Bayesian analysis

Identifying essential genes by mutagenesis

Karl Broman

Biostatistics & Medical Informatics, UW–Madison

kbroman.org
github.com/kbroman
@kwbroman

Course web: kbroman.org/AdvData

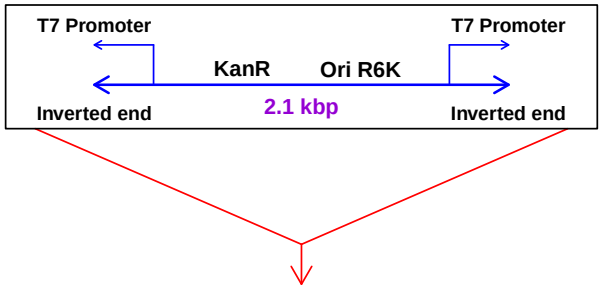
Mycobacterium tuberculosis

- 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

Goal: identify the essential genes

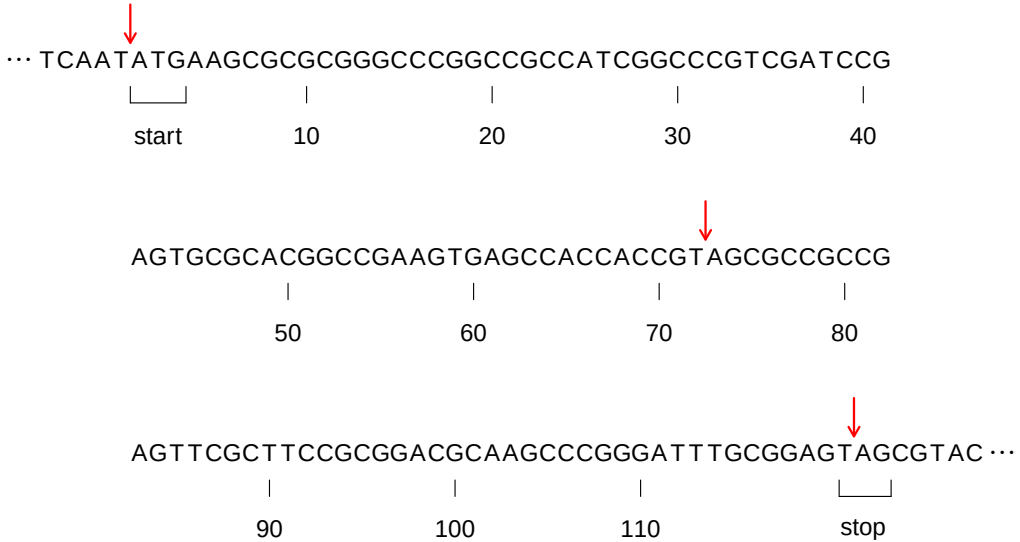
Method: random transposon mutagenesis

Himar1 transposon



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

Sequence of the gene MT598

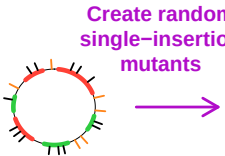


Random transposon mutagenesis

Red = essential

Green = non−essential

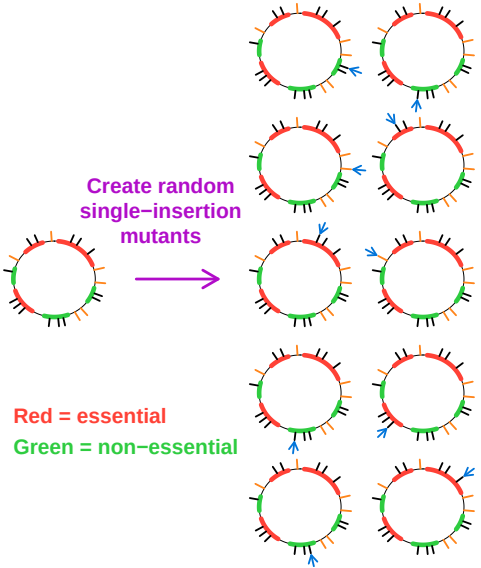
Random transposon mutagenesis



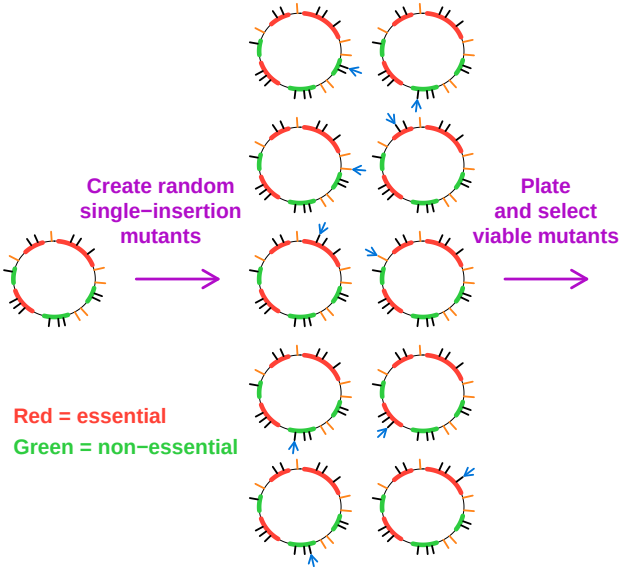
Red = essential

Green = non−essential

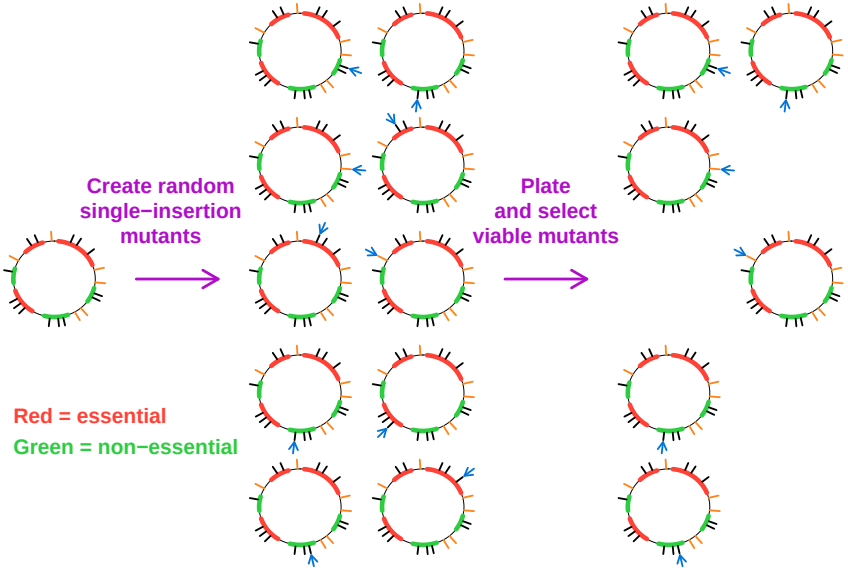
Random transposon mutagenesis



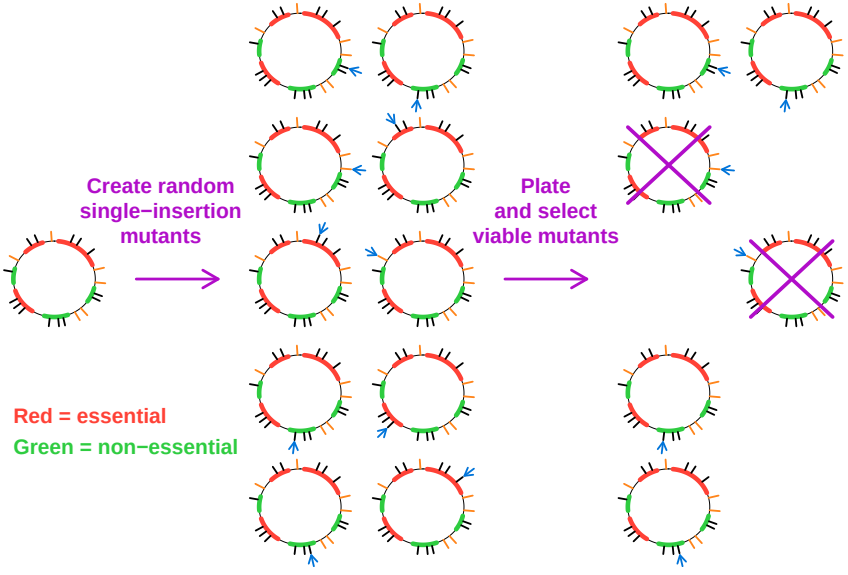
Random transposon mutagenesis



Random transposon mutagenesis



Random transposon mutagenesis



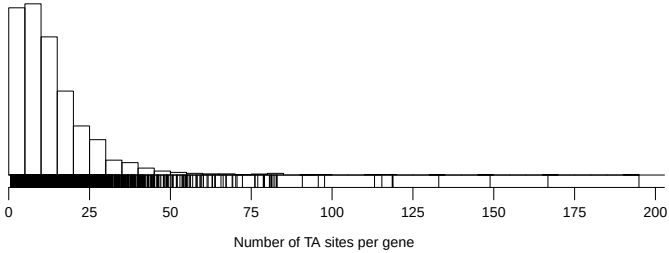
Random transposon mutagenesis

- 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
- Complication: Insertions in the very distal portion of an essential gene may not be sufficiently disruptive.
  Thus, we omit from consideration insertions sites within the last 20% and last 100 bp of a gene.

The data

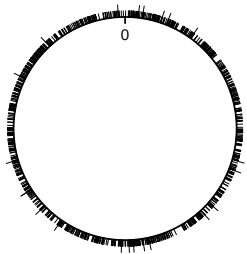
- Number, locations of genes
- Number of insertion sites in each gene
- n viable mutants with exactly one transposon insertion
- Location of the transposon insertion in each mutant

TA sites in M. tuberculosis



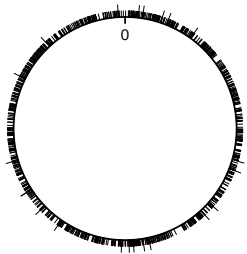
- 74,403 sites
- 65,649 sites within a gene
- 57,934 sites within proximal portion of a gene
- 4204/4250 genes with at least one TA site

1425 insertion mutants



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

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. tuberculosis?

Which genes are likely essential?

Model

Transposon inserts completely at random

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

Model

N genes x_i = no. TA sites in gene i

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

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

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

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

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

Notes:

- Depends only on which $y_i > 0$ and not on the specific values
- The MLE is $\hat{\theta}_i = 1\{y_i > 0\}$

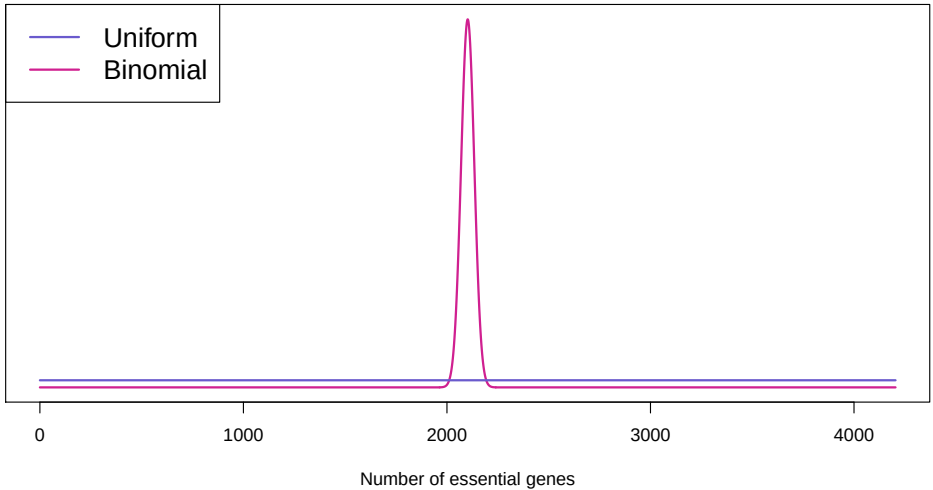
The prior

$\theta_+ \sim$ uniform on { 0, 1, ..., 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$(\theta_i = 1) = 1/2$
- 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 status of particular genes (e.g. known viable knock-outs)


Uniform
Binomial
0
1000
2000
3000
4000
Number of essential genes

A Gibbs sampler

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

Gibbs sampler:

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

This is an example of Markov chain Monte Carlo (MCMC).

MCMC in action

MCMC in action

MCMC in action

MCMC in action

MCMC in action

MCMC in action

MCMC in action

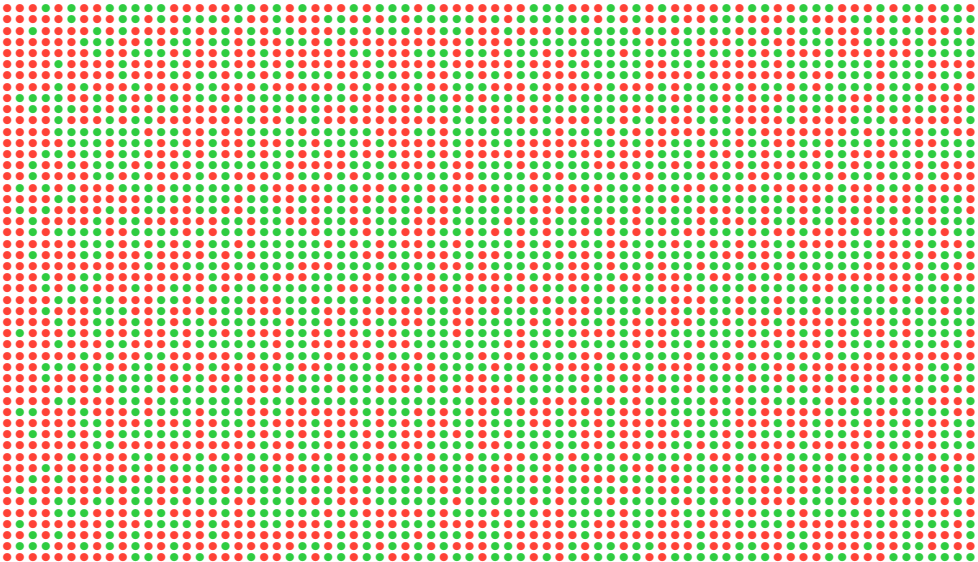
MCMC in action

MCMC in action

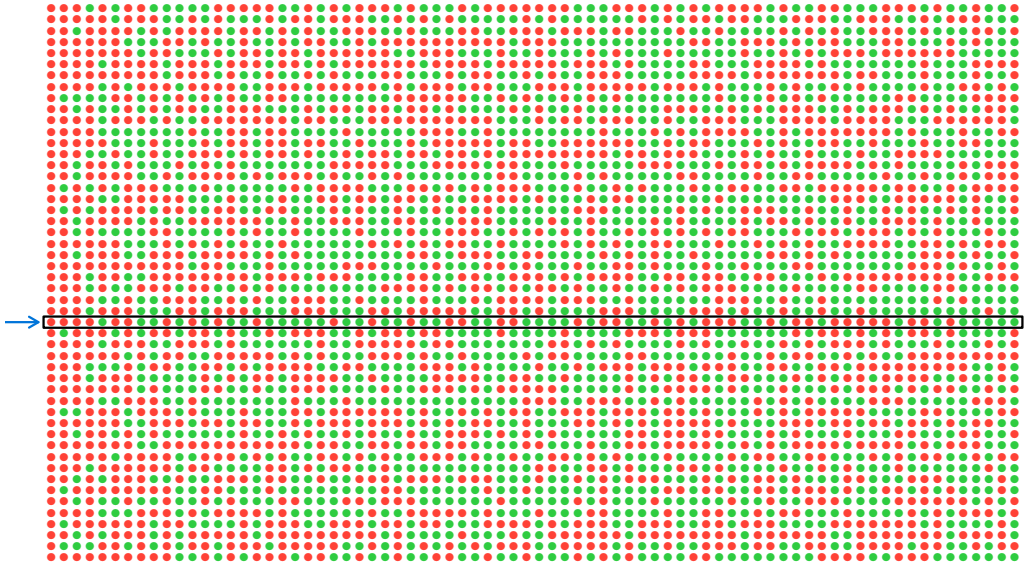
MCMC in action

MCMC in action

MCMC in action



MCMC in action



The conditional probabilities

If $y_i > 0$, then $\Pr(\theta_i = 1 \mid \boldsymbol{y}, \boldsymbol{\theta}_{\text{-i}}^{(\text{s})}) = 1$

If $y_i = 0$,

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

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

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

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: $\mathsf{E}(\theta_+ \mid \mathbf{y})$

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

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

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

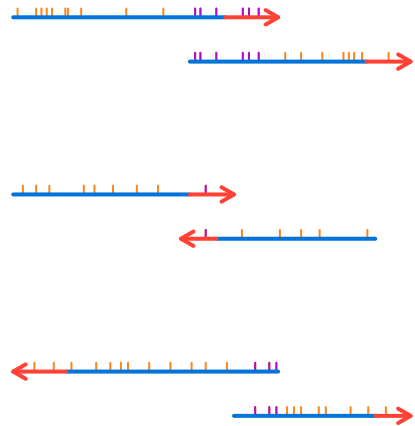
or Rao-Blackwellize:

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

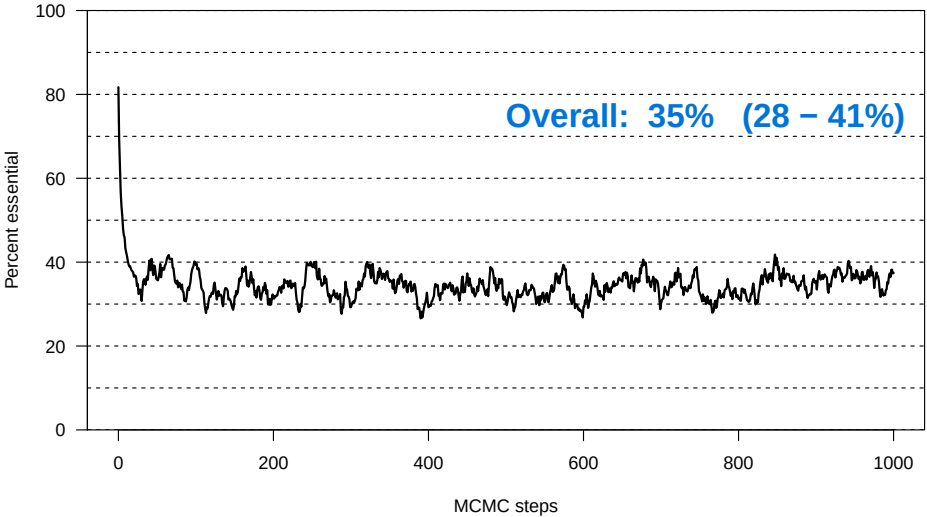
A further complication

Many genes overlap

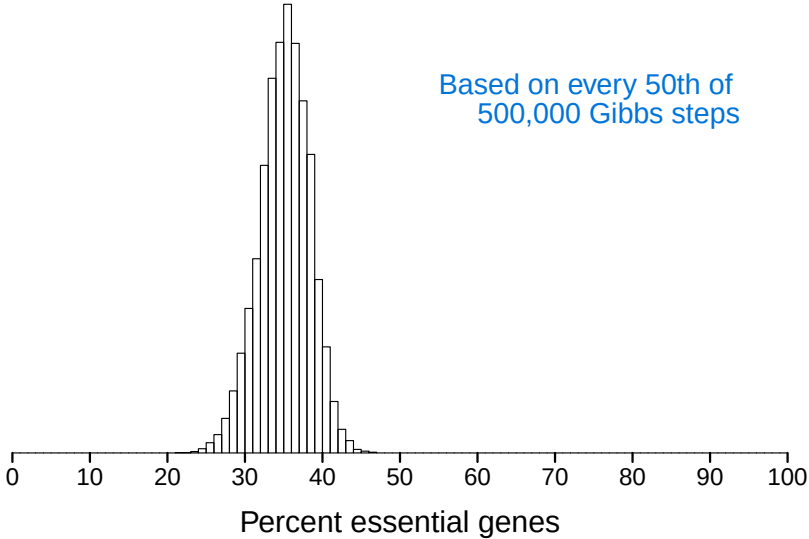
- Of 4250 genes, 1005 pairs overlap (mostly by exactly 4 bp).
- The overlapping regions contain 547 insertion sites.
- Omit TA sites in overlapping regions, unless in the proximal portion of *both* genes.
- The algebra gets a bit more complicated.



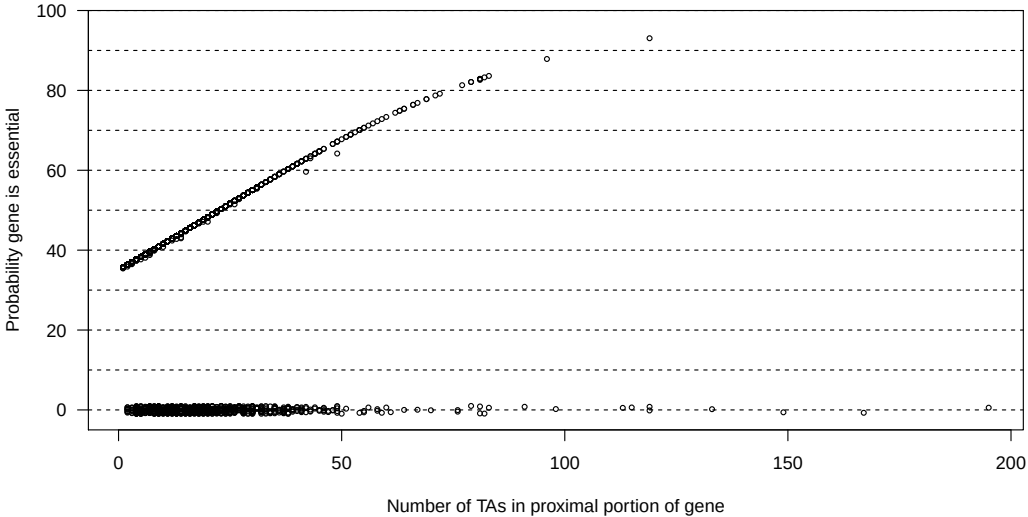
Percent essential genes in M. tb.



Percent essential genes in M. tb.



Probability each gene is essential



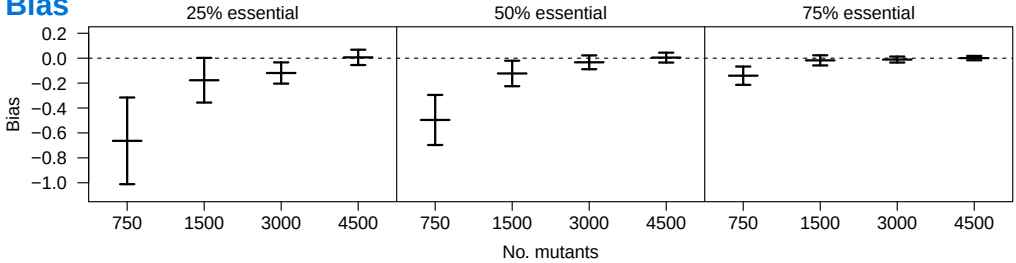
Yet another complication

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

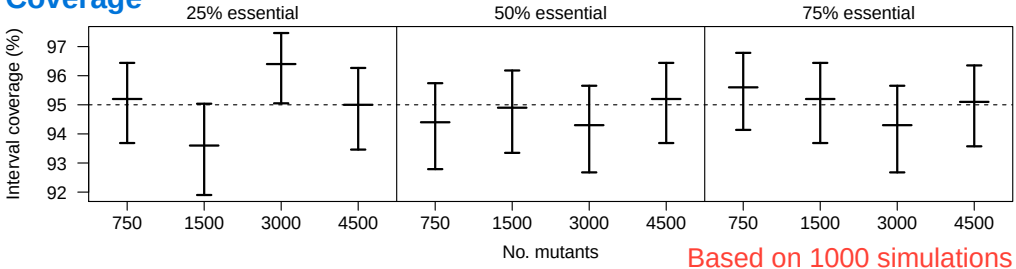
- Insertion at a TA site could disrupt all downstream genes
- If a gene is essential, insertion in any upstream gene would be non-viable
- Re-define the meaning of “essential gene”.
- If operons were known, one could get an improved estimate of the proportion of essential genes.
- If one ignores the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

Frequentist properties

Bias



Coverage



Based on 1000 simulations

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.

References

- Lamichhane et al. (2003) Proc Natl Acad Sci USA 100:7213-7218 doi:10.1073/pnas.1231432100
- Blades and Broman (2002) Tech Report MS02-20 bit.ly/ms0220
- R/negenes package cran.r-project.org/package=negenes