Gene expression profiling of human breast tissue samples using SAGE-Seq

Zhenhua Jeremy Wu[1,2], Clifford A. Meyer[1,2,9], Sibgat Choudhury[3,4,5,9], Michail Shipitsin[3,4,5], Reo Maruyama[3,4,5], Marina Bessarabova[6], Tatiana Nikolskaya[6], Saraswati Sukumar[7], Armin Schwartzman[1, 2], Jun S. Liu[8], Kornelia Polyak[3,4,5] and X. Shirley Liu[1, 2]

Department of [1]Biostatistics and Computational Biology and [3]Medical Oncology, Dana-Farber Cancer Institute, Boston, MA 02115, USA. [2]Harvard School of Public Health, Boston, MA 02115, USA. [4]Department of Medicine, Brigham and Women's Hospital, Boston, MA 02115, USA. [5]Department of Medicine, Harvard Medical School, Boston, MA 02115, USA. [6]Institute for General Genetics, Russian Academy of Sciences, Moscow, Russia. [7]Johns Hopkins Oncology Center, Baltimore, MD 21231, USA. [8]Department of Statistics, Harvard University, Science Center 715, 1 Oxford Street, Cambridge, MA 02138, USA. [9]These authors contributed equally to this work.

Correspondence should be made to: *xsliu@jimmy.harvard.edu* and *kornelia_polyak@dfci.harvard.edu*.

We present a powerful application of ultra high-throughput sequencing, SAGE-Seq, for the accurate quantification of normal and neoplastic mammary epithelial cell transcriptomes. We develop data analysis pipelines that allow the mapping of sense and antisense strands of mitochondrial and RefSeq genes, the normalization between libraries, and the identification of differentially expressed genes. We find that the diversity of cancer transcriptomes is significantly higher than that of normal cells. Our analysis indicates that transcript discovery plateaus at 10 million reads/sample, and suggests a minimum desired sequencing depth around 5 million reads. Comparison of SAGE-Seq and traditional SAGE on normal and cancerous breast tissues reveals higher sensitivity of SAGE-Seq to detect less abundant genes including those encoding for known breast cancer-related transcription factors and G protein-coupled receptors (GPCRs). SAGE-Seq is able to identify genes and pathways abnormally activated in breast cancer that traditional SAGE failed to call. SAGE-Seq is a powerful method for the identification of biomarkers and therapeutic targets in human disease.

[Supplemental material is available online at http://www.genome.org. Software package is downloadable at http://liulab.dfci.harvard.edu/sageExpress/. Data sets are deposit to Sage genie at NCBI, downloading link starting at: http://cgap-stage.nci.nih.gov/SAGE/SAGELibInfo?LID=2463&ORG=Hs (LID from 2463 to 2489)]

Microarrays and sequencing-based technologies have been widely used for gene expression profiling to create global pictures of cellular function (Adams et al. 1991; Schena et al. 1995; Velculescu et al. 1995). Early gene expression data analysis algorithms focused on biases and limitations introduced by each technology. For array-based technologies such as Affymetrix and NimbleGen microarrays, methods have been developed to overcome probe-specific behavior, GC content bias, dye bias, and cross hybridization (Yang and Speed 2002; Johnson et al. 2006; Song et al. 2007). While traditional sequencing-based gene expression methods such as serial analysis of gene expression (Velculescu 2000; Polyak and Riggins 2001) (SAGE) and expressed sequence tag (Adams et al. 1991) (EST) sequencing allow the identification and quantification of both known and novel genes, they were severely limited by sequencing throughput and cost (Adams et al. 1991; Velculescu et al. 1995). As next generation sequencing platforms provide increased throughput at reduced cost (Johnson et al. 2007), their applications to SAGE become a natural choice for comprehensive analysis of gene expression (SAGE-Seq) or other applications (Bloushtain-Qimron et al. 2008) and promise greater sensitivity and specificity (Morrissy et al. 2009). However, SAGE-Seq poses its unique challenges with regard to data normalization, read alignment, identification of differentially expressed genes, and comparison to traditional SAGE.

To address the above questions, we describe data analysis pipelines to process SAGE-Seq data on mammary epithelial cells isolated from normal and cancerous human breast tissue samples deep sequenced on the Illumina Solexa platform. In order to normalize the SAGE-Seq raw data across different libraries, we utilize a nonparametric empirical Bayes method to reduce the sequence sampling bias (Robbins 1956; Gale and Sampson 1995). Appropriate global diversity measurements within and across datasets are evaluated and used to cluster the libraries. We propose a mapping strategy to align SAGE-Seq tags to the genome. We utilize the mapping information to minimize sequencing errors and obtain accurate quantification of sense and anti-sense transcripts corresponding to RefSeq and mitochondrial genes. We develop a method to identify differentially expressed genes with statistical significance and show its utility on differential gene detection

between normal and neoplastic mammary epithelial cells. We also compare traditional SAGE and SAGE-Seq datasets and demonstrate the overwhelming power of SAGE-Seq to detect 20 times more differentially expressed genes with higher statistical confidence. Pathway analysis shows that the greater sequencing depth obtained by SAGE-Seq allows the identification of more than 3 times as many statistically significant Gene Ontology (GO) terms than by traditional SAGE, and improves their statistical significance score. Many of these pathways are newly identified by SAGE-Seq and are completely missed by traditional SAGE. The developed software package for SAGE-Seq data analysis pipelines can be downloaded from http://liulab.dfci.harvard.edu/sageExpress/.

RESULTS

SAGE-Seq library generation

SAGE-Seq libraries in this study were generated from 50,000 - 100,000 uncultured mammary epithelial cells isolated from breast tissue of normal healthy women and from primary invasive ductal breast carcinomas (**Table 1**). Immunomagnetic bead purification of the cells and SAGE library generation was performed essentially as previously described (Shipitsin et al. 2007) except when modifications were necessary for sequencing on the Illumina Solexa platform (see **Methods**). The raw Solexa data consists of millions of sequence tags, but only the first 21 base pair (bp) of each read is useful here. The first 4 bp are all “CATG”, which is the recognition site of the NlaIII mapping restriction enzyme used during the construction of the SAGE libraries. MmeI is used as a tagging enzyme to cut 21 bp 3’ of its recognition site present in the linker immediately 5’ to the NlaIII site. Thus, a SAGE-Seq tag is composed of a 5’ “CATG” followed by a 17 bp unique transcript-specific sequence. The cross lane correlation shows high reproducibility of the abundance measurement in SAGE-Seq libraries (**Supplemental Fig. S1**).

Pipelines for tag mapping and sequencing error minimization

To analyze the expression of individual genes, we used SeqMap (Jiang and Wong 2008) and propose a mapping pipeline (**Supplemental Fig. S3**) to align tags to RefSeq genes. This mapping pipeline allows us to map tags to mitochondrial, sense, and antisense transcripts of RefSeq genes. If a transcript has multiple CATGs, then the one closest to the poly-A tail (3’ end) is called the *best tag* (**Fig. 1A**). If a tag is mapped to multiple RefSeq locations with only one tag being a best tag, the best tag is considered as the uniquely mapped location. Otherwise, the tag is a non-unique tag, and its count is evenly divided among mapped locations. Sense tags are defined as tags that are mapped to the sense strand of exons of known transcripts. Antisense tags are defined as tags that cannot be mapped as sense tags, but are able to map against the antisense strand of known transcribed genes (He et al. 2008).

Mapping results can also be used to identify sequencing errors as shown in **Fig. 1C**. We combined the counts of the tags that are uniquely mapped to the same genes at the same locations to reduce noise and sampling bias due to sequencing error (sequencing error minimization), which reduces the number of false positives in subsequent differentially expressed gene analysis. The tag in the reference genome is used as the consensus tag for sequencing error minimization. For example, suppose there are 190,793 occurrences of the tag 'GCCGTGTCCGCCTGCTA' which maps exactly to the reference genome. If there are 3198 tags that differ from this tag by a single base pair, combined together after sequencing error minimization there is a total of 193,961, therefore the fraction of single basepair mismatches is 1.6% (=3198/193961). This is equivalent to a 0.1% sequencing error rate per base ($17*0.001*(1 - 0.001)^{16}$ =1.7%) . This is consistent with the estimation for high quality reads of Solexa (Shendure and Ji 2008). Using library N1 as an example, we demonstrate that about 76% of the tags can be uniquely mapped using our pipeline; 6% of these tags are mitochondrial tags, 46% are unique RefSeq sense best tags, 14% are unique sense non-best tags, and 10% are unique anti-sense tags (**Fig. 1B, Supplemental spreadsheet 1** for the mapping results of other libraries). All subsequent analyses are conducted on the 46% of unique sense best tags.

Overview of normal and cancer transcriptomes

Gene expression patterns of cell populations in many ways resemble species populations of different species in an ecosystem, where an individual of a species is like a transcript in our study. In typical ecosystems some species are abundant whereas the majority of species are rare (Magurran 2003). Similarly, SAGE-Seq profiling data shows that most of the genes are expressed at low levels (rare transcripts) and a few genes are expressed in large amounts (abundant transcripts) (**Fig. 2A, B**). Interestingly, although rarely expressed tags are the majority, highly expressed unique tags are still dominant when considering their population (expression level). By plotting the accumulative fraction of tag count out of total tag count as a function of unique tag count, we found that although the unique tags with one count are 63% of *S* (overall number of unique tags), they only account for 3% of *N* (the

total tag count) (**Fig. 2B**). The question arises whether these low count tags are spurious tags dominated by sequencing errors or true tags expressed at very low level. As described above, after sequencing error minimization, tags with one mismatch due to sequencing errors can be identified and corrected based on mapping information. Tags with more than 2 mismatches represent only 0.01% of all the reads (see **Methods**). Thus, these low count tags cannot be explained by sequencing errors as they are much more abundant than what could be explained by such errors. They are possibly a mixture of low abundant transcripts and nucleic acid contamination (possibly genomic DNA) as the Sage-Seq preparation protocol does not include a step for the elimination of genomic DNA from the RNA samples.

Nonparametric Empirical Bayes normalization

If each SAGE library were sequenced to the same depth (i.e. the same *N*), tag counts in different libraries would be directly comparable. However, although most of the samples were subjected to one lane of Solexa sequencing, *N* varies from 1M to 13M in different libraries (**Table 1**). Thus, in order to accurately compare gene expression patterns of different libraries, normalization of tag counts is needed. One intuitive way for normalization is to use proportion *p,* defined as *n/N*, where *n* is the count of a unique tag. Known as the maximum likelihood (ML) estimator for population frequency (Fisher 1922). This approach has the drawback that the *p* of any tag that is missing from the sequencing data (undetected tags) is assigned to be zero, while it overestimates the *p* of low and intermediate abundance tags and underestimates highly abundant tags (**Fig. 2C** black symbols). In addition, high throughput quality reads from Illumina Genome analyzer provide a good estimate for the population frequency of tags at different enrichment levels (**Fig. 2A**). This heterogeneous population of tag frequency applied as a prior indicates that the best estimation of tag enrichment should be smaller than the observed count because the population of less abundant transcripts is larger than that of abundant ones (**Fig. 2A**). Thus, a more sophisticated approach for SAGE-Seq data normalization is needed.

We applied the nonparametric empirical Bayes (NEB) method (Good 1953; Robbins 1956; Orlitsky et al. 2003) to normalize libraries with different sequencing depths (see **Methods**). There are two advantages of NEB over ML. First, whereas ML simply considers the undetected tags as zero, NEB estimates the proportion of undetected tags as $P_0 = n_1/N$, where n_1 is the frequency of unique tags with count one. To validate the accuracy of NEB estimator of P_0, we randomly sampled library N1 from 1% to 10% at a step of 1% to generate 10 pseudo libraries with different sequencing depths. We used NEB to estimate the P_0 of the undetected tags in each pseudo library, compared them with their respective proportion in the original library, and found that they were in good agreement (**Supplemental Fig. S2**). Second, NEB adjusts tag counts based on both the observed count and the nature of the frequency distribution of unique tag counts (**Fig. 2A**), which is applied as the empirical prior to reduce the sequence-sampling bias (Gale and Sampson 1995). It also renormalizes the adjusted proportion by the estimated total proportion of detected tags to $1 - P_0$ (See **Supplemental material** section "Algorithms comparison for differentially expressed genes" for comparison between NEB and ML normalization). To show the effect of sampling bias, we randomly sampled 10% (pseudo library 1) and 1% (pseudo library 2) tags from library N1 to generate two pseudo libraries with 10-fold difference in sequencing depth. When comparing the proportion of tags in the two pseudo libraries, we found that the proportions are much more comparable after NEB normalization, whereas ML overestimates *p* for low count tags and underestimates *p* for high count tags in pseudo library 2 with lower sequencing depth (**Fig. 2C**).

Diversity of normal and cancer transcriptomes

One advantage of sequence-based gene expression profiling is that it measures the absolute expression levels of many genes simultaneously. Thus, we can obtain a global view of transcript diversity within the cells and also among libraries. We used two different measures to compare transcript diversity in the libraries we analyzed. First, we used Simpson index of diversity (SID) (Simpson 1949) to characterize transcriptome diversity within each library. SID captures the variance

of the tag count distribution and is independent of sequencing depth (**Methods**). Higher values indicate higher diversity, which means the tag counts are more widely distributed among different genes. We found in our datasets that libraries from cancer samples in general have higher diversity than that from normal (**Figs. 3A, B**, *p* value = 0.07284 by Wilcox rank-sum test, the p value is in the borderline of significance due to limited number of samples). This trend could be due to the fact that tumors express many more genes either because they are composed of more diverse populations of cells or because they lost normal epigenetic controls that maintain tissue and cell type-specific gene expression patterns (**Fig. 4D**).

The second type of diversity is measured across libraries to study the gene expression diversity among libraries derived from different individuals. To ask how similar two libraries *A* and *B* are, we used the Morisita-Horn (MH) similarity index $C_{MH}(A, B)$ (Wolda 1983) (see **Methods**), and calculated their distance as $D \equiv 1 - C_{MH}(A,B)$. The Morisita-Horn index has several advantages over other distance metric measurements. First, MH index is not strongly influenced by *N* and *S*, which is essential to ascertain that the difference of the measurement is not due to differences in sequencing depths. Second, compared to distance based on Pearson cross correlation, MH index has no singularity for data with standard deviation approaching zero.

We found that cancer samples are not only more diverse within each individual (Simpson Index), but are also more diverse (MH index) across different individuals (**Fig. 3C**). This is not entirely surprising, since normal cells have a physiologic role that is essentially the same in different individuals, whereas tumors are genetically diverse and they have no functional role in the body, thus, there is no selection pressure to keep their phenotype within certain limits. The hierarchical clustering of the libraries based on the MH index showed that cancer libraries are more different from each other (larger distance across libraries) and they are also very clearly separated from the normal libraries (**Fig. 3D**).

Data quality of SAGE-Seq compared to traditional SAGE

Following read alignment and sequencing error minimization, we further evaluated the ability of SAGE-Seq to profile genome-wide gene expression and compared it to traditional SAGE. With deeper sequencing coverage, SAGE-Seq gave much higher data correlation between different libraries within the same group (**Supplemental Fig. S4**). In addition, traditional SAGE can only detect genes with proportion from 10^{-5} to 10^{-3}, whereas SAGE-Seq shows a much larger dynamic range (defined as the detected range of gene enrichment), covering about 5 orders of magnitude from 10^{-7} to 10^{-2}. For example, genes encoding for transcription factors are often expressed at intermediate or low levels and SAGE-Seq detected the expression of around 1,300 transcription factors (out of 1658 total in the human genome) in our samples. Most of the transcription factors detected by traditional SAGE are also detected by SAGE-Seq, whereas 384 transcription factors are only detected by SAGE-Seq (**Fig. 4A**). We observed similar phenomena for genes encoding for GPCRs and ABC-transporters (**Fig. 4B, C**), which are known to be differentially expressed between normal and cancer cells and are expressed at relatively low levels (Li et al. 2005; Dean 2009).

To determine how far the sequenced SAGE-Seq libraries are from saturation, we calculated the number of unique best-tag genes (uniquely mapped best tags) detected in relation to the sequencing depth of each library. The number of uniquely mapped best tags is a good indicator of the number of genes detected. Deeper sequencing is expected to detect more genes until a plateau is reached when all the genes are detected. To overcome the lack of data covering a broad range of sequencing depth, we combined all cancer (or normal) libraries and analytically calculated the number of detected unique tag genes at different sequencing depths to obtain the saturation curve (solid lines in **Fig. 4D** and **Methods** for analytic calculation). For sequencing depth below 3 million reads, the number of genes detected increases dramatically with sequencing depth (fast-growth region). The rate continues to grow at a slower rate (slow-growth region) above 5 million reads, until it plateaus around 10 million reads for both normal and cancer samples (**Fig. 4D**). This suggests that

the ideal sequencing depth for SAGE-Seq should be above 10M reads, with a minimum desired sequence depth of 5M per library. Sage-Seq data points (triangles) are all close to or in the slow-growth region, where most of the transcriptome is sequenced. Traditional SAGE data points (circles) are still in the fast-growth region, where more than half of the transcriptome is not detected due to low sequencing depth. **Figure 4D** also indicates that more genes are expressed in cancer (red triangle) than in normal (black triangle) samples, which is consistent with our previous finding that cancer samples have higher transcript diversity within each library and across libraries.

Sampling noise and biological variability

To detect differentially expressed genes between two conditions, it is important to know the sources of gene expression variability. A major source of variability in SAGE-Seq or any techniques using sequencing techniques is sampling variation and most algorithms analyzing traditional SAGE data tackled this using various approaches (Velculescu et al. 1995; Cai et al. 2004). Sequence-based transcriptome profiling can be modeled as a binomial sampling process with replacement that approximates a Poisson distribution, because using current technologies (Kharchenko et al. 2008) sequenced transcripts are a tiny minority of the total amount of cDNA loaded on the sequencers. If the same library is sequenced multiple times, Poisson model dictates that the variance in tag counts of a particular gene is equal to its abundance.

When examining the empirical variance of genes in the normal libraries versus their respective normalized count (normalized by NEB and scaled to the same sequencing depth of *N* = 1M), the observed variance indeed depends on its gene expression level (count) (**Fig. 5A,** red dashed line with slope $\alpha_{obv} \approx 2.0$ in log-log plot). However, if we denoted variance-to-mean slope for random binomial sampling as α_{rbs}, which is expected to be 1.0 according to Poisson distribution (*rbs* stands for random binomial sampling, blue dashed line in **Fig. 5A**), we observed overdispersion ($\alpha_{obv} \approx 2\alpha_{rbs} > \alpha_{rbs} = 1$), which means that the excess variability of the observed data is significantly larger than the variability expected in the random reference model (Poisson model in this case). This suggests a nonlinear

dependency of gene expression variability on the mean expression level, which indicates that in our data sets overdispersion could be the result of variation from biological individuals in addition to the sampling variation (see **Methods**). We also observed overdispersion among a subset of housekeeping genes and subsets of uniquely mapped best tags both in normal and cancer groups (data not shown).

Analysis of differentially expressed genes

One of the major applications of transcriptome profiling is the identification of genes differentially expressed between different samples. After tag alignment and sequencing error minimization, our analysis pipeline for the identification of differentially expressed genes (**Fig. 5B**) first applies nonparametric empirical Bayes method as a normalization step to reduce sampling bias and to bring different libraries to the same sequencing depth (N = 1M. Normalized sequencing depth has no influences to differential gene, which is different from sequencing depth of the library). After normalization, tags with count ≥ 3 *per* million in ≤ 2 out of all the libraries were discarded. This effectively removes a significant portion of non-informative tags, which either contain outliers or have too low counts to detect differential expression with statistical significance, and saves computational time and storage space during subsequent analysis.

The logarithmic transformation is then applied to obtain the expression index and decouple the correlation between the observed variance and the mean expression level of genes (**Fig. 5A**). Quantitatively, the observed variance in our libraries is proportional to the square of the expression level. According to the delta method in statistics, the logarithmic transformation is the right transformation to stabilize the variance (**Methods**). An alternative transformation is *arcsinh*, which is also a logarithm-like transformation but with the advantage of no singularity at zero (Huber W 2002). **Supplemental Fig. S5** shows that after applying a logarithmic transformation of base 2 on the normalized count, for intermediate and high abundance tags, the variance of the expression index is almost independent of its mean. Finally, the SAM (significance analysis of microarray) algorithm is

applied to the expression indices in the two groups of samples to identify differentially expressed genes (Tusher et al. 2001) (**Fig. 5B**). We also tried the standard t-test and found many false positives resulting from the underestimated empirical standard deviation that gives rise to extreme t values. SAM algorithm stabilizes variance to reduce false positives. Other statistical tests could be used in this step instead of using SAM, such as Robinson and Smyth's moderated t-test or Baggerly's t_w test (Keith A. Baggerly 2003; Baggerly et al. 2004; Lu et al. 2005; Robinson and Smyth 2007). Another alterative for the analysis of differentially expressed genes is to use overdispersed models such as overdispersed logistic regression or overdispersed log-linear model (Baggerly et al. 2004; Lu et al. 2005). However, whether these model-based methods can be scaled up to the deeper sequencing depths of SAGE-Seq data needs to be verified through systematic analysis with more data.

We compared the lists of differentially expressed genes between normal and cancer for both SAGE-Seq and traditional SAGE. The expression (i.e. presence) of 10,052 and 4,953 best-tag genes is detected by SAGE-Seq and traditional SAGE, respectively (**Supplemental spreadsheet 2**), with 99% (4,904) overlap. We calculated the FDR (false discovery rate) using the q-value package by Storey (Tibshirani 2003). Traditional SAGE does not sequence deep enough to allow similar *p* value or FDR cutoffs as SAGE-Seq. SAGE-Seq identifies about 4,000 differentially expressed best tag genes at 1% FDR, whereas traditional SAGE detects less than 200 at 10% FDR (**Fig. 5C**). Deeper sequencing gives SAGE-Seq increased statistical power to detect more differentially expressed genes. To compare the two lists of differentially expressed genes, we examined the rank order of genes based on their t-scores. The top 10% genes with highest t-score (495 genes for traditional SAGE and 1,005 for SAGE-Seq) are used as differentially expressed gene lists for comparison between these two methods. SAGE-Seq detected all 26 genes known to be differentially expressed between normal and breast cancer samples based on prior studies, whereas traditional SAGE only identified 4 (**Supplemental Table S1**).

Surprisingly, we only identified 54 genes when comparing the overlap between the top 10% genes identified as differentially expressed by the two methods. Further analysis confirmed that the

top differentially expressed genes detected by traditional SAGE and SAGE-Seq is quite different (**Fig. 5D,** black symbols). Many factors could contribute to this discrepancy such as differences in library preparation protocols and samples. Besides these factors, we observed that the top differentially expressed genes detected by SAGE-Seq are often expressed at moderate or low levels (~100/per million, see **Fig. S7**), which traditional SAGE either completely fails to detect or has too low (2 or 3 count) tag count to show differential expression with statistical power. These differentially expressed tags in SAGE-Seq are unlikely to be from sequencing errors based on the tag counts observed. These data imply that the increased sequencing depth of SAGE-Seq results in the detection of a different set of differentially expressed genes. To demonstrate this, we resort to simulations as the use of defined cell populations with limited numbers of cells isolated from primary breast tissues did not allow the generation of both SAGE-seq and traditional-SAGE libraries from the same sample. We took the 14 SAGE-Seq libraries and sampled them down (binomial sampling) to the sequencing depth level of traditional SAGE (~50k). The top differentially expressed genes of these simulated libraries also show little overlap with the original SAGE-Seq libraries (**Fig. 5D,** red symbols).

Pathways and networks differentially activated between normal and cancer samples

To determine what signaling pathways are identified as differentially activated by SAGE-Seq and traditional SAGE, we applied a combination of gene ontology and pathway analyses for the differentially expressed gene sets using MetaCore (Nikolsky et al. 2009). However, SAGE-Seq identifies 3,587 differentially expressed genes at 1% FDR cut off, whereas the most significantly differentially expressed gene identified by traditional SAGE has an FDR larger than 9%. Thus, we decided to take the top 10% of differentially expressed genes identified by traditional SAGE genes (493) and SAGE-Seq genes at 1% FDR (3,587), since an FDR cutoff gives too few differentially expressed genes in traditional SAGE, (**Supplemental spreadsheet 3**). MetaCore provides a p-value for each tested GO term or pathway name. Using a p value of 10^{-3} as the cutoff for significance, SAGE-Seq identifies 99 pathways to be significant whereas traditional SAGE only 32 with an overlap

of 19 (**Supplemental spreadsheet 4** and **Fig. 5C**). The following pathways and GO processes are commonly enriched between SAGE-Seq and traditional SAGE: apoptosis, cell adhesion, cytoskeleton remodeling, development, immune response, G-protein signaling, signal transduction, and transcription. These are all pathways known to be relevant to breast cancer, and in each category SAGE-Seq identified the term with higher statistical significance. The 80 additional significant GO categories identified by SAGE-seq but not by traditional SAGE are all related to cancer generally or specifically to breast cancer based on published literature, especially categories such as Apoptosis and survival, Cell cycle, Androgen receptor signaling, TGFB signaling, NFKB signaling, *BRCA1* mediated DNA damage, p53 signaling pathway, Development and cell cycle regulation by *ESR1* and *ESR2* (Estrogen receptor), G-protein signaling, and translation and transport pathways. The genes in these pathways are typically expressed at low level and this is consistent with **Figures 4B, C** showing many genes in the GPCRs and ABC-transporters families as detected by SAGE-Seq but missed by traditional SAGE (Li et al. 2005; Dean 2009). Especially worth noting that the NFKB and TGFB pathways, which appeared in multiple GO and pathway branches and are known to be differentially regulated in breast cancer (Shipitsin et al. 2007), are found to be significant in SAGE-Seq but insignificant in traditional SAGE.

DISCUSSION

In this study, we systematically evaluated SAGE-Seq for transcriptome profiling and its ability to identify differentially expressed genes between normal and neoplastic mammary epithelial cells. We are the first to apply nonparametric empirical Bayes (NEB) method to normalize different high-throughput SAGE-Seq libraries in order to correct the sampling bias due to incomplete sampling. NEB normalization can be applied to other types of techniques based on random sampling such as RNA-Seq. We designed a pipeline to align SAGE tags to sense and antisense transcripts and minimize sequencing error through tag alignment, and proposed an approach to detect differentially expressed genes by considering both sampling and biological variability. We compared SAGE-Seq and traditional SAGE to examine the effect of sequencing depth on gene coverage and differentially expressed gene detection. Comparison of SAGE-Seq data between normal and neoplastic mammary epithelial cells revealed that breast cancers have higher within- and across-library diversity than normal breast cells. SAGE-Seq identifies 20 times more differentially expressed genes at 10 fold more stringent cutoff (1% FDR) than traditional SAGE (10% FDR), and 3 times more pathways specifically activated in breast cancer, indicating its higher sensitivity and specificity.

Identifying changes in gene expression associated with physiologic processes is a central issue in biology, especially in the study of human diseases (Zhu et al. 2008). Commonly used methods include expressed sequence tag (EST) sequencing, cDNA microarray hybridization, subtractive cloning, differential display, and serial analysis of gene expression (traditional SAGE) (Adams et al. 1991; Schena et al. 1995; Velculescu et al. 1995). Compared to array-based hybridization methods, SAGE-Seq has many advantages. First, SAGE-Seq has higher sensitivity that allows the detection of less abundant genes with high confidence levels. Second, SAGE-Seq is less subject to technical artifacts such as probe effects and hybridization bias (Yang and Speed 2002). Third, SAGE-Seq does not require the *a priori* knowledge of transcripts to be analyzed, thus, it allows a global analysis of transcriptome present in the cells.

Overdispersion of expression levels of highly expressed genes was observed in microarray data, and as a result, analyses were often conducted at the log intensity level (Irizarry et al. 2003). However, most people attribute the overdispersion to probe hybridization and scanning biases inherent to the microarray platform. We quantitatively identified the relationship between biological variability and mean expression level. The SAGE-Seq data presented here not only show that the Poisson distribution used in many SAGE analysis algorithms is insufficient to capture the biological variance, but also indicate that abundant genes have higher variability among biological samples (**Fig. 5A**). This suggests that cells much better tolerate variations in the levels of highly expressed genes. These findings also imply that efforts on disease marker and drug target discovery might be more fruitful if focused on intermediate or low abundance transcripts, as these show less variation among samples within the same tissue type and differences in their expression might play a more important role in the disease process.

SAGE-Seq with deeper sequencing depth is able to detect many more significant differentially expressed transcripts than traditional SAGE with higher significance. The top differentially expressed genes identified by SAGE-Seq are not the most abundant genes but rather expressed at intermediate or low level (~100 per million). For traditional SAGE, which is sequenced at 20 times less depth, these tags will be at the borderline of being detected. Thus, traditional SAGE has no power to differentiate these genes between different conditions. At the same time, these less abundant genes often are transcription factors and receptors that play important roles of cell regulations and in tumorigenesis (**Fig. 4A, B** and **C**). Thus, even small changes in the expression of these genes might have pronounced effects on the whole cellular environment. It seems that less abundant genes also have less variability (**Fig. 5A**), which enables them to be detected as top differentially expressed genes despite that the absolute changes in their expression levels is not the largest. Thus, high throughput sequencing technologies provide the opportunity to unveil subtle changes in genes expression in more detail and with improved statistical power.

In summary, here we show that SAGE-Seq is a powerful and cost effective method for the gene expression profiling of small numbers of cells isolated from primary human tissue samples and present data analysis tools that enables researchers to decipher the physiological meaning of the immense SAGE-Seq datasets.

METHODS

SAGE-Seq library construction

We posted our detailed protocol for SAGE-Seq library generation at our web-site:

http://research4.dfci.harvard.edu/polyaklab/protocols_linkpage.php

All the SAGE and SAGE-Seq libraries in the manuscript were generated from immunomagnetic bead purified cells freshly isolated from human breast tissue samples, thus, the cell numbers are estimates based on the microscopic examination of the number of captured cells in 10ul volume and they are in the 50,000-100,000 cell range. However, based on FACS analyses and sorting of the same cell type we know their approximate abundance in the tissue sample. All the cells are directly lysed and processed for polyA RNA selection followed by library preparation. The amount of polyA RNA is either measured by Nanodrop or by SYBR green II, but if the number of cells is very limited, we just go straight to library preparation. Based on the estimate that one cell contains 10 pg total RNA, 100,000 cells have ~ 100 ng total RNA and ~1-10 ng polyA RNA (depending on cell type, tumor cells in general have higher RNA content/cell). In addition, 10% of the polyA RNA is saved for semi-quantitative RT-PCR testing of cell purity prior to proceeding with SAGE-Seq sample preparation and this also gives an estimate of the transcribable mRNA present.

Sequencing error minimization

At 0.1% error rate per base, the population of tags with no error is $(1\text{-}0.1\%)^{17} = 0.9831353$ and the population of tags with one error is $17 * 0.001 * (1 - 0.001)^{16} = 0.01673003$. Thus, the population of tags with error in at least 2 bases is: 1 – 0.0831353 – 0.01673003 = 0.0001346471.

Simpson index of diversity

Simpson's measure of diversity(Simpson 1949) (SID) is defined as: $\mathrm{SID} = -\ln D = -\ln \sum \frac{n_i(n_i-1)}{N(N-1)}$,

where n_i is the count of the ith tag and N is the total number of tag count. $\mathrm{SID} = 1$ means one tag

dominates all the tag count of the system, which means that there is no diversity (highest dominance). The larger value of SID is, the higher diversity is (less dominance). SID is not strongly influenced by the sequencing depth, which is confirmed by simulation.

Morisita-Horn similarity index

Morisita-Horn similarity index, $C_{MH}(A,B)$, between two libraries A and B is defined as:

$C_{MH}(A,B) \equiv \frac{2\sum p_i(A)p_i(B)}{\sum\left(p_i^2(A)+p_i^2(B)\right)}$, where $p_i(A)$ and $p_i(B)$ are the proportion of gene (or tag) i for library A and B respectively. MH similarity index is independent of the sequencing depth N.

Sampling variance and biological variance

The high throughput sequencing technology is modeled as a binomial sampling procedure with replacement. Mixing with the biological variability from different samples, another layer of variability is introduced due to sampling. Define p_n, the proportion of tag with count n as, $p_n = \frac{n}{N}$, where N is the total number of tag count. From our hierarchical model, the mean proportion $\bar{p}$ of a gene is an unbiased estimator of the true abundance of genes. However, $s^2(x_i)$, the observed variance of scaled count x_i of gene i is an addition of the true biological variance from individuals and the sampling variance (Supplemental material for analytical proof), $s^2(x) = Np_t(1-p_t) + N(N-1)\sigma_t^2$

Normalization using non-parametric empirical Bayes correction

We implement the empirical Bayes using simple Good-Turing estimator (SGT) (William A. Gale 1995). Assume the total number of all unique tags in the mRNA is *s* and p_i is the true proportion of tag *i*, which is what we want to estimate from data. Empirical Bayes estimation of an observed tag count *r* is:

$$r^* = (r+1)\, n_{r+1} / n_r ,$$

where n_r is the number of tags with count r. Thus, the expected total chance of all tags that are each represented r times (r >=1) is: (r+1) n_r+1 / N, where N is the sequencing depth ($N = n_1 + 2n_2 + 3n_3$ +…). Therefore, the expected total chance of all tags represented in the sample is: $(2n_2+3n_3+\ldots)/N=1-n_1/N$ In SGT, the proportion of undetected tags, P_0 is estimated as

$$P_0=n_1/N,$$

where n_1 represents the number (frequency) of unique tags with count one. The corrected total tag count after SGT, N^*, is $N^*=\Sigma n_r r^*$. The empirical Bayes estimator for proportion of a gene with count r, p_r^*, is renormalized by N^* as: $p_r^*=(1-P_0)n^*/N^*$.

Variance of Good-Turing estimator for unseen tags

The variance of P_0 can be calculated in the following. $\mathrm{Var}(P_0)=\mathrm{Var}(n_1)/N^2$. $n_1=\Sigma_i Np_i(1-p_i)^{N-1}$ under the assumption of binomial sampling approximation. Introducing a new random number x_i: x_i=1 if the ith tag is sequenced with only one tag at sequencing depth N and $x_i=0$ otherwise . Then:

$$E(n_1^2)=E\left[\left(\sum_i x_i\right)^2\right]=E\left(\sum_{i,j} x_i x_j\right)=\sum_i E(x_i^2)+\sum_{i\neq j} E(x_i x_j)=\sum_i Np_i(1-p_i)^{N-1}+\sum_{\mu\neq\nu} Np_\mu(1-p_\mu)^{N-1}Np_\nu(1-p_\nu)^{N-1},$$

and

$$E(n_1)^2=\left[\sum_i E(x_i)\right]^2=\left[\sum_i Np_i(1-p_i)^{N-1}\right]^2=\sum_{i,j} Np_i(1-p_i)^{N-1}Np_j(1-p_j)^{N-1}$$
$$=\sum_i N^2p_i^2(1-p_i)^{2(N-1)}+\sum_{\mu\neq\nu} Np_\mu(1-p_\mu)^{N-1}Np_\nu(1-p_\nu)^{N-1}$$

Thus, $\mathrm{Var}(P_0)=\mathrm{Var}(n_1)/N^2=\left[\sum_i Np_i(1-p_i)^{N-1}-\sum_i N^2p_i^2(1-p_i)^{2(N-1)}\right]/N^2=\frac{P_0}{N}-\sum_i p_i^2(1-p_i)^{2(N-1)}$, which gives:

$$\mathrm{Var}(P_0)<\frac{P_0}{N}.$$

This indicates that the Good-Turing estimator for unseen tags is a stable estimator, which is shown in **Supplemental Fig. S2.**

Saturation curve calculation

Assuming the total number of all unique tags in the mRNA is s and p_i is the true proportion of tag i, which is what we want to estimate from data. Thus, the mean of unseen tags according to binomial sampling is:

$$n_0(N) = \sum_i C_N^0 p_i^0 (1-p_i)^N = \sum_i (1-p_i)^N ,$$

where, the summation is over all the possible unique best tags and *N* is the sequencing depth. Thus the number of detected unique best tag genes is the total number of unique best tag genes minus $n_0(N)$ as shown in **Fig. 4D**.

Variance stabilization

The proof a logarithm transformation to stabilize the variance using delta method in statistics is provided as follows. Suppose a random variable x follows a distribution with mean μ and variance σ^2. Consider a transformation $g(x)$. The Taylor expansion of $g(x)$ around μ up to first order is $g(x) \approx g(\mu) + (x-\mu)g'(\mu)$. Thus the transformed variable $g(x)$ has approximate mean $g(\mu)$ and approximate variance $\mathrm{Var}[g(x)] \approx \sigma^2 [g'(\mu)]^2$. In our data, μ and σ^2 satisfies the observed dependency $\sigma^2 \sim \mu^2$, yielding $\mathrm{Var}[g(x)] \approx \mu^2 [g'(\mu)]^2$. Assuming the transformation g stabilizes the variance, $\mathrm{Var}[g(x)]$ is a constant independent of μ, and thus $g'(\mu) = c/\mu$, where c is a constant. Integrating with respect to μ gives that the form of the stabilization transformation g should be: $g(x) = \log x$.

Databases used

The transcription factor gene list for human is obtained from NCBI. Go to Entrez gene and search for human transcription factor. After filtering out non-human genes, 1,658 human genes encoding for transcription factors are in this list. The 7 normal and 7 cancer raw SAGE-Seq data will be deposited in GEO.

Network and pathway analysis using METACORE

Network and pathway analysis using METACORE was performed essentially as previously described (Nikolsky et al. 2008). Specific details are in Supplemental Methods.

Acknowledgements

We would like to thank Dr. Andrea Richardson (Brigham and Women's Hospital, Boston, MA) for her help with the acquisition of breast tumor samples; Drs. Haiyan Huang, Li Cai, Molin Wang, and David Harrington for valuable discussions; Love Nickerson for English proofreading. This work was supported by the Friends of Dana-Farber Women's Cancer Program (X.S.L.), NIH R01 1HG004069 (X.S.L.), NCI P50 CA89393 CA and R01 CA116235-04S1 (K.P.), the AVON Foundation (K.P.), Terri Brodeur Breast Cancer Research Foundation (S.C.), and the Susan G. Komen Foundation PDF0707996 (M.S., R.M.).

Author Contributions Z.W, did the computational analysis, Z.W., X.S.L., and K.P. designed the study and wrote the manuscript. Z.W. X.S.L., C.M., A.S. and J.L developed the analytical methodology. S.C., M.S., R.M., M.B., and T.N. carried out experiments and analyzed data. S.S. provided normal tissue samples for the study.

Competing Interests Statement The authors have no competing financial interests. Correspondence should be addressed to Kornelia Polyak (kornelia_polyak@dfci.harvard.edu) and X. Shirley Liu (xsliu@jimmy.harvard.edu).

FIGURE LEGENDS

Figure 1. SAGE-Seq tag alignment and sequencing error minimization (**A**) Best tag is defined as the tag next to the 3'-most NlaIII site (CATG) to the polyA tail. (**B**) Tag alignment statistics of sample N1 according the tag alignment pipeline in **Supplemental Fig. S3**. Detailed mapping of other data sets is in **Supplemental spreadsheet 1**. (**C**) Presumed sequencing errors revealed during tag mapping. All of the listed tags are best tags uniquely mapped to the same RefSeq gene "NM_001010" (Not every tag uniquely mapped to this gene at the same location are listed). X-axis is the tag sequences listed in descending order of tag count. The one base difference in sequence most likely due to sequencing error is marked in red. Sequencing error minimization step for this particular example is done in the following way: sum up the count of all these tags and assign it to tag 'CATGGCCGTGTCCGCCTGCTA' and remove all the other tags.

Figure 2. Frequency plot of unique tag count and nonparametric empirical Bayes method. (**A**) Frequency of unique tag counts in libraries N1 (black) and N5 (red). X-axis is the observed tag count and y-axis is the frequency that shows the number of unique tags with a specific count. (**B**) Pie chart depicting the distribution of unique tags in library N1: 62.6% of unique tags has tag count 1, 12.1% count 2, 5.5% count 3, and 19.8% counts larger than 3. The outer plot shows the accumulative fraction of unique tag counts. Although 62.6% unique tags have count 1, they only account for 3% of total tag counts. (**C**) Scatter plot of tag proportion. X-axis is the proportion of tags in pseudo library 1 obtained by randomly sampling 10% of library N1. Y-axis is the average proportion of pseudo library 2 obtained by randomly sampling 1% of library N1. The data points are obtained in the following way. For example, find all the tags in pseudo library 1 with proportion 1e-6, then calculate the mean proportion of these tags in pseudo library 2, which gives for example 1e-5. This gives a data point at (1e-6, 1e-5). The dashed line is y=x. Black symbols indicate the proportion using the maximum likelihood estimator, where overestimation in the low and intermediately expressed tags (< 100/million) and underestimation in the highly expressed tags (> 100/million) are observed. Red

symbols mark the proportion calculated using nonparametric empirical Bayes method with improved more comparable corrected proportions between two libraries with different sequencing depth in both low and highly abundant tags.

Figure 3. Diversity of normal and cancer transcriptomes. (**A**) Simpson index of diversity to measure within-library gene expression diversity. Libraries in the cancer group show higher within-library diversity compared to the normal group. (**B**) Box plot depicting Simpson index of diversity of normal and cancer samples. *p* value = 0.07284 (Wilcoxon rank-sum test). (**C**) Distance defined as "1 - *Morisita-Horn similarity index"* is used to measure gene expression diversity across libraries. Libraries in normal group are more similar to one another, whereas cancer libraries are more diverse. (**D**) Hierarchical clustering using "distance" defined in (**C**) separates normal and cancer libraries.

Figure 4. SAGE-Seq tag mapping and sequencing depths saturation curve. (**A-C**) Differential coverage of expression profiles in three selected gene families: transcription factors (**A**), GPCRs (**B**), and ABC transporters (**c**). Y-axis lists the genes and X-axis is the mean gene expression index (logarithm of the normalized tag count). Red and blue colors mark traditional SAGE and SAGE-Seq, respectively. SAGE-Seq detects many more genes in these gene families than traditional SAGE does. (**d**) Number of unique best-tag genes (y-axis) in relation to sequencing depth (x-axis). The number of best-tag genes is the number of unique genes mapped by best tags, counted as one if multiple tags are mapped to the best tag of the same gene. Black and red colors indicate normal and cancer groups, respectively. Symbols "◯"and "▲"mark traditional SAGE and SAGE-Seq, respectively.

Solid curves (saturation curves) are from simulation by sampling the combination of all libraries in normal (or cancer) group, which depict the trend with increasing sequencing depth. Traditional SAGE identifies much fewer best-tag genes than the SAGE-Seq. SAGE-Seq shows that cancer samples (red triangles) have larger number of unique best-tag genes than normal samples (black triangles). This difference is not detected by traditional SAGE (red circles vs. black circles).

Figure 5. Differentially expressed genes and their variance. (**A**) Mean-to-variance plot for the 7 normal libraries after removing the noise and normalization. Red dashed line is the best linear fit in log-log plot. The slope gives the exponent $\alpha_{obv} \approx 1.9$. Blue dashed line is the mean-to-variance line introduced by sampling. (**B**) Pipeline for the identification of differentially expressed genes: (1) Sequencing error minimization: after tag alignment, tags that are mapped to the same genes at the same locations are combined together. (2) NEB is used to normalize different libraries with different sequencing depth; (3) filtering to remove tags with counts ≥ 3 per million in less than 2 libraries followed by log2 transformation; (4) SAM is used for the detection of differentially expressed genes. (**C**) Detected differentially expressed genes (top panel) and activated pathways (bottom panel) in SAGE-Seq and traditional SAGE. SAGE-Seq identifies ~4000 differential genes at 1% FDR while traditional SAGE identifies less than 200 at a much looser cut off (10% FDR). At p value = 0.001, SAGE-Seq identifies 99 pathways significantly activated in breast cancer while traditional SAGE only shows 32. The 80 pathways only identified by SAGE-Seq and missed by traditional SAGE are all breast cancer related pathways. (**D**) The overlap ratio (defined as the number of overlapping genes divided by the gene number in traditional SAGE in the top x percent differentially expressed genes, where x changes between 0 and 1. The black symbols depict actual data (SAGE-Seq vs. traditional SAGE). It indicates that there is little overlap in the top differentially expressed genes list between SAGE-Seq and traditional SAGE. The red symbols indicate simulation (SAGE-Seq vs. sampled down SAGE-Seq). Sampled down SAGE-Seq means to binomially sample 50k tags from each SAGE-Seq library. 50k is a typical sequencing depth for traditional SAGE. Simulation confirms the same conclusion as that drawn from the actual data: SAGE-Seq gives a different top differentially expressed gene list compared to traditional SAGE. Deeper sequencing reveals that traditional SAGE identifies different sets of top differentially expressed genes than that of SAGE-Seq, confirming our conclusion that traditional SAGE lacks sufficient sequencing depth.

Table 1. SAGE-Seq libraries of normal and cancer groups. N1-N7 denotes mammary epithelial cells isolated from reduction mammoplasty specimens of normal healthy women, whereas C1-C7 indicates primary invasive breast carcinomas. Total and unique tag counts are listed for each library.

Normal							
N (Total tags)	N1	N2	N3	N4	N5	N6	N7
	9,618,916	13,522,703	2,983,207	1,800,069	1,824,933	1,045,874	11,007,864
S (# unique tag)	475,975	533,972	342,066	222,314	173,973	129,323	333,462
Cancer							
N (Total tag)	C1	C2	C3	C4	C5	C6	C7
	4,334,958	4,710,675	4,116,502	3,550,342	3,848,898	4,263,862	3,373,871
S (# unique tag)	477,158	657,887	383,450	374,584	434,886	466,774	372,801

References

Adams, M.D., Kelley, J.M., Gocayne, J.D., Dubnick, M., Polymeropoulos, M.H., Xiao, H., Merril, C.R., Wu, A., Olde, B., Moreno, R.F. et al. 1991. Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* **252**(5013): 1651-1656.

Baggerly, K.A., Deng, L., Morris, J.S., and Aldaz, C.M. 2004. Overdispersed logistic regression for SAGE: modelling multiple groups and covariates. *BMC Bioinformatics* **5**: 144.

Bloushtain-Qimron, N., Yao, J., Snyder, E.L., Shipitsin, M., Campbell, L.L., Mani, S.A., Hu, M., Chen, H., Ustyansky, V., Antosiewicz, J.E. et al. 2008. Cell type-specific DNA methylation patterns in the human breast. *Proc Natl Acad Sci U S A*.

Cai, L., Huang, H., Blackshaw, S., Liu, J.S., Cepko, C., and Wong, W.H. 2004. Clustering analysis of SAGE data using a Poisson approach. *Genome Biol* **5**(7): R51.

Dean, M. 2009. ABC transporters, drug resistance, and cancer stem cells. *J Mammary Gland Biol Neoplasia* **14**(1): 3-9.

Fisher, R.A. 1922. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London* **222**.

Gale, W.A. and Sampson, G. 1995. Good-Turing frequency estimation without tears. *Jthe Journal of Quantitative Linguistics* **2**: 217-237.

Good, I.J. 1953. The Population Frequencies of Species and the Estimation of Population Parameters. *Biometrika* **40**(34): 238.

He, Y., Vogelstein, B., Velculescu, V.E., Papadopoulos, N., and Kinzler, K.W. 2008. The Antisense Transcriptomes of Human Cells. *Science* **322**(5909): 3.

Huber W, V.H.A., Sultmann H, Poustka A, Vingron M. 2002. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* **18**(Supplement 1): 9.

Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U., and Speed, T.P. 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**(2): 249-264.

Jiang, H. and Wong, W.H. 2008. SeqMap: mapping massive amount of oligonucleotides to the genome. *Bioinformatics* **24**(20): 2395-2396.

Johnson, D.S., Mortazavi, A., Myers, R.M., and Wold, B. 2007. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316**(5830): 1497-1502.

Johnson, W.E., Li, W., Meyer, C.A., Gottardo, R., Carroll, J.S., Brown, M., and Liu, X.S. 2006. Model-based analysis of tiling-arrays for ChIP-chip. *Proc Natl Acad Sci U S A* **103**(33): 12457-12462.

Keith A. Baggerly, L.D., Jeffrey S. Morris, C. Marcelo Aldaz. 2003. Differential expression in SAGE: accounting for normal between-library variation. *Bioinformatics* **19**(12): 7.

Kharchenko, P.V., Tolstorukov, M.Y., and Park, P.J. 2008. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat Biotechnol* **26**(12): 1351-1359.

Li, S., Huang, S., and Peng, S.B. 2005. Overexpression of G protein-coupled receptors in cancer cells: involvement in tumor progression. *Int J Oncol* **27**(5): 1329-1339.

Lu, J., Tomfohr, J.K., and Kepler, T.B. 2005. Identifying differential expression in multiple SAGE libraries: an overdispersed log-linear model approach. *BMC Bioinformatics* **6**: 165.

Magurran, A.E. 2003. The commonness, and rarity, of species. In *Measureing biological diveristy*, p. 18. Wiley-Blackwell.

Morrissy, A.S., Morin, R.D., Delaney, A., Zeng, T., McDonald, H., Jones, S., Zhao, Y., Hirst, M., and Marra, M.A. 2009. Next-generation tag sequencing for cancer gene expression profiling. *Genome Res* **19**(10): 1825-1835.

Nikolsky, Y., Kirillov, E., Zuev, R., Rakhmatulin, E., and Nikolskaya, T. 2009. Functional analysis of OMICs data and small molecule compounds in an integrated "knowledge-based" platform. *Methods in Molecular Biology* **563**: 20.

Nikolsky, Y., Sviridov, E., Yao, J., Dosymbekov, D., Ustyansky, V., Kaznacheev, V., Dezso, Z., Mulvey, L., Macconaill, L.E., Winckler, W. et al. 2008. Genome-wide functional synergy between amplified and mutated genes in human breast cancer. *Cancer Res* **68**(22): 9532-9540.

Orlitsky, A., Santhanam, N.P., and Zhang, J. 2003. Always Good Turing: asymptotically optimal probability estimation. *Science* **302**(5644): 427-431.

Polyak, K. and Riggins, G.J. 2001. Gene discovery using the serial analysis of gene expression technique: implications for cancer research. *J Clin Oncol* **19**(11): 2948-2958.

Robbins, H. 1956. An empirical Bayes approach to statistics. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* **1**: 157-163.

Robinson, M.D. and Smyth, G.K. 2007. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics* **23**(21): 2881.

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**(5235): 467-470.

Shendure, J. and Ji, H. 2008. Next-generation DNA sequencing. *Nat Biotechnol* **26**(10): 1135-1145.

Shipitsin, M., Campbell, L.L., Argani, P., Weremowicz, S., Bloushtain-Qimron, N., Yao, J., Nikolskaya, T., Serebryiskaya, T., Beroukhim, R., Hu, M. et al. 2007. Molecular definition of breast tumor heterogeneity. *Cancer Cell* **11**(3): 259-273.

Simpson, E.H. 1949. Measurement of Diversity. *Nature* **163**: 688.

Song, J.S., Johnson, W.E., Zhu, X., Zhang, X., Li, W., Manrai, A.K., Liu, J.S., Chen, R., and Liu, X.S. 2007. Model-based analysis of two-color arrays (MA2C). *Genome Biol* **8**(8): R178.

Tibshirani, J.S.a.R. 2003. Statistical significance for genome-wide studies. *Proceedings of the National Academy of Sciences* **100**: 6.

Tusher, V.G., Tibshirani, R., and Chu, G. 2001. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A* **98**(9): 5116-5121.

Velculescu, V.E., Zhang, L., Vogelstein, B., and Kinzler, K.W. 1995. Serial analysis of gene expression. *Science* **270**(5235): 484-487.

Velculescu, V.E.V., B. and Kinzler, K. W. 2000. Analysing uncharted transcriptomes with SAGE. *Trends Genet* **16**(10): 423-425.

William A. Gale, G.S. 1995. Good-Turing frequency estimation without tears. *Jthe Journal of Quantitative Linguistics* **2**.

Wolda, H. 1983. Diversity, Diversity Indices and Tropical Cockroaches. *Oecologia* **58**(3): 290-298.

Yang, Y.H. and Speed, T. 2002. Design issues for cDNA microarray experiments. *Nat Rev Genet* **3**(8): 579-588.

Zhu, J., Zhang, B., and Schadt, E.E. 2008. A systems biology approach to drug discovery. *Adv Genet* **60**: 603-635.

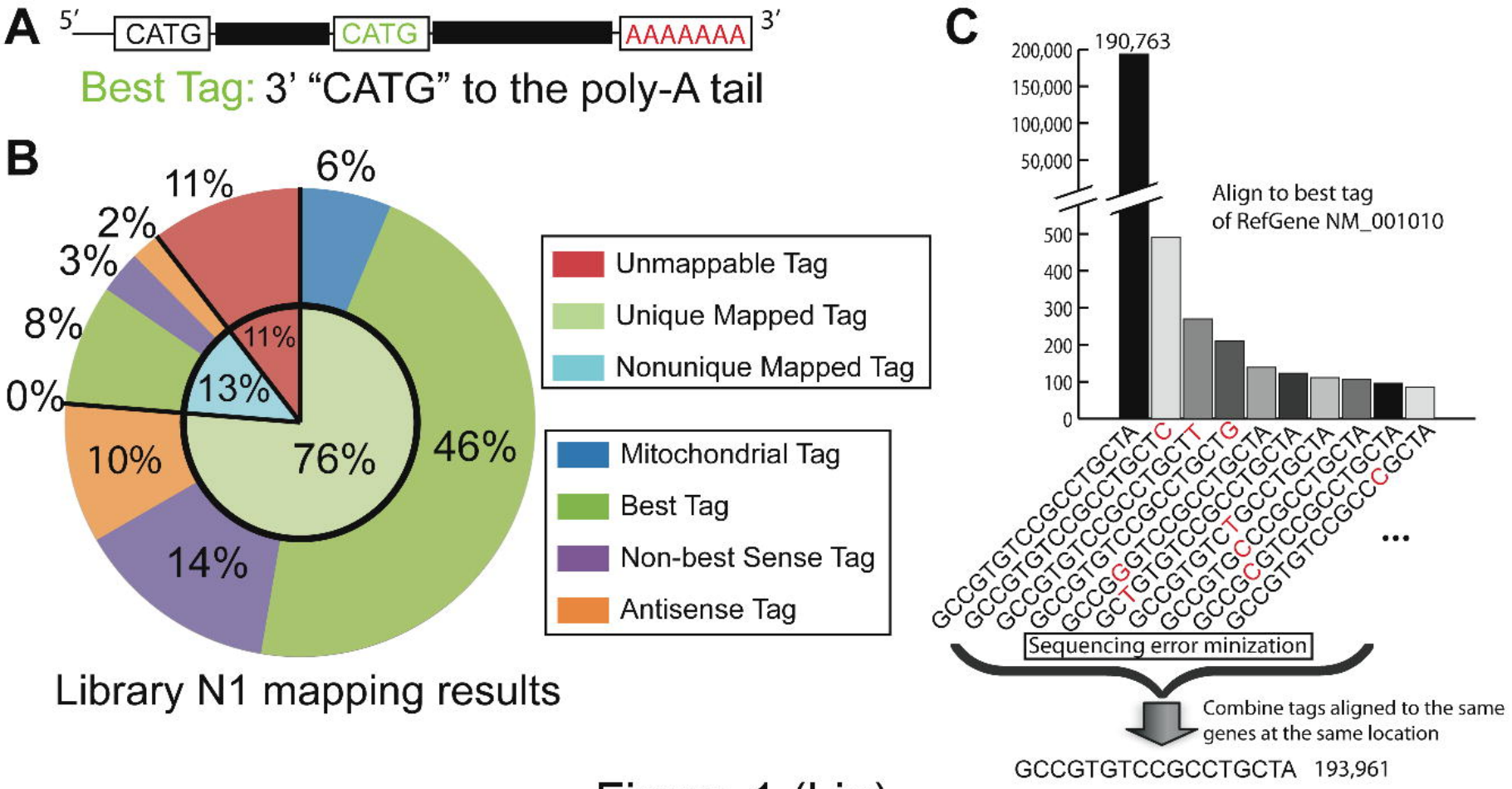
A
5′
CATG
CATG
AAAAAAA
3′
Best Tag: 3′ “CATG” to the poly-A tail
B
6%
11%
2%
3%
8%
0%
10%
14%
46%
76%
11%
13%
Unmappable Tag
Unique Mapped Tag
Nonunique Mapped Tag
Mitochondrial Tag
Best Tag
Non-best Sense Tag
Antisense Tag
Library N1 mapping results
C
200,000
150,000
100,000
50,000
500
400
300
200
100
0
190,763
Align to best tag
of RefGene NM_001010
GCCGTGTCCGCCTGCTA
GCCGTGTCCGCCTGCTC
GCCGTGTCCGCCTGCTT
GCCGTGTCCGCCTGCTG
GCCGGGTCCGCCTGCTA
GCTGTGTCTGCCTGCTA
GCCGTGCCCGCCTGCTA
GCCGCGTCCGCCTGCTA
GCCGTGTCCGCCCGCTA
...
Sequencing error minization
Combine tags aligned to the same
genes at the same location
GCCGTGTCCGCCTGCTA 193,961

Figure 1 (Liu)

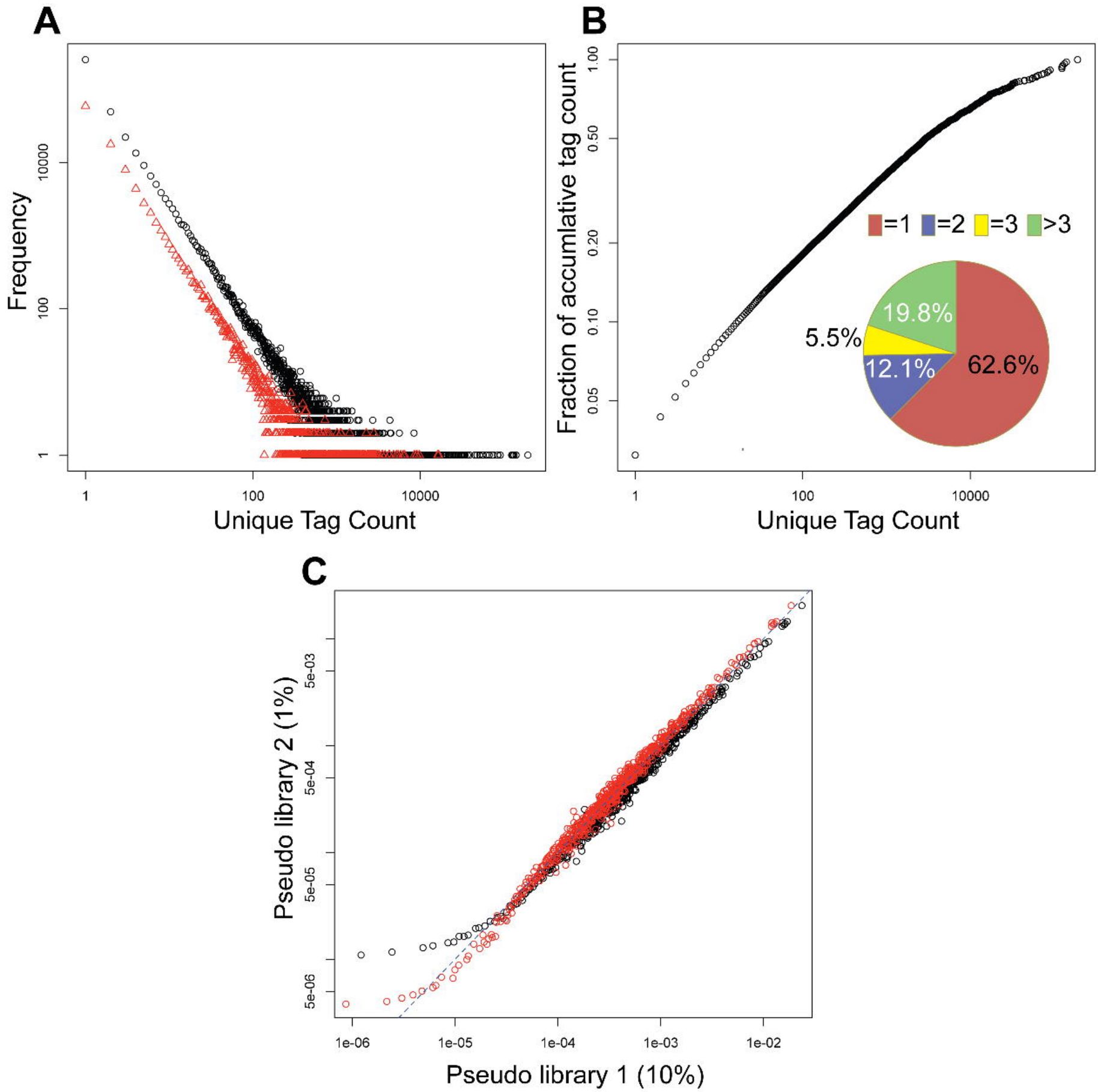
A
Frequency
10000
100
1
1
100
10000
Unique Tag Count
B
Fraction of accumlative tag count
1.00
0.50
0.20
0.10
0.05
=1
=2
=3
>3
19.8%
5.5%
12.1%
62.6%
1
100
10000
Unique Tag Count
C
Pseudo library 2 (1%)
5e-03
5e-04
5e-05
5e-06
1e-06
1e-05
1e-04
1e-03
1e-02
Pseudo library 1 (10%)

Figure 2 (Liu)

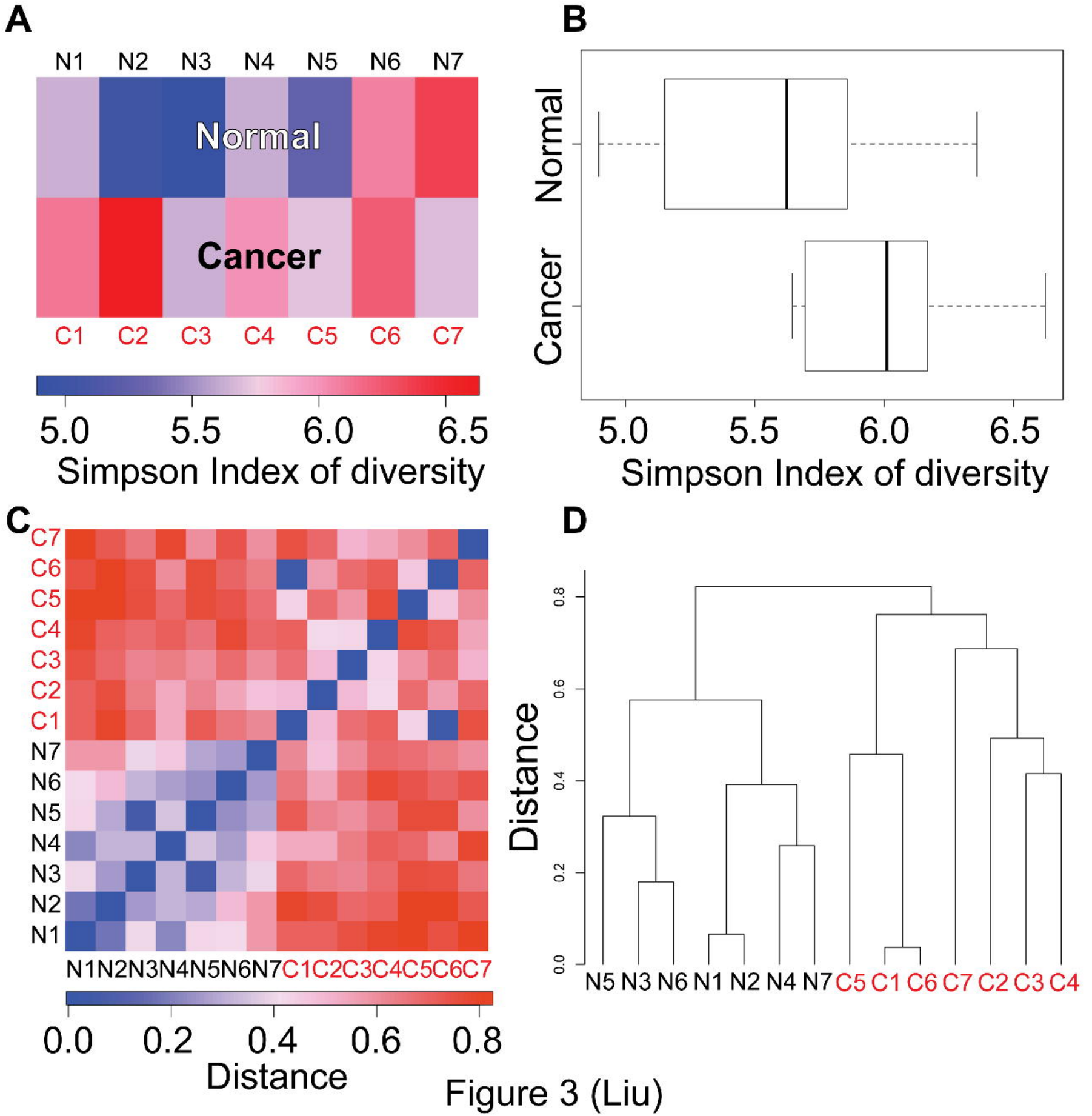
A
N1 N2 N3 N4 N5 N6 N7
Normal
Cancer
C1 C2 C3 C4 C5 C6 C7
5.0 5.5 6.0 6.5
Simpson Index of diversity
B
Normal
Cancer
5.0 5.5 6.0 6.5
Simpson Index of diversity
C
C7 C6 C5 C4 C3 C2 C1 N7 N6 N5 N4 N3 N2 N1
N1N2N3N4N5N6N7C1C2C3C4C5C6C7
0.0 0.2 0.4 0.6 0.8
Distance
D
Distance
0.0 0.2 0.4 0.6 0.8
N5 N3 N6 N1 N2 N4 N7 C5 C1 C6 C7 C2 C3 C4

Figure 3 (Liu)

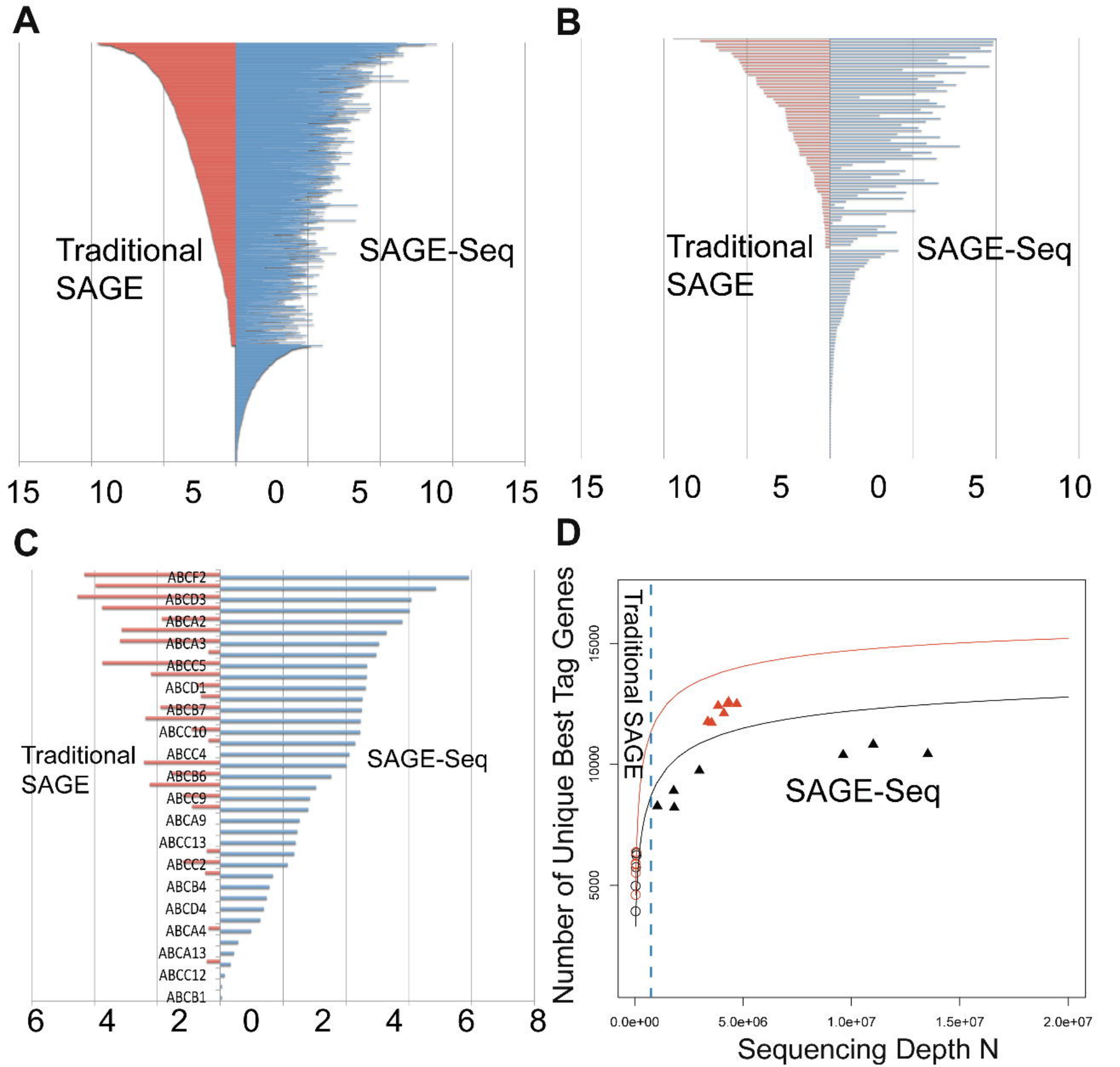
A
Traditional SAGE
SAGE-Seq
15 10 5 0 5 10 15
B
Traditional SAGE
SAGE-Seq
15 10 5 0 5 10
C
ABCF2
ABCD3
ABCA2
ABCA3
ABCC5
ABCD1
ABCB7
ABCC10
ABCC4
ABCB6
ABCC9
ABCA9
ABCC13
ABCC2
ABCB4
ABCD4
ABCA4
ABCA13
ABCC12
ABCB1
Traditional SAGE
SAGE-Seq
6 4 2 0 2 4 6 8
D
Number of Unique Best Tag Genes
5000
10000
15000
Traditional SAGE
SAGE-Seq
0.0e+00
5.0e+06
1.0e+07
1.5e+07
2.0e+07
Sequencing Depth N

Figure 4 (Liu)

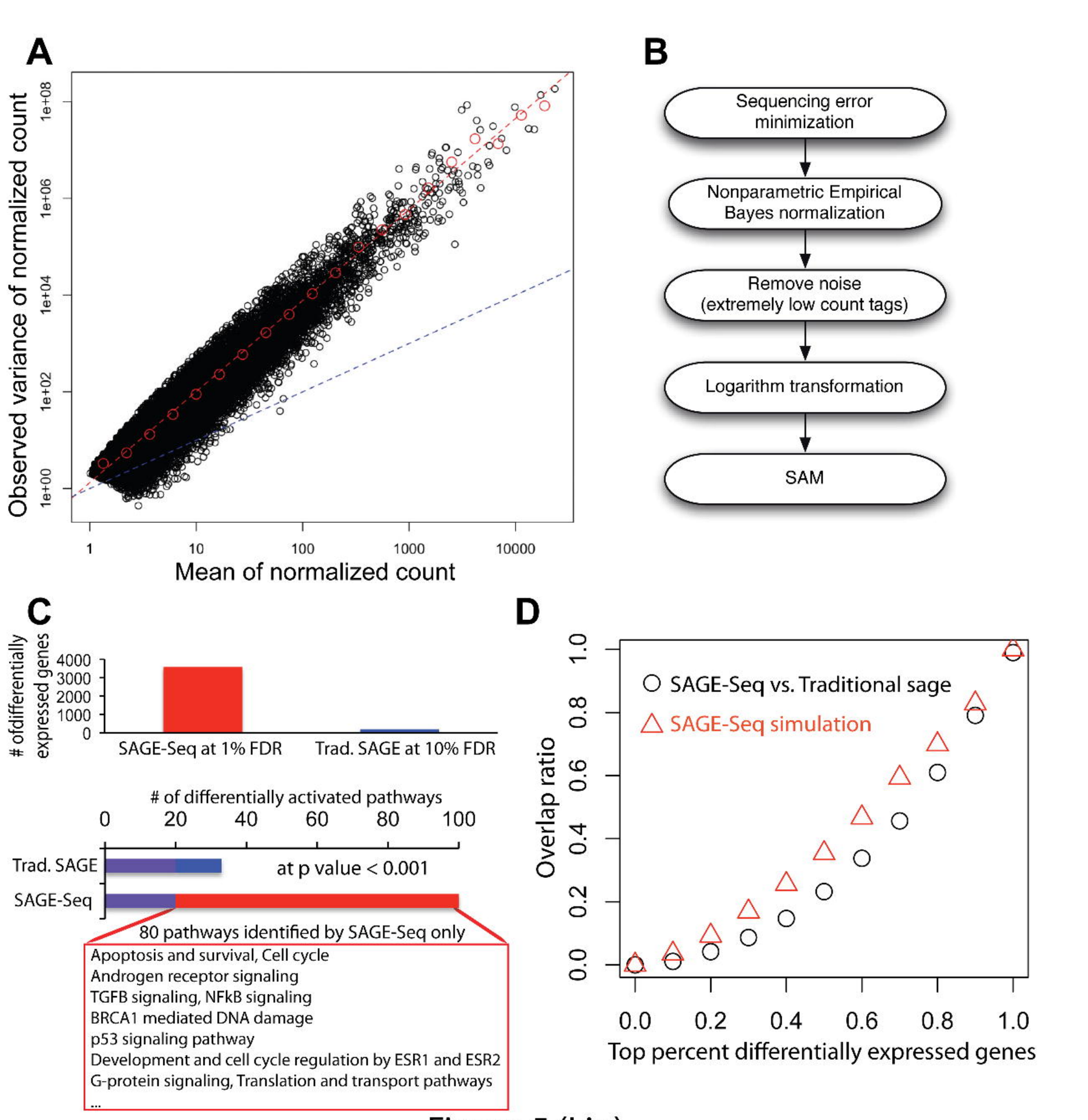
A
Observed variance of normalized count
1e+00
1e+02
1e+04
1e+06
1e+08
1
10
100
1000
10000
Mean of normalized count
B
Sequencing error minimization
Nonparametric Empirical Bayes normalization
Remove noise (extremely low count tags)
Logarithm transformation
SAM
C
ofdifferentially expressed genes
4000
3000
2000
1000
0
SAGE-Seq at 1% FDR
Trad. SAGE at 10% FDR
# of differentially activated pathways
0
20
40
60
80
100
Trad. SAGE
SAGE-Seq
at p value < 0.001
80 pathways identified by SAGE-Seq only
Apoptosis and survival, Cell cycle
Androgen receptor signaling
TGFB signaling, NFkB signaling
BRCA1 mediated DNA damage
p53 signaling pathway
Development and cell cycle regulation by ESR1 and ESR2
G-protein signaling, Translation and transport pathways
...
D
Overlap ratio
0.0
0.2
0.4
0.6
0.8
1.0
SAGE-Seq vs. Traditional sage
SAGE-Seq simulation
Top percent differentially expressed genes

Figure 5 (Liu)