucleic Acids Res.*, **29**, 1690–1694.
14. Beaudoing,E., Freier,S., Wyatt,J.R., Claverie,J.M. and Gautheret,D. (2000) Patterns of variant polyadenylation signal usage in human genes. *Genome Res.*, **10**, 1001–1010.
15. Yelin,R., Dahary,D., Sorek,R., Levanon,E.Y., Goldstein,O., Shoshan,A., Diber,A., Biton,S., Tamir,Y., Khosravi,R. *et al.* (2003) Widespread occurrence of antisense transcription in the human genome. *Nat. Biotechnol.*, **21**, 379–386.
16. Ge,X., Wu,Q., Jung,Y.C., Chen,J. and Wang,S.M. (2006) A large quantity of novel human antisense transcripts detected by LongSAGE. *Bioinformatics.*, **22**, 2475–2479.
17. Werner,A., Schmutzler,G., Carlile,M., Miles,C.G. and Peters,H. (2007) Expression profiling of antisense transcripts on DNA arrays. *Physiol. Genomics*, **28**, 294–300.
18. Katayama,S., Tomaru,Y., Kasukawa,T., Waki,K., Nakanishi,M., Nakamura,M., Nishida,H., Yap,C.C., Suzuki,M., Kawai,J. *et al.* (2005) Antisense transcription in the mammalian transcriptome. *Science*, **309**, 1564–1566.
19. Sun,M., Hurst,L.D., Carmichael,G.G. and Chen,J. (2006) Evidence for variation in abundance of antisense transcripts between multicellular animals but no relationship between antisense transcription and organismic complexity. *Genome Res.*, **16**, 922–933.
20. Perocchi,F., Xu,Z., Clauder-Munster,S. and Steinmetz,L.M. (2007) Antisense artifacts in transcriptome microarray experiments are resolved by actinomycin D. *Nucleic Acids Res.*, **35**, e128.
21. Ruijter,J.M., van Kampen,A.H. and Baas,F. (2002) Statistical evaluation of SAGE libraries: consequences for experimental design. *Physiol. Genomics*, **11**, 37–44.
22. Snedecor,G.W. and Cochran,W.G. (1989) Square root transformation for counts. In Cochran,W.G. (ed.), *Statistical Methods*. Blackwell Publishing, Hoboken, NJ, pp. 287–288.
23. Lin,P.T., Gleeson,J.G., Corbo,J.C., Flanagan,L. and Walsh,C.A. (2000) DCAMKL1 encodes a protein kinase with homology to doublecortin that regulates microtubule polymerization. *J. Neurosci.*, **20**, 9152–9161.
24. Deuel,T.A., Liu,J.S., Corbo,J.C., Yoo,S.Y., Rorke-Adams,L.B. and Walsh,C.A. (2006) Genetic interactions between doublecortin and doublecortin-like kinase in neuronal migration and axon outgrowth. *Neuron*, **49**, 41–53.
25. Shang,L., Kwon,Y.G., Nandy,S., Lawrence,D.S. and Edelman,A.M. (2003) Catalytic and regulatory domains of doublecortin kinase-1. *Biochemistry*, **42**, 2185–2194.
26. Feldker,D.E., Datson,N.A., Veenema,A.H., Proutski,V., Lathouwers,D., de Kloet,E.R. and Vreugdenhil,E. (2003) GeneChip analysis of hippocampal gene expression profiles of short- and long-attack-latency mice: technical and biological implications. *J. Neurosci. Res.*, **74**, 701–716.
27. Cloonan,N., Forrest,A.R., Kolle,G., Gardiner,B.B., Faulkner,G.J., Brown,M.K., Taylor,D.F., Steptoe,A.L., Wani,S., Bethel,G. *et al.* (2008) Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat. Methods*, **5**, 613–619.
28. Marioni,J., Mason,C., Mane,S., Stephens,M. and Gilad,Y. (2008) RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.*, **18**, 1509–1517.
29. Wilhelm,B.T., Marguerat,S., Watt,S., Schubert,F., Wood,V., Goodhead,I., Penkett,C.J., Rogers,J. and Bahler,J. (2008) Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature*, **453**, 1239–1243.
30. Lu,J., Lal,A., Merriman,B., Nelson,S. and Riggins,G. (2004) A comparison of gene expression profiles produced by SAGE, long SAGE, and oligonucleotide chips. *Genomics*, **84**, 631–636.
31. Liu,F., Jenssen,T.K., Trimarchi,J., Punzo,C., Cepko,C.L., Ohno-Machado,L., Hovig,E. and Patrick,K.W. (2007) Comparison of hybridization-based and sequencing-based gene expression technologies on biological replicates. *BMC Genomics*, **8**, 153.
32. Smyth,G.K. (2005) Limma: linear models for microarray data. Gentleman,R.C., Carey,V.J., Dudoit,S., Irizarry,R. and Huber,W. (eds), *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer, New York, pp. 397–420.
33. Nielsen,K.L., Hogh,A.L. and Emmersen,J. (2006) DeepSAGE–digital transcriptomics with high sensitivity, simple experimental protocol and multiplexing of samples. *Nucleic Acids Res.*, **34**, e133.
34. Kim,J.B., Porreca,G.J., Song,L., Greenway,S.C., Gorham,J.M., Church,G.M., Seidman,C.E. and Seidman,J.G. (2007) Polony multiplex analysis of gene expression (PMAGE) in mouse hypertrophic cardiomyopathy. *Science*, **316**, 1481–1484.
35. Torres,T.T., Metta,M., Ottenwalder,B. and Schlotterer,C. (2008) Gene expression profiling by massively parallel sequencing. *Genome Res.*, **18**, 172–177.
36. Mortazavi,A., Williams,B.A., McCue,K., Schaeffer,L. and Wold,B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**, 621–628.
37. Nagalakshmi,U., Wang,Z., Waern,K., Shou,C., Raha,D., Gerstein,M. and Snyder,M. (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, **320**, 1344–1349.
38. Sultan,M., Schulz,M.H., Richard,H., Magen,A., Klingenhoff,A., Scherf,M., Seifert,M., Borodina,T., Soldatov,A., Parkhomchuk,D. *et al.* (2008) A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*, **321**, 956–960.
39. Canales,R.D., Luo,Y., Willey,J.C., Austermiller,B., Barbacioru,C.C., Boysen,C., Hunkapiller,K., Jensen,R.V., Knight,C.R., Lee,K.Y. *et al.* (2006) Evaluation of DNA microarray results with quantitative gene expression platforms. *Nat. Biotechnol.*, **24**, 1115–1122.
40. Evans,S.J., Datson,N.A., Kabbaj,M., Thompson,R.C., Vreugdenhil,E., de Kloet,E.R., Watson,S.J. and Akil,H. (2002) Evaluation of Affymetrix Gene Chip sensitivity in rat hippocampal tissue using SAGE analysis. Serial Analysis of Gene Expression. *Eur. J. Neurosci.*, **16**, 409–413.
41. Feldker,D.E., de Kloet,E.R., Kruk,M.R. and Datson,N.A. (2003) Large-scale gene expression profiling of discrete brain regions: potential, limitations, and application in genetics of aggressive behavior. *Behav. Genet.*, **33**, 537–548.
42. Grigoriadis,A., Mackay,A., Reis-Filho,J.S., Steele,D., Iseli,C., Stevenson,B.J., Jongeneel,C.V., Valgeirsson,H., Fenwick,K., Iravani,M. *et al.* (2006) Establishment of the epithelial-specific transcriptome of normal and malignant human breast cells based on MPSS and array expression data. *Breast Cancer Res.*, **8**, R56.
43. Ishii,M., Hashimoto,S., Tsutsumi,S., Wada,Y., Matsushima,K., Kodama,T. and Aburatani,H. (2000) Direct comparison of GeneChip and SAGE on the quantitative accuracy in transcript profiling analysis. *Genomics*, **68**, 136–143.
44. Siddiqui,A.S., Delaney,A.D., Schnerch,A., Griffith,O.L., Jones,S.J. and Marra,M.A. (2006) Sequence biases in large scale gene expression profiling data. *Nucleic Acids Res.*, **34**, e83.
45. Van Ruissen,F., Ruijter,J.M., Schaaf,G.J., Asgharnegad,L., Zwijnenburg,D.A., Kool,M. and Baas,F. (2005) Evaluation of the similarity of gene expression data estimated with SAGE and Affymetrix GeneChips. *BMC. Genomics*, **6**, 91.
46. Shi,L., Reid,L.H., Jones,W.D., Shippy,R., Warrington,J.A., Baker,S.C., Collins,P.J., de Longueville,F., Kawasaki,E.S., Lee,K.Y. *et al.* (2006) The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat. Biotechnol.*, **24**, 1151–1161.
47. Thygesen,H.H. and Zwinderman,A.H. (2006) Modeling Sage data with a truncated gamma-Poisson model. *BMC. Bioinformatics.*, **7**, 157.
48. Kochetov,A.V., Sarai,A., Vorob'ev,D.G. and Kolchanov,N.A. (2002) [The context organization of functional regions in yeast genes with high-level expression]. *Mol. Biol. (Mosk)*, **36**, 1026–1034.
49. Dohm,J.C., Lottaz,C., Borodina,T. and Himmelbauer,H. (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.*, **36**, e105.