Assessment of a highly multiplexed RNA sequencing platform and comparison to existing high-throughput gene expression profiling techniques

Eric Reed[1,2], Elizabeth Moses[2], Xiaohui Xiao[2], Gang Liu[2], Joshua Campbell[1,2], Catalina Perdomo[2], Stefano Monti[1,2]

1. Bioinformatics Program, Boston University, Boston, MA, USA
2. Section of Computational Biomedicine, Boston University School of Medicine, Boston, MA, USA

Correspondence
Stefano Monti
smonti@bu.edu

Abstract

The need to reduce per sample cost of RNA-seq profiling for scalable data generation has led to the emergence of highly multiplexed RNA-seq. These technologies utilize barcoding of cDNA sequences in order to combine samples into single sequencing lane to be separated during data processing. In this study, we report the performance of one such technique denoted as sparse full length sequencing (SFL), a ribosomal RNA depletion-based RNA sequencing approach that allows for the simultaneous sequencing of 96 samples and higher. We offer comparisons to well established single-sample techniques, including: full coverage Poly-A capture RNA-seq and microarray, as well as another low-cost highly multiplexed technique known as 3' digital gene expression (3' DGE). Data was generated for a set of exposure experiments on immortalized human lung epithelial (AALE) cells in a two-by-two study design, in which samples received both genetic and chemical perturbations of known oncogenes/tumor suppressors and lung carcinogens. SFL demonstrated improved performance over 3' DGE in terms of coverage, power to detect differential gene expression, and biological recapitulation of patterns of differential gene expression from in vivo lung cancer mutation signatures.

Keywords:
RNA Sequencing, Gene Expression, Microarray, Multiplexing, Platform Comparison

Introduction

Since its inception in 2008, RNA sequencing has become the gold-standard for whole-transcriptome high-throughput data generation(Mortazavi et al., 2008). In addition to RNA transcript expression quantification, RNA-seq allows for more advanced analyses including *de novo* transcriptome assembly(Robertson et al., 2010) and characterization of alternative splicing variants(Bryant et al., 2012). Furthermore, RNA-seq is species agnostic, such that the same library preparation technique may be utilized for humans, mouse, rat, kidney bean, etc. These represent clear advantages over hybridization-based microarray platforms in which individual microarray platforms are designed to quantify specific transcripts for a specific species(Wang et al., 2009). However, one persistent drawback of RNA-seq has been its relatively high cost. The use of classic RNA-seq techniques for experimental designs that require profiling of many samples – especially when the marginal information value of each sample is relatively low, such as in medium- and high-throughput screening applications – can thus present a disqualifying cost burden.

Large-scale projects based on transcriptional profiling of chemical exposure experiments include the Toxicogenomics Project-Genomics Assisted Toxicity Evaluation System (Open TG-GATEs)(Igarashi et al., 2015), the DrugMatrix database(Ganter et al., 2006), and the Connectivity Map (CMap)(Subramanian et al., 2017), among others. Both the TG-GATEs and the DrugMatrix projects used microarrays for expression profiling, which was at the time significantly less costly than full coverage RNA-sequencing, yet still requiring multi-million budgets. Alternatively, the CMap project utilizes the Luminex-1000 (L1000) profiling platform, a bead-based analog expression assay which quantifies 1,058 human transcripts, which are used to impute the expression of 11,350 additional transcripts(Subramanian et al., 2017). This technique is among the least expensive expression assays available, but it is restricted to human screens and it directly profiles only a limited panel of genes. Given the flexibility of RNA-sequencing platforms, highly multiplexed techniques represent a viable alternative for generating transcriptional data from exposure screens, as well as from other experiments that require a large sample size. Therefore, evaluation of the technical validity of specific techniques serves to inform research strategies for a variety of biological inquiries.

The need to reduce the per sample cost of RNA-seq has led to the adoption of barcoding technologies, where cDNA sequences from individual samples are tagged and their libraries are combined and multiplex sequenced in a single lane(Wang et al., 2011). More recently, these techniques have been optimized to allow multiplex sequencing of 96 samples per lane or higher(Hou et al., 2015; Shishkin et al., 2015). Here, we report the results of our effort at optimizing and evaluating one such technique denoted as sparse full length (SFL) sequencing (Shishkin et al., 2015), a ribosomal RNA depletion-based RNA sequencing approach that allows for the simultaneous sequencing of 96 samples and higher. We offer comparisons to well established single-sample techniques, including: full coverage Poly-A capture RNA-seq and microarray, as well as another low-cost highly multiplexed technique known as 3' digital gene expression (3' DGE)(Asmann et al., 2009). Assessments include comparisons of coverage between the three RNA-sequencing techniques, as well as signal-to-noise and biological recapitulation of gene-level differential signals between treatment groups for the same samples profiled across SFL, microarray, and 3'DGE. For this evaluation study, we generated a set of exposure experiments on immortalized human lung epithelial (AALE) cells (Lundberg et al., 2002)in a two-by-two study design, in which samples received both genetic and chemical perturbations of known oncogenes/tumor suppressors and lung carcinogens (Figure 1). The goal of this report is not only to assess the performance of our optimized highly multiplexed technique, but to inform future research in terms of the strengths and pitfalls of available cost-effective high throughput transcriptomic profiling techniques.

MATERIAL AND METHODS

Samples

Exposure experiments were performed on immortalized human bronchial epithelial cells (AALE). Cells were exposed to both chemical and genotypic perturbations with three replicates per perturbation combination. Cells were thawed from liquid nitrogen and grown up in SAGM small airway epithelial cell growth media (Lonza, Portsmouth NH). Cells were subcultured using Clonetics ReagentPack subculture reagents (Lonza, Portsmouth NH). In preparation for exposure, cells were plated into 24-well plates and allowed to reach confluency for 24 hours. Cell culture media was then replaced, and compounds added at a concentration of 24 μ g/ml CSC, 173 μ M BaP, 490 μ M NNK or DMSO. NNK and BaP compounds were obtained from Sigma-Aldrich (St. Louis MO) and CSC obtained from Murty Pharmaceuticals (Lexington, KY). Genotypic perturbations included CRISPR knockouts of *FAT1*, and *CDKN2A*, as well as overexpression of *NRF2* (*NFE2L2*), *FGFR1*, *NRG1* and *PIK3CA*. Cells transfected with a pSpCas9-EGFP (*GFP*) plasmid (PX458) in the absence of sgRNAs were used as controls for the CRISPR perturbations while overexpression of an empty vector containing the reporter HcRed served as control for the overexpression experiments. The same samples were profiled across SFL, microarray, and 3' DGE for a subset of combinations of exposures, though all samples were profiled by SFL. In addition, full coverage poly-a capture Full coverage poly-A RNA-seq was performed on a separate set of samples for a subset of genotypic exposures, including CRISPR knockouts of *FAT1*, as well as overexpression of *NRF2*, *NRG1* and *PIK3CA*. These samples did not receive any chemical exposures (Figure 1). Note that in a few cases there was not enough material to perform 3' DGE, as indicated by the sample numbers of certain perturbation combinations.

Library Preparation

Library preparation for SFL sequencing was carried out based on the published protocol (Shishkin et al., 2015). An edited version of this protocol is available in the Supplementary Methods. RNA was isolated using a standard Qiazol and Qiacube protocol from Qiagen (Valencia, CA). RNA purity was assessed using a NanoDrop spectrophotometer and no samples were excluded from downstream analysis. The dual-barcoded SFL libraries were pooled from 96 individual samples and then sequenced on the Illumina® NextSeq 550 to generate more than 400 million Single-Read 75-bp reads. Poly-A RNA Sequencing libraries were prepared from total RNA samples using Illumina® TruSeq® RNA Sample Preparation Kit v2 and then sequenced on the Illumina® HiSeq 2500 to generate more than 5 million single-end 50-bp reads per sample. Microarray procedures were performed as described in GeneChip™ WT PLUS Reagent Kit manual and GeneChip™ WT Terminal Labeling and Controls Kit protocol (Thermo Fisher Scientific). The labeled fragmented DNA was generated from 100 ng of total RNA and was hybridized to the GeneChip™ Human Gene 2.0 ST Array. Microarrays were scanned using Affymetrix GeneArray Scanner 3000 7G Plus. 3'DGE library preparation was performed by *Broad Institute, Cambrige, MA, USA*, similar to (Soumillon et al., 2014). Final libraries were purified using AMPure XP beads (Beckman Coulter) according to the manufacturer's recommended protocol and sequenced on an Illumina NextSeq 500 using paired-end reads of 17bp (read1) + 46bp (read2). Read1 contains the 6-base well barcode along with the 10-base UMI. Across all platforms, the number of samples that were successfully profiled per perturbation combination is shown in Figure 1.

Data Pre-processing

Affymetrix GeneChip Human Gene 2.0 ST Microarray CEL files were annotated to unique Entrez gene IDs, using a custom CDF file from BrainArray (hugene20st_Hs_ENTREZG_21.0.0) and RMA-normalized. For SFL, adapter sequences were trimmed from raw sequence files using *Cutadapt v1.12*. Quality assessment of trimmed SFL sequence files as well as raw full coverage RNA-seq sequencing files was performed with *FastQC v0.11.5*. Both SFL and RNA-seq reads were aligned to human genome (UCSC RefSeq hg19) with *STAR v2.5.2b*(Dobin et al., 2013) and expression quantification in RefSeq genes was carried out with *featureCounts (subread) v1.5.0*(Liao et al., 2014). For 3' DGE, pre-quantified gene expression count matrices were obtained from the *Broad Institute, Cambrige, MA, USA*. These reads had been aligned to the transcriptome (UCSC RefSeq hg19), using *BWA aln v0.7.10*(Li and Durbin, 2009), such that reads with the same UMI and sample barcode were only counted once per gene. All further data processing and analysis were carried out in *R*.

Coverage Assessment

Read coverage across the 82 samples, shared between SFL and 3' DGE, as well as all 18 full coverage RNA-seq samples was assessed for library size as well as percentage of the library size that was aligned, uniquely aligned (i.e. reads that only align once in the genome), and counted in the 18,488 genes with shared annotations across all three platforms. Unlike SFL and full coverage RNA-seq, 3'DGE reads are aligned directly to mRNA sequences, such that the reported numbers of counted reads and uniquely aligned reads are the same. To assess the relative distribution of reads across the total set of 18,488 shared genes, we plotted the cumulative proportion of the sum of reads aligning to individual genes across all samples ranked by relative expression across all three platforms.

Signal-to-Noise Assessment

Signal-to-noise was compared among SFL, 3'DGE and microarrays based on four-group ANOVA analysis and two-group differential analysis. In the ANOVA analysis, the signal-to-noise was assessed across like samples undergoing exposure to CSC or DMSO vehicle, as well as genotypic perturbations of *NRF2* overexpression or HcRed control. Thus, the analysis included four independent groups of samples, receiving each combination of chemical (CSC or DMSO) and genotypic (*NRF2* or HcRed) perturbations, with three replicates in each group. Only genes with mean expression ≥ 1 across all 12 samples in both SFL and 3' DGE were included in the analysis (9,813 total genes). Expression levels across SFL and 3'DGE were normalized via trimmed mean of M values (TMM)(Robinson and Oshlack, 2010) scaling and $\log_2$ counts-per-million transformation. Signal-to-noise was assessed for each gene in each platform by performing a classic ANOVA analysis across the four groups. Additionally, two-group differential analysis was performed for each stratified chemical and genotypic perturbation. That is, differential expression of CSC- vs. DMSO-treated samples, within either HcRed or *NRF2* treatment, as well as differential expression of *NRF2*- vs. HcRedtreated samples, within either DMSO or CSC exposure, was performed. For all three platforms, differential analysis was carried out based on normalized expression values with *LIMMA v3.30.7*(Ritchie et al., 2015). All p-values reported from two-group differential analysis are two-sided. In both ANOVA and LIMMA analyses, nominal p-values for each gene were corrected for multiple comparisons using the Benjamini-Hochberg procedure(Benjamini and Hochberg, 1995).

Biological Signal Recapitulation

Two-group differential analysis signatures were compared by pre-ranked gene set enrichment analysis (GSEA) to gene sets derived from published signatures of smoking exposure in the airway from healthy volunteers (Beane et al., 2007; Spira et al., 2004), as well as to gene sets analytically derived from The Cancer Genome Atlas (TCGA) for patients with lung squamous cell carcinoma (LUSC) or lung adenocarcinoma (LUAD). The two smoking gene sets consist of genes reported as either up- or down-regulated in response to smoking in at least one of the two publications, while TCGA gene sets were derived by probing differential expression of individual genes between patients with or without point mutations or copy number alterations (CNA) in genes of interest. These include mutations for the same panel of genes profiled for genotypic perturbations. In addition we include *KEAP1* mutations, a repressor of *NRF2*(Kansanen et al., 2013, 1). Specifically, point mutation signatures were derived from LUSC and LUAD, independently, by performing differential analysis of subjects with and without point mutations in genes of interest, matched for age, sex, and cancer stage. Likewise, CNA gene signatures were assessed for amplification and deletions of genes of interest by differential analysis, using subjects with zero, one, or two additional copies or deletions of a gene of interest, respectively. All models for mutations and CNA were adjusted for tumor purity, as reported(Campbell et al., 2016). Differential signatures were derived using *LIMMA v3.30.7*. Genes associated with specific mutations or CNA were defined as those with significance and magnitude of the linear model's genetic alteration coefficient at FDR Q-value < 0.05 and |log2 fold-change| > $\log_2(1.5)$, respectively.

Each of our genotypic perturbation signatures was compared by GSEA to the corresponding TCGA-derived gene sets. For example, the *PIK3CA* overexpression signatures were compared to the gene sets derived from *PIK3CA* mutation and copy number alterations in the TCGA data.

Materials

Coverage Assessment

Comparison of coverage of the three sequencing platforms, full coverage poly-A RNA-seq, SFL, and 3' DGE, is summarized in Table 1 and Figure 2. Comparison between SFL and 3'DGE included 82 samples each, while full coverage poly-A RNA-seq included all 18 available samples. Unsurprisingly, full coverage poly-A RNA-seq generated the largest library size, i.e., number of reads per sample, while the SFL and 3'DGE libraries were of comparable size (Figure 2A). Furthermore, full coverage poly-a RNA-seq yielded the highest percentage of reads aligned to the genome, followed by SFL and 3' DGE (Figure 2Bi). For SFL there was a clear drop-off when going from percentage of aligned reads to percentage of uniquely aligned reads due to ribosomal RNA (rRNA) contamination of the SFL samples (Figure 2Bii). The majority of reads aligning to ribosomal regions specifically align to RNA28S (Figure S1). For 3' DGE, unique UMIs are aligned directly to transcript sequences and not to the whole genome, such that the number of uniquely aligned reads and reads counted in transcripts are the same (Figure 2Bii-iii)(Morrissy et al., 2009). The percentage of reads that are counted in transcripts is greatest for full coverage poly-A RNA-seq (mean percentage of total library size: 65.2%), followed by 3' DGE (33.3%), and SFL (24.5%). However, while the functional library size (i.e., the number of counted reads) is greater for 3' DGE than for SFL, more genes were quantified by SFL than by 3' DGE (Figure 2Biv) (counts > 0 across all samples for 22,233 genes shared across all three platforms,). A median of 60.9% and 50.5% genes were quantified by SFL and 3'DGE, respectively. This is further illustrated in Figure 2C, where it is shown that the reads are more evenly distributed across the 22,233 genes by SFL than by 3'DGE, with the cumulative distribution of reads counted in individual genes nearly identical in SFL and full coverage poly-A RNA-seq.

In summary, despite lower overall counted library size due to ribosomal RNA contamination, SFL demonstrates greater coverage in low-to-medium expressed genes than 3' DGE, comparable to full coverage poly-A RNA-seq.

Signal-to-Noise Evaluation

Differential gene expression across experimental groups of matched samples was performed in SFL, microarray, and 3' DGE over the 9,713 shared genes, after removal of low-coverage genes (mean read counts < 1), across SFL and 3'DGE, and the corresponding signal-to-noise scores were compared pairwise (Figure 3). Samples shared across the three platforms include 3 replicates for each of four experimental groups, corresponding to *NRF2* overexpression or HcRed vehicle, as well as CSC chemical exposure or DMSO vehicle (Figure 1). Signal-to-noise was assessed by a four-group comparison with classic ANOVA (Figure 3A-D), as well as by stratified two-group differential analyses using LIMMA (Figure 3 E-F).

We compared the $\log_{10}$ F-statistics between ANOVA models across all three platforms (Figure 3A). Overall, the distribution of F-statistics is most similar between SFL and microarrays, with a Pearson correlation of 0.291. The corresponding mean difference between $\log_{10}$ F-statistics is 0.026, and is significantly skewed towards the SFL results (p-value < 0.01). The mean differences of the $\log_{10}$ F-statistics between SFL and 3'DGE, and between 3'DGE and microarray are 0.328 and 0.302, respectively, and the corresponding Pearson correlations are 0.160 and 0.216, respectively. These results are consistent with the discovery rates estimated for different FDR Q-value thresholds (Figure 3B). For example, at the FDR Q-value threshold of 0.05, the discovery rates of SFL, microarray, and 3'DGE are 0.214 (2083 genes), 0.209 (2038 genes), and 0.032 (310 genes), respectively.

Loess regression of the $\log_{10}$ F-statistics as a function of mean gene expression shows that the statistical signal increases with mean normalized expression. This trend is consistently positive for both SFL and 3'DGE, while leveling off at the most highly expressed genes in microarrays (Figure 3C). Furthermore, SFL signal is greater than 3' DGE signal at all levels of mean expression (Figure 3C). In agreement with the results from coverage comparison, the distribution of mean normalized expressions in 3'DGE is smaller than that of SFL, while SFL is comparable to that of microarray (Figure 3D). Adherence to assumption of normality, assessed through a Shapiro-Wilk test, is also associated with higher mean normalized expression (Figure S3).

The results of the comparisons of the two-group differential analyses across all three platforms were generally congruous with those of the four-group ANOVA analyses (Figure 3E-F, Figure S4). In all four two-group comparisons, the correlation of test statistics is closest between microarray and SFL results, followed by 3' DGE versus microarray results, and 3'DGE versus SFL. For example, in the DMSO-stratified, *NRF2* versus HcRed analysis, estimates of the Pearson correlations of test statistics are 0.64, 0.43, and 0.38, respectively (Figure 3E). The discovery rate of 3' DGE is the lowest across all four differential analyses (Figure 3F).

In summary SFL demonstrated greater statistical power than 3'DGE to detect differentially expressed genes, and its results more closely matched those in microarrays.

Biological Signal Recapitulation Evaluation

To evaluate the ability of each platform to recapitulate biologically relevant results, we utilized previously published signatures of smoking exposure in lung(Beane et al., 2007; Spira et al., 2004), as well as differential signatures derived from the TCGA LUSC and LUAD datasets associated with mutations of the genes over-expressed in our experiments. From each of these signatures two gene sets were extracted, one of genes positively associated and one of genes negatively associated to the variable of interest. These gene sets were then tested via pre-ranked gene set enrichment analysis against each of our differential analysis results (CSC *vs.* DMSO, stratified by *NRF2* or HcRed perturbation; *NRF2* vs. HcRed, stratified by CSC or DMSO perturbation). The enrichment results with respect to both the smoking exposure signatures and the TCGA mutations are summarized in Figure 4A, and further detailed in Figure 5S, and confirm the highest sensitivity of microarrays, followed by SFL and 3'DGE.

The set of genes up-regulated in "smokers *vs.* non-smokers" was found to be significantly (q<0.05) enriched in all "CSC *vs.* DMSO" signatures, within both genotypic stratifications for all three platforms. Conversely, the set of down-regulated genes in "smokers *vs.* non-smokers" was only enriched in the microarray signature of "*NRF2* over-expressed; CSC *vs.* DMSO" (Figure S5).

The enrichment results of TCGA-derived gene sets with respect to differential signatures of genotypic perturbations were in agreement with the gene-level results, in that they consistently demonstrated smaller discovery rates by 3'DGE than by SFL or by microarrays (Figure 4A). For example, the significantly enriched gene sets in "DMSO-treated; *NRF2 vs.* HcRed" differential signatures across all three platforms are highlighted in boxes in Figure S5. The number of gene sets enriched in microarray, SFL, and 3' DGE platforms are six, four, and zero, respectively.

In addition to comparing which gene sets were significantly enriched in individual differential signatures, we compared the relative statistical signal of these enrichments. To this end, we transformed the permutation-based p-values to z-scores, using the direction of individual enrichment scores. For each two-platform comparison, we fit a regression model through the origin. Since consistent results across platforms would result in a model fit close to the identity line, *y=x*, we tested whether the slope coefficient equaled 1 (i.e. $B_1 = 1$). Figure 4B shows these results for each of the three comparisons of the *NRF2* and *KEAP1* mutation-based gene sets enrichment against the "DMSO-treated; *NRF2 vs.* HcRed" signatures. In all three comparisons, microarrays have the strongest statistical enrichment signal, followed by SFL and 3'DGE. In none of the three comparisons does the 95% confidence interval of these regression coefficients included slope = 1. However, the comparison of SFL to microarray results was closest, $B_1 = 0.65$; p-value = 0.01, followed by the comparison between 3'DGE and microarray, and between 3'DGE and SFL, with $B_1 = 0.38$ and 0.29, respectively. Comparison of the enrichment results for other differential signatures show similar trends (Figure S6, top).

Next, we compared enrichment results with respect to all genotypic perturbation signatures between SFL and 3'DGE (Figure 5; Figure S7A). Each comparison (i.e., each point in the plot) denotes gene set enrichment results with respect to genotypic perturbations within each of the four chemical exposures, DMSO, CSC, BaP, and NNK. Gene sets were tested for enrichment against concordant differential signatures, e.g., the *PIK3CA* mutation-derived gene set was tested against the "*PIK3CA vs.* HcRed" signatures. As in the previous analysis, the permutation-based enrichment p-values were z score-transformed. In the "DMSO-treated; genotypic perturbation *vs.* control" signatures, we observe that the gene set enrichment is generally more significant for SFL than for 3'DGE ($B_1 = 0.57$; p-value < 0.01; Figure 5). The results obtained in CSC- and NNK-treated signatures, demonstrate concordance to these results ($B_1 = 0.61$; p-value = 0.23 and $B_1 = 0.63$; p-value = 0.08, respectively). The BaP-treated results are less comparable since only one genotypic perturbation signature, "*FAT1 vs.* GFP", is available for this stratification (Figure S6A).

Finally, we compared our differential signatures to available full coverage poly-A RNA-seq genotypic perturbations (Figure S7B), although these results proved less comparable because of differences in experimental set-up. In particular, in many of the full coverage poly-A RNA-seq experiments the genotypic perturbations were performed on naïve rather than DMSO-treated cell lines (Figure 1).

In summary, differential analysis of molecular and genotypic perturbations with SFL recapitulates biologically meaningful signal of gene sets derived from high coverage in vivo data sets. This performance is comparable to both 3'DGE and microarray.

Discussion

The goal of this study was to evaluate the performance of SFL sequencing, a low-cost method for performing highly multiplexed RNA-seq, and to compare it to other high-throughput gene expression profiling platforms. The development of such methods would be instrumental to the generation of large-scale perturbation screens based on in-vitro models. The reduction of the cost per profile would make it feasible to significantly increase the number of replicates and conditions to be profiled, including multiple time points, concentrations, and biological models, and thus would support a more in-depth investigation of the heterogeneity of the biological response to different exposures. It would also support the development of more accurate predictive models of the adverse or therapeutic outcomes of various exposures. Finally, insights gained from our study will also inform the design of protocols for single cell RNA-sequencing (Eberwine et al., 2014), given their reliance on highly-multiplexed libraries.

In addition to SFL, the platforms included in this analysis were 3'DGE, an alternative highly multiplexed sequencing platform, Affymetrix GeneChip Human Gene 2.0 ST Microarray, an analog expression platform, and full coverage poly-A capture RNA-seq. Performance was assessed in terms of coverage, signal-to-noise, and recapitulation of expected biological signal derived from independently generated, publicly available data collected from human subjects. Coverage was assessed by comparing the three digital expression platforms, while signal-to-noise and biological recapitulation was assessed by comparing SFL, 3'DGE, and microarrays. Chemical and molecular perturbations were carried out in the same samples, and concurrently profiled by SFL, 3'DGE, and microarrays. We also leveraged previously generated full coverage poly-A RNA-seq profiles from similar perturbations of AALE cell lines.

For coverage assessment, performance was evaluated in terms of the distribution of total reads, or library size, that were aligned to the human genome, and further quantified in annotated genes. The best performance was expected in full coverage poly-A RNA-seq, given that this is the most well established technique and has by far the highest sequencing depth. This was confirmed, as full coverage poly-A RNA-seq was measured to have the highest per sample library size, percentage of aligned reads, percentage of uniquely aligned reads, and percentage of counted reads (Figure 2). The coverage performance of SFL suffered as a result of rRNA contamination, where as many as 53% of the total library size per sample was assigned to ribosomal regions of the genome (Figure S2).

3'DGE is a poly-A capture technique, therefore ribosomal depletion is not a possible pitfall. 3'DGE generates a short nucleotide tags from transposon-based fragmentation, which are enriched for 3' adjacent sequences of a given transcript(Soumillon et al., 2014). Since many transcripts of the same gene generate identical sequence tags, unique molecular identifiers (UMIs) are used to distinguish between unique tags and duplicate tags generated from PCR amplification. Although mRNA fragment duplication occurs with any RNA-seq protocol, the impact of this artifact on downstream analyses is negligible for techniques, such as SFL, which generate more complex sequence libraries(Parekh et al., 2016).

Due to their size, 3'DGE sequences were aligned directly to human mRNAs, rather than the whole genome. Therefore, percentages of reads aligned and reads counted (Figure 2B i,iii) reflect the percentages of these non-unique UMIs that align to at least one gene and the number of unique UMIs that align to only one gene, respectively. We observe that the percentage of counted reads is greater for 3' DGE than SFL, which is explained by a loss of reads to rRNA contamination in SFL. However, we observe notably more genes quantified by SFL than by 3'DGE (Figure 2A, Figure 2B iv), which indicates that more reads are assigned to fewer genes in 3'DGE compared to SFL, as well as to full coverage RNA-seq (Figure 2C). Although rRNA contamination is a potential drawback of any ribosomal depletion RNA-sequencing technique, the extent of ribosomal contamination is variable, and could be potentially improved by further optimization of the library preparation protocol.

The difference in distribution of reads across shared genes between SFL and 3'DGE likely explains the difference in statistical signal that we observe between the two platforms. In particular, our signal-to-noise evaluation shows consistently higher gene-level statistical signal from SFL and microarray experiments than from 3'DGE experiments (Figure 3). These differences appear to be driven by the differences in the relative quantification of genes, given that statistical signal is positively associated with mean gene expression for each platform, and 3'DGE experiments showed lower gene-level quantification than SFL and microarrays (Figure 3C-D). We observe similar cross-platform relationships in the two-group differential analyses (Figure 3E-F).

The gene set-based enrichment results are consistent with those from signal-to-noise analyses. In every comparison of enrichment scores between SFL and 3'DGE, we observe generally higher gene set enrichment with respect to the SFL-derived signatures (Figure 4, Figure 5, Figure S6, Figure S7). The gene sets were selected to represent known biological responses to the profiled perturbations, and thus their enrichment with respect to the perturbation signatures are expected to be true positives. The enrichment results confirm this expectation. For example, in the signatures of *NRF2* overexpression, we consistently observe enrichment of the gene sets derived from *NRF2* amplifications and *KEAP1* deletions, each of which should increase *NRF2* activity (Figure S5)(Kansanen et al., 2013). Similarly, we observe significant concordant enrichment of the gene sets derived from *NRF2* and *KEAP1*-dysregulated lung tumors in the signature of CSC exposure, suggesting that the *NRF2* pathway is activated by CSC exposure in vitro (Figure S5), which has been previously reported(Adair-Kirk et al., 2008). Interestingly, these results demonstrate that the activation of the *NRF2* pathway in normal airway epithelial cells *in vitro* (by ectopic expression of the gene or by CSC treatment) is concordant with the activation of *NRF2* by somatic genome alterations in lung tumors, a finding that, to the best of our knowledge, has not been previously observed.

The comparatively high-performance of full coverage poly-A RNA-seq and the microarray platforms is unsurprising considering that these platforms represent well-established protocols in which a single sample is profiled per assay.

In summary, in this study we observe higher performance of SFL than 3'DGE, as measured by coverage, signal-to-noise, and biological recapitulation of known signal, with the performance of SFL often matching that of well-established "gold standards" (full coverage RNA-seq or microarrays). On the other hand, the fact the 3'DGE is shown to allocate a large number of reads to relatively fewer, highly expressed genes, makes this platform more suitable for problems where high accuracy in the differential quantification of highly expressed genes is needed. Furthermore, the ready availability of 3'DGE as a core-provided option, which allows for the out-sourcing of library preparation, sequence read pre-processing and gene quantification, is an additional value-added of the platform. Ultimately, the best-suited platform for a specific project will depend on the study goals, design, and availability of different resources. We believe our study presents useful results to make a more informed choice.

The utility of highly multiplexed RNA-seq crucially depends on the trade-off between cost and data quality, and on the nature of the experiments for which the platform would be ideally suitable. These will in general be experiments where the marginal information content of a single profile is relatively low, and thus justifies trading-off some data quality for reduced cost.

Conflicts of Interest

None to report.

Funding

This work was supported by a Superfund Research Program grant P42ES007381 to SM, an Evans Foundation pilot grant to SM and LUNGevity Career Development award to JC.

Acknowledgment

Dr. Alexander A. Shishkin for his feedback during the development of the SFL protocol. SCRB-Seq libraries were prepared by the Broad Technology Labs and sequenced by the Broad Genomics Platform.

Supplementary Material

Supplementary data and figures are available in the file, SupplementaryMaterial.pdf.

References

Adair-Kirk, T. L., Atkinson, J. J., Griffin, G. L., Watson, M. A., Kelley, D. G., DeMello, D., et al. (2008). Distal Airways in Mice Exposed to Cigarette Smoke: Nrf2-Regulated Genes Are Increased in Clara Cells. *Am. J. Respir. Cell Mol. Biol.* 39, 400–411. doi:10.1165/rcmb.2007-0295OC.

Asmann, Y. W., Klee, E. W., Thompson, E. A., Perez, E. A., Middha, S., Oberg, A. L., et al. (2009). 3' tag digital gene expression profiling of human brain and universal reference RNA using Illumina Genome Analyzer. *BMC Genomics* 10, 531. doi:10.1186/1471-2164-10-531.

Beane, J., Sebastiani, P., Liu, G., Brody, J. S., Lenburg, M. E., and Spira, A. (2007). Reversible and permanent effects of tobacco smoke exposure on airway epithelial gene expression. *Genome Biol.* 8, R201. doi:10.1186/gb-2007-8-9-r201.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Methodol.* 57, 289–300.

Bryant, D. W., Priest, H. D., and Mockler, T. C. (2012). “Detection and Quantification of Alternative Splicing Variants Using RNA-seq,” in *RNA Abundance Analysis*, eds. H. Jin and W. Gassmann (Totowa, NJ: Humana Press), 97–110. doi:10.1007/978-1-61779-839-9_7.

Campbell, J. D., Alexandrov, A., Kim, J., Wala, J., Berger, A. H., Pedamallu, C. S., et al. (2016). Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. *Nat. Genet.* 48, 607–616. doi:10.1038/ng.3564.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinforma. Oxf. Engl.* 29, 15–21. doi:10.1093/bioinformatics/bts635.

Eberwine, J., Sul, J.-Y., Bartfai, T., and Kim, J. (2014). The promise of single-cell sequencing. *Nat. Methods* 11, 25–27. doi:10.1038/nmeth.2769.

Ganter, B., Snyder, R. D., Halbert, D. N., and Lee, M. D. (2006). Toxicogenomics in drug discovery and development: mechanistic analysis of compound/class-dependent effects using the DrugMatrix ® database. *Pharmacogenomics* 7, 1025–1044. doi:10.2217/14622416.7.7.1025.

Hou, Z., Jiang, P., Swanson, S. A., Elwell, A. L., Nguyen, B. K. S., Bolin, J. M., et al. (2015). A cost-effective RNA sequencing protocol for large-scale gene expression studies. *Sci. Rep.* 5. doi:10.1038/srep09570.

Igarashi, Y., Nakatsu, N., Yamashita, T., Ono, A., Ohno, Y., Urushidani, T., et al. (2015). Open TG-GATEs: a large-scale toxicogenomics database. *Nucleic Acids Res.* 43, D921–D927. doi:10.1093/nar/gku955.

Kansanen, E., Kuosmanen, S. M., Leinonen, H., and Levonen, A.-L. (2013). The Keap1-Nrf2 pathway: Mechanisms of activation and dysregulation in cancer. *Redox Biol.* 1, 45–49. doi:10.1016/j.redox.2012.10.001.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760. doi:10.1093/bioinformatics/btp324.

Liao, Y., Smyth, G. K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30, 923–930. doi:10.1093/bioinformatics/btt656.

Lundberg, A. S., Randell, S. H., Stewart, S. A., Elenbaas, B., Hartwell, K. A., Brooks, M. W., et al. (2002). Immortalization and transformation of primary human airway epithelial cells by gene transfer. *Oncogene* 21, 4577–4586. doi:10.1038/sj.onc.1205550.

Morrissy, A. S., Morin, R. D., Delaney, A., Zeng, T., McDonald, H., Jones, S., et al. (2009). Next-generation tag sequencing for cancer gene expression profiling. *Genome Res.* 19, 1825–1835. doi:10.1101/gr.094482.109.

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. doi:10.1038/nmeth.1226.

Parekh, S., Ziegenhain, C., Vieth, B., Enard, W., and Hellmann, I. (2016). The impact of amplification on differential expression analyses by RNA-seq. *Sci. Rep.* 6. doi:10.1038/srep25533.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43, e47. doi:10.1093/nar/gkv007.

Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S. D., et al. (2010). De novo assembly and analysis of RNA-seq data. *Nat. Methods* 7, 909–912. doi:10.1038/nmeth.1517.

Robinson, M. D., and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* 11, R25. doi:10.1186/gb-2010-11-3-r25.

Shishkin, A. A., Giannoukos, G., Kucukural, A., Ciulla, D., Busby, M., Surka, C., et al. (2015). Simultaneous generation of many RNA-seq libraries in a single reaction. *Nat. Methods* 12, 323–325. doi:10.1038/nmeth.3313.

Soumillon, M., Cacchiarelli, D., Semrau, S., van Oudenaarden, A., and Mikkelsen, T. S. (2014). Characterization of directed differentiation by high-throughput single-cell RNA-Seq. doi:10.1101/003236.

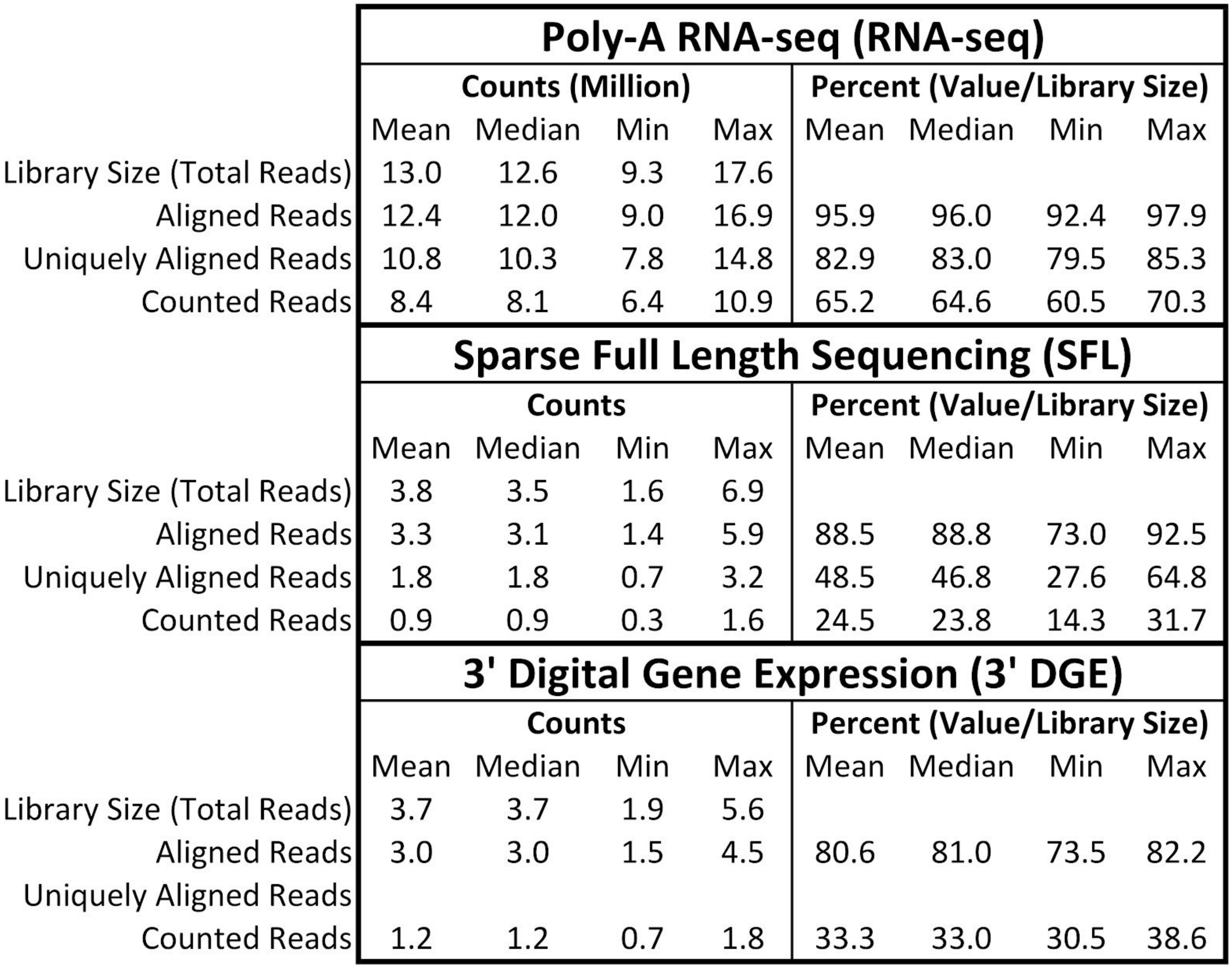
Spira, A., Beane, J., Shah, V., Liu, G., Schembri, F., Yang, X., et al. (2004). Effects of cigarette smoke on the human airway epithelial cell transcriptome. *Proc. Natl. Acad. Sci. U. S. A.* 101, 10143–10148. doi:10.1073/pnas.0401422101.

Subramanian, A., Narayan, R., Corsello, S. M., Peck, D. D., Natoli, T. E., Lu, X., et al. (2017). A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell* 171, 1437-1452.e17. doi:10.1016/j.cell.2017.10.049.

Wang, L., Si, Y., Dedow, L. K., Shao, Y., Liu, P., and Brutnell, T. P. (2011). A Low-Cost Library Construction Protocol and Data Analysis Pipeline for Illumina-Based Strand-Specific Multiplex RNA-Seq. *PLoS ONE* 6, e26426. doi:10.1371/journal.pone.0026426.

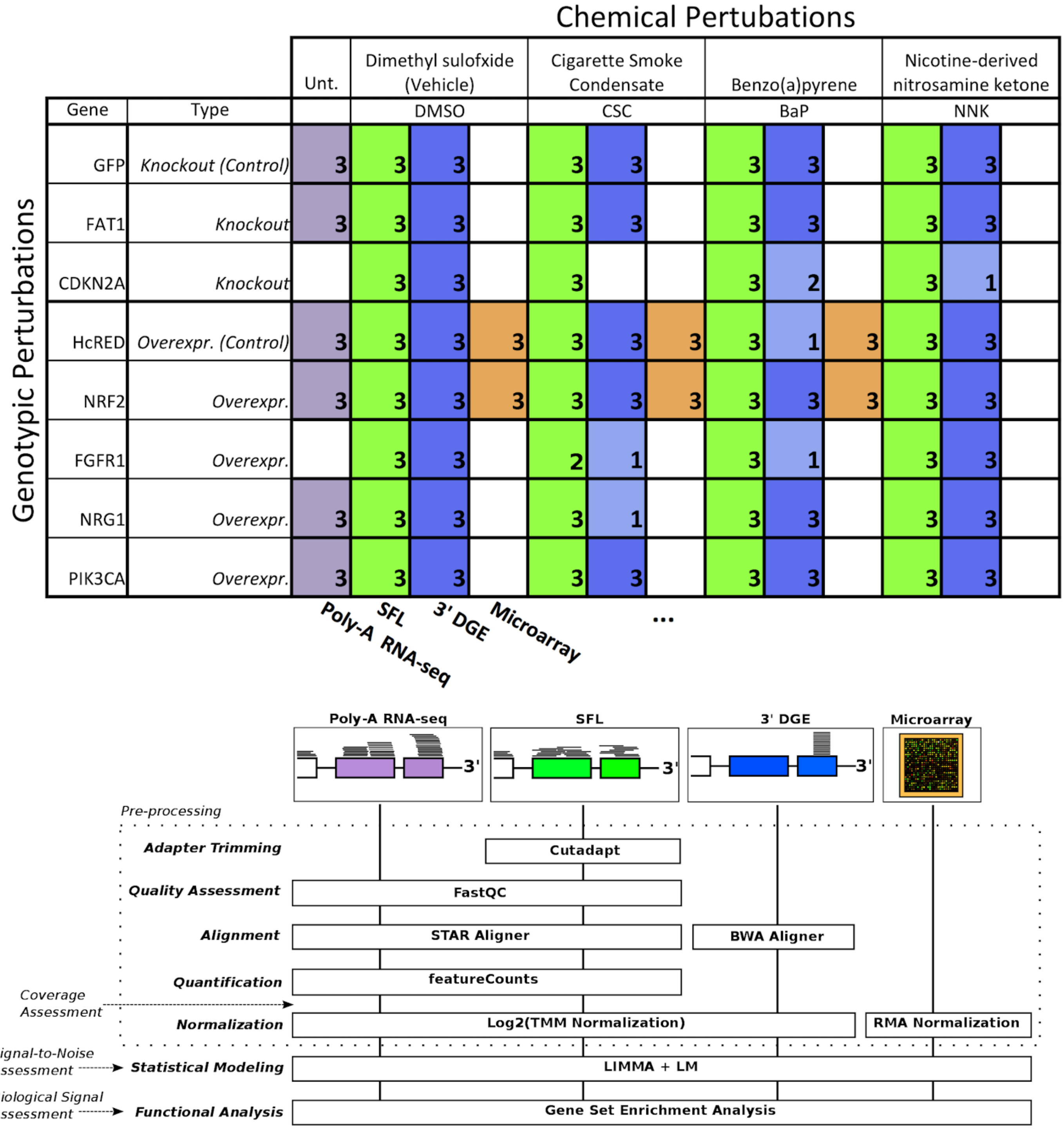
Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10, 57–63. doi:10.1038/nrg2484.

Tables and Figures



	Poly-A RNA-seq (RNA-seq)							
	Counts (Million)				Percent (Value/Library Size)			
	Mean	Median	Min	Max	Mean	Median	Min	Max
Library Size (Total Reads)	13.0	12.6	9.3	17.6				
Aligned Reads	12.4	12.0	9.0	16.9	95.9	96.0	92.4	97.9
Uniquely Aligned Reads	10.8	10.3	7.8	14.8	82.9	83.0	79.5	85.3
Counted Reads	8.4	8.1	6.4	10.9	65.2	64.6	60.5	70.3
	Sparse Full Length Sequencing (SFL)							
	Counts				Percent (Value/Library Size)			
	Mean	Median	Min	Max	Mean	Median	Min	Max
Library Size (Total Reads)	3.8	3.5	1.6	6.9				
Aligned Reads	3.3	3.1	1.4	5.9	88.5	88.8	73.0	92.5
Uniquely Aligned Reads	1.8	1.8	0.7	3.2	48.5	46.8	27.6	64.8
Counted Reads	0.9	0.9	0.3	1.6	24.5	23.8	14.3	31.7
	3' Digital Gene Expression (3' DGE)							
	Counts				Percent (Value/Library Size)			
	Mean	Median	Min	Max	Mean	Median	Min	Max
Library Size (Total Reads)	3.7	3.7	1.9	5.6				
Aligned Reads	3.0	3.0	1.5	4.5	80.6	81.0	73.5	82.2
Uniquely Aligned Reads								
Counted Reads	1.2	1.2	0.7	1.8	33.3	33.0	30.5	38.6

Table 1: Comparison of Coverage Between Poly-A RNA-seq, SFL, and 3'DGE



Chemical Pertubations

Genotypic Perturbations

Gene	Type	Unt.	Dimethyl sulofxide (Vehicle)			Cigarette Smoke Condensate			Benzo(a)pyrene			Nicotine-derived nitrosamine ketone		
			DMSO			CSC			BaP			NNK		
		Poly-A RNA-seq	SFL	3' DGE	Microarray	...								
GFP	*Knockout (Control)*	3	3	3		3	3		3	3		3	3	
FAT1	*Knockout*	3	3	3		3	3		3	3		3	3	
CDKN2A	*Knockout*		3	3		3			3	2		3	1	
HcRED	*Overexpr. (Control)*	3	3	3	3	3	3	3	3	1	3	3	3	
NRF2	*Overexpr.*	3	3	3	3	3	3	3	3	3	3	3	3	
FGFR1	*Overexpr.*		3	3		2	1		3	1		3	3	
NRG1	*Overexpr.*	3	3	3		3	1		3	3		3	3	
PIK3CA	*Overexpr.*	3	3	3		3	3		3	3		3	3	

Figure 1: Design of Cross-Platform Experiments and High-throughput Data Processing

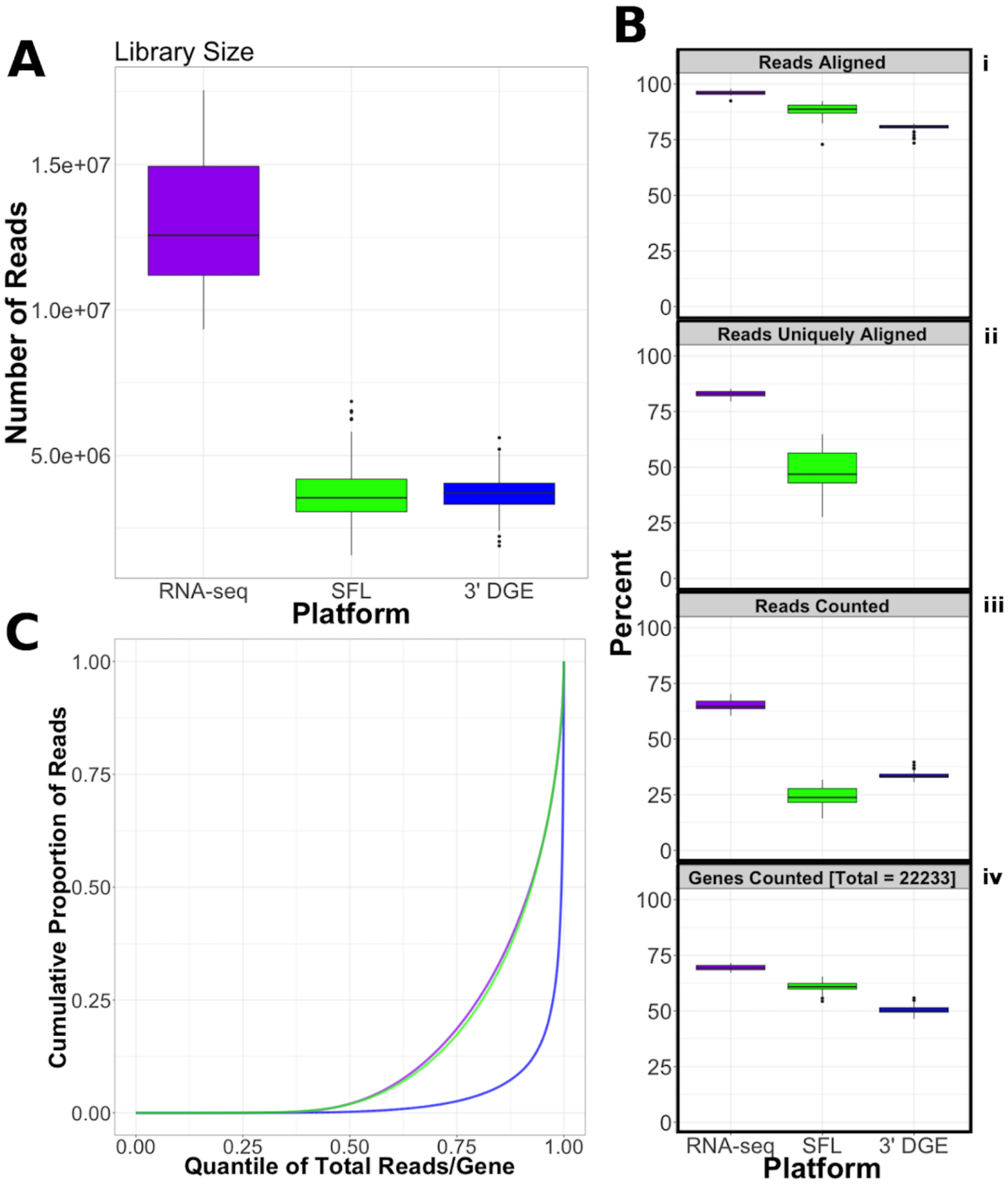


Figure 2: Comparison of Coverage Between Poly-A RNA-seq, SFL, and 3'DGE

A) Boxplots of distribution of library size for each platform.

B) The top 3 boxplots show the percentage of reads aligned (i), uniquely aligned (ii), and counted(iii) relative to the total library size for each platform. The bottom boxplot (iv) shows the proportion of genes with counts > 1, for protein-coding genes annotated across all 3 platforms (18,488). For Figure 2Bii, "Reads Uniquely Aligned" is not shown for 3' DGE because "Reads Uniquely Aligned" and "Reads Counted" are the same values as a result of the data pre-processing protocol, specific to 3' DGE (see Methods).

C) Cumulative distribution of reads assigned to individual genes. The x-axis indicates the quantile for each gene in terms of ranking by relative expression. The y-axis shows the cumulative proportion of total counted reads assigned to these genes, i.e the running sum of reads divided by the total number of reads across all genes.

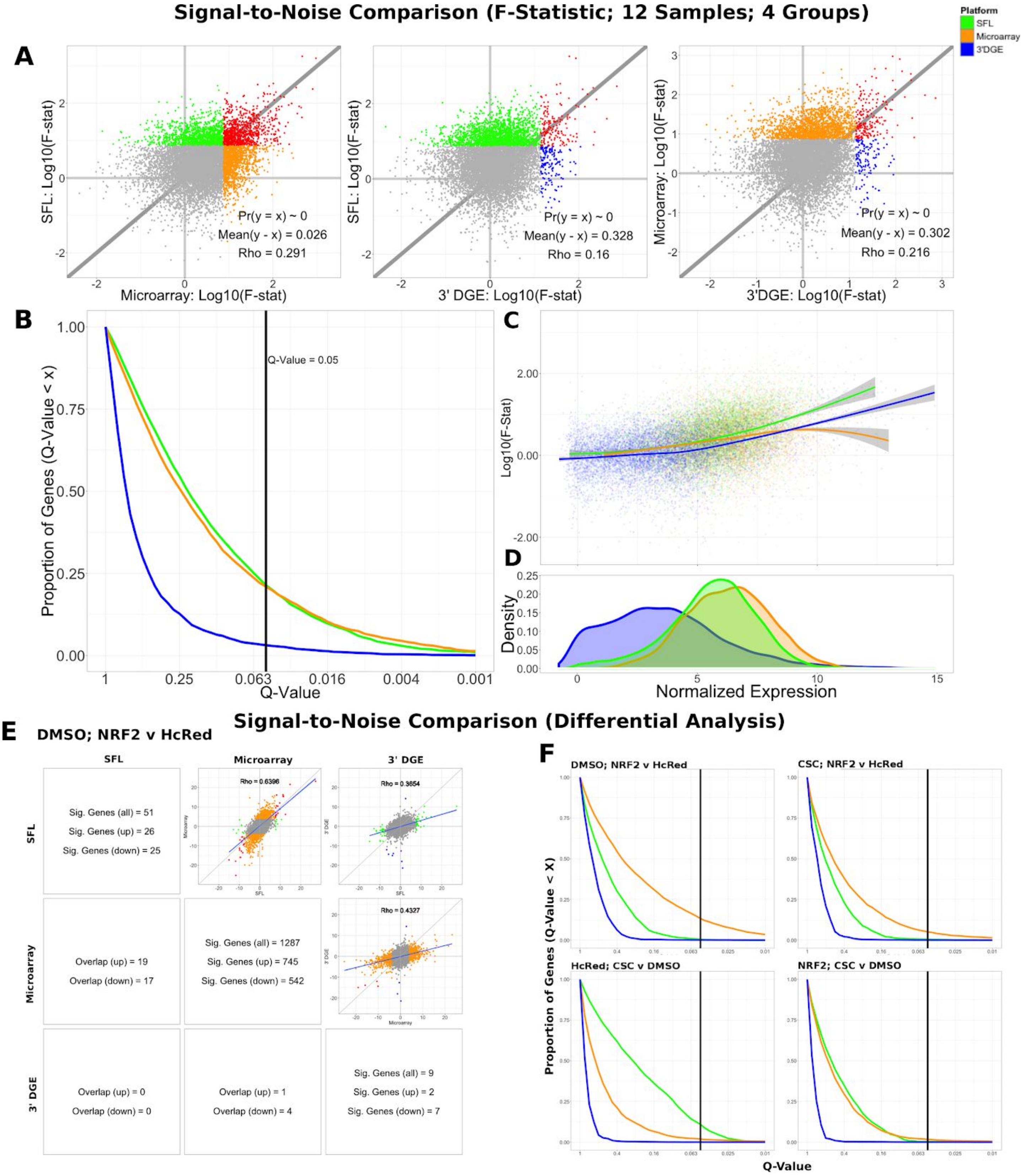


Figure 3: Signal-to-Noise Comparison Between SFL, Microarray, and 3' DGE

A) Scatterplots comparing the log10(F-Statistics) from ANOVA models comparing four n=3 groups (HcRed:DMSO, HcRed:CSC, NRF2:DMSO, and NRF2:CSC). The grey line shows y=x. The platform with the higher mean log10(F-Statistic) is plotted on the y-axis. Also, included are the p-value and difference in mean between each bi-platform comparison from paired t-testing, as well as the squared correlation coefficient. P-values ~ 0 are less than 0.01. Color of indicate genes discovered by individual platforms (green, orange, or blue), neither platform (grey), and both platforms (red).

B) Plot of the Discovery Rate versus FDR Q-Value from threshold for each platform from four group ANOVA models. The x-axis is plotted on a -log10 scale. The vertical line is indicative of a Q-value threshold of 0.05.

C) Loess fit of the log10(F-Statistic) versus median normalized expression from four group ANOVA models.

D) Distribution of mean normalized expression across all three platforms.

E) Comparison of gene discovery (FDR Q-Value < 0.05) by differential analysis with limma, comparing normalized gene expression between DMSO:NRF2 and DMSO:HcRed, including the raw discovery rates, discovered gene overlap, and linear fits, comparing test statistics from each platform. Genes that are discovered by more than 1 platform are shown in red in the scatterplots. Additional comparisons are shown in Figure S5.

F) Plot of the Discovery Rate versus FDR Q-Value from threshold for each platform from two group differential analyses. The x-axis is plotted on a -log10 scale. The vertical line is indicative of a Q-value threshold of 0.05.

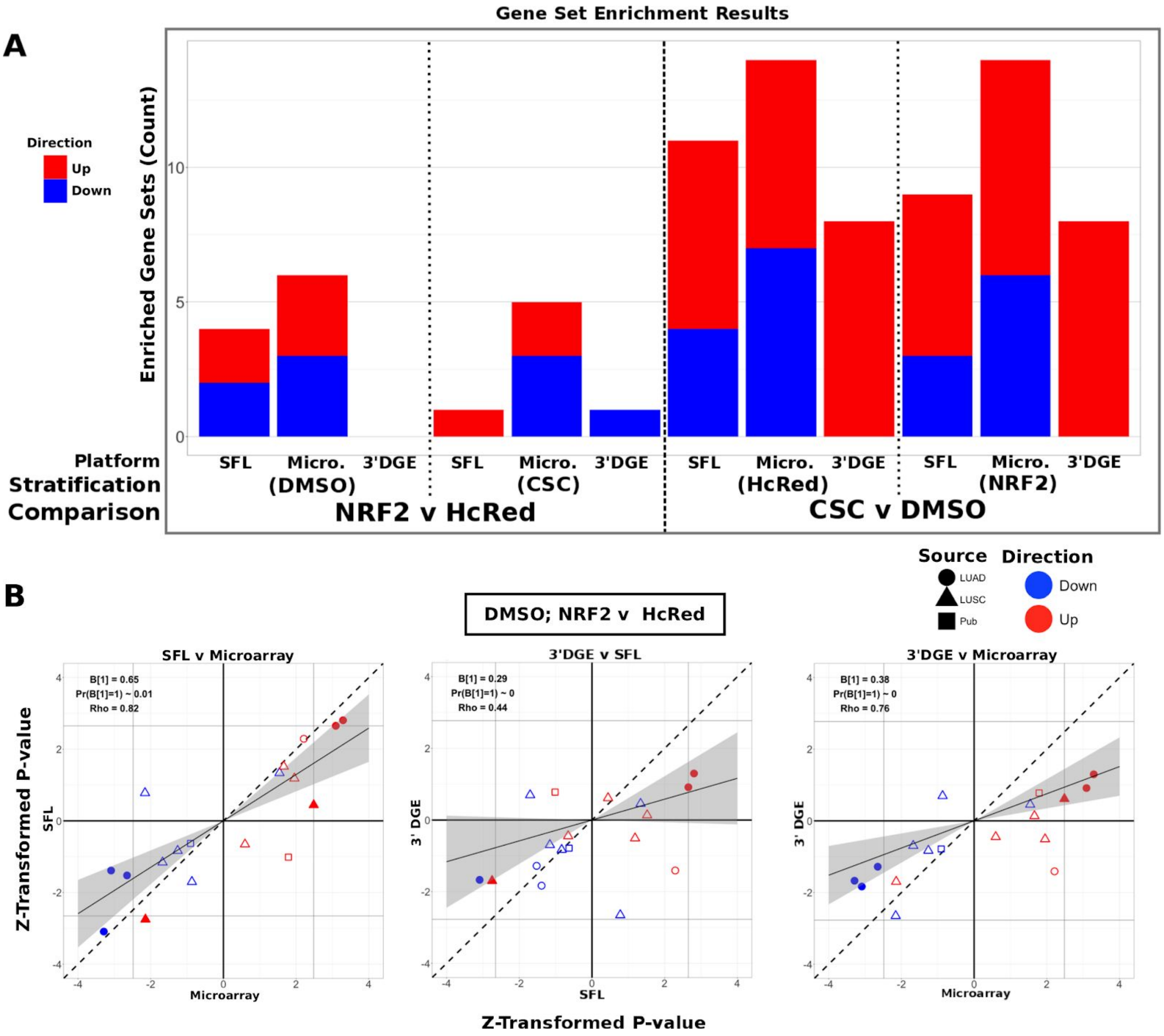


Figure 4: Comparison of Gene-set enrichment of Smoking and Gene Mutation Signatures across SFL, 3'DGE and Microarray

A) Stacked bar charts of counts of statistically significantly TCGA-derived gene-sets with respect genotypic perturbations (left) and chemical perturbations (right) differential signatures across like samples within SFL, Microarray, and 3' DGE. Each column corresponds to differential signatures comparing genotypic or chemical perturbation groups, stratified by a single chemical or genotypic perturbation group, respectively, e.g. the left-most column shows the enrichment results with respect to the "DMSO-treated; NRF2 *vs.* HcRed" signature within the samples (*stratum*) in SFL data. Specific results for TCGA-derived genes sets are shown in Figure S5.

B) Comparison of the gene set enrichment results between SFL, microarray and 3' DGE with respect to the "DMSO-treated; NRF2 *vs.* HcRed" differential signature. Shown are the z-transformed enrichment p-values of the TCGA-derived gene sets corresponding to mutations of NRF2 and CNA of KEAP1. The z-score values corresponding to the FDR=0.05 significance thresholds are shown as vertical and horizontal gray lines for the y and x-axes, respectively. Points of gene sets whose enrichment meets this threshold in either of the two platforms are filled in. Colors and shape of points denote direction and source of the gene set, respectively. Additional results for chemical and genotypic perturbation signatures are shown if Figure S5.

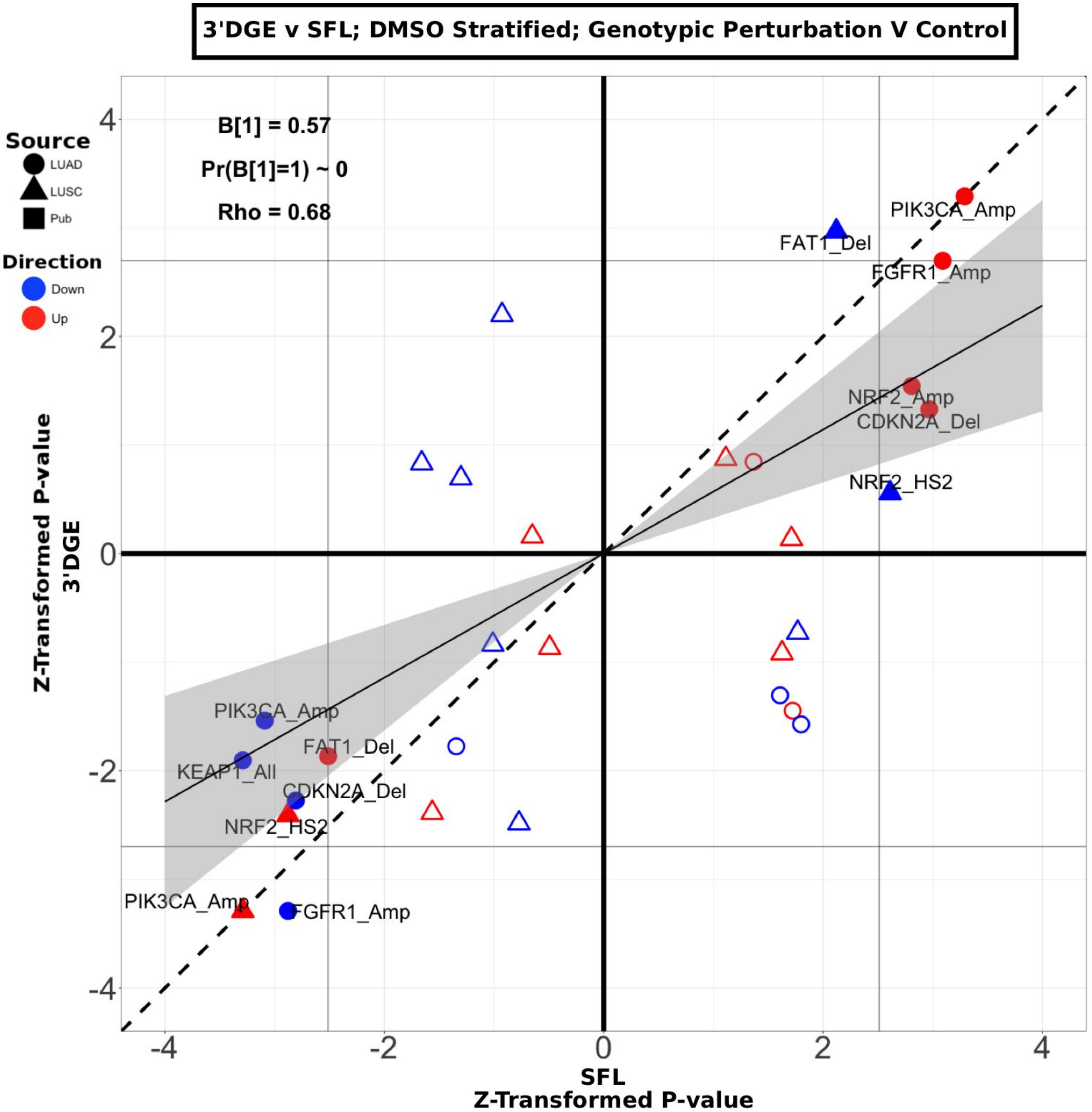


Figure 5: Comparison of Gene-set enrichment of Gene Mutation Signatures across SFL and 3'DGE

Comparison of the gene set enrichment results between SFL and 3' DGE with respect to the "DMSO-treated; genotypic perturbation *vs.* control" differential signatures. Points indicate gene set enrichment against concordant signatures, e.g., PIK3CA mutation and CNA gene sets against the "PIK3CA *vs.* HcRed" differential signatures. Shown are the z-transformed p-values from permutation-based testing by pre-ranked GSEA. The z-score values corresponding to the FDR=0.05 significance thresholds are shown as vertical and horizontal gray lines for the y and x-axes, respectively. The names of the gene sets whose enrichment meets this threshold in either of the two platforms are shown and their points are filled in. Colors and shape of points denote direction and source of the gene set, respectively. Additional results for CSC, NNK, and BaP stratified genotypic perturbation signatures , as well as comparisons between full coverage RNA-seq and either SFL and 3'DGE are shown in Figure S6.