Description of data tables and analysis methods for hypoxia project

The tables include all genes that have average induction or repression values of 1.5 or greater. They are listed in genomic order for gene finding simplicity. Neither are all genes in the tables reproducibly induced nor should it be seen as a complete list of all possibly induced genes as the array represents approximately 90-97% of all H37Rv genes depending on the quality of the individual array.

Data tables are structured using the following columns:

Spot: The spot column describes the position of the DNA spot within the microarray. Some genes are included multiple times on the array in different sections.

Name: The name category lists the H37Rv number given to each ORF by the Sanger Centre. The genes are numbered in the order found within the genome starting from the origin of replication. O-strain indicates a sequence of a putative gene not found in the H37Rv genome but found in strain CDC 1551 (Oshkosh strain) sequenced by The Institute for Genomic Research (TIGR).

Gene: The gene category indicates the annotation of predicted genes on the first version of Tuberculist provided by the Sanger Centre.

CH1 average corrected intensity (Cye3): The CH1 (channel 1) intensity value is the average (from all arrays listed) Cye3 fluorescent signal at the particular spot as measured by an Axon GenePix 4000 microarray scanner. The DNA spots were located using ScanAlyze written by Michael Eisen. Cye3 dUTP was used to label cDNA synthesized from reference RNA. Reference RNA was isolated from *M. tuberculosis* H37Rv at 20% oxygen growing at an OD of 0.1, which did not receive a hypoxic shift.

CH2 average corrected intensity (Cye5): The CH2 (channel 2) intensity value is the average (from all arrays listed) Cye5 fluorescent signal at the particular spot as measured by an Axon GenePix 4000 microarray scanner. The DNA spots were located using ScanAlyze written by Michael Eisen. Cye5 dUTP was used to label cDNA synthesized from hypoxic shift sample RNA. Hypoxic shift RNA was isolated from *M. tuberculosis* H37Rv growing at an OD of 0.1, shifted from 20% oxygen to 0.2% and grown for 2 hours.

For tables of induced genes

CH2/CH1 (Average corrected ratio): The CH2/CH1 ratio is the average ratio calculated from CH2/CH1 ratios from the individual array experiments listed on the right hand side of the table. It is NOT calculated from the Average CH1 and CH2 values. If only one individual array ratio is above the cutoff ratio (1.5) the gene is not included as the data is not reproducible.

StD: The standard deviation is calculated from the ratios of the individual array experiments.

SEM: The standard error of the mean is calculated by dividing the standard deviation by the square root of the number of data points.

For tables of repressed genes

CH1/CH2 (Average corrected ratio): The CH1/CH2 ratio is the average ratio calculated from CH1/CH2 ratios from the individual array experiments listed on the right hand side of the table. It is NOT calculated from the Average CH1 and CH2 values. If only one individual array ratio is above the cutoff ratio (1.5) the gene is not included as the data is not reproducible.

StD: The standard deviation is calculated from the ratios of the individual array experiments.

SEM: The standard error of the mean is calculated by dividing the standard deviation by the square root of the number of data points.

Gene product: The gene product information describes known or predicted function of gene products as annotated by the Sanger Centre on TubercuList.

PCR F: PCR flag numbers indicate quality of the DNA printed on the microarray as determined by gel electrophoresis of gene specific PCR products. A PCR F number of 1.1 indicates a good product and other numbers indicate a questionable PCR product of poor yield, more than one fragment amplified or the amplifed fragment was not the predicted size. However, a good PCR product does not guarantee a good spot on each array.

Individual array ratios (CH2/CH1): The adjusted ratios listed under the individual array names are calculated from raw data generated by the ScanAlyze program by performing two adjustments as described in the following steps.

1) Determination of Cye3 and Cye5 signal intensity normalization factor: The overall signal intensity of all gene specific spots should be approximately the same in both the CH1 (Cye3) and CH2 (Cye5) category under most biological conditions assuming that the majority of genes are not regulated. The intensity values of CH2 are adjusted with a normalization factor to accommodate unequal fluorescence and subjective scanning of the Cye3 and Cye5 channels. The normalization factor is calculated by comparing the intensity from gene specific spots in both

channels. To avoid skewing the normalization factor, due to highly regulated genes, the spots that have a fluorescence ratio in the top 5% and the bottom 5% are ignored in calculating the factor. The ratio of the total fluorescence within gene specific spots (excluding the most regulated gene spots) of Cye5 over Cye3 is calculated and used to adjust the Cye5 channel intensity of all spots on the array. The sum of the adjusted Cye5 channel intensities should then equal the sum of the Cye3 channel intensities.

2) Adjustment of low background values: To eliminate relatively large fluctuations in what are essentially background intensity values from genes that are not expressed to a measurable level, a minimum background (noise) value is calculated for each channel. To calculate an arbitrary value for the noise, the lowest 20% intensity values are averaged for each channel. This calculation is performed with the conservative assumption that less than 80% of the genome is expressed under any particular condition, therefore the lowest 20% of intensity values should be within the noise. After calculation of the noise values for each channel, the noise values are used to replace any value lower than the noise value.
This adjustment produces lower ratios but removes much of the variation in ratios from genes highly induced from a condition of very low (background) expression.

Example: If CH2 is 2000 in two different experiments and in experiment one the CH1 value is 100 and in experiment 2 it is 5, but the noise value is 100 in both, then without correction exp. 1 has a ratio of 20 while exp. 2 has a ratio of 400. However, if adjusted both ratios become 20 with a significantly lower average ratio but a better standard deviation.

ND is used to indicate no data for the spot if it was either flagged as a bad spot based on visual inspection or both channel 1 and 2 values were within the noise value. In these cases the data is meaningless and discarded.

Biol. replicate: Arrays listed under the same biological replicate category were hybridized with cDNA made from the same RNA pools. While arrays listed under different biological replicate categories were hybridized with cDNA made from RNA isolated from different biological experiments.

Array name: Indicates the specific name for each array experiment.