Functional analysis of differential mRNAs in cancer peripheral blood: reflection of population shifts in myeloid-origin and lymphoid-origin cells

Functional analysis of cancer peripheral blood differential mRNAs

Guini Hong, Hongdong Li, Wenjing Zhang, Zheng Guo
Bioinformatics Centre, School of Life Science,
University of Electronic Science and Technology of China,
Chengdu, China
hongguini08@gmail.com, biomantis_lhd@163.com,
youleyouyu@126.com,guoz@ems.hrbmu.edu.cn

Beibei Chen, Hui Xu, Lu Ao, Zheng Guo
College of Bioinformatics Science and Technology,
Harbin Medical University
Harbin, China
bbc0420@hotmail.com, xhh198951@163.com,
lukeyfj@gmail.com, guoz@ems.hrbmu.edu.cn

***Abstract*—Functional enrichment analysis is usually adopted after the identification of differentially expressed (DE) genes in studies focusing on cancer peripheral blood (PB) gene expression. However, whether the disturbed functional signals reflect the expression changes in blood cells or the cell population shifts under cancer condition remains unclear. By deconvolving the gene expression profiles of multiple cancer datasets, we showed that the proportion of myeloid-origin cells increased whereas the proportion of lymphoid-origin cells decreased in cancer PB. The DE genes between cancer PB samples and controls were highly consistent with DE genes between myeloid-origin and lymphoid-origin cells, indicating that cell population shifts contributed predominantly to the differential signals in cancer PB. All of the functional categories enriched for cancer PB DE genes were enriched for DE genes between myeloid-origin and lymphoid-origin cells, suggesting that functional signals in cancer PB probably reflect the changes of population shifts in blood cells, thus the enriched functional categories might not be able to reflect the cell type specific expression changes. Therefore, caution should be taken in translational biomarker discovery based on human PB gene expression profiles.**

Keywords—Cancer peripheral blood; differentially expressed genes; functional enrichment analysis; myeloid-origin and lymphoid-origin cells

I. INTRODUCTION

Recently, blood gene signatures as noninvasive clinical biomarkers have been successfully identified based on the detection of mRNA expression profiles in human peripheral blood samples from many diseases, such as Alzheimer [1], stroke [2], inflammation-related disease [3] and cancer [4-6]. After finding the candidate discriminatory genes or significantly differentially expressed (DE) genes from a peripheral blood (PB) cancer dataset, researchers usually perform the functional enrichment analysis to detect the significant functional categories enriched for these candidate genes [4-7]. However, as PB is known to be a mixture of various types of blood cells [8], whether such functional signals observed in PB could reflect the blood cell-intrinsic expression changes is doubtful. Moreover, it has been reported that the number of blood cells originated from the myeloid precursor (referred to as myeloid-origin cells for simplicity) tend to increase whereas the number of blood cells originated from the lymphoid precursor (referred to as lymphoid-origin cells for simplicity) tend to decrease under cancer condition [9-11]. This means that, in addition to the expression changes intrinsic in blood cells, the cell population shifts could also contribute to the DE genes observing in cancer PB samples compared to controls. Therefore, the functional analysis of significant DE genes may only provide important information about the blood cells but not directly about the cancer, the latter of which is often interpreted by the researchers focused on PB gene expression study [4-7].

In this report, we first showed that the proportion of myeloid-origin cells increased and the proportion of lymphoid-origin cells decreased in cancer PB samples compared to controls. Then, we reported that directions of regulation (up- or down-regulations) of DE genes in cancer PB samples were highly consistent with that of the DE genes detected between myeloid-origin and lymphoid-origin cells. These cancer PB DE genes with elevated expression levels identified from different cancers enriched in functional categories including many immune- and inflammation-related biological processes. The down-regulated genes in PB samples of different cancers involved in functional categories mainly related to protein synthesis, translation control and cellular metabolic. These functional categories were all enriched for the DE genes identified between myeloid-origin cells and lymphoid-origin cells but not enriched for DE genes identified between breast cancer and normal CD4+ T cells. Thus, we concluded that the commonly disturbed functional categories observed in different cancers overwhelmingly reflected the shifts of the cell

978-1-4799-7294-4/14/$31.00 ©2014 IEEE

populations and they may be probably not reflective of the cell type specific expression changes. Therefore, the connection of functional categories enriched by DE genes identified from cancer blood directly to cancer might be misleading as the mRNA expression profiles in PB samples were more likely to reflect the expression changes induced by population shifts in blood cells as a response to the cancer condition.

II. METHOD

A. Microarray Data

We collected two PB microarray datasets for each of the three cancer types studied from the GEO database [12] respectively (Table I). Because no enough PB datasets were available for ovarian cancer, only one dataset was analyzed in this report. The normalized data were downloaded from GEO and the original platform annotation files released from GEO were used to annotate the CloneIDs to GeneIDs.

The dataset for breast CD4+ T cells contained only the gene expression profiles for purified CD4+ T cells from breast cancer and healthy controls, which were from the GEO series (GEO accession number 'GSE36765').

The two datasets for normal leukocyte measured the gene expression levels of different leukocyte subtypes from normal human peripheral blood. In each dataset, the gene expression profiles of normal human leukocyte subtypes were divided into two groups: one group was composed of the profiles of myeloid-origin cells including monocytes, neutrophils and eosinophils while the other group was composed of the profiles of lymphoid-origin cells including T cells, NK cells and B cells. In Table I, "Case" refers to the myeloid-origin group, while "Control" refers to the lymphoid-origin group.

B. Estimation of Proportions of Myeloid-origin and Lymphoid-origin Cells in Peripheral Blood

To determine whether the myeloid-origin and lymphoid-origin cell proportions differ in the PB of cancer patients, we quantified the proportions of myeloid-origin and lymphoid-origin cells by a process of deconvolution [13]. If B represents the known matrix of microarray expression profiles measured for a disease, comprising both disease and control samples; X represents the proportions of myeloid-origin and lymphoid-origin cells; and A represents the known matrix of expression levels of genes in the myeloid-origin and lymphoid-origin cells, then

$$A X \approx \mathrm{B} \tag{1}$$

The object of deconvolution is to find the solution of the deconvolution equation, which will give the cell-type proportions for myeloid-origin and lymphoid-origin cells. Thus, based on a cancer PB microarray dataset and expression profiles of the marker genes specially expressed on myeloid-origin and lymphoid-origin cells documented in the Immune Response in Silico (IRIS) database [14], we could estimate the proportions for myeloid-origin and lymphoid-origin cells in each sample in the cancer PB dataset.

After the proportions of myeloid-origin and lymphoid-origin cells in each sample of a dataset were calculated by the Bioconductor package CellMix [15], we used the two-sample *t*-test method to evaluate whether the proportions were significantly different between diseases and controls. A p-value < 0.05 was considered significant.

TABLE I. DATASETS ANALYZED IN THIS STUDY

Dataset	***Case:Control***[a]	***GEO accession Number***	***Platform***
Colorectal cancer	10:9	GSE11545	GPL2986
	19:11	GSE10715	GPL570
Lung cancer	8:38	GSE42830	GPL10558
	73:80	GSE20189	GPL571
Breast cancer	11:9	GSE11545	GPL2986
	54:67	GSE16443	GPL2986
Ovarian cancer	9: 9	GSE11545	GPL2986
Breast CD4+ T cells	10:4	GSE36765	GPL570
Normal leukocyte	13:20	GSE28491	GPL570
	17:20	GSE28490	GPL570

[a] The number of case and control samples.

C. Detection of Differentially Expressed Genes

The two-sample *t*-test method was used to select DE genes with an FDR (false discovery rate) [16] < 10%. In a dataset, a DE gene was considered up-regulated (down-regulated) if its relative difference of expression between the tumour groups and controls was larger (smaller) than zero.

D. Definition of Consistent Differentially Expressed Gene List

If a DE gene had the same direction of regulation in two datasets for a same disease, this gene was considered as a consistently expressed DE gene. Combining all the consistently expressed DE genes together, we obtained the consistent DE gene list for this disease.

E. Functional Enrichment Analysis

For each dataset, two interesting gene lists, the up- and down-regulated gene lists were analyzed separately [17] for finding significant functional categories using GO-function [18]. The significant GO functional categories were identified after multiple testing adjustments with an FDR < 10%. If n genes are selected as interesting genes (up- or down regulated genes) from N genes in a dataset, and k of them were annotated to a GO functional category with m genes, the probability of observing at least k genes by chance can be appropriately modelled by the hypergeometric distribution model, as follows:

$$P = 1 - \sum_{i=0}^{k-1} \binom{n}{i} p_0^{\,i} (1-p_0)^{n-i} \tag{2}$$

978-1-4799-7294-4/14/$31.00 ©2014 IEEE

where p_0 was estimated using the cumulative uniform distribution model, based on the assumption that the enrichment *P*-values follow a uniform distribution, i.e., every enrichment *P*-value has an equal probability to occur between zero and one. A binomial *P*-value < 0.05 was considered significant.

III. RESULTS

A. Shifts in Populations of Myeloid-origin and Lymphoid-origin Cells

To determine whether DE genes identified from cancer PB might be affected by cell population shifts of myeloid-origin and lymphoid-origin cells, we estimated the proportions of the myeloid-origin and lymphoid-origin cells in colorectal and lung cancer datasets using gene expression deconvolution methods. We found that the average proportions of myeloid-origin cells were significantly higher in PB samples with cancer compared to controls, while the average proportions of lymphoid-origin cells were significantly lower in lung cancer patients (Fig. 1, p-value < 0.05, t-test). This indicated that the proportions of the myeloid-origin and lymphoid-origin cells increased and decreased respectively under cancer condition.

B. Comparison of DE Genes in Cancer PB to DE Genes in Myeloid-origin Cells Compared to Lymphoid-origin Cells

We evaluated whether the differential mRNAs observed in tumor bloods could be explained by the population shift of blood cells. First, we defined consistent DE gene list for cancer compared to controls and for myeloid-origin compared to lymphoid-origin cells respectively. For breast cancer, with an FDR < 10%, only seven DE genes were identified from one (GEO acc.no.GSE16443) of the two datasets, resulting in no overlapping DE genes between the two breast cancer datasets. Therefore, this dataset for breast cancer was excluded from the following analysis and we integrated the consistent DE genes from the two datasets for colorectal cancer, lung cancer and leukocyte cells respectively. Totally, the consistent DE gene list for colorectal and lung cancer included 309 and 207 genes respectively; whereas the consistent DE gene list for myeloid-origin cells compared to lymphoid-origin cells included 4587 genes. If a cancer DE has the same direction of regulation as in myeloid-origin compared to lymphoid-origin cells, it was considered to be explained by the shift populations of myeloid-origin and lymphoid-origin cells [19]. Among the 309 consistent DE genes identified for CRC, 157 were also differentially expressed between myeloid-origin and lymphoid-origin cells, and 89.2% of them were deregulated with the same directions. In lung cancer, 157 of the 207 consistent DE genes were differentially expressed between myeloid-origin and lymphoid-origin cells, with the consistent score as high as 100%. The result demonstrated that the changes in populations of myeloid-origin and lymphoid-origin blood cells could contribute to a significant proportion of the observed differential signals in the cancer blood transcriptome.

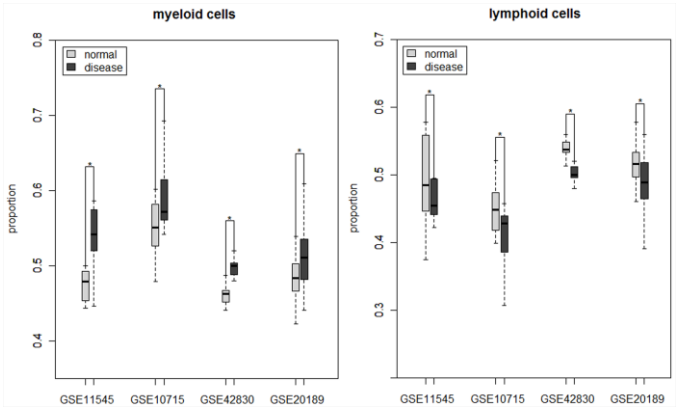

Fig. 1. Average proportions of myeloid-origin and lymphoid-origin cells in multiple cancer datasets.

C. Functional Consistency of Differential Genes in Cancer Peripheral Blood

As PB is the pipeline of the immune system carrying different types of leukocytes [20], it's possible that the DE genes in cancer PB are enriched in biological functional categories shifted by changes in the leukocytes. Thus, we used GO-function to find the significant GO functional categories that consistently enriched by up- or down-regulated PB DE genes identified for cancer. A GO functional category was considered consistent across the six cancer blood datasets when it was detected as significant in at least two datasets (*P*-value = 3.28×10^{-2}, binomial test). As shown in Table II, with an FDR < 10%, we obtained 17 GO functional categories consistently enriched for up-regulated genes identified from colorectal, lung, breast and ovarian cancer respectively, including functional categories related to immune- and inflammation-related processes (Table II), which indicated that the immmune system responds under cancer condition. The down-regulated genes were enriched in 55 functional categories related to the metabolic, cellular component and gene expression/translation processes (Table II).

Then, the 17 and 55 significant functional categories were also analyzed using the two leukocyte datasets. For the functional categories consistently enriched by up-regulated genes, all of the 17 functional categories were non-randomly enriched by the DE genes between the myeloid-origin and lymphoid-origin cell identified from the two leukocyte datasets with an FDR < 10%, which could not be expected to happen by random chance (*P*-value = 2.5×10^{-3}, binomial test), hinting that gene expression patterns in cancer PB samples may reflect the corresponding changes in shift of myeloid-origin and lymphoid-origin cells. All the 55 functional categories consistently enriched for down-regulated genes across different cancer datasets were non-randomly enriched by the DE genes in the two leukocyte expression datasets with an FDR < 10% (*P*-value = 2.5×10^{-3}, binomial test).

Notably, we found that, with an FDR < 10%, only one of the 17 functional categories enriched for different cancer PB DE genes, was enriched for the DE genes identified in CD4+ T cells of cancer patients compared to normal CD4+ T cells. Similarly, only one of the 55 functional categories enriched for

978-1-4799-7294-4/14/$31.00 ©2014 IEEE

TABLE II. SIGNIFICANT GO TERMS CONSISTENTLY ENRICHED IN VARIOUS CANCER DATASETS

Gene list	GO term name
up-regulated genes	**Immunization module** immune system process; response to stress; defense response; regulation of body fluid levels; cell activation; hemostasis; coagulation; blood coagulation **Membrane organization module** membrane organization; cellular membrane organization **Localization module** Transport; vesicle-mediated transport; cellular localization; establishment of localization in cell **Catabolic process module** catabolic process; cellular catabolic process; carbohydrate catabolic process
down-regulated genes	**Cellular component module** cellular protein complex disassembly; protein complex disassembly; macromolecular complex disassembly; cellular macromolecular complex disassembly; cellular component disassembly; cellular catabolic process; macromolecular complex subunit organization; ribonucleoprotein complex biogenesis; cellular component biogenesis at cellular level; cellular component organization or biogenesis at cellular level; cellular component disassembly at cellular level; cellular macromolecular complex subunit organization **Gene expression/translation module** nuclear mRNA splicing, via spliceosome; RNA splicing, via transesterification reactions with bulged adenosine as nucleophile; RNA splicing, via transesterification reactions; mRNA processing; RNA splicing; RNA processing; mRNA metabolic process; rRNA processing; rRNA metabolic process; ncRNA processing; ncRNA metabolic process; translational elongation; translational termination; translation; ribosome biogenesis; gene expression **Cellular metabolic process module** cellular macromolecule biosynthetic process; cellular protein metabolic process; nucleic acid metabolic process; cellular macromolecule metabolic process; protein metabolic process; nucleobase, nucleoside, nucleotide and nucleic acid metabolic process; macromolecule metabolic process; cellular nitrogen compound metabolic process; nitrogen compound metabolic process; primary metabolic process; cellular metabolic process; metabolic process **Viral reproduction module** viral transcription; viral genome expression; viral infectious cycle; cellular process involved in reproduction; viral reproductive process; viral reproduction **Single-multicellular organism process module** endocrine pancreas development; pancreas development; endocrine system development; sensory perception of chemical stimulus; sensory perception of smell; sensory perception; neurological system process; system process; multicellular organismal process

down-regulated cancer PB genes was identified as significant by DE genes for CD4+ T cells. As these 72 functional categories were all enriched for the DE genes between myeloid-origin and lymphoid-origin cells and rarely enriched for the DE genes between cancer and normal CD4+ T cells, we suggested the possibility that the functional categories identified in cancer PB gene expression profiles were not reflective of the blood cell-intrinsic expression changes but the the changes in proportions of myeloid-origin and lymphoid-origin cells. Therefore, when interpreting the functional categories enriched for PB cancer DE genes, simply interpreting the biological processes enriched by DE genes identified from cancer PB to have connection to cancer directly might be improper.

A schematic overview of the research designs is represented in Fig.2.

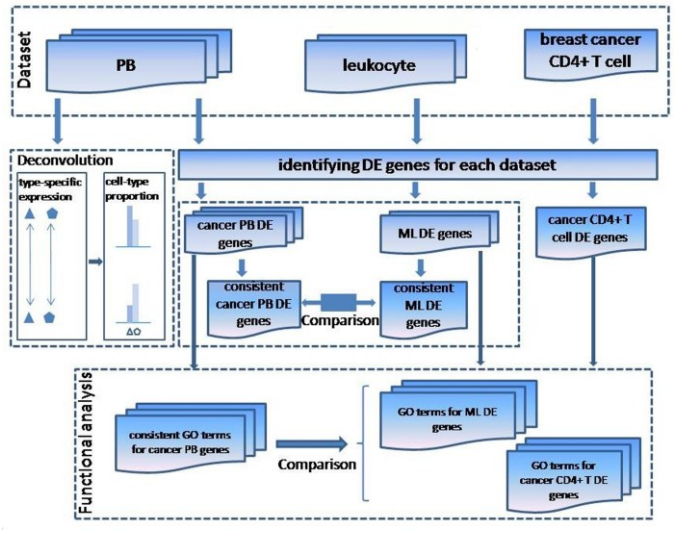

Fig. 2. Schematic overview of the research designs.

IV. DISCUSSION

The functional enrichment analysis is often applied after identifying differentially expressed genes for cancer in human peripheral blood, for which the expression signals are mainly derived from leukocytes. However, as revealed in this paper, the functional categories consistently identified by different cancer PB DE genes were all disturbed by the DE genes between myeloid-origin and lymphoid-origin cells rather than the DE genes identified for the CD4+ T cells, indicating that these functional categories observed in PB were more likely to reflect the changes in blood cell populations. Therefore, the interpretation of enriched functional categories should not focus on leukocyte-specific gene expression alterations as significant functional categories observed in PB are probably disturbed by changes in the cell proportion.

We also showed that the consistent DE genes detected from colorectal and lung cancer PB datasets had the same directions of regulation as in myeloid-origin compared to lymphoid-origin cells, indicating the expression patterns of genes in cancer PB were more likely to reflect the shifts in cell populations of myeloid-origin and lymphoid-origin cells. Notably, similar population shifts in blood cells have also been observed in many inflammation-associated disease [21]. As has been discussed by many researchers, age, body mass, sex, smoking, drinking and inflammatory status could influence gene expression in blood [22]. Promising translational biomarkers developed from blood samples should be stable to the biological variations. Therefore, the applicable blood biomarkers for cancer should be able to distinguish cancer from inflammation related disease samples especially those inflammation diseases occurred in the same organ, and stable to biological variations, which deserve further investigation.

ACKNOWLEDGMENT

978-1-4799-7294-4/14/$31.00 ©2014 IEEE

This work was financially supported by the National Natural Science Foundation of China (grant numbers 91029717, 81071646, 81372213, 81201702 and 81201822).

REFERENCES

[1] Lunnon K., Z. Ibrahim, P. Proitsi, A. Lourdusamy, S. Newhouse, M. Sattlecker, "Mitochondrial dysfunction and immune activation are detectable in early alzheimer's disease blood," J Alzheimers Dis, 2012. 30(3): p. 685-710.

[2] Sharp F.R., G.C. Jickling, B. Stamova, Y. Tian, X. Zhan, D. Liu, "Molecular markers and mechanisms of stroke: Rna studies of blood in animals and humans," J Cereb Blood Flow Metab, 2011. 31(7): p. 1513-31.

[3] Bhattacharya S., S. Tyagi, S. Srisuma, D.L. Demeo, S.D. Shapiro, R. Bueno, "Peripheral blood gene expression profiles in copd subjects," J Clin Bioinforma, 2011. 1(1): p. 12.

[4] Zander T., A. Hofmann, A. Staratschek-Jox, S. Classen, S. Debey-Pascher, D. Maisel, "Blood-based gene expression signatures in non-small cell lung cancer," Clin Cancer Res, 2011. 17(10): p. 3360-7.

[5] Sharma P., N.S. Sahni, R. Tibshirani, P. Skaane, P. Urdal, H. Berghagen, "Early detection of breast cancer based on gene-expression patterns in peripheral blood cells," Breast Cancer Res, 2005. 7(5): p. R634-44.

[6] Isaksson H.S., B. Sorbe, and T.K. Nilsson, "Whole blood rna expression profiles in ovarian cancer patients with or without residual tumors after primary cytoreductive surgery," Oncol Rep, 2012. 27(5): p. 1331-5.

[7] Aaroe J., T. Lindahl, V. Dumeaux, S. Saebo, D. Tobin, N. Hagen, "Gene expression profiling of peripheral blood cells for early detection of breast cancer," Breast Cancer Res, 2010. 12(1): p. R7.

[8] Liew C.C., J. Ma, H.C. Tang, R. Zheng, and A.A. Dempsey, "The peripheral blood transcriptome dynamically reflects system wide biology: A potential diagnostic tool," J Lab Clin Med, 2006. 147(3): p. 126-32.

[9] Cho H., H.W. Hur, S.W. Kim, S.H. Kim, J.H. Kim, Y.T. Kim, "Pre-treatment neutrophil to lymphocyte ratio is elevated in epithelial ovarian cancer and predicts survival after treatment," Cancer Immunol Immunother, 2009. 58(1): p. 15-23.

[10] Go S.I., A. Lee, U.S. Lee, H.J. Choi, M.H. Kang, J.H. Kang, "Clinical significance of the neutrophil-lymphocyte ratio in venous thromboembolism patients with lung cancer," Lung Cancer, 2014. 84(1): p. 79-85.

[11] Hong W.S., S.I. Hong, C.M. Kim, Y.K. Kang, J.K. Song, M.S. Lee, "Differential depression of lymphocyte subsets according to stage in stomach cancer," Jpn J Clin Oncol, 1991. 21(2): p. 87-93.

[12] Barrett T., D.B. Troup, S.E. Wilhite, P. Ledoux, D. Rudnev, C. Evangelista, "Ncbi geo: Archive for high-throughput functional genomic data," Nucleic Acids Res, 2009. 37(Database issue): p. D885-90.

[13] Abbas A.R., K. Wolslegel, D. Seshasayee, Z. Modrusan, and H.F. Clark, "Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus," PLoS One, 2009. 4(7): p. e6098.

[14] Abbas A.R., D. Baldwin, Y. Ma, W. Ouyang, A. Gurney, F. Martin, "Immune response in silico (iris): Immune-specific genes identified from a compendium of microarray expression data," Genes Immun, 2005. 6(4): p. 319-31.

[15] Gaujoux R. and C. Seoighe, "Cellmix: A comprehensive toolbox for gene expression deconvolution," Bioinformatics, 2013. 29(17): p. 2211-2.

[16] Benjamini Y. and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," J.Roy. Stat. Soc. B., 1995. 57: p. 289-300.

[17] Hong G., W. Zhang, H. Li, X. Shen, and Z. Guo, "Separate enrichment analysis of pathways for up- and downregulated genes," J R Soc Interface, 2014. 11(92): p. 20130950.

[18] Wang J., X. Zhou, J. Zhu, Y. Gu, W. Zhao, J. Zou, "Go-function: Deriving biologically relevant functions from statistically significant functions," Brief Bioinform, 2012. 13(2): p. 216-27.

[19] Hong G, Chen B, Li H, Zhang W, Zheng T, Li S, "Differential mrnas in peripheral blood of lung cancer and inflammation-associated pulmonary diseases are defined by differential mrnas between myeloid and lymphoid cells," PLoS One, in press.

[20] Chaussabel D., V. Pascual, and J. Banchereau, "Assessing the human immune system through blood transcriptomics," BMC Biol, 2010. 8: p. 84.

[21] Guglielmetti L., A. Cazzadori, M. Conti, F. Boccafoglio, A. Vella, R. Ortolani, "Lymphocyte subpopulations in active tuberculosis: Association with disease severity and the qft-git assay," Int J Tuberc Lung Dis, 2013. 17(6): p. 825-8.

[22] Menke A., M. Rex-Haffner, T. Klengel, E.B. Binder, and D. Mehta, "Peripheral blood gene expression: It all boils down to the rna collection tubes," BMC Res Notes, 2012. 5: p. 1.

978-1-4799-7294-4/14/$31.00 ©2014 IEEE