

# Estimating Evolutionary Parameters for Protein Low Complexity Regions using an Approximate Bayesian Computation

Alexander Turco

November 27, 2022

# Overview

Background information

Research Questions/Explorations

Experimental Approach

Results

Conclusion and Future Work

## What are LCR's?

## *Saccharomyces cerevisiae* SRP40 Protein LCRs

>CAA82171.1(25-125) complexity=0.92 (15/1.90/2.20)

```

sssssssssssssssssssssssgsssssssssssssdssdssdsessssssssss
sssssdssssesdssssgsssssssssdesssesede

```

>CAA82171.1(149-282) complexity=1.33 (15/1.90/2.20)

```

essssessssgsssssesgsgesdsdsssssssssdsesdsesdsqsssssssdsss
dsdssssdsdssdsdssssssssssdsdssdsdsssdsssgssdsssssdsdssdestssds
dsdssdsdsgssse

```

>CAA82171.1(298-316) complexity=2.18 (15/1.90/2.20)

tpassnestpsasssssan

# Shannon's Entropy - MAYBE

$$H = -L \sum p_i \log_2(p_i)$$

# LCR's Present in Unique Ways

## Homorepeats

Consecutive iterations of a single residue



# LCR's Present in Unique Ways

## Homorepeats

Consecutive iterations of a single residue



## Direpeats

Consecutive iterations of two ordered, different residues



# LCR's Present in Unique Ways

## Homorepeats

Consecutive iterations of a single residue



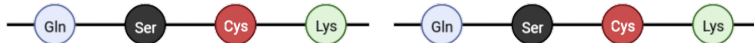
## Direpeats

Consecutive iterations of two ordered, different residues

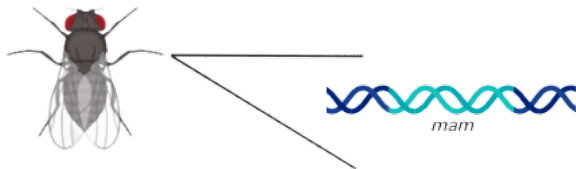


## Tandem Repeats

Sequence of residues which are repeated a number of times



# LCR's are Hypermutable



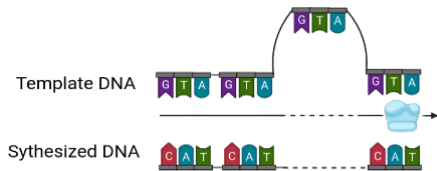
<b>mam domain</b>	<b>Size (bp)</b>	<b>Amino Acid Substitutions</b>	<b>Amino Acid/ Total Substitutions</b>
Unique	933	26	0.15
Repetitive	810	47	0.42

<sup>1</sup>Newfeld, Smoller, and Yedvobnick, 1991

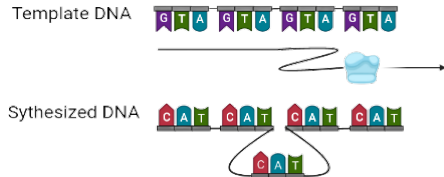


# Proposed Mechanisms of LCR Evolution

## 1. *Polymerase Slippage/Slipped Strand Mismatching*



**Polymerase Slips Forward**



**Polymerase Slips Backwards**

<sup>2</sup>Levinson and Gutman, 1987; Sehn, 2015

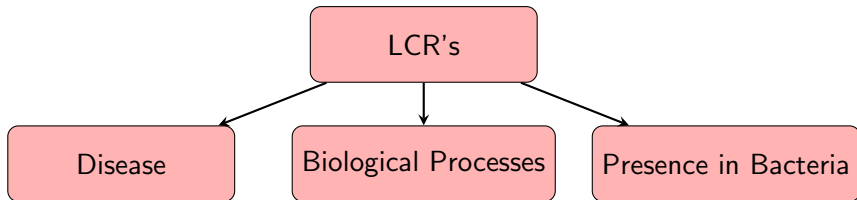
# Proposed Mechanisms of LCR Evolution

## 2. *Unequal Recombination*

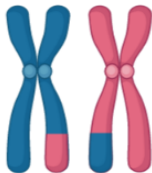


<sup>3</sup> Levinson and Gutman, 1987; Sehn, 2015

# Why Care about LCRs and Their Evolution?



*Huntington's Disease*



*Genetic Recombination*



*Neisseria meningitidis*

# What will this Study Explore?

- ▶ Estimation of evolutionary parameters (mutation rate, indel rates)
- ▶ Various models of insertions and deletions
- ▶ Summary statistics which best explain data

# What Approach will be Taken?

# Why use an ABC-MCMC

# MCMC for ABC

- 1 Propose a move from  $\theta$  to  $\theta'$  according to a transition kernel  $q(\theta, \theta')$ .
- 2 Generate simulated dataset  $D'$  using  $\theta'$  and calculate  $S'$ .
- 3 If  $\rho(S', S) \leq \epsilon$  continue to 4, otherwise remain at  $\theta$  and go to 1.
- 4 Calculate

$$\alpha(\theta, \theta') = \min(1, \frac{\pi(\theta')q(\theta', \theta)}{\pi(\theta)q(\theta, \theta')})$$

- 5 Accept  $\theta'$  with probability  $\alpha$ , otherwise stay at  $\theta$ .
- 6 Return to 1.

---

<sup>3</sup>Marjoram et al., 2003

# MCMC for ABC: Modified Algorithm

- 1 Propose a move from  $\theta$  to  $\theta'$  according to the normal distribution



# MCMC for ABC: Modified Algorithm

- 1 Propose a move from  $\theta$  to  $\theta'$  according to the normal distribution
- 2 Create a random protein sequence and mutate over many generations using  $\theta'$  to generate simulated Dataset  $D'$

# MCMC for ABC: Modified Algorithm

- 1 Propose a move from  $\theta$  to  $\theta'$  according to the normal distribution
- 2 Create a random protein sequence and mutate over many generations using  $\theta'$  to generate simulated Dataset  $D'$
- 3 Calculate summary statistics for simulated dataset  $D'$

# MCMC for ABC: Modified Algorithm

- 1 Propose a move from  $\theta$  to  $\theta'$  according to the normal distribution
- 2 Create a random protein sequence and mutate over many generations using  $\theta'$  to generate simulated Dataset  $D'$
- 3 Calculate summary statistics for simulated dataset  $D'$
- 4 If  $d(S', S) \leq \epsilon$ , go to next step, otherwise stay at  $\theta$  and return to 1

# MCMC for ABC: Modified Algorithm

- 1 Propose a move from  $\theta$  to  $\theta'$  according to the normal distribution
- 2 Create a random protein sequence and mutate over many generations using  $\theta'$  to generate simulated Dataset  $D'$
- 3 Calculate summary statistics for simulated dataset  $D'$
- 4 If  $d(S', S) \leq \epsilon$ , go to next step, otherwise stay at  $\theta$  and return to 1
- 5 Accept  $\theta'$  with probability ?? Ask Brian ab this again lol

# MCMC for ABC: Modified Algorithm

- 1 Propose a move from  $\theta$  to  $\theta'$  according to the normal distribution
- 2 Create a random protein sequence and mutate over many generations using  $\theta'$  to generate simulated Dataset  $D'$
- 3 Calculate summary statistics for simulated dataset  $D'$
- 4 If  $d(S', S) \leq \epsilon$ , go to next step, otherwise stay at  $\theta$  and return to 1
- 5 Accept  $\theta'$  with probability ?? Ask Brian ab this again lol
- 6 Return to step 1

# Simulation Step - MAYBE

# Results

# Conclusion/Future Work



# Blocks in Beamer

## Standard Block

This is a standard block.

## Alert Message

This block presents alert message.

## An example of typesetting tool

Example: MS Word,  $\text{\LaTeX}$