AWARD NUMBER: W81XWH-14-1-0263

TITLE: Early Detection of NSCLC Using Stromal Markers in Peripheral Blood

PRINCIPAL INVESTIGATOR: Dingcheng Gao

CONTRACTING ORGANIZATION: Weill Cornell Medical College of Cornell University
New York, NY 10065

REPORT DATE: November 2017

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
November 2017	Final	1 Sep 2014 - 31 Aug 2017

4. TITLE AND SUBTITLE

Early Detection of NSCLC Using Stromal Markers in Peripheral Blood

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-14-1-0263

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Dingcheng Gao

E-Mail: dig2009@med.cornell.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Weill Cornell Medical College of Cornell University
1300 York Ave, New York, Ny 10065

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

A recent screening trial showed that the use of low dose computed tomography (LDCT) resulted in a 20% reduction in lung cancer mortality, however there was a 96% false positive rate associated with LDCT. Thus, there is an immediate clinical need to develop a diagnostic biomarker that would select patients with CT detected nodules for further testing. The ease with which blood can be sampled makes it a logical choice in which to discover diagnostic biomarkers, however the clinical utility of tumor derived proteins, miRNA or circulating tumor cells as blood-based biomarkers has been limited. In this proposal, instead of tumor-derived biomarkers, we will focus on host response to tumor growth. It has been well documented that tumor growth systemically stimulates and mobilizes BM-derived hematopoietic cells to the tumor bed to establish a permissive microenvironment. Preliminary studies in our lab have shown that in lung cancer patients, the circulating myeloid cells are transcriptionally altered and the alteration is tumor dependent. The specific transcriptomic signature of circulating myeloid cells may provide us unique resources for lung cancer biomarker discovery. Therefore, we hypothesized that the circulating BM-derived myeloid cells carry specific transcriptomic signature, which may be useful for early lung cancer diagnosis. The specific aims are: **Aim 1.** To identify a NSCLC-dependent transcriptomic signature in circulating myeloid cells. **Aim 2.** To validate the diagnostic value of the specific gene signatures of circulating myeloid cells in NSCLC patients with lung nodules.

15. SUBJECT TERMS

None provided

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			USAMRMC
Unclassified	Unclassified	Unclassified	Unclassified	21	19b. TELEPHONE NUMBER *(include area code)*

Table of Contents

Page

1. **INTRODUCTION:**

A recent screening trial showed that the use of low dose computed tomography (LDCT) resulted in a 20% reduction in lung cancer mortality, however there was a 96% false positive rate associated with LDCT. Thus, there is an immediate clinical need to develop a diagnostic biomarker that would select patients with CT detected nodules for further testing. The ease with which blood can be sampled makes it a logical choice in which to discover diagnostic biomarkers, however the clinical utility of tumor derived proteins, miRNA or circulating tumor cells as blood-based biomarkers has been limited. In this proposal, instead of tumor-derived biomarkers, we will focus on host response to tumor growth. It has been well documented that tumor growth systemically stimulates and mobilizes BM-derived hematopoietic cells to the tumor bed to establish a permissive microenvironment. Preliminary studies in our lab have shown that in lung cancer patients, the circulating myeloid cells are transcriptionally altered and the alteration is tumor dependent. The specific transcriptomic signature of circulating myeloid cells may provide us unique resources for lung cancer biomarker discovery. Therefore, we proposed to identify a NSCLC-dependent transcriptomic signature in circulating myeloid cells and then validate the diagnostic value of the specific gene signatures of circulating myeloid cells in NSCLC patients. The proposed study, if succeed, will provide novel strategies and approaches for early detection of lung cancer.

2. **KEYWORDS:**

None small cell lung cancer (NSCLC), biomarker, circulating myeloid cells, flow cytometry, RNA-sequencing, expression profiling.

3. **ACCOMPLISHMENTS:**

- **What were the major goals of the project?**

Specific Aim 1: To identify a NSCLC-dependent transcriptomic signature in circulating myeloid cells. (Proposed to be accomplished during the first year)

Major Task 1. Lung cancer signature gene optimization

Subtask 1: Patient recruitment including pre- and post- surgery patients, and COPD patients

Subtask 2: Flow cytometry sorting of circulating myeloid cells.

Subtask 3: RNA-Sequencing

Subtask 4: RNA-seq data analysis

Subtask 5: Feasible RT-PCR array assay development

Specific Aim 2: To validate the diagnostic value of the specific gene signatures of circulating myeloid cells in patients with lung nodules. (Proposed to be accomplished during the second year)

Major Task 2: Lung cancer signature diagnostic value validation

Subtask 1: Recruitment of patients with positive lung nodules by CT-Scan

Subtask 2: Flow cytometry sorting of circulating myeloid cells

Subtask 3: RT-PCR array and data analysis with clinical outcomes

- **What was accomplished under these goals?**

For this Final report, we would like summarize the main achievements first. We accomplished the patient recruitment for pre- and post- surgical blood collection, flow cytometry sorting of circulating myeloid cells (CD11b+CD33+ cells). We performed RNA-sequencing analysis of the samples aiming to identify reliable biomarkers for early lung cancer detection. We have been working closely with our bioinformatics collaborators during the No-Cost-Extension period of one year. However, we still encountered challenges of batch differences, patient gender differences, great variations in identified genes expressions even within sorted subpopulations. The biological variations between NSCLC patients and benign controls indicate the necessity of establishing a BioBank with large cohort of patients that requires continuous working of long period of time. Another possible scenario is that clinic undetectable residual disease after surgical removal of the primary lung cancer may affect the circulating myeloid cells as well. Indeed, the clinic following of the NSCLC patients found that an increase of ratio between CD8+ T cells and CD33+ myeloid cells after surgery had a positive correlation with patient survival.

Major Task 1. Lung cancer signature gene optimization

Subtask 1: Patient recruitment including pre- and post- surgery patients, and COPD patients

We recruited 34 NSCLC patients and collected their peripheral blood. Of these patients, there are 11 patients missed the collection of post-surgical blood collection; 3 patients were pathologically diagnosed with squamous cell carcinoma (SCC). Totally we obtained 23 paired samples (both pre- and post-surgical blood samples) from patients with stage I-III adenocarcinoma cell (ACC) tumors. The patient demographics are shown in Table 1. In addition, peripheral blood was also collected from 6 patients with benign nodules to serve as non-tumor control group.

All blood samples were sorted via flow cytometry into IMMCs and polymorphonuclear neutrophils. An unfractionated whole white blood cells aliquot was also retained. Total RNA was extracted and RNA sequencing (poly-A selected, single-read, 51 bp, 6 samples per lane) was performed using an Illumina sequencer.

Table 1. Demographics of NSCLC patients

NSCLC	(*n* = 23)
Age (y)	
Average	70
Max	85
Min	49
Gender	
Female	16
Male	7
Tobacco use	
Current	3
Former	18
Never	2
Cancer stage	
Stage I	18
Stage II	2
Stage III	3
Surgery to post-collection	
1-2 mo	16
2-4 mo	5
4-7 mo	2

Subtask 2: Flow cytometry sorting of circulating myeloid cells.

With the peripheral blood, we have performed cytometry analysis of the CD11b+CD33- neutrophils, CD11b+CD33+ monocytic myeloid cells and CD8+ cytotoxic T cells (Figure 1A). Consistently, we found no significant difference between the pre- and post- surgical samples and

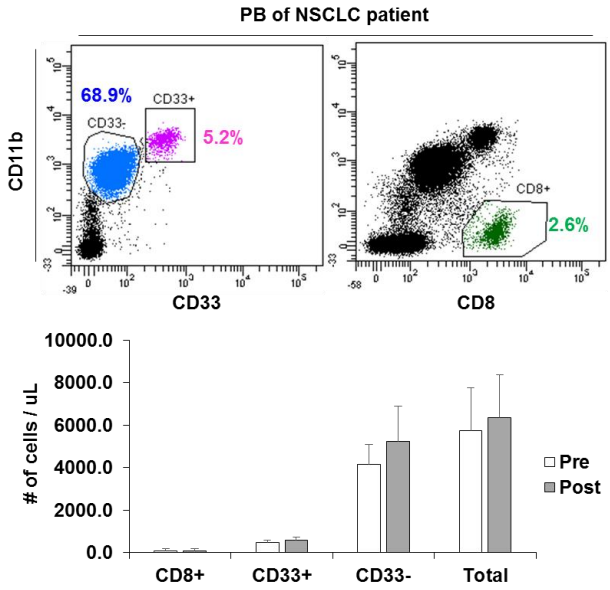

Figure 1. Flow cytometry analysis of circulating myeloid cells and cytoxic T cells in the peripheral blood (PB) of NSCLC patients. A. Flow cytometry plot showing the detection of CD11b+CD33- (neutrophils), CD11b+CD33+ (monocytic myeloid cells) and CD8+ T cells in the peripheral blood analysis. B. Number of subpopulations in the PB of pre- and post-surgical samples of NSCLC patients (n=17).

the absolute number and the percentage of these subpopulations showed broad variations between patients (Figure 1B).

Next, we sorted the myeloid subpopulations (CD11b+CD33- cells and CD11b+CD33+ cells) by flow cytometry. The purity of cells after sorting was confirmed by re-analyzing the post-sorting samples (Figure 2).

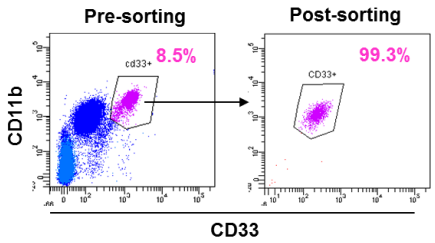

Figure 2. The purity of CD11b+CD33+ myeloid cells in the pre- and post-sorting samples.

Subtask 3: RNA-Sequencing

We extracted RNA from sorted cells using the mirVana kit (Life Technologies). Using the TruSeq RNA sample preparation kit (Illumina, Inc) cDNA libraries was constructed. We performed 51bp single read with HiSeq machines in the Genome Sequencing Facility at WCMC. Short reads (after FastQC quality control) were mapped to hg19 using TopHat and expression levels quantified using CuffLinks. Gene expression level (FPKM) was determined using DEseq and LIMMA. We applied the RNA-seq by single-read for 51 cycles and pooled 6 samples per lane. This strategy has given us reliable sequence-reading with deep enough coverage of the transcriptome.

Subtask 4: RNA-seq data analysis.

As reported previously, we encountered difficulties in the bioinformatics analysis of the RNA-seq data. Therefore, we required expert help from the Applied Bioinformatics Core (ABC) at Weill Cornell Medicine. With the RNA-seq data obtained from 23 NSCLC pateins with malignant tumors before and after surgical resection of the tumor, and 6 patients with benign nodules, we aimed to identify tumor specific transcriptomes of circulating myeloid cells. The final analysis report was shown as follows:

Differential gene expression analysis of circulating myeloid cells in NSCLC

Prepared for Ding Cheng Gao, Weill Cornell Medicine
by the Applied Bioinformatics Core (ABC), Weill Cornell Medicine

1 Background

Previous studies have established a critical function of myeloid cells in non-small cell lung cancer (NSCLC) progression: it has been hypothesized that tumor cell paracrine signaling systemically stimulates and mobilizes myeloid cells from bone marrow compartments to the tumor bed, where they are reprogrammed to promote tumor growth by stimulating tumor angiogenesis, suppressing tumor immunity, and promoting metastasis to distant sites. Two sub-populations of myeloid cells have been implicated: immature monocytic myeloid cells (IMMC) and the polymorphonuclear neutrophils. It has been hypothesized that the altered gene expression of those sub-populations as they traffic through the circulatory system might include a specific gene signature indicative of the malignant nature of the tumor.

1.1 Sample descriptions

Peripheral blood was collected from 23 NSCLC patients with malignant tumors before and after surgical resection of the tumor. Post-surgery blood samples were drawn at ≥ 3 weeks after resection to eliminate the potential effects of anesthesia and surgery. Peripheral blood was also collected from six patients with benign nodules to serve as an additional control group. All blood samples were sorted via flow cytometry into IMMCs and polymorphonuclear neutrophils. An unfractionated whole white blood cells aliquot was also retained. Total RNA was extracted and RNA sequencing (poly-A selected, single-read, 51 bp, 6 samples per lane) was performed using an Illumina sequencer.

1.2 Objective

The ABC aimed to determine tumor-dependent differences in the transcriptomes of circulating myeloid cells.

2 Pre-processing

The sequences were aligned to the human reference genome (GRCh38p3) using `STAR v2.4.2a`, a universal RNA-seq aligner [2]. To improve accuracy of the mapping, the genome was created with a splice junction database based on Gencode v23 basic annotation[3]. Sequences that mapped to more than one locus were excluded from downstream analysis[1].

Prior to the detection of differentially expressed genes, the quality of the sequences was assessed based on several metrics using `FastQC v0.11.3`[1] and `QoRTs v0.3.5`[4]. All samples passed the numerous quality control checks (Table 1). A more detailed explanation of the quality control assessment can be found in Appendix A.

3 Differential Gene Expression

Since the sequence data was determined to be of good quality (Appendix A), several downstream analyses, including differential gene expression analysis, were performed.

[1]Reads that map equally well to more than one location cannot be confidently assigned. Most alignment programs randomly assign these reads to a single random location, leading to irreproducibility.

Table 1: Summary of the QC assessment. For details, see Appendix A.

QC metric	Description	Pass?
Adapter contamination	Percentage of reads mapping to adapter sequences	✓
Per base sequence quality	Median Phred quality score across length of read	✓
Per sequence GC content	Frequency of reads with different proportions of GC content	✓
Mapping stats	Percent and number of reads aligning to reference genome	✓
Mapping locations	Percentage of reads mapping to different genomic categories	✓
Gene-body coverage	Evenness of coverage across exonic regions	✓
Cumulative gene diversity	Measure of library complexity	✓

3.1 Counting the number of reads per gene

Uniquely mapped sequences were intersected with composite gene models from Gencode v23 basic annotation using `featureCounts v1.4.6-p4`, a tool for assigning sequence reads to genomic features[6]. Composite gene models for each gene consisted of the union of exons of all transcript isoforms of that gene. Uniquely mapped reads that unambiguously overlapped with no more than one Gencode composite gene model were counted for that gene model; the remaining reads were discarded. The counts for each gene model correspond to gene expression values, and were used for differential gene expression analysis.

3.2 Hierarchical clustering on all cell types

The pairwise relationship between all samples based on read counts per gene was assessed using the Simple Error Ratio Estimate (SERE), a single-parameter test procedure designed specifically for RNA-seq count data to quantify global sample differences[12]. Using all of the genes in the Gencode v23 basic annotation set, this technique determined the overall similarity between the transcriptomes of the different cell types. The samples clustered according to their cell type, with the exception of four samples, indicating cell-specific expression patterns for neutrophils, monocytes, and the unfractionated aliquot (Figure 1). Within a cell type, there was little to no differentiation between pre- and post- surgical resection or between pre/post and benign samples. The relationship between samples within individual cell types was further explored using principal component analysis plots, described below.

3.3 Principal components analysis on each cell type

Principal components analysis (PCA) was performed on each cell type individually to visualize sample-to-sample distances and to test if the samples could be separated based on the presence or absence of tumor malignancy (pre-surgical resection versus post-/benign samples). The raw gene expression values were transformed using a regularized-logarithm transformation (with the 'rlog' function from the `DESeq2 v1.20.0` package[7]). The transformation was necessary since PCA works best for data which generally has the same range of variance at different ranges of the mean values (i.e., when the data is homoskedastic), which is not the case for RNA-seq data. Rlog-transformed data becomes approximately homoskedastic, which can then be used directly for computing distances between samples and making PCA plots. The top 500 genes showing the highest variance were selected, and the principal components were computed and plotted.

Within each cell type, there was a clear and distinct separation between two sub-populations. Meta-data received post hoc revealed the two sub-populations to be linked with the biological sex of the samples, indicating that the variation resulting from biological sex is larger than the variation resulting from differences in presence or absence of tumor (see Figure 2 for the PCA plot on the monocyte samples. Additional cell types can be found as supplemental figures).

To further assess sample relationships, additional PCA plots were created examining male and female samples separately for each cell type. There was no discernible separation between pre- and post-surgical resection and between pre/post and benign samples when examining either male or female samples alone (Figures 3 - 4; additional cell types can be found as supplemental figures.).

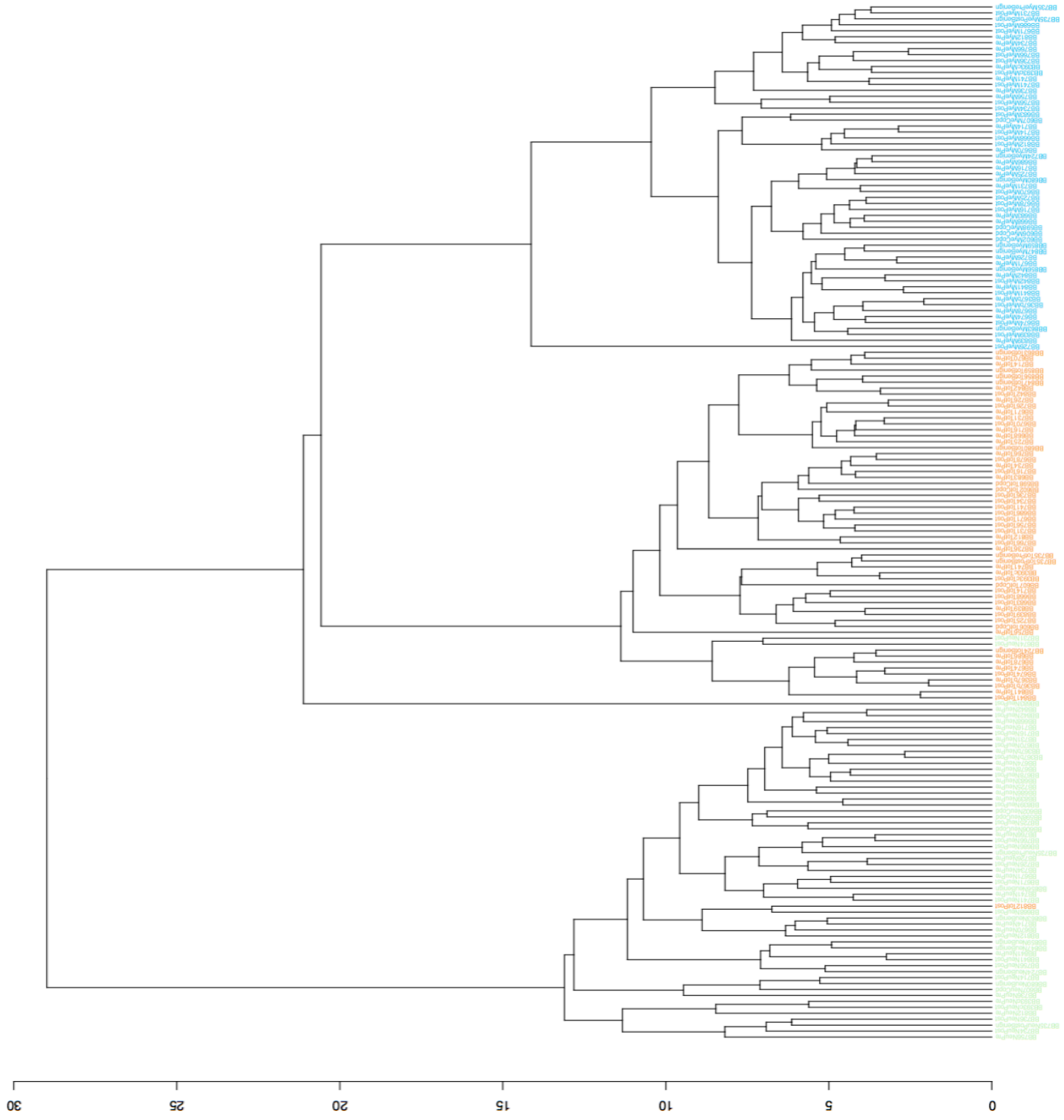

Figure 1: Hierarchical clustering on all cell types. The pairwise relationship between all samples based on read counts per gene was assessed using the Simple Error Ratio Estimate (SERE). The samples clustered according to their cell type (green: neutrophils; orange: unfractionated; blue: monocytes), with the exception of four samples.

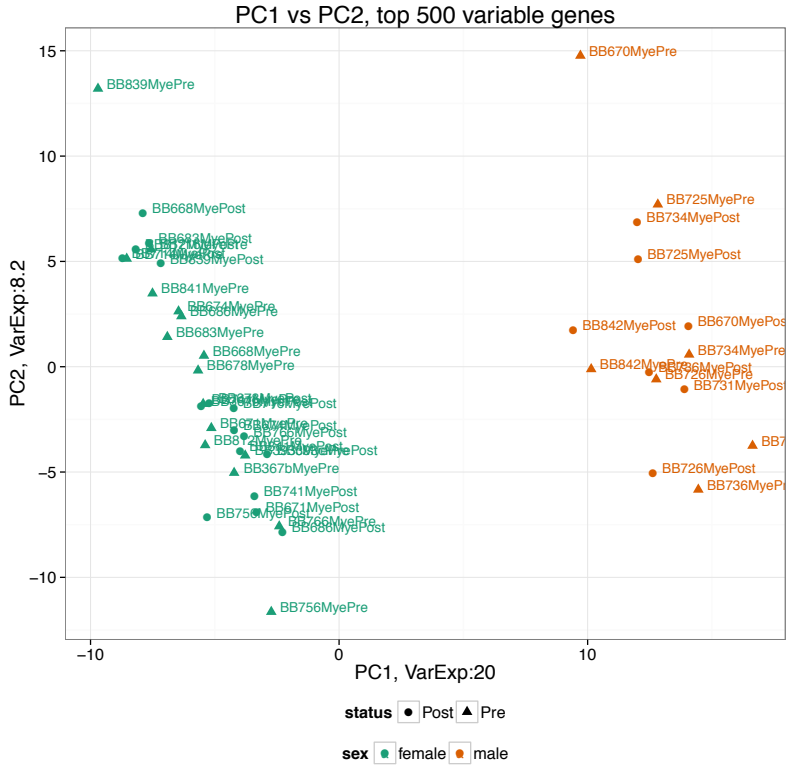

Figure 2: Principal component analysis of monocyte samples. The majority of the variation can be explained by differences due to biological sex (male samples shown in blue; female samples shown in pink).

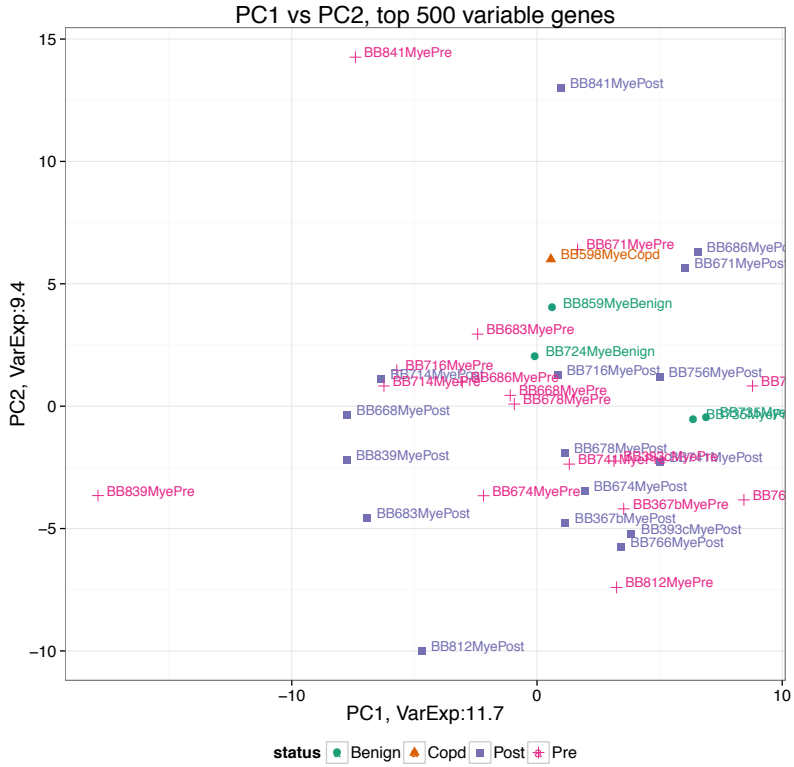

Figure 3: Principal component analysis of male-only monocyte samples. There is no discernible separation between pre- and post-surgical resection (blue, pink) or between pre/post and benign samples (turquoise, orange).

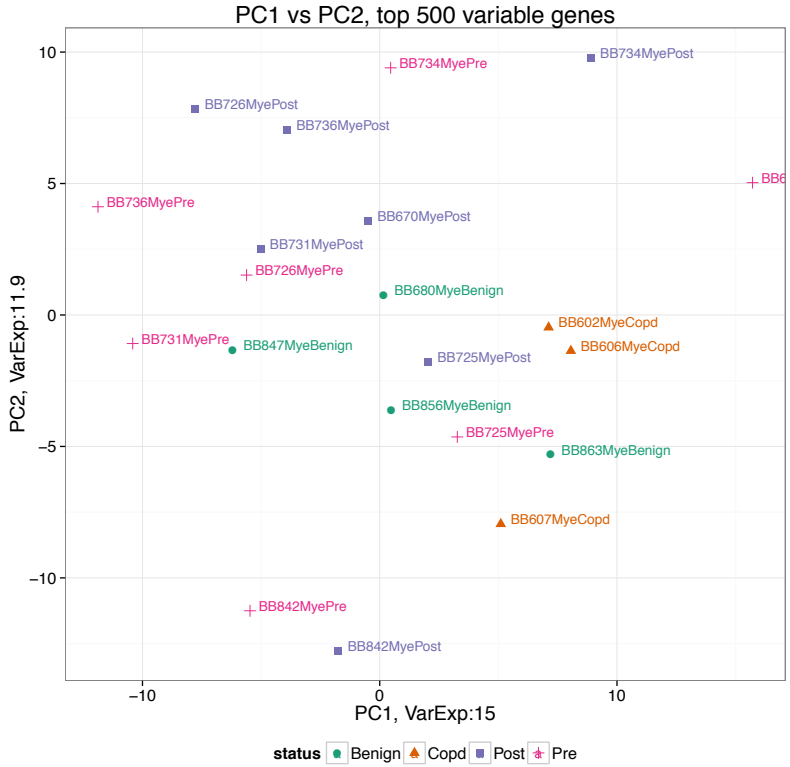

Figure 4: Principal component analysis of female-only monocyte samples. There is no discernible separation between pre- and post-surgical resection or between pre/post and benign samples.

3.4 Differential gene expression analysis

Based on the available sample types (see Section 1.1), the following hypotheses were tested independently for each cell type:

1. Removal of malignant lung tumors alters the transcriptome of circulating myeloid cells.
2. The transcriptome of circulating myeloid cells from patients whose malignant lung tumors were removed resembles the transcriptome of circulating myeloid cells from patients with benign nodules in the lung.
3. The transcriptome of circulating myeloid cells from patients with benign lung nodules differs from the transcriptome of circulating myeloid cells from patients with malignant tumors.

To address hypothesis #1, the pre-resection samples were compared to the post-resection samples using a paired design to account for differences between patients, including biological sex.

To address hypotheses #2 and #3, the benign samples were compared to the pre- and post-resection samples respectively. Because the samples in these comparisons were not paired, patient-specific differences may occlude the effect of the tumor presence.

Differential gene expression analysis was performed for each comparison using three different tools with default normalization and parameters: `limma v3.24.14`[9], `edgeR v3.10.2`[10], and `DESeq2 v1.20.0`[7]. Initially, only genes with adjusted p-values <0.05 were considered as differentially expressed. Table 2 lists the numbers of differentially expressed genes detected by each tool.

Because there were few or zero differentially expressed genes detected in any of the contrasts for any of the cell types, the analysis was repeated allowing for an adjusted p-value as high as 0.20 (Table 3). This high p-value cutoff had been used by a previous collaborator, probably because the standard threshold yieldd no result. Differentially expressed genes at such high thresholds are unlikely to be reproducible.

Increasing the adjusted p-value cutoff to 0.20 did not appreciably change the results for most contrasts and cell types. Moreover, for the contrasts and cell types that did change, there was significant disparity in the total number of differentially expressed genes across the three DE tools. Additionally, `limma`, which is typically the most conservative of the tools, turned out to be the most liberal. As a result, little confidence can be placed in the list of genes that were returned as differentially expressed.

Table 2: Numbers of differentially expressed genes detected (adj. p < 0.05) with three separate tools. The order of results is unfractionated, monocytes, and neutrophils separated by a vertical bar '|'.

hypothesis	contrast	limma	edgeR	DESeq2
1	pre vs. post	0|0|0	3|0|1	1|2|0
2	pre vs. benign	0|0|0	0|0|0	2|12|2
3	post vs. benign	0|7|2	0|0|1	10|20|13

Table 3: Numbers of differentially expressed genes detected (adj. p < 0.20) with three separate tools. The order of results is unfractionated, monocytes, and neutrophils separated by a vertical bar '|'.

hypothesis	contrast	limma	edgeR	DESeq2
1	pre vs. post	0|0|6,921	48|40|6	126|43|3,414
2	pre vs. benign	0|0|0	0|0|0	5|12|4
3	post vs. benign	0|7|3	0|27|5	15|45|748

DE analysis to determine sex specific genes

Because the PCA plots showed a clear and distinct separation between male and female patients, we used this comparison as a proof of principle that sex specific DE genes could be identified using our methodology (using `limma`, `edgeR`, and `DESeq2`) by contrasting males versus females. Using an adjusted p-value < 0.05, all three tools were able to detect differentially expressed genes between males and females (Table 4). Table 5 displays 25 sex specific DE genes that were common to all cell types.

Table 4: Numbers of differentially expressed genes detected (adj. p < 0.05) with three separate tools. The order of results is unfractionated, monocytes, and neutrophils separated by a vertical bar '|'.

contrast	limma	edgeR	DESeq2
male vs. female	50|45|23	87|88|49	100|125|80

Table 5: Table of sex specific genes detected in all three cell types. The last seven genes (XIST, MAP7D2, TSIX, HSPA7, FCGR2C, PRSS21, NLRP2) are expressed more highly in females than in males.

KDM5D	LINC00278	RPS4Y1	ANOS2P	UTY
RP11-424G14.1	ZFY	TTTY15	PRKY	DDX3Y
USP9Y	TXLNGY	EIF1AY	TMSB4Y	RP11-632C17
RP5-857K21.7	JUP	BCORP1	XIST	MAP7D2
TSIX	HSPA7	FCGR2C	PRSS21	NLRP2

3.5 Comparison to published results

A previously published study[5] examined RNA expression levels in peripheral blood cells between paired samples collected from NSCLC patients before and after tumor removal using Illumina gene expression microarrays. Using differential gene expression analysis and support vector machines with recursive feature elimination, Kossenkov et al. published a list of the 50 highest-ranked genes which were different before and after surgical resection, and which were sufficient to separate pre- and post-surgery samples with 100% accuracy.

Before determining whether or not that list of 50 genes was capable of separating the NSCLC samples in this study, we attempted to reproduce the findings of Kossenkov et al. using the methods employed in Section 3.4. Accordingly, the ABC downloaded the microarray data from GEO (GSE13255) and differential gene expression analysis was performed using `limma`. `limma` detected 130 differentially expressed genes (adj. p < 0.05), which included at least 60% of genes which were on the published gene list[2]. This list was also capable of separating the published pre- and post-surgery samples with 100% accuracy (Figure 5).

Since we could recapitulate Kossenkov et al.'s analysis using their data, we decided to check to see if their gene list was sufficient to separate the pre- and post- surgical resection samples from this study. As shown in Figure 6,

[2]There may be a higher overlap than noted due to technical difficulties in converting Illumina microarray probe IDs into gene symbols.

their gene list was incapable of separating pre- and post- surgical resection samples in any of the cell types. Note that this does not necessarily imply that there is a problem with the samples studied here; on the contrary, it might mean that the gene list identified using the Kossenkov data set is idiosyncratic to that data set, and not necessarily applicable to NSCLC samples in general.

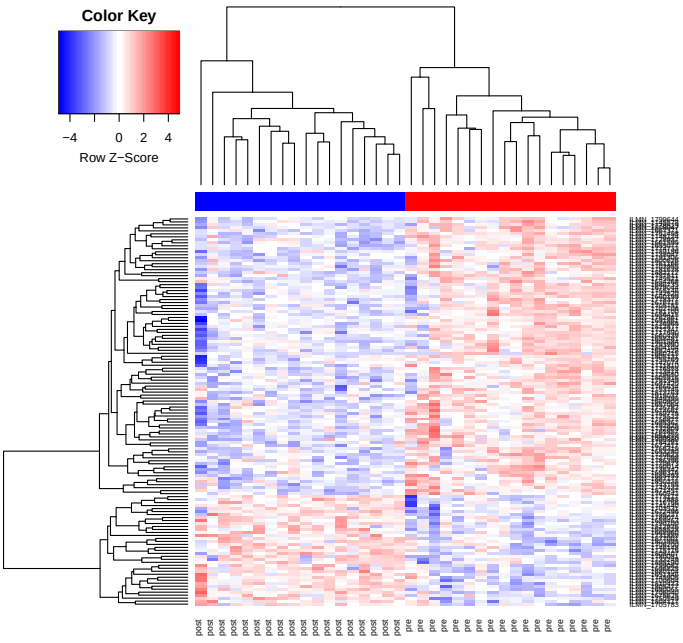

Figure 5: Heatmap of differentially detected genes on data from Kossenkov et al. The genes detected as DE using `limma v3.24.14` were sufficient to separate their published pre- and post-surgery samples with 100% accuracy.

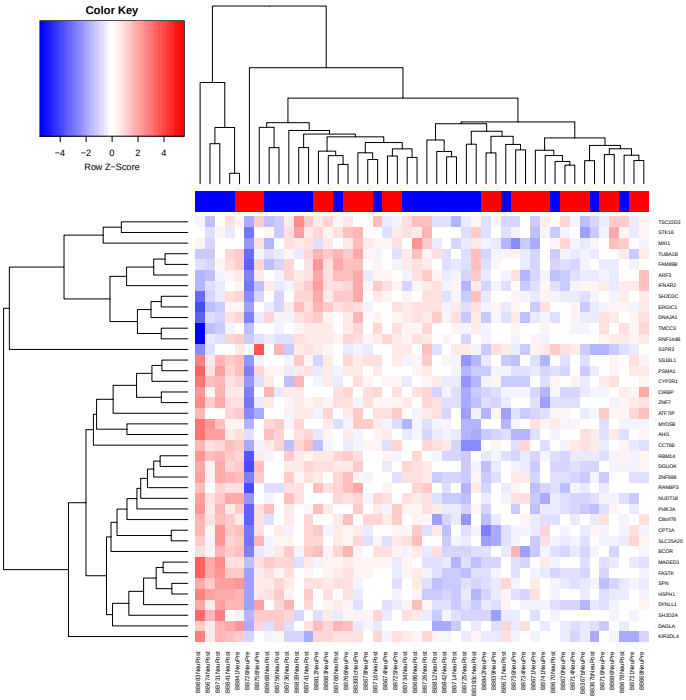

Figure 6: Heatmap of 50 genes published by Kossenkov et al. with gene expression values from the the monocyte cell population studied here. The genes do not distinguish between pre- and post-surgical resection. Additional related figures can be found in Supplemental Materials.

3.6 Coefficients of variation

Since classic DE analysis and clustering did not distinguish between pre/post-surgery samples we tested whether looking at changes in the coefficients of variation may separate the sample types **??**. Coefficients of variation (CVs; σ/μ) were calculated for each gene with sufficient expression (defined as at least 1 count-per-million in at least 3 samples) across all samples for each cell type. Because the standard deviation of log-expression is approximately the same as the CV of unlogged expression, the CV was calculated by taking the standard deviation of the log of the counts-per-million of each gene. Coefficients of variation were plotted to detect differences between pre- and post-surgical resection samples (Figure 7). The differentiation of CV(pre)-CV(post) and CV(post)-CV(pre) was used to identify potentially relevant genes in the pre-vs-post pairs. Across all cell types, there were nearly zero genes with high expression variation in the tumor (pre-surgery) samples. Genes highlighted in red are attached to this report.

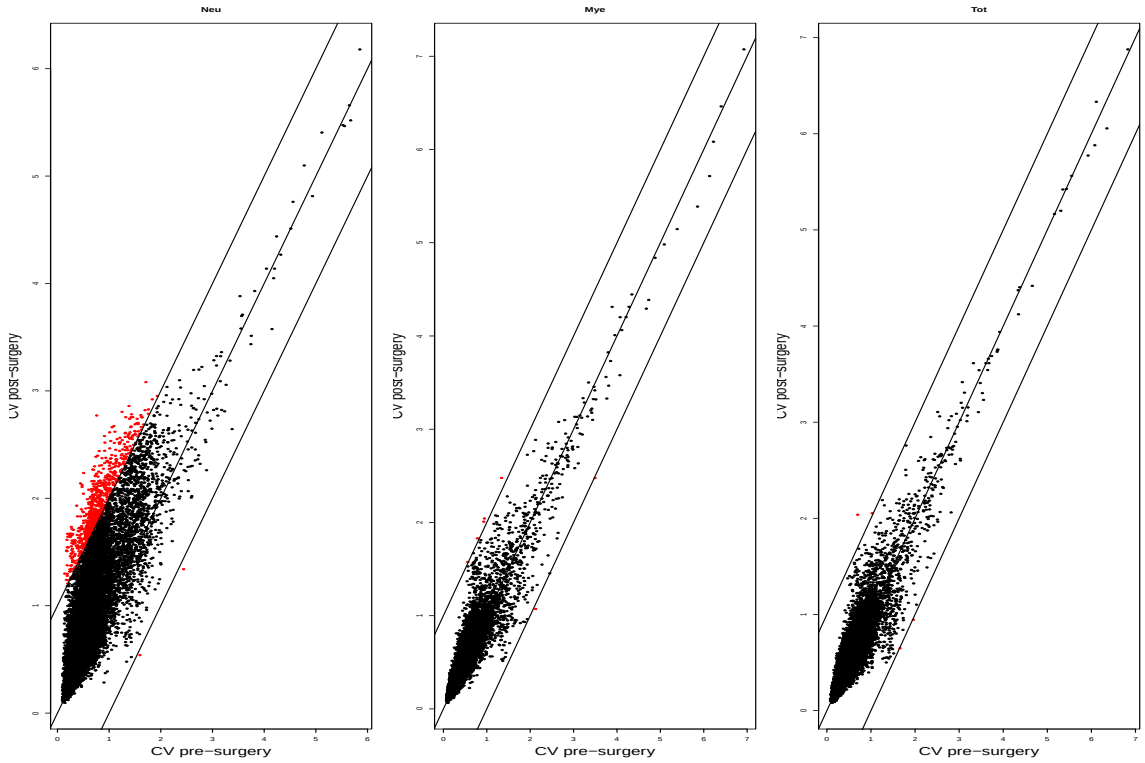

Figure 7: Expression variation in pre- and post-surgical resection for each cell type. Scatterplots represent coefficients of variation (CVs) for tumor and matched post surgical resection. Genes with CV > 1 in both pre and post samples are colored red.

3.7 Gene set enrichment analysis on neutrophils

Upon request, gene set enrichment analysis was run on the neutrophil data set using the results from `limma` (adj. $p < 0.20$) on the pre vs. post contrast with the Broad's GSEA software. The results from `limma` were used mainly because `limma` is highly regarded in the literature. Monocytes and the unfractionated total were excluded from this analysis because they failed to yield appreciable lists of differentially expressed genes, even with an adj. $p < 0.20$. A pre-ranked list of genes was supplied to the GSEA by taking the fold change of each gene multiplied by the inverse of the respective p-value[8]. The complete Molecular Signatures Database (MSigDB) database, a collection of annotated gene sets, as well as the Gene Ontology (GO) subset, was used for gene grouping and enrichment testing. Figures 8-9 visualizes the top 10 enriched gene sets in either direction based on MSigDB or GO testing, respectively. Notably, the enrichment profiles of the enriched gene sets from the GO collection are 'bumpy'; this is because the genes that are contributing to the enrichment profiles appear further down on the ranked list and are therefore less reliable. See Figures 10-11 for comparison between an enrichment profile from the MSigDB collection versus the GO collection.

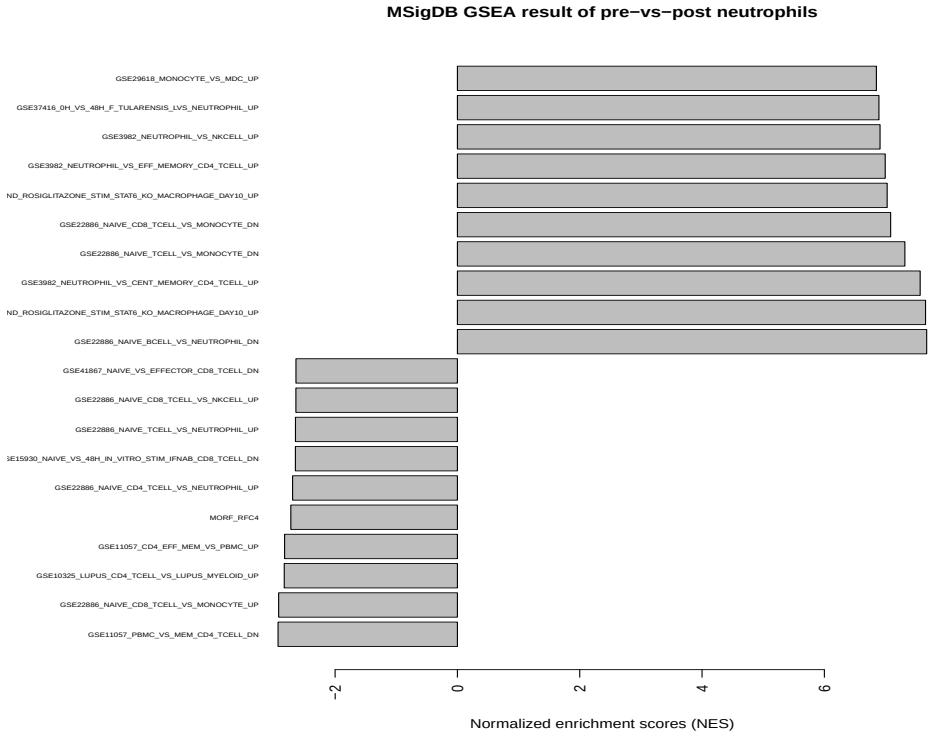

Figure 8: MSigDB GSEA result of pre- vs. post-surgical resection on neutrophil cells. Top 10 gene sets in either direction are shown.

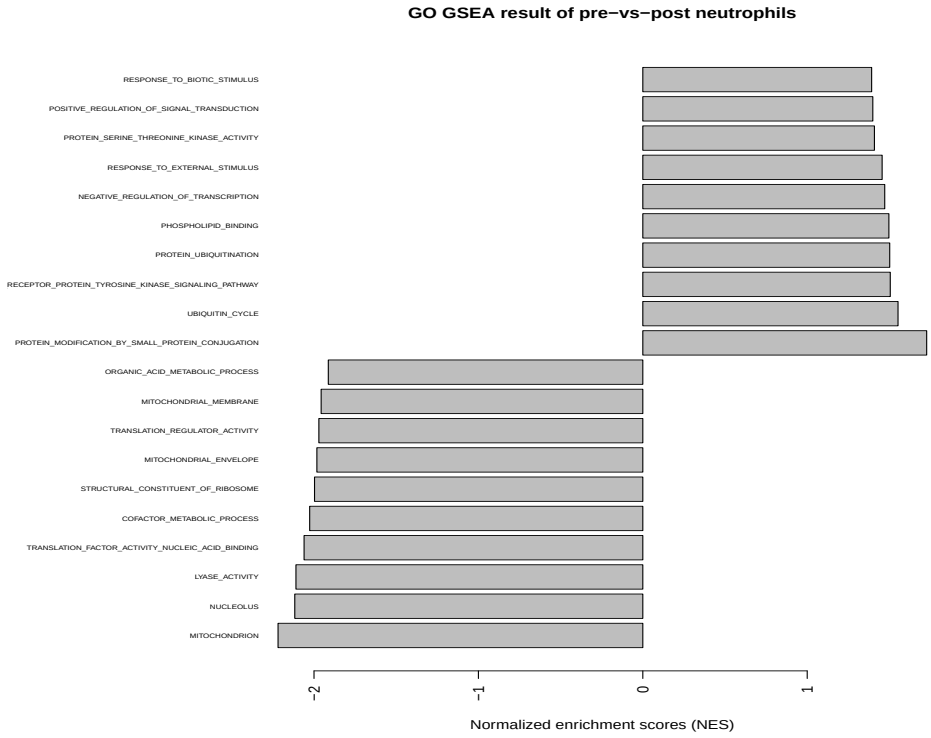

Figure 9: MSigDB GSEA result of pre- vs. post-surgical resection on neutrophil cells. Top 10 gene sets in either direction are shown.

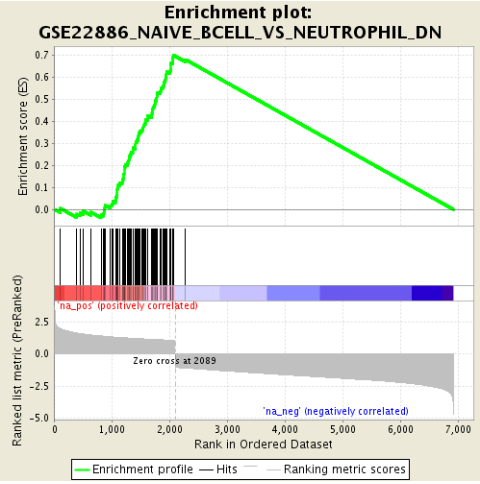

Figure 10: Enrichment profile from MSigDB collection.

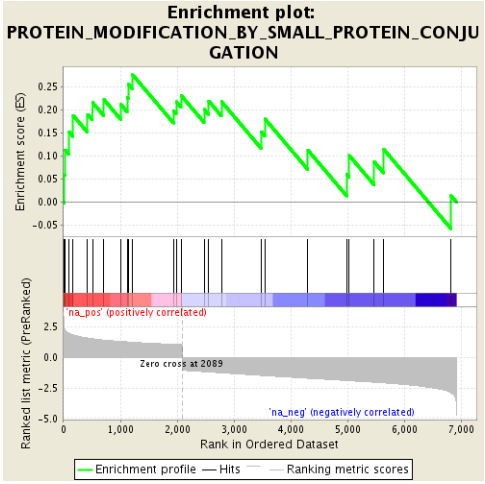

Figure 11: Enrichment profile from GO collection.

4 Summary

The ABC performed quality control, processing, and differential gene expression analysis to determine tumor-dependent differences in the transcriptomes of circulating myeloid cells using RNA-seq from monocytes, neutrophils, and an unfractionated aliquot. Prior to differential gene expression analysis, multiple quality metrics were assessed. All metrics consistently indicated that the data was of high quality. For each cell type, differential gene expression analysis was then performed using three distinct tools (`limma`, `edgeR`, and `DEseq2`) with two adjusted p-value cutoffs (0.05 and 0.20) on the following contrasts: 1) pre vs. post surgery, 2) pre surgery vs. benign, and 3) post surgery vs. benign. Regardless of the p-value cutoff employed, few genes were found to be differentially expressed in any of the cell types for any of the contrasts, except when the adjusted p-value was raised to 0.20 for the pre vs. post contrast in the neutrophil cells. In that case, however, the three DE tools identified vastly different numbers of DE genes which rouses doubts about the reproducibility and merit of these gene lists. To ensure that our method of identifying DE genes was performing as expected, it was applied to 1) determine sex specific differences between males and females, and 2) on a public dataset consisting of microarray expression data from NSCLC patients before and after surgical tumor removal. In the former case, genes known to be sex specific (e.g., located on the Y chromosome) were successfully detected (adj. $p < 0.05$); in the latter case, 130 differentially expressed genes were detected (adj. $p < 0.05$), which were sufficient to separate patients before and after surgical tumor removal with 100% accuracy. Because DE analysis was not particularly fruitful, the coefficients of variation between pre and post samples were contrasted for each cell type in order to detect possibly relevant differences between the two sample types. Regardless of the cell type, there were nearly zero genes with high expression variation in the tumor samples, although the neutrophil post-surgery samples showed some genes with greater than expected variation. Lastly, gene set enrichment analysis was performed on the pre vs. post tumor removal samples from the neutrophil cell type using the results from `limma` (adj. $p < 0.20$). Although sets of genes were determined to be enriched in pre- or post-tumor removal, because of the disparity between the three DE tools as noted above, the results are unlikely to be reliable.

References

[1] Andrews S, 2010. URL `http://www.bioinformatics.babraham.ac.uk/projects/fastqc/`.

[2] Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, and Gingeras TR. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**(1):15–21, 2013. doi:10.1093/bioinformatics/bts635.

[3] Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, Barnes I, Bignell A, Boychenko V, Hunt T, Kay M, Mukherjee G, Rajan J, Despacio-Reyes G, Saunders G, Steward C, Harte R, Lin M, Howald C, Tanzer A, Derrien T, Chrast J, Walters N, Balasubramanian S, Pei B, Tress M, Rodriguez JM, Ezkurdia I, van Baren J, Brent M, Haussler D, Kellis M, Valencia A, Reymond A, Gerstein M, Guigó R, and Hubbard TJ. GENCODE: The reference human genome annotation for the ENCODE Project. *Genome Research*, **22**(9):1760–1774, 2012. doi:10.1101/gr.135350.111.

[4] Hartley SW and Mullikin JC. QoRTs: a comprehensive toolset for quality control and data processing of RNA-Seq experiments. *BMC Bioinformatics*, **16**:224, 2015. doi:10.1186/s12859-015-0670-5.

[5] Kossenkov A, Vachani A, Chang C, Nichols C, Billouin S, Horng W, Rom W, Albelda S, Showe M, and Showe L. Resection of Non-Small Cell Lung Cancers Reverses Tumor-Induced Gene Expression Changes in the Peripheral Immune System. *Clin Cancer Res*, **17**(18):5867–77, 2011. doi:10.1158/1078-0432.CCR-11-0737.

[6] Liao Y, Smyth GK, and Shi W. FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, **30**(7):923–930, 2014. doi:10.1093/bioinformatics/btt656.

[7] Love MI, Huber W, and Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, **15**:550, 2014. doi:10.1186/s13059-014-0550-8.

[8] Plaisier S, Taschereau R, Wong J, and Graeber T. Rankrank hypergeometric overlap: identification of statistically significant overlap between gene-expression signatures. *Nucl. Acids Res.*, **38**:e169, 2010. doi:10.1093/nar/gkq636.

[9] Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, and Smyth GK. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, **43**(7):e47, 2015. doi:10.1093/nar/gkv007.

[10] Robinson MD, McCarthy DJ, and Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**:139–140, 2010. doi:10.1093/bioinformatics/btp616.

[11] Romiguier J, Ranwez V, Douzery EJ, and Galtier N. Contrasting GC-content dynamics across 33 mammalian genomes: Relationship with life-history traits and chromosome sizes. *Genome Res.*, **20**(8):1001–1009, 2010. doi:10.1101/gr.104372.109.

[12] Schulze S, Kanwar R, Glzenleuchter M, Therneau T, and Beutler A. SERE: single-parameter quality control and sample comparison for RNA-Seq. *BMC Genomics*, **13**:524, 2012. doi:10.1186/1471-2164-13-524.

[13] The ENCODE Consortium, 2011. URL `http://genome.ucsc.edu/ENCODE/protocols/dataStandards/`.

[14] Zhao W, He X, Hoadley K, Parker J, Hayes D, and Perou C. Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling. *BMC Genomics*, **15**:15, 2014. doi:10.1186/1471-2164-15-419.

Major Task 2: Lung cancer signature diagnostic value validation

Since no reliable biomarkers were identified from the RNA-seq data, we decided to terminate the further pursuing of specific transcriptome markers from the circulating myeloid cells. This might due to the following reasons:

1. The high variation of gene expression in the circulating myeloid cells in different cancer patients. Tumor-bearing is only one of factors that may affect the transcriptome of the circulating myeloid cells. Other factors such as patient gender and general infectious condition will also alter the transcriptome signature of circulating myeloid cells, which challenge the identification of tumor-specific signature.
2. System batch errors in sample preparation should be taken into considerations. Samples in the same batch gave positive outcomes. However, the vastly different numbers and the lack of overlaps of candidate genes signature between batches roused doubts about the reproducibility and merit of these gene lists.
3. Minimum residue tumors might exist in NSCLC patients after surgical removal of the primary tumor. The residue, clinical undetectable disease may continuously affect the circulating myeloid cells. This would cause the undifferentiated issues between the pre- and post-surgical comparisons. To address this, we analyzed the changes in the ratio of cytotoxic T cells versus circulating myeloid cells (CD8+/CD11b+CD33+ cells) after surgery, which will indicate the immune responses in the patient. Clearly, these patients could be divided into three groups: 1) the ratio was higher in the pre-surgical blood sample than that in the post-surgical samples; 2) the ratio was increased in the post-surgical sample; 3) the ratio was staying low in both pre- and post-surgical samples (Figure 3). It is still an open question whether such changes could predict the residue disease, while we are still following the clinic outcomes of these patients.

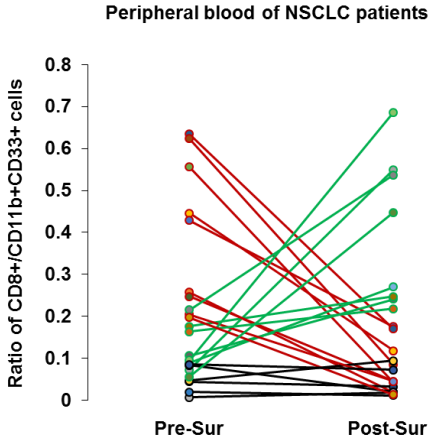

Figure 3. The ratio of cytotoxic T cells versus CD33+ myeloid cells in the peripheral blood of NSCLC patients before and after surgical removal of the primary tumor. Numbers of subpopulations of blood cells were analyzed by flow cytometry (n=24). Red lines indicate a decline of the ratio, Green lines indicate an increase of the ratio, and the black lines indicate relatively no significant changes in the ratio.

- **What opportunities for training and professional development has the project provided?**

  Nothing to Report.

- **How were the results disseminated to communities of interest?**

  Nothing to Report.

- **What do you plan to do during the next reporting period to accomplish the goals?**

  Nothing to Report.

4. **IMPACT:**

- **What was the impact on the development of the principal discipline(s) of the project?**

The persistent poor survival of lung cancer patients is largely attributable to the late stage at diagnosis. New biomarkers for early detection are urgently required in the clinic. However, discovery of biomarkers using peripheral blood is challenging because tumor-specific markers are usually expressed in low concentrations, diluted in a milieu of other abundant proteins and likely to be missed. We proposed to overcome this hurdle, instead of focusing on tumor-derived biomarkers, we will analyze the host responses to the tumor growth. The abundance of circulating myeloid cells, which we know play important roles in tumor growth, may provide a unique source for novel NSCLC biomarker discovery.

The overall outcome of the project is discouraging for future pursuing NSCLC biomarkers with the transcriptome of circulating myeloid cells. The signatures in these cells show great variations and may highly impacted by other conditions such as the patient gender and general infectious conditions other than the primary tumor-bearing itself. Also, the surgery only may not mean the clearance of the tumor. Residue disease may continuously affect the overall transcriptomes of the circulating myeloid cells. Distinguishing patient groups need a large cohorts with long-term clinic outcome data. Continuous efforts will be necessary for such purpose.

- **What was the impact on other disciplines?**

Nothing to Report.

- **What was the impact on technology transfer?**

Nothing to Report.

- **What was the impact on society beyond science and technology?**

Nothing to Report.

5. **CHANGES/PROBLEMS:**

Nothing to Report.

6. **PRODUCTS:**

Nothing to Report.

7. **PARTICIPANTS & OTHER COLLABORATING ORGANIZATIONS**

- **What individuals have worked on the project?**

Name:	*Dingcheng Gao*
Project Role:	*PI*

Researcher Identifier (e.g. ORCID ID):	
Nearest person month worked:	*12*
Contribution to Project:	*Dr. Gao has overseen the ongoing project, performed work in lung cancer biomarker discovery by combining flow cytometry and RNA-sequencing techniques.*
Funding Support:	

Name:	*Nasser Altorki*
Project Role:	*Co-PI*
Researcher Identifier (e.g. ORCID ID):	
Nearest person month worked:	12
Contribution to Project:	*Dr. Altorki has guided the collection of patient samples, cooperate with pathologist and lab members for biobanking management with 5% efforts.*
Funding Support:	

Name:	*Oliver Elemento*
Project Role:	*Co-PI*
Researcher Identifier (e.g. ORCID ID):	
Nearest person month worked:	12
Contribution to Project:	*Dr. Elemento has been in charge of bioinformatics analysis of the project with 3.8% efforts.*
Funding Support:	

Name:	*Cathy Spinelli*
Project Role:	*Clinical coordinator*
Researcher Identifier (e.g. ORCID ID):	
Nearest person month worked:	12
Contribution to Project:	*Ms. Spinelli has supported the collection and biobanking of patient blood samples with 5% efforts.*
Funding Support:	

Has there been a change in the active other support of the PD/PI(s) or senior/key personnel since the last reporting period?

Nothing to Report

What other organizations were involved as partners?

Nothing to Report