

*Understanding the role of germline
selection on patterns of linkage in
HapMaP III data*

Dibyoy Roy

DIMACS-REU Program

7/16/10

Background

- DNA has 4 nucleotides (Adenine, Guanine- Purines; Cytosine, Thymine- Pyrimidines) that comprise its code
- Most mutations (95%) are transition mutations (changing purine to other purine, pyrimidine to other pyrimidine), and the effect is not severe compared to transversions (changing purine to pyrimidine or vice versa)
- As a result, it will be adequate to use a binary system in the created sequences, in place of the four nucleotide system.

A

T

C

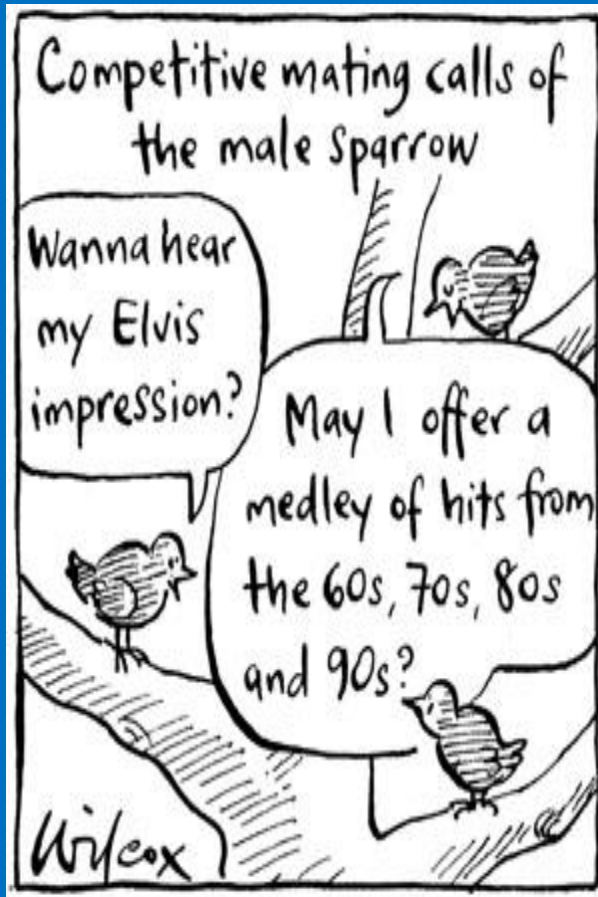
G

Initial Setup(cont.)

- A reference sequence will then be created (a string of zeros) and at initial time ($t=0$) in the simulation, this reference sequence will have the “optimal fitness” level for surviving in its “environment”.
- The main focus of the project will then be to model the evolution of this population under selection pressure due to changes in the environmental condition. Such changes can be modeled by altering the “optimum fitness” string.

Example of Reference Sequence

Let's Go Into Detail (cont.)



- The initial population consists of individuals who deviate from optimality by some random amount, represented by a random variable X .
- X will be defined as the number of differences between the optimum sequence and the sequence of a given individual.
- X (fitness) will help determine which sequences will be allowed to “mate” with each other.

Example of a Sequence (Individual 19, Generation 1)

•	0	0	0	0	1	1	0	1	1	1	0	1	1	0	0	1	0	0	0	0	0	1	0	1	
•	1	0	0	0	0	0	0	0	1	0	0	0	1	1	1	0	0	0	0	0	0	0	1	0	
•	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	
•	0	1	0	1	1	0	1	0	0	0	0	1	1	1	1	0	0	1	1	0	1	0	1	0	
•	0	1	1	0	0	0	1	0	1	1	1	0	0	0	0	1	0	0	0	0	1	1	0	1	
•	0	0	0	0	0	0	0	0	1	1	0	0	0	0	1	0	1	0	0	0	0	0	1	0	0
•	0	0	1	1	0	0	0	0	0	0	1	1	0	0	1	0	1	1	0	0	1	1	1	1	0
•	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	1	1	0
•	1	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	1	0	0	0	0	0	0	
•	0	0	0	1	0	0	0	1	1	0	1	1	0	1	0	0	0	0	0	0	0	0	1	0	
•	0	0	0	0	1	1	0	0	0	1	0	0	0	1	0	0	1	0	1	1	0	1	1	0	
•	0	0	1	0	0	1	1	0	1	1	0	0	1	0	0	0	1	0	0	0	0	1	0	0	
•	0	0	1	0	0	0	1	0	0	1	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0
•	0	1	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	1
•	1	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	0	0	0	1	0	1	0	0
•	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0
•	0	1	0	0	0	0	1	0	0	1	0	0	0	0	0	1	1	0	1	1	0	0	1	0	0
•	1	0	0	0	0	1	1	0	1	0	0	0	0	0	1	0	0	0	0	1	0	1	0	0	1
•	1	1	0	0	0	0	0	0	0	1	0	0	1	0	1	1	1	1	0	0	1	0	0	1	1
•	0	0	0	0	0	1	0	1	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0

Calculating Fitness (X)

- Thus, if $A(j,k,t)$ is a binary variable (it only takes values 0 or 1) representing the nucleotide at position k ($k=1,2,...M$) in individual j ($j = 1,2,...N$) at time t , and if $A_0(k, t)$ ($k = 1,2,...M$) is the “optimum sequence” at time t , then,

- $$X(j,t) = 1 - \frac{[\sum_k \text{XOR} (A(j,k,t), A_0(k,t))]}{M} \quad (1)$$

- Where XOR is the Exclusive OR function, represented by the rule:
 $\text{XOR}(0,0) = \text{XOR}(1,1) = 0$ and $\text{XOR}(0,1) = \text{XOR}(1,0) = 1 \quad (2)$

- Note that as defined, X takes values in $(0,1)$.

Calculating Fitness (cont.)

- 1 1 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 1 0 0 0 0 0 0
- 0 0 0 1 0 0 0 1 1 0 1 1 0 1 0 0 0 0 0 0 0 0 0 1 0
- 0 0 0 0 1 1 0 0 0 1 0 0 0 1 0 0 1 0 1 1 0 1 1 0 0
- 0 0 1 0 0 1 1 0 1 1 0 0 1 0 0 0 1 0 0 0 0 0 1 0 0

- Positions 201-300 selected as loci under selective pressure
- 0: 72
- 1: 28
- $X(j,t) = 1 - [\sum_k \text{XOR} (A(j,k,t), A_0(k,t))] / M$
- $X(19,1) = 1 - (28/100) = 0.72$

Variables

- N- Number of individuals
- M- Number of units in sequences of individual
- The initial population will be set up so that there will be variation amongst the individuals with a given locus in each individual having a probability **p** of differing from the reference sequence.
 - This will create a “population” with individual of differing fitness levels like most real populations.



<http://www.coolopticalillusions.com/illusions/abc123.gif>

Variables (cont.)

- The reference sequence will also be changed at **f** number of sites, which represents the “environmental changes” creating a new definition of “optimal fitness” for the “population”.
 - This is representative of which genotype would be most advantageous in a population due to selective pressure.
- The “offspring” will have a mutation rate, **u**, that will randomly change some of the “nucleotides” in their sequence from 0 to 1 or vice versa.
 - These mutations may increase or decrease fitness depending on how the variable **f** changes.

Environmentally Altered Reference Sequence!

•	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
•	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
•	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
•	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
•	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
•	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
•	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
•	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
•	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
•	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
•	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
•	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
•	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0
•	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
•	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
•	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
•	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
•	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
•	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
•	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0

Outline

- Given the population at time t , the population at time $t+1$ is generated as follows:
 1. Compute the fitness X for each individual at time t .
 2. Select individuals from the population such that the probability to be selected is proportional to fitness.
 3. A pair of individuals so selected at t will produce a pair of individuals at $t+1$. This will be done by swapping a random piece of the genome between these individuals.
 4. Induce mutations in the newly created individuals with probability u per nucleotide for a $0 \leftrightarrow 1$ change in their sequence.
 5. Repeat for further generations.

Procedure

- After 30 generations, environmentally altered reference sequence utilized as new reference sequence
 - Reference sequence was changed drastically once instead of gradually throughout generations
 - Represented some drastic change in environment as opposed to slowly changing environment
- Only 1/5 of offspring in each generation would swap selected loci (201-300) whereas others would swap other sections of genome
- Mutations would also be applied to offspring randomly somewhere throughout genome

Observations

- High fitness individuals (large X) would mate and create offspring with similarly high fitness levels
- These lineages would continue to be the dominant individuals within their generation with some inevitable variation between generations
- Once the reference sequence was changed, the more dominant individuals were still somewhat strong but their fitness (X) had undergone a definitive alteration
- For the rest of the generations, stronger individuals again were usually part of a stronger lineage
 - There was most likely little change in lineage strength due to the fact that the individuals were matched due to X and not allowed to be matched with certain preferences

Future Alterations to Model

- Incorporating mutation rate, u , into loop
- Integrate gradually changing reference sequence as opposed to only changing once

Biological Significance

- The HapMap Project and 1000 genomes project map out many parts of the genome that are essential to understanding many of the genetic causes of diseases and phenomena that occur
- Many populations have extremely different genetic afflictions that cause researchers to study how certain genes become significant



<http://www.lsa.umich.edu/psych/diversity/images/diversity.jpg>

Biological Significance (cont.)

- Through modeling and simulation, researchers are able to visualize how specific genes become significant within a population
- This simulation was meant to show a relatively simplistic method to understanding how previously enigmatic problems could be approached using real data



<http://depts.washington.edu/joubert/graphics/JS-world-map.png>

Sources

1. <http://cplus.about.com/od/glossar1/g/xor.htm>
2. http://evolution.berkeley.edu/evolibrary/article/0_0_0/evo_25
3. <http://hapmap.ncbi.nlm.nih.gov/whatishapmap.html.en>
4. <http://www.1000genomes.org/page.php?page=about>
5. http://www.mun.ca/biology/scarr/Transitions_vs_Transversions.html
6. <http://members.logical.net/~marshall/AbramsRefEnv.pdf>