Supplemental Tables

Dataset S1: Supporting tables

S1.1: List of *M. marinum* E11 genes without orthologues in *M. marinum* M. A gene is included if its annotation was not transferred based on synteny and has a blast hit below 90 % identity and 80% overlap compared to the proteome of *M. marinum* M. S1.2: Essential genes. Listed are genes that are most likely essential for survival of *M. marinum*, determined as described in the supplemental materials and methods. S1.3: Top 40 genes that show the largest average attenuation. The (log2-transformed) effect size is calculated relative to the pool of transposon mutants that were present prior to infection, as described in the experimental procedures. The list is arranged by genome number. Genes involved in ESX-1 secretion, cholesterol transport (MCE4) and PDIM biosynthesis are indicated. S1.4: Comparison of *M. marinum* and *M. tuberculosis* genes involved in cholesterol metabolism. Listed are *M. tuberculosis* genes that are predicted to be specifically required for growth on cholesterol according to J.E. Griffin, *et al*. (PLoS. Pathog. **7**:p e1002251.) and their *M. marinum* E11 orthologues with the effect size of those genes that show a significant effect in one or more host cells as determined by ANOVA multiple hypothesis testing ($p<0.05$). S1.5: Virulence effects. Listed are all CDS of the *M. marinum* E11 genome and plasmid and the effect of disruption by transposon insertion on survival in phagocytic cells derived from five different hosts. S1.6: Top 40 genes that show the most variable effects. Listed are genes that show the highest degree of variation in the different host cells, based on the standard deviation of the effect size. The (log2-transformed) effect size is calculated relative to the pool of transposon mutants that were present prior to infection, as

described in the experimental procedures. All genes show a significant difference by multiple hypothesis ANOVA testing with a p-value below 0.05. S1.7: Orthology relationships between *M. marinum* E11, M and *M. tuberculosis* H37Rv strains.

Table S1

Characteristics of individual transposon mutants used for competition experiments and the TraDIS data for the genes affected

Gene name	H37Rv orthologue	Tn insertion (from ATG)	gene length	Effect size (log2-transformed)				
				Ac	CLC	Dd	RAW	THP-1
cpsA	*Rv3484*	1443	1536	-0.42	-0.91	-0.36	-1.52	-1.85
PE_PGRS15_2	*Rv0872c*	108	1914	-0.14	-0.13	-0.14	-0.04	-0.11
eccCb1	*Rv3871*	239	1776	-0.40	-0.88	-0.68	-1.10	-1.06
ppm1A	*Rv2051c*	389	2115	0.36	-0.37	-0.43	1.22	0.74
				p-values				
				0.091	0.011	0.066	0.00019	0.000016
				0.26	0.21	0.26	0.40	0.23
				0.075	0.0024	0.0067	0.000061	0.000022
				0.24	0.29	0.17	0.0095	0.062

Supplemantal Figures

Figure S1. *M. marinum* infects phagocytic cells derived from different hosts.

A. Images of infected mouse (RAW264.7), human (THP-1) and fish (CLC) phagocytes as visualized by confocal fluorescence microscopy. Bacteria expressing GFP are green and the actin skeleton of the host cells is labeled red. **B**. *A. castellanii* infected by GFP-expressing *M. marinum* E11 wild-type or ESX-1-deficient *eccCb1* transposon mutant.

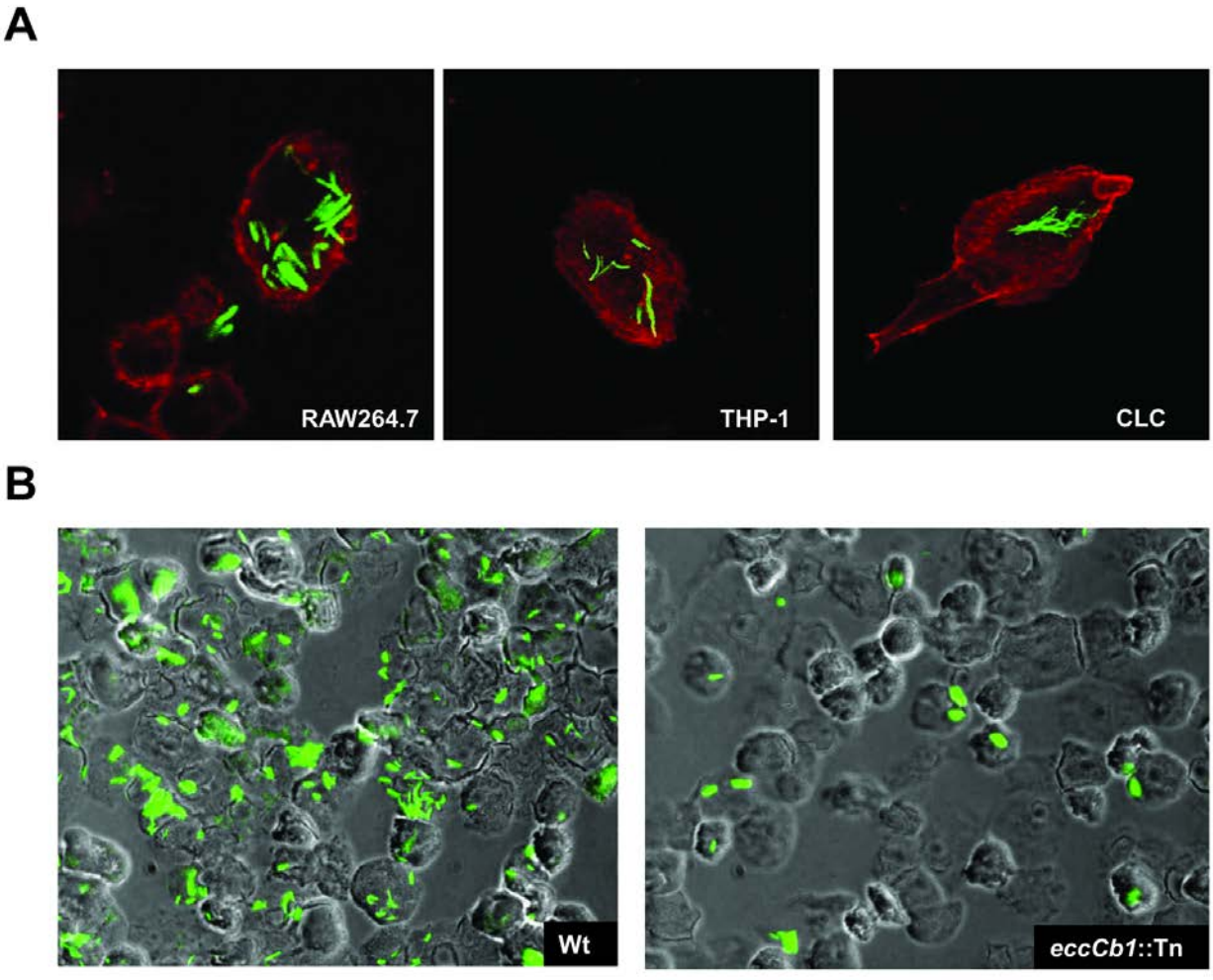


Figure S2. Average reads per TA per gene

For each gene in the genome, the read coverage of all TA sites of the input runs (sequenced DNA of transposon mutant pool prior to infection) were averaged and rounded, producing a number of average reads per TA per gene.

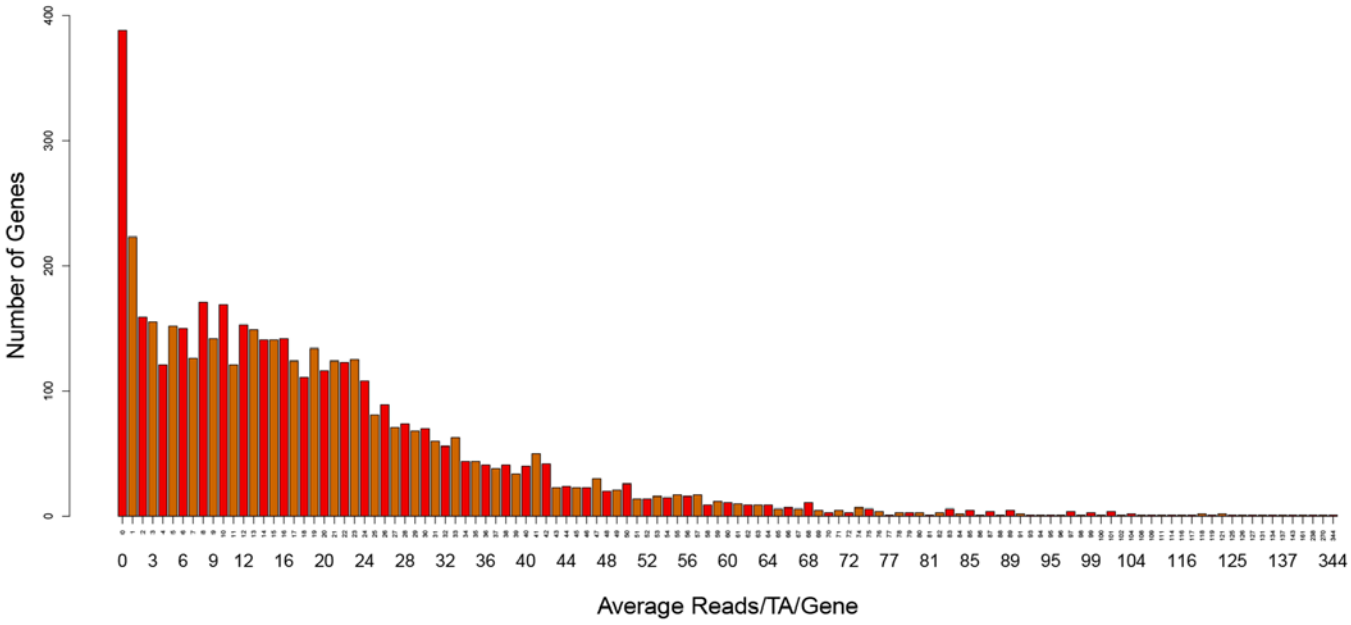


Figure S3. Essential genes within the *esx-3* cluster

Artemis view of the *esx-3* gene cluster (MMARE11_04890 to MMARE11_05080) showing all sequencing reads of the combined input pools. Each green line represents one sequenced transposon insertion at that particular TA site.

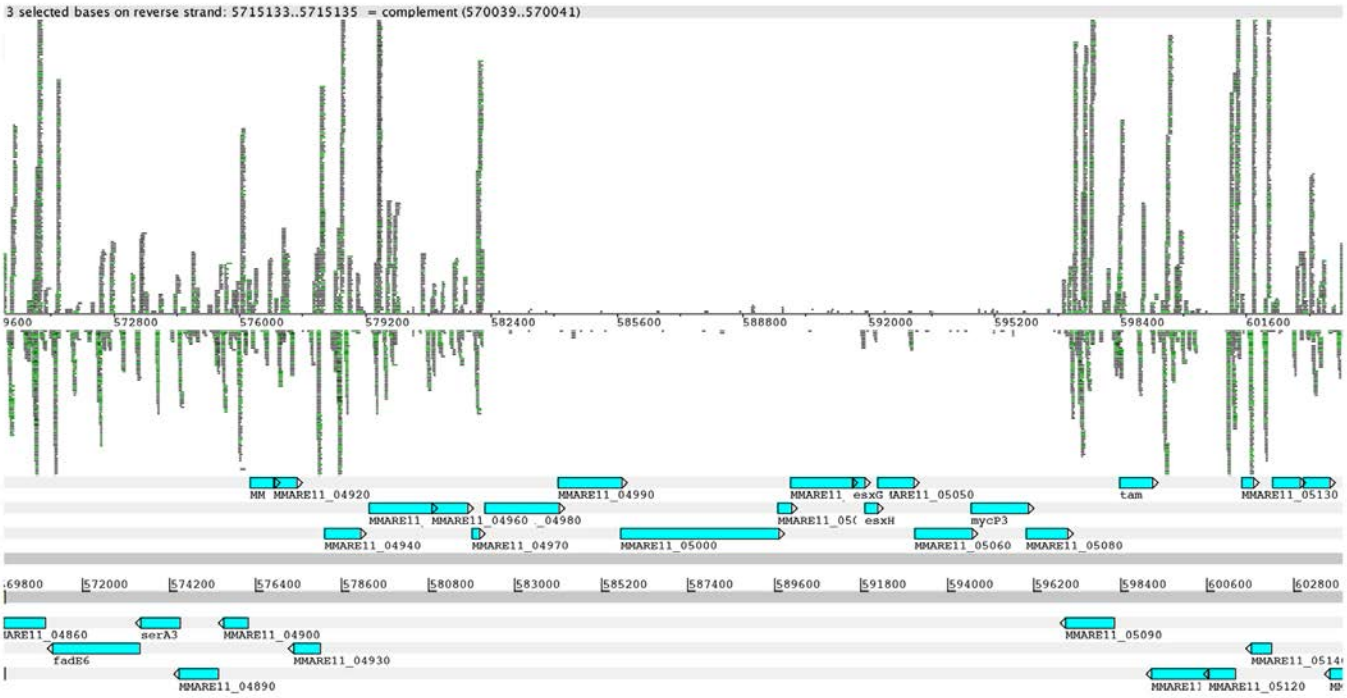


Figure S4. Growth of *M. marinum* in different host cells.

Host cells were infected with the pool of *M. marinum* transposon mutants and after 72 hours, cells were lysed and bacteria were plated in serial dilutions to determine CFU.

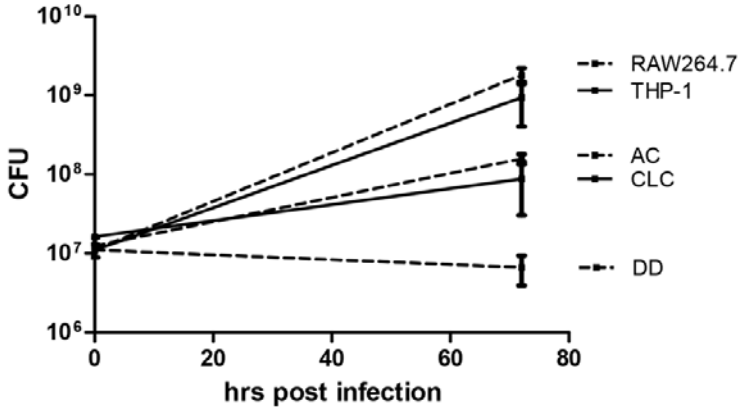


Figure S5. Read coverage correlation

For each TA site and each gene in the genome, the read coverage of all input runs (sequenced DNA of transposon mutant pool prior to infection) were averaged and rounded, producing average input read coverage. Similarly, all output runs (sequenced DNA of transposon mutant pool after infection of the different cells) were averaged and rounded, producing average output read coverage.

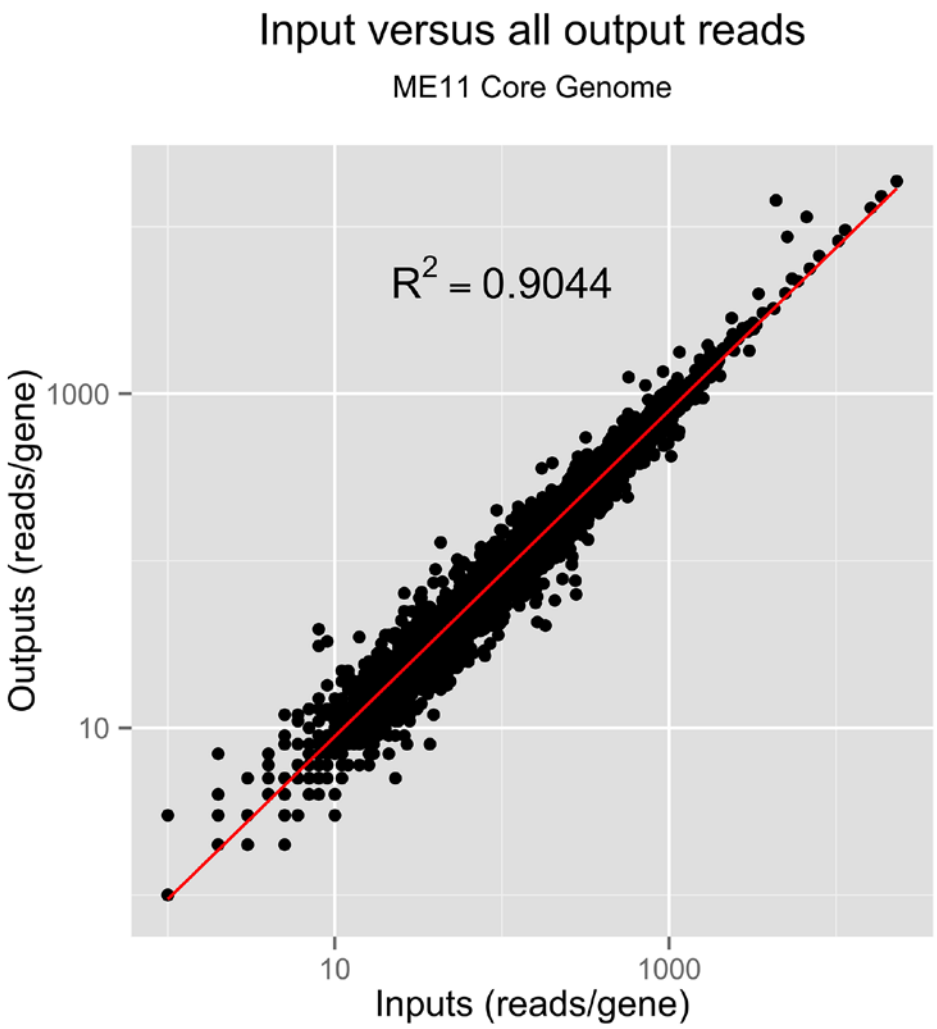


Figure S6. Reproducibility of individual experiments

Clustering of samples using the data from which the technical variance due to sequence primer identity and sequence pool were filtered by linear modelling, using the same linear model that was used to calculate the biological effects. Here only TA-sites for which the false discovery rate on the biological effect was ≤ 0.1 were used for clustering. In this way, only TA-sites in which a biological signal was detected participate in the sample. The fraction of such TA-sites was 19.1% (17733 out of 93043). The distance between samples was calculated using the euclidian method, and clustering was performed using the ward method. Only sample dd.1.1 seems to be an outlier, for unknown reasons.

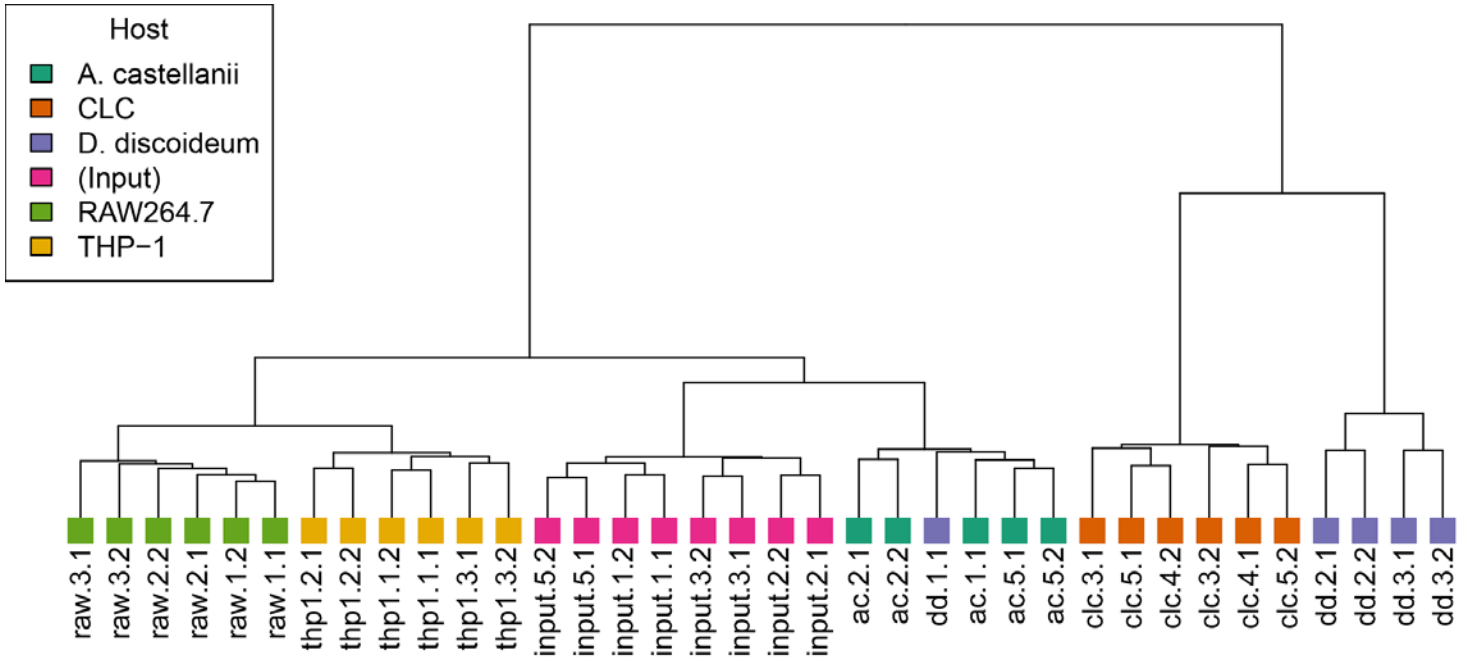


Figure S7. Effects of disturbing vitamin B12 biosynthesis in *A. castellanii*

KEGG pathway for vitamin B12 biosynthesis. Log2-transformed effect in *A. castellanii* of transposon insertions are indicated with red (growth advantage) and blue (growth disadvantage) colors.

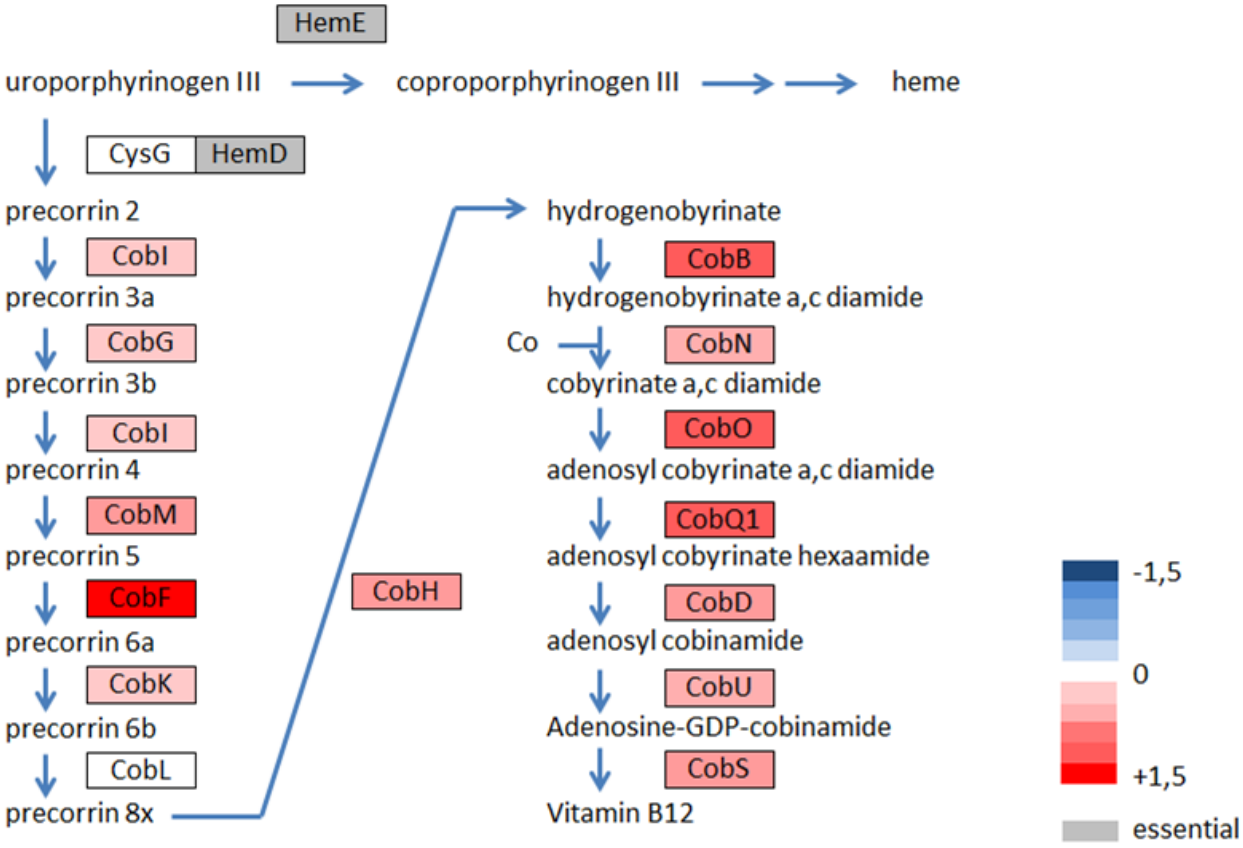
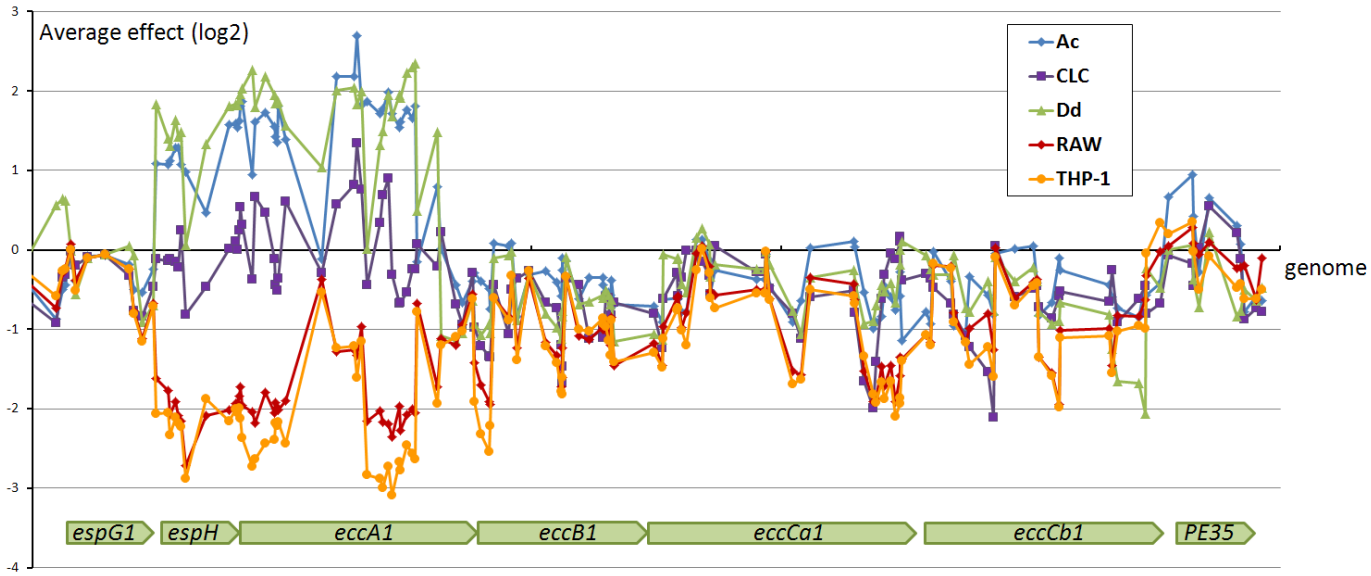


Figure S8. Effects in *esx-1* gene cluster

Effects of individual transposon insertions in several genes of the *esx-1* locus on infection of different host phagocytes. Each transposon site is indicated and the effect size is calculated as described in the supplemental materials and methods. Log2-transformed effects indicate growth advantage (>0) or disadvantage (<0) in respective host cell.



Supplemental materials and methods

Strains used for co-infection experiments

For validation experiments, single transposon mutants in *eccCb1* (MMARE11_52720) (E. J. M. Stoop**,** ***et al*****.** Dis Models Mech **4**:p 526-536, 2011), *espG5* (MMARE11_25960) (A.M. Abdallah ***et al*****.** Mol. Microbiol **73**:p 329-340, 2009), *ppm1A* (MMARE11_29520), *pe_pgrs15_2* (MMARE11_24200) and *cpsA* (MMARE11_47760) were used. Transposon mutants in *ppm1A* and *pe_pgrs15_2* were found by random mutant sequencing whereas the transposon mutant in *cpsA* was found in a pool of transposon mutants using a specific primer-mediated approach.

Fluorescence microscopy

For fluorescence microscopy of infected cells, cells were seeded on glass coverslips and infected with a GFP-expressing *M. marinum* E11 wild-type strain at an MOI of 5 for two hours. Infected cells were washed and incubated for 72 hours followed by fixation with 2% paraformaldehyde and 0.2% glutaraldehyde. Fixed infected THP-1, RAW264.7 and CLC cells were permeabilized with 0.1% Triton X-100 in PBS for 10 minutes, washed with PBS and incubated in 1% BSA in PBS for 30 minutes. After washing in PBS, cells were incubated for 45 minutes with 150 nM Alexa Fluor phalloidin (Molecular Probes) and 1% BSA in PBS to stain F-actin. After washing with PBS, cells were preserved in VectaShield mounting medium with DAPI (Vector Labs). Samples were examined using a Leica SP2 confocal microscope (Leica Microsystems). Fixed infected *A. castellani* and *D. discoideum* cells were analyzed using a Zeiss Axiovert 200M inverted microscope (Zeiss).

Illumina assembly and annotation process

For the assembly of the *M. marinum* E11 genome and the plasmid, three different libraries were used: an Illumina paired end library, a PCR-free library with a fragment size of 550bp and a large insert Nextera mate-pair library (3kb fragment size). First the short reads were assembled with velvet (D.R. Zerbino and E. Birney, Genome Res. **18**:p 821-829, 2008). The assembly was iteratively improved by scaffolding it with SSPACE (M. Boetzer *et al*., Bioinformatics. **27**:p 578-579, 2011) using the 3kb library, closing the sequencing gaps with gapfiller (M. Boetzer and W. Pirovano, Genome Biol. **13**:p R56, 2012) and IMAGE (I.J. Tsai, T. D. Otto and M. Berriman**,** Genome Biol. **11**:p R41, 2010), ordering the scaffolds against the M strain reference (ABACAS (S. Assefa *et al*. Bioinformatics. **25**:p 1968-1969, 2009)) and using REAPR (M. Hunt *et al***.** Genome Biol. **14**:p R47, 2013) and ICORN (T. D. Otto *et al*., Bioinformatics. **26**:p 1704-1707, 2010) to correct the assembly. Most of the assembly improvements are explained in further depth in (M. T. Swain *et al*., Nat. Protoc. **7**:p 1260-1284, 2012). We further manually improved the assembly in repetitive regions, such as the duplicated rRNA or non-ribosomal peptide synthases genes, by using the *M. marinum* M strain sequence of the orthologous genes. The assembly is in two sequences (chromosome plus plasmid) and has 40 sequencing gaps. For the function annotation we transferred the annotation from the *M. marinum* M strain (RATT (T. D. Otto *et al*., Nucleic Acids Res. **39**:p e57, 2011)) and used PROKKA (http://bioinformatics.net.au/prokka-manual.html) for *ab initio* annotation. A bespoke PERL script merged both annotations choosing the RATT transferred gene models if in a region the PROKKA gene model overlapped the RATT model. The core genome has 5335 genes and the plasmid 98.

TraDIS library preparation and sequencing process

Briefly, two micrograms of genomic DNA was sheared to an average size of 300 bp. DNA was purified using QiaQuick PCR purification kit (Qiagen) according to the manufacturer's recommendations, and subsequently Illumina DNA fragment library preparation was performed using NEBNext® DNA Library Prep Reagent Set for Illumina® (New England BioLabs Inc) following the manufacturer's instructions. Ligated fragments were run in 2% agarose gel and fragments corresponding to an insert size of 250–350 bp were excised. DNA was extracted from the gel slice using QiaQuick gel extraction kit (Qiagen). To amplify the transposon insertion sites, 22 cycles of PCR were performed using a transposon-specific forward primer and a custom Illumina reverse primer. Amplified libraries were finally purified with AMPure beads (Beckman Coulter) as per the manufacturer's instructions. A small aliquot (2 µl) was analyzed on Invitrogen Qubit and Agilent Bioanalyzer DNA1000 chip, following the manufacturer's instructions. The amplified DNA fragment libraries were sequenced on single end Illumina flow cells using an Illumina Genome Analyzer IIx sequencer for 105 cycles of sequencing, using a custom sequencing primer and 2× hybridization buffer. This primer was designed such that the first 10 bp of each read was transposon sequence. Data were processed with the Illumina Pipeline Software v1.82.

Analysis of nucleotide sequence data

Sequence reads from the Illumina FASTQ files were first filtered, taking only the reads which sequences start with the specific TraDIS primer sequence. The filtered reads were mapped to the E11 genome using SMALT, producing a BAM file for each TraDIS run. Samtools's mpileup function was then performed on each of the BAM files. The resulting

pileup format enabled us to computationally survey mapped read bases at each genomic position, and subsequently measure the read coverage. OrthoMCL was then ran on the proteomes of the genomes of *M. marinum* E11, *M. tuberculosis* and *M. marinum* M. We then ran a set of in-house scripts that took as input the E11 genome with its annotations, pileup file of TraDIS reads mapped to the E11 genome and OrthoMCL output file, producing a document containing exact positions of all TA sites and the following information tied to each site: which CDS each TA site falls within (if any), the corresponding gene name and product, orthology relation to *M. tuberculosis* H37Rv and *M. marinum* M strains and the read coverage of the TA site in each of our 36 runs. A file with orthologous relationships between *M. marinum* E11, M and *M. tuberculosis* H37Rv genes is available in Dataset S1.7.

Statistical analysis of TraDIS read numbers

The number of hits (sequences counted) per TA-site was normalized to a hit rate per million sequences by dividing the sequence counts for a TA-site by the total number of hits in each sample run and multiplying by 10^6. To be able to apply least-squares fitting and ANOVA, we transformed the hit rate data using a number of variance stabilizing transforms as proposed by N. Anscombe (Biometrika **35**:p 246-254, 1948). The transform proposed for negative-binomial distributed data, defined as

tails than normal distributed data, but they were symmetric. Untransformed hit rates, as well as the transform proposed for Poisson-distributed data or a logarithmic transform were less suitable, because the resulting distributions of residuals deviated strongly from a normal distribution.

The simplest linear model fitting the transform hit rate data was a model with terms accounting for the biological effect (5 host organisms and the ``input'' samples), and for technical effects, being the sequencing primer used (two different primers), and the experiment batch (experiments were performed in 5 batches). There was no evidence for interaction effects between these variables. The largest technical effects were caused by the sequencing primer used. Since we observed in a number of genes that different TA-sites on the gene sometimes displayed opposite biological effects, we modeled the TA-sites individually, and calculated overall effects per gene by taking the median of effect sizes over all TA-sites covered by the gene. These are also the biological effects displayed in the figures. To obtain an idea of the significance of the effects per gene, we calculated the geometric mean over all TA-sites covered by the gene of the p-values resulting from an ANOVA on the model terms. Clearly, these means themselves cannot be interpreted as p-values.

A measure of the essentiality of a gene was calculated as described by J. E. Griffin *et al*., (PLoS. Pathog. **7**:p e1002251, 2011), by comparing expected and observed maximal lengths of consecutive stretches of TA-sites without a hit (L. Gordon, M. F. Schilling and M. S. Waterman, Probab Th Rel Fields **72**:p 279-287, 1986). All analyses were performed in R (R Core Team, Foundation for Statistical Computing, Vienna, Austria, 2013).