Supplementary Material for "Statistical Analysis of Genetic Interactions in TnSeq data"

Michael A. DeJesus
Subhalaxmi Nambi
Clare M. Smith
Richard E. Baker
Christopher M. Sassetti
Thomas R. Ioerger

February 10, 2017

Contents

1 Derivation of Conditional Probability Distributions

To calculate the posterior distribution over the mean count for a given gene, strain, and condition, we combine the Normal likelihood function with conjugate prior distributions for the mean and variance.Specifically, the prior on the mean is chosen to be Normal with hyperparameters μ_0 and κ_0, and the prior on the variance is chosen to be Inverse-Gamma (IG) with hyperparameters ν_0 and σ_0^2. Taken together, the joint distribution of the data, and unknown parameters is:

$$
\begin{aligned}
p(\mu_g^{ij}, {\sigma_g^{ij}}^2, Y_g^{ij}) &= p(Y_g^{ij}|\mu_g^{ij}, {\sigma_g^{ij}}^2)\ p(\mu_g^{ij}|{\sigma_g^{ij}}^2, \theta)\ p({\sigma_g^{ij}}^2|\theta) \\
&= \text{Normal}(Y_g^{ij}|\mu_g^{ij}, {\sigma_g^{ij}}^2) \times \text{Normal}(\mu_g^{ij}|{\sigma_g^{ij}}^2, \theta) \times \text{IG}({\sigma_g^{ij}}^2|\theta) \qquad (1) \\
&= \prod_{y \in Y_g^{ij}} \frac{1}{\sqrt{2\pi{\sigma_g^{ij}}^2}} \exp \frac{(y-\mu_g^{ij})^2}{2{\sigma_g^{ij}}^2} \times \frac{1}{\sqrt{2\pi{\sigma_g^{ij}}^2}} \exp \frac{(\mu_g^{ij}-\mu_0)^2}{2\tau_0^2} \\
&\quad \times \frac{(\nu_0\sigma_0^2/2)^{\nu_0/2}}{\Gamma(\nu_0/2)} ({\sigma_g^{ij}}^2)^{-\nu_0/2-1} \exp -\frac{\nu_0\sigma_0^2/2}{1/{\sigma_g^{ij}}^2} \\
&\propto ({\sigma_g^{ij}}^2)^{-n/2} \exp -\frac{1}{2{\sigma_g^{ij}}^2}\left[(n-1)s^2 + n(\bar{Y}_g^{ij} - \mu_g^{ij})^2\right] \\
&\quad \times {\sigma_g^{ij}}^{-1} ({\sigma_g^{ij}}^2)^{-(\nu_0/2+1)} \exp -\frac{1}{2{\sigma_g^{ij}}^2}\left[\nu_0\sigma_0^2 + \kappa_0(\mu_g^{ij}-\mu_0)^2\right]
\end{aligned}
$$

where θ represents the hyperparameters (i.e. $\theta = \{\mu_0, \kappa_0, \nu_0, \sigma_0^2\}$), and $\bar{Y}_g^{ij}$ and s^2 represent the sample mean and sample variance respectively. The posterior for ${\sigma_g^{ij}}^2$ is derived as follows:

$$
\begin{aligned}
p({\sigma_g^{ij}}^2|Y_g^{ij}) &\propto p(Y_g^{ij}|{\sigma_g^{ij}}^2) \times p({\sigma_g^{ij}}^2) \\
&\propto \left[\int_{-\infty}^{\infty} p(Y_g^{ij}|\mu_g^{ij}, {\sigma_g^{ij}}^2) p(\mu_g^{ij}|{\sigma_g^{ij}}^2) d\mu_g^{ij}\right] \times p({\sigma_g^{ij}}^2) \\
&\propto \int_{-\infty}^{\infty} ({\sigma_g^{ij}}^2)^{-n/2} \exp -\frac{1}{2{\sigma_g^{ij}}^2}\left[(n-1)s^2 + n(\bar{y} - \mu_g^{ij})^2\right] \\
&\quad \times {\sigma_g^{ij}}^{-1} ({\sigma_g^{ij}}^2)^{-(\nu_0/2+1)} \exp -\frac{1}{2{\sigma_g^{ij}}^2}\left[\nu_0\sigma_0^2 + \kappa_0(\mu_g^{ij}-\mu_0)^2\right]\ d\mu \\
&\propto \frac{(\nu_n\sigma_n^2/2)^{\nu_n/2}}{\Gamma(\nu_n/2)} ({\sigma_g^{ij}}^2)^{-\nu_n/2-1} \exp -\frac{\nu_n\sigma_n^2/2}{1/{\sigma_g^{ij}}^2} \\
&\propto \text{Inverse-Gamma}({\sigma_g^{ij}}^2 \mid \nu_n/2,\ \nu_n\sigma_n^2/2) \qquad (2)
\end{aligned}
$$

where the values of ν_n and σ_n^2 are defined as:

$$
\begin{aligned}
\nu_n &= \nu_0 + n \\
\sigma_n^2 &= \frac{1}{\nu_n}\left[\nu_0\sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_n}(\bar{Y}_g^{ij} - \mu_0)^2\right]
\end{aligned}
$$

Given the value of ${\sigma_g^{ij}}^2$, the posterior distribution of the mean is therefore:

$$
p(\mu_g^{ij}|Y_g^{ij}, {\sigma_g^{ij}}^2) \propto p(Y_g^{ij}|\mu_g^{ij}) \times p(\mu_g^{ij}|{\sigma_g^{ij}}^2)
$$

$$\propto \prod_{y \in Y_g^{ij}} \frac{1}{\sqrt{2\pi\sigma^2}} \exp \frac{(y-\mu_g^{ij})^2}{2\sigma_g^{ij^2}} \times \frac{1}{\sqrt{2\pi\sigma_g^{ij^2}}} \exp \frac{(\mu_g^{ij}-\mu_0)^2}{2\tau_0^2}$$

$$\propto \frac{1}{\sqrt{2\pi\sigma_g^{ij^2}}} \exp \frac{(\mu_g^{ij}-\mu_n)^2}{2\sigma_g^{ij^2}}$$

$$\propto \text{Normal}(\mu_g^{ij} \mid \mu_n, \sigma_g^{ij^2}/\kappa_n) \tag{3}$$

where the values of μ_n and κ_n are defined as:

$$\kappa_n = \kappa_0 + n \qquad\qquad \mu_n = \frac{\kappa_0\mu_0 + n\bar{Y}_g^{ij}}{\kappa_n}$$

Note that the posterior parameter for the mean, μ_n, is combination of the sample mean, $\bar{Y}_g^{ij}$ (average insertion count over TA sites in gene), and the prior, μ_0, weighted by the amount of data. This can be thought of as having a similar effect as pseudo-counts, as it provides additional source of information (based on prior knowledge) when there is a small amount of data, or a "smoothing" effect, when there are a few counts that are significantly different than the rest.

2 Experimental Methods

2.1 Strains and plasmids used in the study

M. tuberculosis H37Rv was grown at 37 C in 7H9 and 7H10 media (BD Biosciences) supplemented with 10% OADC (oleic acid, albumin, dextrose, catalase) enrichment plus hygromycin (50 μg/mL) and kanamycin (25 μg/mL) as needed. For generation of mutant strains of *M. tuberculosis*, a double stranded PCR fragment was constructed containing 500-bp upstream region, 1.2 kb hygromycin and 500 bp downstream region of Δ*rv2680*, Δ*rv1432* and Δ*rv1565c* respectively. The PCR product was then electroporated in *M. tuberculosis* recombineering strain. Briefly, the *M. tuberculosis* H37Rv recombineering strain bearing plasmid pNIT:ET [2] was induced for 18 h with 1 μM isovaleronitrile. The culture was treated with 0.2 M glycine for 8 h before making electrocompetent cells and transformed. After selection on 7H10 plates containing hygromycin (50 μg/ml), the presence of the allelic exchange mutation was assayed by PCR amplification of the hyg^{R} cassette flanked by the N- and C-terminal junctions. All primers used are listed in Table S3.

2.2 Transposon library generation

Transposon libraries were made using the pMycoMarT7 transposon as described [5]. Briefly, roller bottles containing 100 mL of 7H9-OADC-Tw80 broth media were inoculated with wild type or the mutant strains. Cells were grown at 37°C until they reached an OD of $\sim$ 1.0. Cells were collected by centrifugation and washed two times with PBS (Sigma) to remove any residual tween-80. Cells were then resuspended in 9 mL of MP buffer and then transduced with 5.0×10^{10} to 1.0×10^{11} pfu/mL pMycoMarT7 phage. To stop the transduction, the resulting mixture was centrifuged and washed with PBS (Sigma) containing 0.05% tween and plated onto 7H10 plates containing OADC, 0.05% tween, and 20 μg/mL kanamycin. Each library consisted of approximately 1×10^5 independent mutants.

2.3 Mouse infections and DNA sample preparation

To initiate the genetic interaction screen, transposon libraries of wild type, Δrv2680, Δrv1432, and Δrv1565c strains were introduced intravenously by tail vein injection into eight groups of three C57BL/6 mice (8-10

weeks old). Housing and experimentation were in accordance with the guidelines set forth by the Department of Animal Medicine of University of Massachusetts Medical School and Institutional Animal Care and Use Committee. Two groups of three received approximately 10^6 colony-forming units (cfu) of the same library. At days 0 and 32, the mice were sacrificed and the surviving mutants were isolated by plating the spleen homogenates on 7H10 plates containing OADC, tween-80 and 20 μg/mL kanamycin. The wild type and mutant libraries, isolated from mice, were collected separately by scraping the plates and then extracting the chromosomal DNA. Each genomic DNA library was sheared by sonication with a Covaris S220, creating DNA fragments approximately 500 bp in length. Damaged ends were repaired (Epicentre DNA End RepairKit), and A-tailed with DNA Taq polymerase (Denville) to allow ligation of barcoded adapters with T-overhangs. The transposon chromosomal junctions were amplified for 20 cycles (95°C, 30 sec; 58°C, 30 sec; 72°C, 45 sec) using a transposon specific primer (T7-1) and an adapter specific primer (JEL-AP1). The PCR product for each library was run on a 2% gel at 50V. Genomic DNA located between 300 and 500 bp was then extracted from the gel. Following gel-extraction, the PCR fragments were further amplified for 10 cycles (95°C, 30 sec; 58°C, 30 sec; 72°C, 45 sec) with primers that were compatible with Illumina sequencing (Table S4). Two to three replicates (from different groups of mice) were sequenced for each library at each time point.

2.4 Sequencing of TnSeq libraries

The TnSeq samples were sequenced on an Illumina HiSeq 2500, operated in paired-end mode, with a read-length of 125+125 bp. The sequencing data was processed using TPP in TRANSIT [1]. Reads were mapped to TA sites in the H37Rv genome, and the raw read counts were then reduced to unique template counts, based on barcodes attached during sample preparation [3]. Sequencing statistics are shown in Table S1.

2.5 Affinity Pull-down Assay

Flag-tagged version of Rv1432 was expressed ectopically in Δ*leuD* Δ*panCD* double auxotroph Mtb and used for immunoprecipitation to characterize interacting proteins, as described in [4]. Lysates from expression strain were extracted with 2% octyl glucoside, Rv1432 was immunoprecipitated using anti-Flag M2 affinity gel, and bound proteins were then trypsinized and subjected to mass spectrometric analysis (Proteomics Facility, University of Massachusetts Medical school). The following peptides mapping to Rv1431 were observed. None of these were observed in other controls, indicating a unique affinity of Rv1431 and Rv1432 and not promiscuous binding.

Mtb protein names	Gene ID	Spectral counts	Peptides Identified
Rv1431	gi|15608569	7	GFLKPDLPDVDHDTWLTQ**P** (R) (K) ISFLPA**M** (R) (R) GDAEATMETSV**V** (K) (K) LYDGATAEIMT**D** (K) (R) CNFADGD**L** (R) (K) TGLFEAGYVTVEDML**S** (R)

3 Table S1: Sequencing Statistics

Table S1: Sequencing Statistics, including total reads and template counts for each sample sequenced. "Density" is the fraction of 74,603 TA sites represented ("TAs hit") in each sample. "NZmean" is the average template count at TA sites with at least 1 insertion.

Sample	Replicate	Total Reads	Template Count	TA sites hit	Density	NZmean
Rv day 0	1	2116076	1174196	32675	0.438	35.9
Rv day 0	2	3252442	1703655	33620	0.451	50.7
Rv day 32	1	1372643	752715	17101	0.229	44.0
Rv day 32	2	11597363	3970806	22224	0.298	178.7
Rv day 32	3	6218458	3017972	24798	0.332	121.7
Rv2680 day 0	1	4927079	2618295	30847	0.413	84.9
Rv2680 day 0	2	5171981	2808698	30938	0.415	90.8
Rv2680 day 32	1	5401019	2598055	20375	0.273	127.5
Rv2680 day 32	2	1748576	894862	20677	0.277	43.3
Rv2680 day 32	3	3930245	1800913	25331	0.340	71.1
Rv1565c day 0	1	4798074	2470014	24611	0.330	100.4
Rv1565c day 0	2	1615187	841104	21006	0.282	40.0
Rv1565c day 32	1	1839034	881357	19830	0.266	44.4
Rv1565c day 32	2	2851939	1365958	20542	0.275	66.5
Rv1565c day 32	3	1894375	1007683	20496	0.275	49.2
Rv1432 day 0	1	4353904	2690648	22270	0.299	120.8
Rv1432 day 0	2	3816435	2239185	21329	0.286	105.0
Rv1432 day 32	1	9074011	4906895	24438	0.328	200.8
Rv1432 day 32	2	4126462	2358200	20604	0.276	114.5
Rv1432 day 32	3	7686187	4529036	20724	0.278	218.5

4 Table S2: Results of the statistical analysis for genetic interactions

Results of the statistical analysis for genetic interactions, listing ΔlogFC and statistical significance for each gene. There is a separate sheet of results for each of the knockout strains. Spreadsheet is sorted so that significant interactions are at the top. The "Type of Interaction" column is defined as follows:

$$\text{Type} = \begin{cases} \text{"Aggravating"} & \text{if } \Delta\text{logFC} < 0 \\ \text{"Alleviating} & \text{if } \Delta\text{logFC} > 0 \text{ and } \mid \text{logFC}^{KO} \mid < \mid \text{logFC}^{WT} \mid \\ \text{"Suppressive"} & \text{if } \Delta\text{logFC} > 0 \text{ and } \mid \text{logFC}^{KO} \mid > \mid \text{logFC}^{WT} \mid \end{cases}$$

5 Table S3: Genes with significant change in ΔlogFC in two or more of the knockouts

Table of genes with significant change in $\Delta \log \text{FC}$ in two or more of the genes studied.

6 Table S4: Primers used in this study

Table of genes primers used in the experimental methods of this study.

References

[1] M.A. DeJesus and T.R. Ioerger. Reducing type i errors in tn-seq experiments by correcting the skew in read count distributions. In *7th International Conference on Bioinformatics and Computational Biology (BICoB 2015)*, 2015.

[2] J. E. Griffin, A. K. Pandey, S. A. Gilmore, V. Mizrahi, J. D. McKinney, C. R. Bertozzi, and C. M. Sassetti. Cholesterol catabolism by Mycobacterium tuberculosis requires transcriptional and metabolic adaptations. *Chem. Biol.*, 19(2):218–227, Feb 2012.

[3] J.E. Long, M. DeJesus, D. Ward, R.E. Baker, T.R. Ioerger, and C.M. Sassetti. Identifying essential genes in *Mycobacterium tuberculosis* by global phenotypic profiling. In Long Jason Lu, editor, *Methods in Molecular Biology: Gene Essentiality*, volume 1279. Springer, 2015.

[4] S. Nambi, J. E. Long, B. B. Mishra, R. Baker, K. C. Murphy, A. J. Olive, H. P. Nguyen, S. A. Shaffer, and C. M. Sassetti. The Oxidative Stress Network of Mycobacterium tuberculosis Reveals Coordination between Radical Detoxification Systems. *Cell Host Microbe*, 17(6):829–837, Jun 2015.

[5] Christopher M. Sassetti, Dana H. Boyd, and Eric J. Rubin. Comprehensive identification of conditionally essential genes in mycobacteria. *PNAS*, 98(22):12712–12717, 2001.