Supplementary material to Advanced significance analysis of microarray data based on weighted resampling: A comparative study and application to gene deletions in *Mycobacterium bovis*

Zoltan Kutalik Jacqueline Inwald Steve V. Gordon R. Glyn Hewinson Kwang-Hyun Cho Olaf Wolkenhauer[*]

May 16, 2003

1 Materials and methods

DNA microarray technology allows the large-scale analysis of whole genomes for comparative genomics. Using this technology we can therefore rapidly screen the genomes of *M.bovis* strains for deletions, using an *M.tuberculosis* H37Rv array and exploiting the $> 99.9\%$ sequence identity at the nucleotide level between the two bacilli. Ideally the microarray signal intensities should differ significantly in case of gene deletion. We carried out replicated measurement to boost the reliability of the results.

Bacterial strains and growth conditions

The *M.tuberculosis* H37Rv control strain (genomic DNA) was obtained from Colorado State University. All the *M.bovis* strains were obtained from the Veterinary Laboratory Agency collection. All the VLA strains were grown at 37°C in Middlebrook 7H9 broth (Difco) containing 0.05% Tween 80 (Sigma), 10% oleic acid/dextrose/albumin/catalase (OADC) enrichment and 2.5mM Sodium pyruvate.

Labeling of genomic DNA

Mid log phase bacteria were harvested by centrifugation at 4800 r.p.m. for 10min, the pellet was resuspended in 400μl TE buffer. The bacteria were heat killed at 80°C for 5 h. Genomic DNA was extracted using the following method. 50μl of 10mg/ml Lysozyme solution (Sigma) was added and incubated overnight in a Thermomixer (Eppendorf 5436) at 37°C, speed 10. 75μl of 10% SDS/ 50μg Proteinase K (Boehringer) was added and the tube incubated at 65°C, speed 10 for 10min followed by the addition of 100μl 5M NaCl and 100μl of CTAB and incubated at 65°C, speed 10 for 10min. The cell solution was transferred to Phase Lock Gel I Heavy (Eppendorf) and 750μl of a Chloroform: Isoamyl alcohol (24:1, BDH) solution was added, the tube was vortexed and centrifuged at $13,000$ r.p.m. for 2min. The top aqueous layer was transferred to a new tube and 0.6vol of Propan-2-ol was added and the tube placed at -20°C for 30min. The tube was centrifuged at $13,000$ r.p.m. for 15 min, the supernatant discarded and the pellet was washed with 70% cold ethanol. The pellet was air dried and resuspended in 100μl of water. DNA concentration was estimated using a spectrophotometer (GeneQuant).

Whole genomic DNA was labelled by the incorporation of Cy3 or Cy5 dCTP (Amersham) by a randomly primed polymerisation reaction. DNA (2μg) was mixed with 3μg Random primers (GIBCO-BRL) in a reaction volume of 41.5μl, heated to 95°C for 5min and snap cooled on ice. Then 5μl of 10× React 2 buffer (GIBCO-BRL), 1.5μl of Cy-dye labelled dCTP, 1μl of dNTP (5mM each dATP, dTTP, dGTP and 2mM dCTP, and 5U of Klenow (GIBCO-BRL) were added. Test strain DNA (*M.bovis*) was

[*] *To whom correspondence should be addressed.*

labelled with Cy3-dCTP and control DNA (*M.tuberculosis* H37Rv) was labelled with Cy5-dCTP. The labelling reactions were incubated at 37°C for 90min. Labelled DNA from the test and control strains were mixed and purified using a Qiagen MiniElute Kit, eluted in water.

Construction of the *M.tuberculosis H37Rv* Microarray

Whole genome microarrays were constructed by robotic spotting on to poly-L-lysine-coated glass microscope slides (MicroGrid II, Bio-Robotics, UK) of PCR amplicons (size range 60-1000bp) derived from portions of each of the predicted ORF's of the sequenced *M.tuberculosis* H37Rv. Primer pairs for each ORF were designed with Primer 3 software and selected by BLAST analysis to have minimal cross homology with other ORF's. All procedures used, including post-processing of deposited arrays, were as described by others (see references in main part of the paper).

The microarray was incubated in prehybridization solution (3.5X SSC, 0.1%SDS and 10mg/ml BSA) at 65°C for 20min. The slide was rinsed in water (MilliQ) for 20min and propan-2-ol for 1min before dried by centrifugation at 1500 r.p.m. for 5min.

The purified Cy3/Cy5-labelled DNA was adjusted to 16μl in 4× SSC and 0.3% SDS. This hybridization solution was heated to 95°C for 2min, briefly centrifuged and applied to the array under a 22-mm 2 cover slip. The slide was sealed in a humid hybridization cassette (Array-IT) and incubated at 5°C for 16-20 h in the dark. The slide was washed for 2min at 65°C in 1x SSC/0.05% SDS, followed by two washes for 2min at room temperature in 0.06× SSC and then dried by centrifugation at 1500 r.p.m. for 5min. The microarrays were scanned with an Affymetrix 428 scanner. The scanned images were further analyzed with ImaGene V4.1.

2 Detailed table of detected RD deletions method by method

Table 2 lists in detail the detected gene deletions in *M. bovis* comparing the different methods. Plus signs refer to recognized gene deletions, while minus signs indicate that that particular method failed to pick up that gene as a deleted one. Stars indicate that that particular gene was excluded from further analysis during the normalization process. WR refers to our weighted resampling, RT to regularized t-test, SAM to SAM method. Gene names in bold face show the differences between the methods. Here we applied all methods to produce lower than 5% FDR.

Gene Id	**WR**	**RT**	**SAM**
glgY (5H17)	-	-	-
Rv1964 (9N7)	+	-	+
Rv1965 (9O7)	+	+	+
mce3 (6L5)	+	+	+
Rv1967 (9P7)	+	+	+
Rv1968 (6M5)	+	+	-
Rv1969 (6N5)	+	+	+
lprM (6O5)	+	+	+
Rv1971 (6P5)	*	*	*
Rv1972 (6I6)	+	+	+
Rv1973 (6J6)	+	+	+
Rv1974 (9I8)	+	+	+
Rv1975 (6K6)	+	+	+
Rv1976c (6L6)	+	+	+
Rv1977 (6M6)	+	+	+
Rv0021c (1E3)	-	-	-
Rv1255c (7N7)	+	+	+
Rv1256c (7O7)	+	+	+
Rv1257c (7P7)	+	+	+
Rv1506c (5E13)	+	+	+
Rv1507c (5F13)	+	+	+
Rv1508c (5G13)	+	+	+
Rv1509 (5H13)	+	+	+
Rv1510 (10C20)	+	+	+
gmdA (10D20)	+	+	+
epiA (5A14)	+	+	+
Rv1513 (10E20)	+	+	+
Rv1514c (5B14)	+	-	+
Rv1515c (5C14)	+	-	+
Rv1516c (10F20)	+	+	+
Rv2645 (1K18)	+	+	+
Rv2646 (1L18)	+	+	+
Rv2647 (9I14)	-	-	-
Rv2648 (10B2)	+	+	+
Rv2649 (10C2)	+	+	+
Rv2650c (10D2)	+	-	+
Rv2651c (10E2)	+	+	+
Rv2652c (1M18)	+	+	+
Rv2653c (10F2)	+	+	+
Rv2654c (10G2)	+	-	-
Rv2655c (1N18)	+	+	+
Rv2656c (10H2)	+	+	+
Rv2657c (10A3)	-	-	-
Rv2658c (1O18)	+	+	+
Rv2659c (1P18)	+	+	+
Rv2660c (10B3)	-	-	-
alr (4B15)	-	-	-
Rv3424c (9C24)	-	-	-
PPE (4C15)	-	-	-
PPE(9D24)	+	+	+
Rv3427c (4D15)	+	+	+
Rv3428c (4E15)	+	+	+
Rv3616c (2G19)	-	-	-
ephA (2H19)	+	+	+
Rv3618 (2A20)	+	+	+
Rv3619c (11H1)	-	-	-
Rv3620c (11A2)	-	-	-
PPE (2B20)	+	+	+
PE (11B2)	+	-	+
lpqG (2C20)	-	-	-
cysA3(10L7)	-	-	-
sseC (10M7)	-	-	-
moaE(10N7)	+	-	+
Rv3120 (3H23)	+	+	+
Rv3121 (3A24)	+	+	+
Rv2347c (8L15)	-	-	-
Rv2348c (8M15)	+	+	+
plcC (8N15)	+	+	+
plcB (8O15)	+	+	+
plcA (8P15)	+	+	+
PPE (8I16)	+	+	+
cobL (6P7)	-	-	-
Rv2073c (6I8)	+	+	+
Rv2074 (6J8)	+	+	+

3 Scatter plot of log-intensities

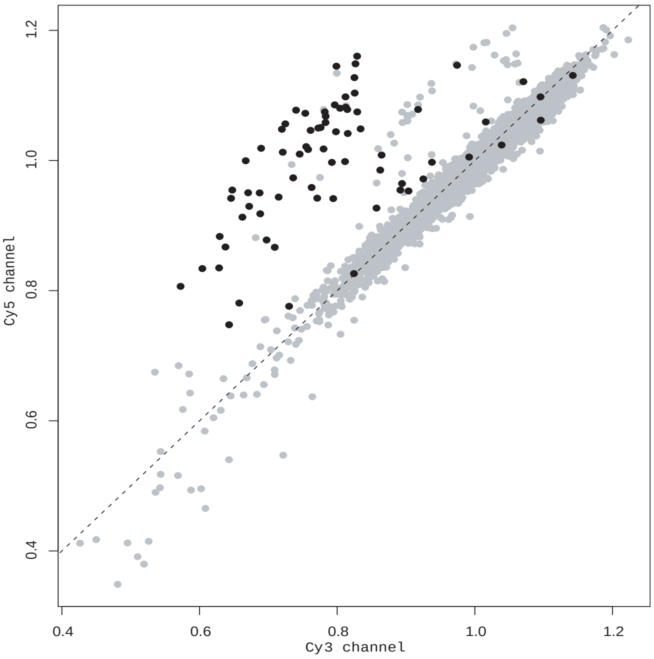

Figure 1: Scatter plot of the average log-intensities for all genes. This figure shows how well hidden significant differences of channel intensities can be. On the axes Cy3 and Cy5 channel activities are measured each dot represents a gene. Black ones mark RD gene deletions from *M.bovis* while grey dots show the genes on rest of the genome.

4 Appendix

4.1 Variance Stabilization

The only theorem we use for proving the equivalence with the modified t-test is the Lagrange equality (Rudin 1976), stating that for any differentiable function f and for any two points, $x < y$; there exists a number $\xi_{x,y} \in (x, y)$ such that

$$f(x) - f(y) = f'(\xi_{x,y})(x - y) \ .$$

Applying this to function h

$$\begin{aligned} \Delta h_g = h(\bar{x}(g)) - h(\bar{y}(g)) &= h'(\xi_g)(\bar{x}(g) - \bar{y}(g)) \\ &= \frac{\bar{x}(g) - \bar{y}(g)}{\sqrt{v(\xi_g)}} \ . \end{aligned}$$

Moreover, in this case - where quadratic regression was used - we know explicitly that $\xi_g = \frac{\bar{x}+\bar{y}}{2}$. Thus Δh_g is a regularized t-score, where the weight of prior knowledge is equal to one and is extracted from the data by quadratic regression.

4.2 Wilk's Lambda Score

The Wilk's lambda is defined as follows

$$\Lambda = \frac{W}{T} \tag{1}$$

where

$$T = \sum_{i=1}^{n} (x_i - (\bar{x} + \bar{y})/2)^2 + \sum_{i=1}^{n} (y_i - (\bar{x} + \bar{y})/2)^2 \tag{2}$$

and

$$W = \sum_{i=1}^{n} (x_i - \bar{x})^2 + \sum_{i=1}^{n} (y_i - \bar{y})^2 \ . \tag{3}$$

Theorem 1. *For two-class comparison Wilk's lambda is a strictly monotone function of the absolute value the t-statistic. Therefore its scoring of significance is equivalent to that of the t-test.*

Proof. We can rewrite (2) into

$$\begin{aligned} T &= \sum_{i=1}^{n} ([x_i - \bar{x}] + (\bar{x} - \bar{y})/2)^2 \\ &+ \sum_{i=1}^{n} ([y_i - \bar{y}] + (\bar{y} - \bar{x})/2)^2 \end{aligned} \tag{4}$$

which is by expanding and then simplifying

$$T = \sum_{i=1}^{n} (x_i - \bar{x})^2 + \sum_{i=1}^{n} (y_i - \bar{y})^2 + n(\bar{x} - \bar{y})^2/2 \ . \tag{5}$$

Now substitute (5) into (1) we obtain

$$\Lambda^{-1} = \frac{T}{W} = 1 + \frac{n(\bar{x} - \bar{y})^2/2}{(n-1)(s_x^2 + s_y^2)} \tag{6}$$

Finally by rearranging (6)

$$\sqrt{2(n-1)(\Lambda^{-1} - 1)} = \frac{\mid \bar{x} - \bar{y} \mid}{\sqrt{s_x^2/n + s_y^2/n}}$$

where the right hand side is the absolute value of the t-statistic. Thus Wilk's lambda is a monotone (decreasing) function of the t-statistic. □

References

Rudin, W.: 1976, *Principles of Mathematical Analysis*, McGraw-Hill.