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

About me

- BS in Mathematics, U. Wisconsin-Milwaukee
  - A good amount of chemistry
  - NSF summer research experience U. Tennessee-Knoxville
- PhD in Statistics, U. California, Berkeley
  - Advisor: Terry Speed
  - Took biochemistry — my last real course
  - Friend who was a postdoc in dog genetics
- Postdoc, Marshfield Medical Research Fdn (Wisconsin)
  - Advisor: Jim Weber
  - Large genotyping facility
- Biostatistics, Johns Hopkins Bloomberg School of Public Health

What is statistics?

We may at once admit that any inference from the particular to the general must be attended with some degree of uncertainty, but this is not the same as to admit that such inference cannot be absolutely rigorous, for the nature and degree of the uncertainty may itself be capable of rigorous expression.

— Sir R. A. Fisher

What is statistics?

- Data exploration and analysis
- Inductive inference with probability
- Quantification of uncertainty
- Experimental design

A comparison

Mathematics vs. Statistics vs. Biostatistics

How I spend my time

25%	Reading, writing, reviewing
15%	Programming
15%	Teaching
15%	Analyzing data
10%	Talking to people about data
20%	Making and drinking coffee

- Gene mapping in mice, rats, humans, dogs
- Other oddball stuff (like what I'm taking about today)

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

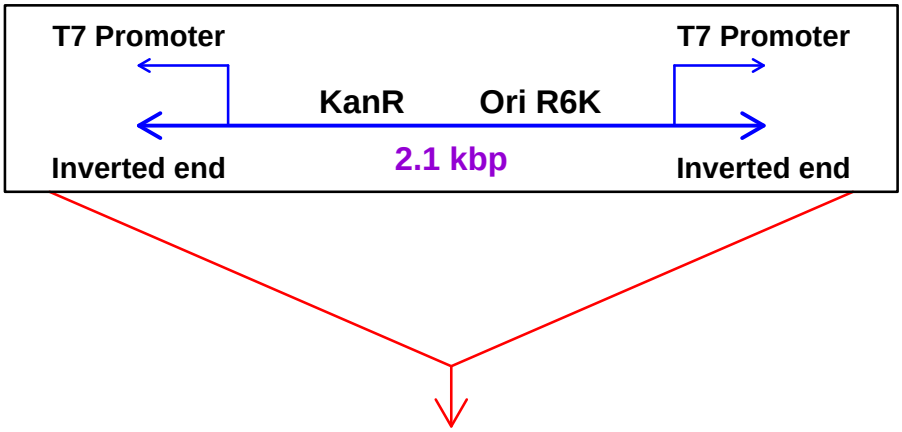
Identify the essential genes

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

Method

Random transposon mutagenesis

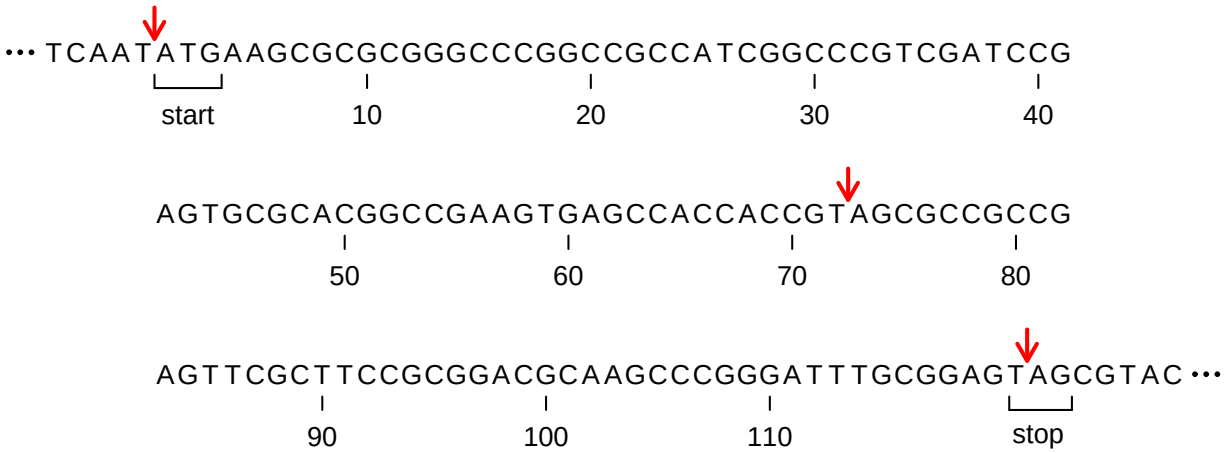
Himar1, a mariner-derived transposon



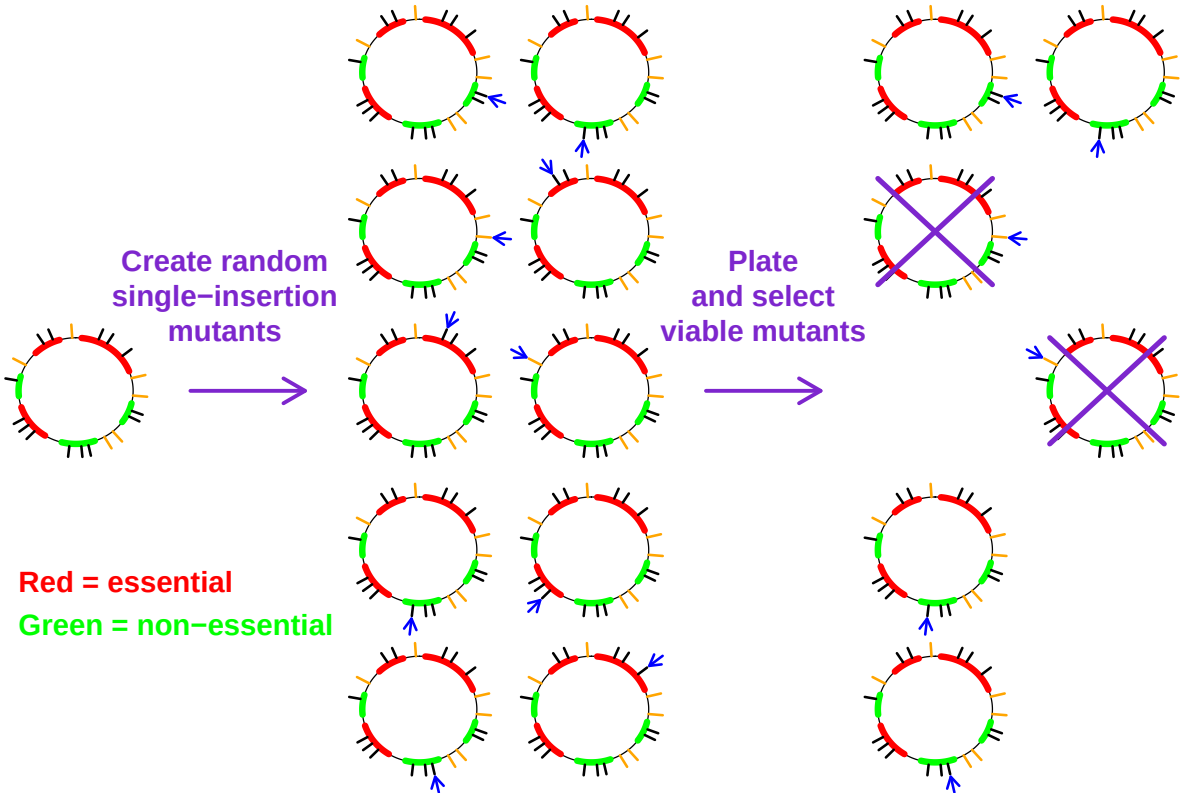
5′-TCGAAGCCTGCGAC**TA**ACGTT**TA**AAGTTTG-3′
3′-AGCTTCGGACGCTG**AT**TGCAA**AT**TTCAAAC-5′

Note: ≥ 30 stop codons in each reading frame

Sequence of the gene MT598



Random transposon mutagenesis



Random transposon mutagenesis

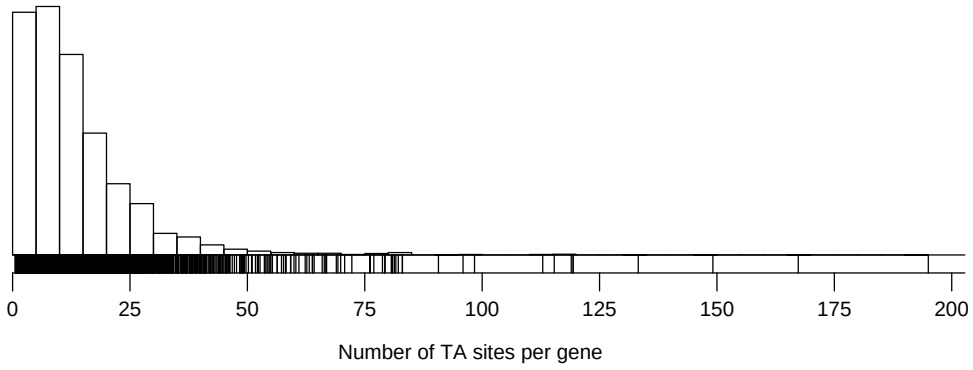
- Location of transposon insertion determined by sequencing across junctions
- Viable insertion within a gene $\Longrightarrow$ gene is non-essential
- Essential genes: we will never see a viable insertion
- Complication: Insertions in the very distal portion of an essential gene may not be sufficently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

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

The data

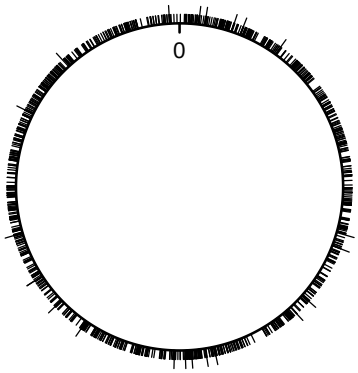
- Number, locations of genes.
- Number of insertion sites in each gene.
- n viable mutants with exactly one transposon insertion.
- Location of the transposon insertion in each mutant.

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?

(i.e., what would we see if we had 10^{100} mutants?)

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.

The model

- Transposon inserts completely at random
  (Each TA site equally likely to be hit)
- Genes are either completely essential or completely non-essential.

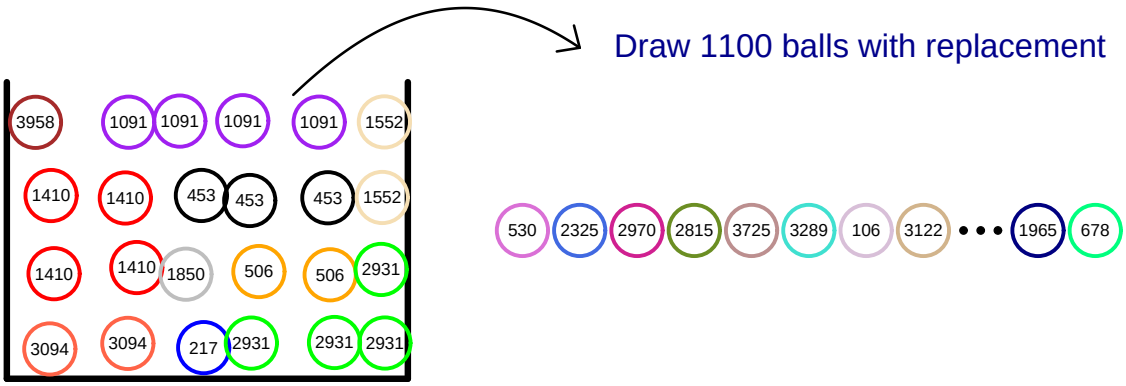
N genes x_i = no. TA sites in gene i

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

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

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

A picture of the model



Urn with balls labelled 1−4204

If essential: 0 balls

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

Part of the data

gene	no. TA sites	no. mutants
1	31	0
2	29	0
3	34	1
4	3	0
5	39	0
⋮	⋮	⋮
21	11	0
22	49	2
23	20	0
24	1	0
25	12	0
⋮	⋮	⋮
4204	4	0
total	**57934**	**1025**

A related problem

How many species of insects are there in the Amazon?

- Sample n insects at random.
- Classify according to species.
- How many total species exist?

My problem is a lot easier!

- Have a bound on the total number of classes.
- Know the relative proportions (up to a set of 0/1 factors).

Statistics, Part 2

Find an estimate of $\boldsymbol{\theta}$.

We're especially interested in $\theta_+ = \sum_i \theta_i$ and 1 – θ_+/N.

Frequentist approach

- View the parameters $\{\theta_i\}$ as fixed, unknown values.
- Find some estimate (function of the [random] data) that has good properties.
- Think about repeated realizations of the random process.

Bayesian approach

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

The likelihood

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

$$= \binom{n}{\mathbf{y}} \prod_i (x_i\ \theta_i)^{y_i} \ / \ (\sum_j x_j\ \theta_j)^n$$

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

Note: Depends only on which $y_i > 0$, and not directly on the particular values of y_i.

Frequentist method

Maximum likelihood estimates (MLEs):

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

$$L(\boldsymbol{\theta} \mid \mathbf{y}) \propto \begin{cases} (\sum_i x_i\, \theta_i)^{-n} & \text{if } \theta_i = 1 \text{ whenever } y_i > 0 \\ \\ 0 & \text{otherwise} \end{cases}$$

And so the MLEs are

$$\hat{\theta}_i = \begin{cases} 1 & \text{if } \ y_i > 0 \\ 0 & \text{if } \ y_i = 0 \end{cases}$$

Further, $\hat{\theta}_+ = \sum_i 1\{y_i > 0\}$.

This is a rather stupid estimate!

Bayes: The prior

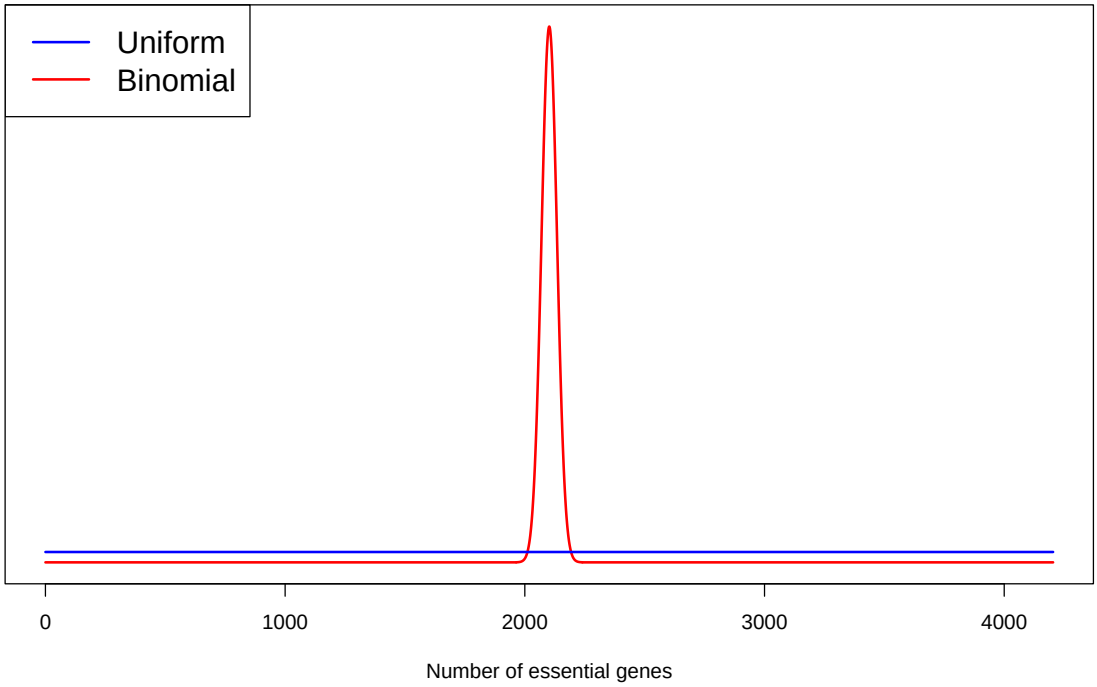
$\theta_+ \sim$ uniform on $\{0, 1, \ldots, n\}$

$\boldsymbol{\theta} \mid \theta_+ \sim$ 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).

Uniform vs. Binomial



A Gibbs sampler

Goal: Estimate $\Pr(\boldsymbol{\theta}|\mathbf{y}) = \Pr(\mathbf{y} \mid \boldsymbol{\theta}) \Pr(\boldsymbol{\theta}) / \sum_{\boldsymbol{\theta}} \Pr(\mathbf{y} \mid \boldsymbol{\theta}) \Pr(\boldsymbol{\theta})$

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 $\boldsymbol{\theta}_{-i}^{(s)} = (\theta_1^{(s+1)}, \ldots, \theta_{i-1}^{(s+1)}, \theta_{i+1}^{(s)}, \ldots, \theta_n^{(s)})$.
  - 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 $y_i > 0$, then $\Pr(\theta_i = 1 \mid \mathbf{y}, \boldsymbol{\theta}_{-i}^{(s)}) = 1$

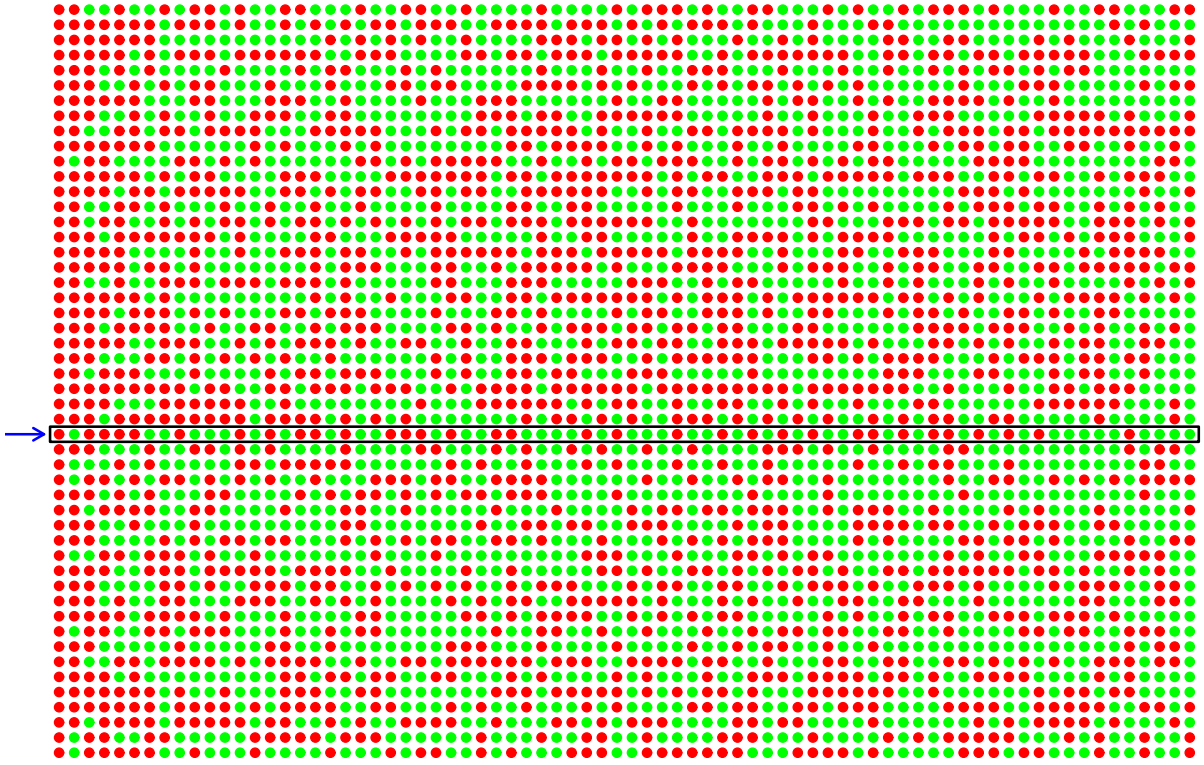
If $y_i = 0$,

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

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

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

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: $E(\theta_i \mid \mathbf{y}) = \Pr(\theta_i = 1 \mid \mathbf{y})$

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

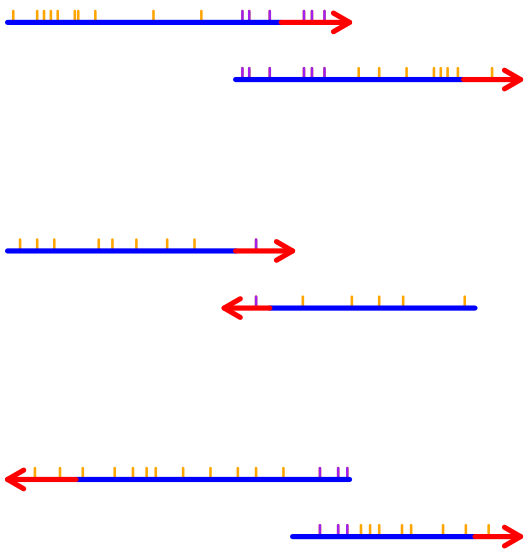
or Rao-Blackwellize:

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

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.

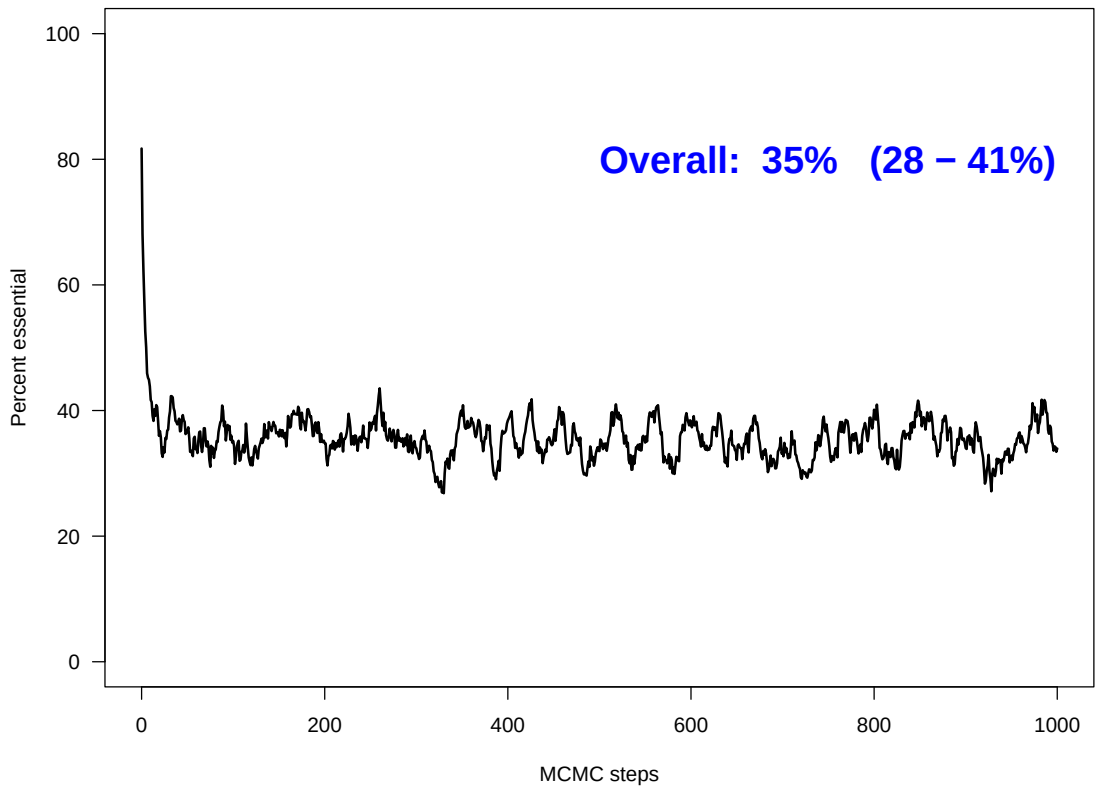


M. tb. mutagenesis data

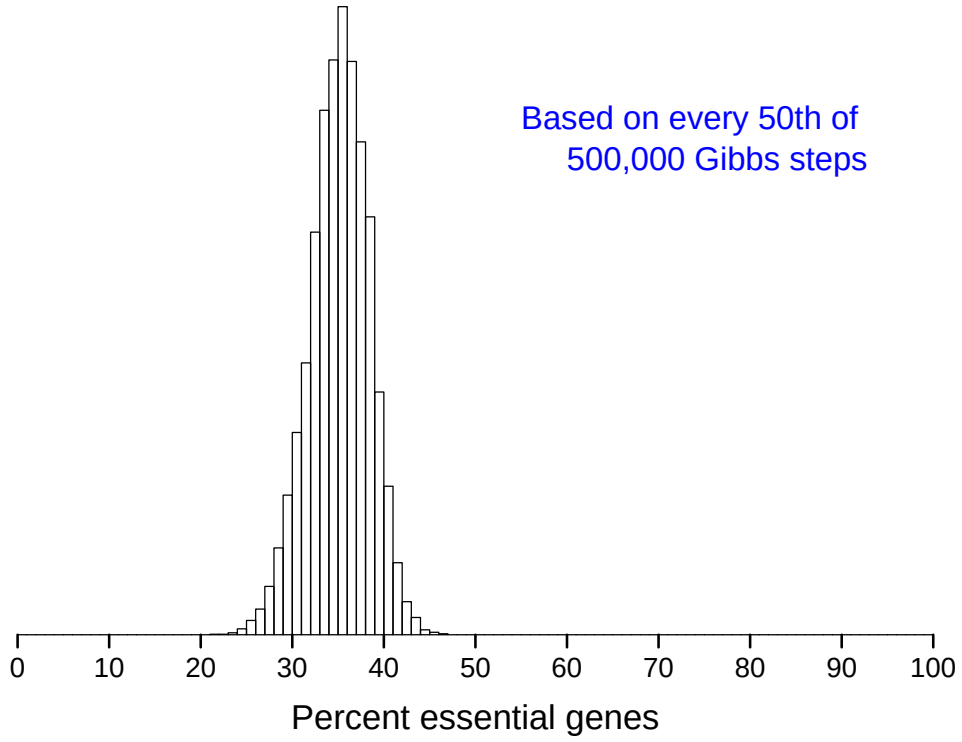
- 74,403 TA sites total
- 57,934 sites within proximal portion of a gene
- 77 sites shared by two genes
- 4204/4250 genes with at least one such site

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

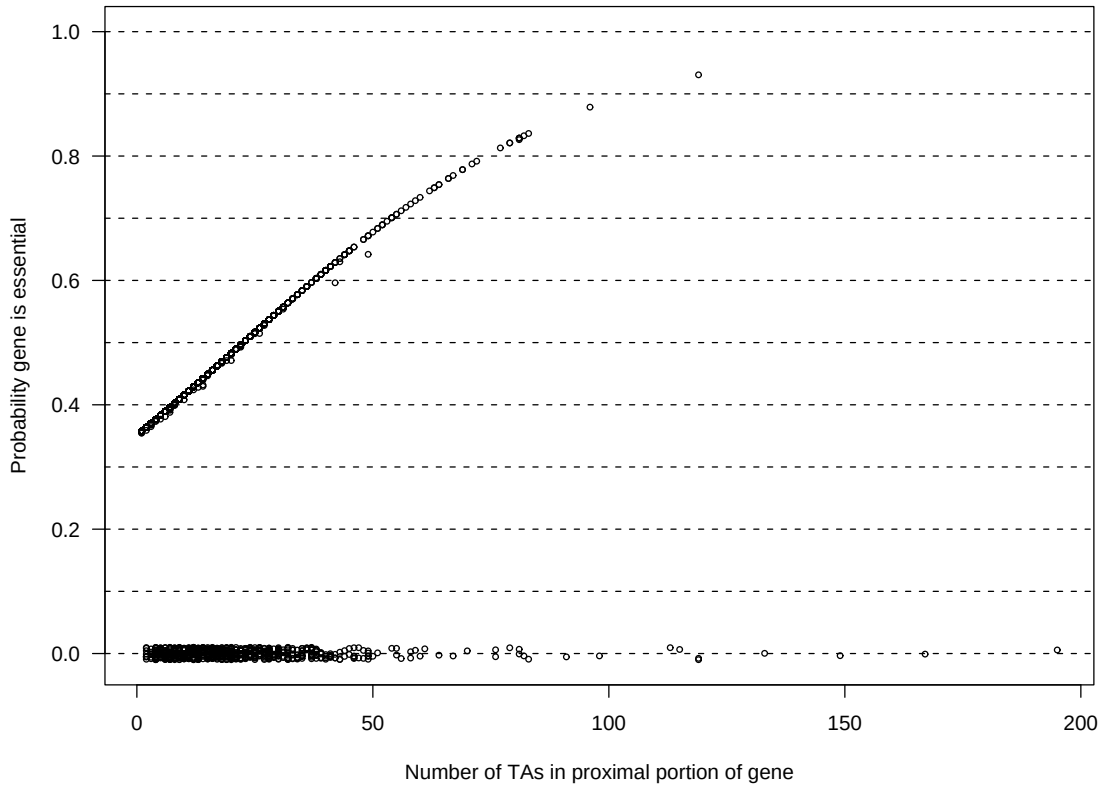
Percent essential genes in M. tb.



Percent essential genes in M. tb.



Probability that each gene is essential



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)

References

- Lamichhane G, Zignol M, Blades NJ, Geiman DE, Dougherty A, Broman KW, Bishai WR (2003) A post-genomic method for predicting essential genes at subsaturation levels of mutagenesis: application to Mycobacterium tuberculosis. Proc Natl Acad Sci USA 100:7213–7218.

  The scientific paper.

- Blades NJ, Broman KW (2002) Estimating the number of essential genes in a genome by random transposon mutagenesis. Technical Report MS02-20, Department of Biostatistics, Johns Hopkins University, Baltimore, MD.

  A technical report with the gory details.

- Carlin BP, Louis TA (2000) Bayes and empirical Bayes methods for data analysis, 2nd edition. CRC Press.

  A good textbook on Bayesian statistics.

- Gelman A, et al. (2003) Bayesian data analysis, 2nd edition. CRC Press.

  Another good textbook on Bayesian statistics; an especially good chapter on Markov chain Monte Carlo.

- Bunge J, Fitzpatrick M (1993) Estimating the number of species: A review. Journal of the American Statistical Association 88(421):364–373.

  A good place to start regarding the number of species problem.