Identifying essential genes in *M. tuberculosis* by random transposon mutagenesis

Karl W Broman

Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

www.biostat.jhsph.edu/~kbroman

Joint work with Natalie Blades, Gyanu Lamichhane, and William Bishai

Typical drug regimens

Tuberculosis

- INH 15g
- RIF 37g
- PZA 141g
- ETB 151g

- ~60 clinic visits

- Cost: > $15,000

Other bacterial pneumonias

- Azithromycin 1.5g

- Self-supervised

- Cost: $35

Mycobacterium tuberculosis

- 4.4 Mbp circular genome, completely sequenced
- 4250 known or inferred genes
- 44% of genome has no match to mammals or other bacteria
- $>$250 lipid biosynthesis genes (E. Coli: $\sim$50)
- Mycolic acids: unique, essential
- Cell division time: 24 hr

Bacterial gene products

Essential genes

- Cell division
- DNA replication
- Transcription
- Protein synthesis
- Cell wall formation

Non-essential genes

- Virulence
- Stress response
- DNA modification
- Mobile elements
- Small molecular biosynthesis
- Regulatory genes

Aim

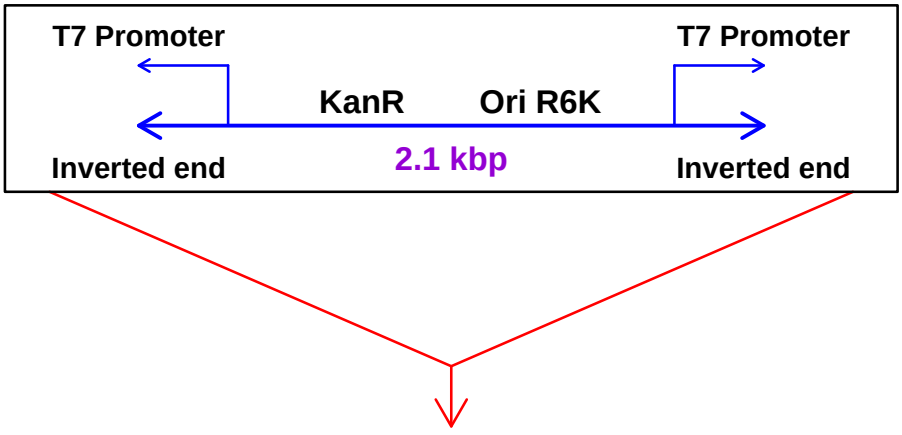
Identify the essential genes

(knock-out $\Longrightarrow$ non-viable mutant)

Method

Random transposon mutagenesis

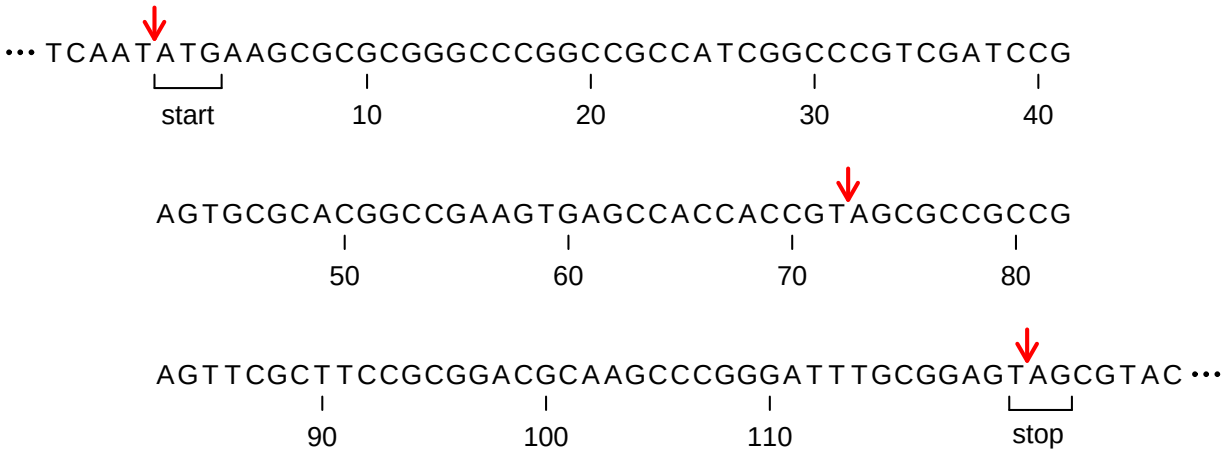
Himar1, a mariner-derived transposon



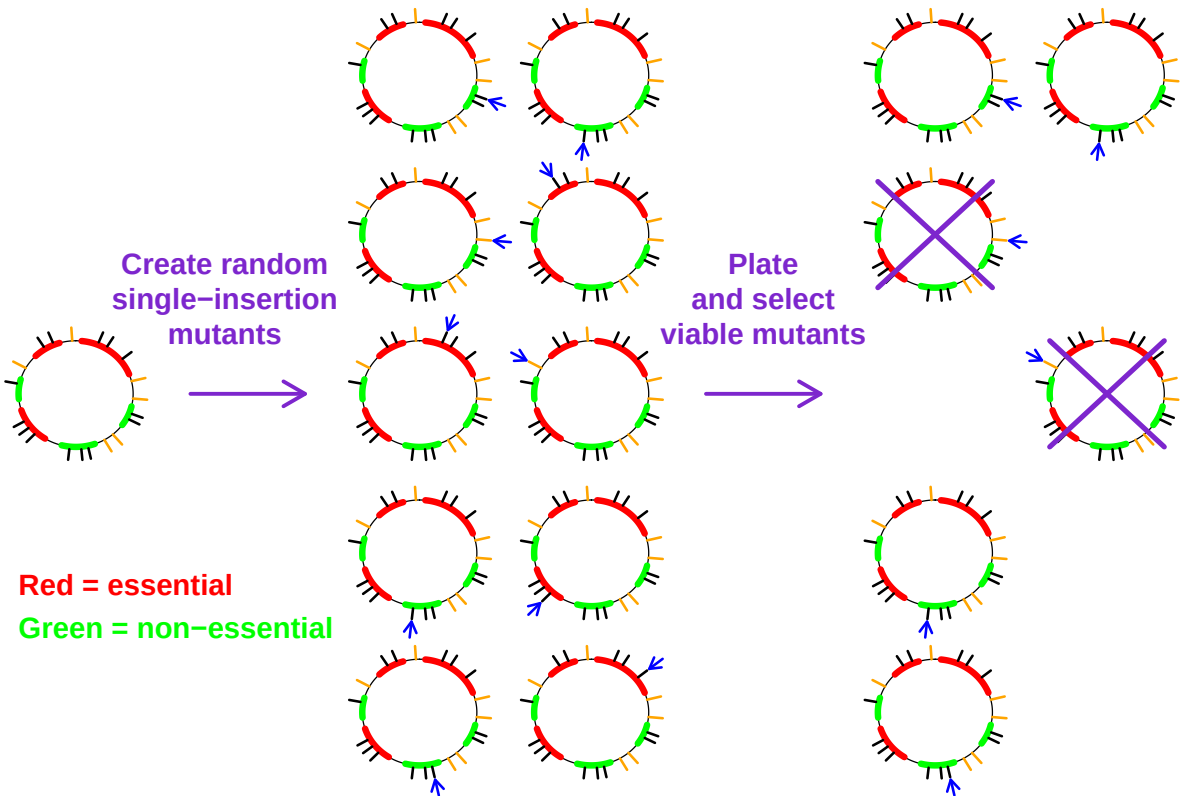
5′-TCGAAGCCTGCGAC**TA**ACGTT**TA**AAGTTTG-3′
3′-AGCTTCGGACGCTG**AT**TGCAA**AT**TTCAAAC-5′

Note: $\geq$ 30 stop codons in each reading frame

Sequence of the gene MT598



Random transposon mutagenesis

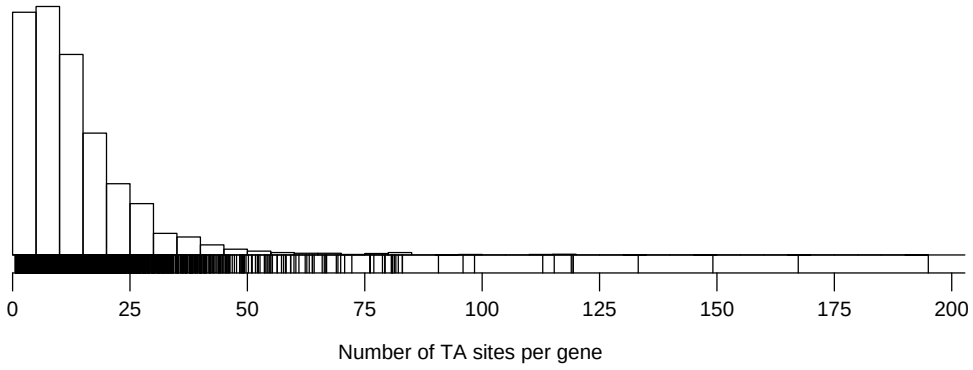


Random transposon mutagenesis

- Location of transposon insertion determined by sequencing across junctions
- Viable insertion within a gene $\Longrightarrow$ gene is non-essential
- Essential genes: we will never see a viable insertion
- Note: We only consider insertion sites within proximal 80% or n–100 basepairs of a gene.

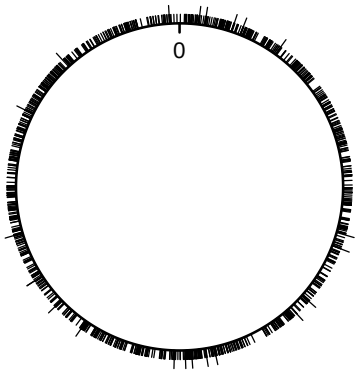
  Insertions in the very distal portion of an essential gene may not be sufficently disruptive.

TA sites in M. tuberculosis



- 74,403 sites
- 65,649 sites within a gene
- 57,934 sites within proximal portion of a gene
- 4204/4250 genes with at least one TA site

1425 insertion mutants



- 1425 insertion mutants
- 1025 within proximal portion of a gene
- 21 double-hits
- 770 unique genes hit

Questions:

- Proportion of essential genes in M. tb.?
- Which genes are likely essential?

Statistical method

Model: Transposon inserts completely at random

- Each TA site equally likely
- Genes are either completely essential or completely non-essential

Prior:

- Number of ess'l genes $\sim$ Uniform$\{0, 1, \ldots, 4204\}$
- Given no. ess'l genes, each possible subset is equally likely

Bayes by Markov chain Monte Carlo (MCMC):

Approximate calculation of

- Pr(gene i is essential $\mid$ data)
- Distribution of no. essential genes given the data

Data and model

N genes — x_i = no. TA sites in gene i

n mutants — y_i = no. mutants with insertion in gene i

$$\theta_i = \begin{cases} 1 \\ 0 \end{cases} \text{ if gene i is } \begin{matrix} \text{non-essential} \\ \text{essential} \end{matrix}$$

$$O_i = \begin{cases} 1 \\ 0 \end{cases} \text{ if } y_i \text{ is } \begin{matrix} > 0 \\ = 0 \end{matrix}$$

Model: $\mathbf{y} \sim$ multinomial(n, $\mathbf{p}$) where $p_i = x_i\,\theta_i \;/\; \sum_j x_j\,\theta_j$

Goal: Estimate $\theta_+ = \sum_i \theta_i$ or $1 - \theta_+/N$

The likelihood

$$L(\boldsymbol{\theta} \mid \mathbf{y}) = \binom{n}{\mathbf{y}} \prod_i (x_i\,\theta_i)^{y_i} \;/\; (\sum_j x_j\,\theta_j)^n$$

$$\propto \begin{cases} (\sum_i x_i\,\theta_i)^{-n} & \text{if } \theta_i = 1 \text{ whenever } O_i = 1 \\ 0 & \text{otherwise} \end{cases}$$

Notes:

- Depends only on the O_i, and not directly on the y_i.
- MLE: $\hat{\theta}_i = O_i$

The prior

$\theta_+ \sim$ uniform on $\{0, 1, \ldots, N\}$

$\boldsymbol{\theta} \mid \theta_+ \sim$ uniform over all sequences of 0's and 1's with θ_+ 1's.

Notes:

- We are assuming that Pr(θ_i = 1) = 1/2.
- This is quite different from taking θ_i iid Bernoulli(1/2).
- We are assuming that θ_i is independent of x_i and the length of the gene.
- We could make use of information about the essential or non-essential status of particular genes (e.g., known viable knock-outs).

A Gibbs sampler

Goal: Estimate Pr($\boldsymbol{\theta}|\mathbf{y}$)

Gibbs sampler:

- Begin with some initial assignment, $\boldsymbol{\theta}^{(0)}$, ensuring that $\theta_i^{(0)}$ = 1 whenever O_i = 1.
- For iteration s, consider each gene one at a time, and let $\boldsymbol{\theta}_{-i}^{(s)} = (\theta_1^{(s+1)}, \ldots, \theta_{i-1}^{(s+1)}, \theta_{i+1}^{(s)}, \ldots, \theta_N^{(s)})$.
  - Calculate Pr$(\theta_i = 1 \mid \boldsymbol{\theta}_{-i}^{(s)}, \mathbf{y})$.
  - Assign $\theta_i^{(s)}$ = 1 at random with this probability.
- Repeat many times.

The conditional probabilities

If $O_i = 1$, then $\Pr(\theta_i = 1 \mid \mathbf{y}, \boldsymbol{\theta}_{-i}^{(s)}) = 1$

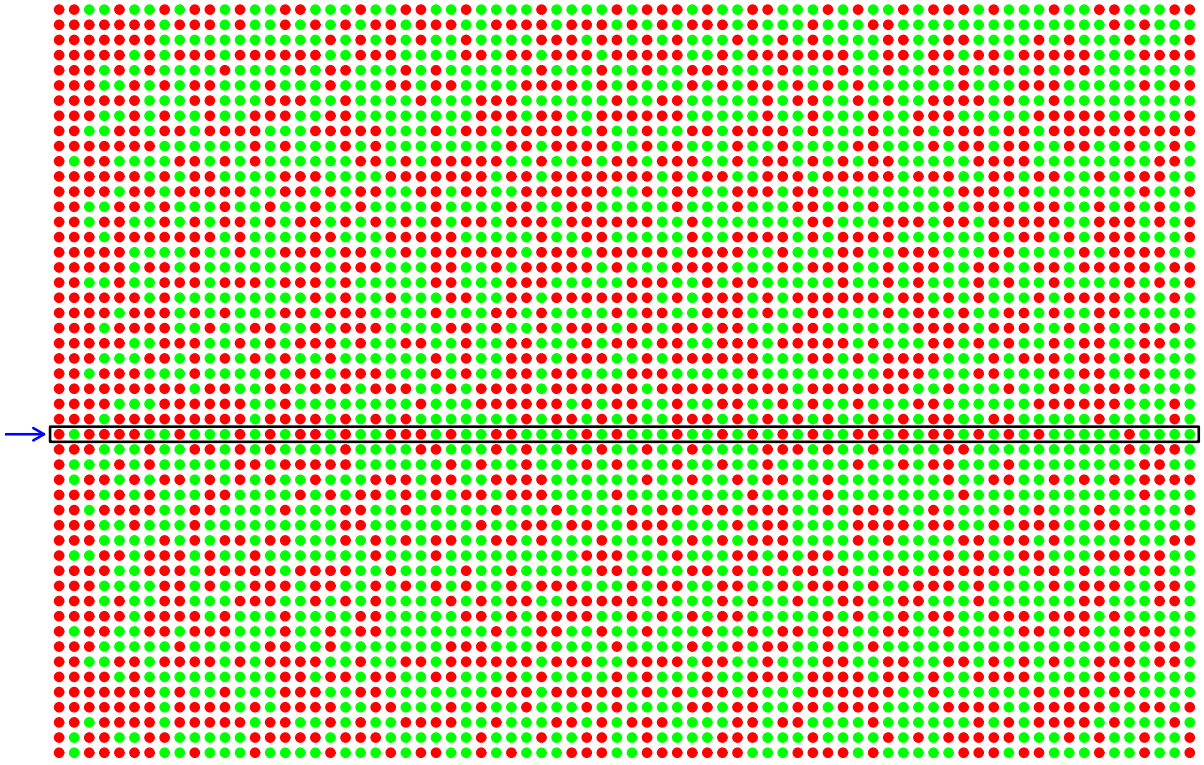
If $O_i = 0$,

$$\begin{aligned} \text{Let } A &= \sum_{j<i} \theta_j^{(s+1)} + \sum_{j>i} \theta_j^{(s)} \\ B &= \sum_{j<i} x_j\, \theta_j^{(s+1)} + \sum_{j>i} x_j\, \theta_j^{(s)} \end{aligned}$$

$$\begin{aligned} \text{Then } \Pr(\boldsymbol{\theta}_{-i}^{(s)}, \theta_i = k) &= \binom{N}{A+k} / N \\ \Pr(\mathbf{y} \mid \boldsymbol{\theta}_{-i}^{(s)}, \theta_i = k) &= (B + k\, x_i)^{-n} \end{aligned}$$

$$\begin{aligned} \text{And so } \Pr(\theta_i = 1 \mid \mathbf{y}, \boldsymbol{\theta}_{-i}^{(s)}) &= \ldots \\ &= \frac{(1 + x_i/B)^{-n}}{(1 + x_i/B)^{-n} + (N - A)/(A + 1)} \end{aligned}$$

MCMC in action



Estimators

The Gibbs sampler produces $\boldsymbol{\theta}^{(0)}, \boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(S)}$

We discard the first 200 or so samples ("burn-in").

Estimated number of non-essential genes: $E(\theta_+ \mid \mathbf{y})$

$$\theta_+^{(s)} = \sum_i \theta_i^{(s)} \qquad \longrightarrow \qquad \hat{\theta}_+ = \frac{1}{S-200} \sum_{s=201}^{S} \theta_+^{(s)}$$

Probability that gene i is non-essential: $E(\theta_i \mid \mathbf{y})$

$$\hat{\theta}_i = \frac{1}{S-200} \sum_{s=201}^{S} \theta_i^{(s)}$$

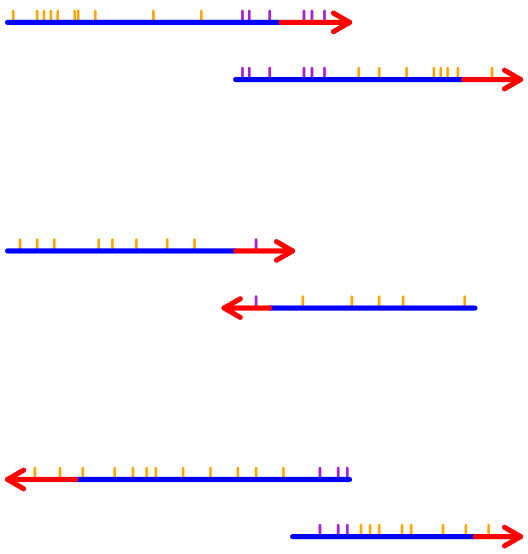
or Rao-Blackwellize:

$$\hat{\theta}_i^{\star} = \frac{1}{S-200} \sum_{s=201}^{S} \Pr(\theta_i = 1 \mid \mathbf{y}, \boldsymbol{\theta}_{-i}^{(s)})$$

A further complication

Many genes overlap

- Of 4250 genes, 1005 pairs overlap (mostly by ex. 4 bp).
- The overlapping regions contain 547 insertion sites.
- Omit TA sites in overlapping regions, unless in the proximal portion of *both* genes.
- Define
  - w_i = no. TA sites shared by genes i and i+1.
  - z_i = no. mutants shared by genes i and i+1.
  - $O_i = 1$ if $y_i = 1$, $z_i = 1$, or $z_{i+1} = 1$
- $L(\boldsymbol{\theta} \mid \mathbf{y}, \mathbf{z}) \propto (\sum_i x_i \theta_i + w_i \theta_i \theta_{i+1})^{-n}$

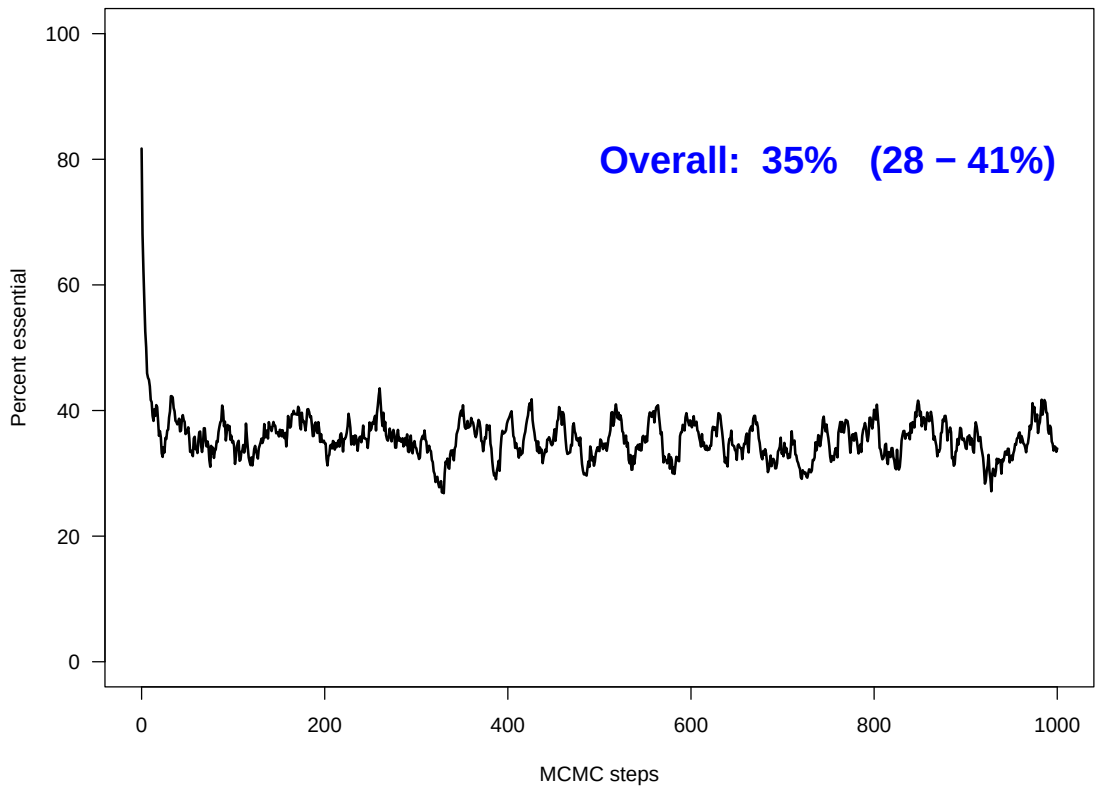


M. tb. mutagenesis data

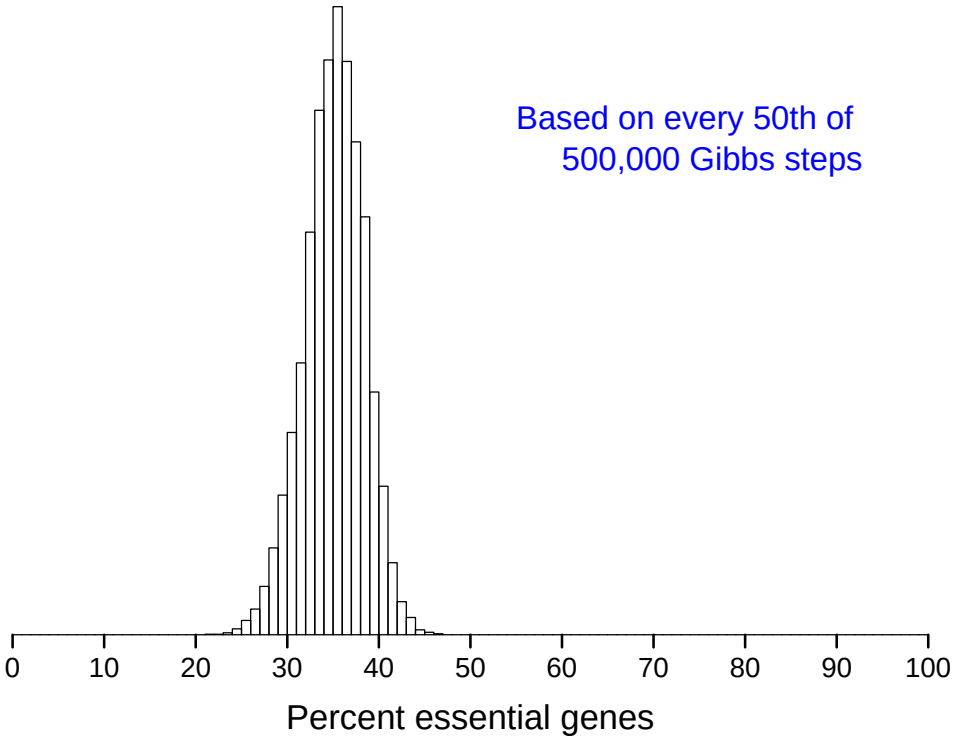
- 74,403 TA sites total
- 57,934 sites within proximal portion of a gene
- 77 sites shared by two genes
- 4204/4250 genes with at least one such site

- 1425 insertion mutants
- 1025 within proximal portion of a gene
- 2 mutants for sites shared by two genes
- 770 unique genes hit

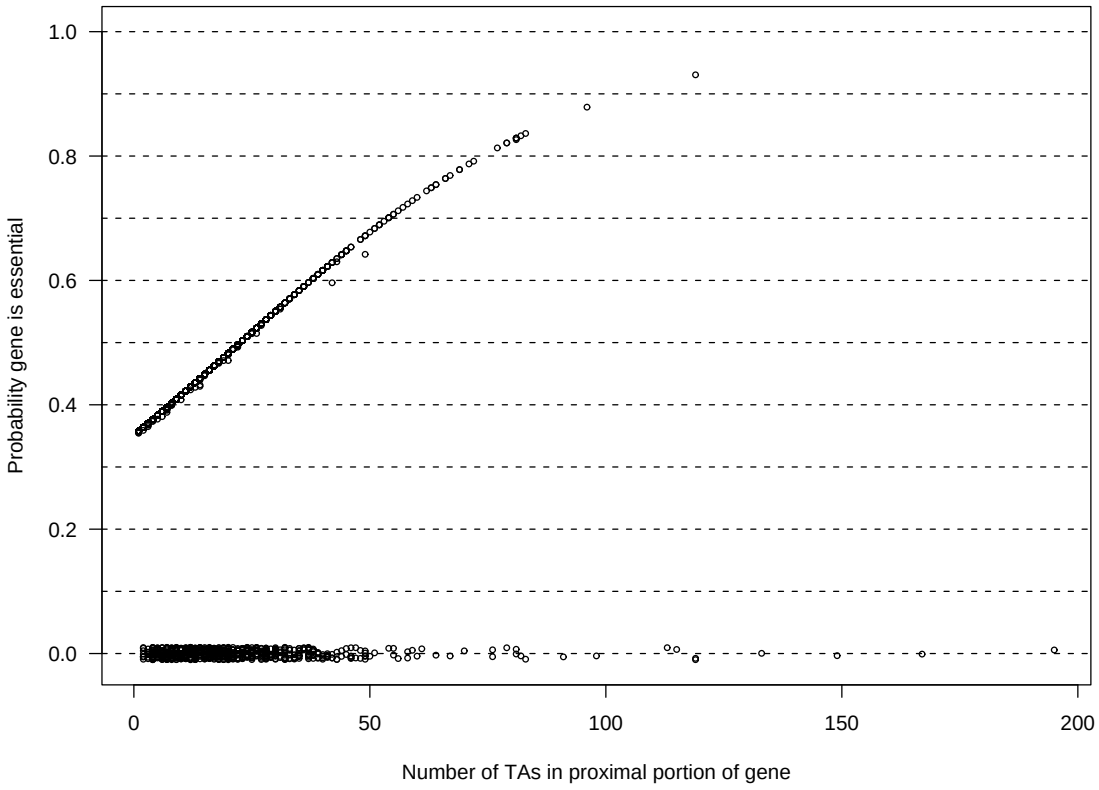
Percent essential genes in M. tb.



Percent essential genes in M. tb.

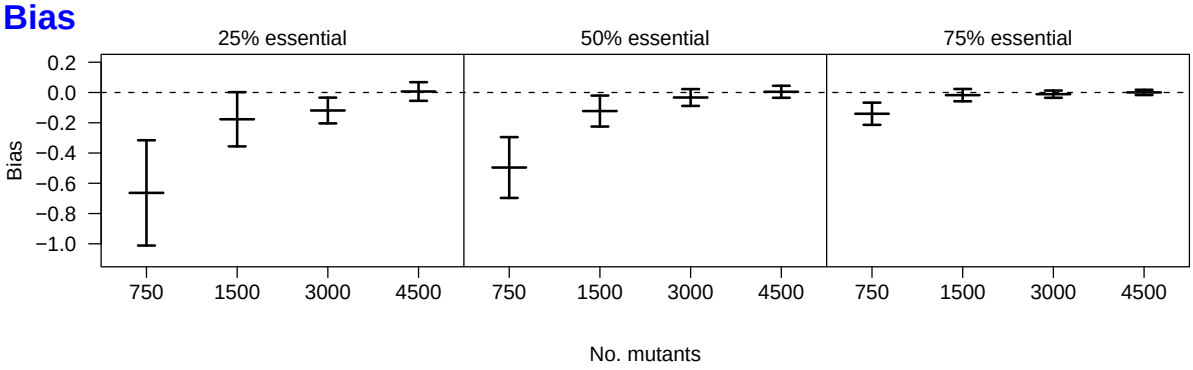


Probability that each gene is essential

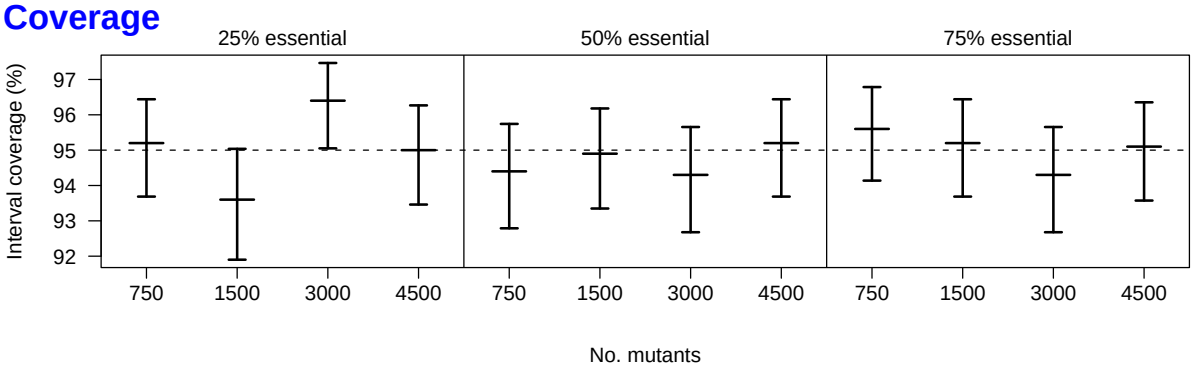


Frequentist properties of $\hat{\theta}_+$

Bias



Coverage



Based on 1000 simulations

Yet another complication

Operon: A group of adjacent genes that are transcribed together as a single unit.

- Insertion at a TA site could disrupt all downstream genes
- If a gene is essential, insertion in any upstream gene would be non-viable
- Re-define the meaning of “essential gene”.
- If operons were known, one could get an improved estimate of the proportion of essential genes.
- If one ignores the presence of operons, estimates should still be unbiased.

Summary

- Bayesian method, using MCMC, to estimate the proportion of essential genes in a genome with data from random transposon mutagenesis.
- Crucial assumptions:
  - Randomness of transposon insertion.
  - Essentiality is an all-or-none quality.
  - No relationship between essentiality and no. insertion sites.
  - The 80% rule.
- For *M. tuberculosis*, with data on 1400 mutants:
  - 28 – 41% of genes are essential
  - 20 genes which have $\geq$ 64 TA sites and for which no mutant has been observed, have $>$ 75% chance of being essential.

Acknowledgements

Bill Bishai

Natalie Blades

Gyanu Lamichhane