

# 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](http://www.biostat.jhsph.edu/~kbroman)

---

Joint work with Natalie Blades, Gyanu Lamichhane,  
and William Bishai

# Typical drug regimens

---

## Tuberculosis

- INH 15g
- RIF 37g
- PZA 141g
- ETB 151g
- ~60 DOT visits
- Cost: > \$15,000

## Other bacterial pneumonias

- Azithromycin 1.5g
- Self-supervised
- Cost: \$35

# *Mycobacterium tuberculosis* genome

---

- 4.4 Mbp circular genome, completely sequenced
- 4250 known or inferred genes
- 44% of genome has no match to mammals or other bacteria
- >250 lipid biosynthesis genes (E. Coli: ~50)
- Mycolic acids: unique, essential
- Cell division time: 24 hr

# Bacterial gene products

---

## Essential genes

- Cell division
- DNA replication
- Transcription
- Protein synthesis
- Cell wall formation

## Non-essential genes

- Virulence
- Stress response
- DNA modification
- Mobile elements
- Small molecular biosynthesis
- Regulatory genes

# Aim

---

Identify the essential genes  
(knock-out  $\Rightarrow$  non-viable mutant)

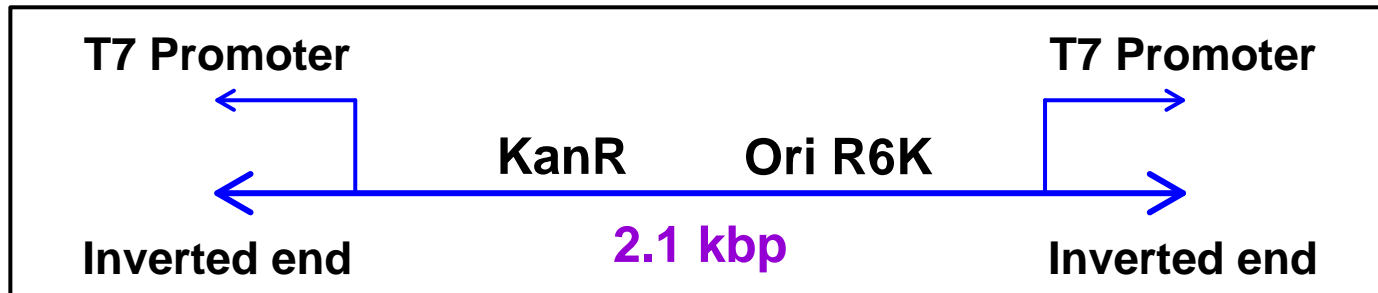
# Method

---

Random transposon mutagenesis

# *Himar1*, a mariner-derived transposon

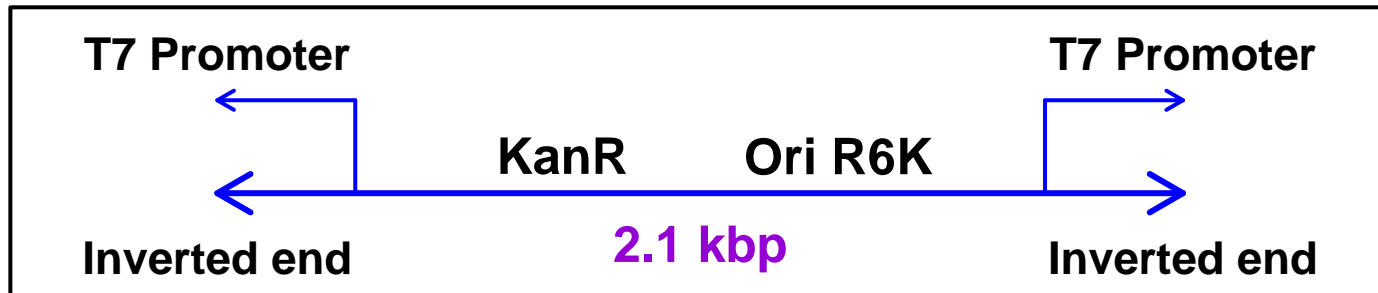
---



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

# *Himar1*, a mariner-derived transposon

---



5' -TCGAAGCCTGCGAC**TA**ACGTT**TA**AAGTTTG-3'  
3' -AGCTTCGGACGCTG**ATT**GCAA**ATT**TCAAAC-5'

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

... TCAATATGAAGCGCGCGGGCCCGGCCCATCGGCCCGTCGATCCG  
 |  
 start 10 20 30 40

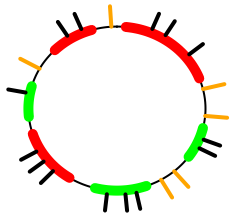
AGTGCGCACGGCCGAAGTGAGCCACCACCGTAGCGCCGCGCG  
 |  
 50 60 70 80

AGTTCGCTTCCGCGGACGCAAGCCCGGGATTTGCGGAGTAGCGTAC ...  
 |  
 90 100 110 stop



# Random transposon mutagenesis

---

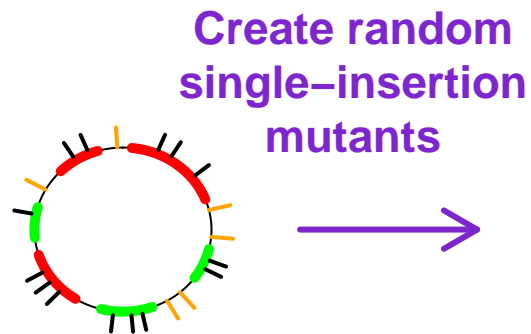


**Red = essential**

**Green = non-essential**

# Random transposon mutagenesis

---

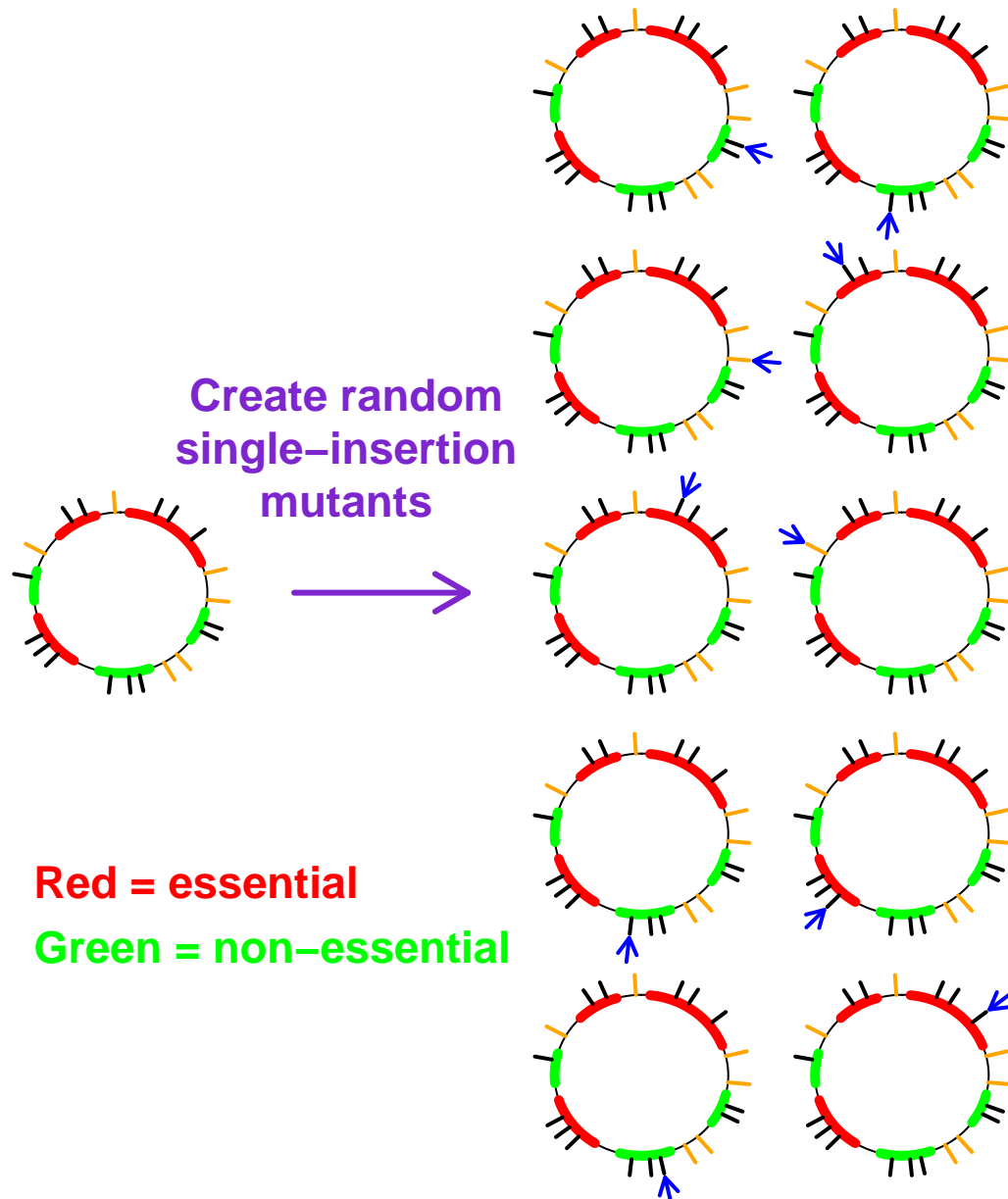


**Red = essential**

**Green = non-essential**

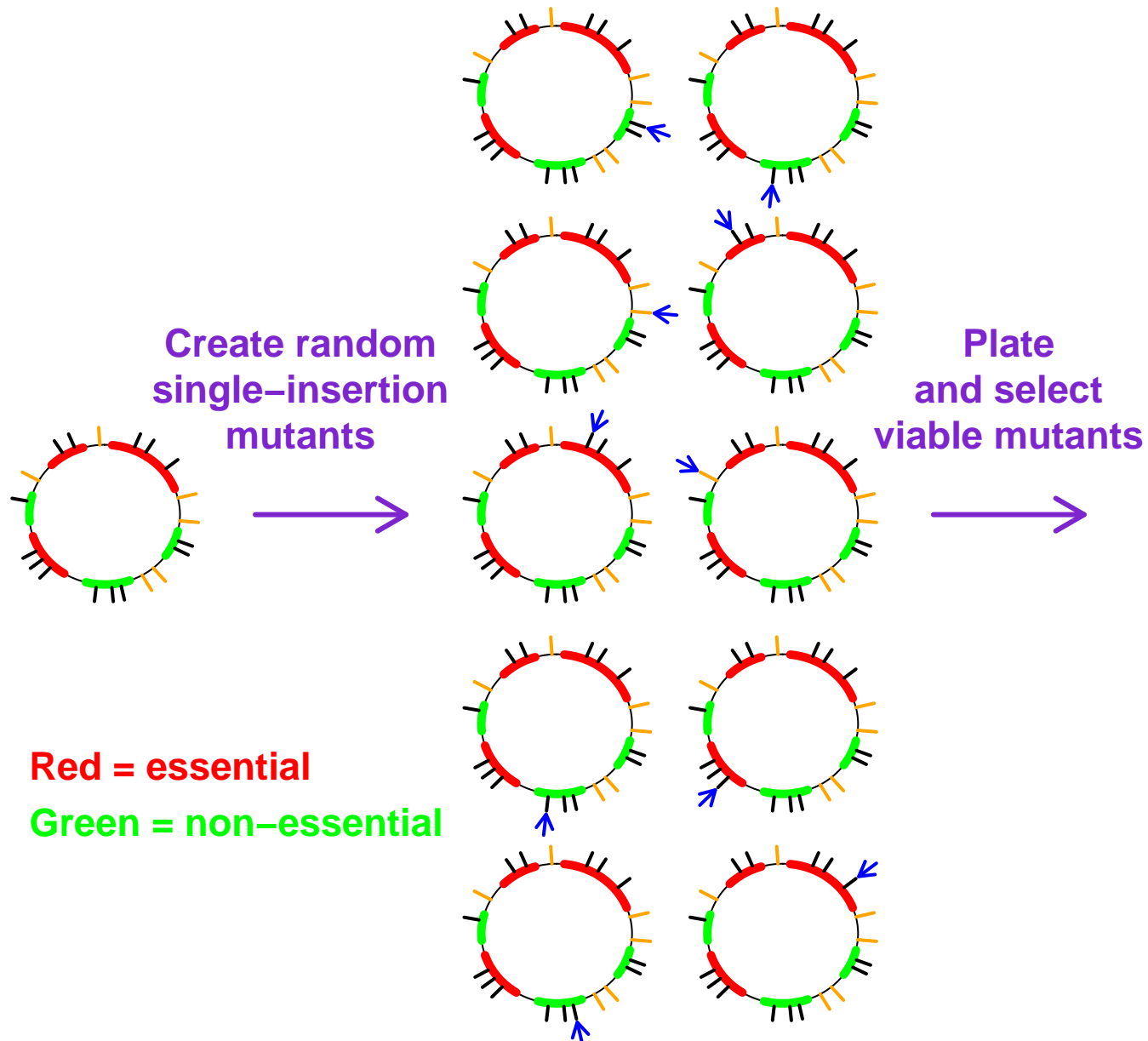
# Random transposon mutagenesis

---

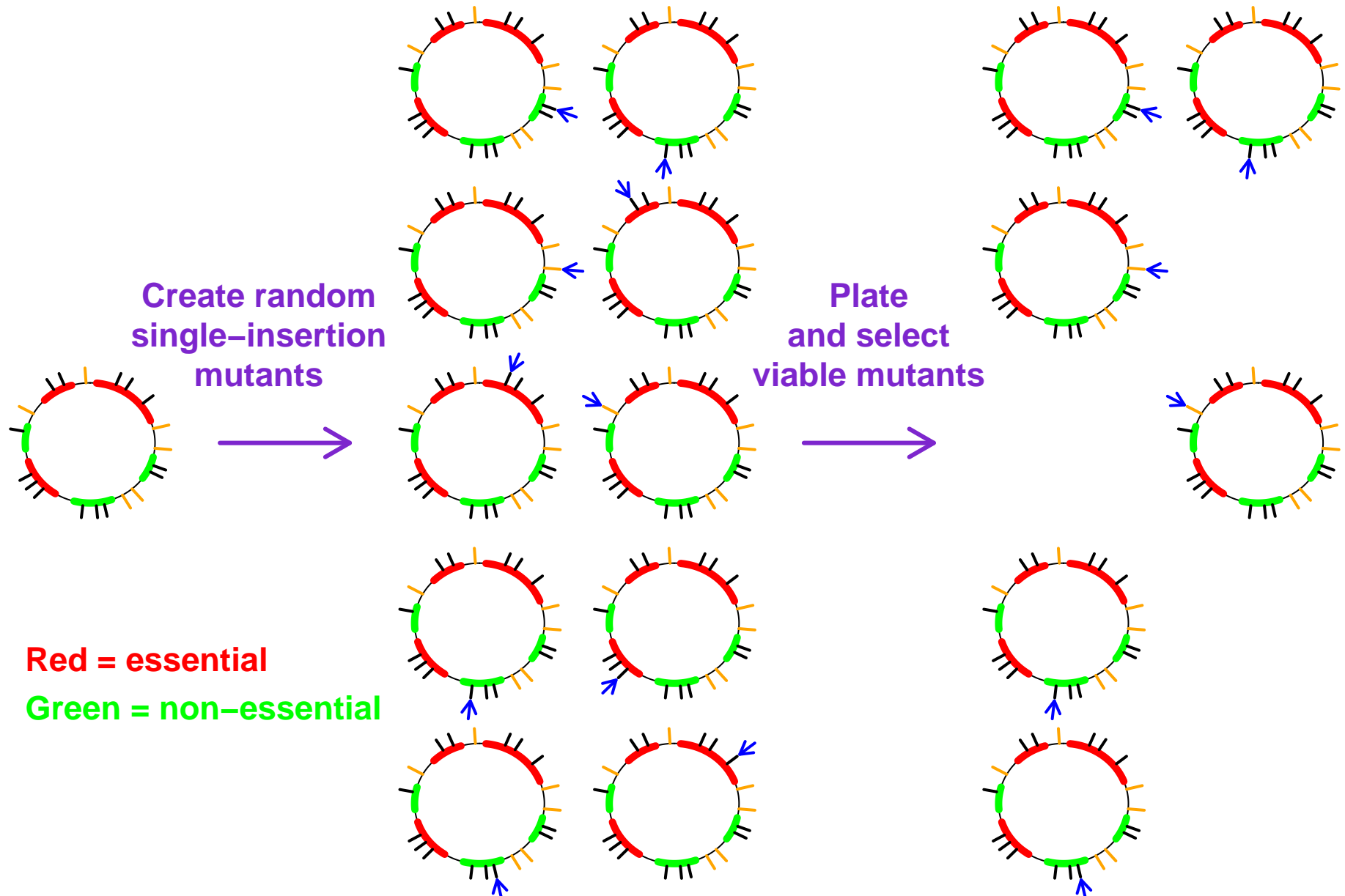


# Random transposon mutagenesis

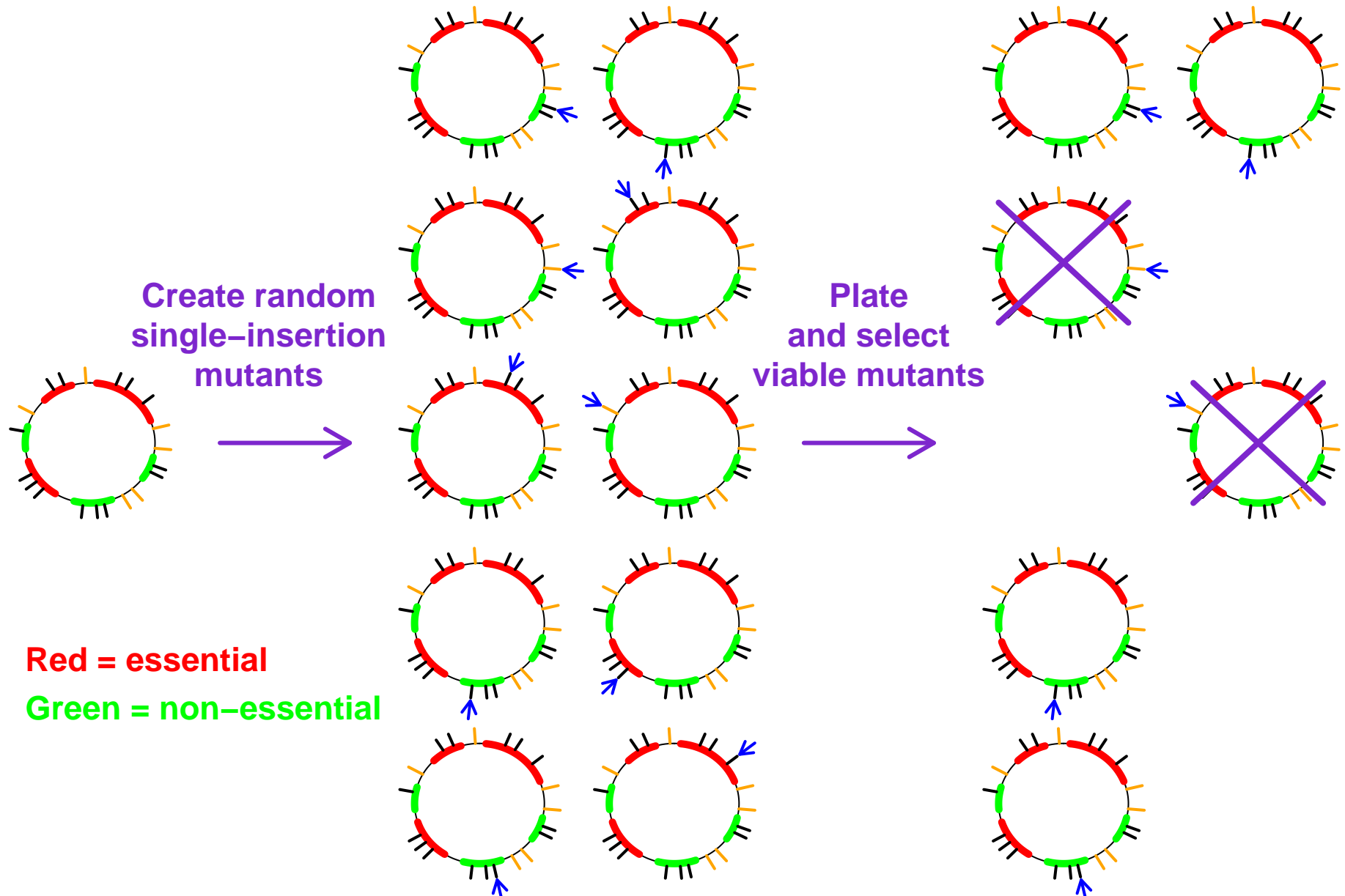
---



# Random transposon mutagenesis



# Random transposon mutagenesis



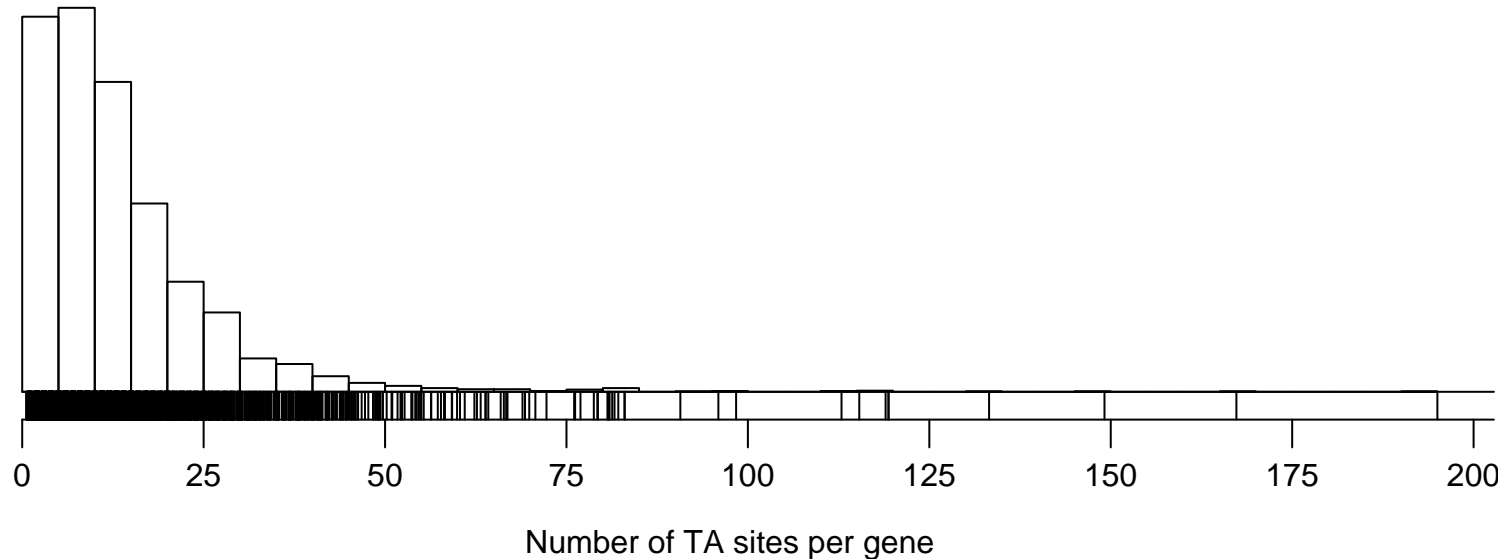
# Random transposon mutagenesis

---

- Location of transposon insertion determined by sequencing across junctions
- Viable insertion within a gene  $\implies$  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

# TA sites in *M. tuberculosis*

---

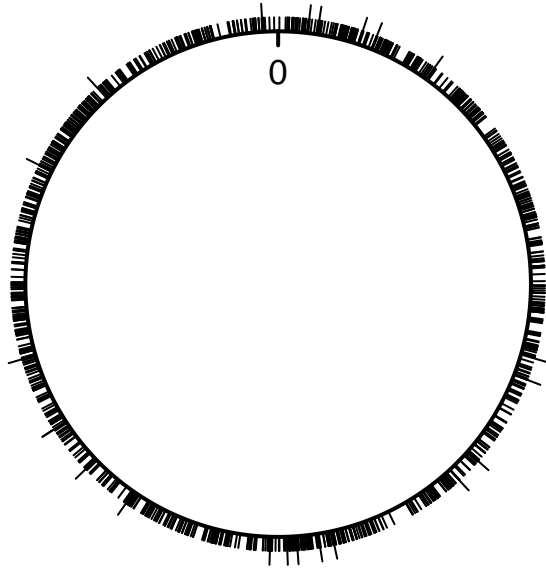


- 74,403 sites
- 65,649 sites within a gene
- 57,934 sites within proximal portion of a gene
- 4204/4250 genes with at least one TA site



# 1425 insertion mutants

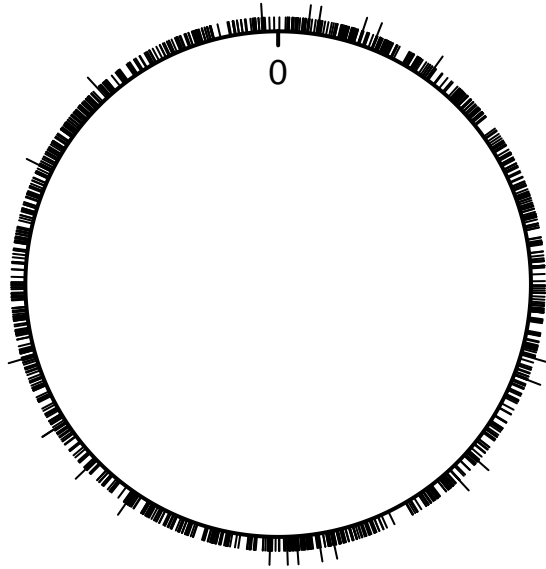
---



- 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. 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  $\sim \text{Uniform}\{0, 1, \dots, 4204\}$
- Given no. ess'l genes, each possible subset is equally likely

**Bayes by Markov chain Monte Carlo (MCMC):**

Approximate calculation of

- $\text{Pr}(\text{gene } i \text{ is essential} \mid \text{data})$
- Distribution of no. essential genes given the data

# MCMC algorithm

---

- Begin with initial assignment of essential status of each gene
- Consider each gene, one at a time
  - Calculate
$$\Pr(\text{gene is ess'l} \mid \text{data, status of other genes})$$
  - Randomly assign it to be essential or non-ess'l according to this probability
- Repeat many times
- Summarize results

# MCMC algorithm

---

- Begin with initial assignment of essential status of each gene
- Consider each gene, one at a time

- Calculate

- $\text{Pr}(\text{gene is ess'l} \mid \text{data, status of other genes})$

← Depends on:

- Randomly assign it to be essential or non-ess'l according to this probability
  - Repeat many times
  - Summarize results
- No. mutants
  - No. TA sites in gene
  - Total no. viable TA sites
  - No. essential genes

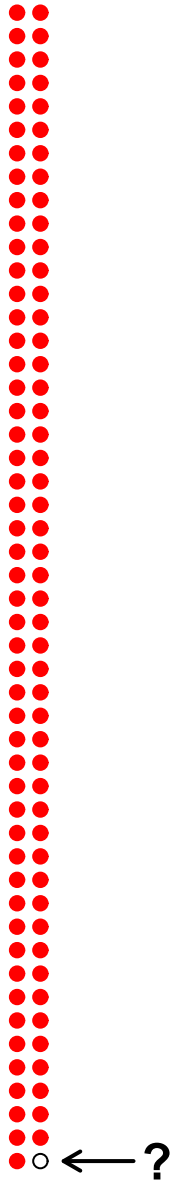
# MCMC in action

---



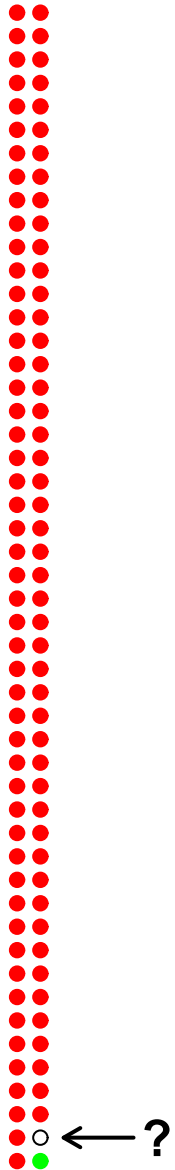
# MCMC in action

---



# MCMC in action

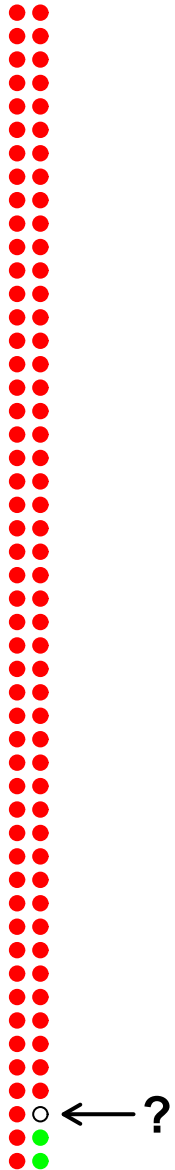
---





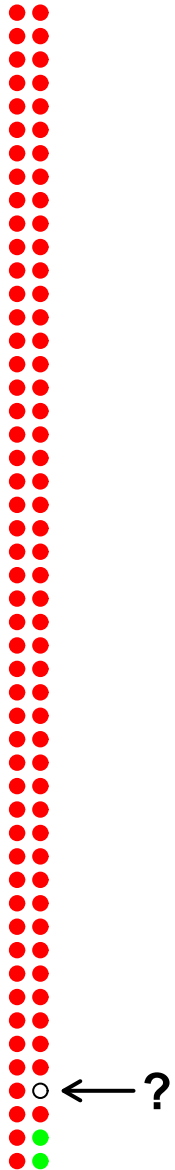
# MCMC in action

---



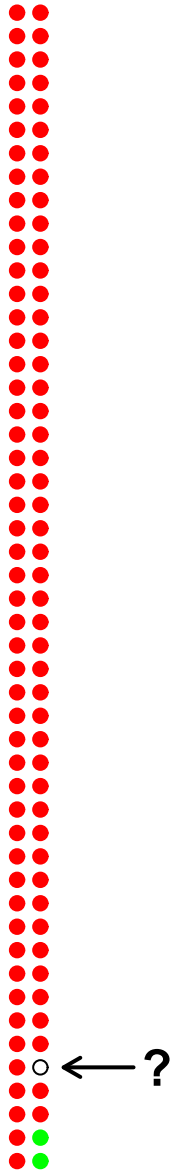
# MCMC in action

---



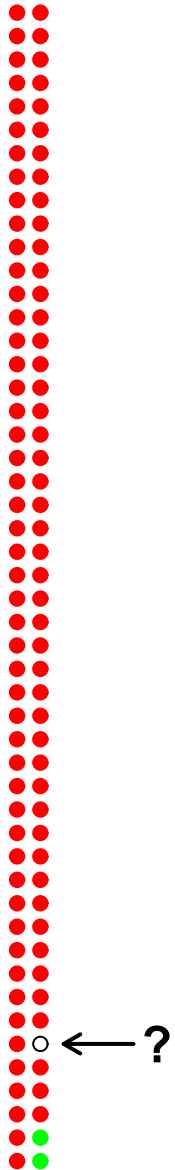
# MCMC in action

---

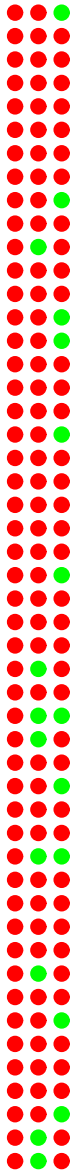


# MCMC in action

---

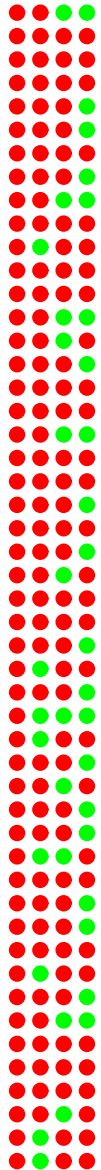






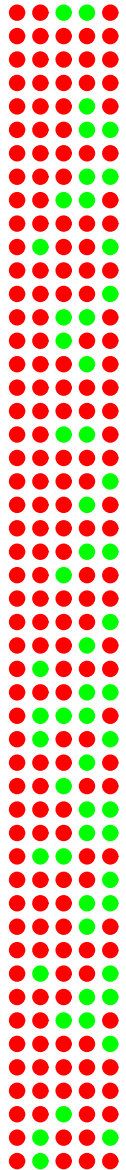
# MCMC in action

---



# MCMC in action

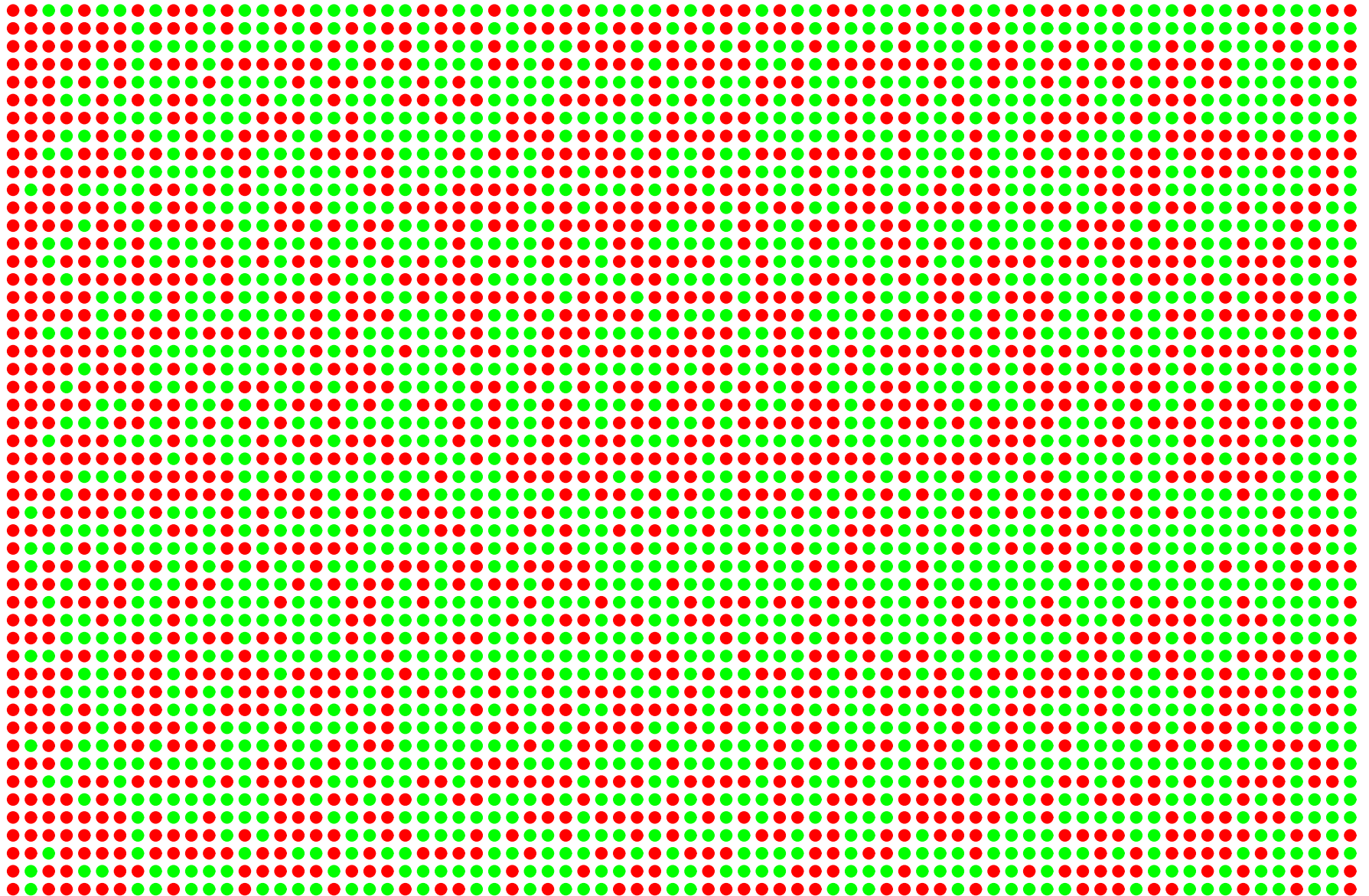
---





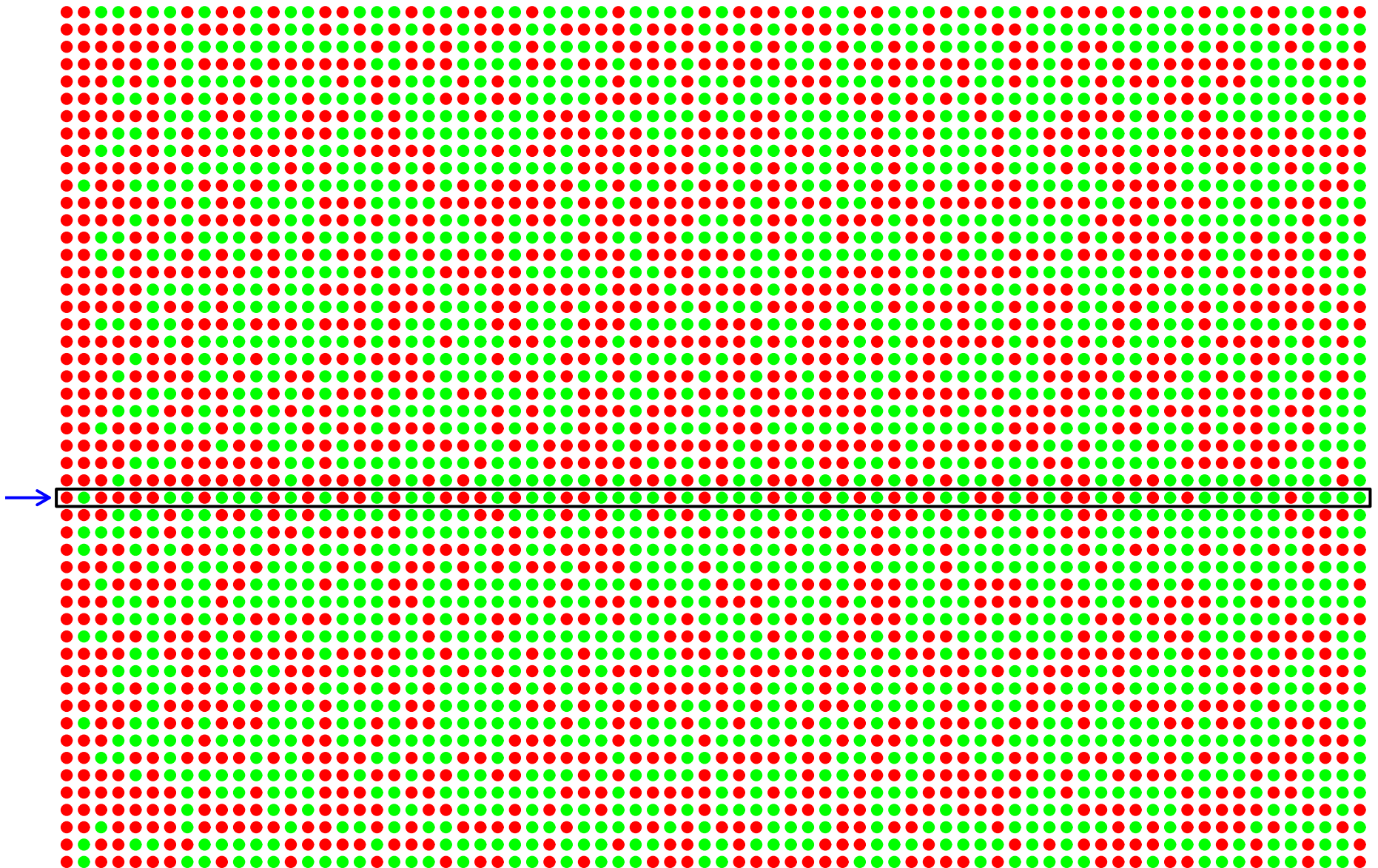
# MCMC in action

---



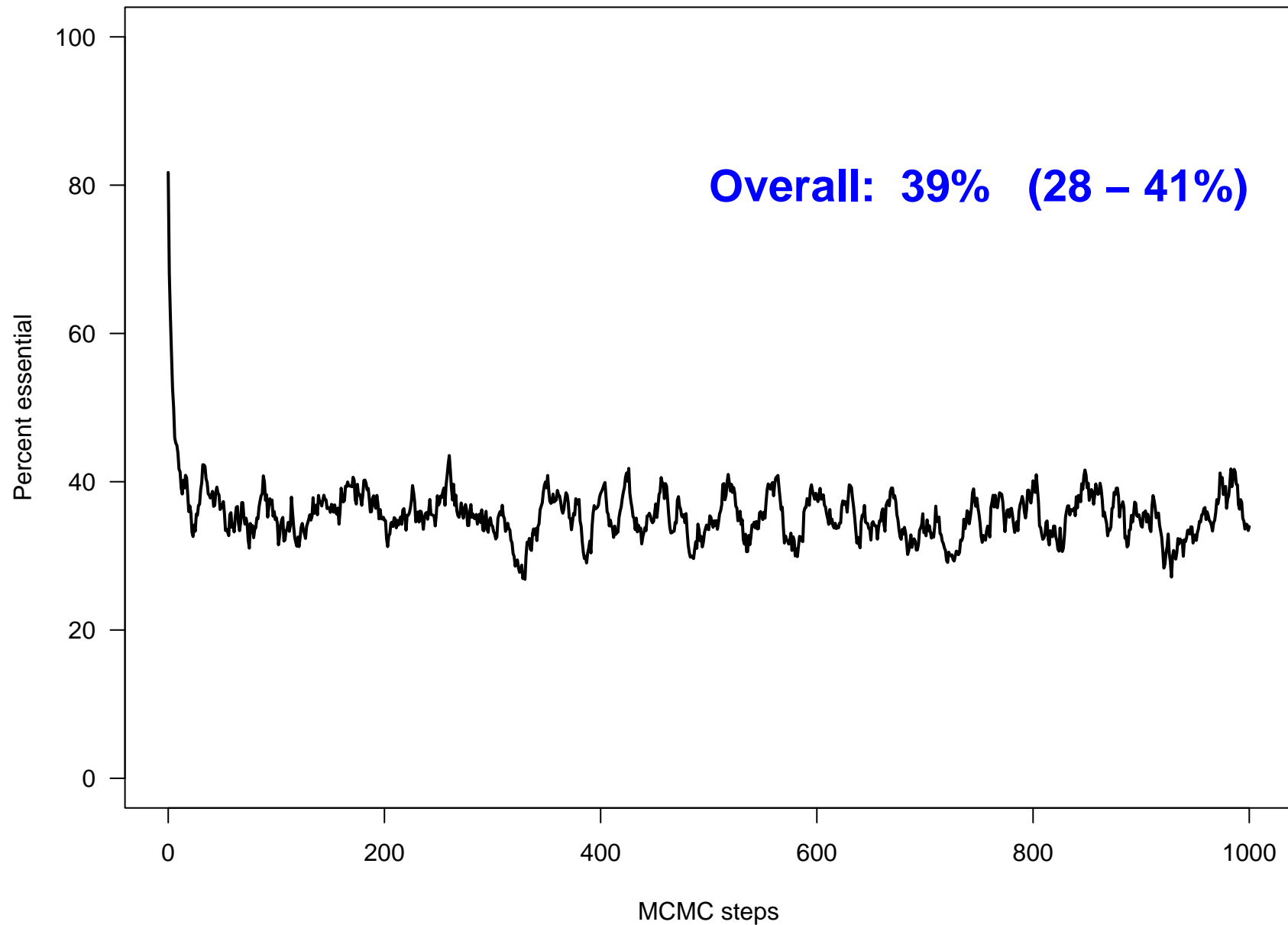
# A particular gene

---



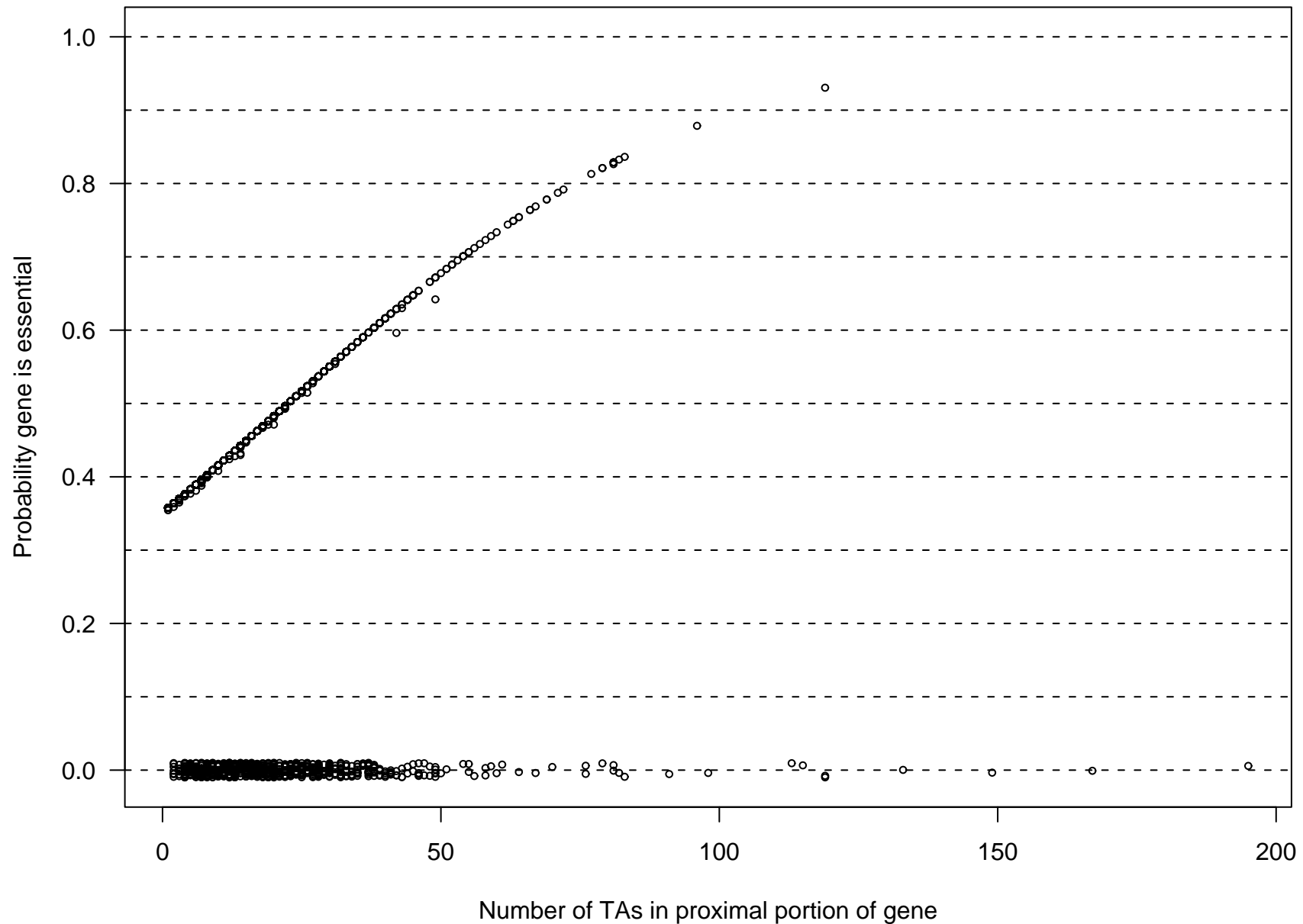
# Percent essential genes in M. tb.

---



# Probability that each gene is essential

---



# Potentially dicey bits

---

- Insertion sites in regions of gene overlap
- Operons
- The 80% rule
- Relationship between essentiality and number of insertion sites
- Randomness of transposon insertion

# Acknowledgements

---



Bill Bishai



Natalie Blades



Gyanu Lamichhane