

Identification of the **essential** genes in the *M. tuberculosis* genome 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

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

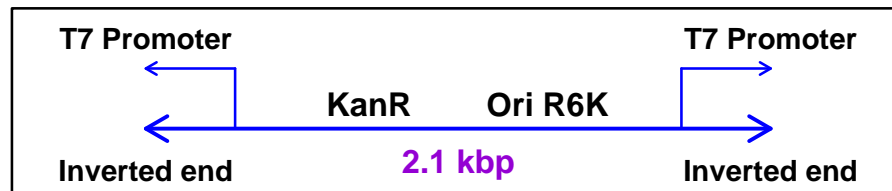
Aim

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: ≥ 30 stop codons in each reading frame

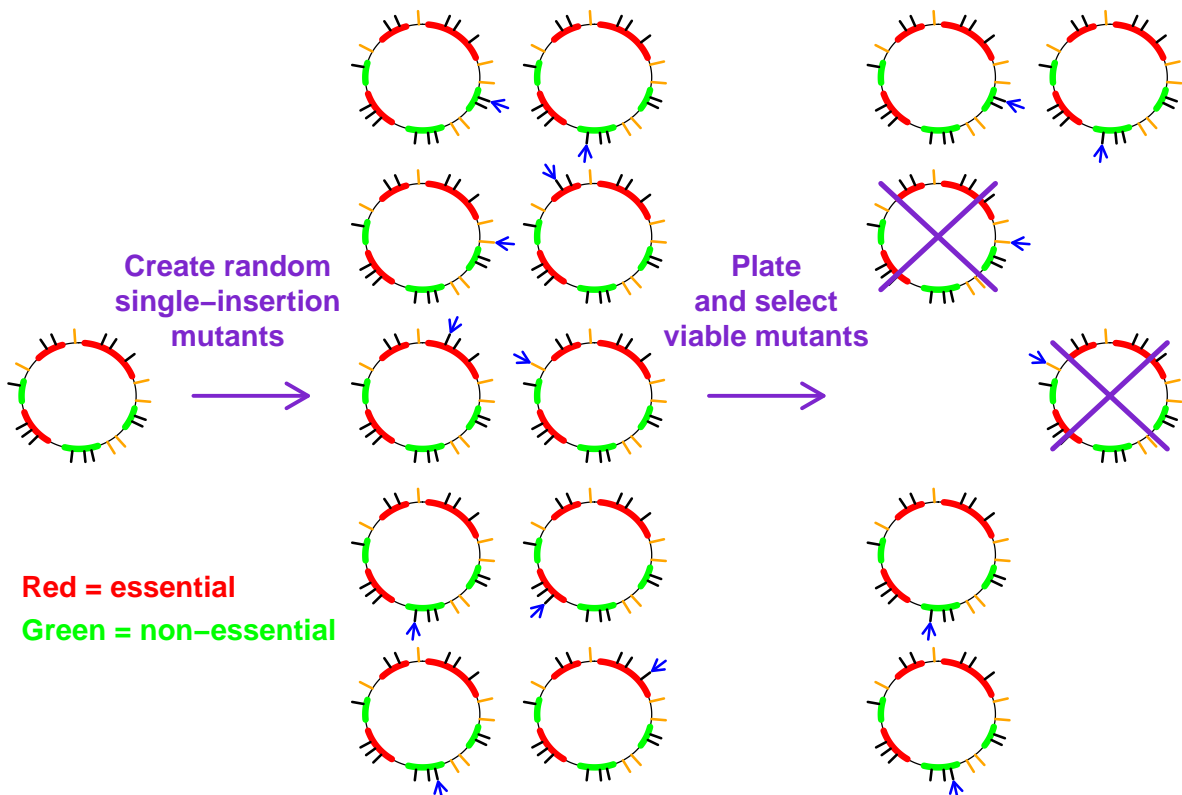
Sequence of the gene MT598

... TCAATATGAAGCGCGCGGGCCCGGCCGCCATCGGCCCGTCGATCCG
 | | | | |
 start 10 20 30 40

AGTGCACACGGCCGAAGTGAGCCACCACCGTAGCGCCGCGG
 | | | | |
 50 60 70 80

AGTTCGCTTCCGCGGACGCAAGCCCGGGATTTGCGGAGTAGCGTAC ...
 | | | | |
 90 100 110 stop

Random transposon mutagenesis



Random transposon mutagenesis

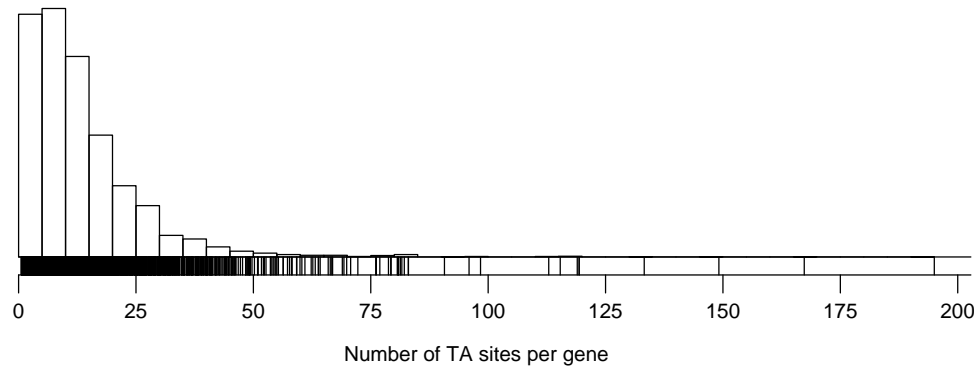
- Location of transposon insertion determined by sequencing across junctions
- Viable insertion within a gene \implies gene is non-essential
- Essential genes: we will never see a viable insertion
- **Complication:** Insertions in the very distal portion of an essential gene may not be sufficiently disruptive.

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

The data

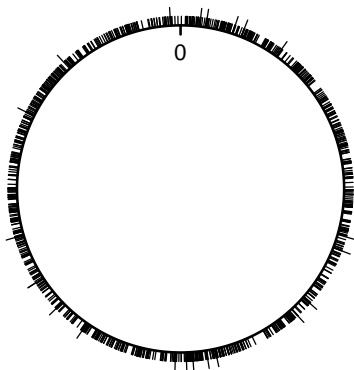
- Number, locations of genes.
- Number of insertion sites in each gene.
- n viable mutants with exactly one transposon insertion.
- Location of the transposon insertion in each mutant.

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?
(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.

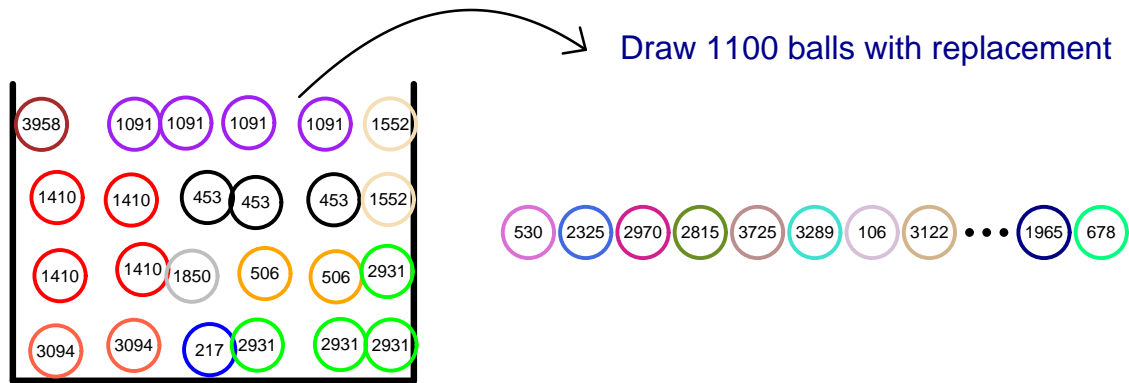
N genes x_i = no. TA sites in gene i

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

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

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

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 θ .

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

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

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

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

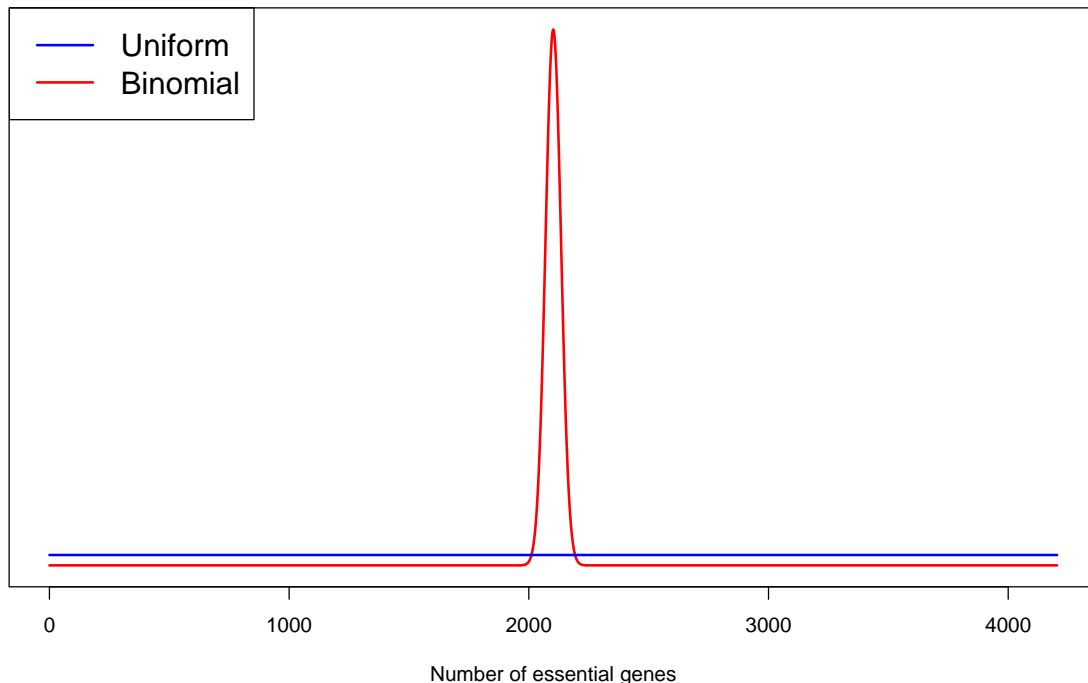
$\theta_+ \sim$ uniform on $\{0, 1, \dots, n\}$

$\theta \mid \theta_+ \sim$ 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 or non-essential status of particular genes (e.g., known viable knock-outs).

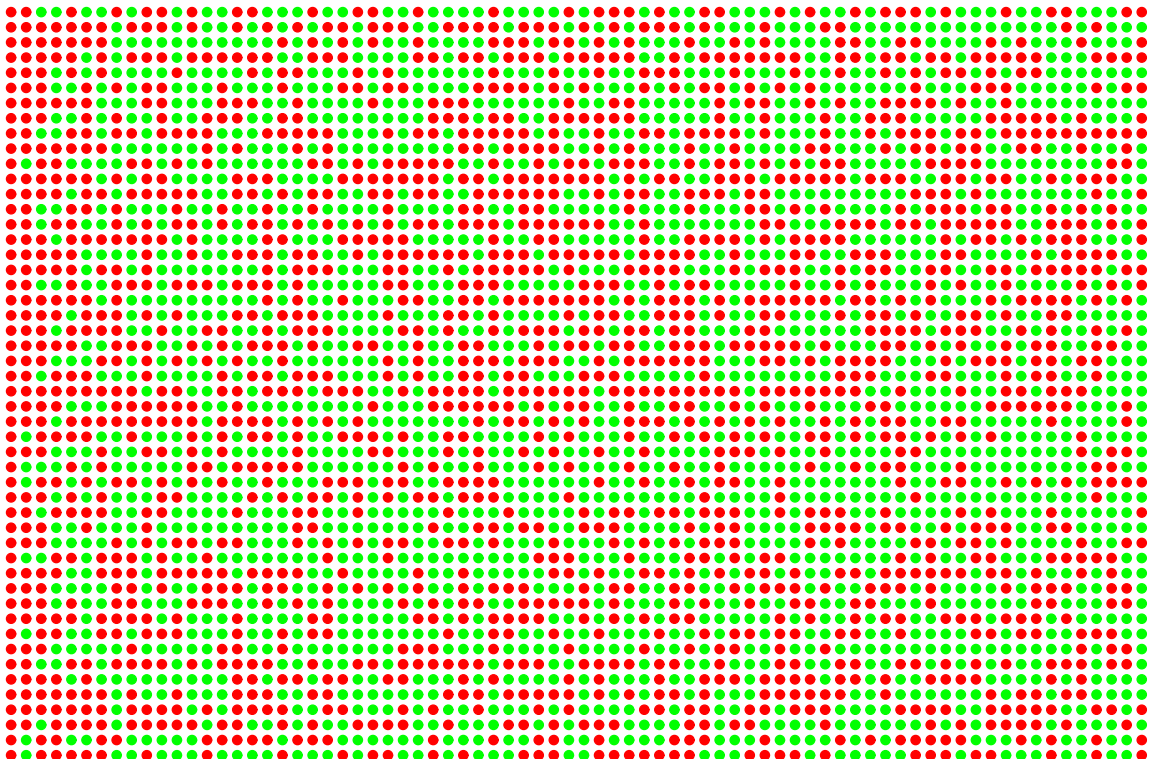
Uniform vs. Binomial



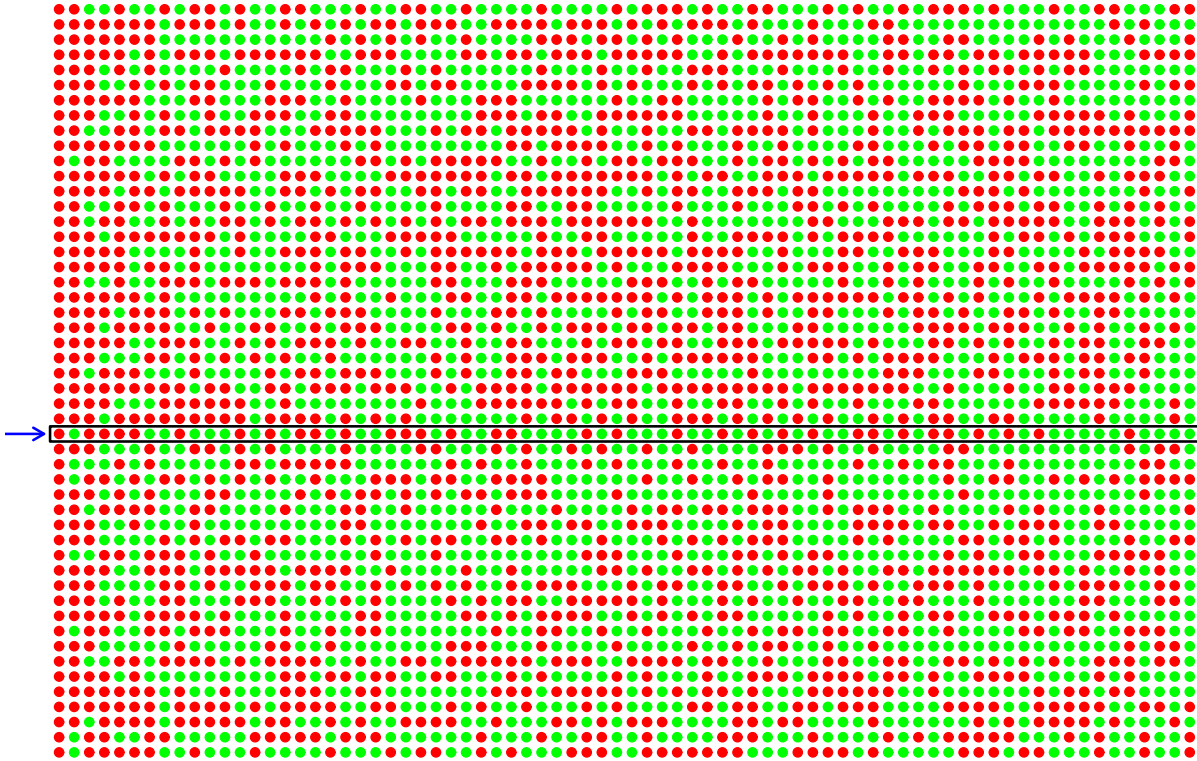
MCMC algorithm

- Begin with initial assignment of essential status of each gene
- Consider each gene, one at a time
 - Calculate $\Pr(\text{gene is ess'l} \mid \text{data, status of other genes})$ ← Depends on:
 - No. mutants
 - No. TA sites in gene
 - Total no. viable TA sites
 - No. essential genes
 - Randomly assign it to be essential or non-ess'l according to this probability
- Repeat many times
- Summarize results

MCMC in action



A particular gene



Estimators

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

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

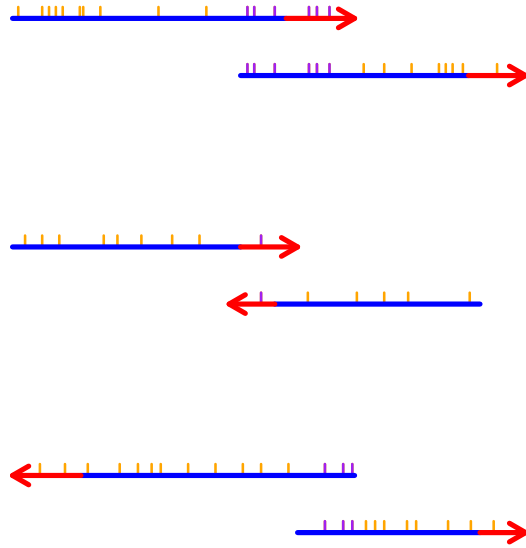
or Rao-Blackwellize:

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

A further complication

Many genes overlap

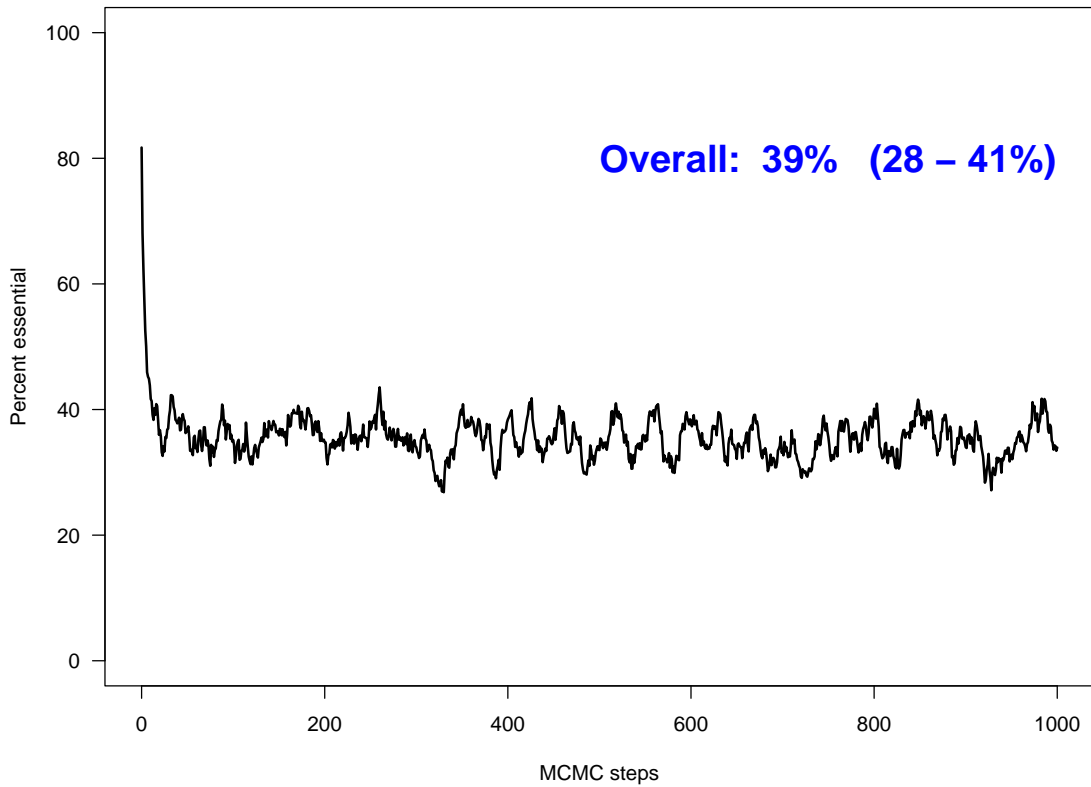
- Of 4250 genes, 1005 pairs overlap (mostly by exactly 4 bp).
- The overlapping regions contain 547 insertion sites.
- Omit TA sites in overlapping regions, unless in the proximal portion of *both* genes.
- The algebra gets a bit more complicated.



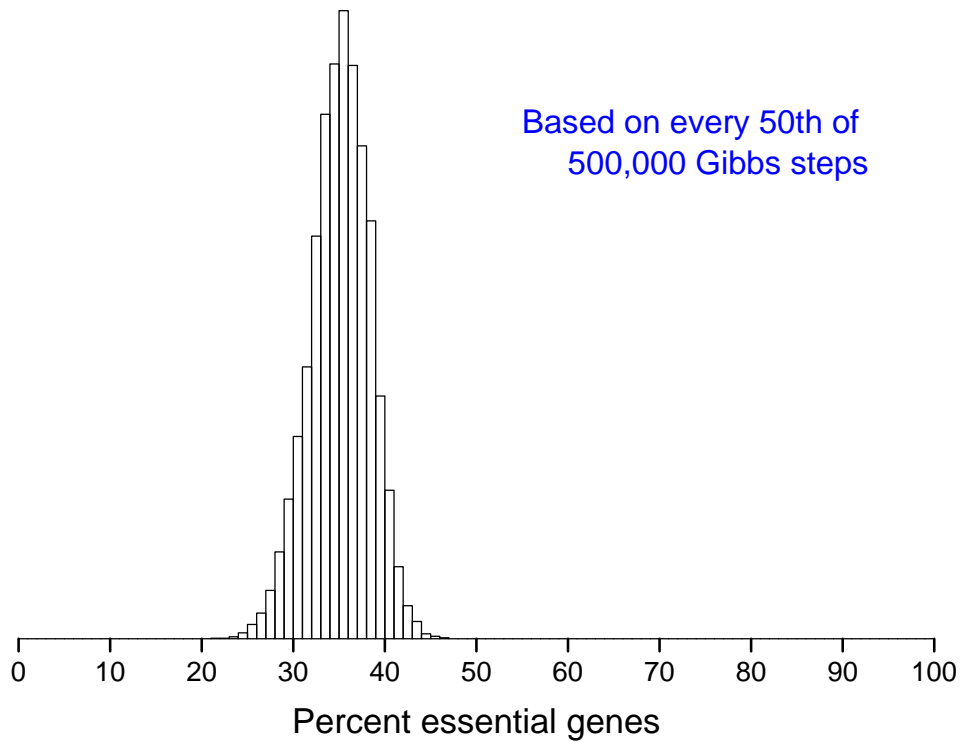
M. tb. mutagenesis data

- 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

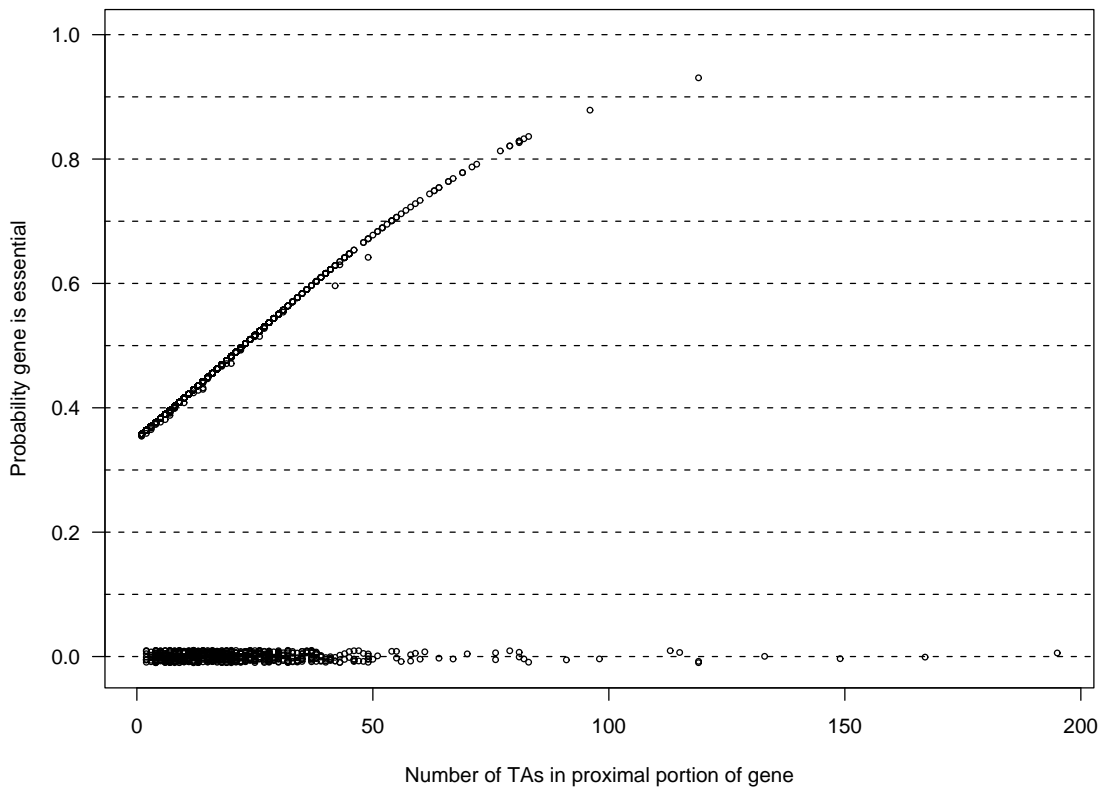
Percent essential genes in M. tb.



Percent essential genes in M. tb.



Probability that each gene is essential



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 ≥ 64 TA sites and for which no mutant has been observed, have $> 75\%$ chance of being essential.

Acknowledgements



Bill Bishai



Natalie Blades



Gyanu Lamichhane

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.