# 16754. Marker to Marker LD in Relation to the Evolutionary History of the Site

**Soller, M.[1], Dolezal, M.A[2]., Bagnato, A. [2], O'Sullivan, N. [3], Fulton, J. [3], Santus, E. [2], Lipkin, E[1].**

[1]The Hebrew University of Jerusalem, 91904 Jerusalem, Israel; [2]Universita` degli Studi di Milano, 20133 Milano, Italy; [3]Hy-Line International, Dallas Center, IA 50063, USA

## The question

- LD between a pair of markers ***A*** and ***B*** as measured by $r^2$, depends on the frequency of the four haplotypes: ***AB, ab, Ab, aB***.
- In this study we ask:
  - What are the aspects of haplotype distribution that determine LD?
  - How does the distribution of haplotype frequencies develop as a function of the evolutionary history of the site, and the distance separating the marker pair?

## Notation

- For generality, we assign
  - Haplotype ***AB*** to the most frequent haplotype,
  - Haplotype ***ab*** to the double alternative to ***AB***,
  - Haplotype ***Ab***, to the least frequent haplotype,.
  - Haplotype ***aB***, to the remaining haplotype.

## Haplotype frequencies according to $r^2$ values

- For each of target $r^2$ = 0.1, 0.2, …. 1.0, ten marker pairs presenting $r^2$ value centered at the target were chosen (100 total marker pairs), and average haplotype frequencies ***f(AB),f(ab), f(aB)*** and ***fAb)*** were calculated for each $r^2$ value.

**Table 1**. Haplotype frequencies by $r^2$.

| $r^2$ | f(AB) | f(ab) | f(aB) | f(Ab) | f(AB+ab) | f(aB+Ab) |
|---|---|---|---|---|---|---|
| 0.1 | 0.476 | 0.333 | 0.155 | 0.037 | 0.809 | 0.191 |
| 0.2 | 0.524 | 0.268 | 0.184 | 0.024 | 0.792 | 0.208 |
| 0.3 | 0.518 | 0.255 | 0.212 | 0.015 | 0.773 | 0.227 |
| 0.4 | 0.581 | 0.248 | 0.143 | 0.029 | 0.829 | 0.171 |
| 0.5 | 0.595 | 0.274 | 0.125 | 0.006 | 0.869 | 0.131 |
| 0.6 | 0.572 | 0.323 | 0.087 | 0.019 | 0.894 | 0.106 |
| 0.7 | 0.708 | 0.229 | 0.051 | 0.012 | 0.937 | 0.063 |
| 0.8 | 0.660 | 0.295 | 0.038 | 0.007 | 0.955 | 0.045 |
| 0.9 | 0.634 | 0.344 | 0.019 | 0.004 | 0.977 | 0.023 |
| 1.0 | 0.693 | 0.307 | 0.000 | 0.000 | 1.000 | 0.000 |

## Puzzle No. 1: Consistent order of frequencies for ab and aB

- In each case, without exception ***f(ab)>f(aB)***.
- Thus, the general structure ***f(AB)>f(ab)>f(aB)>>f(Ab)*** appears to be essential for significant LD.
- For the 777K array, over 90% of pairs between adjacent markers ("adjacent marker pairs" – i.e., markers on the diagonal of the LD matrix) displayed the above haplotype distribution.
- **Puzzle No. 1: How did such a high proportion of adjacent marker pairs come to have this very specific distribution of haplotypes?**

## $r^2$ as a function of Haplotype frequencies

- In the range $r^2$=0.3 to 1.0, ***f(AB+ab)*** increases consistently with $r^2$.
- ***f(Ab)*** is very close to zero and does not vary consistently with $r^2$,
- ***f(aB)*** decreases consistently with $r^2$.

## f(aB) is the main determiner of $r^2$

**Table 2**. $r^2$=0.5 as function of haplotype frequencies. Individual values of 9 randomly chosen marker pairs presenting $r^2$~05.

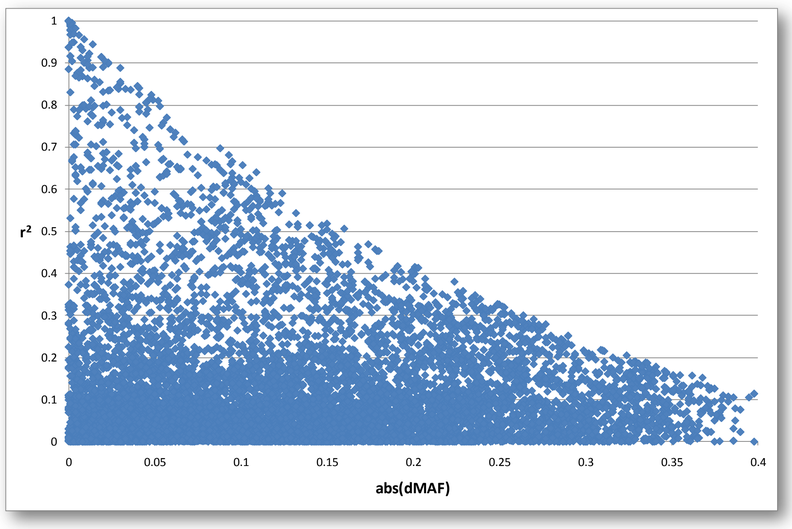
| $r^2$ | AB | ab | aB | Ab | AB+ab |
|---|---|---|---|---|---|
| 0.508 | 0.516 | 0.332 | 0.137 | 0.015 | 0.848 |
| 0.510 | 0.463 | 0.371 | 0.166 | 0.000 | 0.834 |
| 0.497 | 0.656 | 0.207 | 0.137 | 0.000 | 0.863 |
| 0.495 | 0.729 | 0.157 | 0.112 | 0.002 | 0.886 |
| 0.504 | 0.781 | 0.116 | 0.103 | 0.000 | 0.897 |
| 0.503 | 0.591 | 0.259 | 0.151 | 0.000 | 0.850 |
| 0.498 | 0.675 | 0.195 | 0.129 | 0.001 | 0.870 |
| 0.506 | 0.440 | 0.407 | 0.127 | 0.026 | 0.847 |
| 0.501 | 0.564 | 0.279 | 0.157 | 0.000 | 0.839 |

**Within $r^2$=0.5**

- ***f(AB)*** and ***f(ab)*** individually vary widely for virtually identical $r^2$ values, but their sum varies over very small range (0.834-0.897), and does not overlap that for other $r^2$ values.
- ***f(Ab)*** values are very low, often zero, for virtually identical $r^2$ values, (and also have same values for widely different $r^2$ values, data not shown),
- f***(aB)*** varies over very small range (0.103 to 0.166), and values do not overlap those for other $r^2$ values.
- Each level of $r^2$ values has its own characteristic ***f(aB)*** and ***f(AB+ab)***, but not ***f(AB)*** or ***f(ab)*** taken individually.
- Thus, ***f(aB)*** appears to be the main determiner of ***f(AB+ab)***, and of $r^2$.

**Acknowledgement.** This study is funded by the EC 7th Framework Programme (FP7/2007-2013), agreement n° 222664, "Quantomics" (this Publication reflects only the author's views and the EC is not liable for any use that may be made of the information contained herein).

**Figure 1**. $r^2$ plotted against dMAF.



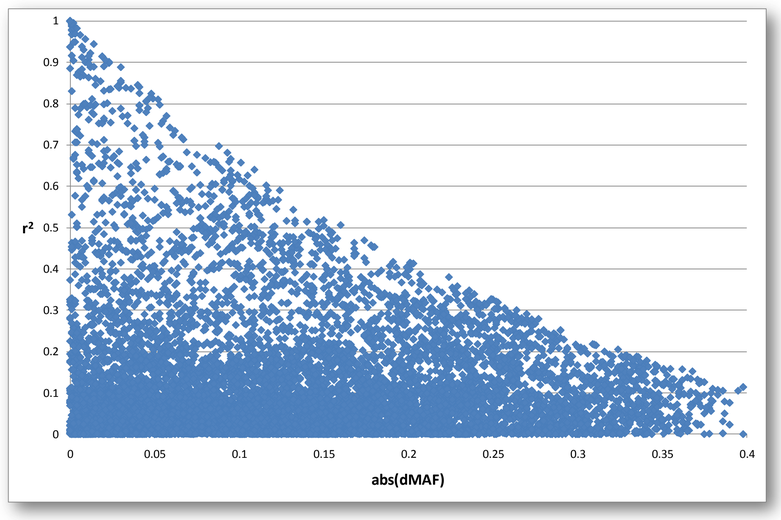

## Puzzle number 2

- For the 777K array, 55% of diagonal $r^2$ values ≥ 0.6.
- For $r^2$>0.6, we need f(aB) < 0.100; and f(Ab) close to zero.
- **Puzzle No. 2: How did such a high proportion of marker pairs come to have these very special proportions of f(aB) and f(Ab)?**
- We unravel these puzzles in the light of the evolutionary history of a closely linked marker pair.

## Unraveling the puzzles

- **We start** with a polymorphic site, ***A, a***.
- **Mutation** of a nearby site, from ***B*** to ***b***.
- This immediately generates three Haplotypes:
  - ✓ ***AB, aB***, and ***ab*** (assuming the new mutation occurs in a chromosome strand carrying allele ***a***).
  - ✓ Frequency: ***f(AB)>f(aB)>>f(ab)***
  - ✓ ***f(Ab)***=0, until crossing over between ***AB*** and ***ab***, which must wait for ***f(ab)*** to reach appreciable values.
- **Drift** enters the picture.
  - ✓ Increase in ***f(ab)*** depends on genetic drift over evolutionary time.
  - ✓ With ***f(Ab)***=0, ***f(b) = f(ab)***.
  - ✓ Choosing markers with minimum MAF → ***ab*** has not been lost from the population, and is present at some minimum frequency.
- **Selection** of markers enters too**!**
  - ✓ Markers for the array are chosen to have MAF in the range 0.05 to 0.5. Since ***f(Ab)*** is either zero or very very low, this means that ***f(ab)*** is also in this range, with mean about 0.2-0.3.
  - ✓ We have defined ***AB*** as the most frequent haplotype, hence frequency will be >0.5 – about 0.6-0.8 (Tables 1 and 2).
  - ✓ Thus, between ***f(AB)*** and ***f(ab)***, there is not much left for ***f(aB)*** – hence the low values for ***f(aB)***.
  - ✓ And that is how we end up with 55% of diagonal LD values >0.6.
- **Recombination?**
  - ✓ All of this holds only for very close markers, so that ***f(Ab)*** remains very low.
  - ✓ Once markers are separated by greater distances, ***f(Ab)*** becomes appreciable, and LD descends to very low values.

**Figure 2**. Diagonal LD against separation distance.




## Oh! Oh! Trouble?

- If high marker-marker LD is mainly due to selection for high MAF → marker-marker LD may not be representative of marker-QTL LD!
- **Not necessarily!**
  - To contribute to genetic variation, QTL must also be at moderate frequencies.
  - Thus, marker-marker LD may also represent marker-QTL LD.

***Thus, in an amazing way it all comes together to provide genomic structures preadapted for Genome selection and GWAS.***