

# 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](http://www.biostat.jhsph.edu/~kbroman)

---

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

## *Mycobacterium tuberculosis*

---

- The organism that causes tuberculosis.
  - Cost for treatment: ~ \$15,000
  - Other bacterial pneumonias: ~ \$35
- 4.4 Mbp circular genome, completely sequenced
- 4250 known or inferred genes

# Aim

---

Identify the essential genes  
(knock-out  $\implies$  non-viable mutant)

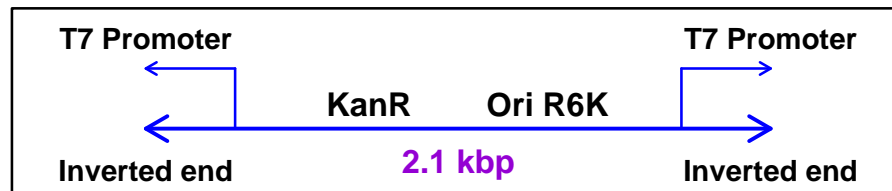
# Method

---

Random transposon mutagenesis

*Himar1*, a mariner-derived transposon

---



5' -TCGAAGCCTGCGAC**TA**ACGTT**TA**AAGTTTG-3'  
3' -AGCTTCGGACGCTG**ATT**GCAA**ATT**TCAAAC-5'

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



# Random transposon mutagenesis

---

- Location of transposon insertion determined by sequencing across junctions
- Viable insertion within a gene  $\implies$  gene is non-essential
- Essential genes: we will never see a viable insertion
- **Complication:** Insertions in the very distal portion of an essential gene may not be sufficiently disruptive.

Thus, we omit from consideration insertion sites within the last 20% and last 100 bp of a gene.

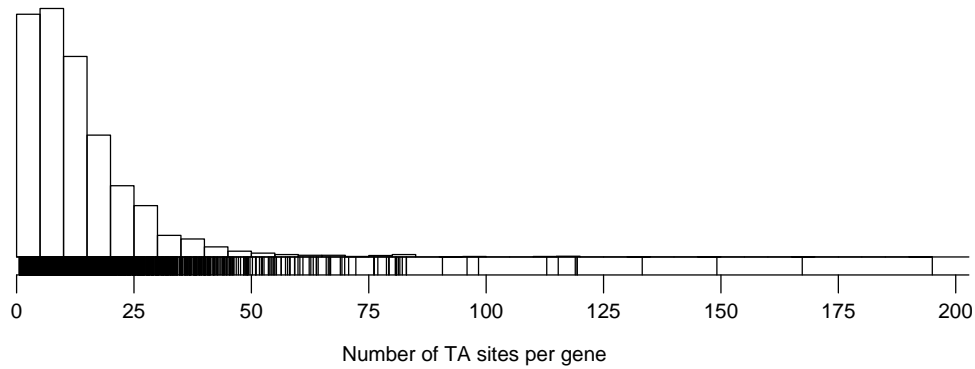
## The data

---

- Number, locations of genes.
- Number of insertion sites in each gene.
- $n$  viable mutants with exactly one transposon insertion.
- Location of the transposon insertion in each mutant.

# TA sites in *M. tuberculosis*

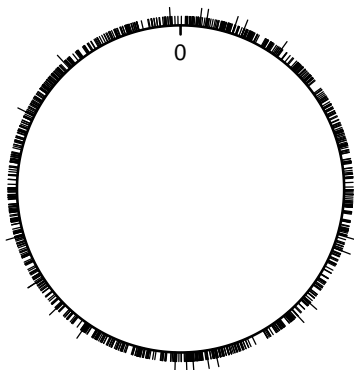
---



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

**Bayes by Markov chain Monte Carlo (MCMC):**

Approximate calculation of

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

## Data and model

---

N genes  $x_i = \text{no. TA sites in gene } i$

n mutants  $y_i = \text{no. mutants with insertion in gene } i$

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

**Model:**  $\mathbf{y} \sim \text{multinomial}(n, \mathbf{p})$  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

---

$$L(\boldsymbol{\theta} \mid \mathbf{y}) = \binom{n}{\mathbf{y}} \prod_i (x_i \theta_i)^{y_i} / \left( \sum_j x_j \theta_j \right)^n$$
$$\propto \begin{cases} (\sum_i x_i \theta_i)^{-n} & \text{if } \theta_i = 1 \text{ whenever } y_i > 0 \\ 0 & \text{otherwise} \end{cases}$$

## Notes:

- Depends only on which  $y_i > 0$ , and not directly on the particular values of  $y_i$ .
- MLE:  $\hat{\theta}_i = 1\{y_i > 0\}$

# The prior

---

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

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

## Notes:

- We are assuming that  $\Pr(\theta_i = 1) = 1/2$ .
- This is quite different from taking  $\theta_i$  iid Bernoulli(1/2).
- We are assuming that  $\theta_i$  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  $y_i > 0$ .
- For iteration  $s$ , consider each gene one at a time, and let  $\boldsymbol{\theta}_{-i}^{(s)} = (\theta_1^{(s+1)}, \dots, \theta_{i-1}^{(s+1)}, \theta_{i+1}^{(s)}, \dots, \theta_N^{(s)})$ .
  - 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  $y_i > 0$ , then  $\Pr(\theta_i = 1 \mid \mathbf{y}, \boldsymbol{\theta}_{-i}^{(s)}) = 1$

If  $y_i = 0$ ,

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

$$\text{Then } \Pr(\boldsymbol{\theta}_{-i}^{(s)}, \theta_i = k) = \binom{N}{A+k} / N$$

$$\Pr(\mathbf{y} \mid \boldsymbol{\theta}_{-i}^{(s)}, \theta_i = k) = (B + k x_i)^{-n}$$

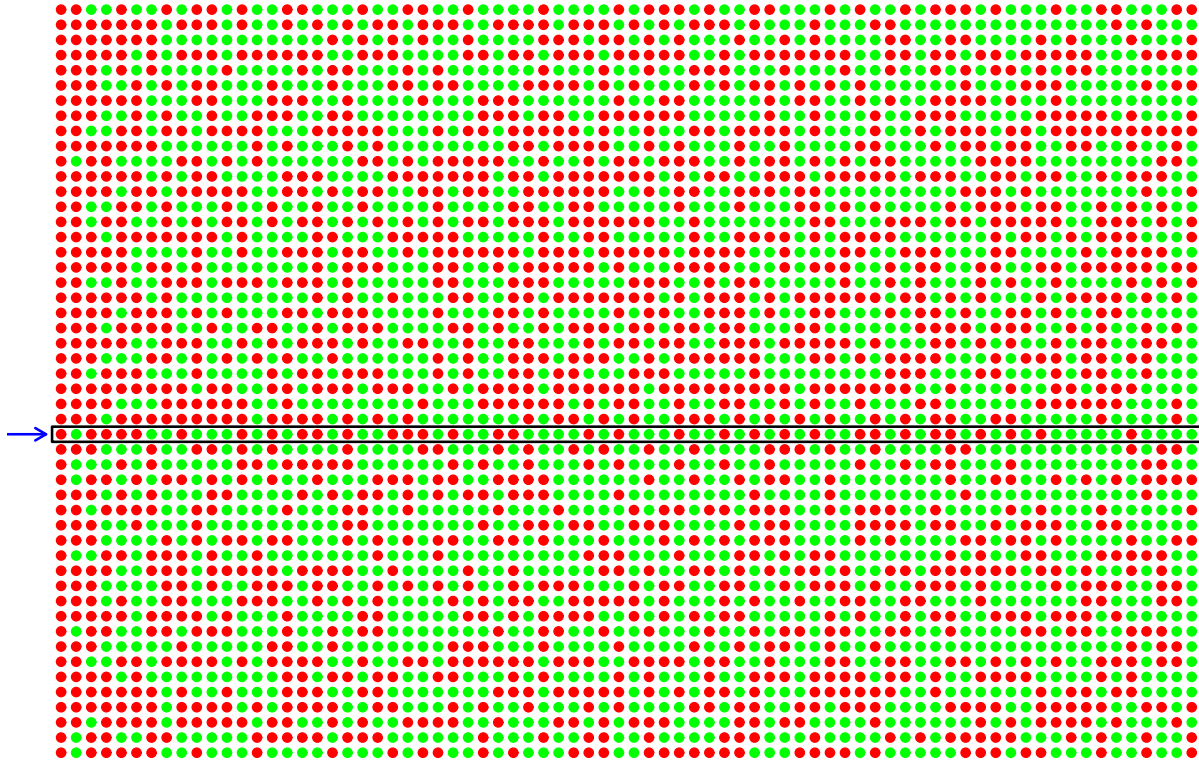
And so  $\Pr(\theta_i = 1 \mid \mathbf{y}, \boldsymbol{\theta}_{-i}^{(s)}) = \dots$

$$= \frac{(1 + x_i/B)^{-n}}{(1 + x_i/B)^{-n} + (N - A)/(A + 1)}$$



# MCMC in action

---



## Estimators

---

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

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

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

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

Probability that gene  $i$  is non-essential:  $E(\theta_i | \mathbf{y}) = \Pr(\theta_i = 1 | \mathbf{y})$

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

or Rao-Blackwellize:

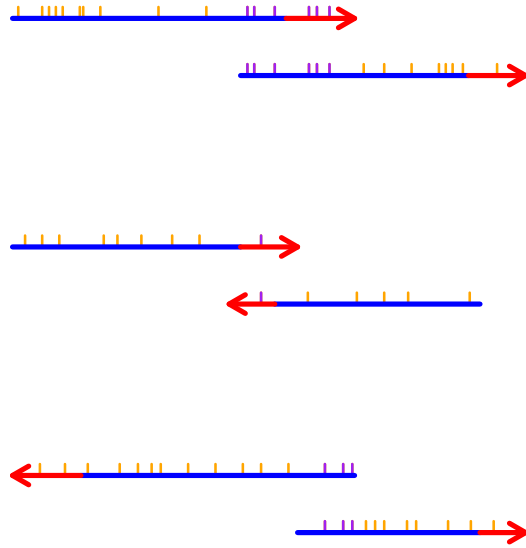
$$\hat{\theta}_i^* = \frac{1}{S-200} \sum_{s=201}^S \Pr(\theta_i = 1 | \mathbf{y}, \theta_{-i}^{(s)})$$

# A further complication

---

## Many genes overlap

- Of 4250 genes, 1005 pairs overlap (mostly by exactly 4 bp).
- The overlapping regions contain 547 insertion sites.
- Omit TA sites in overlapping regions, unless in the proximal portion of *both* genes.
- The algebra gets a bit more complicated.



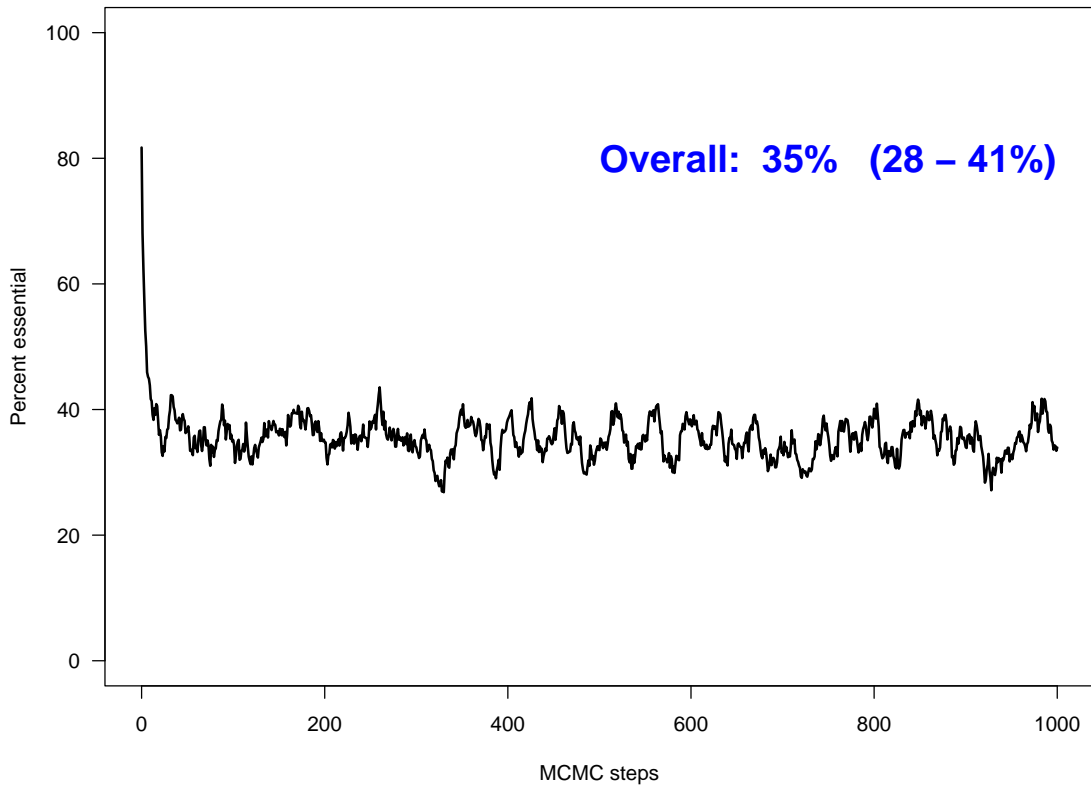
## M. tb. mutagenesis data

---

- 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

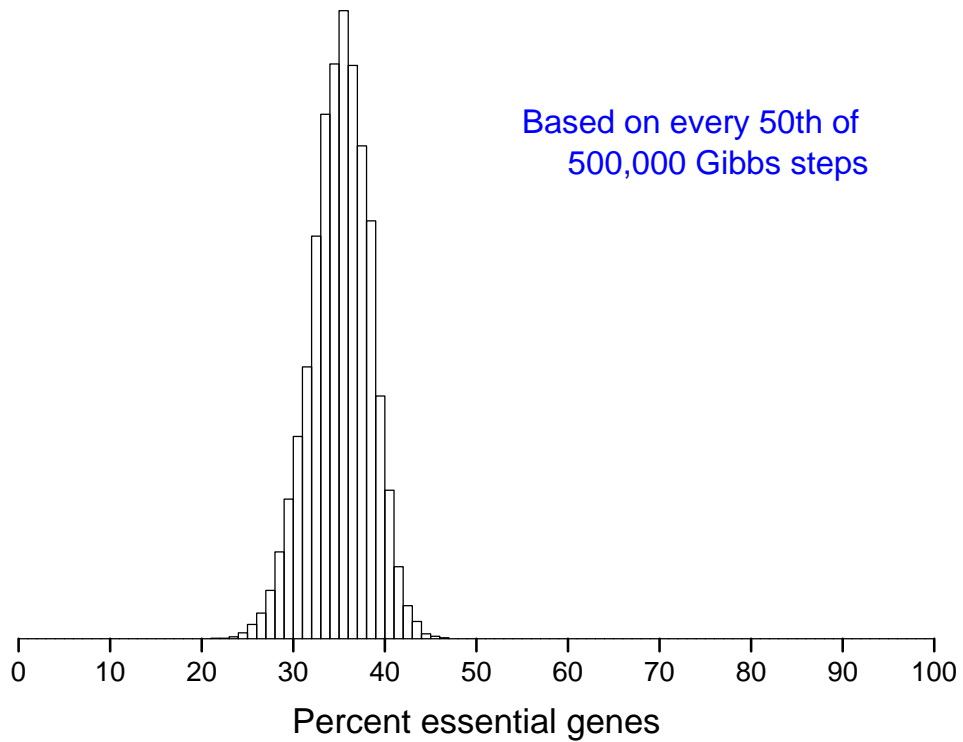
# Percent essential genes in M. tb.

---

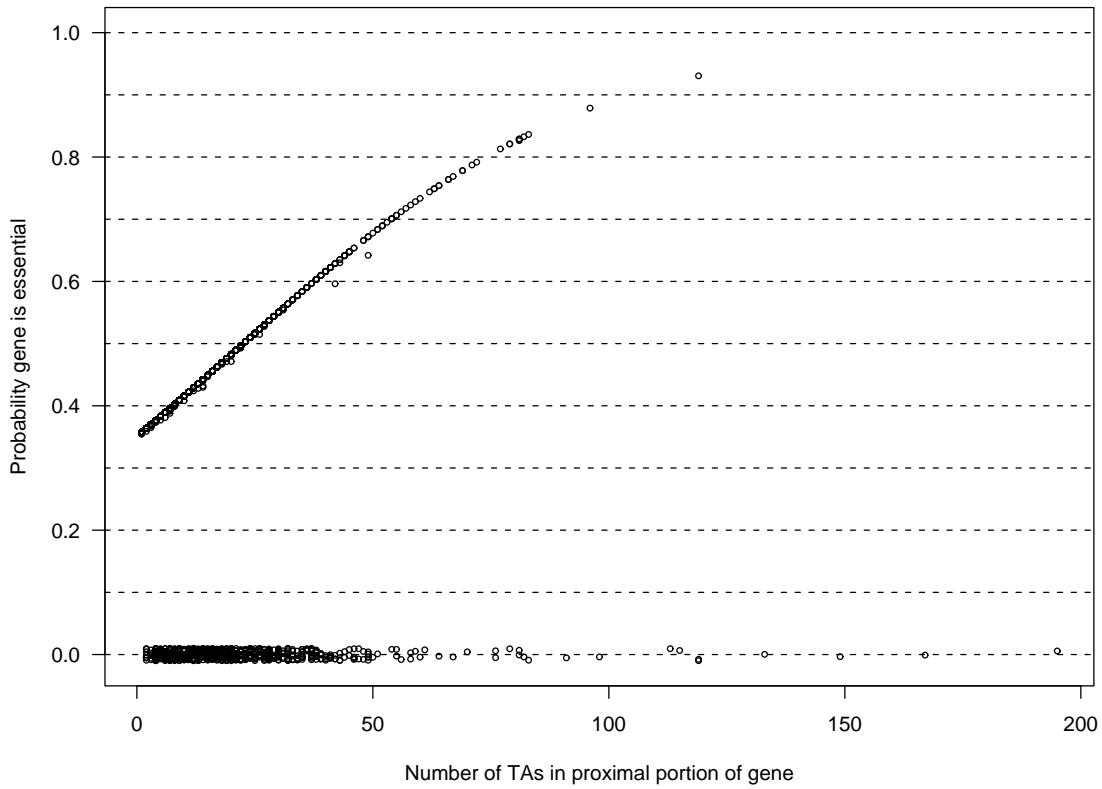


# Percent essential genes in M. tb.

---

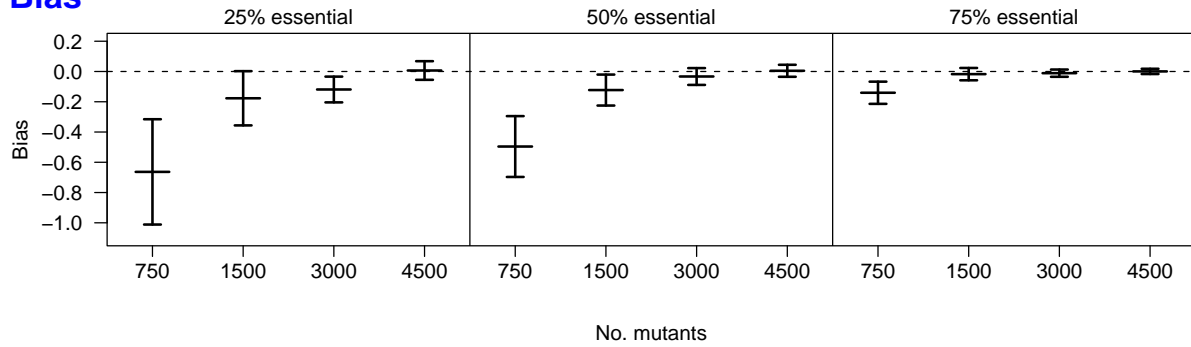


# Probability that each gene is essential

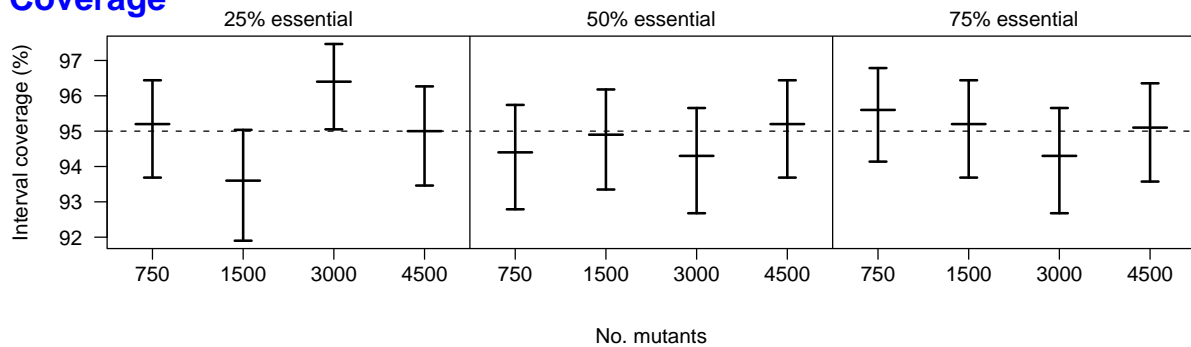


## Frequentist properties of $\hat{\theta}_+$

### 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

(and many others)