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

Karl W. Broman

Department of Biostatistics
Johns Hopkins University

http://www.biostat.jhsph.edu/~kbroman

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

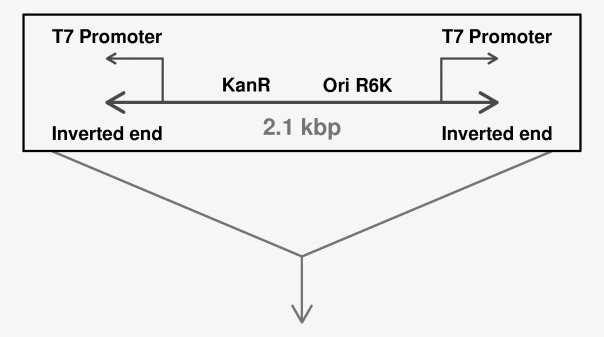
2

Aim

- Identify the essential genes
  - Knock-out $\Rightarrow$ non-viable mutant
- Random transposon mutagenesis
  - Rather than knock out each gene systematically, we knock out them out at random.

3

The *Himar1* transposon

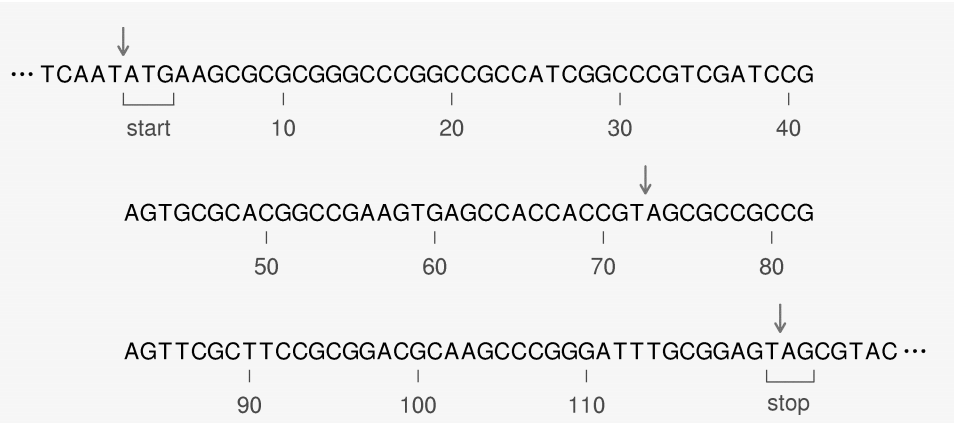



```
5′ - TCGAAGCCTGCGACTAACGTTTAAAGTTTG - 3'
3′ - AGCTTCGGACGCTGATTGCAAATTTCAAAC - 5'
```

Note: 30 or more stop codons in each reading frame

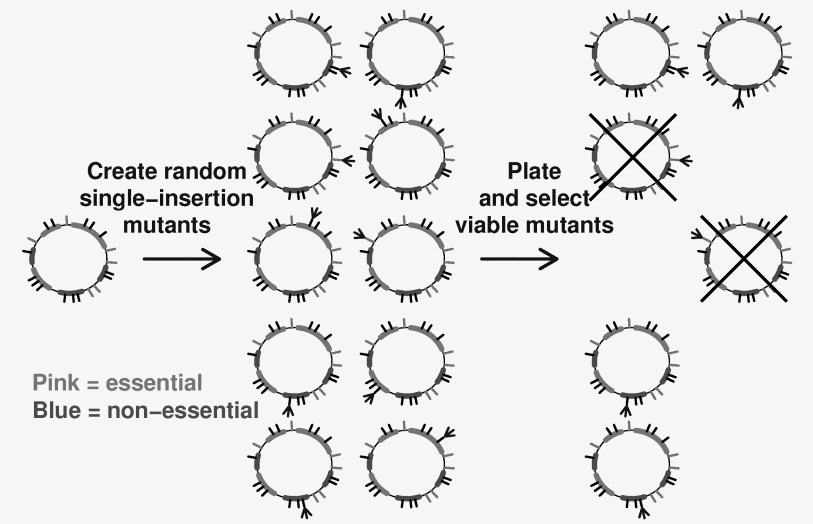
4

Sequence of the gene MT598



5

Random transposon mutagenesis



6

Random transposon mutagenesis

- Locations of tranposon insertion determined by sequencing across junctions.
- Viable insertion within a gene ⇒ gene is not 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.

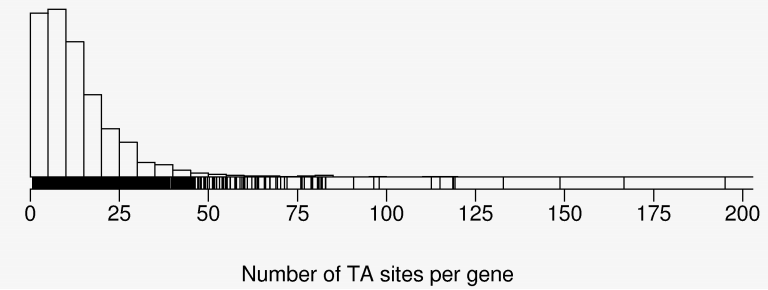
7

The data

- Number, locations of genes
- Number of insertion sites in each gene
- Viable mutants with exactly one transposon
- Location of the transposon insertion in each mutant

8

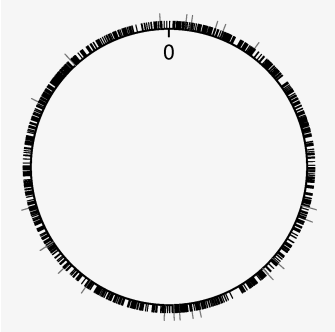
TA sites in *M. tuberculosis*



- 74,403 sites
- 65,659 sites within a gene
- 57,934 sites within proximal portion of a gene
- 4204/4250 genes with at least one TA site

9

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 Mtb?
- Which genes are likely essential?

10

Statistics, Part 1

- Find a probability model for the process giving rise to the data.
- Parameters in the model correspond to characteristics of the underlying process that we wish to determine

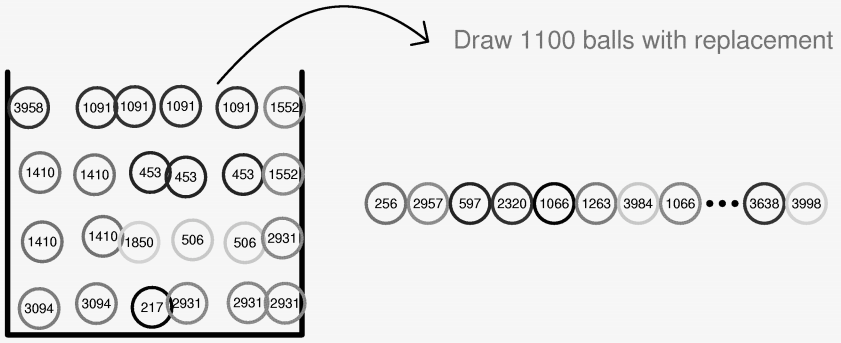
11

The model

- Transposon inserts completely at random (each TA site equally likely to be hit)
- Genes are either completely essential or completely non-essential.
- Let N = no. genes t_i = no. TA sites in gene i
  n = no. mutants m_i = no. mutants of gene i
- $\theta_i = \begin{cases} 1 \\ 0 \end{cases}$ if gene *i* is $\begin{matrix} \text{non-essential} \\ \text{essential} \end{matrix}$

12

A picture of the model



Urn with balls labelled 1–4204

If essential: 0 balls

If non–essential: no. balls = no. TA sites

13

Part of the data

Gene	No. TA sites	No. mutants
1	31	0
2	29	0
3	34	1
4	3	0
:	:	:
22	49	2
:	:	:
4204	4	0
Total	**57,934**	**1,025**

14

A related problem

- How many species of insects are there in the Amazon?
  - Get a random sample of insects.
  - Classify according to species.
  - How many total species exist?
- The current problem is a lot easier:
  - Bound on the total number of classes.
  - Know the relative proportions (up to a set of 0/1 factors).

15

Statistics, Part 2

Find an estimate of $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_N)$.

We're particularly interested in $\theta_+ = \sum_i \theta_i$ and $1 - \theta_+ / N$

Frequentist approach

- View parameters $\{\theta_i\}$ as fixed, unknown values
- Find some estimate that has good properties
- Think about repeated realizations of the experiment.

Bayesian approach

- View the parameters as random.
- Specify their joint prior distribution.
- Do a probability calculation.

16

The likelihood

$$L(\theta \mid \mathbf{m}) = \Pr(\mathbf{m} \mid \theta)$$

$$= \binom{n}{\mathbf{m}} \prod_i (t_i\theta_i)^{m_i} \Big/ \left(\sum_j t_j\theta_j\right)^n$$

$$\propto \begin{cases} \left(\sum_i t_i\theta_i\right)^{-n} & \text{if } \theta_i = 1 \text{ whenever } m_i > 0 \\ 0 & \text{otherwise} \end{cases}$$

Note: Depends on which $m_i > 0$, but not directly on the particular values of m_i.

17

Frequentist method

Maximum likelihood estimates (MLEs):

Estimate the θ_i by the values for which $L(\boldsymbol{\theta} \mid \mathbf{m})$ achieves its maximum.

In this case, the MLEs are $\hat{\theta}_i = \begin{cases} 1 \text{ if } m_i > 0 \\ 0 \text{ if } m_i = 0 \end{cases}$

Further, $\hat{\theta}_+$ = No. genes with at least one hit.

This is a really stupid estimate!

18

Bayes: The prior

θ_+ ~ uniform on {0, 1, …, N}

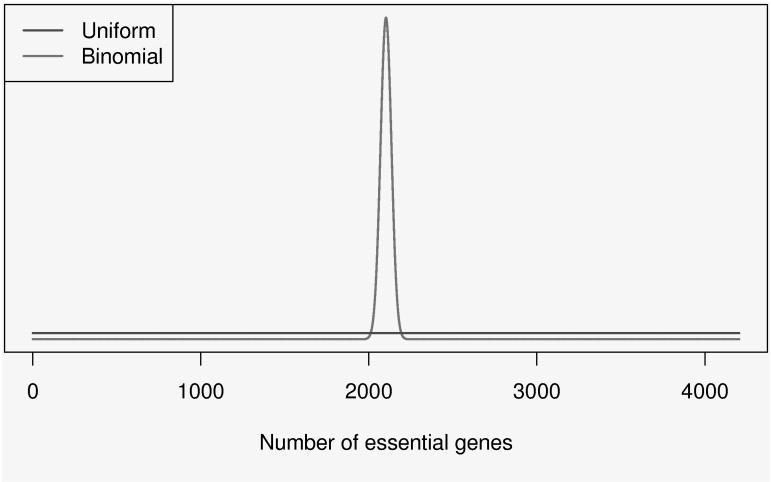
$\boldsymbol{\theta} \mid \theta_+$ ~ uniform on sequences of 0s and 1s with θ_+ 0s

Note:

- We are assuming that $Pr(\theta_i = 1) = 1/2$.
- This is quite different from taking the θ_i to be like coin tosses.
- We are assuming that θ_i is independent of t_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).

19

Uniform vs. Binomial



20

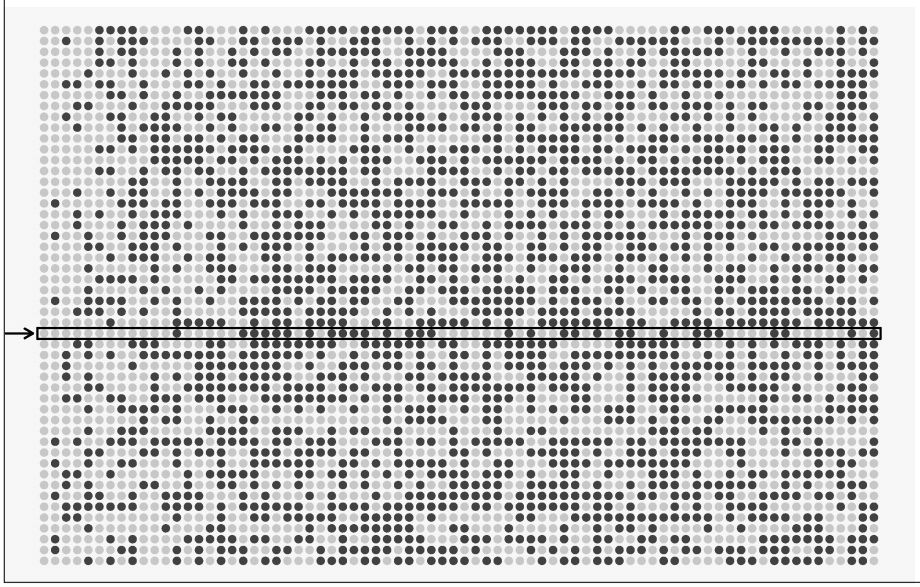
Markov chain Monte Carlo

Goal: Estimate Pr($\boldsymbol{\theta}$ | **m**).

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

21

MCMC in action

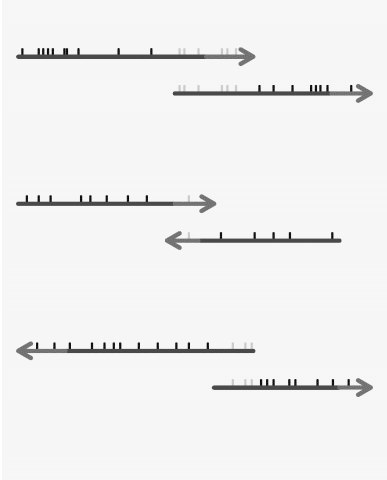


22

A further complication

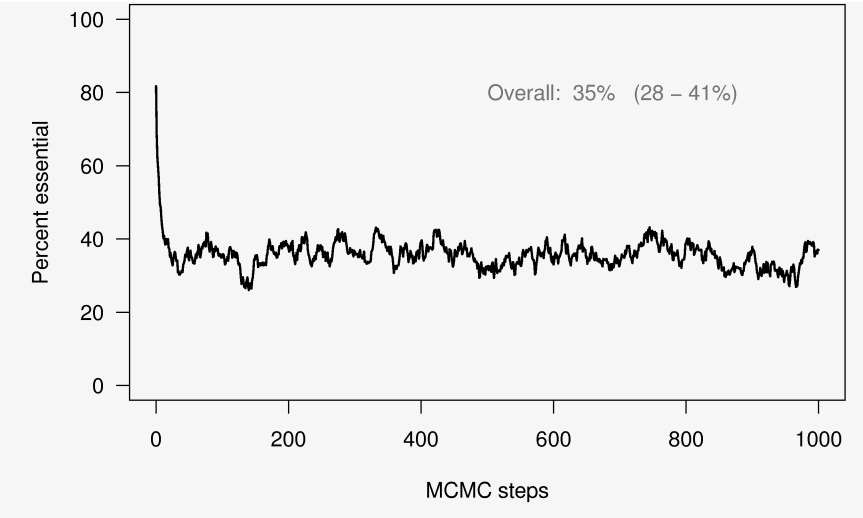
Many genes overlap

- 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.



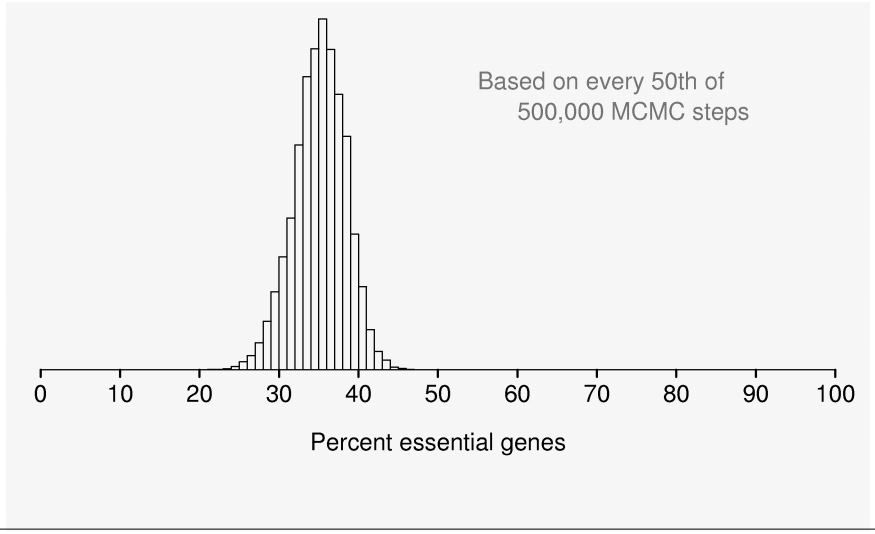
23

Percent essential genes



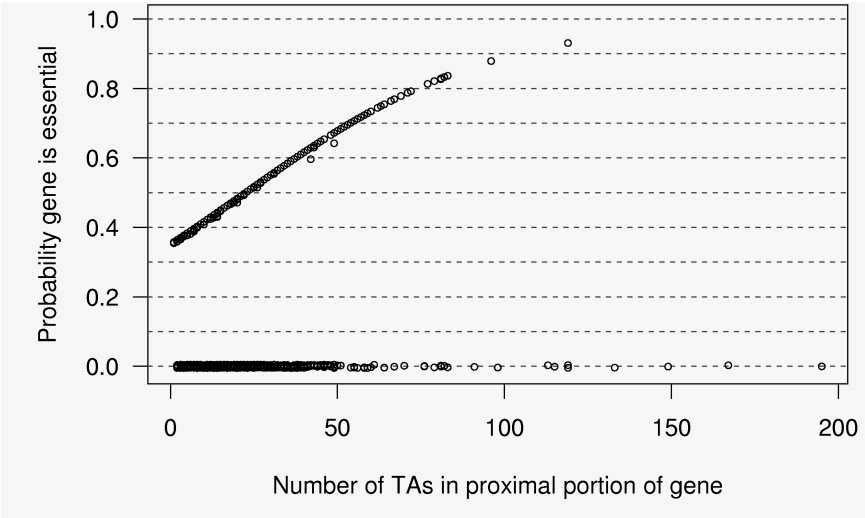
24

Percent essential genes



25

Probability a gene is essential



26

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 are still unbiased.

27

Summary

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

28

Acknowledgements

Natalie Blades (now at The Jackson Lab)

Gyanu Lamichhane, Hopkins

William Bishai, Hopkins

29