Focused human gene expression profiling using dual-color reverse transcriptase multiplex ligation-dependent probe amplification (dcRT-MLPA).

Mariëlle C. Haks¹, Jelle J. Goeman², Cecile Magis-Escurra^{1,3} and Tom H.M. Ottenhoff¹

¹*Department of Infectious Diseases, Leiden University Medical Center, Leiden, The Netherlands;* ²*Biostatistics, Radboud Institute for Molecular Life Science, Radboud University, Nijmegen, The Netherlands;* ³*Present address: Department of Pulmonary Diseases, Radboud University Medical Centre Nijmegen, University Centre of Chronic Diseases Dekkerswald, Groesbeek, The Netherlands.*

Corresponding author: Prof. Dr. Tom H. M. Ottenhoff
Department of Infectious Diseases,
Leiden University Medical Center
Group of Immunology and Immunogenetics of Mycobacterial Infectious Diseases
Albinusdreef 2
2333 ZA Leiden
The Netherlands.
Tel: +31-71-526 3809
E-mail: t.h.m.ottenhoff@lumc.nl

Abstract

To investigate the human immune response to newly developed or existing vaccines, or during infection/disease on a population scale, we have recently developed a dual-color Reverse Transcriptase Multiplex Ligation-dependent Probe Amplification (dcRT-MLPA) assay, which can rapidly profile mRNA expression of multiple host genes. dcRT-MLPA has a dynamic range and sensitivity comparable to real-time QPCR and RNA-Sequencing. Since this assay is high-throughput, it is an exceptionally suitable technique for monitoring host biomarkers in semi-large scale human cohorts, such as cross sectional studies with multiple groups, or longitudinal studies with multiple time points. Multicomponent host biomarker signatures with excellent predictive values can easily be identified using lasso regression analysis, while exploring additional data adjustment methods like RUV-2 may further optimize the identification of informative host biomarker signatures. dcRT-MLPA also allows comparisons of gene expression patterns across different human populations to explore the impact of geographical diversity on for example vaccine induced responses. The use of dcRT-MLPA is not limited to peripheral blood but can be adapted to analyze host biomarkers derived from any tissue or body fluids, further demonstrating the versatility of the dcRT-MLPA platform. Several examples will be given and discussed.

Keywords: dcRT-MLPA; gene expression profiling; host biomarkers.

Identification of host biomarkers by different gene expression profiling platforms.

Biomarkers are defined as ‘characteristics that are objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes or pharmacological responses to a therapeutic intervention’ [1]. Host biomarkers can be powerful ‘surrogate endpoints’ to classify disease status, disease activity, disease progression, as well as prognosis and the effects of interventions (e.g. drugs, vaccines). Host biomarkers can be determined at the cellular-, protein-, metabolic, or transcript level, and can be analyzed in any tissue or body fluid, but peripheral blood is the most commonly used source in clinical practice.

Measurements of quantitative changes in RNA expression levels, to identify host biomarkers at the transcriptomic level, are currently primarily being analyzed using assays at the extreme ends of the RNA expression profiling spectrum: either genome-wide (microarray or RNA-Sequencing) or single gene (real-time QPCR) screening methods. Since genome-wide approaches are by definition hypothesis-free, they are the methods of choice to identify novel host biomarkers and biomarker signatures in small sample sets. However, because host biomarker signatures useful in classifying disease status, disease activity, disease progression, prognosis and the effects of interventions, will most likely encompass restricted sets of multiple gene transcripts, neither genome-wide nor single gene approaches will be ideally suited to achieve this. Furthermore, microarray and RNA-Sequencing assays are technically challenging, require substantial amounts of RNA, are costly and are complicated by complex data analysis, further limiting their applicability as a tool to routinely screen for host biomarkers in large human cohorts (Figure 1).

To monitor and validate multicomponent host biomarker signatures on a population scale, techniques have been developed that combine sets of markers at the transcriptomic level. Fluidigm, for example, provides high-throughput real-time QPCR based on proprietary microfluidic chips [2]. These gene expression assays allow testing of either 48 genes in a 48-well format or 96 genes in a 96-well format. NanoString's nCounter gene expression assay is a novel technology using color-coded molecular barcodes and microscopic imaging to detect and count up to 800 unique transcripts in a single reaction. Although it does not require the conversion of mRNA to cDNA or amplification of the resulting cDNA, its sensitivity rivals PCR-based methods [3]. Both Fluidigm and nCounter are commercially available but are still costly and require specialized equipment not routinely available in many laboratories, especially not in resource-poor settings. Alternatively, dual-color reverse-transcriptase multiplex ligation-dependent probe amplification (dcRT-MLPA) is a robust technique that can rapidly and accurately profile RNA expression of as many as ~90-100 host genes in a single reaction [4]. As the assay is reliable, inexpensive, user friendly, high-throughput (96-well format), and only requires 100 ng RNA per sample, it is exceptionally suited to determine host biomarker signatures in larger sample sets, such as in population and/or longitudinal studies. Furthermore, genes of interest can be selected on a tailor-made basis and are interchangeable, adding flexibility to the selection of host biomarkers being analyzed (Figure 1).

Principle of the dcRT-MLPA technique.

dcRT-MLPA is based on the principle that for each target specific sequence, a specific RT primer is designed that is located immediately downstream of the probe target sequence. Following

reverse transcription, two half-probes are hybridized directly adjacent to a target sequence and ligated together. DcRT-MLPA is designed such that all ligated products are amplified with the use of only two fluorophore-conjugated PCR primer pairs while the length of each amplification product is unique. The PCR amplification step ensures assay sensitivity, which is an essential prerequisite for the relative quantification of scarcely expressed genes. Amplified products are subsequently size-separated using capillary electrophoresis followed by analysis of trace data using the GeneMapper software package and exporting category tables quantifying the fluorescence-intensity of each amplification product (peak area) for further analysis (described in more detail below) (Figure 2).

Comparing data generated with dcRT-MLPA to microarray, RNA-Sequencing, and real-time QPCR.

As both dcRT-MLPA and real-time QPCR are based on amplification of target products by PCR, dcRT-MLPA and real-time QPCR have very similar dynamic range (4-7 log) and sensitivity [4]. To directly compare gene expression profiles using dcRT-MLPA and real-time QPCR, the same RNA samples were profiled using both techniques. Data illustrated in Figure 3A clearly show that the results obtained with these two RNA expression-profiling techniques are highly similar. Both weak and strong variations in gene expression levels detected by QPCR are faithfully replicated by dcRT-MLPA, demonstrating the comparability, accuracy and robustness of dcRT-MLPA to quantify specific changes in RNA expression levels. As expected, also dcRT-MLPA and RNA-Sequencing were comparable in detecting changes in gene expression levels (Figure 3B), indicating that dcRT-MLPA, real-time QPCR and RNA-Sequencing all have very similar assay

features including sensitivity and dynamic range. Clearly, the rather limited dynamic range (2-3 logs) of the microarray platform [5] compromises its ability to accurately quantify RNA expression levels (Figure 3C).

dcRT-MLPA data normalization and downstream data analysis.

Differential gene expression data need to be corrected statistically for unwanted variation. In addition to the biological variation of interest, other biological and technical factors may affect gene expression levels. For example, a well-recognized and infamous factor contributing to unwanted variation is the batch effect. Therefore, pre-processing of gene expression data, including data normalization, has proven essential for the accurate detection of differentially expressed genes. Downstream analysis pipelines for microarray and real-time QPCR data are relatively well-established. Microarray data routinely follows 3 stages of pre-processing: background correction, normalization using for example global adjustment methods like quantile normalization, and summarization; while real-time QPCR data is usually normalized against a single endogenous unregulated control gene transcript. In contrast, optimal data pre-processing methods for more recently developed techniques like RNA-sequencing and dcRT-MLPA have not been fully investigated or resolved yet. Several efforts to systemically compare normalization approaches for RNA-sequencing data have been conducted [6-9] but for dcRT-MLPA, a technique situated in the middle of the RNA expression profiling spectrum, accurate downstream analysis pipelines are much less explored.

Currently, dcRT-MLPA trace data are analyzed using the GeneMapper software package (Applied Biosystems) to quantify the fluorescence-intensity of each transcript. Signals below the

threshold value for noise cutoff in GeneMapper are assigned the threshold value for noise cutoff. Subsequently, results from target genes are calculated relative to the geometric average signal of selected control genes present within the gene set tested (e.g. housekeeping genes ABR, GUSB, GAPDH or B2M) and the percentage s.d. is calculated to determine which control gene is most stably expressed across the evaluated samples. Following normalization of the data against the most stably expressed single control gene, signals below the threshold value for noise cutoff are again assigned the threshold value for noise cutoff. Downstream statistical analyses are subsequently performed on log-transformed data in order to stabilize the variance [4].

Since (i) expression of control (housekeeping) genes can be stable in particular cell types or experimental conditions but vary considerably in others [10-12], and (ii) dcRT-MLPA data are of a much higher dimension and complexity than real-time QPCR data, data normalization against a single control gene might not always ensure accurate inference of expression levels. Thus, alternative adjustment approaches should be further explored for future use. One relatively simple alternative could encompass geometric averaging of multiple internal control genes [13]. Another adjustment approach that might be worth exploring is 'Remove Unwanted Variation, 2-step' (RUV-2). In RUV-2, factor analysis is performed on just the negative control genes and incorporates the resulting factor into a linear regression model. The method is based on the observation that negative control genes (genes whose expression levels are known to be truly unassociated with the biological variation of interest) can not only be used to detect the presence of unwanted variation, but can also be exploited to adjust for unwanted variation [14]. RUV-2 has already been demonstrated to outperform adjustment techniques such as

Combat and ordinary regression when analyzing microarray data [14] and more recently was also shown to outperform state-of-the-art normalization methods currently used for RNA-Sequencing data [8].

Key to the performance of many normalization techniques, including RUV-2, is the selection of negative control genes. Effective negative control genes must be uninfluenced by the biological variation of interest and they must be influenced by the unwanted factors. This not only implies that different experimental situations will require different sets of control genes, but also that the most stably expressed control gene in a given data set may not always be the control gene of choice because it fails to capture the unwanted factors. Clearly, the selection of appropriate negative control genes is not straightforward but housekeeping genes will likely outperform spike-in controls because housekeeping genes may be able to capture unwanted biological variation, whereas spike-in controls can only capture unwanted technical variation [14]. A selection of housekeeping genes that are likely suitable candidates as negative control genes can be found in Lippa *et al* and Eisenberger *et al* [15;16].

Identification of host biomarkers by dcRT-MLPA using different sources of immune cells.

A particularly useful application of dcRT-MLPA is the identification and monitoring of host biomarker signatures to investigate the human immune response on a population scale. We have used dcRT-MLPA to characterize the human innate and adaptive immune response in peripheral blood during infection with mycobacteria (both *Mycobacterium leprae* and *Mycobacterium tuberculosis*, the causative agents of leprosy and tuberculosis (TB), respectively) as well as in the response to curative TB treatment in genetically and geographically diverse

populations [4;17-20]. Host biomarker signatures with the highest discriminatory power between active TB disease (TB cases) and latent infection were identified and validated using a lasso logistic regression model, which is a shrinkage and selection method for logistic regression [21]. Host biomarker signatures had excellent classifying values and genes that were differentially expressed between active TB disease and latent infection normalized during TB treatment [4]. Host biomarkers within the signatures included the high-affinity IgG Fc receptor 1 FCGR1A (CD64), apoptosis-related genes, and T-cell/B-cell markers, suggesting important contributions of adaptive immunity to TB pathogenesis [4]. Moreover, FCGR1A expression levels appeared not affected by HIV status or geo-genetic differences, categorizing it as a remarkably consistent classifier of active TB disease [19].

More recent studies have focused on the identification of prognostic host biomarkers measurable in peripheral blood that can identify people at risk of developing active TB disease. Probing immunological differences between individuals progressing or non-progressing to active TB disease at early time points after contact with a TB case, identified several markers that may predict the onset of active TB at a very early stage after infection [18]. Moreover, in a retrospective case-control study of HIV-infected drug users, lasso regression analysis identified a biomarker signature encompassing IL13 and AIRE with discriminatory power to identify individuals progressing to TB within a high risk population up to 8 months prior to clinical diagnosis. Individuals progressing to active TB disease and expressing IL13 had a gene expression pattern strongly enriched for type I IFN related signaling genes, suggesting that these genes represent processes that contribute to TB pathogenesis [22].

199 Although peripheral blood is the most commonly used cell source for identifying host
200 biomarker transcripts in clinical practice, dcRT-MLPA can be adapted to analyze host
201 biomarkers in purified cell populations or in any tissue or body fluid provided that sufficient
202 RNA can be recovered from the tissue sample. When monitoring host biomarker signatures to
203 investigate the human immune response to infection on a population scale, peripheral blood is
204 often the only source available, but in many cases, including tuberculosis, this is not the primary
205 site of infection. Probing differences in host biomarker signatures between “systemic”,
206 peripheral blood versus “local”, e.g. tissue or fluid from the site of infection probably will
207 significantly increase our understanding of the elicited innate and adaptive immune response
208 against specific pathogens in relation to organ pathology. For example, in tuberculosis the
209 primary site of infection is the lung. When comparing gene expression profiles of peripheral
210 blood versus bronchoalveolar lavage (BAL; site of infection) samples, differences found in
211 peripheral blood between TB cases and latently infected health care workers were not identical
212 in BAL (Figure 4) [4]:the majority of genes exhibiting discriminatory power in peripheral blood
213 between active TB disease and latent infection (B-cell associated markers BLR1, CD19, and
214 FCGR1A; T-cell associated marker IL-7R) were expressed at comparable levels in BAL between
215 study groups (active TB cases infected part of the lung versus health care workers) (Figure 4A)
216 [4], suggesting that the significantly reduced expression of B-cell markers during active disease
217 in the periphery does not simply reflect migration of B-cells from peripheral blood to the site of
218 infection. Conversely, most of the genes that were differentially expressed in BAL samples from
219 active TB disease versus latent infection showed similar expression levels in peripheral blood
220 (Figure 4B), again indicating that the gene expression profiles capture a complex immunological

response against TB that is not merely the result of cell migration from peripheral blood to the site of infection. Of further interest, the gene expression patterns in BAL samples from the non-affected lung portions from TB patients were more similar to those of health care workers than the ones from the affected lung parts (Figure 4), again suggesting the existence of site specific and disease activity related biomarker signatures, whose further dissection could be useful in understanding disease pathogenesis.

Finally, monitoring of host biomarkers by dcRT-MLPA could also be accomplished successfully in samples from skeletal muscle biopsies obtained from volunteers who had followed a 5-day high-fat high calorie (HFHC) diet. In these samples we found significant increases in the expression levels of pro-inflammatory (M1) macrophage markers [23], accompanied by a reduction in the expression of markers of insulin signaling and by the development of insulin resistance. These data show that recruitment of macrophages into muscle may be an early event in development of insulin resistance in response to short-term HFHC-feeding. They also demonstrate the versatility of the dcRT-MLPA platform in probing innate and adaptive immune responses in diverse experimental setups, including in human observational as well as experimental medicine approaches.

Conclusion

dcRT-MLPA allows analysis of host biomarkers in relation to vaccination, infection, disease, prognosis and therapeutic interventions. dcRT-MLPA can also be used to probe differences in immune reactivity in populations in (and between) different areas, e.g. due to environmental or genetic factors which may impact on for example vaccine efficacy. Site specific biomarker

243 signatures were detectable in a study directed towards elucidating local versus systemic
244 signatures in TB. The dcRT-MLPA platform has comparable sensitivity and dynamic range to
245 real-time QPCR and RNA-Sequencing and the assay can be adapted to monitor host biomarker
246 transcripts in purified cell populations and in any tissue or body fluid. Multicomponent host
247 biomarker signatures can be identified with excellent classifying values using lasso regression
248 analysis but exploring additional data adjustment methods like RUV-2 may further optimize the
249 identification and monitoring of host biomarkers in future research.

250

251 **Materials and methods**

252 *Ethics statement*

253 The research was approved by the Ministry of Public Health and Social Welfare (Paraguay).
254 Informed consent was obtained from all participants, and the clinical investigation was
255 conducted according to the principles expressed in the 'Declaration of Helsinki'.

256

257 **Donors**

258 Whole-blood (PAXgene tubes; PreAnalytix GmbH, Switzerland) and bronchoalveolar lavage
259 (BAL) samples were collected from a Paraguayan cohort (10 TB patients at recruitment, and 11
260 latently infected healthcare workers). TB patients were sputum smear positive using the Ziehl-
261 Nelsen (ZN) stain and had not initiated treatment yet at the time of sample collection. Latently
262 infected health care workers had a positive tuberculin skin test (TST+), had never been treated
263 for TB in the past, and did not show evidence of active TB as determined by chest X-ray. Both
264 TB patients and latently infected healthcare workers were HIV- and were older than 18 years of
265 age.

266

267 *PAXgene whole-blood RNA isolation*

268 Total RNA from venipuncture PAXgene blood collection tubes was extracted and purified using
269 the PAXgene Blood RNA kit (PreAnalytix) including on-column DNase digestion according to the
270 manufacturer's protocol. The RNA yield was determined by a NanoDrop ND-1000
271 spectrophotometer (NanoDrop Technologies, Wilmington, DE, USA). To assess the quality and

integrity of the RNA, samples were run on an Agilent 2100 BioAnalyzer (Agilent Technologies, Amstelveen, the Netherlands) using the RNA 6000 Nano Chip kit.

Microarray

Microarray analysis was performed as described previously [24;25]. Briefly, RNA was either Cy3- or Cy5-labeled using the Fluorescent Linear Amplification Kit (Agilent Technologies) according to the instructions of the manufacturer. Quantity and labeling efficiency were verified before hybridization of the samples to whole-genome 4x44 k human expression arrays (Agilent Technologies) and scanned at 5 mm using an Agilent scanner. Analysis of the scanned images was performed with Feature Extraction software (version 6.1.1, Agilent Technologies). Background-corrected values [26] were lowess normalized. Data was subsequently quantile normalized [27] and log2-transformed.

RNA-Sequencing

Libraries were constructed according to Illumina protocol (www.illumina.com), which includes an oligo-dT capture of poly-adenylated transcripts from purified total RNA. RNA-seq reads were mapped to a reference human genome and the expression of known genes (i.e., expected fragments per kilo-base of transcript per million fragments mapped [FPKM]) was quantified with Cufflinks [28] on the basis of Tophat2 [29] mappings.

dcRT-MLPA assay

A dual-color reverse transcriptase multiplex ligation-dependent probe amplification (dcRT-MLPA) assay was performed as described previously [4]. Briefly, for each target-specific sequence, a specific RT primer was designed located immediately downstream of the left and right hand half-probe target sequence. Following reverse transcription of 100 ng RNA using MMLV reverse transcriptase (Promega, Leiden, the Netherlands), left and right hand half-probes were hybridized to the cDNA at 60°C overnight. Annealed half-probes were ligated and subsequently amplified by PCR (33 cycles of 30 s at 95 °C, 30 s at 58 °C and 60 s at 72 °C, followed by 1 cycle of 20 min at 72 °C). Primers and probes were from Sigma-Aldrich Chemie (Zwijndrecht, the Netherlands) and MLPA reagents from MRC Holland (Amsterdam, the Netherlands). PCR amplification products were 1:10 diluted in HiDi formamide containing 400HD ROX size standard and analysed on an Applied Biosystems 3730 capillary sequencer in GeneScan mode (Baseclear, Leiden, the Netherlands). Trace data were analysed using the GeneMapper software package (Applied Biosystems/Life Technologies, Bleiswijk, the Netherlands). The areas of each assigned peak (in arbitrary units) were exported for further analysis in Microsoft Excel spreadsheet software. Signals below the threshold value for noise cutoff in GeneMapper ($\log_2$ -transformed peak area ≤ 7.64) were assigned the threshold value for noise cutoff. Following normalization of the data to the average signal of housekeeping gene GAPDH, signals below the threshold value for noise cutoff in GeneMapper were again assigned the threshold value for noise cutoff.

Taqman real-time PCR

314 First-strand cDNA was synthesized using oligo(dT) primers and SuperScript III First-Strand
315 Synthesis System (Life Technologies) according to the manufacturer's protocol. Quantitative
316 real-time PCR was performed on a 7900HT Fast Real- Time PCR System (Life Technologies) in a
317 total volume of 25 μ l containing 900 nM forward and reverse primer, 250 nM FAM dye-labeled
318 Taqman MGB probe and 10 ng of cDNA in a 1x TaqMan Universal PCR Master Mix, No
319 AmpErase UNG, including a passive reference (ROX) fluorochrome. Optimized thermal cycling
320 conditions used were as follows: 1 cycle of 10 min at 95 °C, followed by 50 cycles of 15 s at 95
321 °C and 1 min at 60 °C. Relative gene expression was calculated using the Δ Ct method and
322 normalized to the expression of GAPDH. All inventoried TaqMan Gene Expression Assays were
323 purchased from Life Technologies and were performed in triplicate.

324

Acknowledgements

We thank Dr. E.M.S. Leyten, Dr. M. van Westreenen and employees and patients of Instituto de Enfermedades Respiratorias y Ambientales (INERAM), Asuncion, Paraguay, for their contribution to the Paraguay cohort design and sample collection. We gratefully acknowledge all the funding that made the work reviewed here possible. We especially acknowledge the Bill and Melinda Gates Foundation (Grand Challenges in Global Health GC6#74), 6th framework Programme TBVAC contract no. LSHP-CT-2003-503367, and 7th framework Programmes NEWTBVAC contract no. HEALTH-F3-2009-241745, IDEA contract no. 241642, ADITEC contract no. 280873, TANDEM contract no. 305279 (the text represents the authors' views and does not necessarily represent a position of the Commission that will not be liable for the use made of such information). The funders had no role in study design, data collection and analysis, decision to publish or in preparation of the manuscript.

Figure legends

Figure 1. Spectrum of methods to determine gene expression profiles.

Features and characteristics of different methods used to quantify the abundance of RNA transcripts. (*Assay pricing is complicated and highly variable and depends on the number of assays ordered, the use of custom versus pre-designed gene panels, and the number of genes within a panel).

Figure 2. Outline of the dual-color RT-MLPA assay.

Dual-color RT-MLPA is based upon the ligation of two half-probes when hybridized adjacently to a cDNA target sequence. Ligation is enabled by 5' phosphorylation of the right hand half-probe. Fragments are amplified using only two fluorophore-conjugated PCR primer pairs recognizing the flanking regions of the half-probes, while the discriminative length of each amplicon is assured by varying the length of the target-specific sequence in each probe set. The length of the transcripts increases in a stepwise fashion by only 2 nucleotides, with a total size range between ~80-220 nucleotides. Trace data are analyzed using the GeneMapper software package and category tables quantifying the fluorescence-intensity of each amplification product (peak area) are exported for further analysis.

Figure 3. Comparison of dcRT-MLPA to real-time QPCR, microarray, and RNA-sequencing.

Comparison between dcRT-MLPA and (A) real-time QPCR, (B) RNA-Sequencing, and (C) microarray. Direct ex vivo RNA expression profiling was performed on peripheral blood from TB cases and latently infected individuals (A and C) or on peripheral blood from individuals before

and after challenge with malaria-infected mosquito bites. Expression profiles were determined of those genes present within the probe set described in Joosten *et al*, 2012 [4] and fold induction was calculated relative to the latently infected (A and C) or before challenge (B) samples.

Figure 4. Identification of host biomarkers in BAL and whole blood associated with TB disease in a small Paraguayan cohort.

dcRT-MLPA was performed on direct ex vivo RNA isolated from whole blood (PAXgene tubes) and bronchoalveolar lavage (BAL) samples derived from the same 10 TB cases and 11 latently infected health care workers. From each TB case, 2 BAL samples were collected, 1 from the uninfected part of the lungs and 1 from an infected part of the lungs as determined by chest X-ray. Median gene expression levels (peak areas normalized to GAPDH and log2 transformed) of the indicated genes are shown as box-and whisker plots (5-95 percentiles; dotted line represents assay cutoff). Significant differences between study groups were determined using Kruskal-Wallis H and Dunn's multiple comparison tests. *0.01<P<0.05, **0.001<P<0.01 and ***P<0.001.

377

378

Reference List

379

- 380 [1] Atkinson A.J., Colburn W.A., DeGruttola V.G., et al. Biomarkers and surrogate endpoints: Preferred
381 definitions and conceptual framework. *Clinical Pharmacology & Therapeutics* 2001;69:89-95.
- 382 [2] Rahimov F, King OD, Leung DG, et al. Transcriptional profiling in facioscapulohumeral muscular
383 dystrophy to identify candidate biomarkers. *Proc Natl Acad Sci U S A* 2012 Oct 2;109(40):16234-9.
- 384 [3] Geiss GK, Bumgarner RE, Birditt B, et al. Direct multiplexed measurement of gene expression with
385 color-coded probe pairs. *Nat Biotechnol* 2008 Mar;26(3):317-25.
- 386 [4] Joosten SA, Goeman JJ, Sutherland JS, et al. Identification of biomarkers for tuberculosis disease
387 using a novel dual-color RT-MLPA assay. *Genes Immun* 2012 Jan;13(1):71-82.
- 388 [5] Bowtell DD. Options available--from start to finish--for obtaining expression data by microarray.
389 *Nat Genet* 1999 Jan;21(1 Suppl):25-32.
- 390 [6] Bullard JH, Purdom E, Hansen KD, Dudoit S. Evaluation of statistical methods for normalization and
391 differential expression in mRNA-Seq experiments. *BMC Bioinformatics* 2010;11:94.
- 392 [7] Dillies MA, Rau A, Aubert J, et al. A comprehensive evaluation of normalization methods for
393 Illumina high-throughput RNA sequencing data analysis. *Brief Bioinform* 2013 Nov;14(6):671-83.
- 394 [8] Risso D, Ngai J, Speed TP, Dudoit S. Normalization of RNA-seq data using factor analysis of control
395 genes or samples. *Nat Biotechnol* 2014 Sep;32(9):896-902.
- 396 [9] Sun Z, Zhu Y. Systematic comparison of RNA-Seq normalization methods using measurement error
397 models. *Bioinformatics* 2012 Oct 15;28(20):2584-91.
- 398 [10] Suzuki T, Higgins PJ, Crawford DR. Control selection for RNA quantitation. *Biotechniques* 2000
399 Aug;29(2):332-7.
- 400 [11] Thellin O, Zorzi W, Lakaye B, et al. Housekeeping genes as internal standards: use and limits. *J*
401 *Biotechnol* 1999 Oct 8;75(2-3):291-5.
- 402 [12] Warrington JA, Nair A, Mahadevappa M, Tsyganskaya M. Comparison of human adult and fetal
403 expression and identification of 535 housekeeping/maintenance genes. *Physiol Genomics* 2000
404 Apr 27;2(3):143-7.
- 405 [13] Vandesompele J, De PK, Pattyn F, et al. Accurate normalization of real-time quantitative RT-PCR
406 data by geometric averaging of multiple internal control genes. *Genome Biol* 2002 Jun
407 18;3(7):RESEARCH0034.
- 408 [14] Gagnon-Bartsch JA, Speed TP. Using control genes to correct for unwanted variation in microarray
409 data. *Biostatistics* 2012 Jul;13(3):539-52.

410 [15] Eisenberg E, Levanon EY. Human housekeeping genes are compact. *Trends Genet* 2003
411 Jul;19(7):362-5.

412 [16] Lippa KA, Duewer DL, Salit ML, Game L, Causton HC. Exploring the use of internal and
413 external controls for assessing microarray technical performance. *BMC Res Notes* 2010;3:349.

414 [17] Geluk A, van Meijgaarden KE, Wilson L, et al. Longitudinal immune responses and gene expression
415 profiles in type 1 leprosy reactions. *J Clin Immunol* 2014 Feb;34(2):245-55.

416 [18] Sutherland JS, Hill PC, Adetifa IM, et al. Identification of probable early-onset biomarkers for
417 tuberculosis disease progression. *PLoS One* 2011;6(9):e25230.

418 [19] Sutherland JS, Loxton AG, Haks MC, et al. Differential gene expression of activating Fcγ
419 receptor classifies active tuberculosis regardless of human immunodeficiency virus status or
420 ethnicity. *Clin Microbiol Infect* 2014 Apr;20(4):O230-O238.

421 [20] Mihret A, Loxton AG, Bekele Y, et al. Combination of gene expression patterns in whole blood
422 discriminate between tuberculosis infection states. *BMC Infect Dis* 2014;14:257.

423 [21] Tibshirani R. Regression shrinkage and selection via the lasso. *J Royal Statist Soc B* 1996;58:267-88.

424 [22] Slood R, Schim van der Loeff MF, van Zwet EW, et al. Biomarkers can identify pulmonary
425 tuberculosis in HIV-infected drug users months prior to clinical diagnosis. *EBioMedicine* 2014;doi:
426 10.1016/j.ebiom.2014.12.001.

427 [23] Boon MR, Bakker LE, Haks MC, et al. Short-term high-fat diet increases macrophage markers in
428 skeletal muscle accompanied by impaired insulin signalling in healthy male subjects. *Clin Sci (Lond)*
429 2015 Jan;128(2):143-51.

430 [24] Maertzdorf J, Repsilber D, Parida SK, et al. Human gene expression profiles of susceptibility and
431 resistance in tuberculosis. *Genes Immun* 2011 Jan;12(1):15-22.

432 [25] Maertzdorf J, Ota M, Repsilber D, et al. Functional correlations of pathogenesis-driven gene
433 expression signatures in tuberculosis. *PLoS One* 2011;6(10):e26938.

434 [26] Ritchie ME, Silver J, Oshlack A, et al. A comparison of background correction methods for two-
435 colour microarrays. *Bioinformatics* 2007 Oct 15;23(20):2700-7.

436 [27] Smyth GK, Speed T. Normalization of cDNA microarray data. *Methods* 2003 Dec;31(4):265-73.

437 [28] Trapnell C, Williams BA, Pertea G, et al. Transcript assembly and quantification by RNA-Seq reveals
438 unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* 2010
439 May;28(5):511-5.

440 [29] Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of
441 transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol*
442 2013;14(4):R36.
443
444

Figure 1

Figure 1

	Microarray	RNA-Sequencing		
Number of genes analysed	Whole genome	Whole genome	<div>Genome wide expression assays</div> <div>Broad screen for new biomarkers on few samples.</div>	
Gene selection	No selection of genes	No selection of genes		
Assay flexibility	Not applicable	Not applicable		
Input total RNA required	± 200 ng - 1 µg	± 500 ng - 1 µg		
Cost per sample	Expensive (€100-450)	Expensive (€300-600)		
Data analysis	Complex	Complex		
RNA quantification (Dynamic range)	Relative (2-3 logs)	Absolute and relative (>5 logs)		
	NanoString	dcRT-MLPA	Fluidigm	
Number of genes analysed	± 800 genes	± 90-100 genes	Maximum 96 genes	<div>Gene signature expression assays</div> <div>Testing multicomponent biomarker signatures in larger cohorts.</div>
Gene selection	Custom gene selection possible	Custom gene selection	Custom gene selection	
Assay flexibility	Genes can be changed	Genes can be changed easily	Genes can be changed easily	
Input total RNA required	± 100-150 ng	± 20-100 ng	± 1-125 ng	
Cost per sample	Expensive (€30->400*)	Cheap (€5-7)	Expensive (€20-30)	
Data analysis	Simple	Simple	Simple	
RNA quantification (Dynamic range)	Absolute and relative (>5 logs)	Relative (>5 logs) (Absolute quantification possible)	Relative (>5 logs) (Absolute quantification possible)	
	Real-time QPCR			
Number of genes analysed	Single gene			<div>Single gene expression assays</div> <div>Testing single biomarkers in larger cohorts.</div>
Gene selection	Custom gene selection			
Assay flexibility	Gene can be selected easily			
Input total RNA required	± 0.01-10 ng			
Cost per sample	Relatively expensive (€5-7)			
Data analysis	Simple			
RNA quantification (Dynamic range)	Relative (>5 logs) (Absolute quantification possible)			

Figure 2

Hybridization

Hybridization of the two half-probes to adjacent target sequences on the cDNA.

Ligation

Ligation of the two half-probes.

PCR amplification

PCR amplification of the ligated half-probes using two different primer pairs, each labeled with a different fluorophore (FAM (blue) or HEX (green)).

Fragment run

Detection and quantification of amplification products by capillary electrophoresis. The distinctive length of each amplicon is generated by varying the length of the target-specific sequence in each probe set.

Data analysis

Trace patterns are analysed using GeneMapper software package. Category tables containing peak areas are exported for further analysis.

Figure 3

Figure A

B

BAL - TB cases - Infected part of the lung
 BAL - TB cases - Uninfected part of the lung
 BAL - Health Care Workers
 Blood - TB cases
 Blood - Health Care Workers