UNIVERSITY OF CALIFORNIA, SAN DIEGO

Cell-free RNA Sequencing from Microliters of Unprocessed Serum

A Thesis submitted in partial satisfaction of the requirements of the degree Master of Science

in

Bioengineering

by

Zixu Zhou

Committee in Charge:

Professor Sheng Zhong, Chair
Professor Stephanie I. Fraley
Professor Hui-Chun Irene Su

2017

The Thesis of Zixu Zhou is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____
Chair

University of California, San Diego

2017

Table of Contents

List of Figures

List of Tables

ABSTRACT OF THE THESIS

Cell-free RNA Sequencing from Microliters of Unprocessed Serum

by

Zixu Zhou

Master of Science in Bioengineering

University of California, San Diego, 2017

Professor Sheng Zhong, Chair

Though numerous treatments for breast cancer have been developed, recurrence

still exists, biomarkers and risk assessment tools are thus desired by the field. Cell-free RNA in blood or serum is now receiving more and more interest as a pool of biomarkers and RNA sequencing was believed to be a powerful tool for its analysis. However, development of the field was hindered by large volume of serum required for RNA extraction.

In this study, by circumventing RNA extraction and with the inspiration from single-cell RNA sequencing, we developed a technology being able to construct RNA sequencing libraries in large scale with microliters of direct unprocessed serum input. Optimizations for sequencing and mapping qualities, library complexity and library construction efficiency were conducted by varying serum input volume, adding custom sequencing primer and applying liquid handling instrument epMotion 5075. The optimized strategy "construct 96 different libraries in one single automated batch with 7ul serum input" was proposed accordingly. A huge diversity of genes could be found in serum with this technology. The developed technology was then applied to 96 breast cancer patients' serums for its performance and potential in clinical cases. Based on the sequencing data: 465, 601 and 1259 genes were identified by three methods respectively to behave differently in recurrence and non-recurrence patients; The two populations could be separated by principle component analysis with all three sets of genes mentioned above; Preliminary recurrence risk assessment was conducted successfully with the classifier random forest.

Introduction

Though numerous of treatments for breast cancer have been developed, recurrence of breast cancer still exists, its risk assessment and potential biomarkers are thus desired by the field. With the potential to provide an easily accessible window for monitoring changes not only in breast cancer, but also at distant metastatic sites, blood as a minimally invasive, rapid and cost-effective biopsy source is nowadays receiving more and more clinical interest. In blood/serum exist three major categories of potential biomarker sources circulating tumor cells, cell-free DNA and cell-free RNA (serum is usually used when studying cell-free DNA and RNA). Lots of research has already been conducted on circulating tumor cells and cell-free DNA, while studies on cell-free RNA are still at its early stage with more effort needed. However, though studies on it were just started, cell-free RNA was already shown with great potential being a promising biomarker pool. One of the advantages of cell-free RNA is that it reveals not only sequence information, but also gene expression and regulation details, an important information none of the other two sources feature [1-2]. Some early stage efforts in the field studying cell-free RNA disease biomarkers in serum were spent on their concentrations [3-4], but due to the fact that yield of cell-free RNA extraction was always highly variable [5] and release of RNA into the blood was reported mostly not due to specific conditions [6], limited achievement was made. Meanwhile, another way studying cell-free RNA: high-throughput RNA sequencing (RNA-Seq), a method making full use of cell-free RNA's advantage on expression and regulation details, is getting to the center of the stage [7]. In a word, studies on expression and regulation details of cell-free RNA

in serum with RNA-Seq are now more and more emphasized in the field.

Among numerous studies in the field, RNA-Seq are mostly performed based on libraries constructed with extracted RNA harvested from milliliters of serum. However, this requirement of milliliters of serum volume could potentially limit the development of the field, and RNA-Seq with low serum starting volume is thus needed. With one of the most important reasons why large serum starting volume is necessary being the huge RNA molecule loss during RNA extraction, low starting serum volume RNA-Seq library construction might be possible if this extraction step could be bypassed, namely, if RNA-Seq library construction could be constructed directly from serum. The question now comes as how should one develop such an application. Studies have reported that most of cell-free RNA in serum locates inside exosomes, and these lipid bilayer structure similar to cell membrane prevent or slow cell-free RNA digestion and degradation [6, 8-9]. This structural similarity between exosomes and cells enlightened us to refer to single-cell sequencing technology. One step further, with the ability to construct full-length total RNA library from low amount of RNA input, “Switching Mechanism At the 5’ end of RNA Template (SMART)-Sequencing (SMART-Seq)” [10] widely used in cell RNA-Seq library construction is not only a representative of the technology but also a good candidate for the application. To summarize, the low starting serum volume goal mentioned above and studies on stable existence of cell-free RNA in serum point together to one potential direction: applying single-cell RNA Sequencing technology “SMART-Seq” on low starting volume of unprocessed serum directly to construct RNA-Seq library for downstream sequencing and analysis.

Thus, here, we firstly introduce a novel technology developed based on single-cell RNA-Seq technology “SMART-Seq” to provide a deeper view into human serum using microliters of unprocessed serum as input, and then apply the technology to the case of breast cancer recurrence to examine its performance and reveal its potential in clinical scenarios.

Results

I. Cell-Free RNA Concentration and Size in Breast Cancer Patient Serum

Four kits/methods, “TRIzol LS”, “exoRNeasy”, “NORGEN w/ DNase” and “QIAzol Method”, were applied to a total of 4 serum samples from 3 different breast cancer patients (serum 6025_8, 6025_9, 6038_1, 6046_3 from patients 6025, 6038 and 6046). These four methods above covered almost all commercially available technology for cell-free RNA extraction. Input serum volumes were 1ml for the first three methods and 200ul for QIAzol method. TRIzol LS was used to extract cell-free RNA from 1ml 6025_9 serum. Cell-free RNA extraction from serum 6038_1 was also conducted with three different methods, exoRNeasy, NORGEN w/ DNase and QIAzol method. Biological replicates were tested as well by applying exoRNeasy to serum 6025_8, 6038_1 and NORGEN w/ DNase to 6038_1, 6046_3 (Table.1). Quantification and size distribution of extracted RNA were determined with Bioanalyzer RNA Pico 6000.

The observed cell-free RNA concentrations in breast cancer patient serum were in the range from 0.2957ng/ml to 4.2ng/ml. TRIzol LS gave cell-free RNA concentration 0.8132 ± 0.364ng/ml in serum 6025_9, exoRNeasy gave 0.2957 ng/ml for serum 6025_8 and 0.6197ng/ml for 6038_1, NORGEN w/ DNase gave 3.2 ± 0.3ng/ml for 6038_1 and 1.8 ± 0.3ng/ml for 6046_3, QIAzol Method gave 4.2 ng/ml for 6038_1 (Table.1). TRIzol LS and exoRNeasy appeared to capture less RNA than NORGEN w/ DNase I and QIAzol Method.

Cell-free RNA extracted from breast cancer patient serums fell in the size range 40-300nt and two high concentration areas could be identified at sizes 70nt and 140nt for

most RNA samples extracted. Size distributions of all extracted RNA were obviously different from that of ultrapure water control, in terms of both concentration (or equivalently Florescence Units, FU) and shape. (Figure.1)

Table.1 RNA Extraction Methods and Corresponding RNA Concentrations Detected in Serums. "Methods" are the short names of the extraction methods, as mentioned in the section "Methods". "Purification Mechanism" shows brief summaries of working principles of the methods. "Serum" gives unique IDs to the serums, starting with a 4-digit patient ID followed by a 1-2 digit follow up number. In the last column contains corresponding cell-free RNA concentrations detected in specific serum with designated RNA extraction method. Averages and standard deviations of cell-free RNA concentrations detected in each serum with specific method were calculated, with number of measurement conducted being N in bracket (e.g. N=2 means a sample of extracted RNA was measured twice).
* All concentrations were derived from by Bioanalyzer quantification results, Qubit RNA HS gave too low for all samples of extracted RNA except the one in first row, which is 352pg/ul (only one measurement made). This value was not included since this was a different quantification tool.

Methods	***Purification Mechanism***	***Serum***	***Cell-free RNA Concentration in Serum (ng/ml) ****
TRIzol LS	Phenol-chloroform phase separation	6025_9	0.8132 ± 0.364 (N = 3)
exoRNeasy Kit	Affinity binding column for EVs then phenol-chloroform and silica membrane spin column purification	6025_8	0.2957
		6038_1	0.6197
NORGEN w/ DNase	Two spin column chromatography, DNase I digestion and silica membrane spin column purification	6038_1	3.2 ± 0.3 (N = 2)
		6046_3	1.8 ± 0.3 (N = 2)
QIAzol Method	Phenol-chloroform phase separation	6038_1	4.2

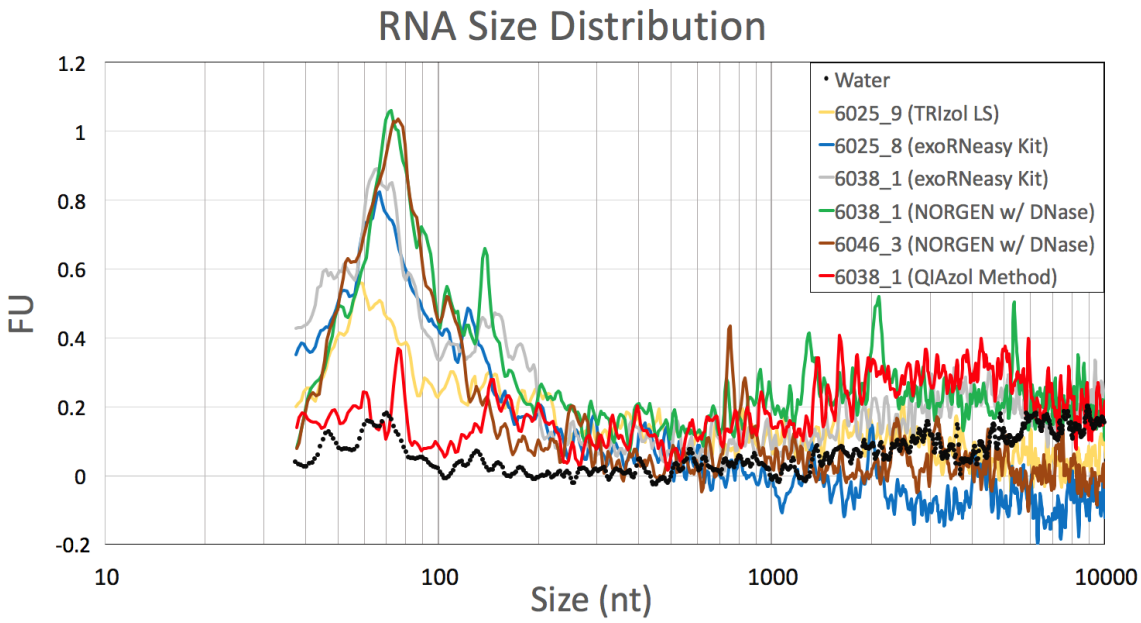


Figure.1 Size Distribution Results of Cell-free RNA Extracted from All 4 Serums with All 4 Different Methods. Marker signals (<37nt) were removed, x-axis is size with the unit of nucleotides, y-axis is fluorescence unit (FU), which is proportional to concentration. Almost all size distributions shared the same shape and most of the cell-free RNA fell in the size range of 40-300nt with two major concentrated area around 70 and 140nt. Water: 1ul ultrapure water was given as direct input to bioanalyzer.

II. Development of Ultra-Low Starting Volume Cell-free RNA Library Construction and Sequencing Technology

1. Optimization of Library Sequencing and Mapping Quality

Sequencing quality of a library was mostly represented by total read counts and mapping quality was mostly represented by uniquely mapped percentage to human genome (will be mentioned as "sample total read counts" and "uniquely mapped percentage"). Two conditions most possibly affecting sequencing and mapping qualities, and most easily to be manipulated during library construction and sequencing were serum input volume and addition of custom sequencing primer.

a. Optimization with Serum Input Volume to Library Construction Protocol

We firstly set off to identify the volume range of serum input that would lead to completion of sequencing library construction. Two serums, 6004_9 and 6057_5, were tested with input volumes being 3, 7, 11 15ul and 3, 7, 15, 30ul respectively. Library construction could not be completed with ≥11ul of serum input as gel-like structure would form in tubes and made following procedures not feasible. Viscosity increase was observed with 7ul serum input and library construction completion could just be achieved. Thus, the serum input volume range where library construction could be completed was ≤ 7ul.

Optimization with serum input volume was thus conducted with the range ≤ 7ul. Tests within the range were carried out for trend between qualities and serum input volume as shown below. 8 libraries were constructed from 3ul, 5ul, 6ul and 7ul input of serum 2010_4, two libraries for each volume, and these 8 libraries will be mentioned as "Set 1". Besides, 4 and 2 libraries were constructed from 3ul and 7ul input of serum

6004_9, 5 and 4 libraries were constructed from 3ul and 7ul input of serum 6057_5 respectively, and these libraries will be mentioned as “Set 2”. Averages and standard deviations were calculated for sample total read counts and uniquely mapped percentage in each condition if applicable (Appendix Table.A1).

Observed from the results of Set 1libraries, both sample total reads and uniquely mapped percentage increased as serum input volume increases (Figure.2). The drop of sample total read counts shown by the left-most column was expected and due to human error. Sequencing results from Set 2 libraries showed also better sequencing and mapping quality when serum input volume is 7ul (Figure.3). Thus, 7ul was the optimized serum volume input with the restriction of library construction completion.

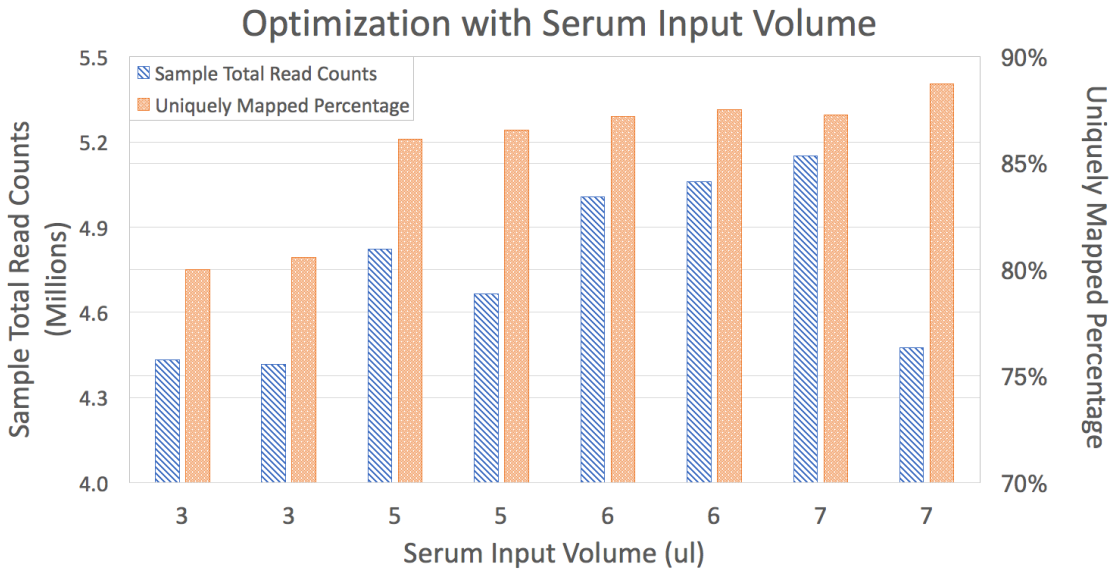


Figure.2 Sequencing and Mapping Qualities Optimization with 3ul-7ul Input of Serum 2010_4. Data used in this plot are from sequencing results of 3ul, 5ul, 6ul and 7ul input of serum 2010_4 (i.e. Set 1). Blue bars are sample total read counts and their readings could be obtained from the vertical axis on the left. Orange bars are uniquely mapped percentages, and their readings could be obtained from the vertical axis on the right.

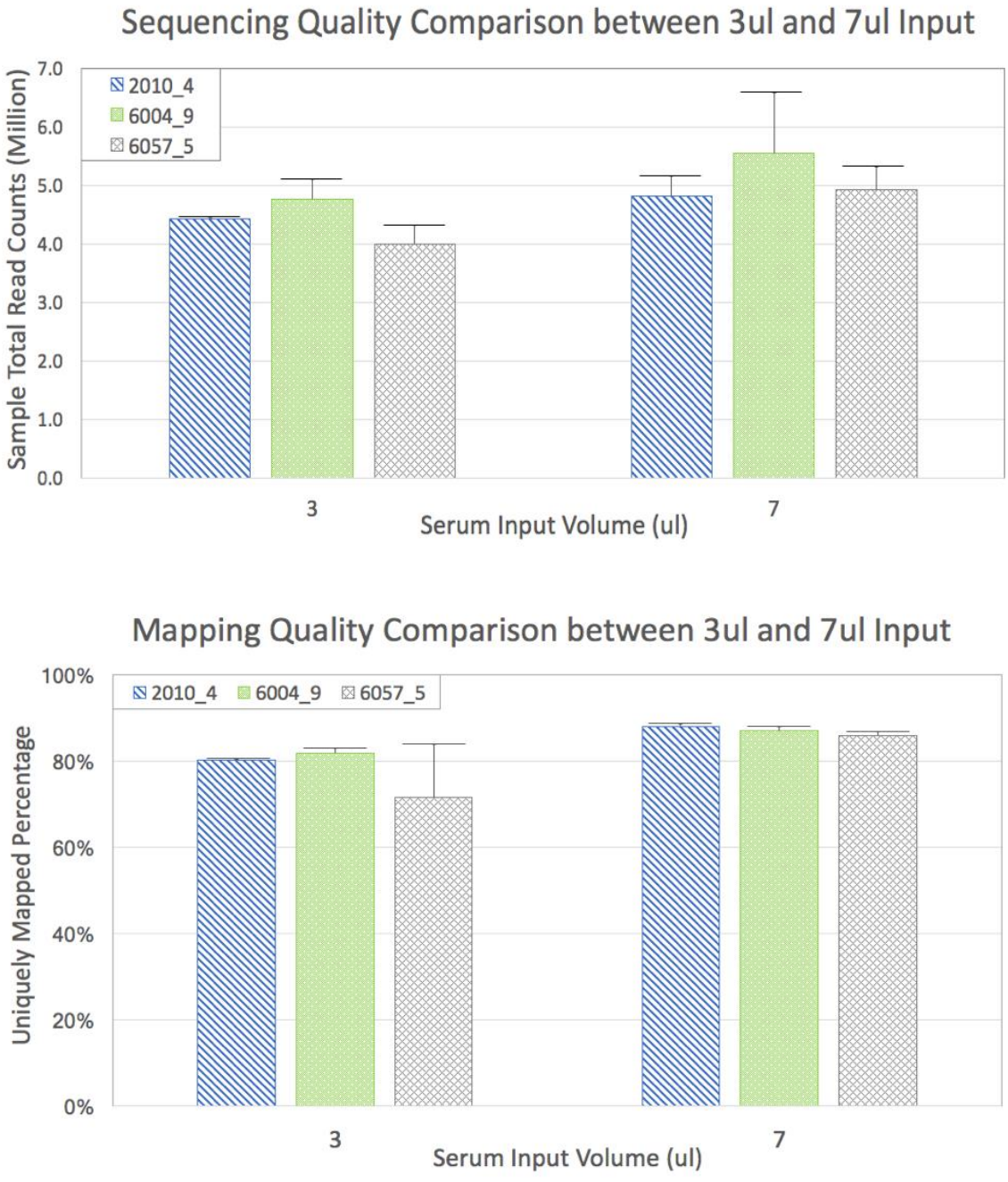


Figure.3 Effect of Increasing Serum Input Volume on Sequencing and Mapping Qualities. Effect of increasing serum input volume on sequencing (upper) and mapping (lower) qualities with results from serum 2010_4, 6004_9 and 6057_5. Blue bars represent qualities from libraries constructed from 3ul or 7ul serum 2010_4 input, green bars represent those from 3ul or 7ul serum 6004_9 input, gray bars represent those from 3ul or 7ul serum 6057_5.

b. Optimization with Addition of Custom R1 Sequencing Primer

We tested both sequencing and mapping qualities with and without custom R1 sequencing primer. 4 and 6 libraries were constructed from 3ul input of serums 6004_9 and sequenced with and without custom sequencing primer respectively, 5 and 6 libraries were constructed from 3ul input of serums 6057_5 and sequenced with and without custom primer respectively. Averages and standard deviations of sample total read counts and uniquely mapped percentages were calculated in each condition (Appendix Table.A2).

From the comparison, addition of custom read 1 sequencing primer almost doubled sample total read counts, yet had no obvious effect on uniquely mapped percentage, i.e. adding custom R1 sequencing primer improved sequencing quality but had no obvious effect on mapping quality (Figure.4).

In a word, sequencing and mapping qualities were optimized with the combination of 7ul library construction serum input volume and custom sequencing primer under the library construction input volume restriction.

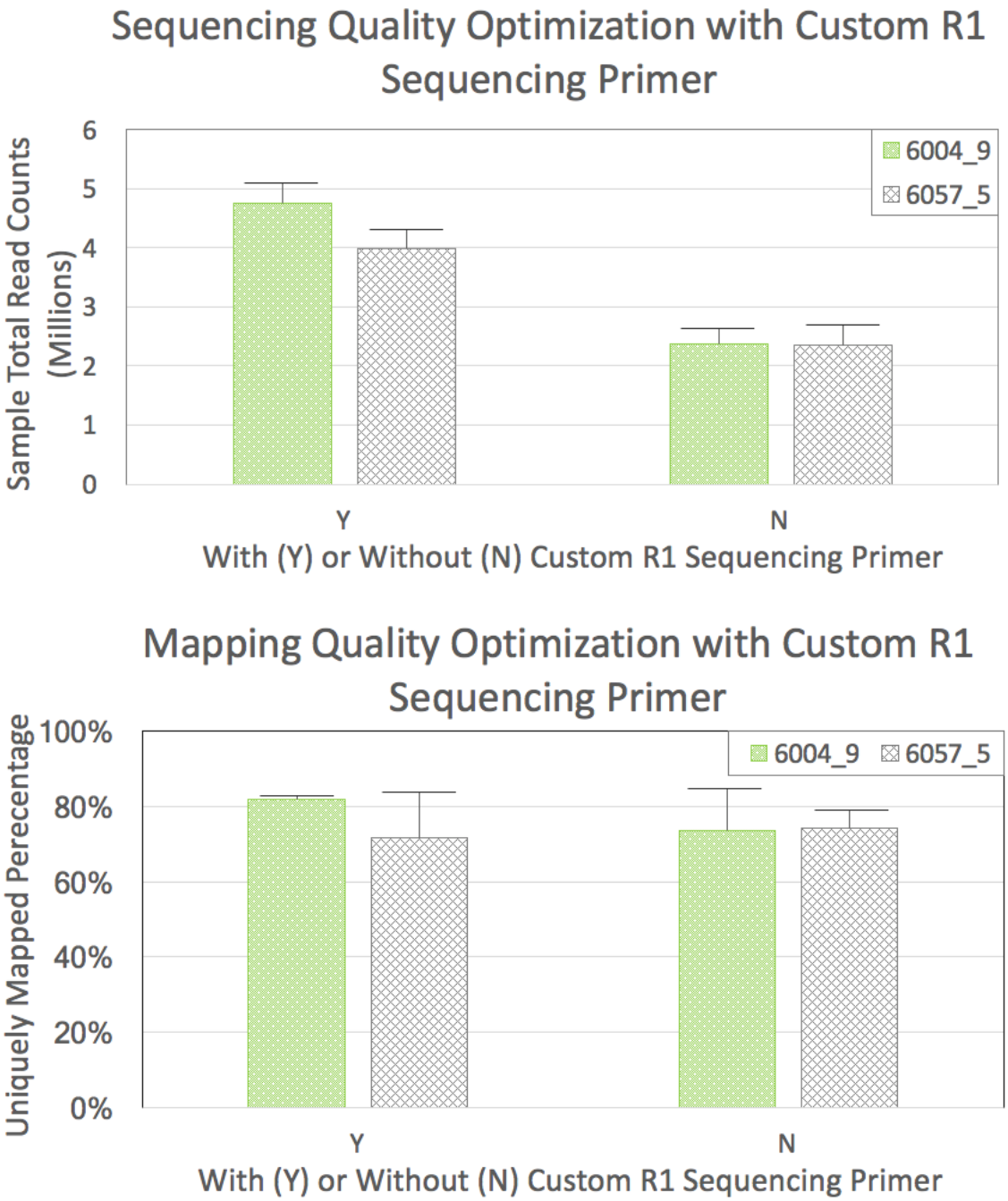


Figure.4 Effect of Adding Custom R1 Sequencing Primer on Sequencing and Mapping Qualities. Effect of adding custom R1 sequencing primer on sequencing (upper) and mapping (lower) qualities with results from serum 6004_9 and 6057_5. Green bars represent qualities from libraries constructed from serum 6004_9, gray bars represent those from serum 6057_5.

2. Optimization of RNA-Seq Library Complexity

Library complexity was mostly indicated by number of genes detected. With serum input volume being the easiest factor to manipulate, and the most possible factor related, library complexity was optimized by identifying a volume where most genes could be detected. Optimization was conducted over the range consisted of two sections: less, and more than 7ul serum input volume inclusively, theses sections will be referred as lower and higher range sections respectively. Sequencing data from each constructed library with serum input volume less than 7ul were used directly for analysis in lower range section and because library construction was not feasible with >7ul serum input, "combined" sequence data (will be mentioned with more details in next paragraph) from libraries with 3-7ul input of same serums were used for analysis in higher section range. A total of 39 libraries constructed from 3ul to 7ul input of the serums 2010_4, 6004_9 and 6057_5 were studied here for the optimization.

Optimization in lower range section was conducted with a subset of 23 libraries of the total 39 libraries: 2 libraries from each of 3ul, 5ul, 6ul and 7ul input of serum 2010_4 (8 libraries in total), 4 and 2 libraries from 3ul and 7ul of serum 6004_9 respectively (6 libraries in total), 5 and 4 libraries from 3ul and 7ul of serum 6057_5 respectively (9 libraries in total). Optimization in higher range section was carried out by combining sequencing data of all 9, 15 and 15 libraries constructed from serums 2010_4, 6004_9 and 6057_5 respectively. To elaborate "combining sequencing data" with details, sequencing data from each serum were cumulatively combined starting and continuing with a random uncombined library built from highest serum input volumes (i.e. combine all sequencing data from 7ul first and then cumulatively combine all sequence data from

6ul, then same for 5ul, and finally 3ul, this method will be mentioned as “sequencing data combination” in the rest of the report). Number of genes detected together with the corresponding total volumes of serum involved were recorded in pair for analysis (see Table.2 for an example).

In serum 2010_4, it was observed that increase of serum input volume in 3ul to 7ul range improved number of genes detected, i.e. library complexity (Figure.5). This improvement was confirmed with those 6 libraries from serum 6004_9 and 9 libraries from 6057_5 mentioned above (Appendix Table.A3 and Figure.6). For all three serums, among all 60675 human genes (ENSEMBL gene annotation (HG38)), no less than 30000 genes in most cases (35353 and 32477 for serum 2010_4, 34416 and 29334 for serum 6004_9 and 28034, 31950, 34002 and 34241 for serum 6057_5) were detected (TPM, transcripts per million>1) with only 7ul direct serum input for library construction. Besides, it was observed that in higher range section (> 7ul), for all three serums, numbers of detected genes (TPM>1) went up almost monotonically with total volume of serum involved. However, the increase was much less obvious after around 26ul for serum 2010_4 or 28ul for the other two was involved, indicating plateaus being reached. More than 47000 genes (47540 for serum 2010_4, 47851 for 6004_9 and 47490 for 6057_5) or around 78% of all human genes were detected (TPM>1) with 26ul of serum 2010_4 involved or 28ul of the other two involved. Numbers of genes detected (TPM>1) with sequencing data of all 9, 15 and 15 libraries combined were 50606, 50949 and 50544 for serums 2010_4, 6004_9 and 6057_5 respectively. (Figure.7, red lines)

Though highest number of genes were detected with the total involved serum volumes being 45ul, 61ul and 61ul for serums 2010_4, 6004_9 and 6057_5, with the

plateaus being reached at around 26ul or 28ul and library protocol restriction (input volume < 7ul), optimization of library complexity gave the strategy: combining sequence data from 4 libraries constructed from 7ul serum volume input to achieve a 28ul total volume of serum involved.

Table.2 An Example of "Sequencing Data Combination". Data were from sequencing data combination of 15 libraries from the serum 6057_5. The column "Total volume of serum involved" gives the summations of the serum input volumes of libraries of which the sequencing data were combined. Difference between two adjacent cells in this column is the serum input volume of the library that would be combined to achieve the bigger volume. The column "Number of genes detected" gives the numbers of genes identified in the corresponding combined sequencing data set. Contribution to number of genes detected made by combining each library is the difference between the consecutive numbers in the column.

Total Volume of Serum Involved (ul)	*Number of Genes Detected*
7	39669
14	44402
21	46662
28	47851
31	48482
34	48720
37	48880
40	49136
43	49316
46	49764
49	49888
52	49992
55	50057
58	50339
61	50544

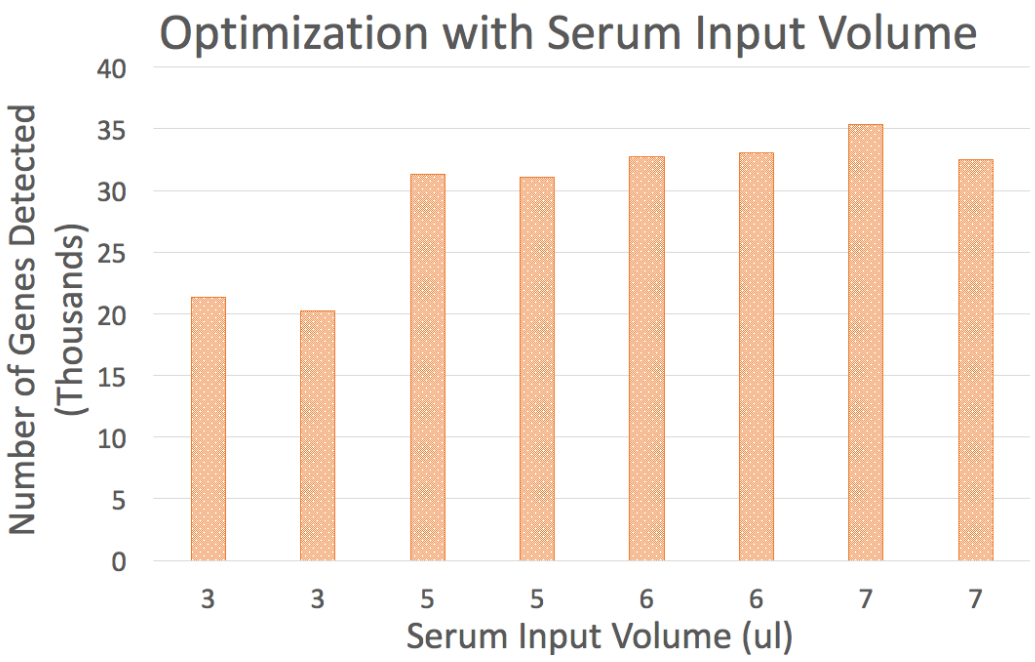


Figure.5 Effect of Serum Input Volume on Library Complexity (2010_4). Effect of serum input volume on library complexity with results from the 8 libraries constructed from serum 2010_4.

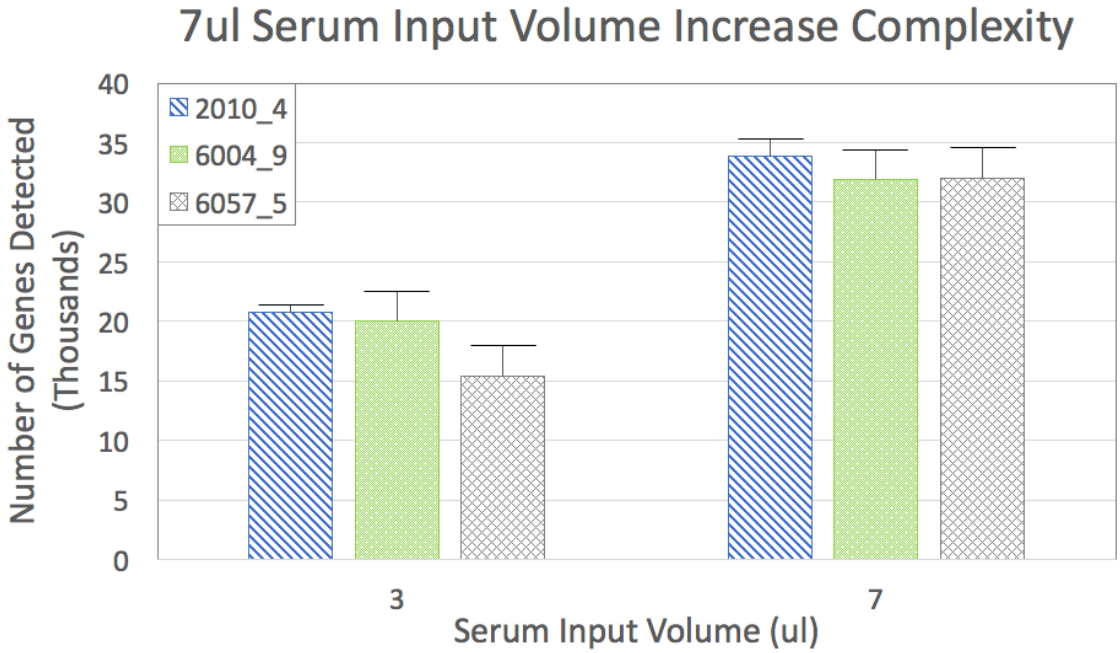
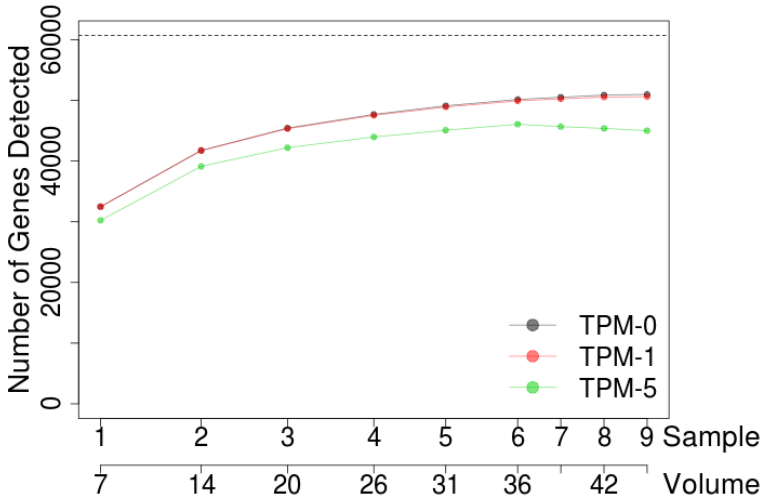
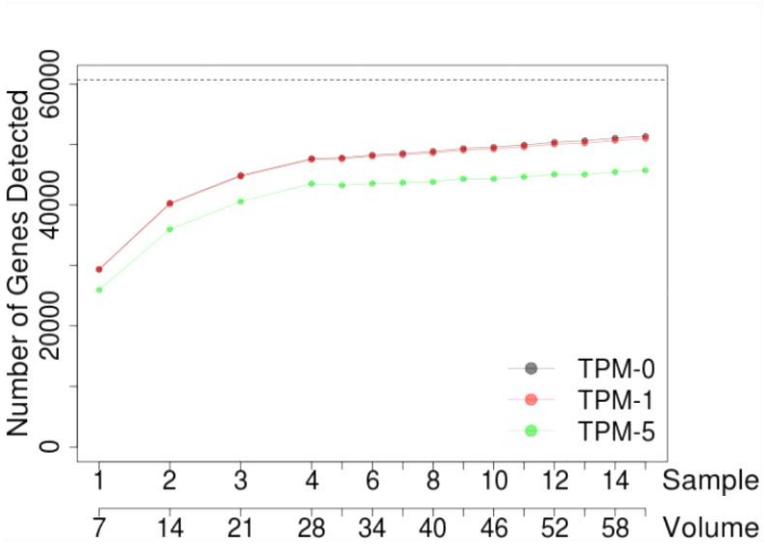


Figure.6 Effect of Serum Input Volume on Library Complexity (2010_4, 6004_9 and 6057_5). Effect of serum input volume on library complexity with results from 4, 6 and 9 libraries constructed from 3ul or 7ul input of serums 2010_4, 6004_9 and 6057_5 respectively. Blue, green and gray bars are numbers of genes detected from libraries constructed from 3ul or 7ul serum 2010_4, 6004_9 and 6057_5 input respectively.





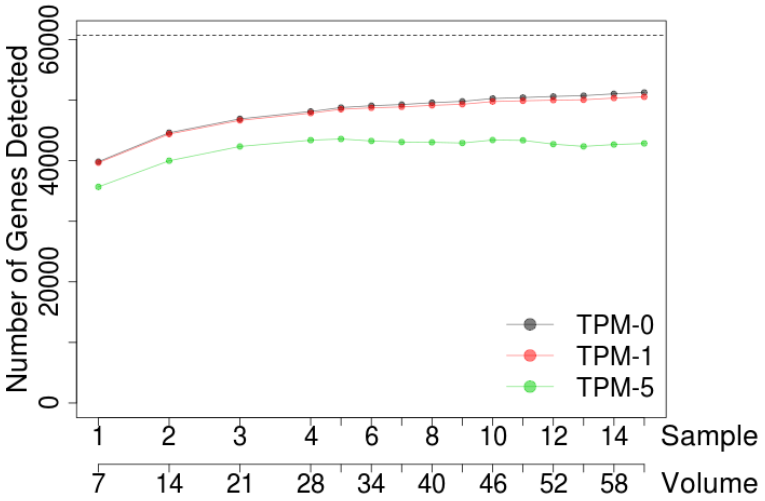


Figure.7 Effect of Sequencing Data Combination on Library Complexity. Effect of sequencing data combination on library complexity, i.e. number of genes detected. Top-left: serum 2010_4; Top-right: serum 6004_9; Bottom-left: serum 6057_5.
There are two x-axes on each chart: the upper one, marked by "Sample" indicates the number of sequencing data set cumulatively combined, the lower one marked by "Volume" is total volume of serum involved (with the unit of microliter) after each cumulative combination. Y-axis is number of genes detected. Though gene was commonly considered as detected with TPM>1, of which the results are shown in red lines, different thresholds, TPM>0 and TPM>5, were also applied, results shown in gray and green lines respectively. Same results and trend were observed for detection criteria TPM>0 and TPM>5, except the plateaus go down to around 42000 genes when TPM>5.

3. Optimization of Library Construction Efficiency

During the construction of in total 163 libraries, different numbers of libraries could be constructed at the same time within one experimental batch, and construction could be conducted with or without robot automation (epMotion). These two concepts will be mentioned as "batch status" and "automation status" in the rest of the report.

There were in total three different batch and automation statuses applied to prepare libraries: 8 libraries per batch without epMotion, 8 libraries per batch with epMotion and 96 libraries per batch with epMotion. Average times spent to construct one library were 2.29, 1.5 and 0.42 hours for the three circumstances respectively. In terms of percentage, 28% and 18.3% of the time spent per library construction under the first circumstance was spend under the second and third (Table.3). It is obvious that larger batch and automation both increased library construction efficiency

Efficiency is thus optimized by applying large scale automated library preparation, in this case, constructing 96 libraries in a batch with epMotion, which is the upper limit of epMotion capacity at this stage.

Table. 3 Effect of Batch and Automation Status on Library Construction Efficiency. "Batch Status" is how many libraries were constructed at the same time with in one batch of experiment. "Automation Status" shows whether epMotion was used to prepared libraries. "Average Time Spent per Sample (hours)" is the calculated amount of time required to construct one library under specific circumstance summarized in the first two columns.

Batch Status	*Automation Status*	*Average Time Spent per Library (hours)*
8 Libraries per Batch	w/o epMotion	2.29
	w/ epMotion	1.5
96 Libraries per Batch	w/ epMotion	0.42

III. Diversity of Detected Genes in Serum

1. Overall Diversity of Detected Genes in Serum

How many different categories of genes can be detected in only micro-liter level of serum? The question was investigated with the three most well studied serums from three different patients covering both recurrence and non-recurrence statuses. The serums were 2010_4, 6004_9 and 6057_6, the first two serums were from two recurrence patients and the last one was from a non-recurrence patient. 9, 15 and 15 libraries constructed with each of the serums were studied with method "sequencing data combination" respectively, the same method mentioned in the previous section "Development of Ultra-Low Starting Volume Cell-free RNA Library Construction and Sequencing Technology" (Table.2).

A large diversity of genes was observed in high consistency over all three serums, regardless of patient recurrence status: the in total 50606, 50949 and 50544 genes detected respectively in serum 2010_4, 6004_9 and 6057_9 fell in the same 42 out of 44 human gene categories (ENSEMBL gene annotation (HG38)). These categories included, most importantly, protein-coding genes, lincRNA, miRNA and snRNA, etc., which were crucial for basic functions and regulations. With sequencing data of all 9, 15 and 15 libraries combined: 95% of all protein coding genes, 92% of all lincRNA, 41% of all miRNA and 67% of all snRNA were detected in serum 2010_4; 95% of all protein coding genes, 93% of all lincRNA, 42% of all miRNA and 65% of all snRNA were detected in serum 6004_9; 94% of all protein coding genes, 93% of all lincRNA, 41% of all miRNA and 63% of all snRNA were detected in serum 6057_5 (see Table.4 for numbers).

To summarize, large and consistent gene diversity was observed over all three serums. Genes were detected in the same 42 out of all 44 human gene categories in all three serums covering both recurrence and non-recurrence statuses. Important categories such as protein coding genes, lincRNA, miRNA and snRNA were studied as examples. Almost all protein coding genes (>94%) and lincRNA (>92%), around 41% of all miRNA and 65% of all snRNA (ENSEMBL gene annotation (HG38)) could be detected in equivalently 45ul of serum 2010_4 and 61ul of the other two serums. High consistency of gene diversity was observed among three serums, though they covered recurrence and non-recurrence statuses.

Table.4 Numbers of Genes Detected in Important Gene Categories. Numbers of genes detected in each of the following gene categories: protein-coding gene, lincRNA, miRNA and snRNA. ENSEMBL gene annotation (HG38) was used as reference genome. Patient 2010 and patient 6004 are recurrence patient and patient 6057 is non-recurrence patient. Total numbers of genes were in the brackets after the name of each gene category.

Serum	*Protein-coding Gene (19826)*		*lincRNA (7668)*		*miRNA (4198)*		*snRNA (1905)*	
	Number	Percent	Number	Percent	Number	Percent	Number	Percent
2010_4	18769	95%	7073	92%	1708	41%	1268	67%
6004_9	18784	95%	7116	93%	1781	42%	1236	65%
6057_5	18716	94%	7096	93%	1718	41%	1197	63%

2. Discovery of Tissue-Specific Genes in Serum

In human serum, is it possible to detect tissue specific genes, especially those specific to tissues where biopsy sampling is difficult or even impossible? “Yes” to the answer may potentially reveal a huge advantage of serum on patient compliance and sampling feasibility over biopsy of those specific tissues. Three tissues where biopsy sampling was difficult or impossible were selected as examples: brain, bone marrow and

peripheral nervous system (PNS). With the example tissues, the question was addressed by firstly combining sequencing data of all 9, 15 and 15 libraries from serums 2010_4, 6004_9 and 6057_5 respectively, then mapping the data to ENSEMBL gene annotation (HG38) and lastly matching the results with TiGER (http://bioinfo.wilmer.jhu.edu/tiger/) tissue-specific gene expression profile for genes specific to example tissues.

Among all 176, 192 and 78 tissue specific genes for brain, bone marrow and PNS, 176, 191 and 78 genes were detected respectively in serum 2010_4 with 46ul serum volume involved in sequencing data combination. Numbers of genes detected were 175, 191 and 78 respectively in serum 6004_9 and 176, 189 and 78 respectively in serum 6057_5 (Table.5). TPMs of genes specific to these three tissues were mostly larger than 4, detailed distributions were shown in Figure.8 for each sample.

Success in detecting genes specific to important tissues and organs revealed the potential of serum as an alternative biopsy for diagnosis and prognosis of diseases related to specific tissues or organs.

Table.5 Numbers of Tissue Specific Genes Detected for Example Tissues/Organs. Total numbers of tissue specific genes for the three tissues/organs (brain, bone marrow and peripheral nervous system (PNS) in TiGER are shown in the first row. Numbers of tissue specific genes detected for the three tissues/organs in three serums are shown in the other three rows.

	Brain Specific	*Bone Marrow Specific*	*PNS Specific*
All in Data Base	176	192	78
2010_4	176	191	78
6004_9	175	191	78
6057_5	176	189	78

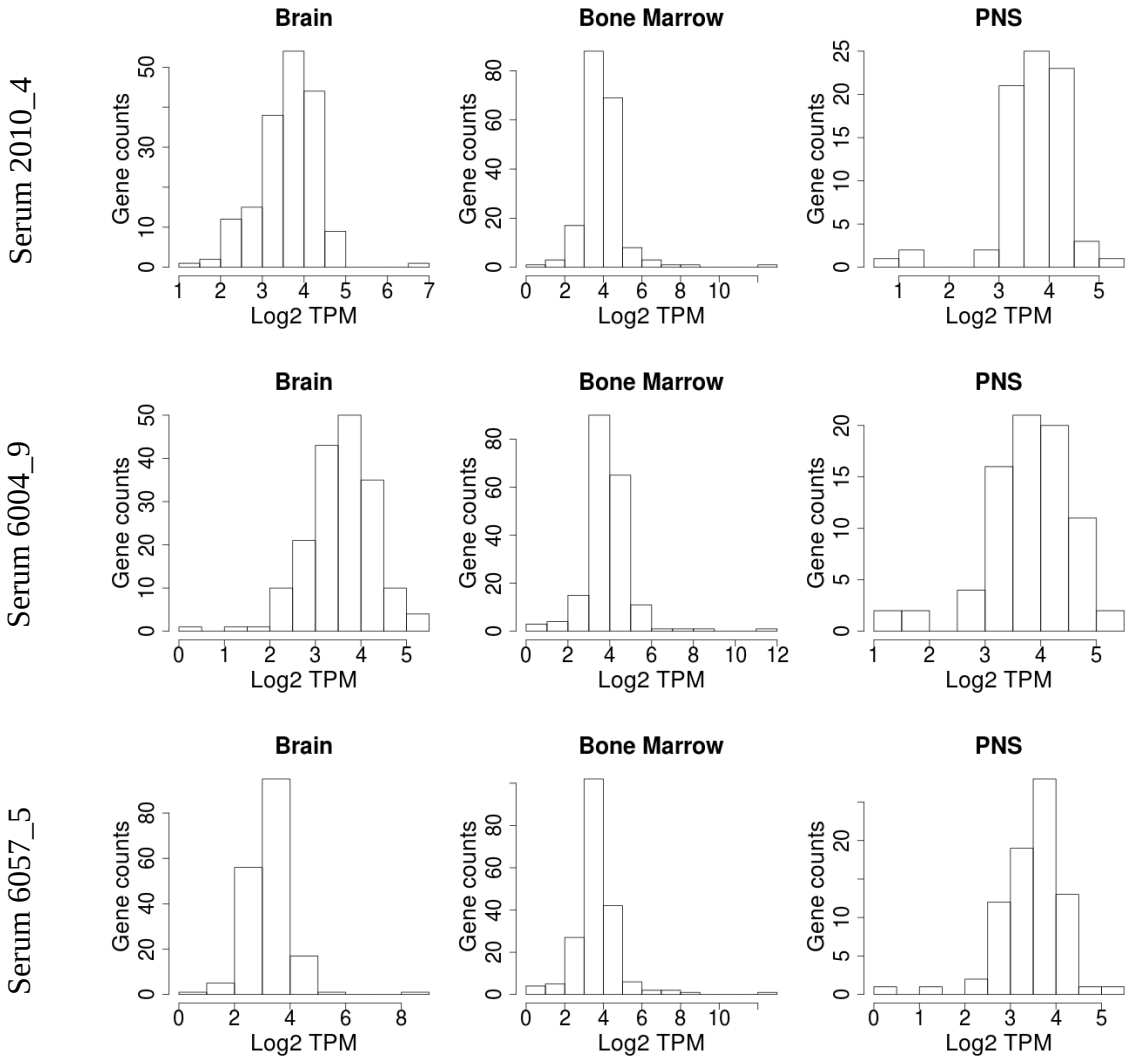


Figure.8 Tissue Specific Genes Detected and Their Gene Expression Levels. Tissue specific genes detected and their gene expression (TPM) levels in brain, bone marrow and peripheral nervous system (PNS) in three different serums: 2010_4 (upper), 6004_9 (middle), 6057_5 (lower), based on combined sequencing data of all 9, 15 and 15 libraries respectively. Different tissues are in titles of each histogram, x-axis is gene expression level in logarithm (log_2TPM), and y-axis is the number of genes with their expression level falling in the range indicated on x-axis.

IV. Application of the Technology on Breast Cancer Recurrence

Finally, how is the performance of the developed technology in terms of disease prognosis? We applied the technology to the case of breast cancer recurrence to answer the question. With RNA-Seq library construction details optimized with the conditions mentioned above, 96 libraries from 96 serums of both recurrence and non-recurrence patients were built in one single batch with 7ul serum input. Sequencing was conducted on the platform HiSeq4000 with custom R1 sequencing primer using two out of eight lanes on a flow cell. Qualities and complexity of the libraries were firstly examined. Principle component analysis (PCA) was conducted on all genes for dominant difference between recurrence and non-recurrence patient populations. Three methods were then applied to identify populations of potential recurrence biomarkers, and their ability separating recurrence and non-recurrence patient populations were also demonstrated. Lastly a classifier was built accordingly for recurrence risk assessment.

1. Sample Description

96 serum samples were carefully selected to achieve the consistency of sampling time. Patient selection was firstly conducted, followed by serum selection from the selected patients. Patients were selected into the study only when such two serums for the patient existed: one collected between 30 days before and 60 days after chemotherapy ends, inclusive (Range 1) and the other collected more than 60 days after chemotherapy ends, exclusive (Range 2). With the rule above, 9 recurrence and 33 non-recurrence patients were identified (Figure.9). Then, all serums collected in Range 1 and Range 2 were selected for recurrence patients, and only one serum collected in Range1 and one collected earliest in Range 2 were selected for non-recurrence patients. Numbers of

serums selected above were 24 and 66 respectively. 6 more serums with collection dates slightly out of ranges were also included into the study, these were serums 6027_2, 6039_2, 2056_2 and 2056_4, 6049_1 and 6049_4 (63, 49 and 49 days before and 204 days after, 149 days before and 241 days after chemotherapy ends respectively). In summary, in the total of 96 serums selected, there were 28 serums from 10 recurrence patients with 2-4 serums per patient, and 68 serums from 34 non-recurrence patients with 2 serums per patient. List of all 96 serums is shown in Table.6, with red and black indicates their being selected from recurrence and non-recurrence patients respectively.

2. A Glance at Sequencing Result: Qualities, Complexity and PCA

Sequencing, mapping qualities and library complexity were firstly examined with sequencing results of the 96 libraries. Good in these three aspects was indicated by 7666371 ± 962851 (around 736 million for 96 samples in total) sample total read counts, 83.97% ± 2.53% uniquely mapped percentage and 27257 ± 6613 genes detected (TPM>1) respectively (Table.6). Mapping and gene detection were conducted with ENSEMBL gene annotation (HG38). PCA were then conducted on 96 libraries with all 60675 genes (ENSEMBL gene annotation (HG38)) to identify the existence of dominant differences between recurrence and non-recurrence populations and the non-obvious separation of the populations indicates their absence (Figure.10), namely, there's no obvious difference between recurrence and non-recurrence patients' serums in terms of gene expression.

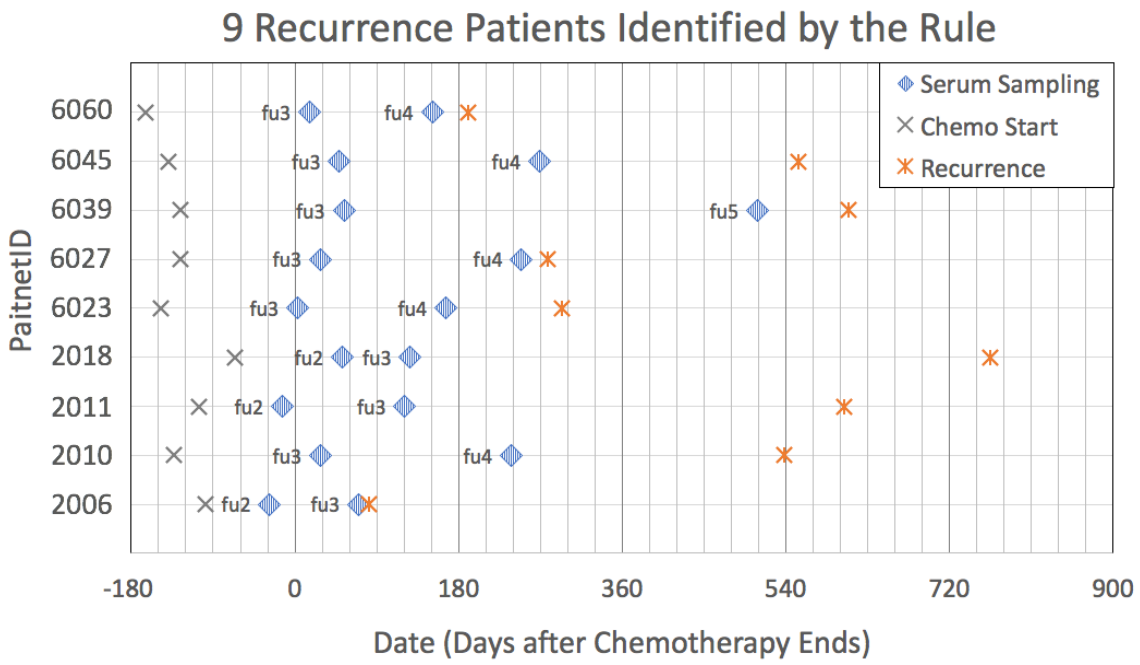


Figure.9 The 9 Recurrence and the 33 Non-recurrence Patients Identified by the Rule. The 9 recurrence (upper) and the 33 non-recurrence (lower) patients identified by the rule. Dates of events (such as serum sampling) are shown on x-axis with the day chemotherapy ends being Day 0, negative values mean the event is before chemotherapy end and positive values means after. Each horizontal line represents a patient with the corresponding PatientID on y-axis. Blue squares on each horizontal line denote serum sampling events of the specific patient with the dates on x-axis, gray crosses denote starts of chemotherapy of the specific patient and orange stars denote cancer recurrence.
In the upper plot, PatientID of the nine patients (i.e. nine horizontal lines) were shown explicitly to the left of the plot, and follow-up numbers of each serum sample were also shown explicitly to the left of each blue square in the format of e.g. fu3 (representing the third follow-up). Due to space limit, these were not shown for the 33 non-recurrence patients. Details could be found in non-recurrence section of Table.6 (in black color). PatientIDs are 2002-6002 from top to bottom, with 6049 skipped, and follow-up numbers are the corresponding numbers in the table.

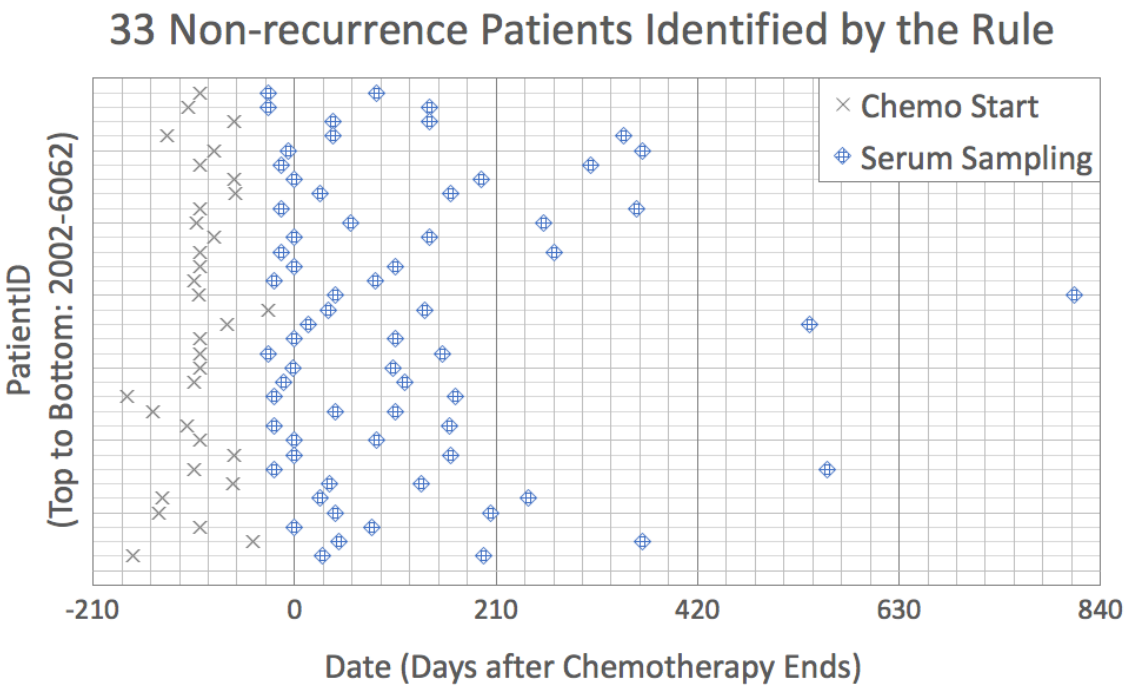


Figure.9 The 9 Recurrence and the 33 Non-recurrence Patients Identified by the Rule (Continued). The 9 recurrence (upper) and the 33 non-recurrence (lower) patients identified by the rule. Dates of events such as serum sampling are shown on x-axis with the day chemotherapy ends being Day 0, negative values mean the event is before chemotherapy end and positive values means after. Each horizontal line represents a patient with the corresponding PatientID on y-axis. Blue squares on each horizontal line denote serum sampling events of the specific patient with the dates on x-axis, gray crosses denote starts of chemotherapy of the specific patient and orange stars denote cancer recurrence.
In the upper plot, PatientID of the nine patients (i.e. nine horizontal lines) were shown explicitly to the left of the plot, and follow-up numbers of each serum sample were also shown explicitly to the left of each blue square in the format of e.g. fu3 (representing the third follow-up). Due to space limit, these were not shown for the 33 non-recurrence patients. Details could be found in non-recurrence section of Table.6 (those in black color). PatientIDs are 2002-6002 from top to bottom, with 6049 skipped, and follow-up numbers are the corresponding numbers in the table.

Table.6 Information of the 96 Serums Selected. Recurrence serums are marked in red and non-recurrence serums are marked in black. IDs in "Serum" identify the serums with a 4-digit PatientID and a 1-2 digit follow up (fu) number. "Sample Total Read Counts" gives total read counts for each library. "Uniquely Mapped Percentage" gives the percentages of reads could be uniquely mapped to human genome in sample total read counts. "Number of Genes Detected (TPM>1)" shows the numbers of genes detected (ENSEMBL gene annotation (HG38)) in the libraries (TPM>1).

Serum	*Sample Total Read Counts*	*Uniquely Mapped Percentage*	*Number of Genes Detected (TPM>1)*
2006_2	7293272	82.88%	30416
2006_3	7870895	82.92%	24107
2010_3	6622312	84.87%	26608
2010_4	8248171	85.37%	32072
2011_2	9233397	85.80%	27425
2011_3	7156841	85.80%	29720
2011_5	7344237	83.32%	30418
2011_10	8130653	84.35%	24800
2018_2	8374268	87.15%	24085
2018_3	7637570	85.16%	29435
2018_4	7110637	82.77%	25272
2018_5	6848186	85.05%	27550
2056_2	7093422	84.26%	25821
2056_4	6609194	83.60%	37714
6023_3	7105263	86.89%	29322
6023_4	8346173	84.26%	19070
6027_2	9045224	82.20%	18095
6027_3	8330933	83.87%	31390
6027_4	7620354	85.44%	33337
6027_5	7365425	84.48%	34499
6039_2	7023051	84.80%	33066
6039_3	5636176	84.33%	21682
6039_5	7389168	83.45%	24183
6045_3	8827999	86.48%	26709
6045_4	7799765	85.40%	33002
6045_6	8161797	84.26%	31291
6060_3	7786528	85.96%	30928
6060_4	11070218	86.23%	31219

Table.6 Information of the 96 Serums Selected (Continued). Recurrence serums are marked in red and non-recurrence serums are marked in black. IDs in "Serum" identify the serums with a 4-digit PatientID and a 1-2 digit follow up (fu) number. "Sample Total Read Counts" gives total read counts for each library. "Uniquely Mapped Percentage" gives the percentages of reads could be uniquely mapped to human genome in sample total read counts. "Number of Genes Detected (TPM>1)" shows the numbers of genes detected (ENSEMBL gene annotation (HG38)) in the libraries (TPM>1).

Serum	***Sample Total Read Counts***	***Uniquely Mapped Percentage***	***Number of Genes Detected (TPM>1)***
2002_2	9025676	84.20%	20757
2002_3	8148798	85.17%	20978
2003_2	9183238	86.14%	30010
2003_3	8961017	85.65%	24133
2005_2	7377949	84.21%	18879
2005_3	6490139	80.58%	14867
2014_3	6536128	84.36%	21559
2014_5	6532672	76.46%	16863
2023_2	8296427	85.84%	27102
2023_5	6965863	84.49%	33694
2024_2	7283528	85.52%	33302
2024_4	8761199	83.18%	21410
2028_2	7288621	75.50%	31026
2028_3	7764460	83.72%	24892
2031_2	6968233	79.16%	11462
2031_3	7802809	82.60%	22478
2038_2	7804645	85.22%	27067
2038_4	6103776	80.29%	32121
2039_3	7986157	85.07%	24251
2039_4	5873688	82.87%	34075
2041_2	7578732	85.50%	24612
2041_3	9298701	82.88%	17323

Table.6 Information of the 96 Serums Selected (Continued). Recurrence serums are marked in red and non-recurrence serums are marked in black. IDs in "Serum" identify the serums with a 4-digit PatientID and a 1-2 digit follow up (fu) number. "Sample Total Read Counts" gives total read counts for each library. "Uniquely Mapped Percentage" gives the percentages of reads could be uniquely mapped to human genome in sample total read counts. "Number of Genes Detected (TPM>1)" shows the numbers of genes detected (ENSEMBL gene annotation (HG38)) in the libraries (TPM>1).

Serum	***Sample Total Read Counts***	***Uniquely Mapped Percentage***	***Number of Genes Detected (TPM>1)***
2048_2	7846486	82.78%	39066
2048_4	7677529	85.27%	24690
2060_2	7935080	84.29%	39942
2060_3	6392908	82.84%	37473
2065_2	6047318	77.38%	35686
2065_3	6628185	85.48%	31037
2074_2	8561951	84.77%	19936
2074_7	6459683	82.57%	37380
2075_2	10165886	84.72%	39152
2075_3	7355997	87.18%	30162
2091_2	7570788	84.30%	38126
2091_5	7458148	86.38%	29498
2097_2	7857570	86.46%	33617
2097_3	7322319	84.11%	25116
6017_2	7772216	80.35%	20170
6017_3	6514249	81.30%	16769
6019_2	6820883	83.94%	26882

Table.6 Information of the 96 Serums Selected (Continued). Recurrence serums are marked in red and non-recurrence serums are marked in black. IDs in "Serum" identify the serums with a 4-digit PatientID and a 1-2 digit follow up (fu) number. "Sample Total Read Counts" gives total read counts for each library. "Uniquely Mapped Percentage" gives the percentages of reads could be uniquely mapped to human genome in sample total read counts. "Number of Genes Detected (TPM>1)" shows the numbers of genes detected (ENSEMBL gene annotation (HG38)) in the libraries (TPM>1).

Serum	***Sample Total Read Counts***	***Uniquely Mapped Percentage***	***Number of Genes Detected (TPM>1)***
6019_3	8798580	87.02%	30726
6021_2	9586903	86.03%	26323
6021_3	8328434	85.52%	29370
6022_3	7002006	75.28%	11197
6022_4	6274345	82.80%	26914
6025_3	8048885	86.34%	16943
6025_4	6955775	79.11%	13748
6028_2	8047948	85.21%	29231
6028_3	7215702	84.07%	28572
6034_2	9147154	86.38%	40194
6034_3	7820560	86.28%	29264
6035_2	5791185	82.92%	31090
6035_3	6715321	85.20%	25802
6038_2	8401167	85.07%	30782
6038_5	7116876	84.96%	26596
6040_2	8724735	85.44%	39442
6040_3	7672233	86.85%	26331

Table.6 Information of the 96 Serums Selected (Continued). Recurrence serums are marked in red and non-recurrence serums are marked in black. IDs in "Serum" identify the serums with a 4-digit PatientID and a 1-2 digit follow up (fu) number. "Sample Total Read Counts" gives total read counts for each library. "Uniquely Mapped Percentage" gives the percentages of reads could be uniquely mapped to human genome in sample total read counts. "Number of Genes Detected (TPM>1)" shows the numbers of genes detected (ENSEMBL gene annotation (HG38)) in the libraries (TPM>1).

Serum	***Sample Total Read Counts***	***Uniquely Mapped Percentage***	***Number of Genes Detected (TPM>1)***
6043_3	8491263	85.49%	36208
6043_4	7630757	80.26%	20857
6044_3	8203775	85.22%	29522
6044_4	8505871	83.53%	18207
6049_1	7315306	82.26%	25571
6049_4	8340748	86.18%	31046
6056_2	6335158	84.46%	32438
6056_3	6891717	82.84%	23884
6057_3	8429401	86.70%	27594
6057_5	7649365	84.60%	21241
6062_3	7646966	75.65%	14702
6062_4	7382725	85.75%	22098

3. *Populations of Potential Breast Cancer Recurrence Biomarker and Its Separation of Recurrence and Non-Recurrence Patients*

Is there any gene or sets of genes we detected in serum could be used to separate recurrence and non-recurrence patient population? We identify them (will be mentioned as "populations of potential biomarkers") with various of methods and then demonstrated their performance separating recurrence and non-recurrence patients using PCA.

Three criteria were used to identify populations of potential biomarkers: differentially expressed genes between recurrence and non-recurrence patients' serums distinguished by DESeq in R with default parameters; genes showed significant difference between the two patient populations from analysis of variance on ranked TPM; genes, between the two patient populations, showed significantly different pattern of expression changes along time. For convenience, genes in the three populations will be mentioned as "differentially expressed genes", "ANOVA select genes" and "follow-up change genes", and the three criteria will be mentioned as "differential expression", "ANOVA selection" and "follow-up change" respectively. Details of the follow-up change criteria will be mentioned in "Method". Sequencing data used in the first two methods were from all 96 libraries and data used in the last method were from two portions: 86 libraries constructed from two serums from 9 and 34 recurrence and non-recurrence patients, one collected in Range 1 and the other one collected earliest possible in Range 2, as well as 2 libraries constructed from 2056_2 and 2056_4 which were slightly out of the ranges. PCAs were then conducted on sequencing data from all 96 libraries with the populations of potential biomarkers identified using the first two criteria

and on data from 88 libraries with the population identified using “follow-up change” criteria.

There were 465 differentially expressed genes with p-value < 0.05, including 62 protein coding genes (protein-coding genes listed in Appendix Table.A4). Among the 465 genes, 353 genes showed lower expression and 112 genes showed higher expression levels in recurrence patients. 601 ANOVA select genes were identified with p-value < 0.01, among which 200 genes were protein-coding genes (protein-coding genes shown below in Table.7), including some reported by others related to breast cancer prognosis, such as TUBG1, SNAT1, AURKC and CYP2D6 [21-24]. The method follow-up change gave differential scores between 26 and -24, indicating a strong trend of the gene being over-expressed in recurrence patients and under-expressed in non-recurrence patients and the opposite respectively. There were 1259 genes with differential scores larger than 14 or smaller than -13 where the p-value is lower than 1.65×10^{-7}, among which 752 genes were protein-coding genes. Clear and obvious separation of recurrence and non-recurrence patients was observed in PCAs with all three populations of potential biomarkers. (Figure.10 and Appendix Figure.A1)

To summarize, three populations of potential breast cancer recurrence biomarkers were generated by three methods respectively. There were 465 “differentially expressed genes”, 601 “ANOVA select genes” and 1259 “follow-up change genes” among which 62, 200 and 752 protein-coding genes were included. Patient separation regarding to recurrence status were clear and obvious on PCAs with all three gene populations.

Table.7 A List of Protein-coding Genes Identified by the Method "ANOVA Selection". A list of protein-coding genes identified by applying ANOVA selection method with p-value lower than 0.01 on sequencing data from all 96 samples.

Gene Name	*Gene Name*	*Gene Name*	*Gene Name*	*Gene Name*
ANKRD65	NAT10	CMYA5	HERC1	WDR83
PRAMEF5	VPS37C	LEAP2	WDR61	CYP4F22
HSPB7	LRRC10B	ZMAT2	DET1	RAB8A
RCC2	OTUB1	EFCAB9	C16orf91	ZNF792
MINOS1-NBL1	BAD	ID4	PKMYT1	TBCB
ECE1	ESRRA	BTN2A1	PPP4C	EID2B
MDS2	MRPL49	FKBPL	CBFB	FBL
TSSK3	B4GAT1	KCTD20	KIAA0895L	B9D2
YRDC	UCP3	TBC1D22B	MVD	GRIK5
FAM159A	NARS2	GUCA1A	TUBB3	IRGC
LRRC42	GRM5	AKIRIN2	SERPINF2	DACT3
SSBP3	EXPH5	KLHL32	OR3A2	KLK8
GADD45A	MPZL3	SMPDL3A	SPATA22	AURKC
PRMT6	PVRL1	TAAR5	P2RX5	GUCD1
SRGAP2B	BLID	IFNGR1	SAT2	ASPHD2
NBPF11	AP000866.1	CCDC28A	KRBA2	ENTHD1
BOLA1	IDI1	RAET1L	TOP3A	RP5-1042K10.14
S100A5	UBE2D1	MRPL18	LGALS9C	TOB2
SDCCAG8	SEC24C	SP8	PIP4K2B	CYP2D6
C2orf50	SNCG	ZNRF2	KAT2A	DENND6B
SFXN5	KLLN	EPDR1	TUBG1	NXF2B
TXNDC9	ACTA2	C7orf57	SMG8	AMOT
DPP10	PDCD4	GUSB	POLG2	TMEM257
POTEI	LRP6	DNAJC30	NAT9	AF131216.1
NIF3L1	PFDN5	WBSCR27	ITGB4	CNOT7
VIL1	HOXC5	CCL26	SPHK1	PSD3

Table.7 A List of Protein-coding Genes Identified by the Method "ANOVA Selection" (Continued). A list of protein-coding genes identified by applying ANOVA selection method with p-value lower than 0.01 on sequencing data from all 96 samples.

Gene Name	*Gene Name*	*Gene Name*	*Gene Name*	*Gene Name*
RNF25	PDE1B	AGFG2	MGAT5B	PLEKHA2
PSMD1	NABP2	LRCH4	DEFB132	OTUD6B
AQP12B	DEPDC4	GNB2	FOXA2	CTHRC1
SRGAP3	IL17D	TRPV6	PXMP4	MED30
TBC1D5	ATP8A2	TAS2R39	PTPRT	PTPRD
THRB	DCLK1	GIMAP2	TNNC2	LURAP1L
SLC25A38	ARL11	ABCF2	ZSWIM1	APTX
TRMT10C	KCTD12	ATXN3L	SNAI1	IPPK
ZXDC	CLN5	SYAP1	GPX4	ALG2
SLITRK3	GRK1	PHEX	STK11	LCN10
CAMK2N2	NFATC4	DCAF8L1	GADD45B	PTDSS2
FRAS1	SLC10A1	PAGE5	ZNF77	OR52M1
ANXA5	RP5-1021I20.4	SLC7A3	SH3GL1	TMEM41B
PAIP1	STARD9	FAM46D	CLPP	DEPDC7

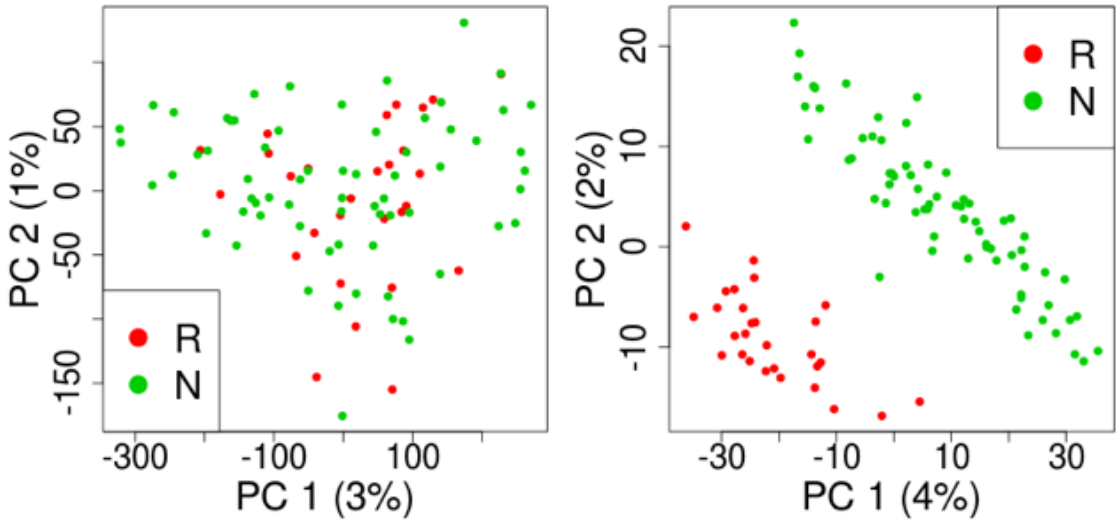


Figure.10 Principle Component Analysis Results with All Genes or the Potential Biomarker Population Identified by the Method "ANOVA Selection". PCA for All 96 Libraries Constructed from Different Serums with All 60675 human genes (ENSEMBL gene annotation (HG38)) (Left), or population of 601 potential biomarkers identified by ANOVA (Right). "R" and red dots denote recurrence patients' serums, "N" and green dots denote non-recurrence patients' serums. Clear and obvious separation could be observed on the right but not the left.

4. Risk Assessment of Breast Cancer Recurrence

Eventually, here comes the question: Is it possible to for people to assess the risk of breast cancer recurrence from microliter level of serum? To answer to question, we built a classifier based on random forest with the input being TPM of potential biomarkers identified by ANOVA selection. 40 out of 96 samples (20 from each of recurrence and non-recurrence populations) were randomly selected as training set, and the rest served as testing set. Receiver operating characteristic (ROC) curve and relative costs curve (RCC) were plotted to evaluate its performance. Repeated random sub-sampling validation was then carried out for 1000 times with exactly the same classifier building and training details mentioned above to examine repeatability and reliability.

Out of the 56 serum samples in testing set, the classifier, built as described above based on random forest, identified all 8 recurrence patients' serums correctly, while labeled 9 out of 48 non-recurrence patients' serum as recurrence patients'. True positive rate (TPR) and false positive rate (FPR) were thus calculated to be 100% and 18.75% (Table.8). In ROC plot, area under curve (AUC) was 0.992, with AUC =1 for perfect tests. The ROC curve also went extremely close to the left and top of the ROC space, indicating the good quality of the classifier. In RCC plot, area above curve (AAC) was 0.335 and the value meant misclassification cost is reasonable. In a word, ROC and RCC curves both indicated relatively good performance of the classifier (Figure.11). Results of the 1000 repeated random sub-sampling validation were presented as shown in the scatter plot and histogram (Figure.12). TPR and FPR of the 1000 classifiers are 87.86% ± 11.88% and 11.06% ± 9.73% respectively. All the classifiers showed reasonable sensitivity and specificity.

In summary, a classifier was successfully built for breast cancer recurrence risk assessment with the population of potential breast cancer recurrence marker identified by "AVONA selection". The true positive rate was 87.86% ± 11.88% and false positive rate was 11.06% ± 9.73%.

Table.8 Prediction Details of the Classifier. A total of 40 out of 96 samples (20 from each of recurrence and non-recurrence populations) were randomly selected as training set, with the rest 56 samples being test set. Prediction results of the 56 samples in testing sets were shown here. All recurrence patients in testing set were identified correctly, while 9 out of 48 non-recurrence patients' serums were identified as recurrence patients'.

Actual / Prediction	Recurrence	Non-recurrence	Predicted Total
Recurrence	8	9	17
Non-recurrence	0	39	39
Actual Total	8	48	56

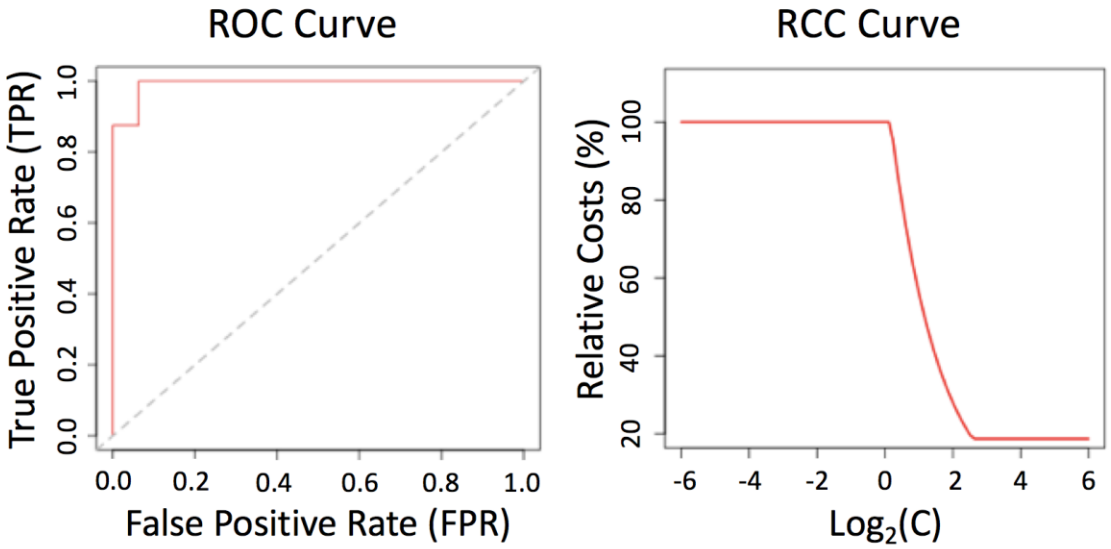


Figure.11 ROC and RCC Curves of the Classifier. The curves were generated by the same classifier mentioned in Table.8. ROC curve: as can be seen, x-axis is FPR and y-axis is TPR, representing sensitivity and 1-specificity respectively. RCC curve: x-axis is $Log_2(C)$ where C is the relative cost of false negative over false positive, y-axis is the misclassification cost at each value of C over naive misclassification cost.

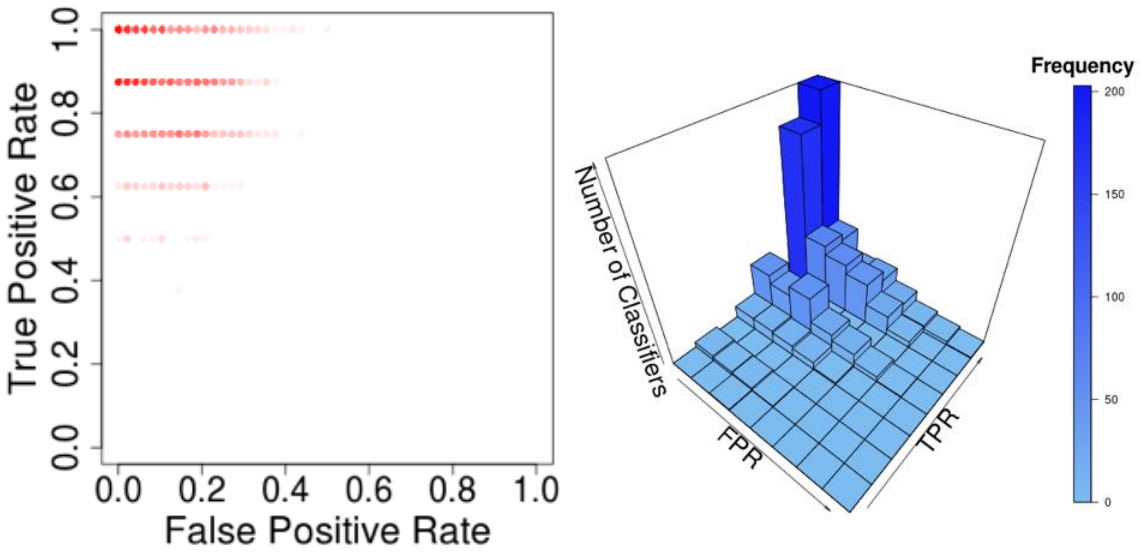


Figure.12 TPR and FPR for 1000 Times Repeated Random Sub-Sampling. Each dot in the plot on the right represents the performance of one classifier out of the total of 1000 classifiers built during repeated random sub-sampling validation. Exact same performance may occur among the 1000 classifiers, and darker red color indicates more classifiers with that performance. A more intuitive plot of performance distribution of the 1000 classifiers is shown on the left with 8 bins on each axis. TPR and FPR of the 1000 classifiers were 87.86% ± 11.88% and 11.06% ± 9.73% respectively.

Methods

I. Serum Preparation and Identification

Serums samples in this study were provided by Dr. H. Irene Su from Moores Cancer Center. Patients enrolled in the study were requested to visit the hospital around every 6 months. Standardized procedures were taken for serum preparation. At each of their enrollment and follow up visits, 10 ml whole blood specimen was drawn into serum separator tube, no anticoagulants were added during collection. With no disturbance allowed, 15-30 minutes were waited for the blood to clot at room temperature. 10-minute 1000-2000 X g refrigerated centrifugation was then carried out, and liquid supernatant, which is serum, was transferred to clean polypropylene tubes with a Pasteur pipette. 1ml aliquots were made and were stored at -80°C. The entire workflow was finished within the same day to avoid potential degradation of compositions in serum.

Serums were identified with codes (IDs) in the format of "PatientID_fu". "PatientID" were unique 4-digit numbers each assigned to a patient. "fu" were 1-2-digit numbers representing at which follow-up visit were serums collected, the smaller the number was, the earlier the collection happened.

II. Cell-free RNA Extraction and Characterization

RNA was extracted from various volumes of serums with the following four kits/methods, following the protocols. Kits/methods were listed below with details:

1. TRIzol LS Reagent (Invitrogen, Cat# 10296010): This reagent uses phenol-chloroform (1:4 v/v ratio) to extract RNA from specifically liquid samples. 40ul volume of Cell-free RNA was extracted from 1ml serum following manufacturer's protocol.

2. exoRNeasy Serum/Plasma Midi Kit (QIAGEN, Cat# 77044): This kit extracts cell-free RNA by firstly isolating exosomal vesicle with membrane-based affinity binding column, then extracting RNA from eluted vesicles with a combination of QIAzol, which is a phenol-chloroform reagent, and silica-based membrane column purification technology. 14ul volume of RNA was extracted from 1ml serum following manufacturer's protocol.

3. Plasma/Serum RNA Purification Kit (NORGEN Biotek, Cat# 56100), followed by RNase-Free DNase I set (QIAGEN, Cat# 79254) and RNeasy MinElute Cleanup Kit (QIAGEN, Cat# 74204): Plasma/Serum RNA Purification Kit uses two columns feature spin column chromatography in series for cell-free RNA extraction. The first handles large volume input and the second conducts concentration. 50ul volume of RNA was extracted from 1ml serum following manufacturer's protocol (with no optional on column DNase I digestion). In liquid DNase I treatment was then conducted with RNase-Free DNase I set and RNeasy MinElute Cleanup Kit following manufacturer's protocols. 14ul volume of purified RNA were harvested from 14ul out of the 50ul extracted RNA above.

4. QIAzol Method [11]: The method extracts cell-free RNA with phenol-chloroform reagent QIAzol. 20ul volume of cell-free RNA was extracted according to the instruction in literature from 200ul serum, DNase I treatment mentioned in the literature was not conducted. As reported by the literature, the method successfully extracted cell-free RNA

from human saliva with high yield, and due to similarity between human saliva and serum, the method was also used here to extract cell-free RNA from serum.

For convenience, the four methods above will be mentioned as “TRIzol LS”, “exoRNeasy”, “NORGEN w/ DNase” and “QIAzol Method” respectively. Bioanalyzer RNA Pico 6000 (Agilent, Cat# 5067-1513) runs were then conducted for quantification and size distribution. At least one well on each Bioanalyzer chip was loaded with 1ul ultrapure water (ThermoFisher, Cat# 10977023) as negative control.

III. Library Construction

Major compositions of serum and cell after lysis buffer treatment are similar to the extent that they both consisted of mostly broken lipid bilayer vesicles, proteins and nucleic acids. Based on this similarity and the goal of low serum input volume, we applied the Switching Mechanism At the 5’ end of RNA Template (SMART) Sequencing (SMART-Seq) [10], which was originally used for RNA library construction with cell input, to construct full-length RNA libraries from serum. General procedures of library construction are shown below, most of the steps were believed could be conducted by a commercially available kit Ovation® SoLo RNA-Seq System (NuGEN, Cat# 0500).

Starting with 3-15ul unprocessed serum in the tube, ultrapure water was added to those tubes with less than 7ul serum to achieve a total volume of 7ul. 5ul cell lysis buffer (NuGEN, Cat# 0500, User Guide.V.A) was then added to break up exosomes and dissociate protein-nucleic acid complexes, the mixture was thoroughly mixed with pipette for best breaking and dissociation efficiency. Next, potential genomic DNA contamination was removed by HL-dsDNase (NuGEN, Cat# 0500, User Guide.V.A).

First strand cDNA was synthesized and fragmented followed by second strand synthesis (NuGEN, Cat# 0500, User Guide.V.A, B, E, F). Afterwards, barcoded adaptors were ligated to cDNA molecules in each library for PCR amplification and sequencing multiplexing (NuGEN, Cat# 0500, User Guide.V.G-I). After 18 rounds PCR amplification (NuGEN, Cat# 0500, User Guide.V.L), rRNA depletion was performed (NuGEN, Cat# 0500, User Guide.V.M-N). Library construction was finally completed with another 8 rounds of PCR amplification for signal enrichment (NuGEN, Cat# 0500, User Guide.V.O-P), and was ready for downstream process. [12]

IV. Sequencing, De-multiplexing and Mapping

Before sequencing was performed, libraries were subject to Bioanalyzer quality check with High Sensitivity DNA Kit (Agilent, Cat# 6057-4626) in terms of size, purity and concentration. Sequencing was then conducted with MiniSeq (Illumina) or HiSeq 4000 (Illumina). Sequencing with MiniSeq was performed according to manufacturer's protocol using MiniSeq High Output Reagent Kit (75-cycle) (Illumina, Cat# FC-420-1001). Single-end sequencing with 75 read-1 cycles and 8 index-1 cycles was performed and each sequencing pool consisted of 8 different libraries. HiSeq 4000 sequencing run was completed with the service provided by Institute of Genomic Medicine (IGM) Genomics Center, UCSD, in general, bridge amplification on cBot first and then sequencing-by-synthesis on 2 out of 8 lanes of HiSeq 4000 flow cell. Single-end sequencing with 50 read-1 cycles and 16 index-1 cycles were conducted and the pool consisted of 96 different libraries. Unless otherwise specified, sequencing was performed on MiniSeq with parameters mentioned above or by default if not mention. Besides,

"custom R1 sequencing primer" (NuGEN, Cat# 0500-S02225) could also be used to replace illumina standard sequencing primer comes with the sequencing reagent kit or cartridge (Figure.13, [12]). Unless otherwise specified, custom R1 sequencing primer was loaded according to the protocol "MiniSeq System Custom Primers Guide" and "HiSeq System Custom Primers Guide" for sequencing.

Sequencing data from MiniSeq runs were automatically de-multiplexed with Basespace (Illumina, basespace.illumina.com) and sequencing data from HiSeq were manually de-multiplexed with the first 8 of the 16 index-1 reads using the tool fastq-multx [13]. FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) was used for quality control. The data was then mapped to ENSEMBL gene annotation (HG38) [14] with the mapping software STAR [15]. Read counts of each gene were given by the software HTSeq-count [16]. Transcripts per million (TPM)values were calculated for each gene for expression levels. Tissue specific genes based on expressed sequence tags (mentioned as "tissue specific genes" for short) were extracted from the database TiGER (http://bioinfo.wilmer.jhu.edu/tiger/).

V. Using Illumina Standard Sequencing Primer in Section "Optimization with Addition of Custom R1 Sequencing Primer"

Illumina standard sequencing primer (Figure.13, [2]) was used only when generating the "without custom R1 sequencing primer" results in section "Optimization with Addition of Custom R1 Sequencing Primer". The primer is integrated in sequencing reagent kit (Illumina, Cat# FC-420-1001). To summarize, results of "without custom R1 sequencing primer" in section "Optimization with Addition of Custom R1 Sequencing

Primer" was generated by sequencing constructed libraries on MiniSeq (Illumina) using MiniSeq High Output Reagent Kit (75-cycle) (Illumina, Cat# FC-420-1001) with illumina standard sequencing primer integrated in the reagent kit. Single-end sequencing with 75 read-1 cycles and 8 index-1 cycles was carried out and each sequencing pool consisted of 8 different libraries.

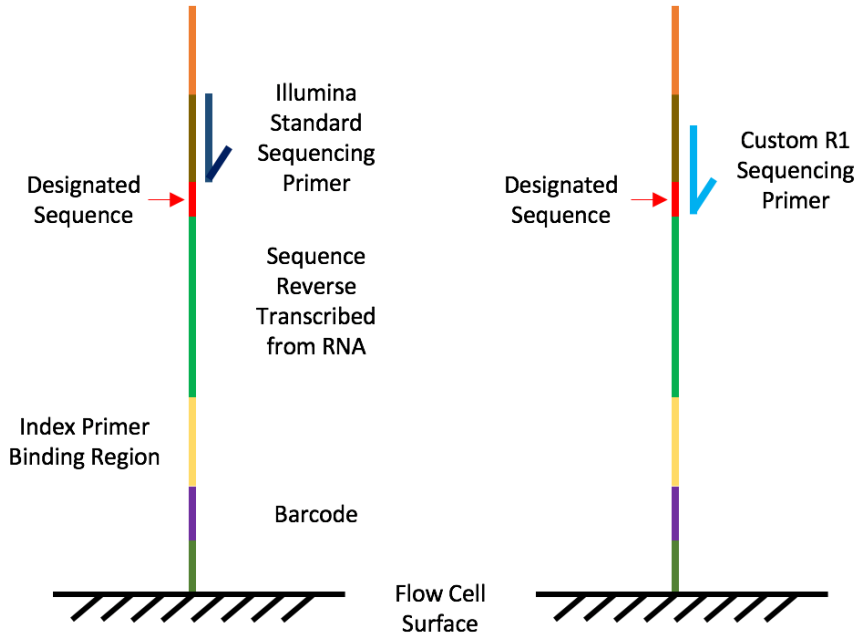


Figure.13 Target Base-pairing Regions of Illumina Standard Sequencing Primer and Custom R1 Sequencing Primer. Target base-pairing regions of illumina standard sequencing primer (dark blue) and custom R1 primer (light blue). As can be seen, the process of library construction added the same short designated sequence (red) between adaptors and sequences reverse transcribed from RNA. The difference between sequencing with illumina standard sequencing primer and with custom R1 primer is that this designated region will be sequenced with the former primer, causing drops in cluster number and cluster passing filter rate but will not be sequenced with the latter one.

VI. Statistical Analysis and Classification

1. Differential Expressed Genes Analysis

Differential expressed genes were identified using DESeq package in R [17]. The input was read counts data for all 60675 human genes (ENSEMBL gene annotation (HG38)) of all 96 samples, the output was p-value for each gene being differentially expressed between recurrence and non-recurrence groups. After this, we extracted the genes with a p-value lower than 0.05 as differential expressed genes.

2. Analysis of Variance

Analysis of variance were carried out using function "aov" and the model in eq. 1 in R [18]. The input was gene TPM for all the 60675 human genes (ENSEMBL gene annotation (HG38)) of all 96 samples and the output was p-value for each gene being significant in one-way analysis of variance. Then, genes with a p-value lower than 0.01 were extracted as differential expressed genes.

ANOVA model:

$$E_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

where E_{ij} is the observed value of the jth gene of group i (in our study, group set is recurrence and non-recurrence group), μ is the combined population mean, α_i is the fixed deviation of the mean of group i from mean μ, and ε_{ij} is a random deviation of the jth gene of group i from $\mu + \alpha_i$.

Corresponding computational model:

$$SS_{total} = n * (E_i - \overline{E})^2$$

$$SS_{bg} = n_R * (\overline{E_R} - \overline{E})^2 + n_N * (\overline{E_N} - \overline{E})^2$$

$$SS_{wg} = SS_{total} - SS_{bg}$$

$$F = \frac{\frac{SS_{bg}}{df_{bg}}}{\frac{SS_{wg}}{df_{wg}}}$$

eq. 1

where SS_{total} is the sum of squares for the gene expression in all the samples, SS_{bg} is the between groups sum of squares, SS_{wg} is the within groups sum of squares, n is the number of total samples, whereas n_R is the number of samples from group R (recurrence group), n_N is the number of samples from group N (non-recurrence group), while E_i is the gene expression level for the ith sample, $\overline{E}$ is the average gene expression level across all the samples, $\overline{E_R}$ and $\overline{E_N}$ are the average gene expression level in group R and N respectively, and F is the F-score used for F-test, df_{bg} and df_{wg} are the degree of freedom for between group and within group respectively.

3. <u>Follow-up Change Analysis</u>

Follow-up change analysis used the change of expression from the first follow-up to the second follow-up (two eligible follow-ups were selected for each patient as described before) as input to calculate a score "s", defined below, (ranges from -44 to 44) for each gene. The score indicates the different expression change trend between recurrence patient and non-recurrence patient.

$$s = (r_{up} - r_{down}) - (n_{up} - n_{down})$$

where s is the differential score, r_{up} and r_{down} are frequency that expression level of the specific gene was observed going up or down with time in recurrence patient population, n_{up} and n_{down} are frequency that expression level of the gene was observed going up or down with time in non-recurrence patient population.

As an example, in the table below (Table.9), 10 serums are collected from 5 patients' (2 recurrence patients and 3 non-recurrence), and the 1st follow-up is before the 2nd follow-up in terms of collecting date. For Gene 1 in recurrence patients, 2 of 2 are observed to go up with time, 0 of them is observed to go down, while in non-recurrence patients, 1 of 3 is observed to go up and 2 of them are observed to go down. Thus, in this case, $r_{up} = 2$, $r_{down} = 0$, $n_{up} = 1$ and $n_{down} = 2$, while we get the final score $s = 3$. (Table.9)

The score s therefore represents how expression level of a gene would change with time (i.e. increase or decrease) in recurrence or non-recurrence scenarios. A larger s for a gene indicates a stronger trend of it being over-expressed in recurrence patients and under-expressed in non-recurrence patients, and a smaller s means the opposite.

Table.9 Gene 1 TPM Information in Example Patients. As an example to show the calculation of score s, there are 5 patients (2 recurrence and 3 non-recurrence), from each of whom, 2 serums are collected (marked as 1st and 2nd fu in the second line) and Gene 1's TPM is measured in all the serums as shown in the third line.

Patient status	*R*		*R*		*N*		*N*		*N*	
fu	1st fu	2nd fu	1st fu	2nd fu	1st fu	2nd fu	1st fu	2nd fu	1st fu	2nd fu
Gene 1 (TPM)	0.0	5.4	0.1	7.5	1.3	1.2	2.4	1.1	1.0	1.5

4. Principal Component Analysis

Principal component analysis (PCA) was conducted using function “prcomp” in R, with the input of log of TPM for each gene.

5. Classification Model

A classifier was built using R package “randomForest” [19]. When building the classifiers, each time 40 samples were randomly selected as training set (20 recurrence samples and 20 non-recurrence samples to balance the training procedure), and the rest were left as testing set. The default parameters were used (ntree=500 and cutoff=0.5) during training process. Repeated random sub-sampling validation was performed by repeatedly re-sample the training set 1000 times with the same classifier building and training details above.

Discussion

With the reported around 3%-23% chance of local recurrence rate [20], risk assessment and prognosis biomarker of recurrence is desired by the field. With various of advantages such as easy accessibility and high information density etc., the importance of cell-free RNA in serum is now being more and more emphasized in prognosis and diagnosis of cancer as well as other diseases. However, development of the field was to a large extent impeded by the large serum volume required to extract and purify RNA, which was the first step of now commonly used RNA-Seq technology. We developed a novel technology being able to construct RNA-Seq library from only microliters of unprocessed serum in large scale and then applied it to breast cancer recurrence case for its ability to separate recurrence and non-recurrence patients and assess recurrence risk. The study could benefit not only academia but also the society with both understandings of recurrence and the technology itself as a potential clinical application.

In this study, cell-free RNA sequencing libraries were firstly constructed directly from microliters of unprocessed human serums. Diversity of genes in serum was then observed. Afterwards, with sequencing data of libraries constructed with 96 serum samples from 10 recurrence and 34 non-recurrence patients, populations of potential biomarkers were identified with various statistical analysis tools, most importantly ANOVA. Finally, risk assessment of breast cancer recurrence was performed with the 96 serums using a Random Forest classifier with ANOVA-identified population of biomarker as feature. Almost all human genes were detected in serum, their huge diversity in terms of mainly gene categories was also observed. Existence of some genes

and categories in serum was previously reported [24], yet existence of others was not, this was expected due to the novelty of the technology: Libraries here were constructed directly from unprocessed serum, while currently the most commonly used RNA-Seq library construction method is based on the input of purified RNA instead of unprocessed biopsy itself, loss of low expressed information during RNA purification step would potentially case the discrepancy above. Various of genes were identified as potential biomarkers, including some of those previously reported by the literature [21-24]. Random forest classifiers built identified breast cancer recurrence with a true positive rate of 87.86% ± 11.88% and false positive rate of 11.06% ± 9.73%.

When compared to the “traditional” “first RNA purification from large volume of sample then library construction” method commonly used now, our technology provided much more complete coverage of human genes. In addition, application of the technology on breast cancer recurrence case further revealed its potential in clinical applications. In addition, there are also some directions worth more efforts. With the basic principle being breaking exosome and ribonucleoprotein complexes to expose cell-free RNA to library construction enzymes, the technology is not specific to the type of liquid biopsy. It is believed that this low starting volume cell-free RNA sequencing technology could be performed on other liquid biopsy providing chemical composition, especially the mechanism of cell-free RNA existence and protection, is the same with serum. Besides, the result that almost all human genes could be found in serum also indicates potential applications of the technology on diagnosis and prognosis for other diseases. Expansion on liquid biopsy type and disease type thus should be carried out to fully release its potential.

However, limitations also exist in the study. Risk assessment of breast cancer recurrence in this report was still based on classifiers where the features were defined by applying statistical analysis tool on the same 96 serums used for training and testing. Though what mentioned above is a common practice when study size is limited, this would to some extent affect the accuracy of classifiers, a stricter rule of feature selection is to define features of the classifiers with only the samples used for training. Inclusion of more patients/serums as well as deeper analysis of the data were needed for better biomarker analysis and recurrence risk assessment. What's more, with limited time, most of the library construct procedures were conducted with the kit Ovation® SoLo RNA-Seq System (NuGEN, Cat# 0500), more effort should be spent to further validate the work with other enzymes and reagents from the market.

To summarize, a technology being able to construct RNA-Seq libraries directly from microliters of unprocessed serum was developed. The technology was then applied to breast cancer recurrence case for its performance and potential in real life clinical problems. Plenty of human genes among lots of categories could be found in serum, including tissue specific genes. Based on sequencing data generated with developed technology, potential recurrence biomarker populations were identified, recurrence and non-recurrence patients were successfully separated, risk assessment of breast cancer recurrence was also conducted.

Conclusion

A technology was successfully developed to construct cell-free RNA sequencing libraries from only microliters of unprocessed serum, its application on breast cancer recurrence also demonstrated the potential on separating recurrence and non-recurrence breast cancer patients and recurrence risk assessment. Library construction method was firstly optimized for sequencing, mapping quality, gene complexity and automation. Large diversity of genes in serum were then surveyed and described in terms of their categories and tissue specificity. Finally, three populations of potential breast cancer recurrence biomarkers were identified with statistical analysis tools and risk assessment of breast cancer recurrence was successfully conducted with a random forest classifier. Though limitations such as possible overfitting caused by relatively small study size do exist, the technology developed here could greatly benefit both academia and clinic fields. Increase of the study size, deeper investigation of the potential biomarkers and classifiers were needed, and more work should also be pursued to fully exploit its potential on various liquid biopsy and diseases.

References

[1] Melo, S. A., Sugimoto, H., O'Connell, J. T., Kato, N., Villanueva, A., Vidal, A., … Kalluri, R. (2014). Cancer Exosomes Perform Cell-Independent MicroRNA Biogenesis and Promote Tumorigenesis. *Cancer Cell*, *26*(5), 707–721. https://doi.org/10.1016/j.ccell.2014.09.005

[2] O'Driscoll, L. (2015). Expanding on exosomes and ectosomes in cancer. *New England Journal of Medicine*, *372*(24), 2359–2362.

[3] Feng, G., Li, G., Gentil-Perret, A., Tostain, J., & Genin, C. (2008). Elevated serum-circulating RNA in patients with conventional renal cell cancer. *Anticancer Research*, *28*(1A), 321–326.

[4] Rykova, E. Y., Wunsche, W., Brizgunova, O. E., Skvortsova, T. E., Tamkovich, S. N., Senin, I. S., … Vlassov, V. V. (2006). Concentrations of Circulating RNA from Healthy Donors and Cancer Patients Estimated by Different Methods. *Annals of the New York Academy of Sciences*, *1075*(1), 328–333. https://doi.org/10.1196/annals.1368.044

[5] "Plasma/Serum RNA Purification Midi Kit (Cat. 56100)." Plasma/Serum RNA Purification Midi Kit (Cat. 56100) | Norgen Biotek Corp. NORGEN Biotek Corp, n.d. Web. 06 July 2017.

[6] Schwarzenbach, H. (2013). Circulating nucleic acids as biomarkers in breast cancer. *Breast Cancer Research*, *15*, 211. https://doi.org/10.1186/bcr3446

[7] Kishikawa, T., Otsuka, M., Ohno, M., Yoshikawa, T., Takata, A., & Koike, K. (2015). Circulating RNAs as new biomarkers for detecting pancreatic cancer. *World Journal of Gastroenterology : WJG*, *21*(28), 8527–8540. https://doi.org/10.3748/wjg.v21.i28.8527

[8] Schwarzenbach, H., Hoon, D. S. B., & Pantel, K. (2011). Cell-free nucleic acids as biomarkers in cancer patients. *Nature Reviews Cancer*, *11*(6), 426–437. https://doi.org/10.1038/nrc3066

[9] Dong, L., Lin, W., Qi, P., Xu, M.-D., Wu, X., Ni, S., … Du, X. (2016). Circulating Long RNAs in Serum Extracellular Vesicles: Their Characterization and Potential Application as Biomarkers for Diagnosis of Colorectal Cancer. *Cancer Epidemiology, Biomarkers & Prevention*, *25*(7), 1158–1166. https://doi.org/10.1158/1055-9965.EPI-16-0006

[10] Ramsköld, D., Luo, S., Wang, Y.-C., Li, R., Deng, Q., Faridani, O. R., … Sandberg, R. (2012). Full-length mRNA-Seq from single-cell levels of RNA and individual

circulating tumor cells. *Nature Biotechnology*, *30*(8), 777–782. https://doi.org/10.1038/nbt.2282

[11] Pandit, P., Cooper-White, J., & Punyadeera, C. (2013). High-Yield RNA-Extraction Method for Saliva. *Clinical Chemistry*, *59*(7), 1118–1122. https://doi.org/10.1373/clinchem.2012.197863

[12] Ovation® SoLo RNA-Seq System. (2017, July 11). Retrieved July 18, 2017, from http://www.nugen.com/products/ovation-solo-rna-seq-system

[13] Aronesty, E. (2013). Comparison of Sequencing Utility Programs. *The Open Bioinformatics Journal*, *7*(1). Retrieved from https://benthamopen.com/ABSTRACT/TOBIOIJ-7-1

[14] Yates, A., Akanni, W., Amode, M. R., Barrell, D., Billis, K., Carvalho-Silva, D., ⋯ Flicek, P. (2016). Ensembl 2016. *Nucleic Acids Research*, *44*(D1), D710–D716. https://doi.org/10.1093/nar/gkv1157

[15] Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., ⋯ Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

[16] Anders, S., Pyl, P. T., & Huber, W. (2015). HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics*, *31*(2), 166–169. https://doi.org/10.1093/bioinformatics/btu638

[17] Anders, S. (2010). Analysing RNA-Seq data with the DESeq package. *Mol Biol*, *43*(4), 1–17.

[18] Chambers, J. M., Freeny, A and Heiberger, R. M. (1992) Analysis of variance; designed experiments. *Chapter 5 of Statistical Models in Seds J. M.*

[19] Liaw, A., Wiener, M., & others. (2002). Classification and regression by randomForest. *R News*, *2*(3), 18–22.

[20] What are the Risks of Breast Cancer Returning? (n.d.). Retrieved July 18, 2017, from https://ww5.komen.org/BreastCancer/SurvivalandRiskofHavingCancerReturnAfterTreatment.htm

[21] Martin, K. J., Patrick, D. R., Bissell, M. J., & Fournier, M. V. (2008). Prognostic Breast Cancer Signature Identified from 3D Culture Model Accurately Predicts Clinical Outcome across Independent Datasets. *PLOS ONE*, *3*(8), e2994. https://doi.org/10.1371/journal.pone.0002994

[22] Herreros, A. G. de, Peiró, S., Nassour, M., & Savagner, P. (2010). Snail Family Regulation and Epithelial Mesenchymal Transitions in Breast Cancer Progression. *Journal of Mammary Gland Biology and Neoplasia*, *15*(2), 135–147. https://doi.org/10.1007/s10911-010-9179-8

[23] Romanelli, A., Clark, A., Assayag, F., Chateau-Joubert, S., Poupon, M.-F., Servely, J.-L., … Marangoni, E. (2012). Inhibiting Aurora Kinases Reduces Tumor Growth and Suppresses Tumor Recurrence after Chemotherapy in Patient-Derived Triple-Negative Breast Cancer Xenografts. *Molecular Cancer Therapeutics*, *11*(12), 2693–2703. https://doi.org/10.1158/1535-7163.MCT-12-0441-T

[24] Dezentjé, V. O., van Blijderveen, N. J. C., Gelderblom, H., Putter, H., van Herk-Sukel, M. P. P., Casparie, M. K., … Guchelaar, H.-J. (2010). Effect of Concomitant CYP2D6 Inhibitor Use and Tamoxifen Adherence on Breast Cancer Recurrence in Early-Stage Breast Cancer. *Journal of Clinical Oncology*, *28*(14), 2423–2429. https://doi.org/10.1200/JCO.2009.25.0894

Appendix

Table.A1 Averages and Standard Deviations Calculated to Analyze the Effect of Increasing Serum Input Volume on Sequencing and Mapping Qualities. "Serum Sample" shows the serum sample with the format of "PatientID_fu". "Serum Input Volume (ul)" shows volume of serum input to construct the library. "Sample total Reads Counts" tab indicates data below were calculated from sample total read counts values and same for "Uniquely Mapped Percentage" tab. "Average" and "SD" under each of the two tabs give averages and standard deviations calculated in each condition. "N =" shows the number of samples used to calculate "Average" and "SD". For example, the very first row means there were two libraries constructed with 3ul input of serum 2010_4, and their average sample total read counts was 4424293.5 with SD being 7170.5, and their average uniquely mapped percentage being 80.30% with SD being 0.0027 (or 0.27%).

Serum Sample	Serum Input Volume (ul)	Sample Total Read Counts		Uniquely Mapped Percentage		N =
		Average	SD	Average	SD	
2010_4	3	4424293.50	7170.50	80.30%	0.0027	2
	7	4812662.50	337278.50	87.99%	0.0073	2
6004_9	3	4756960.25	343721.22	81.87%	0.0103	4
	7	5548050.50	1028693.50	87.13%	0.0086	2
6057_5	3	3989951.60	323523.02	71.70%	0.1212	5
	7	4929965.50	382189.14	85.98%	0.0084	4

Tabel.A2 Averages and Standard Deviations Calculated to Analyze the Effect of Custom R1 Sequencing Primer on Sequencing and Mapping Qualities. Upper, a): Averages and standard deviations calculated for analysis of custom primer effect on sequencing quality. "Serum Sample" shows the serum sample with the format of "PatientID_fu". "Y" and "N" means "Yes"/ "With" and "No"/ "Without" respectively in the column "With Custom Primer". "Average of Sample Total Read Counts" and "SD" are averages and standard deviations calculated from sample total read counts in each condition. "N =" shows the number of samples used to calculate "Average of Sample Total Read Counts" and "SD".
Lower, b) Averages and standard deviations calculated for analysis of custom primer effect on mapping quality. Table content same with that in a), except "Average of Sample Total Read Counts" and "SD" are replaced with "Average of Uniquely Mapped Percentage" and "SD". These numbers are averages and standard deviations of uniquely mapped percentage calculated from uniquely mapped percentage in each condition.

Serum Sample	With Custom Primer	Average of Sample Total Read Counts	SD	N =
6004_9	Y	4756960.25	343721.22	4
	N	2366266.50	270566.63	6
6057_5	Y	3989951.60	323523.02	5
	N	2359712.50	333421.14	6

Serum Sample	With Custom Primer	Average of Uniquely Mapped Percentage	SD	N =
6004_9	Y	81.87%	0.0103	4
	N	73.73%	0.1115	6
6057_5	Y	71.70%	0.1212	5
	N	74.16%	0.0484	6

Tabel.A3 Averages and Standard Deviations Calculated to Analyze the Effect of Increasing Serum Input Volume on Library Complexity. Library complexity was represented by number of genes detected. “Serum Sample” shows the serum sample with the format of “PatientID_fu”. “Serum Input Volume (ul)” shows volume of serum input to construct the library. “Average of Number of Genes Detected” and “SD” are averages of number of genes detected (TPM>1) and the corresponding standard deviations calculated in each condition. “N =” shows the number of samples used to calculate “Average” and “SD”. For example, the very first row means there were two libraries constructed with 3ul input of serum 2010_4, and their average number of genes detected was 20788 with SD being 538.

Serum Sample	Serum Input Volume (ul)	Average of Number of Genes Detected	SD	N =
2010_4	3	20788.00	538.00	2
	7	33915.00	1438.00	2
6004_9	3	20040.75	2499.80	4
	7	31875.00	2541.00	2
6057_5	3	15419.00	2583.95	5
	7	32056.75	2487.41	4

Table.A4 A List of Protein-coding Genes Identified by the Method "Differential Expression". A list of protein-coding gene identified by applying differential expression method with p-value lower than 0.05 on sequencing data from all 96 samples.

Gene Name	*Gene Name*	*Gene Name*	*Gene Name*	*Gene Name*	*Gene Name*	*Gene Name*
B4GALT2	C1orf53	HSPB3	RAET1L	TUT1	RPL36AL	UCKL1
KLF18	DIRC1	HIGD2A	MRPL32	C11orf86	C14orf169	PEX11G
BEST4	CRYGD	RAB24	SPDYE5	LHPP	DET1	GEMIN7
MAGOH	CDHR4	OR2Y1	AF131216.1	TAS2R20	FA2H	APOL1
LRRC42	STX19	HIST1H4D	FAM86B1	YEATS4	TXNDC17	DNAL4
WLS	GAP43	HLA-A	FABP9	RP11-310K10.1	KRBA2	NPTXR
FAM72D	CHST13	FKBPL	CNTNAP3B	OR5AU1	KAT2A	CYP2D6
S100A5	ZDHHC19	CCDC167	FBP2	OR4E2	NNAT	KRTAP13-1
TMEM79	UGT2A2	TREML1	INPP5E	CMA1	DOK5	

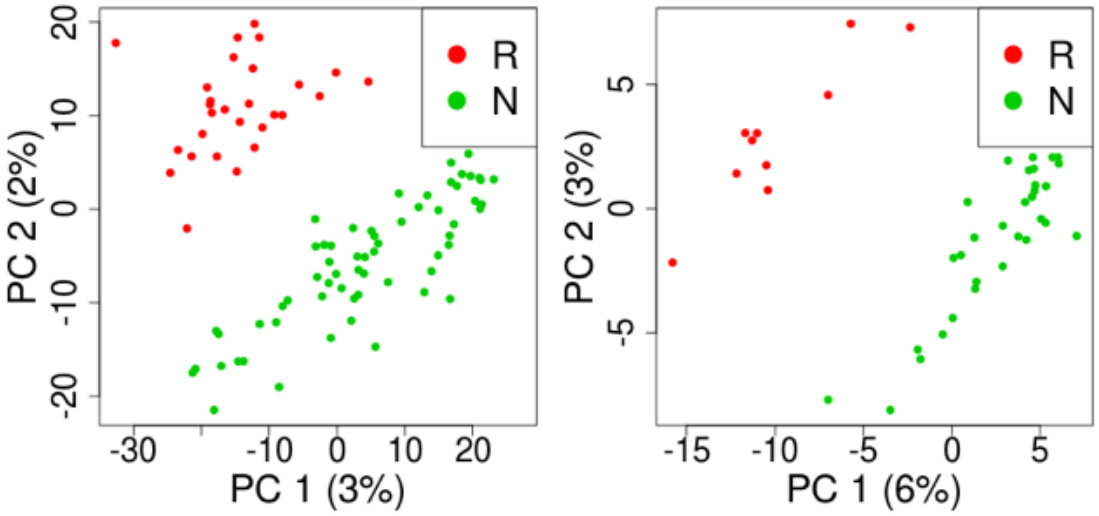


Figure.A1 Principle Component Analysis Results for All Serum Samples/Patients Using Different Potential Biomarker Populations as Features. Left: PCA for all 96 libraries constructed from different serums with population of 465 potential biomarkers identified by differential expression method. Right: PCA of 44 patients (10 recurrence and 34 non-recurrence) using 88 libraries constructed as input with population of 1259 potential biomarkers identified by follow-up change method. “R” and red dots are recurrence patients and serums, “N” and green dots are non-recurrence patients and serums. Clear and obvious separation could be observed in both results.