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

Karl W Broman

Department of Biostatistics
Johns Hopkins University

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: $\sim$ \$15,000
  - Other bacterial pneumonias: $\sim$ \$35
- 4.4 Mbp circular genome, completely sequenced
- 4250 known or inferred genes

Aim

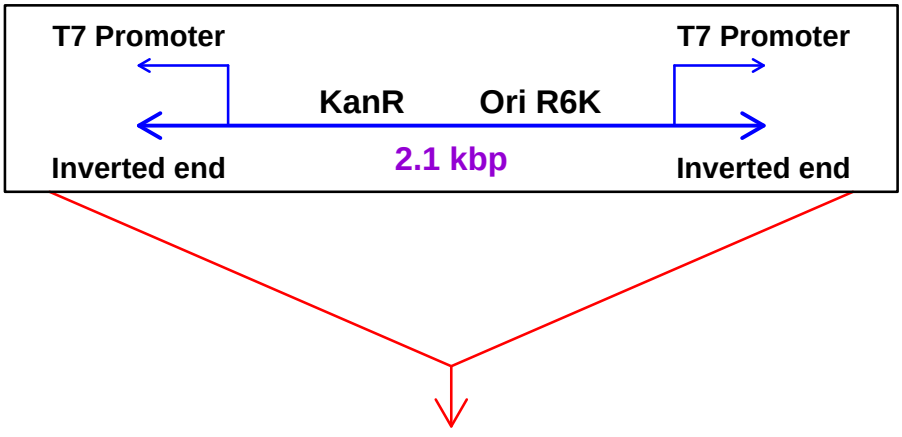
Identify the essential genes

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

Method

Random transposon mutagenesis

The *Himar1* transposon

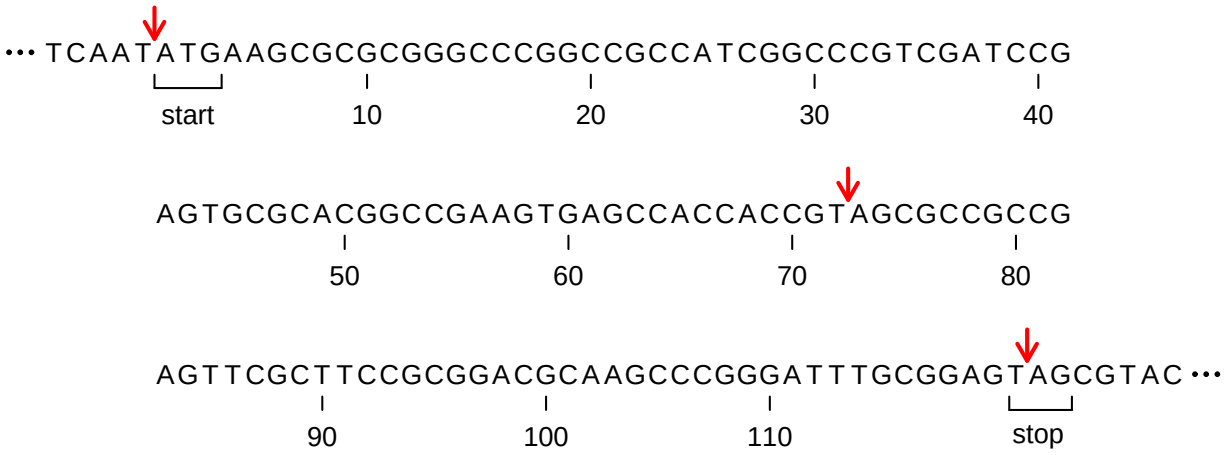


5′-TCGAAGCCTGCGAC**TA**ACGTT**TA**AAGTTTG-3′

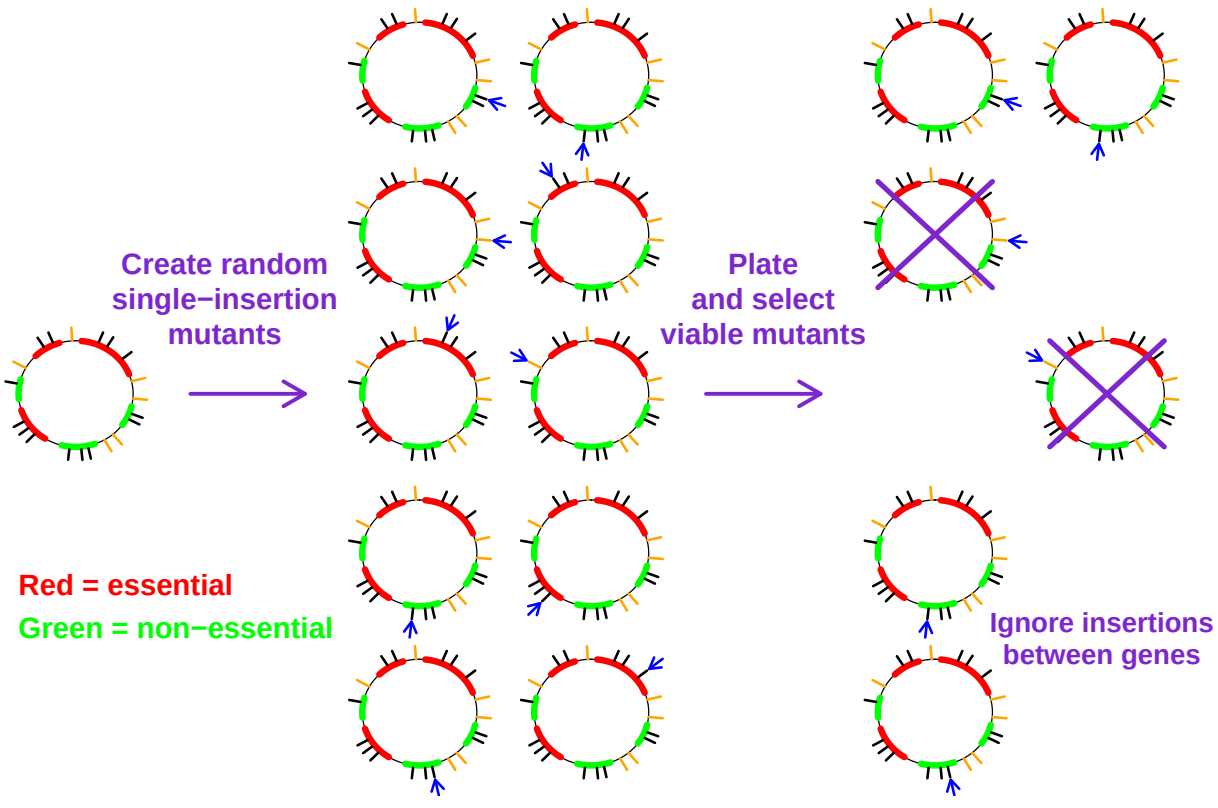
3′-AGCTTCGGACGCTG**AT**TGCAA**AT**TTCAAAC-5′

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

Sequence of the gene MT598



Random transposon mutagenesis



Random transposon mutagenesis

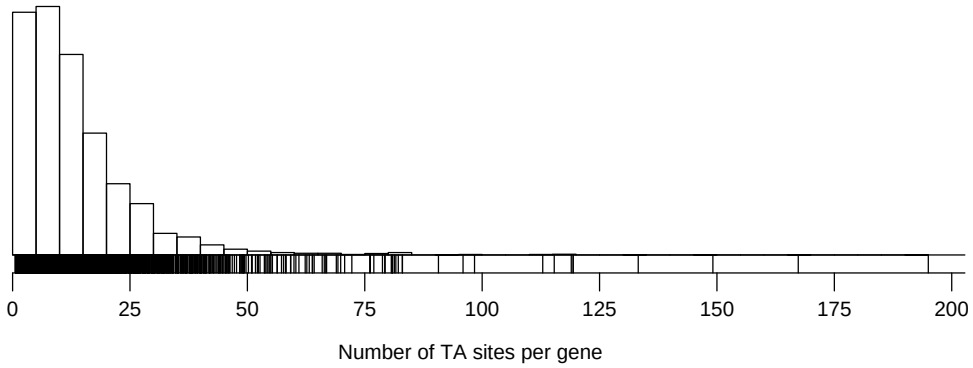
- Location of transposon insertion determined by sequencing across junctions
- Viable insertion within a gene $\Longrightarrow$ gene is non-essential
- Essential genes: we will never see a viable insertion
- Note: We only consider insertion sites within proximal 80% or n–100 basepairs of a gene.

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

The data

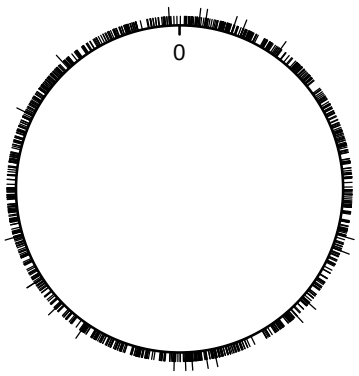
- Number, locations of genes.
- Number of insertion sites in each gene.
- *n* viable mutants with exactly one transposon insertion.
- Location of the transposon insertion in each mutant.

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 $\sim$ Uniform$\{0, 1, \ldots, 4204\}$
- Given no. essential genes, each possible subset is equally likely

Bayes by a Gibbs sampler:

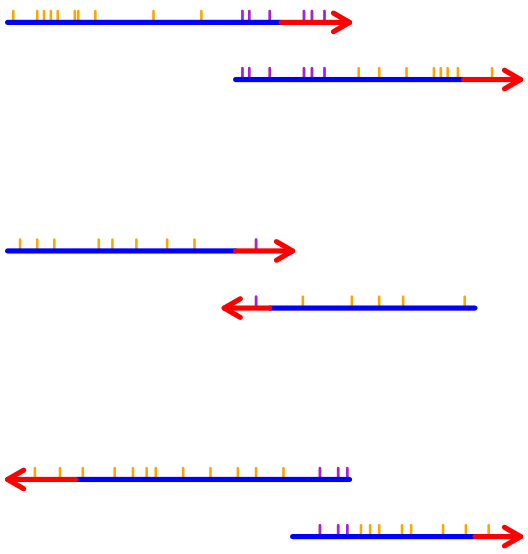
Estimate:

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

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.

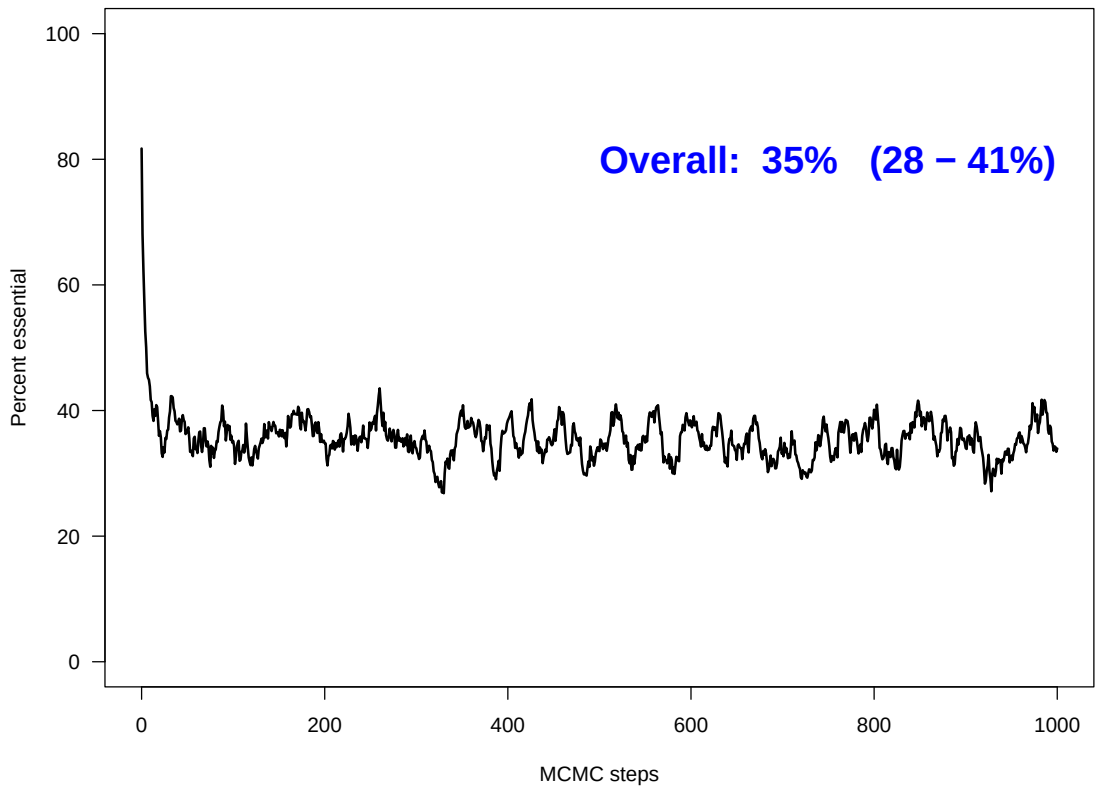


M. tb. mutagenesis data

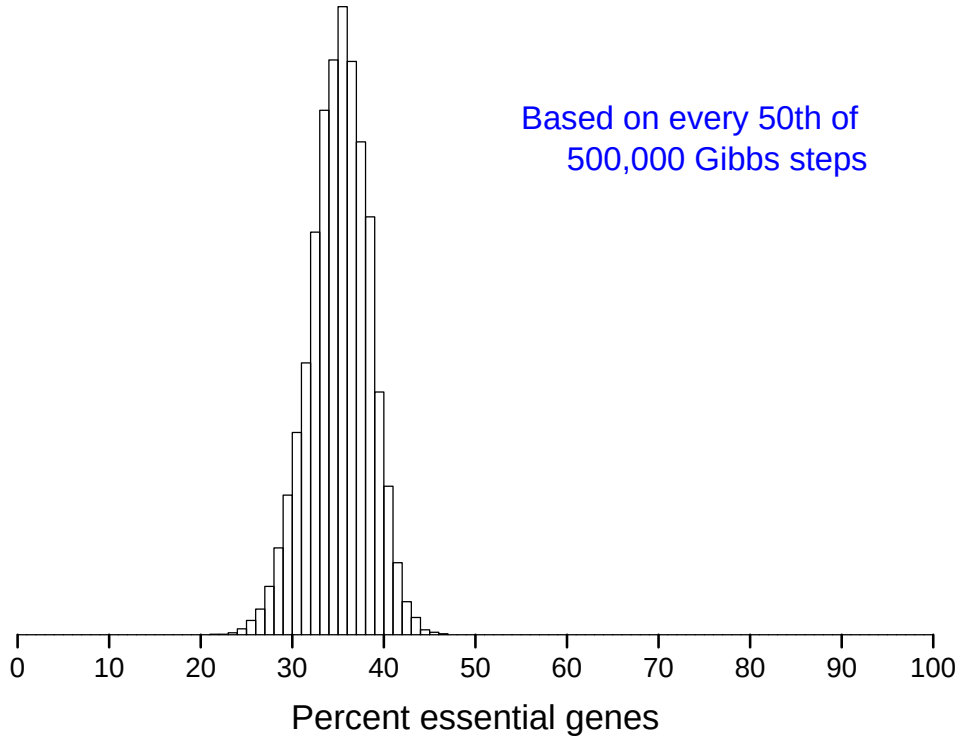
- 74,403 TA sites total
- 57,934 sites within proximal portion of a gene
- 77 sites shared by two genes
- 4204/4250 genes with at least one such site

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

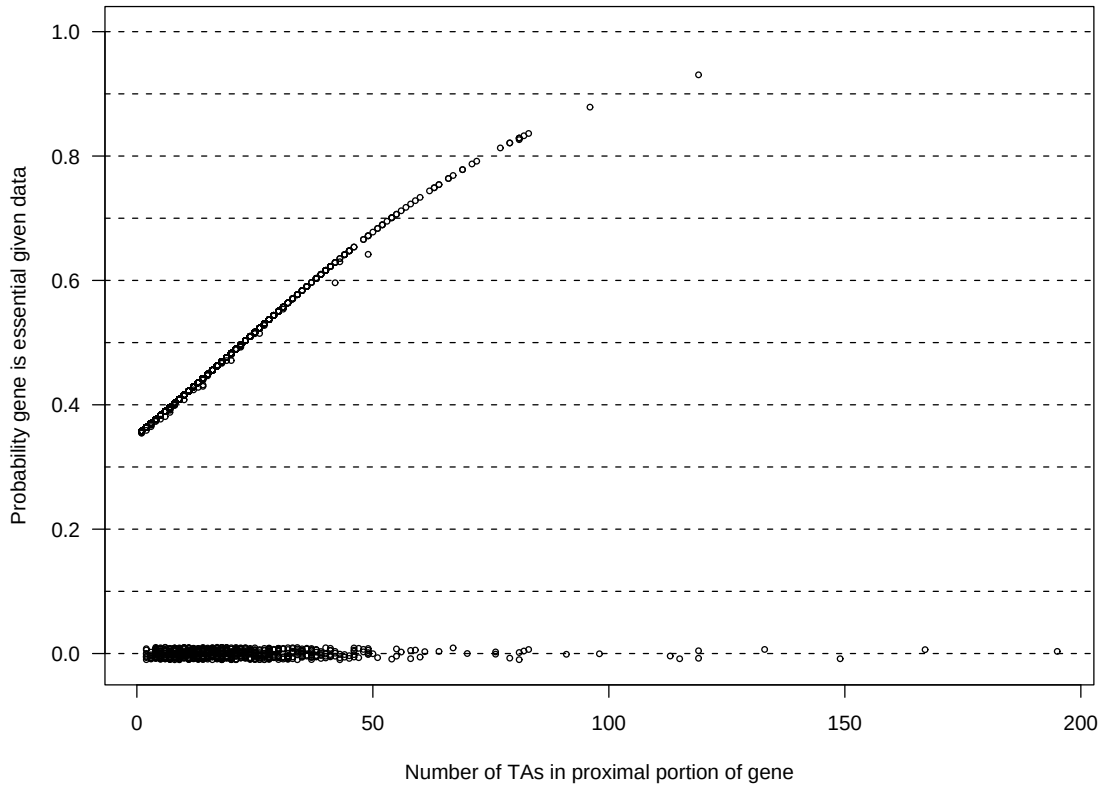
Percent essential genes in M. tb.



Percent essential genes in M. tb.



Posterior probability that gene is essential



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