**DOT/FAA/AM-20/09**
**Office of Aerospace Medicine**
**Washington, D.C. 20591**

# Comparison Study of Microarray and RNA-seq for Differential Expression

Susan K. Munster[1], Vicky L. White[1],

David C. Hutchings[2], Dennis M. Burian[1],

Scott J. Nicholson[1]

[1]Federal Aviation Administration Civil Aerospace Medical Institute, Oklahoma City, OK

[2]Venesco, LLC, Chantilly, VA

Civil Aerospace Medical Institute
Federal Aviation Administration
Oklahoma City, OK 73125

October 2018

Final Report

| 1. Report No. DOT/FAA/AM-20/09 | 2. Government Accession No. | 3. Recipient's Catalog No. |
|---|---|---|
| 4. Title and Subtitle<br>Comparison Study of Microarray and RNA-seq for Differential Expression | | 5. Report Date<br>October, 2018<br>6. Performing Organization Code |
| 7. Author(s)<br>Munster SK[1], White VL[1], Hutchings, DC[2], Burian DM[1], Nicholson SJ[1] | | 8. Performing Organization Report No. |
| 9. Performing Organization Name and Address<br>[1]FAA Civil Aerospace Medical Institute, P. O. Box 25082<br>Oklahoma City, OK 73126<br>[2]Venesco LLC, 14801 Murdock St., Ste. 125, Chantilly, VA 20151 | | 10. Work Unit No. (TRAIS)<br>11. Contract or Grant No. |
| 12. Sponsoring Agency Name and Address<br>Office of Aerospace Medicine<br>Federal Aviation Administration<br>800 Independence Ave., S.W.<br>Washington, DC 20591 | | 13. Type of Report and Period Covered<br>14. Sponsoring Agency Code AAM-1 |
| 15. Supplemental Notes<br>CAMI Aerospace Medical Research Division Project No. 2017-AAM-612-GEN-10020 | | |

16. Abstract

As costs for performing RNA-seq approach the costs affiliated with utilizing microarray technology, it is worthwhile to determine which methodology is the most efficient, cost effective, and accurate. In this study, we examined the relative capacities of microarrays and total RNA-seq to detect differential gene expression between total RNA samples containing the following RNA mixtures: blood, brain, 2:1 blood:brain, and 1:2 blood:brain. Two microarray hybridization library amplification methods were compared to optimize microarray results, which were then compared to total RNA-seq results for the same samples. Bioinformatic analysis was performed using commonly accepted analysis software for microarray and RNA-seq data to determine which method produced the greatest number of significant differentially expressed genes. It was determined that total RNA-seq outperformed microarrays in efficiency, result quality, and total numbers of detected genes and that cost was similar to utilizing microarrays for quantifying RNA expression levels from mixtures of brain and blood tissue.

| 17. Key Words | 18. Distribution Statement |
|---|---|
| Microarray, total RNA-seq, differential expression | Document is available to the public through the Defense Technical Information Center, Ft. Belvior, VA 22060; and the National Technical Information Service, Springfield, VA 22161 |

| 19. Security Classif. (of this report) | 20. Security Classif. (of this page) | 21. No. of Pages | 22. Price |
|---|---|---|---|
| Unclassified | Unclassified | 59 | |

**Form DOT F 1700.7** (8-72) Reproduction of completed page is authorized

CONTENTS

LIST OF FIGURES

LIST OF TABLES

## ABBREVIATIONS

| Abbreviation | Explanation |
|---|---|
| cDNA | complementary DNA |
| DE | differential expression |
| FDR | false discovery rate |
| LFC | $\log_2$ fold change |
| all-blood | a sample comprised of 100% blood-derived RNA |
| 2/3-blood | a sample mixture comprised of 67% blood-derived RNA and 33% brain-derived RNA |
| 1/3-blood | a sample mixture comprised of 33% blood-derived RNA and 67% brain-derived RNA |
| all-brain | a sample comprised of 100% brain-derived RNA |
| PCA | principal components analysis |
| PC1 | principal component 1 |
| PC2 | principal component 2 |

[Sequence data has been submitted to NCBI's Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra) and to NCBI's Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo) under BioProject accession number PRJNA492492]

# COMPARISON STUDY OF MICROARRAY AND RNA-SEQ FOR DIFFERENTIAL EXPRESSION

## INTRODUCTION

In the past four decades, since Sanger sequencing (Sanger *et al.,* 1977) was developed and Chang (1983) first introduced the concept of microarrays, the study of genetics has undergone revolutionary changes in what genetic data can be collected and how it is analyzed. Microarrays became a commonly used technology through the efforts of Southern, whose laboratory developed techniques to automate scanning of deoxyribnucleic acid (DNA) sequencing gel data by computer (Elder *et al*., 1986). Fodor founded Affymetrix, which produces high-density microarrays for quantitation and identification of DNA and ribonucleic acid (RNA) samples (Fodor *et al.,* 1991). Brown's publications enabled laboratories to construct their own spotted microarrays (Shalon *et al.,* 1996). Schena's use of Brown's microarrays proved the utility of microarrays for examining gene expression (Schena *et al.,* 1995). Initially, microarray chips were limited to dozens or hundreds of probes, but currently the most advanced microarray chips incorporate millions of probes, covering the entire human genome or transcriptome; the GeneChip® Human Transcriptome Array 2.0 contains more than six million probes (Affymetrix, n.d.) and the Clariom™ D Human microarray (ThermoFisher, 2017) contains more than 6.8 million probes.

RNA-seq is a more recent innovation that utilizes methods and technologies developed for next generation sequencing. Some of the earliest DNA sequencing was performed using Sanger sequencing, initially capable of determining the sequence of approximately 200 nucleotides at a time (Sanger *et al.,* 1977). Later, 454 sequencing was developed by Margulies *et al.* (2005). This method utilized shotgun sequencing of up to 25 million bases in a four-hour run through pyrosequencing. Currently, RNA-seq incorporates molecular techniques to fragment RNA sequences into small reads, convert RNA to complementary DNA (cDNA), and then determine the sequence of those reads by utilizing "sequencing by synthesizing" technologies

that allow the detection of bases added to complementary strands of cDNA as they are synthesized (Bentley *et al.,* 2008) with a high level of accuracy and efficiency.

The purpose of this study was to determine which of these two methods, microarrays or RNA-seq, provide the most accurate, effective, and economical results for biomarker detection. In the past, the Federal Aviation Administration's (FAA) Civil Aerospace Medical Institute's (CAMI) Functional Genomics Research Laboratory has used microarrays for biomarker studies, but there are some concerns related to the quality, dynamic range, and limit-of-detection of data generated using microarrays. One concern is that the maximum microarray detection threshold is determined by the number and specificity of complementary oligonucleotides printed onto the array for each target. This approach could interfere with detection of highly expressed genes, where thousands of copies may be present. Microarrays also are known to be statistically noisy (Mantione *et al.,* 2014; Wang *et al.,* 2014), which could dampen their ability to detect genes at low concentrations. Wang *et al.* (2009), found that there was reasonable concordance between measured expression levels from microarrays and RNA-seq for moderately expressed genes, but poor concordance for genes that had high or low levels of expression. RNA-seq is capable of detecting low-expressing genes expressed with little interference from statistical noise, and also genes expressed at very high levels, with thousands of copies present or more. These findings would indicate that RNA-seq is potentially a more robust methodology for biomarker detection using differential expression.

There were two objectives to the study and was split into two parts in order to optimize biomarker discovery methodology. First, two different amplification methods for microarray hybridization were compared to determine differences, if any, between the Affymetrix GeneChip® WT PLUS (ThermoFisher P/N 902280) amplification method and the NuGEN Ovation® Pico WTA system V2 (NuGEN P/N 3302-12) amplification method, with respect to RNA differential expression (DE) results used for potential biomarker discovery. Our laboratory has previously used the NuGEN amplification kit with success but determined it was worthwhile to compare the effectiveness of the NuGEN kit with the newer Affymetrix WT PLUS kit, which could potentially be better optimized for use with the Affymetrix microarrays typically used. The second objective was to determine which transcript detection method, RNA-seq or Affymetrix GeneChip® Human Transcriptome Array 2.0 (ThermoFisher P/N 902162), was most effective at identifying and quantifying levels of RNA present in samples. The Affymetrix GeneChip Human

Transcriptome Array 2.0 (HTA 2.0) was selected for comparison to RNA-seq analysis as it contains over six million probes specifically recognizing more than 65,000 genes (Affymetrix, n.d.), including non-coding (ncRNA) and regulatory RNA, and represented the most complete coverage of the human transcriptome available in microarray format. Total RNA-seq was chosen as the best comparison to microarrays, as it detects ncRNA and other non-poly-A RNA species. To determine which method of gene expression data collection produced results best suited for biomarker discovery in our hands, we compared DE results from Affymetrix's HTA 2.0 microarrays to total RNA-seq.

Blood-derived RNA was chosen as a sample source for multiple reasons. Blood samples are widely used in many studies for RNA quantification and analysis. One advantage of utilizing blood-derived RNA is that blood samples are easily collected and are minimally invasive, as compared to extracting RNA from other tissues, such as tumors, which are of clinical interest in medical studies but require surgery to obtain samples (Shabihkhani *et al*., 2014). Additionally, protocols to collect blood and stabilize any RNA present are thoroughly documented and simple to utilize. Using well-studied collection methods, such as PAXgene® blood RNA collection tubes and PAXgene's® blood RNA kit, can also minimize risks of processing or handling - induced changes in RNA expression (Feezor *et al*., 2004). Blood also typically contains a wide variety of RNA originating from various sources, including cell-free RNA (Koh *et al.,* 2014) as well as intracellular RNA and is therefore well suited to DE studies. Brain-derived RNA was also used in this research to provide a basis for comparison against blood RNA, based upon the differences in genetic expression between blood and brain (GTEx Consortium 2015).

Analysis of RNA allows researchers to evaluate gene expression of individuals at a given moment in time. Studies using RNA enable evaluation of which genes are being differentially expressed and how gene expression alters due to various stimuli. Specifically, DE analysis allows researchers to study and identify which genes are being over- or under- expressed under a given condition. In the design of this study, which was modeled after that of the MicroArray Quality Control Project (MAQC; Shi *et al.,* 2006), samples were purposefully mixed in defined ratios to ensure that they would demonstrate differential expression when compared, similar to biomarker discovery studies.

Blood-derived RNA and brain-derived RNA were both chosen for this study because they would be expected to produce a sizeable number of DE genes for analysis when comparing the

transcript levels of genes found in both tissues (Melé *et al.,* 2015). To this end, four samples were created containing varying proportions of blood and brain RNA. The samples contained 100% blood RNA/0% brain RNA (all-blood), 67% blood RNA/33% brain RNA (2/3-blood), 33% blood/67% brain RNA (1/3-blood), and 0% blood RNA/100% brain RNA (all-brain). Differential expression comparisons between these samples could reasonably be expected to detect $\log_2$ fold changes of one, as when 2/3-blood RNA expression levels are compared to 1/3-blood RNA, or more, as when all-blood RNA expression levels are compared to all-brain RNA for example.

The four blood/brain samples were amplified using Affymetrix's GeneChip® WT PLUS and using NuGEN's Ovation® Pico WTA system V2. The cDNA produced from these two amplification methods was then hybridized onto Affymetrix HTA 2.0 microarray chips and evaluated to determine which amplification method produces the greatest number of DE genes for use in biomarker discovery studies. The data from the optimum amplification method was then compared to data produced from total RNA-seq analysis of the four blood/brain samples. This contrast was used to determine whether RNA-seq or microarrays identify the most DE genes useful for biomarker studies, as well as taking fold-change of DE genes, quality control, and cost into account in determining the ideal method to use in future studies.

## MATERIALS AND METHODS

### Sample preparation and processing

Total RNA from both blood and brain were used in this study. A brain-derived RNA sample, FirstChoice® Human Brain Reference RNA (1 mg/mL; P/N AM6050), was purchased from Life Technologies™ for use in this study. To obtain blood-derived RNA, blood samples were voluntarily given by FAA employees with informed consent (IRB Protocol No. 10028) and samples were collected using PAXgene® blood RNA tubes (Fisher Scientific, P/N 23-021-01). Blood-derived RNA was extracted using a PAXgene® Blood RNA kit IVD (Qiagen/BD, P/N 762164). The quality of purified blood RNA samples was evaluated by micro-capillary electrophoresis using RNA Nano Chips (Agilent Technologies©, P/N 5067-1521) on an Agilent 2100 Bioanalyzer (Agilent, P/N G2939BA) and RNA concentrations from blood samples were determined at A260 using a NanoDrop™ 2000 (Thermo Scientific™; P/N ND-2000). Such quality measurements for brain RNA were taken from the manufacturer's certificate of analysis

(lot number 105P055201A). Blood RNA samples from three individuals were then pooled, the concentrations were measured, divided into aliquots, and stored at -80ºC. Three of the pooled blood-derived RNA aliquots were concentrated by speed-vac to reach desired concentration and combined with the brain-derived RNA in the following ratios: 100% blood / 0% brain (all-blood), 67% blood / 33% brain (2/3-blood), 33% blood / 67% brain (1/3-blood), and 0% blood / 100% brain (all-brain). The samples were prepared containing a total of 1500 ng in each sample. From these master samples, aliquots were made in duplicate for RNA-seq (500 ng total per blood/brain mixture), Affymetrix amplification (50 ng total per blood/brain mixture), and NuGEN amplification (50 ng total per blood/brain mixture). The samples were stored at -80ºC.

**Microarray preparation and processing**

Affymetrix GeneChip® WT PLUS and NuGEN Ovation® Pico WTA system V2 kits utilize different amplification methods to produce cDNA for downstream hybridization and quantification, so it is possible that these varying methods could affect the quality of the data produced in DE comparisons. Affymetrix's GeneChip® WT PLUS kit utilizes the Eberwine protocol to first produce cDNA from RNA samples, utilize that cDNA to produce amplified RNA by *in vitro* transcription, which in turn, is the substrate to synthesize cDNA for downstream analysis (Van Gelder *et al.,* 1990). NuGEN's Ovation® Pico WTA System V2 uses SPIA technology, an RNA/DNA chimeric mix of primers and reverse transcriptase to generate and amplify cDNA (NuGEN 2016).

Affymetrix Poly-A RNA controls (GeneChip® WT PLUS Reagent Kit, P/N 703147) were added prior to amplification to aliquots for both Affymetrix and NuGEN kits. Samples for Affymetrix and NuGEN amplification each started with 50 ng of RNA in 2.5 µL of sample, an acceptable input RNA quantity for both kits. Amplification for both Affymetrix GeneChip® WT Plus and NuGEN Ovation® Pico WTA System V2 was performed according to manufacturer's instructions. After amplification, sample quality was assessed using an Agilent 2100 Bioanalyzer and concentration was determined using a NanoDrop™ 2000. The cDNA products of these methods were frozen at -20ºC until needed for fragmentation and hybridization onto HTA 2.0 microarray chips. In order to produce enough cDNA for hybridization, the NuGEN amplification had to be performed twice. Samples from the first NuGEN amplification and second amplification were mixed 50/50 by mass of RNA for hybridization. In order to reach required

quantities (ng) of RNA in specified volumes, combined samples were concentrated by speed-vac to 25 µL. Nuclease-free water was added as needed if sample volumes were less than 25 µL after drying.

For both amplification methods, the purified cDNA was fragmented, biotin labeled, and hybridized onto HTA 2.0 Affymetrix microarrays using Affymetrix's GeneChip® Hybridization, Wash, and Stain Kit (P/N 900720). Samples were loaded onto microarrays and incubated for 18 hours at 45ºC rotating at 60 rpm, according to the manufacturer's instructions. All microarrays were stained and washed using a GeneChip® fluidics station 450 (Affymetrix, P/N 00-0079) using protocol FS450-0001, in accordance with the kit manual from Affymetrix. Microarray expression intensities were then read using an Affymetrix 7G GeneChip® Scanner 3000 (P/N 00-00212).

**Microarray data QC**

In R (R Core Team, 2018) using the oligo package (Carvalho & Irizarry, 2010), quality control assessment was performed on individual samples using fitProbeLevelModel() on each .cel file individually. It was also run on all Affymetrix-amplified samples pooled together as well as all NuGEN-amplified samples pooled together. Additional QC evaluation was performed using the R package arrayQualityMetrics (Kauffmann *et al.,* 2009) and command arrayQualityMetrics() to evaluate the raw .cel data for Affymetrix-amplified samples and for NuGEN-amplified samples separately.

**Microarray data filtering**

Primary data analysis for HTA 2.0 microarray expression data was performed in R using oligo and limma (Ritchie *et al.,* 2015) packages. Bioconductor's chip-based package, pd.hta.2.0 (MacDonald, 2017) was used to read in and interpret HTA 2.0 chip information. Raw .cel files were normalized using the rma() command from the oligo package. Filtering was performed using kOverA(), genefilter(), and nsFilter() from the genefilter (Gentleman *et al.,* 2017) package. Two methods of filtering were applied for both the NuGEN- and the Affymetrix- amplified arrays. The first filtering method was based on the normalized expression values of the HTA 2.0 microarray's antigenomic transcript clusters. Expression levels of the antigenomic transcript clusters should be indicative of background noise. For both NuGEN and Affymetrix

preparations, the quantiles of the antigenomic transcript clusters were determined and the normalized "core" datasets were filtered using limma to only retain transcript clusters in which at least one sample had a normalized expression value that was greater than or equal to the third quantile value of the antigenomic transcript clusters.

The second method of filtering transcript clusters utilized the median values for the normalized core transcript clusters. For both Affymetrix and NuGEN filtered samples, the normalized data set was sub-setted, removing any antigenomic or control transcript clusters, and quantile values for the core transcript clusters were determined. Both datasets were filtered to require that at least one sample had a normalized expression value equal to or greater than the median value of the core transcript clusters. Median filtering was performed because there were concerns about comparing transcript cluster results from data sets of notably different sizes, as was seen with antigenomic filtering, where the Affymetrix-amplified antigenomic filtered dataset is roughly half the size of the NuGEN amplified antigenomic filtered dataset.

**Microarray differential expression comparisons**

Using the limma package in R, contrast matrices were constructed for all data sets to compare all-blood v. 2/3-blood, 2/3-blood v. 1/3-blood, 1/3-blood v. all-brain, and all-blood v. all-brain for each amplification and filtering method. The limma command model.matrix() was used to create a design matrix for each comparison, which was then adjusted using lmFit() to create a linear model fit for the data. The command makeContrasts() was used to carry out pair-wise comparisons between sample concentrations. makeContrasts() results were fit using contrasts.fit() and eBayes(). Finally, topTable() performed a T-test comparing the expression levels at varying sample concentration to analyze for DE with significance. Total numbers of differentially expressed transcript clusters, based on adjusted p-value (Benjamini-Hochberg), were found using decideTests() and vennDiagrams(). Results of these t-tests represent the $\log_2$ fold change difference between samples. $\log_2$ fold changes, which were significantly different from each other, determined by adjusted p-values, were utilized for creating volcano plots. R was used to generate additional analyses including principal component analysis using prcomp() and violin plots (Wickham, 2016) of expression value ranges.

**RNA-seq preparation**

Duplicate blood/brain-derived RNA mixtures containing 500 ng of RNA were analyzed by RNA-seq separately. For both sets of samples, library preparation was carried out using TruSeq Stranded Total RNA with Ribo-Zero Globin (Illumina 20020613) to remove globin and ribosomal RNA from the input RNA, shear the RNA, synthesize first- and second-strand cDNA, add indexed sequencing adaptors, and enrich the purified resulting fragments to produce the final dual-indexed library. The quantity of the resulting libraries was determined using the KAPA Library Quantitation kit. Dual indexed libraries were multiplexed and sequenced on a single high output NextSeq 500 flowcell (Illumina) to produce 150 base paired-end reads (2x150).

**RNA-seq data QC and processing**

Initial QC was performed using fastQC (Andrews, 2010) on raw read files. Reads were trimmed using trimmomatic (Bolger *et al*., 2014) requiring paired ends, leading and trailing PHRED scores of at least 30, a minimum length of 100 bp, and using the flag "MAXINFO" to allow the program to automatically balance read length with sequence error rate. The provided adapter file "TruSeq3-PE-2.fa" was used and all other command line settings used with trimmomatic were done according to recommended settings. After trimming, all files were rechecked with fastQC to ensure high PHRED scores throughout reads and adapter removal from sample data.

Read alignment was done using the Rsubread package (Liao *et al*., 2013) in R using UCSC's hg38 human genome assembly (rsync://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/hg38.fa.gz, downloaded on 1 Jun. 2018). Hg38 was used with Rsubread's buildindex() to build the alignment index, which was then used to align all of the RNA-seq samples via the align() command, set to output BAM files allowing for 10 subreads extracted from each read with a consensus threshold for the forward and reverse reads of 3. The command align() was also set to allow a maximum of one mismatch and one indel. The minimum length for reads was set to 100 bp. Reads were aligned both with multi-mapping and without multi-mapping. After alignment, Rsubread's featureCounts() (Liao *et al.,* 2013) was used to generate count tables. The "hg38.fa.gz" annotation index was used with the command featureCounts(), with "useMetaFeatures=TRUE," and designating that the sample is a paired-end sample but not requiring both ends to be mapped to include in the counts table.

Two sets of counts tables were generated, one with and one without multi-mapping of reads permitted.

**RNA-seq differential expression analysis**

After generating feature count tables in R using Rsubread, limma and edgeR (Robinson *et al.,* 2010), Bioconductor packages were used in R to detect significantly differentially expressed genes. Similar to the analysis done for microarray DE analysis, limma and edgeR were used to construct model.matrix() and normalized using calcNormFactors(). Dispersion was estimated and generalized linear modelling was performed using estimateGLMCommonDisp(), estimateGLMTrendedDisp(), estimateGLMTagwiseDisp(), and glmFit(). Contrast matrices were created using makeContrasts() to compare the expression levels of genes between the different blood/brain mixtures. Results were fit to a linear model using glmLRT() and $\log_2$ fold changes of blood/brain contrasts were extracted, along with false discovery rate (FDR) to indicate significance, using topTags(), producing counts of DE genes. Volcano plots were created using genes that were significant by FDR.

**Comparisons between microarray and RNA-seq differential expression**

After determining the optimum amplification and filtering method for microarrays, Affymetrix HTA 2.0 transcript cluster notations had to be converted to Entrez gene IDs to be comparable to RNA-seq results. The conversion between Affymetrix HTA 2.0 transcript cluster IDs to Entrez IDs was a multi-step process. First, "HTA-2_0-na36_hg19-transcript-csv" was downloaded on 4 June 2018 from the ThermoFisher/Affymetrix website. Using this file, a list of HTA 2.0 transcript cluster IDs with their corresponding Entrez ID was created. Problematically, some of the HTA 2.0 transcript cluster IDs were annotated to more than one Entrez ID and were removed from the list. Because microarray transcript clusters that annotated to multiple Entrez IDs had been discarded from further analyses, the decision was made to use the RNA-seq dataset without any multi-mapping reads included, to maintain parity of sample treatment between the two analysis methods.

The list of HTA 2.0 transcript cluster IDs, which annotated to single Entrez IDs, was then used to add Entrez IDs to any transcript cluster IDs that matched in the datasets. Any transcript clusters that did not annotate to an Entrez ID were removed from the data sets. These new

microarray Entrez ID datasets had a further problem to resolve: Any individual transcript cluster ID that annotated to multiple Entrez IDs had been removed, but there were also multiple transcript cluster IDs that annotated to the same Entrez ID. Because each of these transcript clusters could be expected to map to different locations within the Entrez IDs' gene region, it was determined to take the means across values for these transcript clusters and thereby merge all data into a single Entrez ID. This was considered to be equivalent to combining counts of different reads all falling within the same gene in RNA-seq.

Normalized expression values from microarrays cannot be directly compared to normalized count values from RNA-seq; $\log_2$ fold changes, or what the MAQC consortium (MAQC; Shi *et al.,* 2006) referred to as log ratio comparison, can be directly compared. $\text{Log}_2$ fold changes (LFC) from Entrez IDs common to both the optimum amplification and filtering method for microarrays and RNA-seq analysis without multi-mapping genes were collated together and compared. Principal components analysis was performed and violin plots were generated on the overlapping LFC dataset. Comparisons were also made using LFC without limiting datasets to only overlapping Entrez IDs. Violin plots comparing LFC were performed for datasets containing only significant DE changes by FDR/adjusted p-value and on a smaller biomarker subset, containing only significant DE changes by FDR/adjusted p-value and which had an LFC greater than |1.0|. Additionally, comparison of significant DE gene counts between RNA-seq and microarray datasets for each blood/brain mixture comparison were made.

## RESULTS

The overall purpose of our study was to determine the most accurate, efficient, and economical method to collect DE data from RNA samples. Processing time required and data quality were two key factors to determine which methodology to be the optimum choice for our laboratory. Affymetrix-amplified samples were prepared first and produced adequate quality and quantities of cDNA to proceed to hybridization onto HTA 2.0 microarray chips. Overall, the Affymetrix GeneChip® WT PLUS amplification took roughly 16 hours of laboratory time, spread across 3 days, to complete, not including an overnight incubation between the first and second days of the amplification. This is similar to the estimates from the kit manual, which estimates 2 days of laboratory time to complete the procedure and prep samples to proceed to hybridization.

NuGEN's Ovation® Pico WTA System V2 kit manual estimates the required laboratory time to complete amplification at roughly 5 hours, without the final bead purification step. In our laboratory, it required approximately 10 hours of laboratory work to complete, including the bead purification step. The quality of the amplified material produced was adequate, but the first amplification we performed using the NuGEN Ovation® Pico WTA System V2 amplification kit did not produce enough of each sample to proceed to hybridization and had to be repeated. Any time advantage from using the NuGEN amplification kit was negated by having to repeat the amplification.

R was used to perform QC evaluation of the raw and normalized microarray expression data. In all instances, samples passed the "fitProbeLevelModel" QC test performed on the raw .cel files. Outliers were detected by MAplots on raw .cel files from the arrayQualityMetrics reports (Appendix 1:Supplementary Tables 1, 2) for two Affymetrix-amplified samples (2/3-blood A, 1/3-blood A) and for three NuGEN-amplified samples (2/3-blood B, 1/3-blood A, all-brain B). No outliers were detected from boxplots or distances between arrays in arrayQualityMetrics reports. After normalization and log-transformation, however, no outliers were detected for Affymetrix-amplified or NuGEN-amplified samples (Appendix 1:Supplemetary Tables 3, 4). All samples were retained for analysis for two reasons. First, while raw .cel file expression values did produce outliers in one QC test, the technical replicates for each of these samples did not produce outliers, and no other QC metrics indicated outlier results. Second, when expression values were log transformed and normalized, none of the Affymetrix- or NuGEN- amplified samples produced any outliers.

After normalization, datasets were filtered to remove non-informative transcript clusters by two different methods. Filtration was performed at the third quartile value for the "antigenomic" subset of each amplification method (Affymetrix amplified and NuGEN amplified) and at the "median" value of the "core" transcript clusters for each dataset. Antigenomic filtering required at least one sample to have a value of 6.0 or 3.7 for Affymetrix or NuGEN samples respectively. Median filtering required at least one sample to have a value of 3.9 or 3.1 for Affymetrix or NuGEN samples respectively. Antigenomic filtering resulted in the retention of 17,389 of 70,523 transcript clusters for Affymetrix amplified samples and 34,101 of 70,523 transcript clusters for NuGEN-amplified samples. After this filtering step, any remaining control transcript clusters were removed, resulting in final datasets containing 15,386 transcript

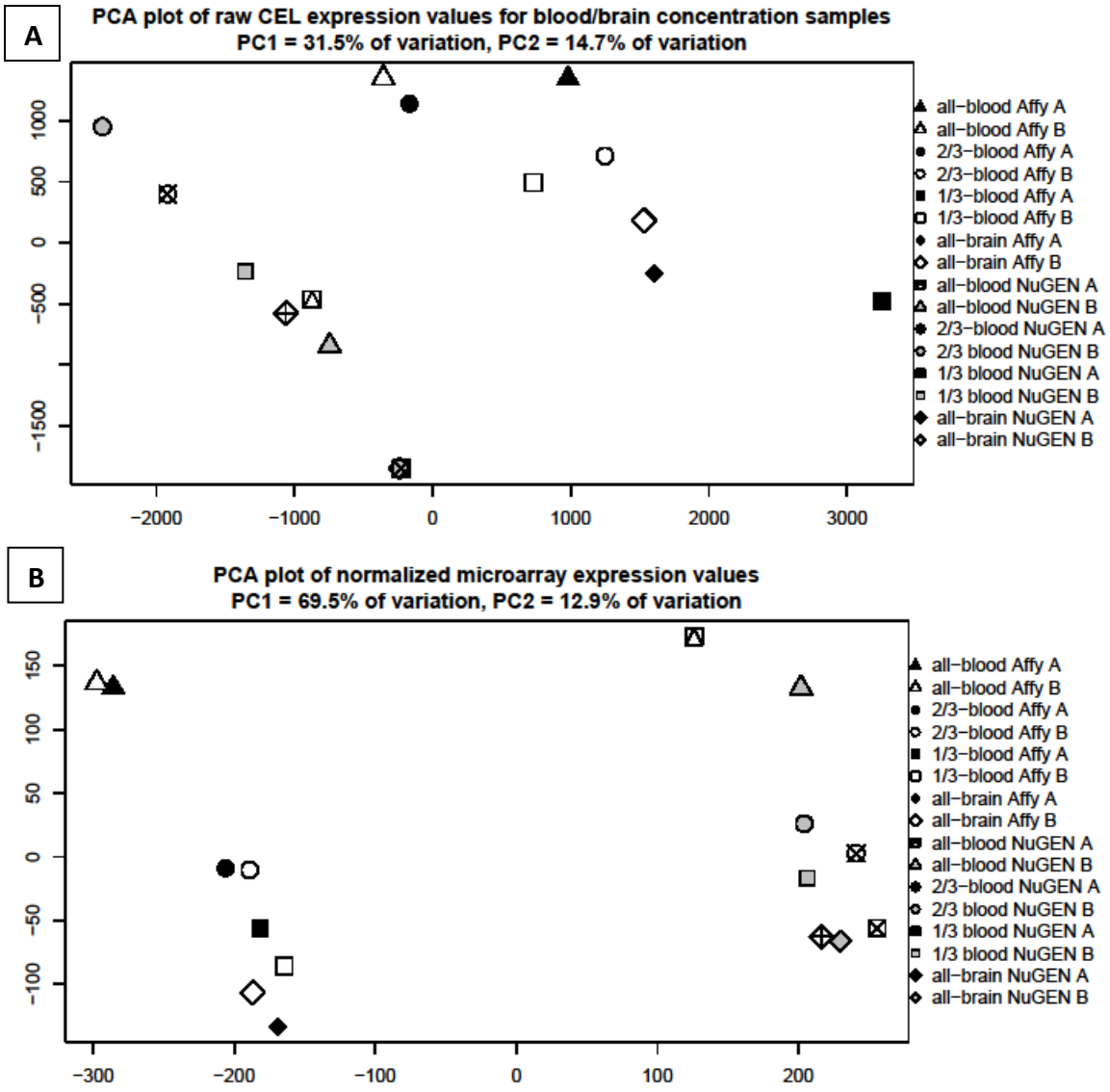
clusters from Affymetrix amplified antigenomic filtered samples and 31,862 transcript clusters from NuGEN amplified antigenomic filtered samples (Table 1). Median filtering was also performed on the normalized Affymetrix and NuGEN datasets and resulted in the retention of 40,004 transcript clusters (30,519 discarded) from the Affymetrix amplification and 41,752 transcript clusters (28,771 discarded) from the NuGEN amplification. Removal of any remaining control / antigenomic transcript clusters resulted in 37,696 transcript clusters remaining (2,308 discarded) in the Affymetrix amplified dataset and 39,384 transcript clusters retained (2,368 discarded) in the NuGEN amplified dataset (Table 1).

**Table 1.** Microarray transcript clusters and filtering results by amplification and filtering method.

| Amplification Method | Transcript Clusters | Filtering Method | Transcript Clusters after Filtering | Transcript Clusters after Filtering and Removal of Controls |
|---|---|---|---|---|
| **Affymetrix** | 70,523 | Antigenomic | 17,389 | 15,386 |
| **Affymetrix** | 70,523 | Median | 40,004 | 37,696 |
| **NuGEN** | 70,523 | Antigenomic | 34,101 | 31,862 |
| **NuGEN** | 70,523 | Median | 41,752 | 39,384 |

Principal components analysis (PCA) of raw .cel expression values demonstrate a distinct separation between the two amplification methods along the first and second principal components, which accounted for 31.5% and 14.7% of variation respectively (Figure 1A). Improved results were seen with PCA after normalization, with preparation method separating principal component 1 (PC1) and sample composition separating the second component. Separation along PC1 from both raw and normalized data indicates that preparation method comprises the main variance component, also indicated by the number of filtered transcript clusters (Table 1). Principal component 2 (PC2) indicates that both sample preparation methods produce a general trend from all-blood through the 2/3-blood and 1/3-blood dilutions to the all-brain samples and that the Affymetrix-prepared replicates are more consistent than the NuGEN-prepared replicates (Figure 1B). PCA of Affymetrix-amplified samples only (Figure 1C) shows separation along PC1 roughly based on blood/brain concentration and accounts for 56.6% of the variation. Distribution of NuGEN-amplified samples in the PCA plot (Figure 1D) demonstrates

separation based on blood/brain concentration along both PC1 and PC2 axes, incorporating 38.8% and 27.1% of the variation respectively.



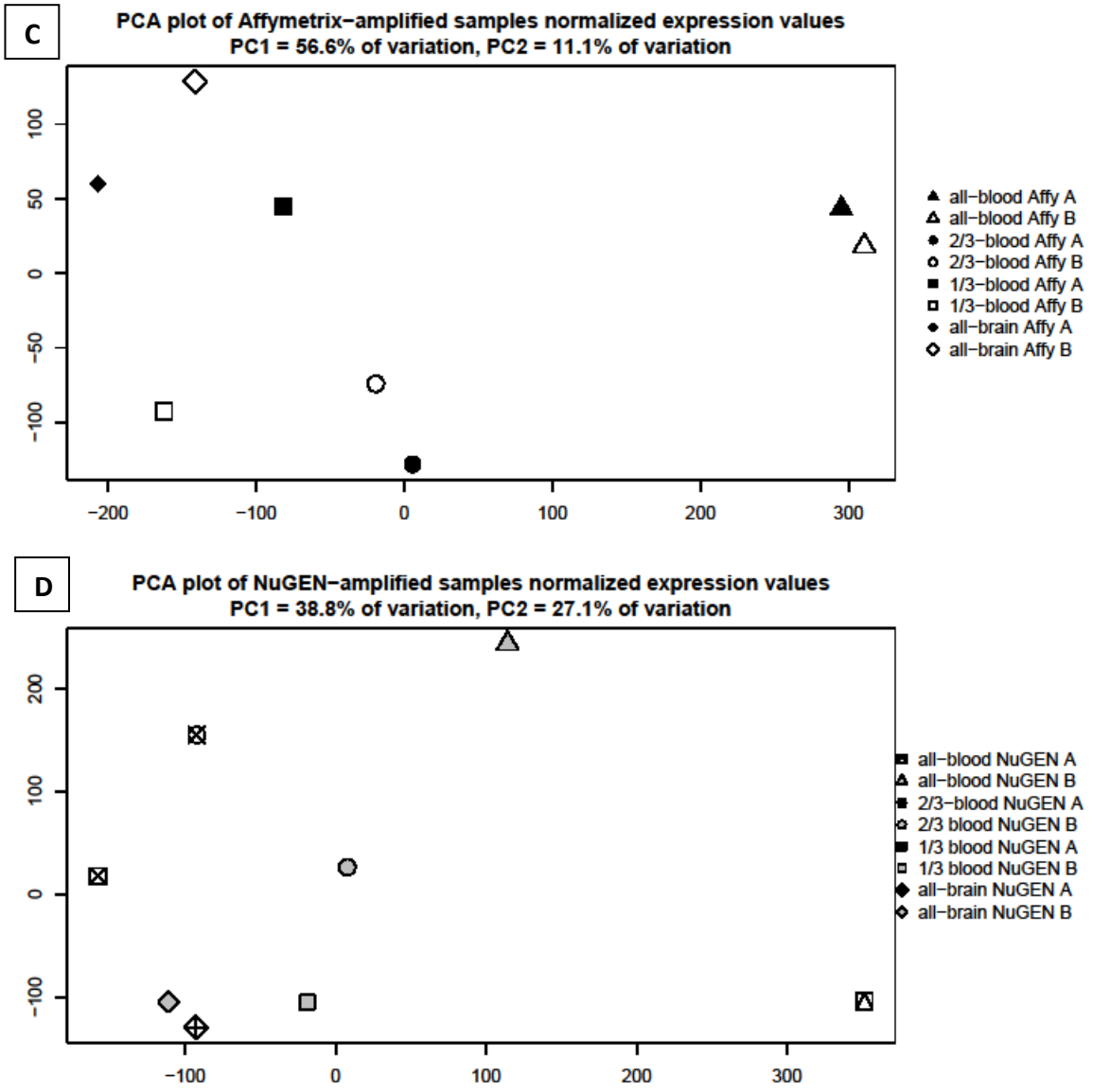



**Figure 1.**

A. PCA plot of raw .cel expression values from NuGEN and Affymetrix amplified blood/brain samples.
B. PCA plot of normalized Affymetrix-amplified and NuGEN-amplified blood/brain mixtures expression values.
C. PCA plot of normalized Affymetrix-amplified expression values for blood/brain mixtures after normalization.
D. PCA plot of normalized NuGEN-amplified expression values for blood/brain mixtures after normalization.

Violin plots display normalized expression value ranges for Affymetrix and NuGEN amplified data sets for both antigenomic (Figure 2) and median (Figure 3) filtering. Affymetrix-amplified samples demonstrate consistently higher normalized expression values for both antigenomic and median filtering methods. Differential expression was determined between blood/brain concentrations and the numbers of significant DE transcript clusters produced were

compared (Table 2). In all cases, Affymetrix amplification produced a greater number of significant DE transcript clusters than NuGEN amplification. Comparisons between antigenomic and median filtering demonstrate an important difference between transcript cluster filtering methods. For Affymetrix-amplified samples, the all-blood vs. 2/3-blood and the all-blood vs. all-brain contrasts had more DE transcript clusters with median filtering, indicating that the more expansive median filtering method used was better suited for lower-expressing transcript clusters. The 2/3-blood vs. 1/3-blood and the 1/3-blood vs. all-brain contrasts resulted in more DE transcript clusters with antigenomic filtering. For the NuGEN-amplified samples, the antigenomic and median filtering produced similar results, with the exception of the 2/3 blood vs. 1/3 blood contrast.

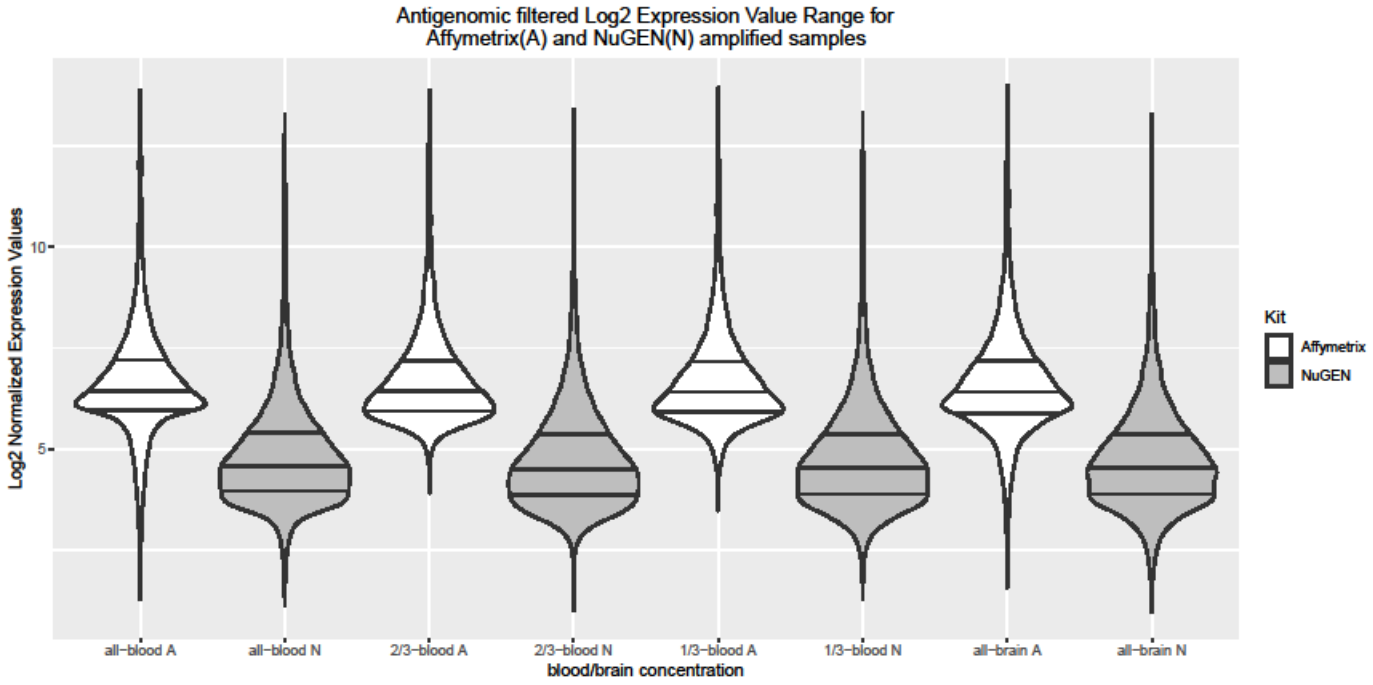



**Figure 2.** Affymetrix (A) and NuGEN (N) amplified antigenomic filtered dataset violin plot showing normalized expression value ranges. The middle line indicates the mean (50$^{th}$ percentile) value in the dataset. The lines above and below the middle are the 75$^{th}$ percentile and the 25$^{th}$ percentile for the dataset and the width indicates the proportion of observations at that percentile.

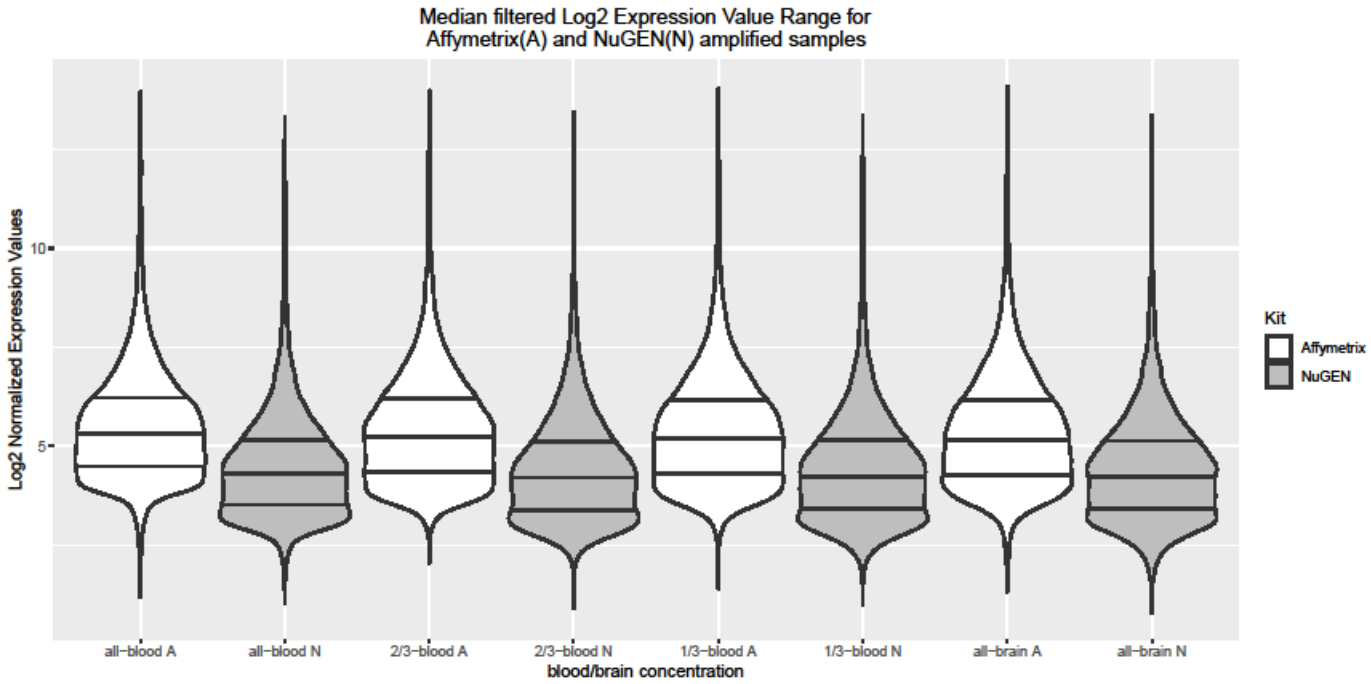



**Figure 3.** Affymetrix (A) and NuGEN (N) amplified median filtered dataset violin plot showing normalized expression value ranges. The middle line indicates the mean (50$^{th}$ percentile) value in the dataset. The lines above and below the middle are the 75$^{th}$ and 25$^{th}$ percentiles for the dataset and the width indicates the proportion of observations at that percentile.

**Table 2.** Tissue comparisons list the number of DE transcript clusters (adjusted p-value <0.05) under varying amplification and filtering methodologies. Total number of transcript clusters for each amplification and filtering method listed in parentheses.

| Tissue | Affymetrix Antigenomic (15,386) | Affymetrix Median (37,696) | NuGEN Antigenomic (31,862) | NuGEN Median (39,384) |
|---|---|---|---|---|
| **all-blood v. 2/3-blood** | 8,080 | 16,989 | 3,103 | 3,088 |
| **2/3-blood v. 1/3-blood** | 3,933 | 3,427 | 0 | 0 |
| **1/3-blood v. all-brain** | 2,427 | 2,118 | 207 | 198 |
| **all-blood v. all-brain** | 11,685 | 25,640 | 10,031 | 10,882 |

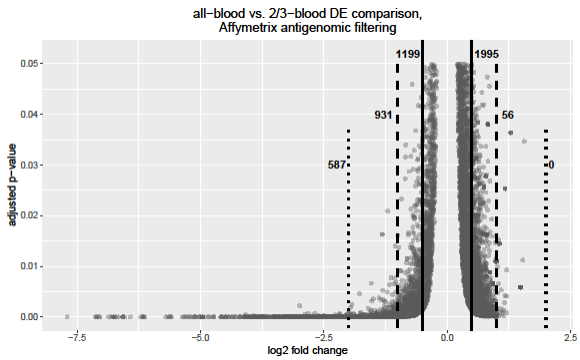
In this study, the 2/3-blood vs. 1/3-blood contrast was considered the most stringent test of differential gene expression. Affymetrix amplification produced over 3,400 DE transcript clusters for each filtering method, while NuGEN detected no significant DE transcript clusters for this comparison from either antigenomic or median filtering (Table 2). Additionally, Affymetrix amplified samples also tended to produce greater LFC differences than NuGEN amplified samples (Table 3). This trend is repeated when comparing antigenomic filtered datasets, even though Affymetrix amplified samples were reduced to roughly half the number of

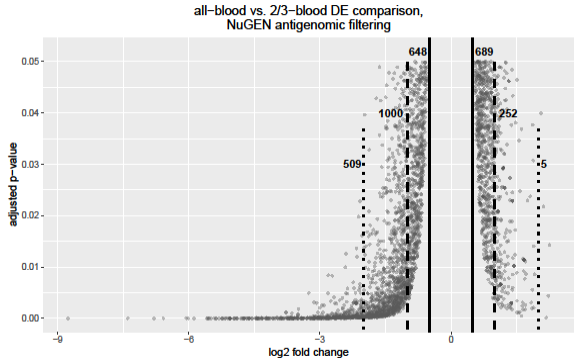
transcript clusters as NuGEN amplified samples (15,386 vs. 31,862 respectively). Volcano plots also demonstrate Affymetrix amplified samples produce many more significant DE transcript clusters than when using NuGEN amplification methods (Figures 4 and 5) regardless of the filtering method used.

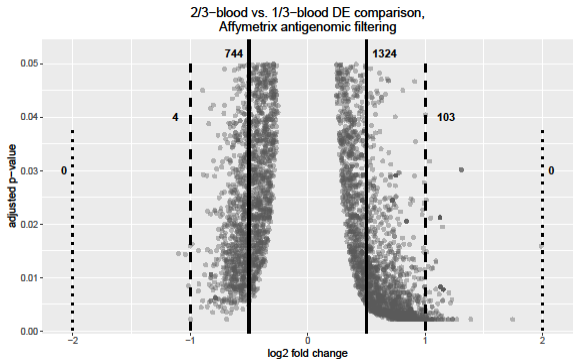
**Table 3.** Comparison of LFC ranges between differing blood/brain concentrations. All transcript clusters (TC) in this count are significant by adjusted p-value (< 0.05).

| Amplification & Filtering Method | Tissue Comparison | < -2.0 LFC | -1.0 to -2.0 LFC | -0.5 to -1.0 LFC | 0.5 to 1.0 LFC | 1.0 to 2.0 LFC | > 2.0 LFC |
|---|---|---|---|---|---|---|---|
| **Affymetrix Antigenomic (15,386 TC)** | **all-blood v. 2/3-blood** | 587 | 931 | 1,199 | 1,995 | 56 | 0 |
| **Affymetrix Antigenomic (15,386 TC)** | **2/3-blood v. 1/3 blood** | 0 | 4 | 744 | 1,324 | 103 | 0 |
| **Affymetrix Antigenomic (15,386 TC)** | **1/3-blood v. all-brain** | 0 | 1 | 42 | 757 | 604 | 183 |
| **Affymetrix Antigenomic (15,386 TC)** | **all-blood v. all-brain** | 1,179 | 1,542 | 1,472 | 2,203 | 1,533 | 1,067 |
| **NuGEN Antigenomic (31,862 TC)** | **all-blood v. 2/3-blood** | 509 | 1,000 | 648 | 689 | 252 | 5 |
| **NuGEN Antigenomic (31,862 TC)** | **2/3-blood v. 1/3 blood** | 0 | 0 | 0 | 0 | 0 | 0 |
| **NuGEN Antigenomic (31,862 TC)** | **1/3-blood v. all-brain** | 0 | 2 | 0 | 3 | 98 | 104 |
| **NuGEN Antigenomic (31,862 TC)** | **all-blood v. all-brain** | 910 | 1,751 | 1,781 | 2,100 | 2,305 | 935 |
| **Affymetrix Median (37,696 TC)** | **all-blood v. 2/3-blood** | 597 | 1,100 | 1,753 | 4,567 | 173 | 20 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Affymetrix Median (37,696 TC)** | **2/3-blood v. 1/3 blood** | 0 | 12 | 684 | 1,546 | 109 | 0 |
| **Affymetrix Median (37,696 TC)** | **1/3-blood v. all-brain** | 0 | 2 | 61 | 872 | 635 | 183 |
| **Affymetrix Median (37,696 TC)** | **all-blood v. all-brain** | 1,299 | 2,276 | 2,662 | 6,770 | 2,608 | 1,268 |
| **NuGEN Median (39,384 TC)** | **all-blood v. 2/3-blood** | 509 | 1,002 | 659 | 661 | 252 | 5 |
| **NuGEN Median (39,384 TC)** | **2/3-blood v. 1/3 blood** | 0 | 0 | 0 | 0 | 0 | 0 |
| **NuGEN Median (39,384 TC)** | **1/3-blood v. all-brain** | 0 | 2 | 0 | 2 | 91 | 103 |
| **NuGEN Median (39,384 TC)** | **all-blood v. all-brain** | 910 | 1,838 | 2,007 | 2,498 | 2,446 | 935 |



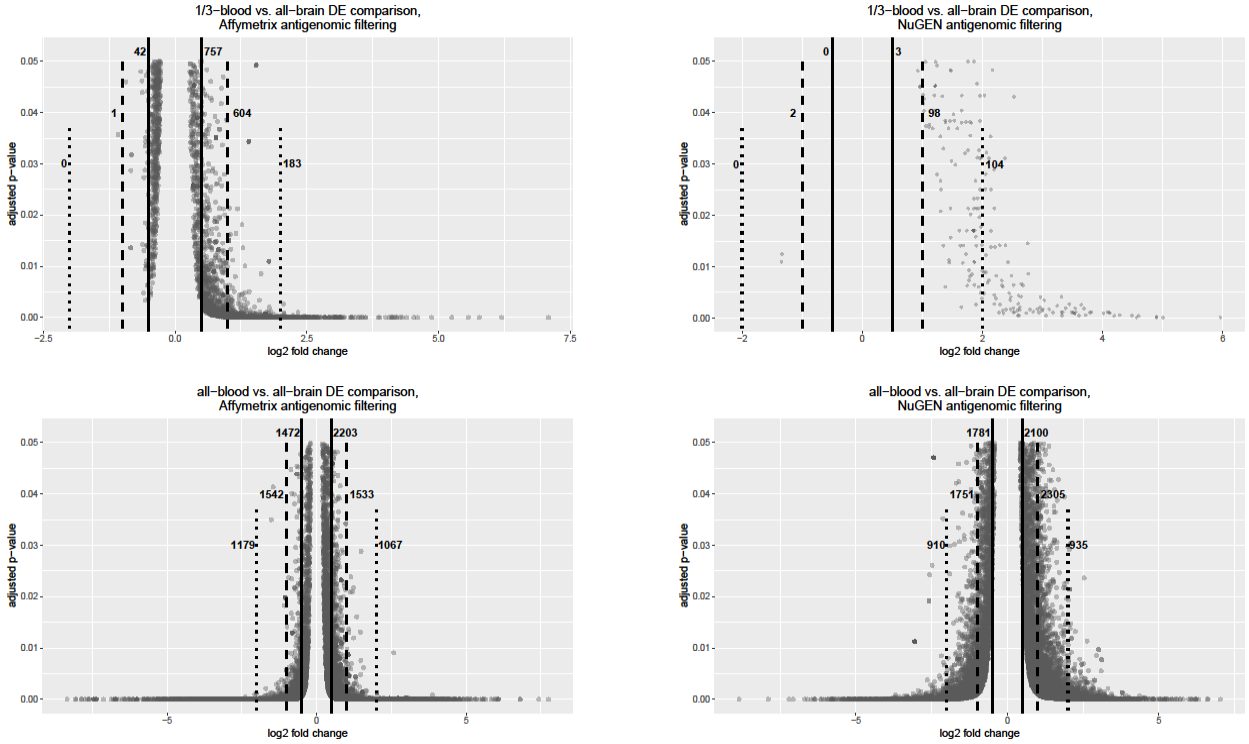
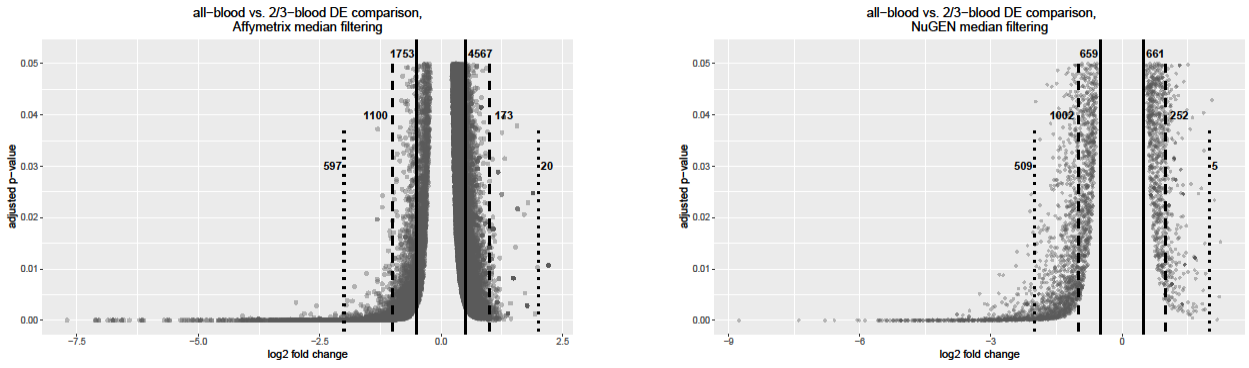



**Figure 4.** Log fold changes are shown for Affymetrix-amplfied and NuGEN-amplified antigenomic filtered samples. The numbers list on each graph, from left to right, are the number comparisons with a LFC < -2, LFC between -2 and -1, LFC between -0.5 and -1, LFC between 0.5 and 1.0, LFC between 1 and 2, and LFC > 2. Dotted lines show the cutoff for LFC > |2.0|. Dashed lines show the cutoff for LFC < |1.0|. Solid lines show the cutoff for LFC < |0.5|. A volcano plot is not presented for 2/3-blood vs. 1/3-blood NuGEN because no DE genes were detected in this comparison.



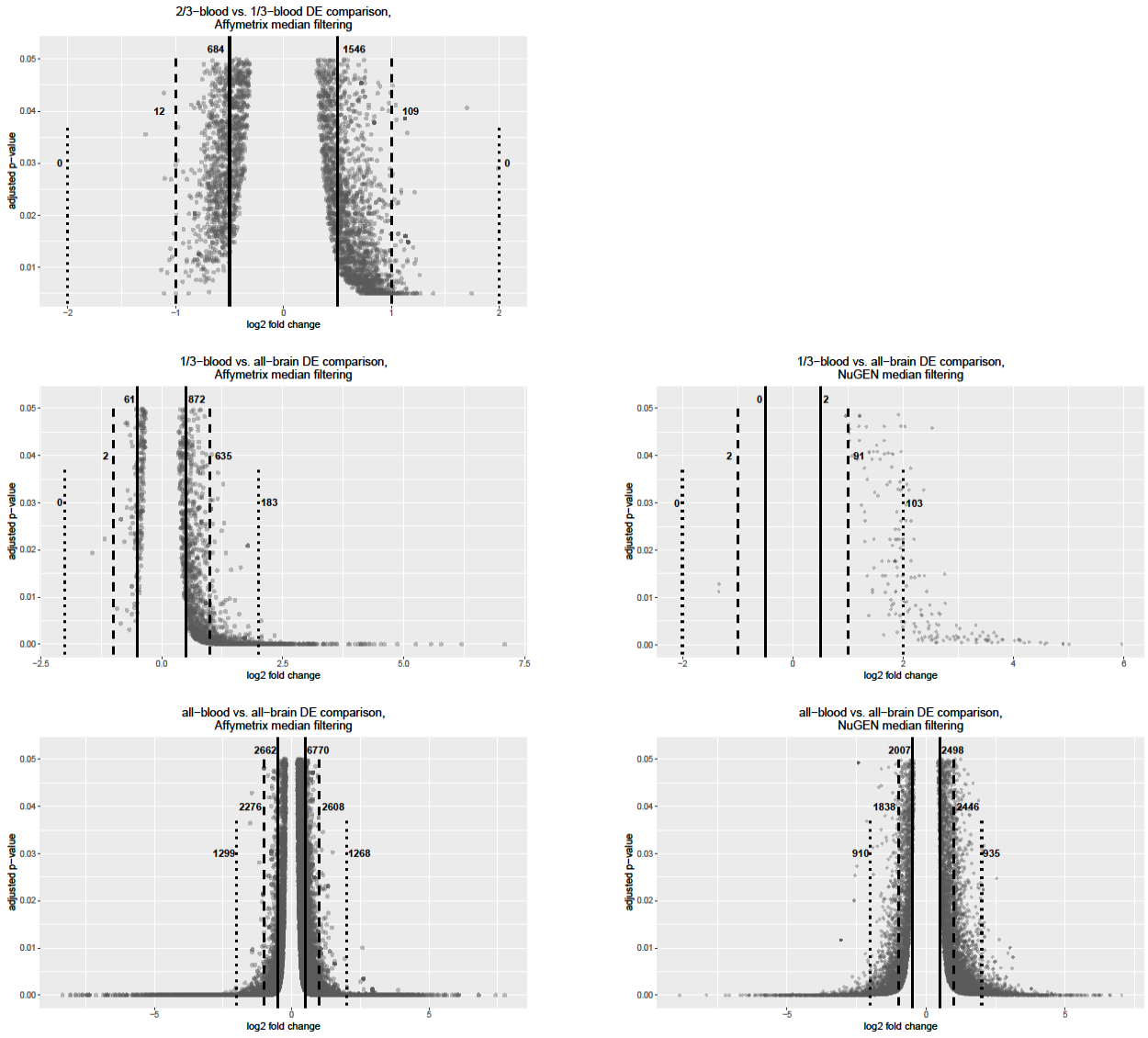


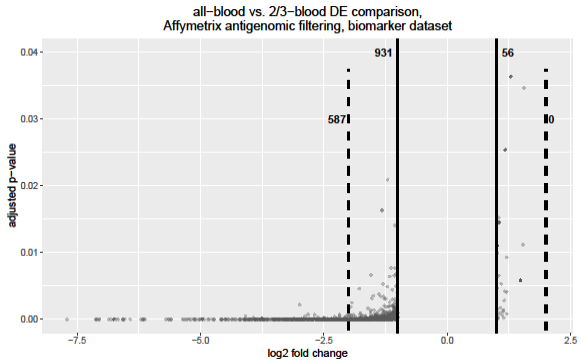

**Figure 5.** Log fold changes are shown for Affymetrix-amplfied and NuGEN-amplified median filtered samples. The numbers list on each graph, from left to right, are the number comparisons with a LFC < -2, LFC between -2 and -1, LFC between -0.5 and -1, LFC between 0.5 and 1.0, LFC between 1 and 2, and LFC > 2. Dotted lines show the cutoff for LFC > |2.0|. Dashed lines show the cutoff for LFC < |1.0|. Solid lines show the cutoff for LFC < |0.5|. A volcano plot is not presented for 2/3-blood vs. 1/3-blood NuGEN because no DE genes were detected in this comparison.

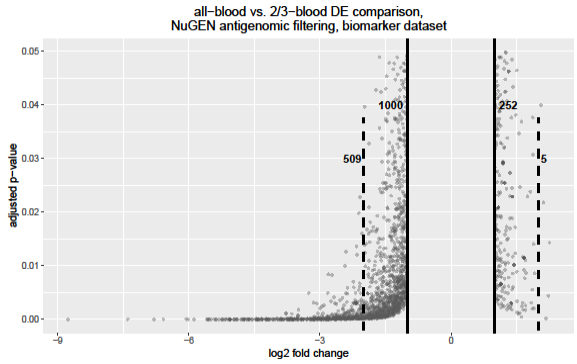
The performance of Affymetrix and NuGEN amplified and filtered data subsets for biomarker discovery was also examined, limited to transcript clusters with a statistically significant (adjusted p-value < 0.05) expression level difference between different concentrations and with a log-fold change greater than |1.0|. Table 4 lists the number of DE biomarker candidates from each amplification and filtration method subset. The results of the Affymetrix and NuGEN amplification comparison are somewhat mixed. NuGEN still produces no DE transcript clusters for the 2/3-blood v. 1/3-blood comparison. However, NuGEN antigenomic and

median both produce more DE transcript clusters than Affymetrix antigenomic for the all-blood v. 2/3-blood comparison and the NuGEN median subset produces more DE transcript clusters than the Affymetrix antigenomic subset for the all-blood v. all-brain comparison. Affymetrix-amplified samples produce more DE transcript clusters at the 1/3-blood v. all-brain comparison for both filtration methods. Volcano plots of the biomarker subsets (Figures 6 and 7) also demonstrate the differences between Affymetrix-amplified and NuGEN-amplified samples. Generally, the numbers of DE transcript clusters between the two filtration methods are similar, despite the different sizes of datasets after filtration. Affymetrix-amplified samples were chosen for further analysis because they were able to detect differentially expressed transcript clusters at all-blood/brain contrasts and, for most of the comparisons performed, Affymetrix amplification produced more DE transcript clusters than NuGEN amplification.

**Table 4.** Numbers of transcript clusters (TC) that could be considered to be potential biomarkers reported above. All reported transcript clusters have an LFC greater than |1.0| and an adjusted p-value of 0.05 or less.

| Tissue | Affymetrix Antigenomic | Affymetrix Median | NuGEN Antigenomic | NuGEN Median |
|---|---|---|---|---|
| **all-blood v. 2/3-blood** | 1,574 | 1,890 | 1,766 | 1,768 |
| **2/3-blood v. 1/3-blood** | 107 | 121 | 0 | 0 |
| **1/3-blood v. all-brain** | 788 | 820 | 204 | 196 |
| **all-blood v. all-brain** | 5,321 | 7,451 | 196 | 6,129 |



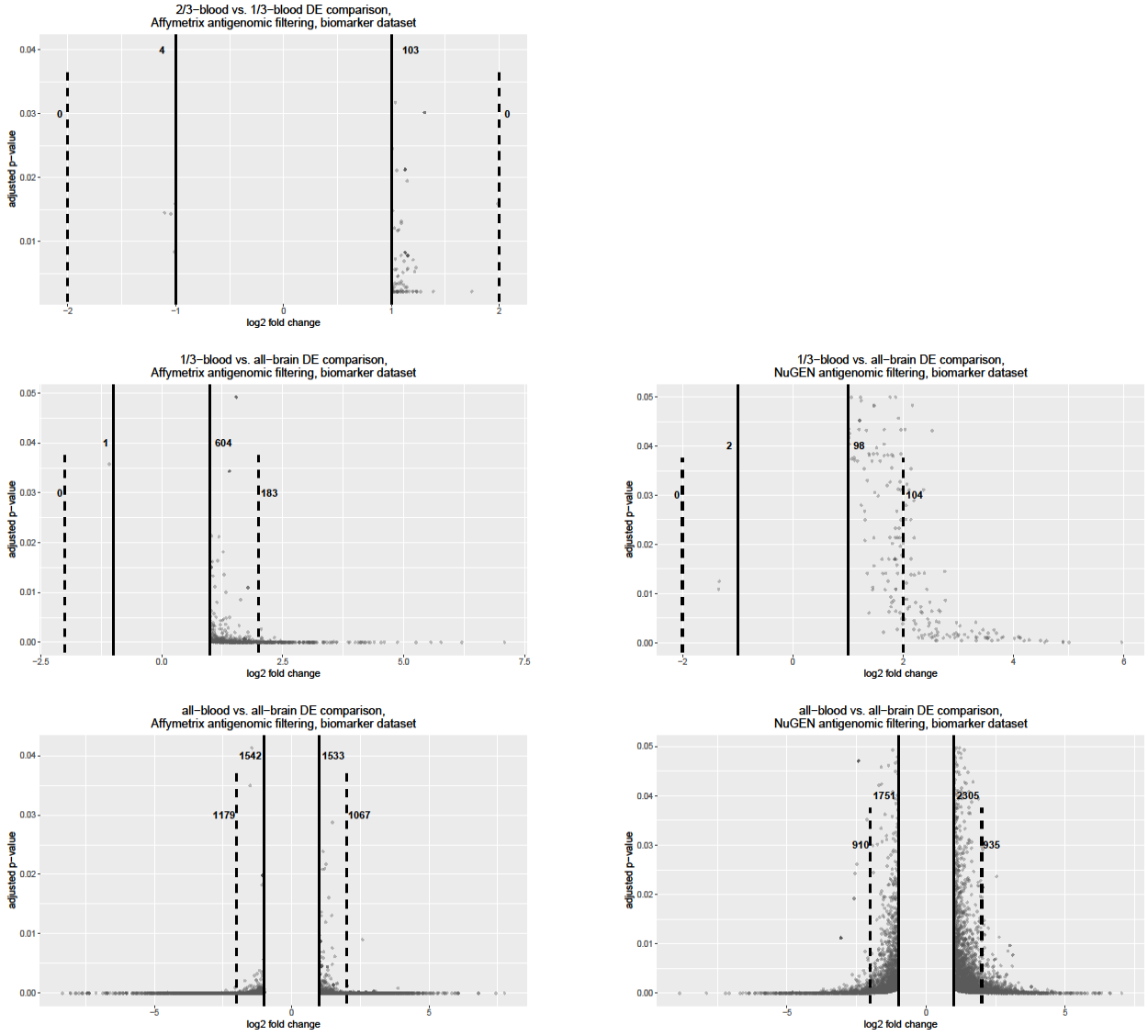



**Figure 6.** Log fold changes are shown for the biomarker subset of Affymetrix-amplfied and NuGEN-amplified antigenomic filtered samples. All samples have an adjusted p-value of < 0.05 and a LFC > |1.0|. The numbers list on each graph, from left to right, are the number comparisons with a LFC < -2, LFC between -2 and -1, LFC between 1 and 2, and LFC > 2. Dashed lines show the cutoff for LFC > |2.0|. Solid lines show the cutoff for LFC < |1.0|. A volcano plot is not presented for 2/3-blood vs. 1/3-blood NuGEN because no DE genes were detected in this comparison.

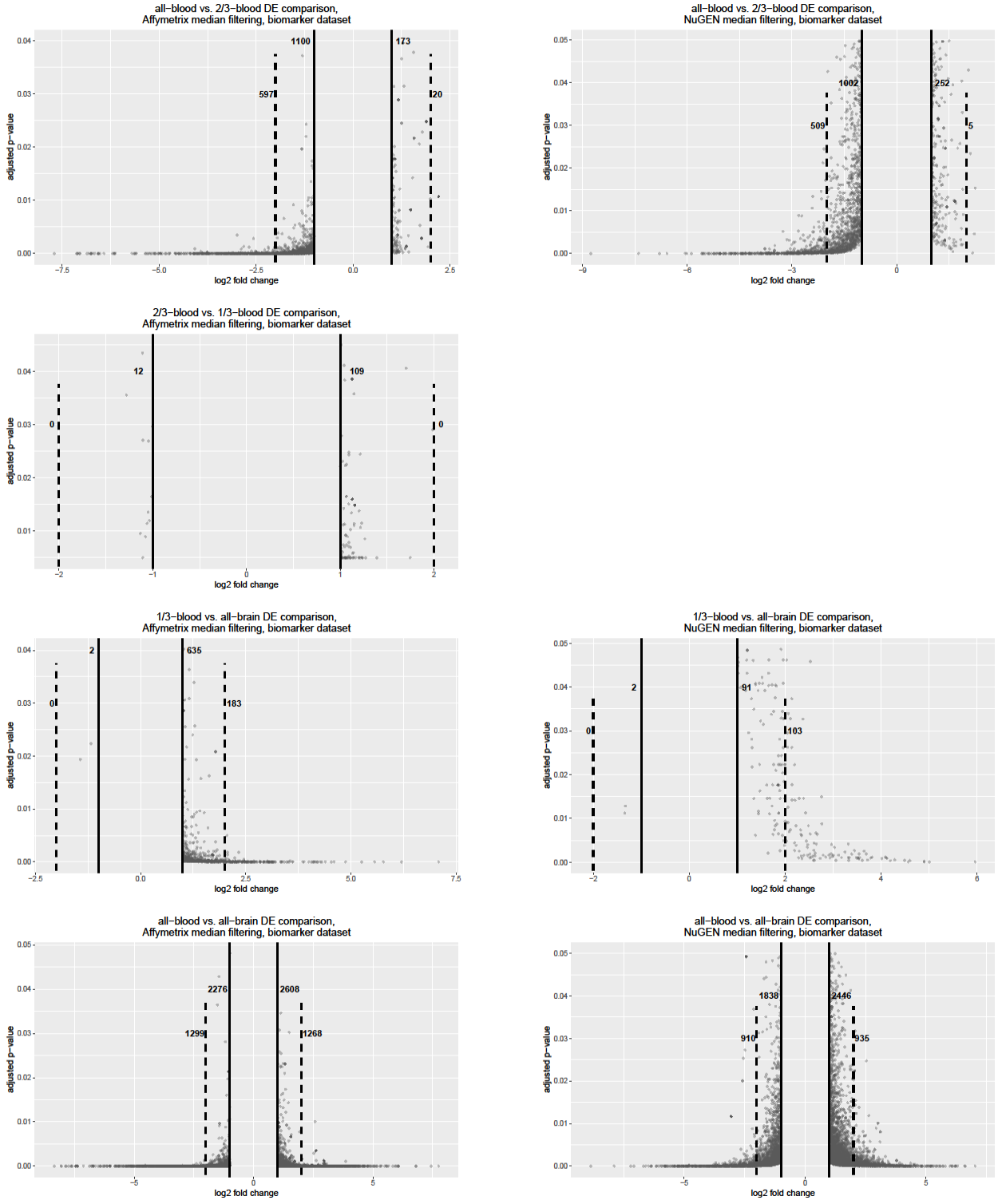



**Figure 7.** Log fold changes are shown for the biomarker subset of Affymetrix-amplfied and NuGEN-amplified median filtered samples. All samples have an adjusted p-value of < 0.05 and a LFC > |1.0|. The numbers list on each graph, from left to right, are the number comparisons with a LFC < -2, LFC between -2 and -1, LFC between 1 and 2, and LFC > 2. Dashed lines show the cutoff for LFC > |2.0|. Solid lines show the cutoff for LFC < |1.0|. A volcano plot is not presented for 2/3-blood vs. 1/3-blood NuGEN because no DE genes were detected in this comparison.

The effectiveness of antigenomic and median filtering methods on Affymetrix sample data were also compared for biomarker detection. When looking at counts of differentially expressed transcript clusters (Table 2), antigenomic filtering produces more DE transcript clusters than median filtering at the 2/3-blood v. 1/3-blood and the 1/3-blood v. all-brain concentrations. Median filtering produced more DE transcript clusters for the all-blood v. 2/3-blood and the all-blood v. all-brain contrasts. Table 4 shows that, for the biomarker subsets, median filtering does produce greater numbers of DE transcript clusters for all-blood/brain contrasts, but that for the most sensitive contrast, 2/3-blood v. 1/3-blood, and also for 1/3-blood v. all-brain, the numbers of DE transcript clusters are very similar. There is considerable overlap between the antigenomic and median filtered biomarker subsets (Table 5). Only one contrast, median all-blood v. all-brain, had less than a 96% overlap with the other filtration method due to the increased number of transcript clusters considered in the median-filtered data set. Table 6 notes the LFC ranges for the biomarker subsets for both filtration methods. The numbers of DE transcript clusters at each LFC range were similar for the 2/3-blood v. 1/3-blood and 1/3-blood v. all-brain comparisons. Median filtering did produce notably more biomarker candidate DE transcript clusters than antigenomic filtration for the all-blood v. 2/3-blood and all-blood v. all-brain comparisons. Therefore, in this study, median filtering of Affymetrix amplified samples will be used for comparison to RNA-seq data.

**Table 5**. The above table lists the percentage of transcript clusters in each biomarker subset (Affymetrix antigenomic or Affymetrix median) that is also found in the other biomarker subset.

| Tissue Contrast | Percentage of Affymetrix antigenomic transcript clusters also found in Affymetrix median subset | Percentage of Affymetrix median transcript clustes also found in Affymetrix antigenomic subset |
|---|---|---|
| **all-blood v. 2/3-blood** | 100.00% | 96.84% |
| **2/3-blood v. 1/3-blood** | 99.95% | 99.81% |
| **1/3-blood v. all-brain** | 99.91% | 99.59% |
| **all-blood v. all-brain** | 100.00% | 78.69% |

**Table 6.** LFC ranges for Affymetrix antigenomic and median biomarker subsets, in which adjusted p-value < 0.05 and LFC > |1|.

| Filtering Method | Tissue Contrast | < -2 LFC | < -1, > -2 LFC | > 1, < 2 LFC | > 2 LFC | Total |
|---|---|---|---|---|---|---|
| **Antigenomic** | **all-blood v. 2/3-blood** | 587 | 931 | 56 | 0 | 1,574 |

| | | | | | | |
|---|---|---|---|---|---|---|
| **Antigenomic** | **2/3-blood v. 1/3-blood** | 0 | 4 | 103 | 0 | 107 |
| **Antigenomic** | **1/3-blood v. all-brain** | 0 | 1 | 604 | 183 | 788 |
| **Antigenomic** | **all-blood v. all-brain** | 1,179 | 1,542 | 1,533 | 1,067 | 5,321 |
| **Median** | **all-blood v. 2/3-blood** | 597 | 1,100 | 173 | 20 | 1,890 |
| **Median** | **2/3-blood v. 1/3-blood** | 0 | 12 | 109 | 0 | 121 |
| **Median** | **1/3-blood v. all-brain** | 0 | 2 | 635 | 183 | 820 |
| **Median** | **all-blood v. all-brain** | 1,299 | 2,276 | 2,608 | 1,268 | 7.451 |

Total RNA-seq produced between 55.3 and 88.4 million reads for the blood/brain samples, as shown in Table 7. Initial quality control was performed using fastQC and did demonstrate the need to trim reads for quality. After using trimmomatic, the number of reads ranged from 40.3 to 76.3 million and fastQC was run again to verify the results of trimming. Reads were found to have adapters removed and have median PHRED scores of 30 or greater across all reads. Reads were then mapped to the human genome (hg38), resulting in the retention of between 34.3 and 72.5 million reads (85.2% to 89.5%). Mapping to annotated genes (hg38) resulted in a decline in the number of reads being retained, with between 13.7 and 36.9 million reads being annotated within annotated features of the hg38 genome. Finally, numbers of reads were totaled by Entrez ID using featureCounts (RSubread) and resulted in 17,979 Entrez IDs without multi-mapping and 18,374 Entrez IDs with multi-mapping (Table 7).

**Table 7.** Read counts for raw RNA-seq data, after trimming, mapping, annotation, and featureCount totals per Entrez ID.

| | **all-blood A** | **2/3-blood A** | **1/3-blood A** | **all-brain A** | **all-blood B** | **2/3-blood B** | **1/3-blood B** | **all-blood B** |
|---|---|---|---|---|---|---|---|---|
| **Raw Reads** | 56.3 M | 59.7 M | 59.0 M | 55.3 M | 88.4 M | 87.3 M | 73.7 M | 78.9 M |
| **Reads after trimming** | 40.3 M | 43.0 M | 44.4 M | 40.5 M | 76.3 M | 74.1 M | 62.6 M | 67.1 M |
| **Reads mapped (no multi-mapping** | 34.3 M | 37.0 M | 38.6 M | 34.5 M | 67.2 M | 65.7 M | 55.4 M | 60.1 M |
| **% Reads mapped (no multi-mapping)** | 85.2% | 85.9% | 86.8% | 85.3% | 88.1% | 88.6% | 88.5% | 89.5% |
| **Reads mapped (with multi-mapping)** | 37.7 M | 40.7 M | 42.2 M | 37.9 M | 72.5 M | 70.7 M | 60.0 M | 64.3 M |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **% Reads mapped (with multi-mapping)** | 93.5% | 94.7% | 94.9% | 93.6% | 95.1% | 95.3% | 95.3% | 95.8% |
| **Reads uniquely mapped (both)** | 34.3 M | 37.0 M | 38.6 M | 34.5 M | 67.2 M | 65.7 M | 55.4 M | 60.1 M |
| **Reads multi-mapped (no multi-mapping)** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Reads multi-mapped (with multi-mapping)** | 3.3 M | 3.8 M | 3.6 M | 3.3 M | 5.4 M | 5.0 M | 4.3 M | 4.3 M |
| **Reads not mapped (no multi-mapping)** | 6.0 M | 6.1 M | 5.9 M | 6.0 M | 9.1 M | 8.4 M | 7.2 M | 7.1 M |
| **Reads no mapped (with multi-mapping)** | 2.6 M | 2.3 M | 2.3 M | 2.6 M | 3.7 M | 3.5 M | 2.9 M | 2.8 M |
| **Assigned fragments (no multi-mapping)** | 13.7 M | 17.7 M | 20.5 M | 20.0 M | 27.2 M | 31.3 M | 29.8 M | 34.8 M |
| **% Assigned fragments** | 34.0% | 41.1% | 46.0% | 49.5% | 35.6% | 42.2% | 47.7% | 51.9% |
| **Assigned fragments (with multi-mapping)** | 14.9 M | 19.1 M | 21.8 M | 21.4 M | 29.3 M | 33.4 M | 31.7 M | 36.9 M |
| **% Assigned fragments (with multi-mapping)** | 36.9% | 44.3% | 49.0% | 52.9% | 38.4% | 45.1% | 50.7% | 54.9% |
| **Entrez IDs with fragments assigned (both)** | 28,395 | 28,395 | 28,395 | 28,395 | 28,395 | 28,395 | 28,395 | 28,395 |
| **Entrez IDs after normalization (no multi-mapping)** | 17,979 | 17,979 | 17,979 | 17,979 | 17,979 | 17,979 | 17,979 | 17,979 |
| **Entrez IDs after normalization (with multi-mapping)** | 18,374 | 18,374 | 18,374 | 18,374 | 18,374 | 18,374 | 18,374 | 18,374 |

In order to directly compare the results of microarray to RNA-seq methodology, Affymetrix HTA 2.0 transcript cluster IDs needed to be converted to Entrez IDs, which were

used in RNA-seq annotation. The Affymetrix median filtered dataset started with 37,698 transcript clusters. After converting these to Entrez IDs, we were left with 18,385 microarray Entrez IDs. There are multiple reasons why the number of Entrez IDs for the microarrays is so much smaller than the number of Affymetrix transcript cluster IDs. Roughly one-third of the Affymetrix transcript cluster IDs did not annotate to any Entrez ID and were excluded. There were also a large number of Affymetrix transcript cluster IDs that annotated to more than one Entrez ID for an individual transcript cluster and were also excluded. Similarly, any RNA-seq reads that annotated to more than one location in the genome were similarly excluded (no multi-mapping). Lastly, there were a number of different Affymetrix transcript cluster IDs that annotated to a single Entrez ID. In this case, the mean of LFC across these transcript clusters was taken so that a single value could be used. By using the Affymetrix median data set, after conversion to Entrez IDs, we were left with a data set with 18,385 Entrez IDs, which was similar in size to the RNA-seq dataset with 17,979 Entrez IDs. In comparison, the Affymetrix antigenomic filtered dataset, after conversion to Entrez IDs, contained only 8,963 Entrez IDs.

Statistically speaking, the normalized $\log_2$ expression values between the Affymetrix median dataset and the RNA-seq data set should not be compared directly, but their LFC can be compared (MAQC; Shi *et al.,* 2006). To compare LFC, an overlapping subset of Entrez IDs from the non-multimapping RNA-seq dataset and the Affymetrix-amplified median filtered dataset was found. There were 14,552 Entrez IDs common to both datasets. This combined dataset was used to perform a principal components analysis (Figure 8), which indicates clear differences within and between microarray and RNA-seq data. The overlap of the two datasets was also used to explore the range of LFC observed in the RNA-seq and Affymetrix median datasets through violin plots (Figure 9). RNA-seq demonstrates notably greater LFC ranges than the Affymetrix median microarray dataset in each comparison.

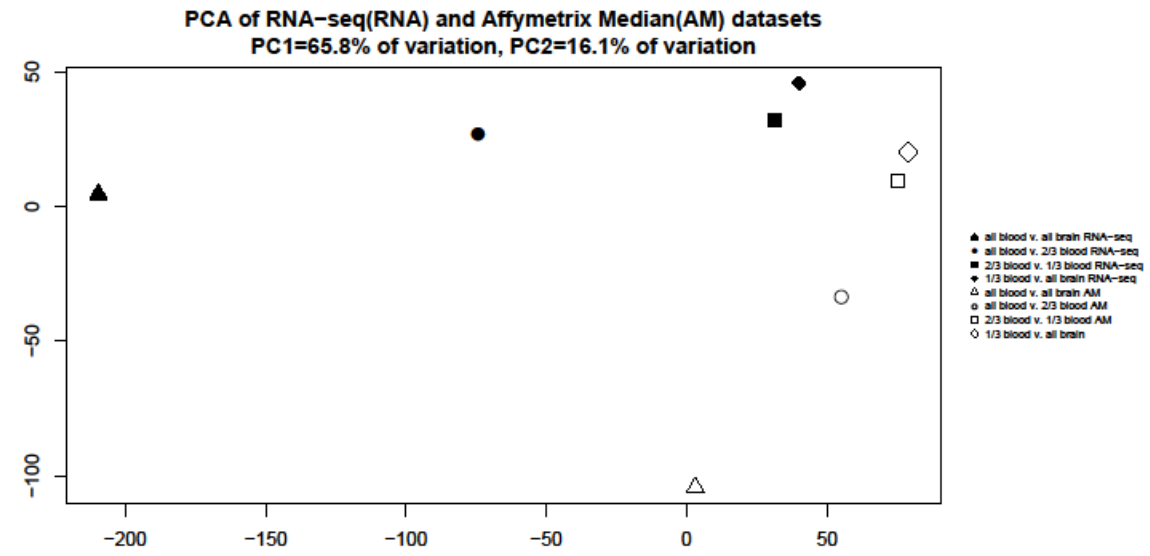



**Figure 8.** PCA plot of RNA-seq and Affymetrix-amplified median filtered LFC for entrez ID's common to both datasets.

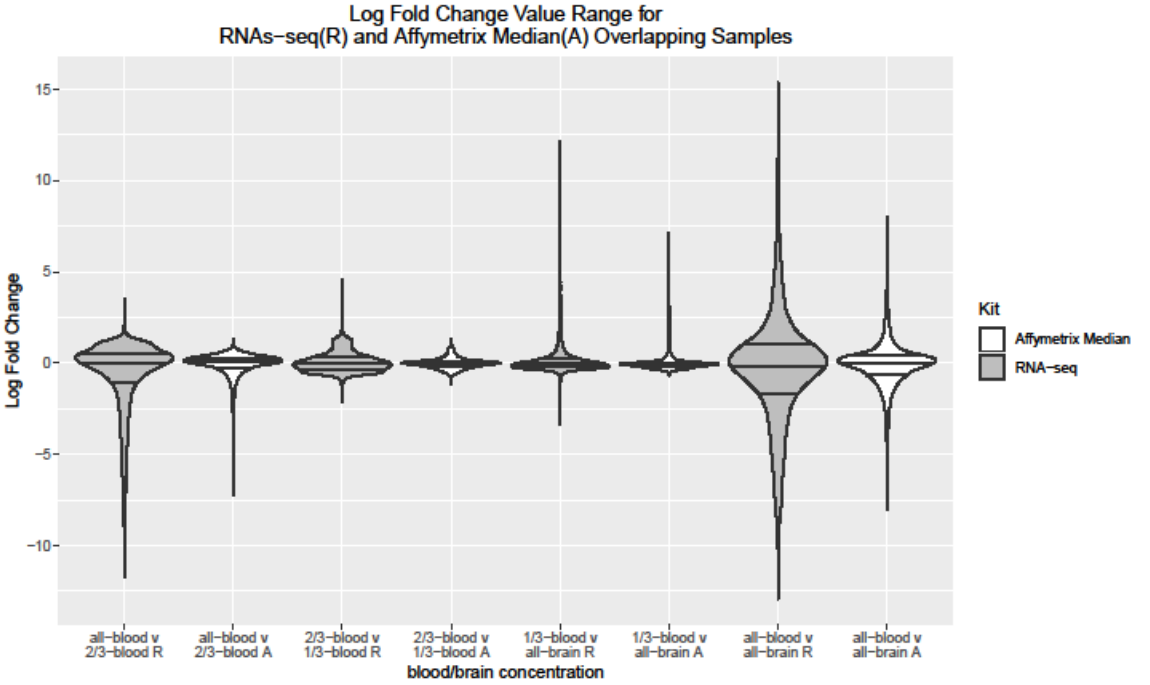



**Figure 9.** Violin plot of LFC ranges for entrez IDs common to both RNA-seq® and Affymetrix-amplified median(A) filtered datasets. The middle line indicates the mean (50$^{th}$ percentile) value in the dataset. The lines above and below the middle are the 75$^{th}$ percentile and the 25$^{th}$ percentile for the dataset and the width indicates the proportion of observations at that percentile.

$Log_2$ fold change datasets not limited to overlapping Entrez IDs were also compared and demonstrated the same trend. Violin plots (Figure 10) of the significant transcript clusters/Entrez IDs (adjusted p-value or FDR <0.05) continue to indicate that the RNA-seq dataset displays a

greater LFC range than the Affymetrix median datasets. Finally, a violin plot of the biomarker subset, containing only those transcript clusters/Entrez IDs that are significant (adjusted p-value or FDR <0.05) and have an LFC > |1|, display the differences in LFC ranges observed (Figure 11). Volcano plots of the biomarker subsets are shown in Figure 12. Transcript cluster/Entrez ID counts of LFC ranges are shown in Table 8. All indicate that RNA-seq identifies more DE genes and more LFC greater than |2.0| in comparison to the Affymetrix-amplified median filtered dataset.

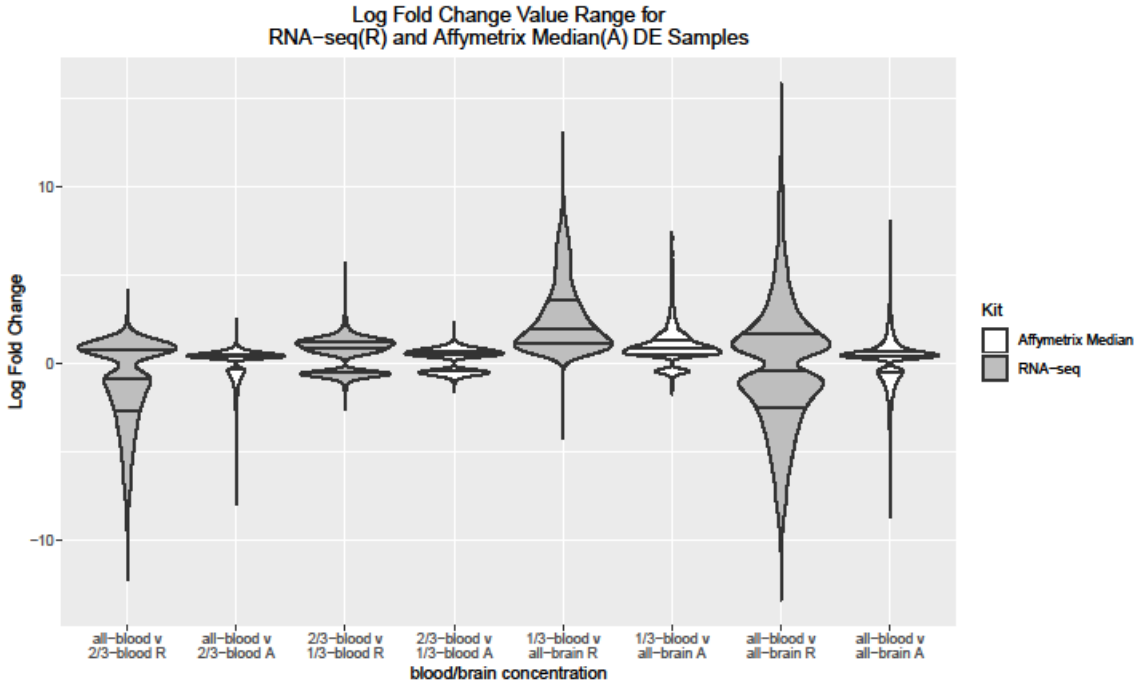



**Figure 10**. Violin plot of LFC ranges for RNA-seq(R) and Affymetrix-amplified median(A) filtered datasets. Datasets are limited to transcript clusters/entrez IDs which were significant by adjusted p-value < 0.05. The middle line indicates the mean (50$^{th}$ percentile) value in the dataset. The lines above and below the middle are the 75$^{th}$ percentile and the 25$^{th}$ percentile for the dataset and the width indicates the proportion of observations at that percentile.

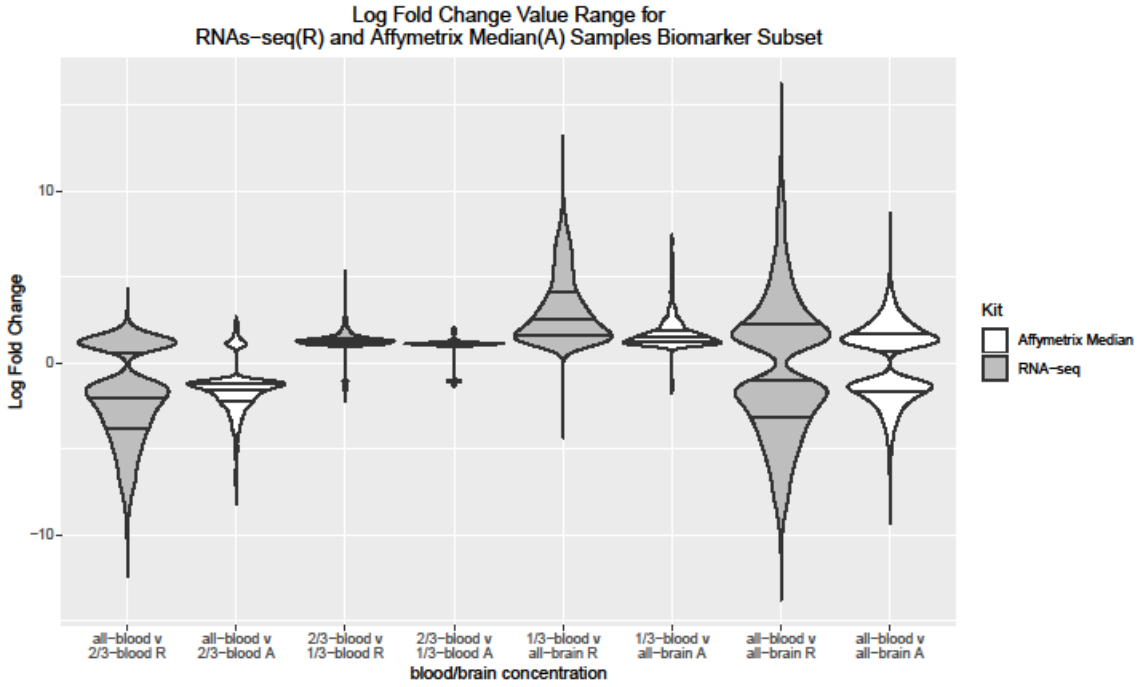
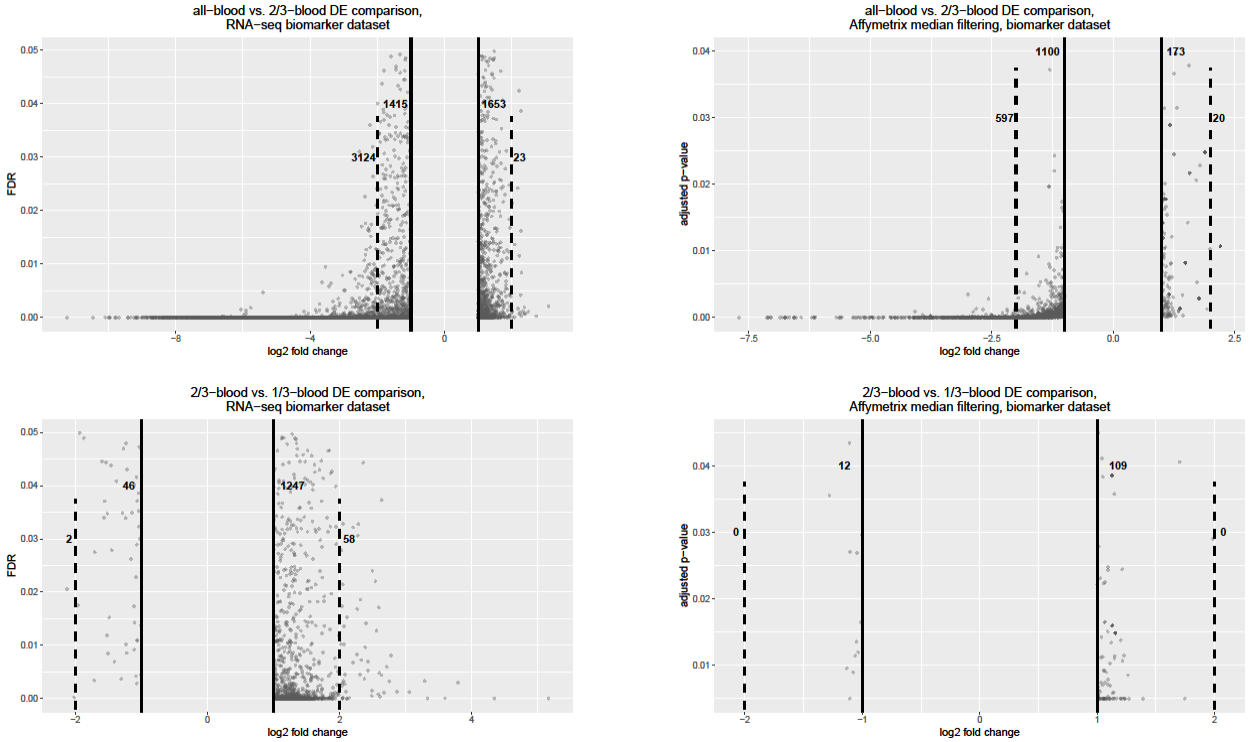



**Figure 11**. Violin plot of LFC ranges for RNA-seq(R) and Affymetrix-amplified median(A) filtered data subsets. Data subsets are limited to transcript clusters/entrez IDs which have a LFC > |1.)| and an adjusted p-value < 0.05. The middle line indicates the mean (50$^{th}$ percentile) value in the dataset. The lines above and below the middle are the 75$^{th}$ percentile and the 25$^{th}$ percentile for the dataset and the width indicates the proportion of observations at that percentile.



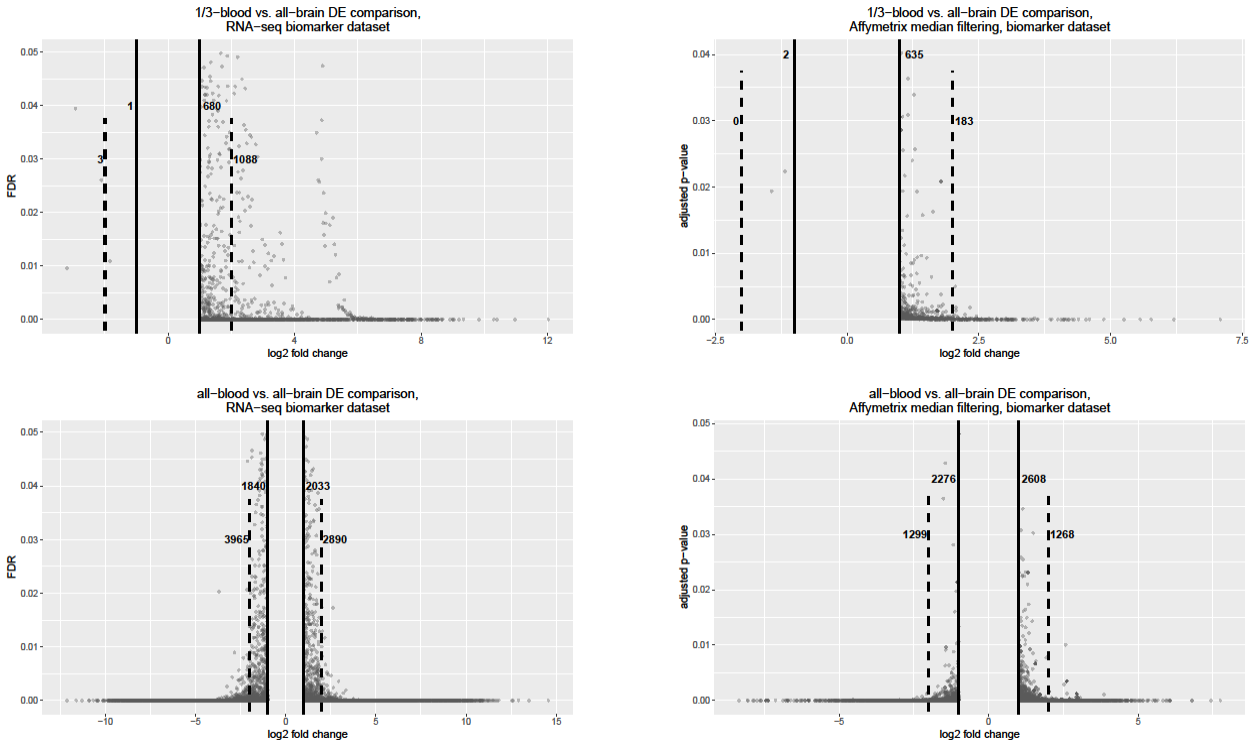



**Figure 12.** Log fold changes are shown for the biomarker subset of RNA-seq and Affymetrix-amplfied median filtered samples. All samples have an FDR/adjusted p-value of < 0.05 and a LFC > |1.0|. The numbers list on each graph, from left to right, are the number comparisons with a LFC < -2, LFC between -2 and -1, LFC between 1 and 2, and LFC > 2. Dashed lines show the cutoff for LFC > |2.0|. Solid lines show the cutoff for LFC < |1.0|.

**Table 8.** LFC ranges for Affymetrix median and RNA-seq biomarker subsets, in which adjusted p-value/FDR < 0.05 and LFC > |1|.

| Sample Preparation Method | Tissue Comparison | < -2.0 LFC | -1.0 to -2.0 LFC | 1.0 to 2.0 LFC | > 2.0 LFC | Total |
|---|---|---|---|---|---|---|
| **RNA-seq** | **all-blood v. 2/3-blood** | 3,124 | 1,415 | 1,653 | 23 | 6,215 |
| **RNA-seq** | **2/3-blood v. 1/3-blood** | 2 | 46 | 1,247 | 58 | 1,353 |
| **RNA-seq** | **1/3-blood v. all-brain** | 3 | 1 | 680 | 1,088 | 1,772 |
| **RNA-seq** | **all-blood v. all-brain** | 3,965 | 1,840 | 2,033 | 2,890 | 10,728 |
| **Affymetrix amplified, median filtered** | **all-blood v. 2/3-blood** | 597 | 1,100 | 173 | 20 | 1,890 |
| **Affymetrix amplified, median filtered** | **2/3-blood v. 1/3-blood** | 0 | 12 | 109 | 0 | 121 |
| **Affymetrix amplified, median filtered** | **1/3-blood v. all-brain** | 0 | 2 | 635 | 183 | 820 |
| **Affymetrix amplified, median filtered** | **all-blood v. all-brain** | 1,299 | 2,276 | 2,608 | 1,268 | 7,451 |

Cost and processing time were also considerations in determining which methodology to use in processing and analyzing samples. When comparing the Affymetrix GeneChip® WT Plus amplification kit to the NuGEN Ovation® Pico WTA System V2 amplification kit, it was initially expected, based on the kit manuals, for the NuGEN amplification to be the more rapid of the two; however, the total preparation time of each method was largely equivalent. Because the NuGEN amplification had to be repeated, the actual costs to use NuGEN in this instance were even higher and no time was saved. Table 9 lists the costs associated with Affymetrix and NuGEN amplification. NuGEN amplification has a significantly higher cost, roughly 2.5 times that of the Affymetrix kit per sample.

In order to perform the microarray hybridization, an additional one and a half laboratory days were needed. The cost of the hybridization was $3,500.00 for ten samples (Table 9). To compare to the costs for RNA-seq processing, an average was taken from multiple RNA-seq processing facilities in the U. S., with similar requirements to those used for our sample and found an average of $6,400 to process ten RNA-seq samples under the same conditions used in this study. In comparison, to perform Affymetrix amplification and microarray hybridization, the cost would be around $4,500 and an additional three and a half workdays in the laboratory.

**Table 9**. Associated costs for amplification methods and hybridization used as well as an average estimate of RNA-seq processing costs.

| Amplification | Item | Reference # | Kit Quantity | Cost | Total per method: |
|---|---|---|---|---|---|
| Affymetrix | GeneChip ® WT Plust Reagent Kit | 902280 | 10 reactions | $1014.00 | $1014.00 |
| NuGEN | Eukaryotic Poly A RNA | 900433 | 100 reactions | $306.00 | |
| NuGEN | Ovation Pico ® WTA System V2 | 3302-12 | 12 reactions | $1038.40 | |
| NuGEN | Encore Biotin Module | 4200-12 | 12 reactions | $692.00 | |
| NuGEN | Genechip ® Oligo B | | | $295.00 | |
| NuGEN | Beckman Agencourt RNAClean XP Beads | 41105518 | | $700.00 | $3031.40 |
| Microarray Hybridization | GeneChip Hybridization, Wash, and Stain Kit | 900720 | 30 reactions | $528.00 | |
| Microarray Hybridization | GeneChip HTA 2.0 Array, 10 pk | 902309 | 10 arrays | $2950.00 | $3478.00 |
| Total cost for Affymetrix amplification and hybridization of 10 samples | | | | | $4492.00 |
| Total cost for NuGEN amplification (2 kits) and hybridization of 10 samples | | | | | $9540.80 |
| RNA-seq estimate | Lab A | | | $4950.00 | |
| RNA-seq estimate | Lab B | | | $6520.50 | |
| RNA-seq estimate | Lab C | | | $6832.00 | |
| RNA-seq estimate | Lab D | | | $7665.00 | |
| RNA-seq estimate | Lab E | | | $6224.25 | $6438.35 (average for 10 samples |

## DISCUSSION

The first goal of this study was to determine which amplification method, Affymetrix GeneChip® WT Plus or NuGEN Ovation® Pico WTA System V2, produced greater numbers of DE genes for biomarker detection using microarrays. PCA plots were expected to show technical replicates plotted near each other and with a distribution of samples based on their concentrations. The expected observation was to see all-blood plotted near 2/3-blood, 2/3-blood near 1/3-blood, and 1/3-blood near all-brain. Distinct separation between kits was not expected on PCA plots. When raw .cel expression values were evaluated using principal component analysis, differences between the two amplification kits were apparent (Figure 1A). Variations from the expected pattern were observed in the plotting of technical replicates and in sample

concentrations. Additionally, Affymetrix-amplified samples and NuGEN-amplified samples demonstrated obvious separation along both PC1 and PC2. After normalization, the separation of samples based on amplification kit accounts for a greater proportion of the variation observed (Figure 1A, B). Separation based on kit was observed with raw .cel file expression values on both PC1 and PC2 axes, accounting for 31.5% and 14.7% of the variation respectively (Figure 1A). After normalization, separation based on amplification kit was demonstrated on PC1 and accounts for 69.5% of the variation (Figure 1B). While it is expected that there should be separation of samples based on blood/brain concentrations, it was not expected, after normalization, to see such obvious differences in expression from identical samples amplified by different kits.

When plotted separately, Affymetrix-amplified samples (Figure 1C) show the expected distribution of samples across PC1, accounting for 56.6% of the observed variation. NuGEN-amplified samples (Figure 1D) do not adhere as clearly to the expected pattern of distribution between samples, where the expected sequence would be all-blood next to 2/3-blood, 2/3-blood next to 1/3-blood, and 1/3-blood next to all-brain. PC1 accounts for 38.8% of the variation for NuGEN-amplified samples. In Figure 1D, the all-brain samples and the 2/3-blood samples align diagonally, along both the PC1 and PC2 axes, rather than only along the PC1 axis as expected, nor do the 2/3-blood and 1/3-blood samples fall between the all-blood and the all-brain as expected. Comparing Figure 1C to 1D, Affymetrix-amplified samples demonstrate a greater consistency than NuGEN-amplified samples for plotting of technical replicates and for distribution of samples based on blood/brain concentration.

Commonly, microarray datasets are filtered for the purpose of removing transcript clusters with normalized expression values similar to dataset background noise. Ha *et al.* (2009) suggest removing any transcript clusters with a normalized $\log_2$ expression value less than 5.0. Quackenbush (2002) states that statistical power is improved in DE detection when datasets are filtered at two standard deviations above the dataset background mean. Affymetrix HTA 2.0 microarrays include a small group of antigenomic transcript clusters that are useful in determining the background for a microarray dataset. For our datasets, the mean value for the Affymetrix amplified antigenomic transcript clusters was 4.047 and the upper and lower values for two standard deviations from the mean were 3.630 and 4.465. The mean value for the NuGEN amplified antigenomic transcript clusters was 2.725 and two standard deviations away

from the mean were 2.475 and 2.975. The Affymetrix median filtering value of 3.9 did come close to the upper value, 4.465, of two standard deviations from the Affymetrix antigenomic mean and did indeed become the method of filtering that was chosen. The antigenomic third quartile value used for the Affymetrix-amplified samples, at 6.0, was considerably higher and may have been too selective for detecting DE, as it produced lower numbers of DE transcript clusters. Both values chosen to filter the NuGEN-derived data, 3.1 for median filtering and 3.7 for antigenomic filtering, were higher than two standard deviations above the NuGEN antigenomic mean. Neither was able to improve detection of DE transcript clusters for further analysis. For this study, median filtering was chosen as the optimum method to be used for DE, but, as was demonstrated by both datasets, it is worthwhile to also look at the values for two standard deviations above the mean for further guidance in selection of a filtering threshold.

Volcano plots (Figure 4 & 5) of LFC from both filtering methods demonstrated that the Affymetrix GeneChip® WT Plus amplification system clearly produced more differentially expressed transcript clusters than the NuGEN Ovation® Pico WTA System V2 with Affymetrix HTA 2.0 microarrays. Affymetrix amplification also produces a greater range of expression values than NuGEN amplification for both filtering methods, as shown in violin plots (Figure 2 & 3). Counts of differentially expressed genes (Table 2) and their fold change ranges (Table 3) also indicate that Affymetrix amplification outperforms NuGEN amplification when used with HTA 2.0 microarrays for DE detection. In this analysis, the most stringent test of DE would be the 2/3-blood vs. 1/3-blood contrast, as the $\log_2$ fold change is expected to be |1.0|. The NuGEN amplification system produced no significant differentially expressed transcript clusters in this contrast. For the other contrasts, all-blood v. all-brain, all-blood v. 2/3-blood, and 1/3-blood v. all-brain, Affymetrix consistently outperformed NuGEN's amplification method for detection of DE transcript clusters. These results indicated that Affymetrix amplification was optimum for use in DE detection with HTA 2.0 microarrays. When considering only those transcript clusters with an adjusted p-value of less than 0.05 and removing any transcript clusters that did not have a LFC of at least |1.0| (biomarker subset), we saw the same pattern repeated, where Affymetrix amplification consistently produced greater numbers of DE transcript clusters (Table 4) and a greater LFC range (Table 6). Median filtering was selected over antigenomic filtering for Affymetrix-amplified samples for multiple reasons. While median filtering did not always produce more DE transcript clusters than antigenomic filtering for the whole dataset (Table 2),

median filtering did produce more DE transcript clusters for the biomarker subset (Table 4), where only those transcript clusters with an adjusted p-value < 0.05 and LFC > |1.0| were considered. Nearly all (99.9% or more) of the transcript clusters in the antigenomic filtered dataset were also in the median filtered dataset, so little would be lost by utilizing the median filtered dataset. If the antigenomic filtered dataset was used, some transcript clusters would be lost, mainly from the all-brain v. all-blood comparison (Table 5). Lastly, median filtering produced an equal number or more DE transcript clusters with LFC greater than |1.0| in all contrasts tested (Table 6).

The second goal of this study was to compare DE detection between microarray datasets and RNA-seq datasets. The PCA performed on LFC for Entrez IDs common to RNA-seq and the Affymetrix median datasets indicate differences between microarray and RNA-seq processing (Figure 8). The RNA-seq samples are aligned based on concentration along PC1, with little variation along PC2, demonstrating that differences within the RNA-seq dataset derive mainly from concentration variations, as would be expected. The Affymetrix median dataset aligns along both the PC1 and PC2 axes. There is separation based on concentration, as would be expected, but it is unclear why the microarray samples would demonstrate separation along both axes. It is possible that there are other factors affecting the microarray dataset that are not affecting the RNA-seq dataset. PC1 accounts for 65.8% of the variation and PC2 accounts for 16.1% of the variation. It is not known if the additional variation observed from the microarray dataset is affecting data quality. Violin plots of the overlapping Entrez IDs common to both RNA-seq and Affymetrix median microarrays consistently demonstrate a greater LFC range for RNA-seq as compared to microarray data, which would enable researchers to better detect differential expression (Figure 10).

When comparing between the full RNA-seq and the Affymetrix-amplified median filtered datasets, RNA-seq consistently demonstrated advantages over microarrays in DE detection. RNA-seq produced greater counts of DE Entrez IDs (Table 8) within the restrictive biomarker candidate subset, where only significant Entrez IDs/transcript cluster IDs with an FDR/adjusted p-value less than 0.05 and an LFC of greater than |1.0| are considered. Comparison of the Affymetrix-amplified median filtered dataset to the RNA-seq dataset also clearly demonstrated another advantage in biomarker detection. Greater LFC ranges in DE RNA-seq Entrez IDs were observed (Table 8). In the 2/3-blood v. 1/3-blood comparison, RNA-seq

produces a total of 1,353 significant and DE Entrez IDs with LFC > |1.0| or more. The Affymetrix median dataset, in the same comparison, only produced 121 significant and DE transcript cluster IDs, none of which had an LFC greater than |2.0|. Volcano plots (Figure 12) and count tables (Table 8) also clearly show the RNA-seq biomarker subset produced far more Entrez IDs with LFC > |2.0| than the Affymetrix median biomarker subset, enabling clearer biomarker detection.

The work presented by Wang *et al.* (2009) correlates with the improvement in sensitivity of DE genes detected by RNA-seq over the optimum microarray methodology determined by our study. Wang *et al.* 2009 found concordance between microarray and RNA-seq results for moderately expressed genes but found RNA-seq to have greater sensitivity for high and low expressed genes as compared to microarrays. RNA-seq is capable of sequencing the entire transcriptome from a sample, including unknown transcripts, depending on the depth of sequencing used (Kukurba & Montgomery, 2015). For microarrays to detect and quantify a gene, the gene must be known and a probe for it must be present on the microarray, although with well-characterized species, this is a less significant detraction. The improvement in sensitivity is still obvious when comparing RNA-seq to microarray results for our study using human blood and brain RNA.

## CONCLUSION

Affymetrix's GeneChip® WT PLUS outperformed NuGEN's Ovation® Pico WTA system V2 for RNA amplification in data quality and cost. The time required to perform both amplifications was roughly equivalent as well. For microarray data filtering, in this instance, it was determined that median filtering provided the greatest sensitivity in differential expression detection for Affymetrix-amplified data, but the NuGEN-amplified dataset does indicate that median filtering may not always prove successful and should be compared with other means of data filtering to ensure optimum results. Total RNA-seq also provided a higher quality of results for differential expression detection, as evidenced by wider log fold change ranges and increased numbers of DE genes detected. While there was a notable price increase to use total RNA-seq for analysis, as compared to Affymetrix amplification and microarrays, there was also a decrease in the time costs to the laboratory and a significant improvement in the quality of data collected.

## REFERENCES


Affymetrix. (n.d.). GeneChip™ human transcriptome array 2.0. Retrieved from ThermoFisher Scientific website: https://www.thermofisher.com/order/catalog/product/902162

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. Version 0.11.5.

Bentley, D. R., Balasubramanian, S., & Swerdlow, H. P. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218), 53.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15), 2114-2120.

Carvalho, B. S., & Irizarry, R. A. (2010). A framework for oligonucleotide microarray preprocessing. *Bioinformatics*, 26(19), 2363-2367.

Chang, T. W. (1983). Binding of cells to matrixes of distinct antibodies coated on solid surface. *Journal of Immunological Methods*, 65(1-2), 217-223.

Elder, J. K., Green, D. K., & Southern, E. M. (1986). Automatic reading of DNA sequencing gel autoradiographs using a large format digital scanner. *Nucleic Acids Research*, 14(1), 417-424.

Feezor, R. J., Baker, H. V., & Mindrinos, M. (2004). Whole blood and leukocyte RNA isolation for gene expression analyses. *Physiological Genomics*, 19(3), 247-254.

Fodor, S. P., Read, J. L., Pirrung, M. C. (1991). Light-directed, spatially addressable parallel chemical synthesis. *Science*, 251(4995), 767-773.

Gentleman, R., Carey, V., Huber, W. (2017). genefilter: genefilter: methods for filtering genes from high-throughput experiments. R package version 1.60.0.

GTEx Consortium. (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*, 348(6235), 648-660.

Ha, K. C., Coulombe-Huntington, J., & Majewski, J. (2009). Comparison of Affymetrix Gene Array with the Exon Array shows potential application for detection of transcript isoform variation. *BMC Genomics*, 10(1), 519.

Kauffmann, A., Gentleman, R., & Huber, W. (2009). arrayQualityMetrics—a bioconductor package for quality assessment of microarray data. *Bioinformatics*, 25(3), 415-416.

Koh, W., Pan, W., Gawad, C. (2014). Noninvasive in vivo monitoring of tissue-specific global gene expression in humans. *Proceedings of the National Academy of Sciences*, 111, 7361-7366, (2014).

Kukurba, K. R., & Montgomery, S. B. (2015). RNA sequencing and analysis. *Cold Spring Harbor Protocols*, 951-969.

Liao, Y., Smyth, G. K., & Shi, W. (2013). The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Research*, 41(10), e108.

MacDonald, J. W. (2017). pd.hta.2.0: Platform Design Info for Affymetrix HTA-2_0. R package version 3.12.2.

Mantione, K. J., Kream, R. M., Kuzelova, H. (2014). Comparing bioinformatic gene expression profiling methods: microarray and RNA-seq. *Medical Science Monitor Basic Research*, 20, 138-141.

Margulies, M., Egholm, M., Altman, W. E. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(15), 376-380.

Melé, M., Ferreira, P. G., Reverter, F. (2015). The human transcriptome across tissues and individuals. *Science*, 348(6235), 660-665.

NuGEN. (2016). User Guide Ovation® Pico WTA System V2. Retrieved from NuGEN website: https://www.nugen.com/sites/default/files/M01224_v5_User_Guide:_Ovation_Pico_WTA_System_V2_2214.pdf

Quackenbush, J. (2002). Microarray data normalization and transformation. *Nature Genetics*, 32, 496.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing.

Ritchie, M. E., Phipson, B., & Wu, D. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research,* 43(7), e47.

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1), 139-140.

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12), 5463-5467.

Schena, M., Shalon, D., Davis, R. W. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235), 467-470.

Shabihkhani, M., Lucey, G. M., Wei, B. (2014). The procurement, storage, and quality assurance of frozen blood and tissue biospecimens in pathology, biorepository, and biobank settings. *Clinical Biochemistry*, 47(4-5), 258-266.

Shalon, D., Smith, S. J., & Brown, P. O. (1996). A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Research*, 6(7), 639-645.

Shi, L., Reid, L. H., Jones, W. D. (2006). The MicroArray Quality Control (MAQC) project shows inter-and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology*, 24(9), 1151-1160.

ThermoFisher Scientific. (2017). Clariom D solutions for human, mouse, and rat. Retrieved from ThermoFisher Scientific website: https://www.thermofisher.com/order/catalog/product/902922?SID=srch-srp-902922

Van Gelder, R. N., Von Zastrow, M. E., Yool, A. (1990). Amplified RNA synthesized from limited quantities of heterogeneous cDNA. *Proceedings of the National Academy of Sciences*, 87(5), 1663-1667.

Wang, C., Gong, B., Bushel, P. R. (2014). A comprehensive study design reveals treatment- and transcript abundance-dependent concordance between RNA-seq and microarray data. *Nature Biotechnology*, 32(9), 926-932.

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1), 57.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer

Supplementary Table 1

# arrayQualityMetrics report for raw Affymetrix dataset



### - Array metadata and outlier detection overview

| | array | sampleNames | *1 | *2 | *3 | index | concentration |
|---|---|---|---|---|---|---|---|
| ☐ | 1 | 1000BloodA | | x | | 1 | 2/3_blood |
| ☐ | 2 | 1000BloodB | | | | 2 | 2/3_blood |
| ☐ | 3 | 1500BloodA | | | | 3 | all_blood |
| ☐ | 4 | 1500BloodB | | | | 4 | all_blood |
| ☐ | 5 | 1500BrainA | | | | 5 | all_brain |
| ☐ | 6 | 1500BrainB | | | | 6 | all_brain |
| ☐ | 7 | 500BloodA | | x | | 7 | 1/3_blood |
| ☐ | 8 | 500BloodB | | | | 8 | 1/3_blood |

The columns named *1, *2, ... indicate the calls from the different outlier detection methods:

1. outlier detection by Distances between arrays
2. outlier detection by Boxplots
3. outlier detection by MA plots

The outlier detection criteria are explained below in the respective sections. Arrays that were called outliers by at least one criterion are marked by checkbox selection in this table, and are indicated by highlighted lines or points in some of the plots below. By clicking the checkboxes in the table, or on the corresponding points/lines in the plots, you can modify the selection. To reset the selection, reload the HTML page in your browser.

At the scope covered by this software, outlier detection is a poorly defined question, and there is no 'right' or 'wrong' answer. These are hints which are intended to be followed up manually. If you want to automate outlier detection, you need to limit the scope to a particular platform and experimental design, and then choose and calibrate the metrics used.

## Section 1: Between array comparison

### - Figure 1: Distances between arrays.

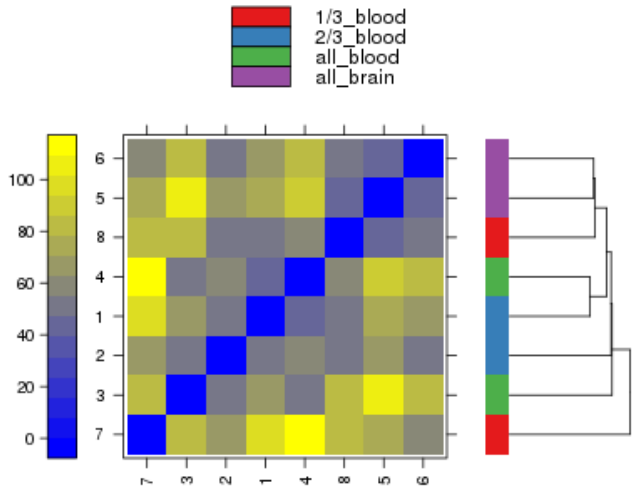



**Figure 1** (PDF file) shows a false color heatmap of the distances between arrays. The color scale is chosen to cover the range of distances encountered in the dataset. Patterns in this plot can indicate clustering of the arrays either because of intended biological or unintended experimental factors (batch effects). The distance $d_{ab}$ between two arrays *a* and *b* is computed as the mean absolute difference ($L_1$-distance) between the data of the arrays (using the data from all probes without filtering). In formula, $d_{ab}$ = mean | $M_{ai}$ - $M_{bi}$ |, where $M_{ai}$ is the value of the *i*-th probe on the *a*-th array. Outlier detection was performed by looking for arrays for which the sum of the distances to all other arrays, $S_a = \Sigma_b\, d_{ab}$ was exceptionally large. No such arrays were detected.

**+ Figure 2: Outlier detection for Distances between arrays.**
**- Figure 3: Principal Component Analysis.**

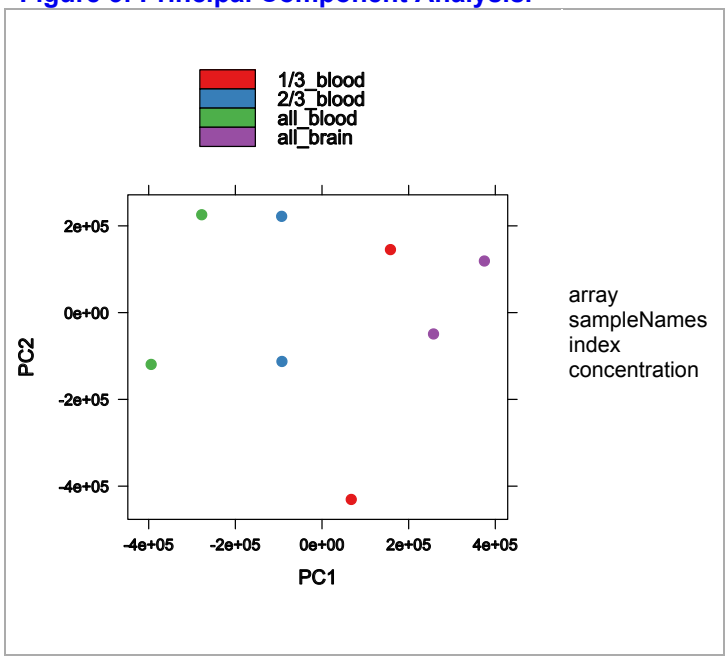



**Figure 3** (PDF file) shows a scatterplot of the arrays along the first two principal components. You can use this plot to explore if the arrays cluster, and whether this is according to an intended experimental factor, or according to unintended causes such as batch effects. Move the mouse over the points to see the sample names.
Principal component analysis is a dimension reduction and visualisation technique that is here used to project the multivariate data vector of each array into a two-dimensional plot, such that the spatial arrangement of the points in the plot reflects the overall data (dis)similarity between the arrays.

---

## Section 2: Array intensity distributions

**- Figure 4: Boxplots.**

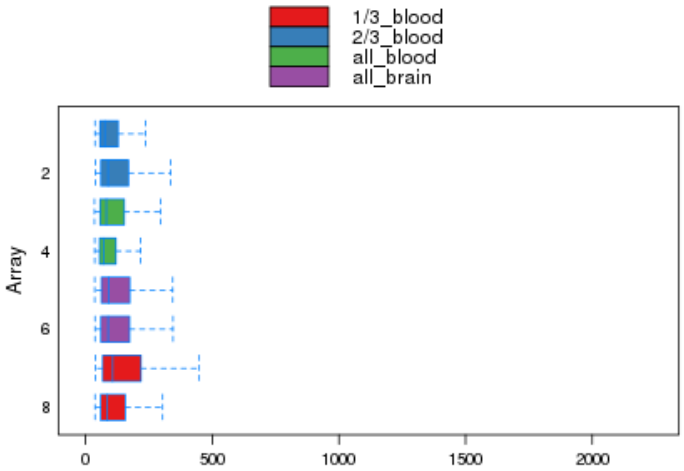



**Figure 4** (PDF file) shows boxplots representing summaries of the signal intensity distributions of the arrays. Each box corresponds to one array. Typically, one expects the boxes to have similar positions and widths. If the distribution of an array is very different from the others, this may indicate an experimental problem. Outlier detection was performed by computing the Kolmogorov-Smirnov statistic $K_a$ between each array's distribution and the distribution of the pooled data.

**+ Figure 5: Outlier detection for Boxplots.**
**- Figure 6: Density plots.**

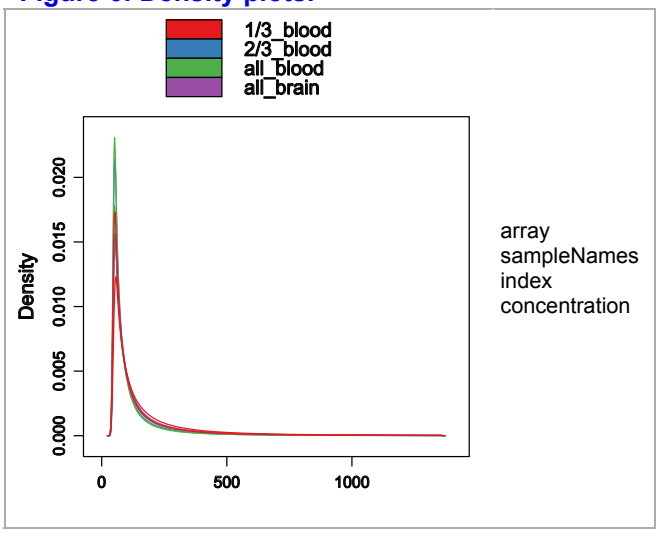



**Figure 6** (PDF file) shows density estimates (smoothed histograms) of the data. Typically, the distributions of the arrays should have similar shapes and ranges. Arrays whose distributions are very different from the others should be considered for possible problems. Various features of the distributions can be indicative of quality related phenomena. For instance, high levels of background will shift an array's distribution to the right. Lack of signal diminishes its right right tail. A bulge at the upper end of the intensity range often indicates signal saturation.

---

## Section 3: Variance mean dependence

**- Figure 7: Standard deviation versus rank of the mean.**

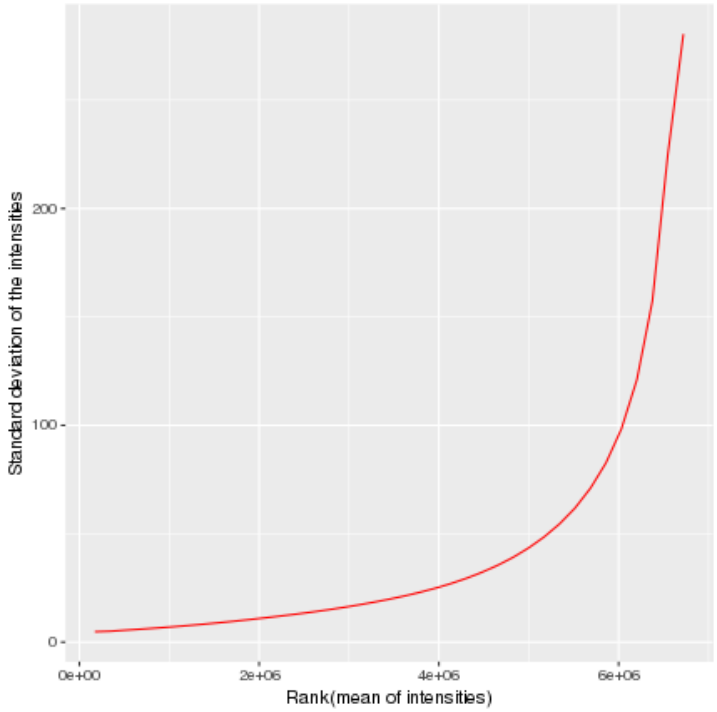



**Figure 7** (PDF file) shows a density plot of the standard deviation of the intensities across arrays on the *y*-axis versus the rank of their mean on the *x*-axis. The red dots, connected by lines, show the running median of the standard deviation. After normalisation and transformation to a logarithm(-like) scale, one typically expects the red line to be approximately horizontal, that is, show no substantial trend. In some cases, a hump on the right hand of the x-axis can be observed and is symptomatic of a saturation of the intensities.

---

## Section 4: Individual array quality

### - Figure 8: MA plots.

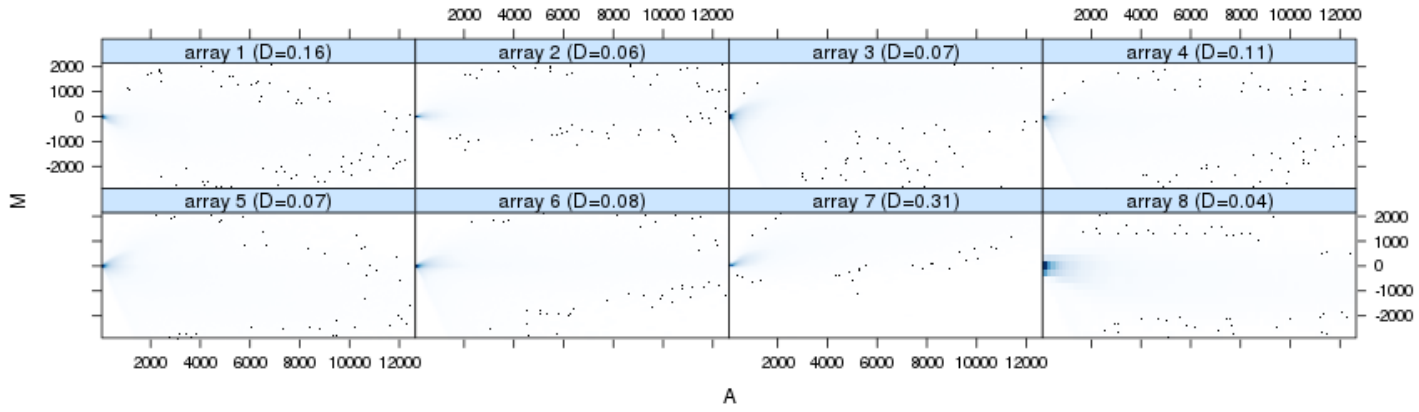



**Figure 8** (PDF file) shows MA plots. M and A are defined as:
$M = \log_2(I_1) - \log_2(I_2)$
$A = 1/2\ (\log_2(I_1)+\log_2(I_2))$,
where $I_1$ is the intensity of the array studied,and $I_2$ is the intensity of a "pseudo"-array that consists of the median across arrays. Typically, we expect the mass of the distribution in an MA plot to be concentrated along the M = 0 axis, and there should be no trend in M as a function of A. If there is a trend in the lower range of A, this often indicates that the arrays have different background intensities; this may be addressed by background correction. A trend in the upper range of A can indicate saturation of the measurements; in mild cases, this may be addressed by non-linear normalisation (e.g. quantile normalisation).
Outlier detection was performed by computing Hoeffding's statistic $D_a$ on the joint distribution of A and M for each array. The value of $D_a$ is shown in the panel headings. 2 arrays had $D_a$>0.15 and were marked as outliers. For more information on Hoeffing's *D*-statistic, please see the manual page of the function `hoeffd` in the `Hmisc` package.

## - Figure 9: Outlier detection for MA plots.

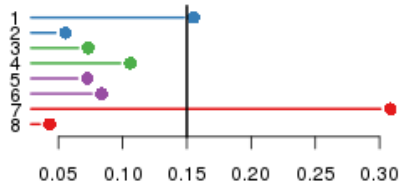



**Figure 9** (PDF file) shows a bar chart of the $D_a$, the outlier detection criterion from the previous figure. The bars are shown in the original order of the arrays. A threshold of 0.15 was used, which is indicated by the vertical line. 2 arrays exceeded the threshold and were considered outliers.

---

This report has been created with arrayQualityMetrics 3.34.0 under R version 3.5.0 (2018-04-23).

---

(Page generated on Fri Aug 10 16:52:18 2018 by hwriter )

Supplementary Table 2

# arrayQualityMetrics report for raw NuGEN dataset



### - Array metadata and outlier detection overview

| | array | sampleNames | *1 | *2 | *3 | index | concentration |
|---|---|---|---|---|---|---|---|
| ☐ | 1 | 1000Blood A | | | | 1 | 2/3_blood |
| ☐ | 2 | 1000Blood B | | | x | 2 | 2/3_blood |
| ☐ | 3 | 1500Blood A | | | | 3 | all_blood |
| ☐ | 4 | 1500Blood B | | | | 4 | all_blood |
| ☐ | 5 | 1500Brain A | | | | 5 | all_brain |
| ☐ | 6 | 1500Brain B | | | x | 6 | all_brain |
| ☐ | 7 | 500Blood A | | | x | 7 | 1/3_blood |
| ☐ | 8 | 500Blood B | | | | 8 | 1/3_blood |

The columns named *1, *2, ... indicate the calls from the different outlier detection methods:

1. outlier detection by Distances between arrays
2. outlier detection by Boxplots
3. outlier detection by MA plots

The outlier detection criteria are explained below in the respective sections. Arrays that were called outliers by at least one criterion are marked by checkbox selection in this table, and are indicated by highlighted lines or points in some of the plots below. By clicking the checkboxes in the table, or on the corresponding points/lines in the plots, you can modify the selection. To reset the selection, reload the HTML page in your browser.

At the scope covered by this software, outlier detection is a poorly defined question, and there is no 'right' or 'wrong' answer. These are hints which are intended to be followed up manually. If you want to automate outlier detection, you need to limit the scope to a particular platform and experimental design, and then choose and calibrate the metrics used.

## Section 1: Between array comparison

### - Figure 1: Distances between arrays.

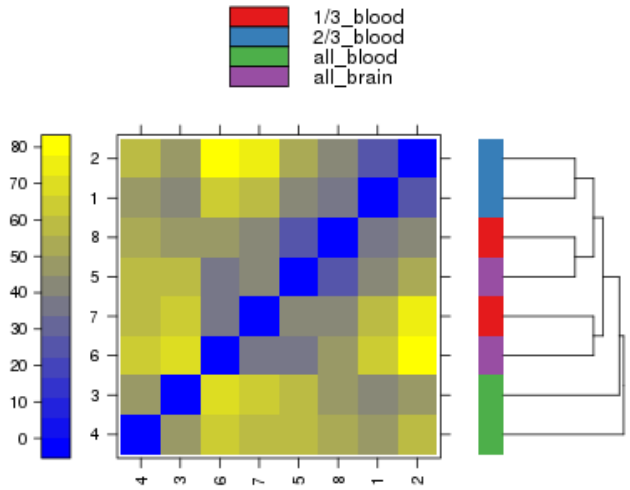



**Figure 1** (PDF file) shows a false color heatmap of the distances between arrays. The color scale is chosen to cover the range of distances encountered in the dataset. Patterns in this plot can indicate clustering of the arrays either because of intended biological or unintended experimental factors (batch effects). The distance $d_{ab}$ between two arrays $a$ and $b$ is computed as the mean absolute difference ($L_1$-distance) between the data of the arrays (using the data from all probes without filtering). In formula, $d_{ab}$ = mean | $M_{ai}$ - $M_{bi}$ |, where $M_{ai}$ is the value of the $i$-th probe on the $a$-th array. Outlier detection was performed by looking for arrays for which the sum of the distances to all other arrays, $S_a = \Sigma_b\, d_{ab}$ was exceptionally large. No such arrays were detected.

**+ Figure 2: Outlier detection for Distances between arrays.**
**- Figure 3: Principal Component Analysis.**

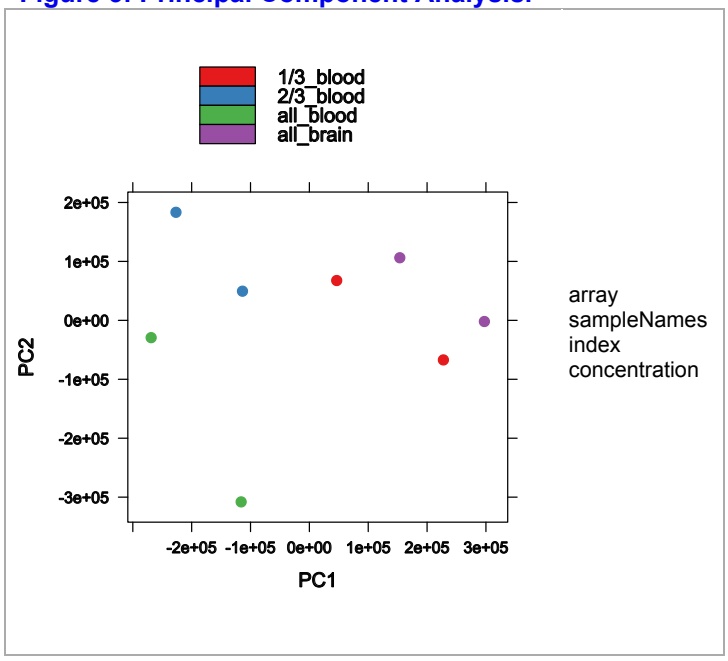



**Figure 3** (PDF file) shows a scatterplot of the arrays along the first two principal components. You can use this plot to explore if the arrays cluster, and whether this is according to an intended experimental factor, or according to unintended causes such as batch effects. Move the mouse over the points to see the sample names.
Principal component analysis is a dimension reduction and visualisation technique that is here used to project the multivariate data vector of each array into a two-dimensional plot, such that the spatial arrangement of the points in the plot reflects the overall data (dis)similarity between the arrays.

---

## Section 2: Array intensity distributions

**- Figure 4: Boxplots.**

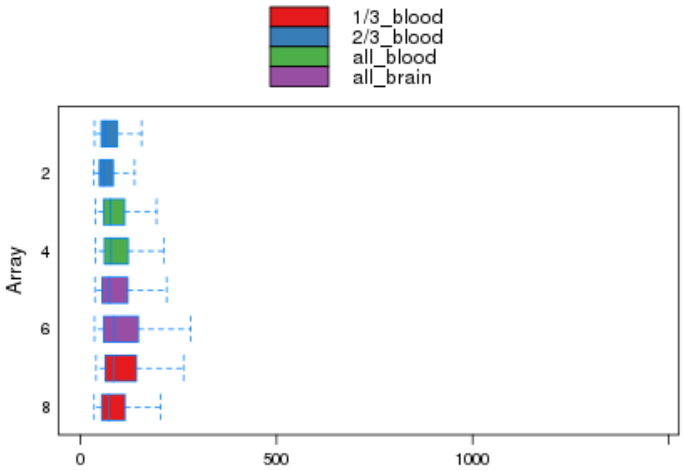



**Figure 4** (PDF file) shows boxplots representing summaries of the signal intensity distributions of the arrays. Each box corresponds to one array. Typically, one expects the boxes to have similar positions and widths. If the distribution of an array is very different from the others, this may indicate an experimental problem. Outlier detection was performed by computing the Kolmogorov-Smirnov statistic $K_a$ between each array's distribution and the distribution of the pooled data.

**+ Figure 5: Outlier detection for Boxplots.**
**- Figure 6: Density plots.**

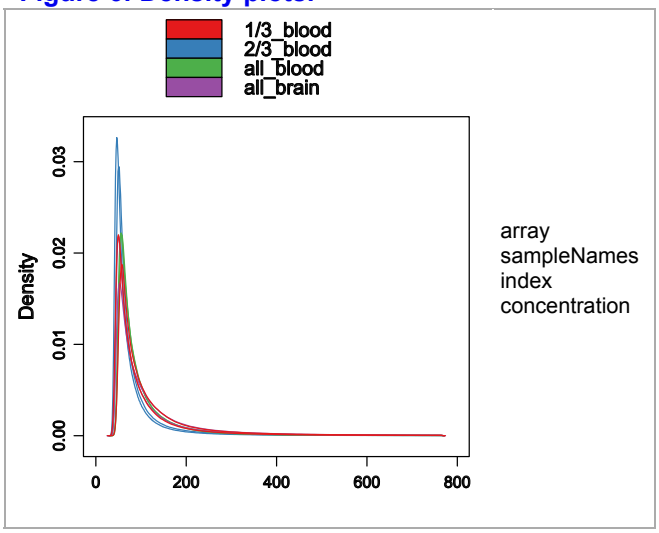



**Figure 6** (PDF file) shows density estimates (smoothed histograms) of the data. Typically, the distributions of the arrays should have similar shapes and ranges. Arrays whose distributions are very different from the others should be considered for possible problems. Various features of the distributions can be indicative of quality related phenomena. For instance, high levels of background will shift an array's distribution to the right. Lack of signal diminishes its right right tail. A bulge at the upper end of the intensity range often indicates signal saturation.

---

## Section 3: Variance mean dependence

**- Figure 7: Standard deviation versus rank of the mean.**

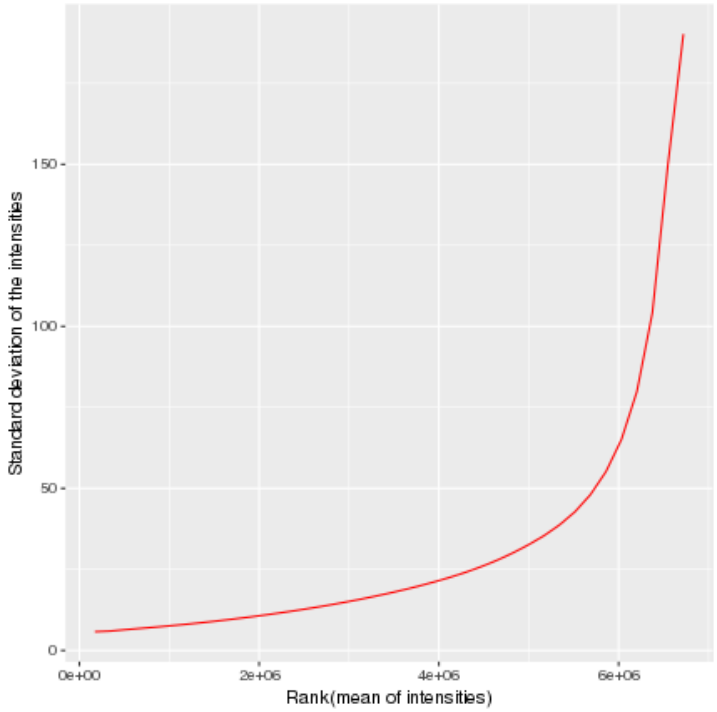



**Figure 7** (PDF file) shows a density plot of the standard deviation of the intensities across arrays on the *y*-axis versus the rank of their mean on the *x*-axis. The red dots, connected by lines, show the running median of the standard deviation. After normalisation and transformation to a logarithm(-like) scale, one typically expects the red line to be approximately horizontal, that is, show no substantial trend. In some cases, a hump on the right hand of the x-axis can be observed and is symptomatic of a saturation of the intensities.

---

## Section 4: Individual array quality

### - Figure 8: MA plots.

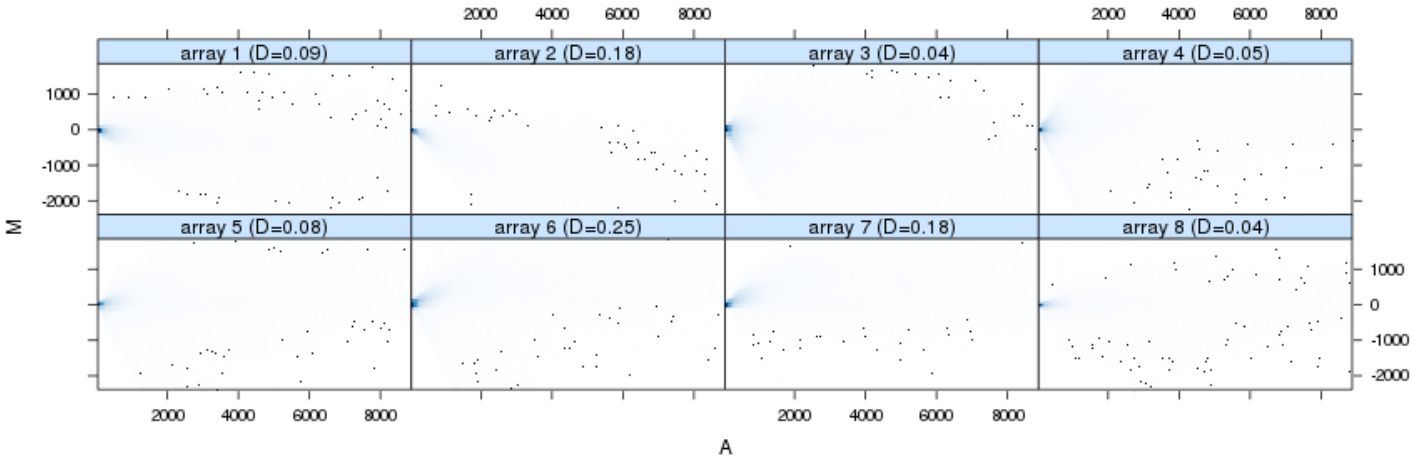



**Figure 8** (PDF file) shows MA plots. M and A are defined as:
$M = \log_2(I_1) - \log_2(I_2)$
$A = 1/2\ (\log_2(I_1)+\log_2(I_2))$,
where $I_1$ is the intensity of the array studied,and $I_2$ is the intensity of a "pseudo"-array that consists of the median across arrays. Typically, we expect the mass of the distribution in an MA plot to be concentrated along the M = 0 axis, and there should be no trend in M as a function of A. If there is a trend in the lower range of A, this often indicates that the arrays have different background intensities; this may be addressed by background correction. A trend in the upper range of A can indicate saturation of the measurements; in mild cases, this may be addressed by non-linear normalisation (e.g. quantile normalisation).
Outlier detection was performed by computing Hoeffding's statistic $D_a$ on the joint distribution of A and M for each array. The value of $D_a$ is shown in the panel headings. 3 arrays had $D_a$>0.15 and were marked as outliers. For more information on Hoeffing's *D*-statistic, please see the manual page of the function `hoeffd` in the `Hmisc` package.

## - Figure 9: Outlier detection for MA plots.

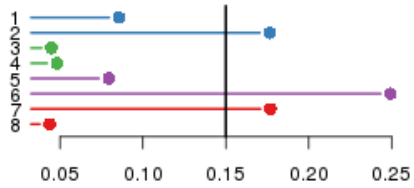



**Figure 9** (PDF file) shows a bar chart of the $D_a$, the outlier detection criterion from the previous figure. The bars are shown in the original order of the arrays. A threshold of 0.15 was used, which is indicated by the vertical line. 3 arrays exceeded the threshold and were considered outliers.

---

This report has been created with arrayQualityMetrics 3.34.0 under R version 3.5.0 (2018-04-23).

---

(Page generated on Mon Aug 13 08:50:48 2018 by hwriter )

Supplementary Table 3

# arrayQualityMetrics report for normalized Affymetrix dataset



### - Array metadata and outlier detection overview

| | array | sampleNames | *1 | *2 | *3 | index | concentration |
|---|---|---|---|---|---|---|---|
| ☐ | 1 | 1000 Blood A | | | | 1 | 2/3_blood |
| ☐ | 2 | 1000 Blood B | | | | 2 | 2/3_blood |
| ☐ | 3 | 1500 Blood A | | | | 3 | all_blood |
| ☐ | 4 | 1500 Blood B | | | | 4 | all_blood |
| ☐ | 5 | 1500 Brain A | | | | 5 | all_brain |
| ☐ | 6 | 1500 Brain B | | | | 6 | all_brain |
| ☐ | 7 | 500 Blood A | | | | 7 | 1/3_blood |
| ☐ | 8 | 500 Blood B | | | | 8 | 1/3_blood |

The columns named *1, *2, ... indicate the calls from the different outlier detection methods:

1. outlier detection by Distances between arrays
2. outlier detection by Boxplots
3. outlier detection by MA plots

The outlier detection criteria are explained below in the respective sections. Arrays that were called outliers by at least one criterion are marked by checkbox selection in this table, and are indicated by highlighted lines or points in some of the plots below. By clicking the checkboxes in the table, or on the corresponding points/lines in the plots, you can modify the selection. To reset the selection, reload the HTML page in your browser.

At the scope covered by this software, outlier detection is a poorly defined question, and there is no 'right' or 'wrong' answer. These are hints which are intended to be followed up manually. If you want to automate outlier detection, you need to limit the scope to a particular platform and experimental design, and then choose and calibrate the metrics used.

## Section 1: Between array comparison

### - Figure 1: Distances between arrays.

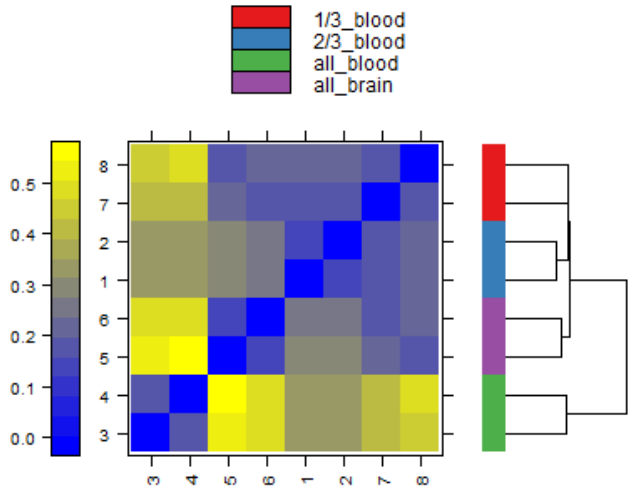



**Figure 1** (PDF file) shows a false color heatmap of the distances between arrays. The color scale is chosen to cover the range of distances encountered in the dataset. Patterns in this plot can indicate clustering of the arrays either because of intended biological or unintended experimental factors (batch effects). The distance $d_{ab}$ between two arrays *a* and *b* is computed as the mean absolute difference ($L_1$-distance) between the data of the arrays (using the data from all probes without filtering). In formula, $d_{ab}$ = mean | $M_{ai}$ - $M_{bi}$ |, where $M_{ai}$ is the value of the *i*-th probe on the *a*-th array. Outlier detection was performed by looking for arrays for which the sum of the distances to all other arrays, $S_a = \Sigma_b d_{ab}$ was exceptionally large. No such arrays were detected.

**+ Figure 2: Outlier detection for Distances between arrays.**
**- Figure 3: Principal Component Analysis.**

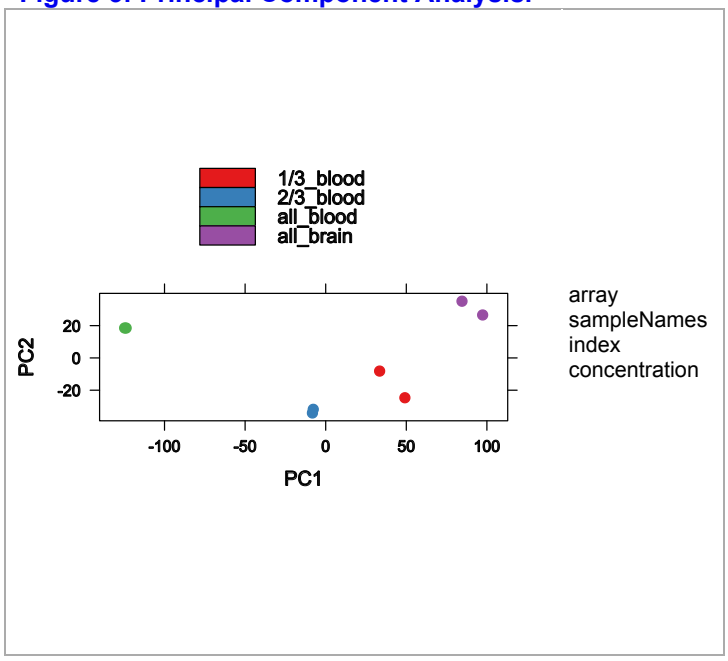



**Figure 3** (PDF file) shows a scatterplot of the arrays along the first two principal components. You can use this plot to explore if the arrays cluster, and whether this is according to an intended experimental factor, or according to unintended causes such as batch effects. Move the mouse over the points to see the sample names.
Principal component analysis is a dimension reduction and visualisation technique that is here used to project the multivariate data vector of each array into a two-dimensional plot, such that the spatial arrangement of the points in the plot reflects the overall data (dis)similarity between the arrays.

---

## Section 2: Array intensity distributions

**- Figure 4: Boxplots.**

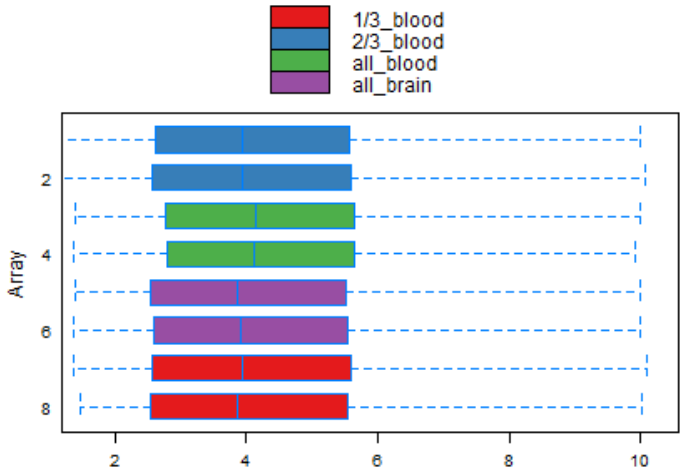



**Figure 4** (PDF file) shows boxplots representing summaries of the signal intensity distributions of the arrays. Each box corresponds to one array. Typically, one expects the boxes to have similar positions and widths. If the distribution of an array is very different from the others, this may indicate an experimental problem. Outlier detection was performed by computing the Kolmogorov-Smirnov statistic $K_a$ between each array's distribution and the distribution of the pooled data.

**+ Figure 5: Outlier detection for Boxplots.**
**- Figure 6: Density plots.**

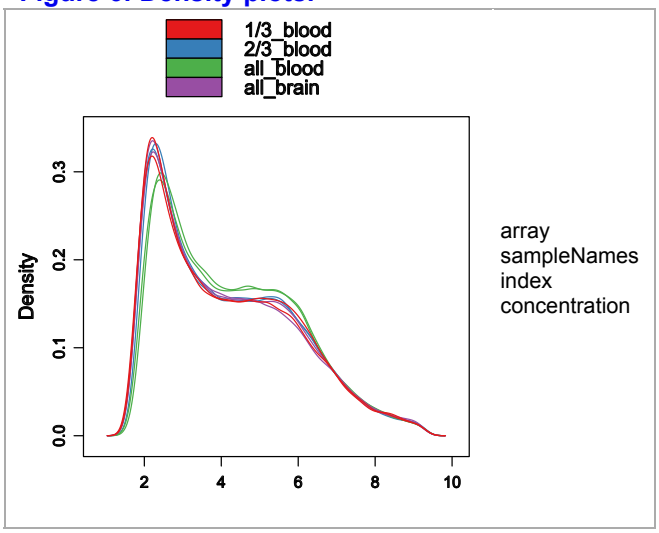



**Figure 6** (PDF file) shows density estimates (smoothed histograms) of the data. Typically, the distributions of the arrays should have similar shapes and ranges. Arrays whose distributions are very different from the others should be considered for possible problems. Various features of the distributions can be indicative of quality related phenomena. For instance, high levels of background will shift an array's distribution to the right. Lack of signal diminishes its right right tail. A bulge at the upper end of the intensity range often indicates signal saturation.

---

## Section 3: Variance mean dependence

**- Figure 7: Standard deviation versus rank of the mean.**

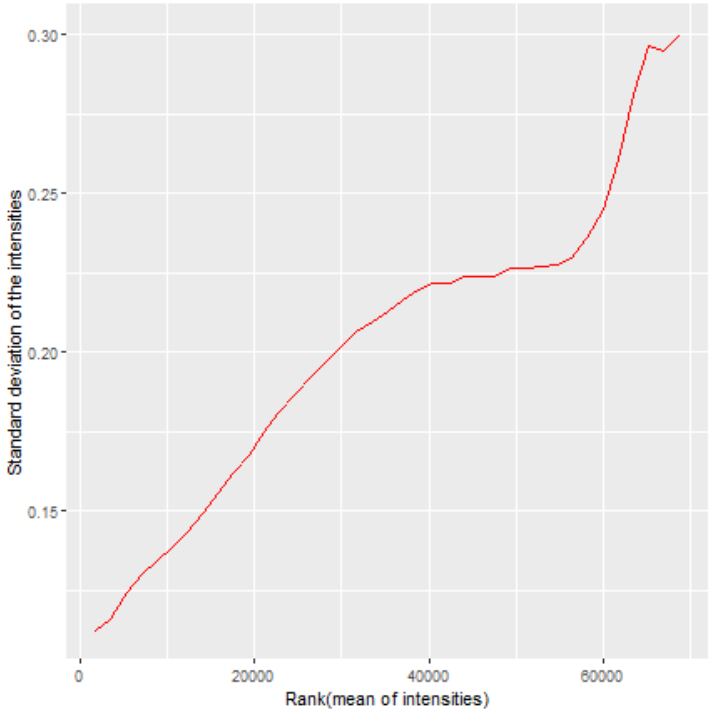



**Figure 7** (PDF file) shows a density plot of the standard deviation of the intensities across arrays on the *y*-axis versus the rank of their mean on the *x*-axis. The red dots, connected by lines, show the running median of the standard deviation. After normalisation and transformation to a logarithm(-like) scale, one typically expects the red line to be approximately horizontal, that is, show no substantial trend. In some cases, a hump on the right hand of the x-axis can be observed and is symptomatic of a saturation of the intensities.

---

## Section 4: Individual array quality

### - Figure 8: MA plots.

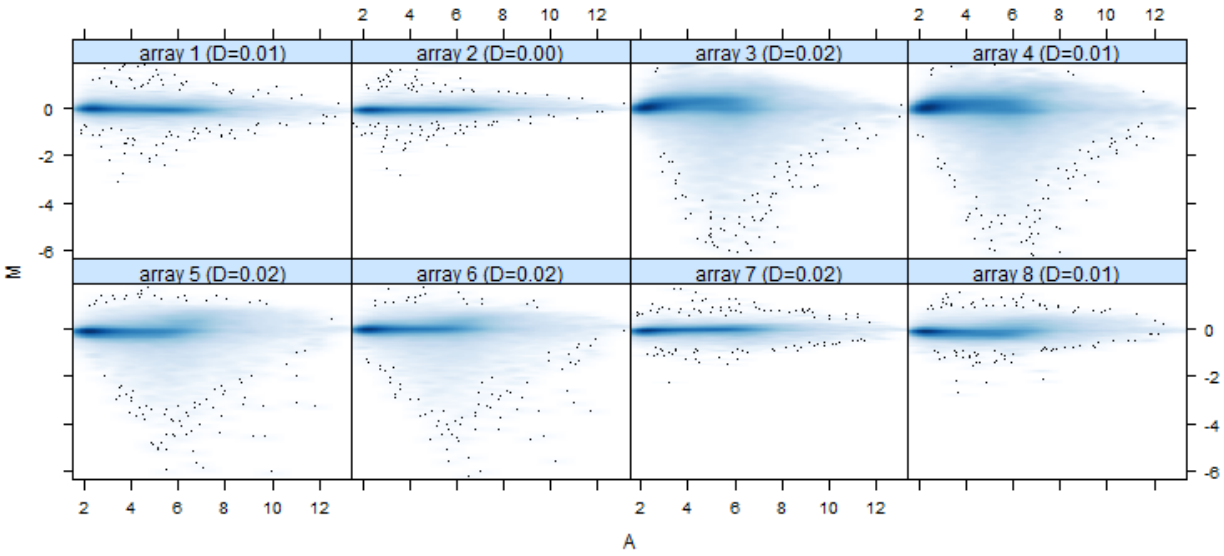



**Figure 8** (PDF file) shows MA plots. M and A are defined as:
$M = \log_2(I_1) - \log_2(I_2)$
$A = 1/2\ (\log_2(I_1)+\log_2(I_2))$,
where $I_1$ is the intensity of the array studied,and $I_2$ is the intensity of a "pseudo"-array that consists of the median across arrays. Typically, we expect the mass of the distribution in an MA plot to be concentrated along the M = 0 axis, and there should be no trend in M as a function of A. If there is a trend in the lower range of A, this often indicates that the arrays have different background intensities; this may be addressed by background correction. A trend in the upper range of A can indicate saturation of the measurements; in mild cases, this may be addressed by non-linear normalisation (e.g. quantile normalisation).
Outlier detection was performed by computing Hoeffding's statistic $D_a$ on the joint distribution of A and M for each array. The value of $D_a$ is shown in the panel headings. 0 arrays had $D_a$>0.15 and were marked as outliers. For more information on Hoeffing's *D*-statistic, please see the manual page of the function `hoeffd` in the `Hmisc` package.

Supplementary Table 4

# arrayQualityMetrics report for normalized NuGEN dataset



### - Array metadata and outlier detection overview

| | array | sampleNames | *1 | *2 | *3 | index | concentration |
|---|---|---|---|---|---|---|---|
| ☐ | 1 | 1000 Blood A | | | | 1 | 2/3_blood |
| ☐ | 2 | 1000 Blood B | | | | 2 | 2/3_blood |
| ☐ | 3 | 1500 Blood A | | | | 3 | all_blood |
| ☐ | 4 | 1500 Blood B | | | | 4 | all_blood |
| ☐ | 5 | 1500 Brain A | | | | 5 | all_brain |
| ☐ | 6 | 1500 Brain B | | | | 6 | all_brain |
| ☐ | 7 | 500 Blood A | | | | 7 | 1/3_blood |
| ☐ | 8 | 500 Blood B | | | | 8 | 1/3_blood |

The columns named *1, *2, ... indicate the calls from the different outlier detection methods:

1. outlier detection by Distances between arrays
2. outlier detection by Boxplots
3. outlier detection by MA plots

The outlier detection criteria are explained below in the respective sections. Arrays that were called outliers by at least one criterion are marked by checkbox selection in this table, and are indicated by highlighted lines or points in some of the plots below. By clicking the checkboxes in the table, or on the corresponding points/lines in the plots, you can modify the selection. To reset the selection, reload the HTML page in your browser.

At the scope covered by this software, outlier detection is a poorly defined question, and there is no 'right' or 'wrong' answer. These are hints which are intended to be followed up manually. If you want to automate outlier detection, you need to limit the scope to a particular platform and experimental design, and then choose and calibrate the metrics used.

## Section 1: Between array comparison

### - Figure 1: Distances between arrays.

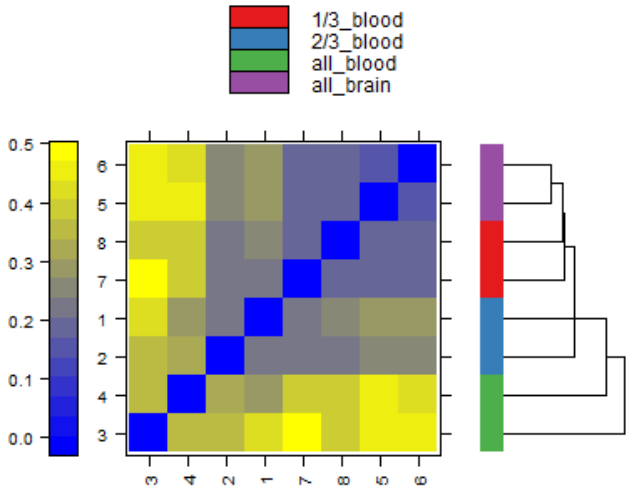



**Figure 1** (PDF file) shows a false color heatmap of the distances between arrays. The color scale is chosen to cover the range of distances encountered in the dataset. Patterns in this plot can indicate clustering of the arrays either because of intended biological or unintended experimental factors (batch effects). The distance $d_{ab}$ between two arrays *a* and *b* is computed as the mean absolute difference ($L_1$-distance) between the data of the arrays (using the data from all probes without filtering). In formula, $d_{ab}$ = mean | $M_{ai}$ - $M_{bi}$ |, where $M_{ai}$ is the value of the *i*-th probe on the *a*-th array. Outlier detection was performed by looking for arrays for which the sum of the distances to all other arrays, $S_a = \Sigma_b d_{ab}$ was exceptionally large. No such arrays were detected.

**+ Figure 2: Outlier detection for Distances between arrays.**
**- Figure 3: Principal Component Analysis.**

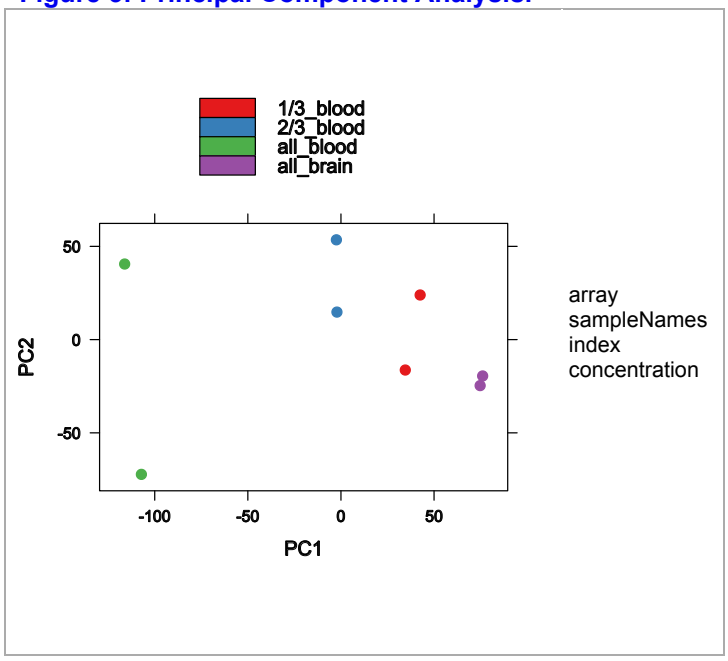



**Figure 3** (PDF file) shows a scatterplot of the arrays along the first two principal components. You can use this plot to explore if the arrays cluster, and whether this is according to an intended experimental factor, or according to unintended causes such as batch effects. Move the mouse over the points to see the sample names.
Principal component analysis is a dimension reduction and visualisation technique that is here used to project the multivariate data vector of each array into a two-dimensional plot, such that the spatial arrangement of the points in the plot reflects the overall data (dis)similarity between the arrays.

---

## Section 2: Array intensity distributions

**- Figure 4: Boxplots.**

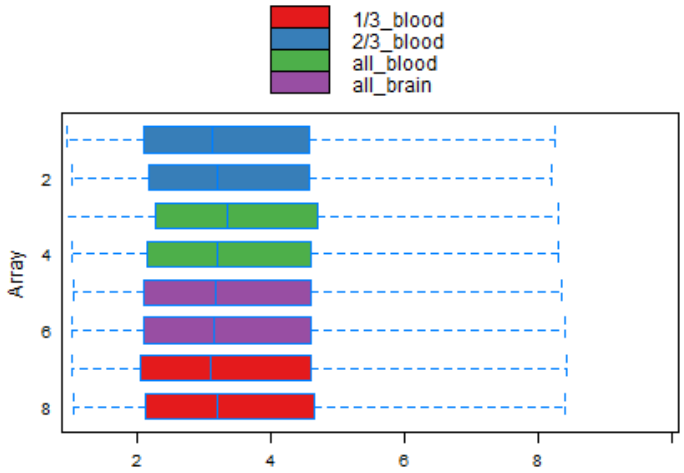



**Figure 4** (PDF file) shows boxplots representing summaries of the signal intensity distributions of the arrays. Each box corresponds to one array. Typically, one expects the boxes to have similar positions and widths. If the distribution of an array is very different from the others, this may indicate an experimental problem. Outlier detection was performed by computing the Kolmogorov-Smirnov statistic $K_a$ between each array's distribution and the distribution of the pooled data.

**+ Figure 5: Outlier detection for Boxplots.**
**- Figure 6: Density plots.**

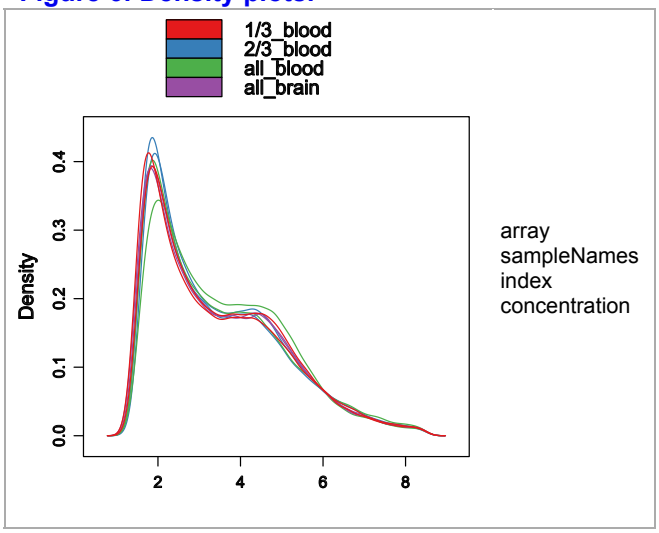



**Figure 6** (PDF file) shows density estimates (smoothed histograms) of the data. Typically, the distributions of the arrays should have similar shapes and ranges. Arrays whose distributions are very different from the others should be considered for possible problems. Various features of the distributions can be indicative of quality related phenomena. For instance, high levels of background will shift an array's distribution to the right. Lack of signal diminishes its right right tail. A bulge at the upper end of the intensity range often indicates signal saturation.

---

## Section 3: Variance mean dependence

**- Figure 7: Standard deviation versus rank of the mean.**

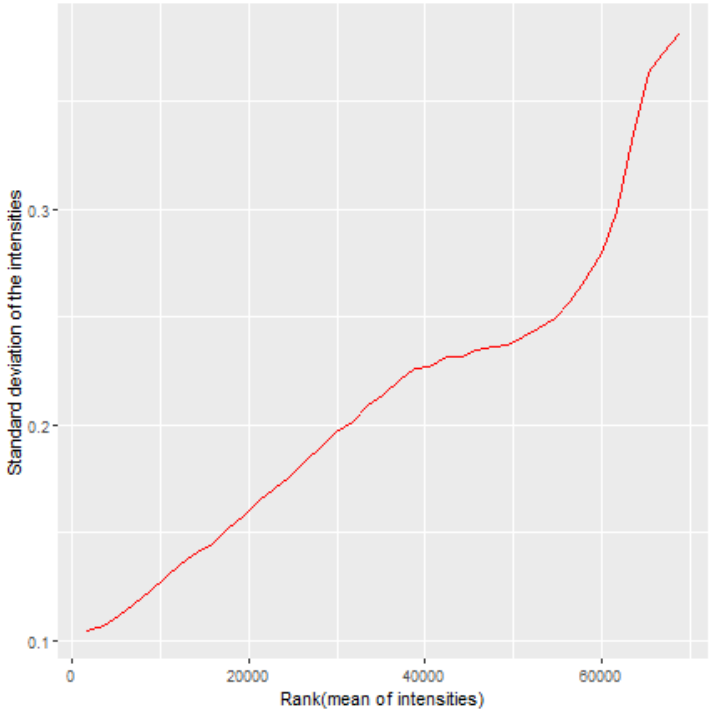



**Figure 7** (PDF file) shows a density plot of the standard deviation of the intensities across arrays on the *y*-axis versus the rank of their mean on the *x*-axis. The red dots, connected by lines, show the running median of the standard deviation. After normalisation and transformation to a logarithm(-like) scale, one typically expects the red line to be approximately horizontal, that is, show no substantial trend. In some cases, a hump on the right hand of the x-axis can be observed and is symptomatic of a saturation of the intensities.

---

## Section 4: Individual array quality

### - Figure 8: MA plots.

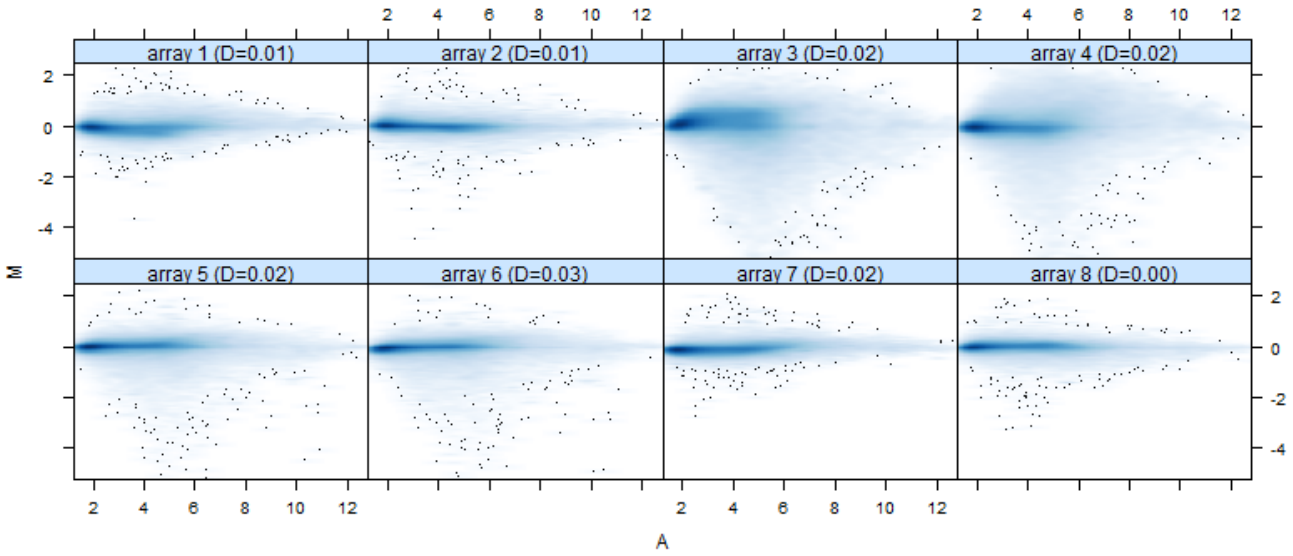



**Figure 8** (PDF file) shows MA plots. M and A are defined as:
$M = \log_2(I_1) - \log_2(I_2)$
$A = 1/2\ (\log_2(I_1)+\log_2(I_2))$,
where $I_1$ is the intensity of the array studied,and $I_2$ is the intensity of a "pseudo"-array that consists of the median across arrays. Typically, we expect the mass of the distribution in an MA plot to be concentrated along the M = 0 axis, and there should be no trend in M as a function of A. If there is a trend in the lower range of A, this often indicates that the arrays have different background intensities; this may be addressed by background correction. A trend in the upper range of A can indicate saturation of the measurements; in mild cases, this may be addressed by non-linear normalisation (e.g. quantile normalisation).
Outlier detection was performed by computing Hoeffding's statistic $D_a$ on the joint distribution of A and M for each array. The value of $D_a$ is shown in the panel headings. 0 arrays had $D_a$>0.15 and were marked as outliers. For more information on Hoeffing's *D*-statistic, please see the manual page of the function `hoeffd` in the `Hmisc` package.