Supplemental Material

A Genomic Portrait of the Genetic Architecture and Regulatory Impact of microRNA Expression in Response to Infection

Katherine J. Siddle, Matthieu Deschamps, Ludovic Tailleux, Yohann Nédélec, Julien Pothlichet, Geanncarlo Lugo-Villarino, Valentina Libri, Brigitte Gicquel, Olivier Neyrolles, Guillaume Laval, Etienne Patin, Luis B. Barreiro, Lluís Quintana-Murci

Supplemental Methods

Supplemental Figures

Additional files provided as separate files.

Tables listed below are provided as one Excel file with multiple worksheets.

Supplemental Methods

miRNA expression analyses

Genome-wide miRNA expression was profiled using the Agilent Human miRNA microarray (Release 16.0) consisting of 56,044 probes representing 1205 human and 144 human viral miRNAs annotated in Build 16 of miRBase (http://www.mirbase.org/) (Griffiths-Jones et al. 2006). To minimize potential batch effects, samples were randomized across array batches while ensuring that infected and non-infected samples from the same individual were kept in the same batch. To confirm the technical reproducibility of our arrays, we performed technical replicates for 11 samples. We found miRNA expression to be highly correlated between these replicates (mean Pearson's r = 0.96, compared to r=0.93 between individuals in the same condition). Unless otherwise indicated, all analyses were carried out using the R statistical framework.

Initial analysis and quality control of the microarrays were performed using Agilent's Feature Extraction Software. Subsequent pre-processing was performed using the Bioconductor package AgiMicroRna (Lopez-Romero 2011). All arrays were normalized using the Robust Multi-Array Average (RMA) method including background correction (Irizarry et al. 2006). Probes for which expression was not detected, or was indistinguishable from background levels in at least 10% of samples (N<7) in both the non-infected and infected state, were removed. Using Principal Component (PC) analysis, we identified two samples as outliers, which were discarded from the analyses, owing to low number of expressed miRNAs. We further corrected for batch effects and removed the first PC, using a linear model, as this was associated with technical variation in sample processing.

To identify differentially expressed miRNAs upon MTB infection, we applied a linear model with a fixed effect for MTB treatment. Moderated statistics were obtained using the empirical Bayes approach, implemented in the Bioconductor package limma, (Smyth 2004) and multiple-testing corrected p-values were calculated using the Benjamini and Hochberg FDR (Benjamini and Hochberg 1995).

Mapping of expression quantitative trait loci (eQTLs)

Associations between SNP genotypes (GEO Accession Number GSE34588) (Barreiro et al. 2012) and miRNA expression levels were calculated using a linear regression model, assuming an additive effect of alleles on expression, in infected and non-infected samples. Given the reduced power to detect associations for rare SNPs and infrequently expressed

miRNAs, we restricted eQTL mapping to SNPs with a minor allele frequency higher than 10% (N=570,803) and miRNAs that were expressed in at least 50% of samples in a given condition (N=266; 250 and 264 in non-infected and infected samples, respectively). We improved the power to detect eQTLs by quantile normalization and through the regression of a given number of PCs (1 and 4 for non-infected and infected samples, respectively), to account for unknown confounders (Supplemental Fig. S9A,B). The number of PCs removed was determined based on maximizing the number of significant associations in a given condition. However, correlations between p-values with and without PC removal were high (>0.7), showing a negligible impact on the relationship between genotypes and expression levels (Supplemental Fig. S9C).

We mapped putative cis-eQTLs using a region of 200 kb centered on the mature miRNA and recorded the lowest p-value obtained by regressing the expression level of each miRNA against the genotype of each SNP within the 200 kb window. We removed all probes that mapped to more than one genomic location. We estimated the FDR by comparing the observed to the null distribution, generated using the lowest p-values observed for each miRNA in 100 permutations of expression values (Pickrell et al. 2010; Barreiro et al. 2012). We detected genotype-treatment interaction effects by Bayesian regression with the software BRIdGE (Maranville et al. 2011), using scripts available on the authors' web pages. We used default effect sizes (0.8, 1, 1.2 & 1.6), a threshold of 0.001, and a posterior probability cut-off of 0.7 for determining significance. We mapped trans-eQTLs by performing a genome-wide association of miRNA expression levels against all genotyped SNPs, as well as against a subset of SNPs (N=4) previously identified as susceptibility loci for TB by GWAs (http://www.genome.gov/26525384) (Hindorff et al. 2009). Multiple testing corrections were performed using a Bonferroni correction at the 95% significance level.

To study the genomic context of these miR-eQTLs, we assessed their overlap with active genomic regions, using ChIP-seq and DNase-seq peak data for human monocytes (RO1746) from the ENCODE project (http://encodeproject.org/ENCODE/) (The ENCODE Project Consortium 2012). We used a Fisher's exact test to calculate enrichments of miR-eQTLs in regions associated with histone modifications or open chromatin.

For the fine-mapping of miR-eQTL regions, we imputed genotypes for SNPs not present on our genotyping array with IMPUTE2 (Howie et al. 2009), using integrated haplotype data from Phase 1 of the 1000 Genomes project (The 1000 Genomes Project Consortium 2012). We defined sets of SNPs in high linkage disequilibrium (LD, $r^2 > 0.8$) with our array-based eQTL SNPs, to test for their presence among dsQTLs (Degner et al. 2012) and mRNA-eQTLs

(Barreiro et al. 2012; Xia et al. 2012). To refine miR-eQTL signals, we repeated eQTL mapping, as described above, using all array-based and imputed SNPs within a 1 Mb region around the initial set of detected miR-eQTLs. Regional associations were visualized using LocusZoom (Pruim et al. 2010).

As the power to detect a significant association between a genotype and the expression of a given gene depends on a combination of factors, including; sample size, minor allele frequency (MAF), expression level, inter-individual variability in expression and the fold change between genotypes, we performed simulations to quantify the power of the present study to map eQTLs varying these parameters (Supplemental Fig. S2). For sample sizes ranging from 20-150 individuals we first simulated genotypes for a SNP with a MAF of either 0.2 or 0.5. We next simulated expression levels for each genotype using a normal distribution with mean and standard deviation derived from our observed data, and varying the fold change between the mean expression levels of the homozygote genotypes. Simulated expression levels were then quantile normalized, across all genotypes, and the association calculated using a linear model as described above. We performed 100 simulations for each set of conditions.

Cell transfection assays of miR-29 inhibitors and mimics

Immature DCs from 4 unrelated individuals were transfected on day 5 using HiPerFect® transfection reagent. Cells were harvested and resuspended at 10^6 cells/ml in complete medium without cytokines. 2 ml of cell suspension was added at once to 1ml of transfection medium (948 µl RPMI, 45 µl HiPerFect® and 4 µl of 20 µM oligonucleotide solution) in 6-well plates. After 6 h incubation at 37°C, 4 ml of complete medium containing 40ng/ml IL4 and 20ng/ml CSF2 was added to the cells and incubated at 37°C over night. miRCURY LNA Power Inhibitors were purchased from Exiqon (miR-29 family 460039, control 199020-00) and miRIDIAN microRNA mimics from Thermo Fisher (miR-29a C-300504-07, control CN-001000-01). At day 6, transfection efficiency was assessed by flow cytometry using a fluorescently labeled control oligonucleotide (Exiqon, 199020-04), and found to be on average 77% (Supplemental Fig. S11). Transfected cells were then infected for 24 h with MTB (H37Rv) as miR-29 induction peaks at this time (Supplemental Fig. S12).

miR-29 quantification

Total RNA was extracted using the miRNeasy kit (Qiagen). To quantify miR-29 expression upon MTB infection, cDNA was synthesized and quantitative real-time PCR (qPCR)

performed using the Qiagen miScript PCR system and primers (miScript II RT Kit: 218161; miScript SYBR® Green PCR kit: 218073; miR-29a-3p MS00003262; miR-29b-3p MS00006566; miR-29c-3p MS00003269; U6 MS00033740). To validate miR-29 perturbation in transfected cells, cDNA was synthesized using the miRCURY LNA™ Universal cDNA Synthesis kit II (Exiqon, 203301) and qPCR performed using the ExiLENT SYBR® Green master mix and specific primers provided by Exiqon (miR-29a-3p 204698; miR-29b-3p 204679; miR-29c-3p 207429; RNU6-1 203907). qPCRs were performed in a 7900 Real-time PCR system (Applied Biosystem). The relative expression of miR-29a, b and c, normalized to the endogenous control RNU6-1 was calculated using the $\Delta\Delta C_t$ method (Livak and Schmittgen 2001).

Gene expression analysis

Genome-wide profiling of non-infected and MTB-infected samples was obtained by hybridizing RNA to Illumina HumanHT-12 v4 Expression BeadChip arrays. RNA quality was assessed with the Agilent Bioanalyzer and all samples were of high integrity (RIN>8). Technical replicates were performed for 4 samples. Gene expression levels among technical replicates were highly correlated compared to biological replicates (mean Pearson's r = 0.997, $p< 1\times10^{-20}$), indicating high reproducibility. Initial microarray analysis was performed using the Bioconductor package lumi (Du et al. 2008). We performed a background correction and variance stabilizing transformation before quantile normalizing the data. We removed all probes that (i) did not map to a unique RefSeq ID, (ii) mapped to poorly characterized genes without an Ensembl ID, (iii) contained one or more HapMap SNPs (MAF>0.1 in the CEU population, i.e., Utah residents with ancestry from northern and western Europe) (The International HapMap Consortium 2010), or (iv) were detected in less than half of the samples. For genes represented by multiple probes, we used the mean probe expression level. These preprocessing steps yielded a final set of 12,722 probes, corresponding to 9610 genes, which were used for downstream analyses. Differential expression analysis was performed using the Bioconductor package limma (Smyth 2004) as described above. Enrichments of miR-29a predicted targets (Friedman et al. 2009) in differentially expressed genes were calculated using a Fisher's exact test. Differences in mean fold changes upon miR-29 perturbation between predicted targets, correlated genes identified by computational analyses and all genes were calculated using a t-test. Enrichments of functional Gene Ontology categories and KEGG pathways among differentially up- and down-regulated genes were

computed using GeneTrail (Backes et al. 2007), as described above, using all detected genes as a background set.

Quantification of cytokine and chemokine levels in supernatants

Supernatants of untreated and MTB-infected DCs were 0.22μm filtered (Millipore) and kept at -80°C. We measured supernatant levels of 25 cytokines/chemokines, in duplicate, using the Human Cytokine Magnetic 25-Plex Panel (Invitrogen) according to manufacturer's instructions. We calculated the average quantity across technical replicates for each protein and used this for subsequent analyses. IL4 and CSF2 were excluded from the analysis as these cytokines were artificially added to the culture medium during DC derivation. CCL11 was removed from the analysis as the median concentration in infected samples was lower than the detection limit specified by the manufacturer. Differences in secretion levels between conditions were calculated using a Wilcoxon paired rank sum test.

References

Backes C, Keller A, Kuentzer J, Kneissl B, Comtesse N, Elnakady YA, Muller R, Meese E, Lenhof HP. 2007. GeneTrail - advanced gene set enrichment analysis. *Nucleic Acids Res* **35**: W186-192.

Barreiro LB, Tailleux L, Pai AA, Gicquel B, Marioni JC, Gilad Y. 2012. Deciphering the genetic architecture of variation in the immune response to Mycobacterium tuberculosis infection. *Proc Natl Acad Sci U S A* **109**: 1204-1209.

Benjamini Y, Hochberg Y. 1995. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *J R Statist Soc B* **57**: 289-300.

Degner JF, Pai AA, Pique-Regi R, Veyrieras JB, Gaffney DJ, Pickrell JK, De Leon S, Michelini K, Lewellen N, Crawford GE, et al. 2012. DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482**: 390-394.

Du P, Kibbe WA, Lin SM. 2008. lumi: a pipeline for processing Illumina microarray. *Bioinformatics* **24**: 1547-1548.

Friedman RC, Farh KK, Burge CB, Bartel DP. 2009. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res* **19**: 92-105.

Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ. 2006. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res* **34**: D140-144.

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A* **106**: 9362-9367.

Howie BN, Donnelly P, Marchini J. 2009. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* **5**: e1000529.

Irizarry RA, Wu Z, Jaffee HA. 2006. Comparison of Affymetrix GeneChip expression measures. *Bioinformatics* **22**: 789-794.

Livak KJ, Schmittgen TD. 2001. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods* **25**: 402-408.

Lopez-Romero P. 2011. Pre-processing and differential expression analysis of Agilent microRNA arrays using the AgiMicroRna Bioconductor library. *BMC Genomics* **12**: 64.

Maranville JC, Luca F, Richards AL, Wen X, Witonsky DB, Baxter S, Stephens M, Di Rienzo A. 2011. Interactions between glucocorticoid treatment and cis-regulatory polymorphisms contribute to cellular response phenotypes. *PLoS Genet* **7**: e1002162.

Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras JB, Stephens M, Gilad Y, Pritchard JK. 2010. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**: 768-772.

Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP, Boehnke M, Abecasis GR, Willer CJ. 2010. LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* **26**: 2336-2337.

Smyth GK. 2004. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* **3**: Article3.

The 1000 Genomes Project Consortium. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**: 56-65.

The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57-74.

The International HapMap Consortium. 2010. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**: 52-58.

Xia K, Shabalin AA, Huang S, Madar V, Zhou YH, Wang W, Zou F, Sun W, Sullivan PF, Wright FA. 2012. seeQTL: a searchable database for human eQTLs. *Bioinformatics* **28**: 451-452.

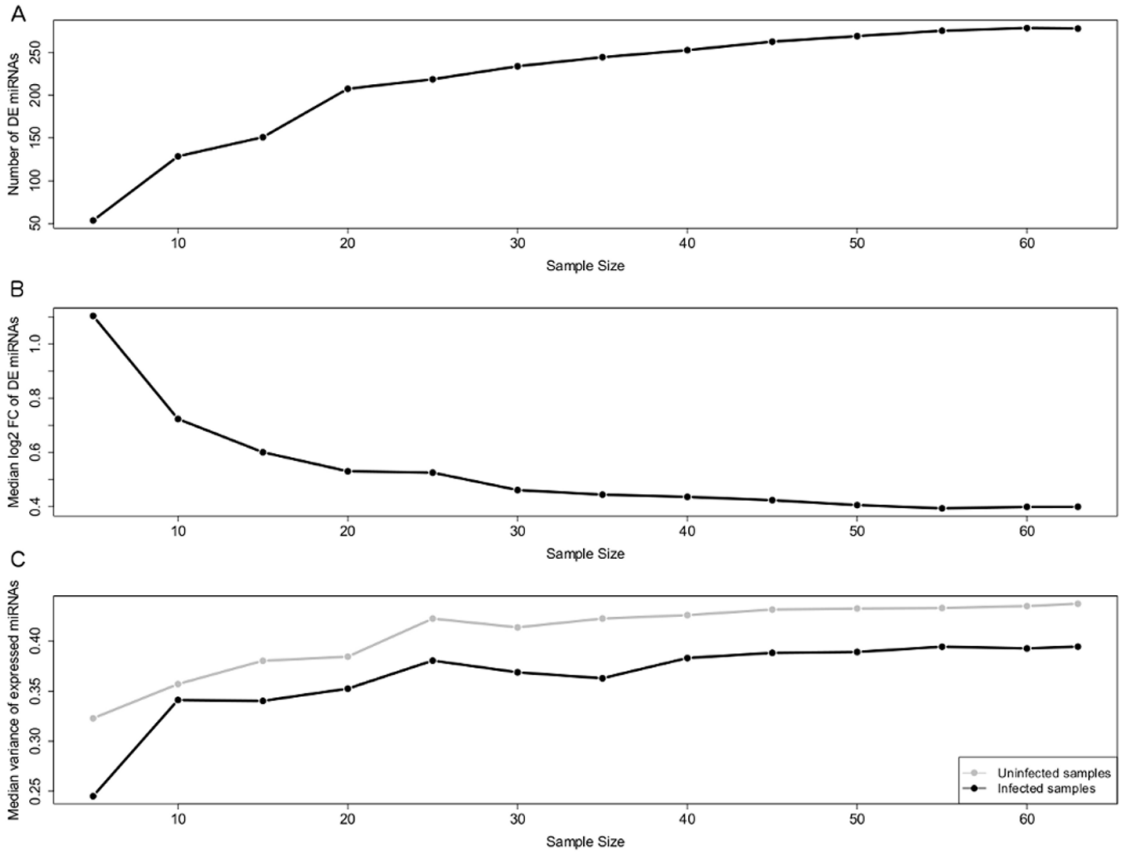

Figure S1. Relationship between sample size and differentially expressed miRNAs. Plots showing the changes in (*A*) mean number, (*B*) median $\log^2$ fold change and (*C*) median variance of differentially expressed miRNAs with increasing sample size. We assessed our resolution for differential expression analyses, compared to the smaller sample sizes more common for cellular studies, by repeating the analysis using randomly resampled subsets of 5 to 63 individuals, across 10 replicates. We observed an increase in the number of significantly differentially expressed miRNAs with larger sample sizes (*A*). At the same time, the average fold change of differentially expressed miRNAs decreased with increasing sample size (*B*), and the average variance increased (*C*). This demonstrates that larger sample sizes improve the detection of differential expression for miRNAs where the change is more subtle and/or inter-individual variation in expression more pronounced. Moreover, the plateau reached in all measures around 30 individuals indicates that the large number of differentially expressed miRNAs detected probably represents close to all changes occurring in miRNA expression profiles upon MTB infection.

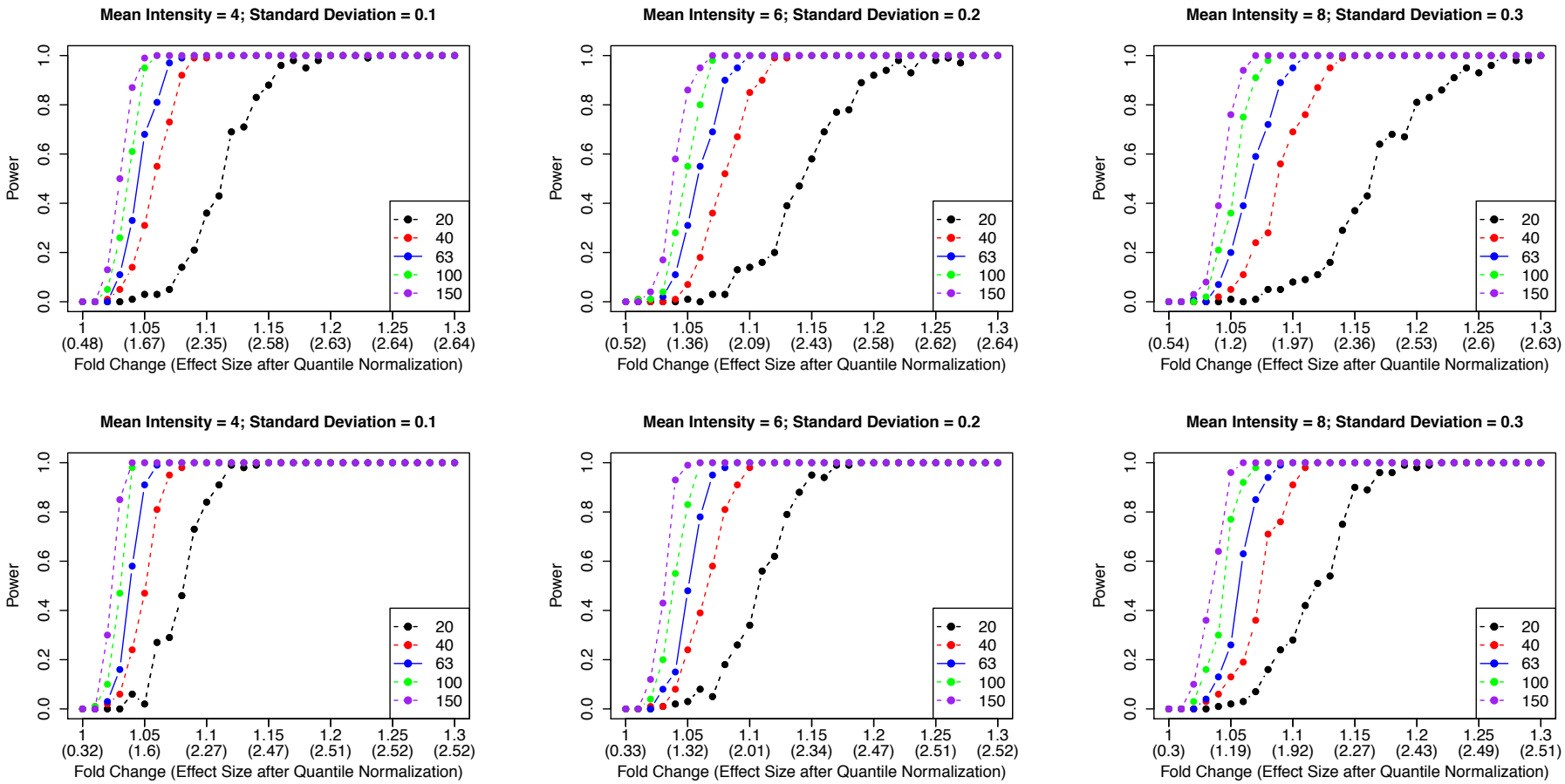

Figure S2. Power estimations for the detection of eQTLs. Plots showing the power to detect an association between a genotype and expression levels at $p<1\times10^{-4}$ for a range of sample sizes from 20-150 individuals, considering two different MAFs (0.2 (top) and 0.5 (bottom)) and three combinations of mean expression level and within genotype standard deviation in expression (4 and 0.1, 6 and 0.2, and 8 and 0.3). These values

correspond approximately to the 25th, 50th and 75th percentiles of our observed values. We simulated fold changes between mean expression levels of the homozygote genotypes from 1 to 1.3. Effect size (given in brackets on the x-axis of the plots) was calculated as the standardized difference between the means of quantile normalized expression levels for the homozygote genotypes. The sample size of the current study (63 individuals) is shown with a solid blue line. We observed that the sample size used in this study gives reasonable power to detect associations when the effect size is greater than 1.5 and has almost 100% power when the effect size exceeds 2. In addition, the current sample size consistently performs almost as well as 100 individuals.

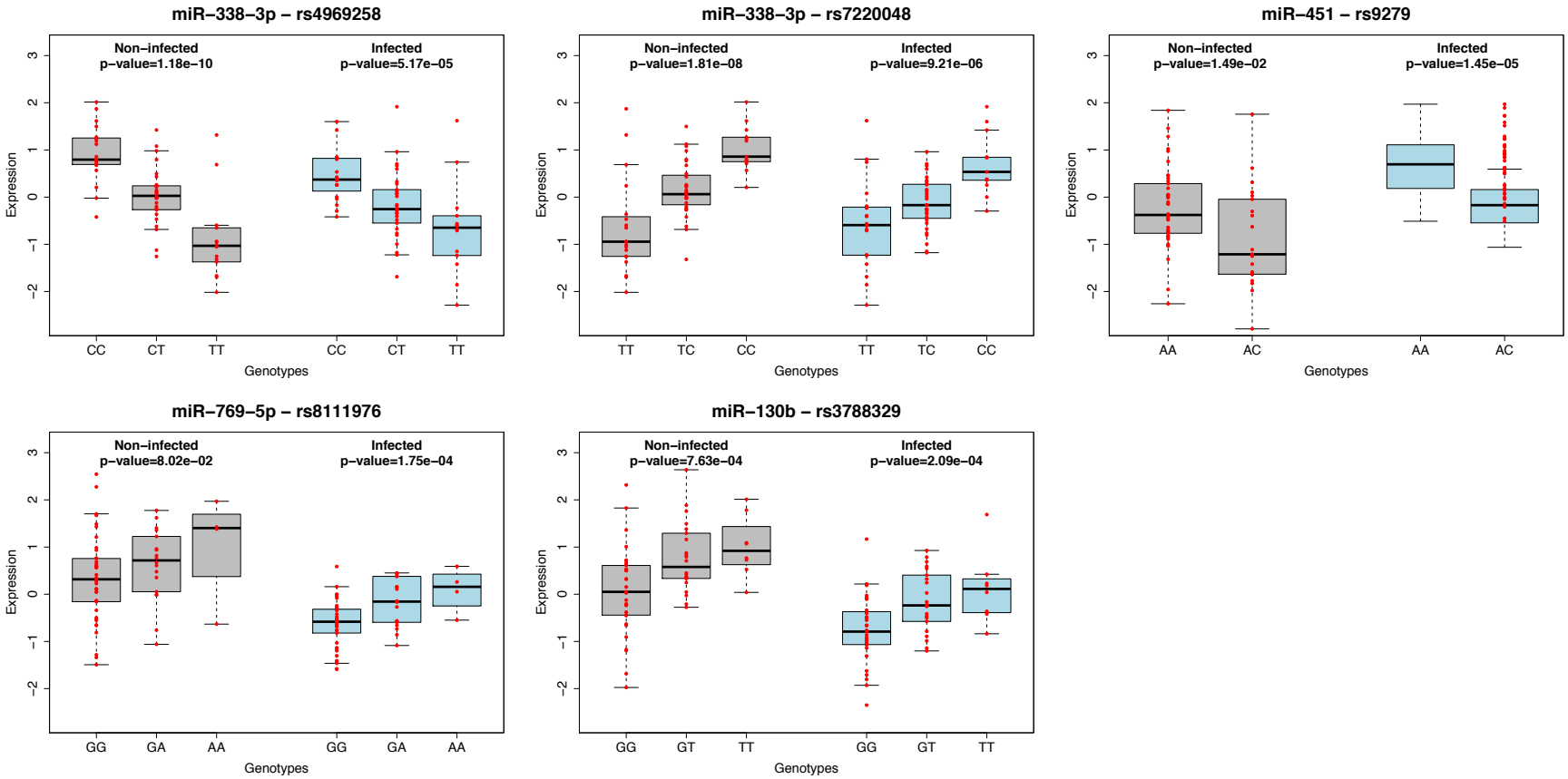

Figure S3. Boxplots of cis-eQTLs identified in this study. Boxplots showing all cis eQTLs identified in non-infected and/or infected DCs that did not satisfy the conditions to be considered as a response eQTL. Only one of these associations (miR-338-3p) reached genome-wide significance in both infected and non-infected samples, although different SNPs showed the strongest association with expression of the miRNA in each condition. BRIdGE analysis allowed us to refine this association by identifying an interaction effect for miR-338-3p/ rs4969258, where the association between the genotype and the molecular phenotype differs in its magnitude. All remaining associations were significant only upon MTB infection, although similar tendencies with regards to the effect of the genotype on miRNA expression before and after infection were observed.

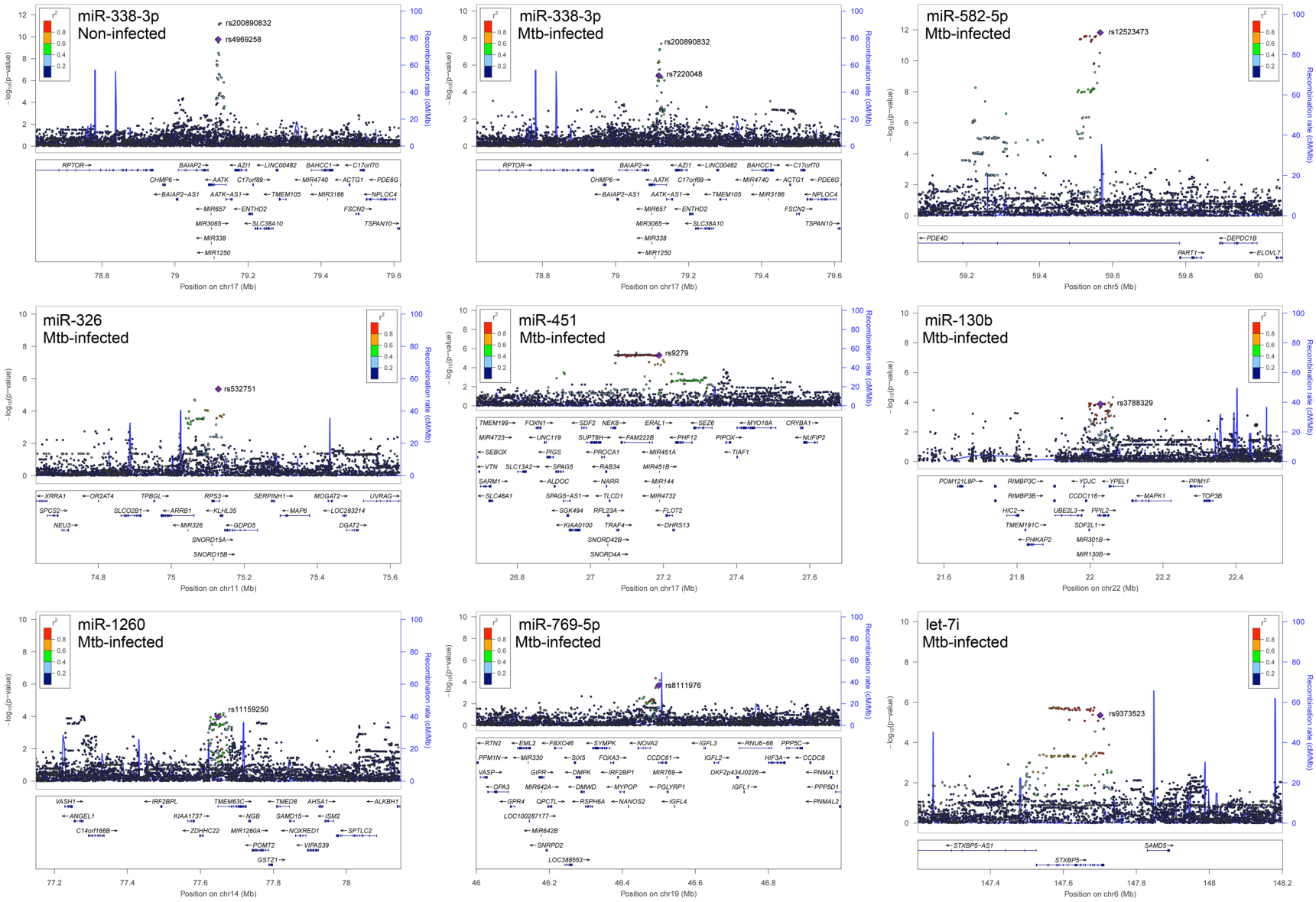
miR-338-3p
Non-infected
rs200890832
rs4969258
miR-338-3p
Mtb-infected
rs200890832
rs7220048
miR-582-5p
Mtb-infected
rs12523473
miR-326
Mtb-infected
rs532751
miR-451
Mtb-infected
rs9279
miR-130b
Mtb-infected
rs3788329
miR-1260
Mtb-infected
rs11159250
miR-769-5p
Mtb-infected
rs8111976
let-7i
Mtb-infected
rs9373523
Position on chr17 (Mb)
Position on chr5 (Mb)
Position on chr11 (Mb)
Position on chr22 (Mb)
Position on chr14 (Mb)
Position on chr19 (Mb)
Position on chr6 (Mb)
Recombination rate (cM/Mb)

Figure S4. Regional association plots of genotyped and imputed SNPs surrounding detected miR-eQTLs. To refine our array-based miR-eQTL signals, SNPs that were not present on our array were imputed using data from the 1000 Genomes project, and eQTL mapping was repeated for all SNPs. Recombination rates and LD values are based on European-descent populations from the 1000 Genomes Project. The strongest array-based miR-eQTL SNP is named on the plot and denoted by a purple diamond. Grey circles indicate that no LD information is available for a given SNP. In almost all cases, the strongest array-based miR-eQTL SNP showed, or was in strong LD with the SNP showing, the strongest association with miRNA expression.

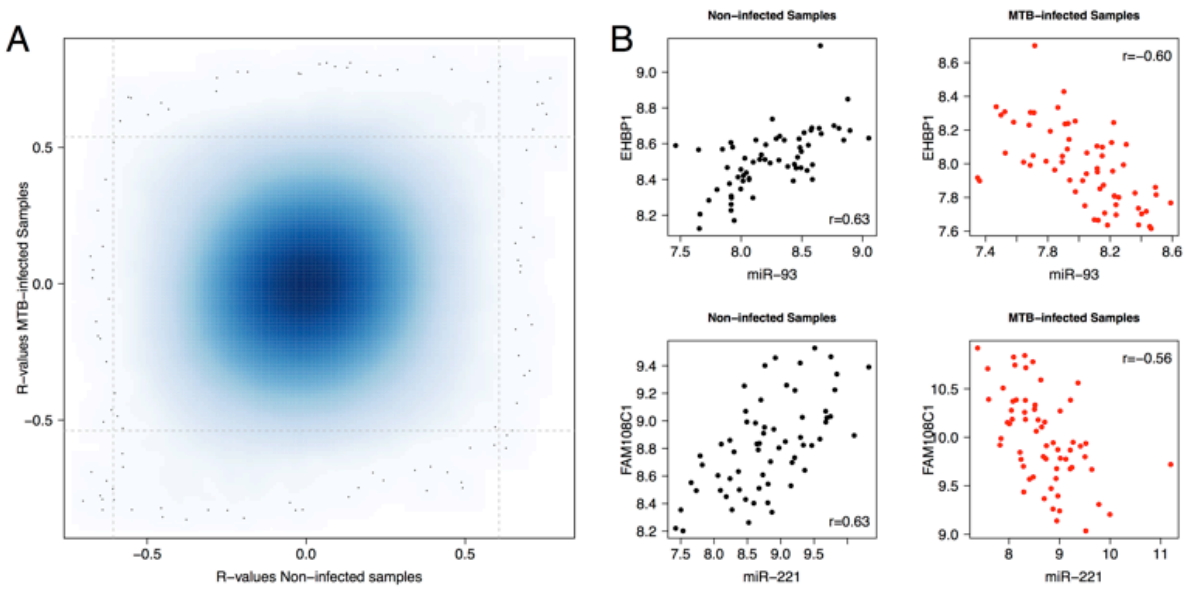

Figure S5. Impact of infection on miRNA-mRNA correlations. (*A*) Smooth scatter plot showing miRNA-mRNA correlations in non-infected and infected samples. Dashed lines represent the significance thresholds used to define the set of significantly correlated miRNA-mRNA pairs reported in the manuscript. Forty miRNA-mRNA pairs were significantly correlated in both analyses. The majority of this overlap was accounted for by correlations with 2 miRNAs – miR-155 and miR-210 – that together accounted for 88% of all overlapping pairs. Interestingly, while the vast majority of these 40 pairs showed the same direction of correlation before and after infection (r=0.88), 2 miRNA-mRNA pairs showed opposing correlations in the 2 conditions. (*B*) Scatter plots for these 2 miRNA-mRNA pairs, showing miRNA and mRNA expression levels for 63 individuals before and after infection. Both of these pairs were significantly positively correlated before and significantly negatively correlated after infection. As neither of these genes is predicted to be a direct target of the miRNA with which they are correlated, these changes are likely to reflect more general changes in the regulatory network upon infection.

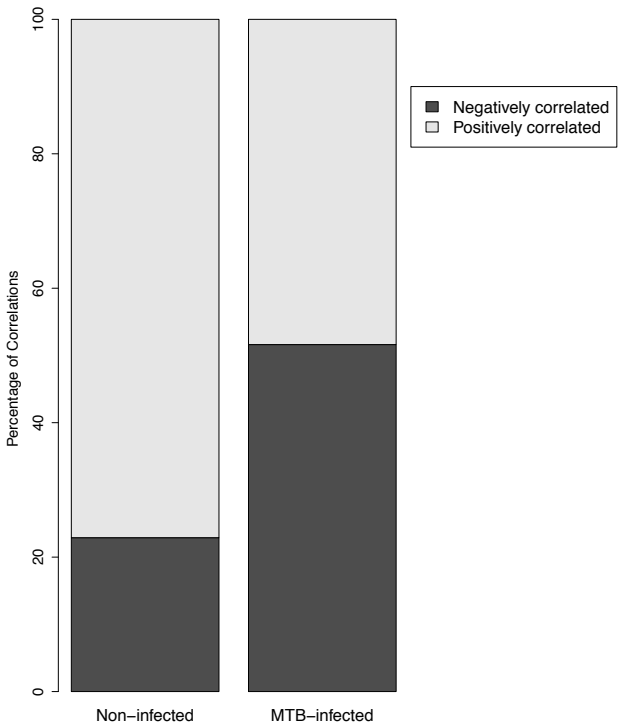

Figure S6. Barplots of significantly correlated miRNA-mRNA pairs. Barplots showing the proportions of negative and positive correlations among significantly correlated miRNA-mRNA pairs (FDR<0.005) in non-infected and infected samples.

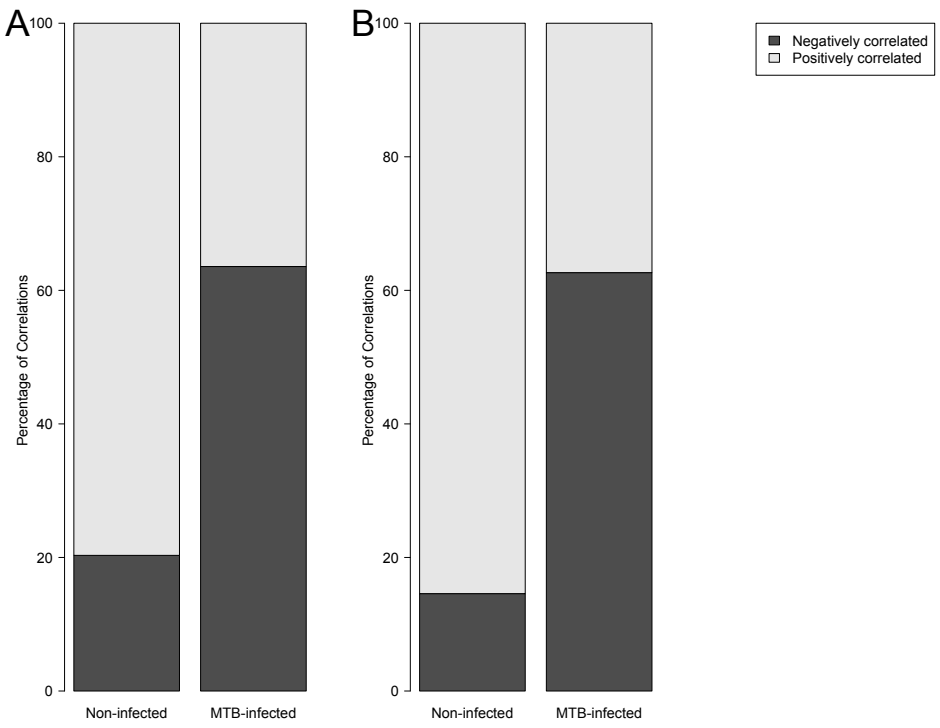

Figure S7. Variation in the percentage of infected cells and miRNA-mRNA correlations. Barplots of the genome-wide correlations between miRNAs and mRNAs (|r|>0.7) (*A*) before and (*B*) after correction for the percentage of infected cells. To assess the effect of inter-individual variability in the percentage of infected cells (ranging from 16-67%), we recalculated the miRNA-mRNA correlations for a subset of 47 samples for which information on the percentage of infected cells was quantified by FACS analysis using GFP-tagged bacteria. The effect of variation in the percentage of infected cells was corrected for using a linear model in non-infected (left panel) and infected (right panel) samples. In both analyses, at an |r|>0.7, we detected a greater number of significant correlations in non-infected samples, with respect to infected samples (not shown here). Moreover, among significantly correlated pairs, the majority of correlations in non-infected samples were positive, while after infection around 60% of miRNA-mRNA correlations were negative, consistent with the results obtained using the full data set. Furthermore, we observed a strong overlap in the significantly correlated miRNA-mRNA pairs across datasets (data not shown), suggesting that the percentage of infected cells does not have a strong influence on our results.

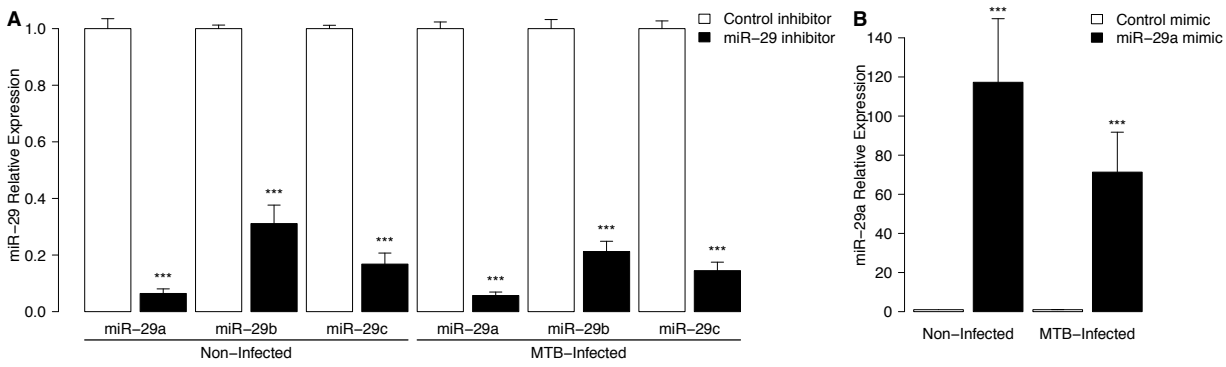

Figure S8. Perturbation of miR-29a expression using gain- and loss-of-function approaches. DCs were transfected with either (*A*) a miR-29a mimic or (*B*) a miR-29 inhibitor at day 5 and infected with MTB at day 6. Cells were lyzed and miR-29 levels quantified using qPCR, normalized on RNU6-1 levels. Fold expression was calculated with respect to miR-29 expression in control transfected DCs. The data represent the mean of duplicates of qPCR calculated across 4 different donors.

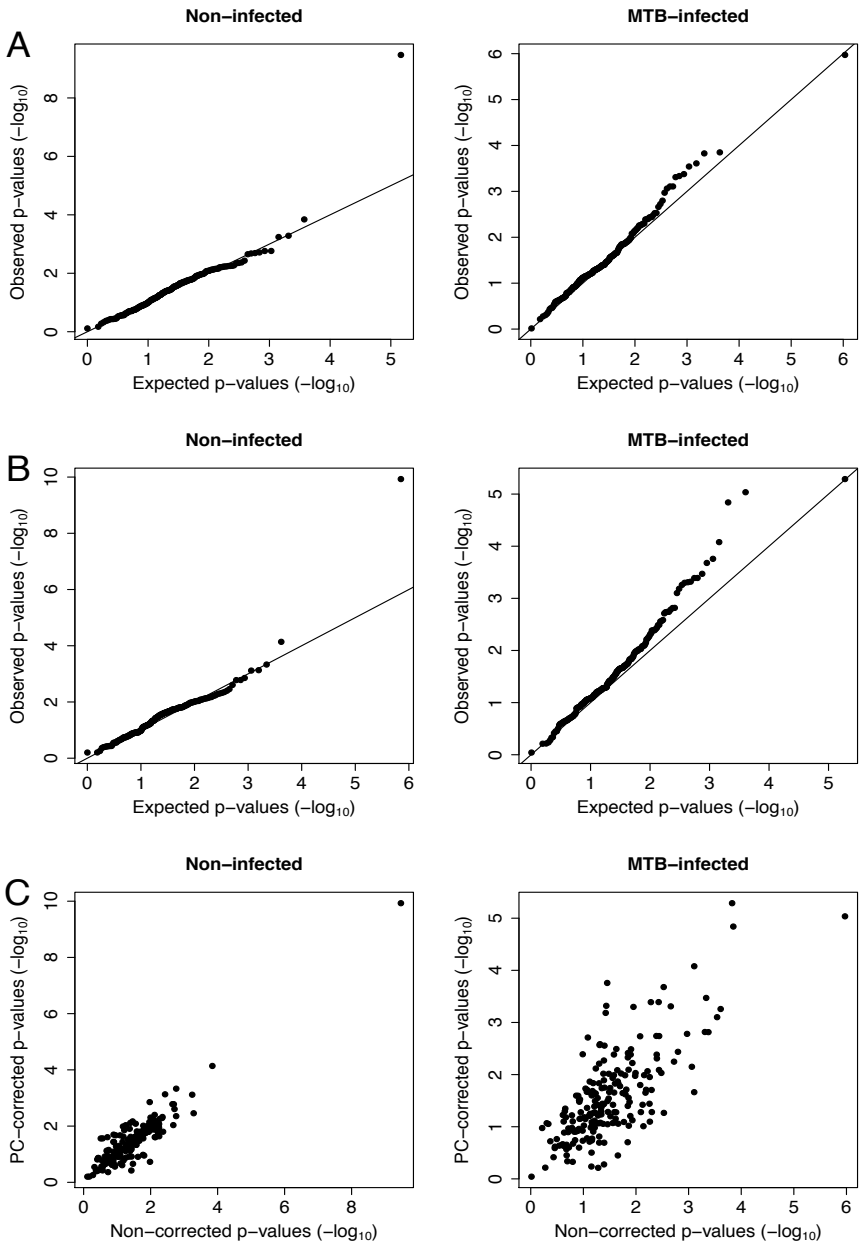

Figure S9. Principal component corrections for cis-eQTL detection. (*A*). QQ plots of minimum p-values for the association between miRNA expression and all SNPs within a 200kb window around the miRNA for non-infected (left panels) and infected (right panels) samples. Expected values were calculated based on 100 permutations of miRNA expression data. (*B*). QQ plots showing improved sensitivity of eQTL detection after accounting for unknown confounders using PC analysis. Specifically, the greatest number of significant associations was found after removing 1 and 4 PCs in non-infected and infected samples, respectively. (*C*). Plots showing the correlation between minimum p-values, before and after PC correction, indicating that this correction did not qualitatively change the relationships between genotypes and miRNA expression variation.

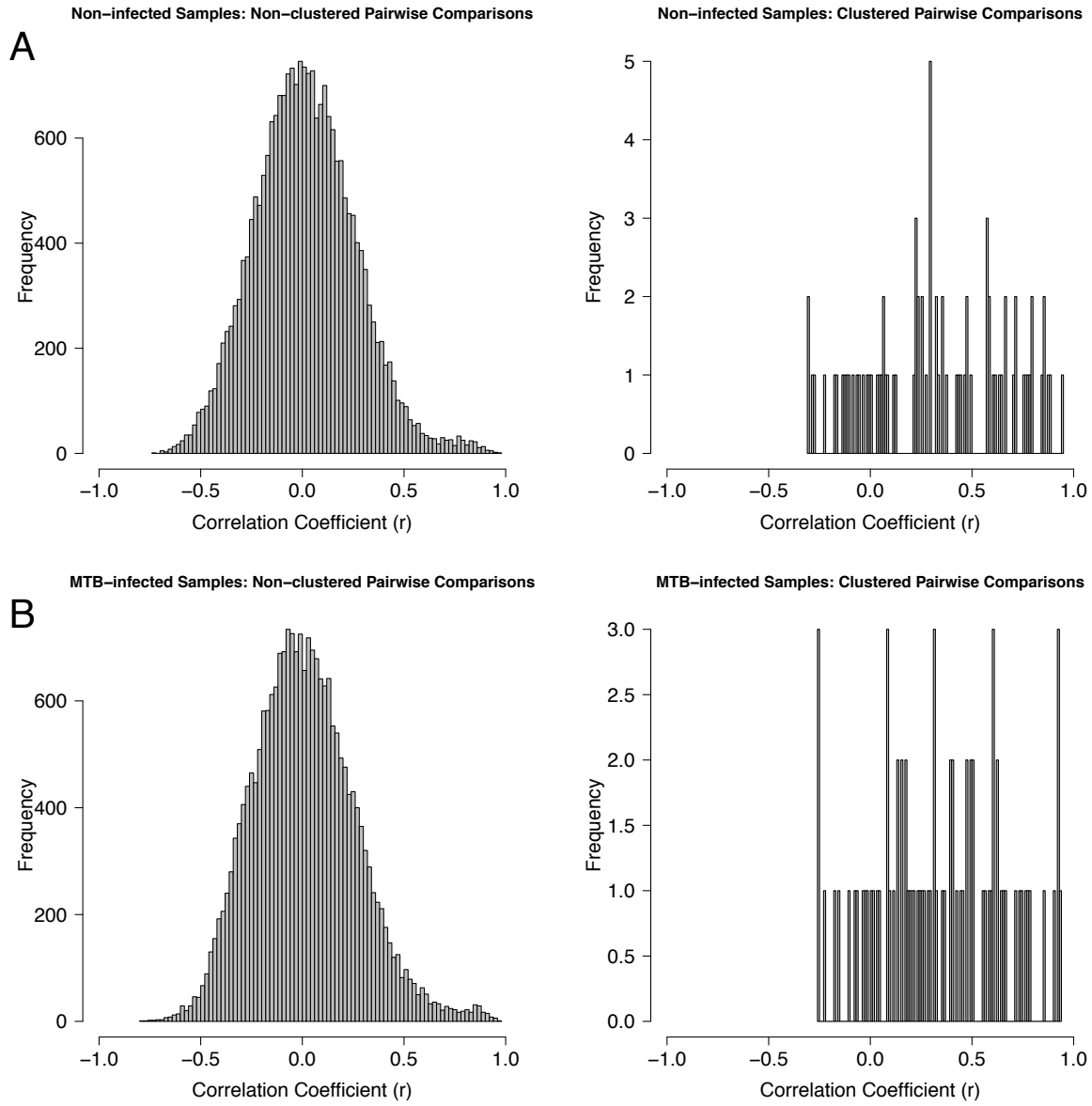

Figure S10. Distribution of correlation coefficients between miRNA pairs. We calculated the Pearson correlation coefficient pairwise among quantile normalized miRNA expression levels. Clustered miRNAs (right), those lying within 10kb of each other, were significantly enriched in positive correlations compared to the genome-wide distribution of correlations (left) in both (*A*) non-infected and (*B*) MTB-infected samples. This supports the hypothesis that independent mature miRNA sequences lying less than 10kb apart are frequently co-transcribed. This observation motivated our decision to consider only the most highly expressed transcript from each precursor in the analysis of miRNA-mRNA correlations, as this is the most likely to have an impact on mRNA expression levels.

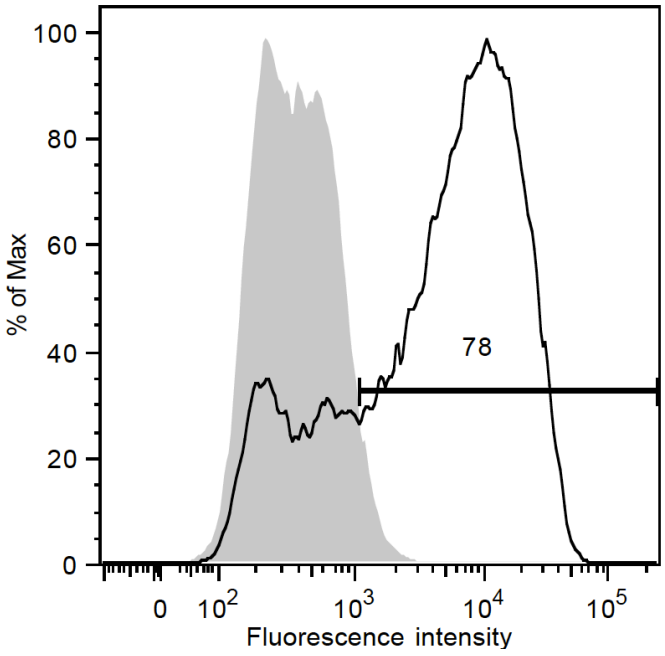

Figure S11. Transfection efficiency of DCs. DCs were transfected at day 5 using a fluorescently labeled negative control LNA Power inhibitor to evaluate the percentage of transfected cells (black). Fluorescence was compared to DCs transfected with an unlabeled control oligonucleotide (grey). Transfection efficiency was assessed by flow cytometry at day 6. Data shown are for one representative sample of four independent experiments performed on four different donors. The average transfection efficiency across all donors was 77%.

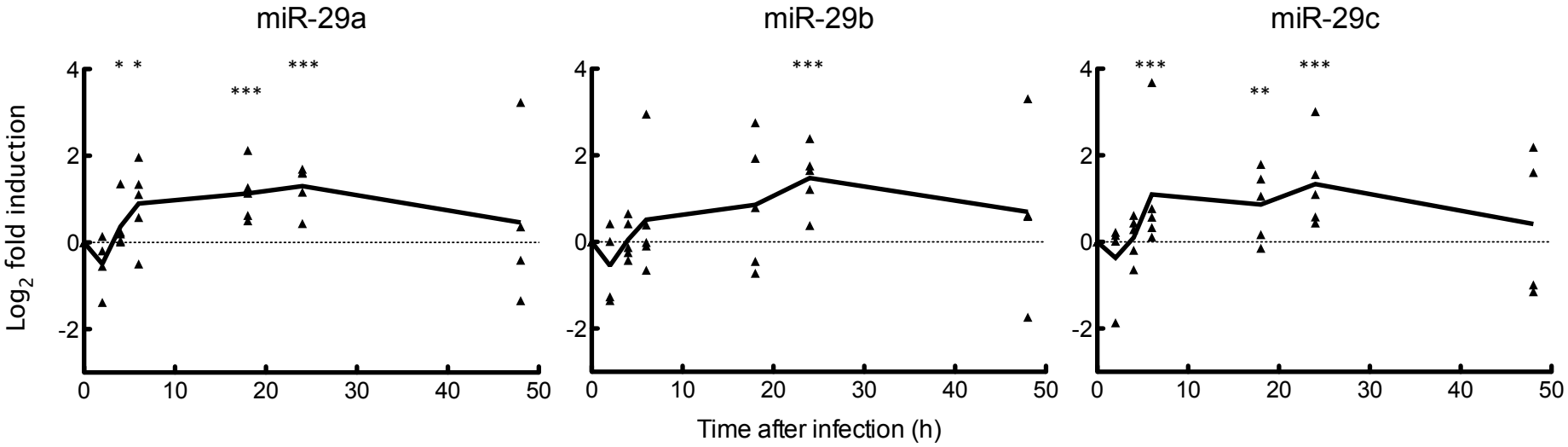

Figure S12. Time course of miR-29 family expression upon MTB infection in DCs. DCs were infected at day 6 with MTB (H37Rv) and lysed at six different time points (i.e., 2, 4, 6, 18, 24 and 48h) to quantify miR-29 expression levels. Expression levels of miR-29a, b and c were determined by qPCR, normalized on RNU6-1 levels, and fold induction was obtained by comparing to the non-infected condition. The data represent the mean of a duplicate qPCR calculated from at least 4 independent experiments, each derived from different donors. The significance of differences observed between MTB-infected and non-infected DCs was tested using a Mann-Whitney test (*$p<0.05$, **$p<0.01$, ***$p<0.001$).