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 $\Rightarrow$ non-viable mutant)

**Method**

Random transposon mutagenesis
**Himar1**, a mariner-derived transposon

![Diagram of Himar1 transposon](image)

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**

```
... TCAATTGAAAGCGCGCGGGGCCGCGCGCGCATCGGCGCGTCGATCCG
     start   10    20    30    40

AGTGCACGGCCGAAGTGAACACCACCGTACGCGCGCGCG
     50   60   70   80

AGTTGCGCTTCCGCAGCAAGCCCGGAGTTTGCGGAGTACGTAC...
     90  100  110
```
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.

\[ N \text{ genes} \quad x_i = \text{no. TA sites in gene } i \]
\[ n \text{ mutants} \quad y_i = \text{no. mutants with insertion in gene } i \]

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

Model: \[ y \sim \text{multinomial}(n, p) \quad \text{where } p_i = \frac{x_i \theta_i}{\sum_j x_j \theta_j} \]

A picture of the model

Draw 1100 balls with replacement

Urn with balls labelled 1–4204

If essential: 0 balls
If non–essential: no. balls = no. TA sites
### Part of the data

<table>
<thead>
<tr>
<th>gene</th>
<th>no. TA sites</th>
<th>no. mutants</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>31</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>29</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>34</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>39</td>
<td>0</td>
</tr>
<tr>
<td>:</td>
<td>:</td>
<td>:</td>
</tr>
<tr>
<td>21</td>
<td>11</td>
<td>0</td>
</tr>
<tr>
<td>22</td>
<td>49</td>
<td>2</td>
</tr>
<tr>
<td>23</td>
<td>20</td>
<td>0</td>
</tr>
<tr>
<td>24</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>25</td>
<td>12</td>
<td>0</td>
</tr>
<tr>
<td>:</td>
<td>:</td>
<td>:</td>
</tr>
<tr>
<td>4204</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td><strong>total</strong></td>
<td><strong>57934</strong></td>
<td><strong>1025</strong></td>
</tr>
</tbody>
</table>

### 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).
Find an estimate of $\theta$.
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

$$L(\theta | y) = \Pr(y | \theta)$$

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

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

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

---

Statistics, Part 2
Frequentist method

Maximum likelihood estimates (MLEs):

Estimate the $\theta_i$ by the values for which $L(\theta \mid y)$ achieves its maximum.

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

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

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

$\theta \mid \theta_+ \sim \text{uniform over all sequences of 0's and 1's with } \theta_+ 1$'s.

Notes:

- We are assuming that $\Pr(\theta_i = 1) = 1/2$.
- This is quite different from taking $\theta_i$ iid Bernoulli(1/2).
- We are assuming that $\theta_i$ 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 knockouts).
A Gibbs sampler

Goal: Estimate \( \Pr(\theta | y) = \frac{\Pr(y | \theta) \Pr(\theta)}{\sum_{\theta} \Pr(y | \theta) \Pr(\theta)} \)

Gibbs sampler:

- Begin with some initial assignment, \( \theta^{(0)} \), ensuring that \( \theta_i^{(0)} = 1 \) whenever \( y_i > 0 \).

- For iteration \( s \), consider each gene one at a time, and let \( \theta_{-i}^{(s)} = (\theta_1^{(s+1)}, \ldots, \theta_{i-1}^{(s+1)}, \theta_i^{(s)}, \ldots, \theta_n^{(s)}) \).
  - Calculate \( \Pr(\theta_i = 1 | \theta_{-i}^{(s)}, y) \).
  - Assign \( \theta_i^{(s)} = 1 \) at random with this probability.

- Repeat many times.
The conditional probabilities

If \( y_i > 0 \), then \( \Pr(\theta_i = 1 \mid y, \theta_{-i}^{(s)}) = 1 \)

If \( y_i = 0 \),

Let \( A = \sum_{j < i} \theta_j^{(s+1)} + \sum_{j > i} \theta_j^{(s)} \)
\( B = \sum_{j < i} x_j \theta_j^{(s+1)} + \sum_{j > i} x_j \theta_j^{(s)} \)

Then \( \Pr(\Theta_{-i}^{(s)}, \theta_i = k) = \binom{n}{A+k}/n \)
\( \Pr(y \mid \Theta_{-i}^{(s)}, \theta_i = k) = (B + k x_i)^{-n} \)

And so \( \Pr(\theta_i = 1 \mid y, \theta_{-i}^{(s)}) = \ldots \)
\[ = \frac{(1 + x_i/B)^{-n}}{(1 + x_i/B)^{-n} + (n - A)/(A + 1)} \]

MCMC in action
The Gibbs sampler produces $\theta^{(0)}, \theta^{(1)}, \ldots, \theta^{(S)}$

We discard the first 200 or so samples (“burn-in”).

**Estimated number of non-essential genes:** $E(\theta_+ | y)$

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

**Probability that gene $i$ is non-essential:** $E(\theta_i | y) = \Pr(\theta_i = 1 | y)$

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

or Rao-Blackwellize:

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

**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.

Overall: 35% (28 – 41%)
Percent essential genes in M. tb.

Based on every 50th of 500,000 Gibbs steps

Probability that each gene is essential

Number of TAs in proximal portion of gene

Probability 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 ≥ 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

  The scientific paper.

  A technical report with the gory details.

  A good textbook on Bayesian statistics.

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

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