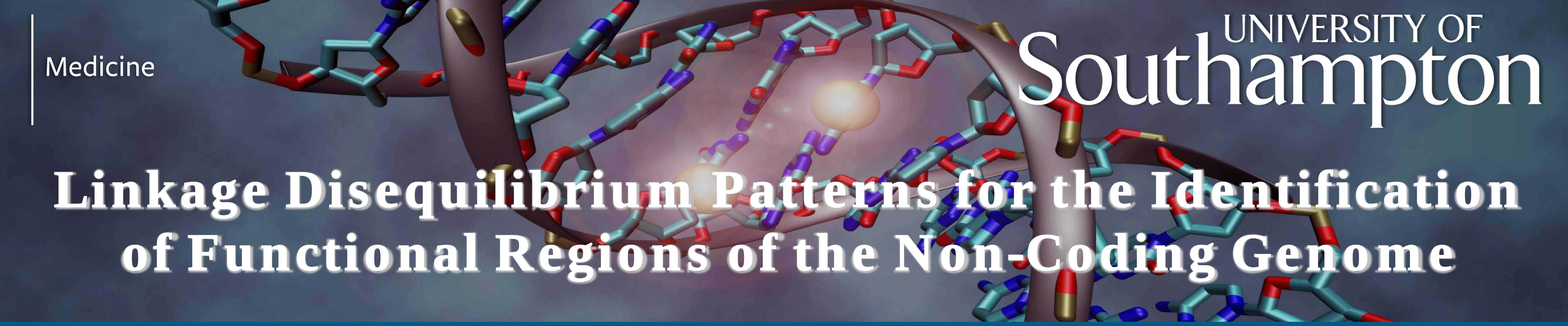

Medicine

UNIVERSITY OF Southampton

# Linkage Disequilibrium Patterns for the Identification of Functional Regions of the Non-Coding Genome

Reuben J. Pengelly

Human Genetics & Genomic Medicine, Faculty of Medicine, University of Southampton, UK

Contact: R.J.Pengelly@soton.ac.uk

## Introduction

Linkage disequilibrium (LD) is the physical association of two alleles, being co-inherited at a rate greater than that expected by chance. This phenomenon underpins many of our methodologies for medical genetic research, such as GWAS studies. Several metrics for the quantification of LD between markers are available (such as r, D and D'), however the fitting of a biologically meaningful statistical model makes more efficient use of the available data, producing a more robust map of LD across the genome [1]. The Malécot model of isolation by distance is one such model:

$$\rho = (1 - L)Me^{-\varepsilon d} + L$$

where $\rho$ is the correlation between 2 markers, $L$ is the component of $\rho$ not due to physical LD, $M$ is linkage at 0 distance, $\varepsilon$ is the rate of degradation of LD between markers and $d$ is the physical distance between markers. Iterative fitting of this model is implemented in the LDMAP software [2]. From an input of population genotype data, LDMAP produces a cumulative map in LD units (LDU; the product of $\varepsilon d$), broadly analogous to centimorgan maps derived from linkage data. As patterns of LD in a population are associated with selection pressures, these LD maps should be able to be interrogated in order to identify regions under selection, highlighting functional regions of the genome [3].

## Aims

We aimed to investigate the relationship between LDU length and various genomic features, in order to assess whether this was a useful metric for the prediction of functional regions of the non-coding genome.

## Methods

We utilised LD maps derived from whole-genome sequencing data for chromosome 22 of the GRCh37/hg19 release of the human reference genome, built using genotypes from Yoruba in Ibidan, Nigeria individuals [4].

Genomic feature coordinates were downloaded from the UCSC Table Browser and were evaluated using custom scripts, *BEDTools* and *R*. We evaluated the LDU/kb ratio of each of the feature classes, and compared these against the chromosome mean.

## Results

LD maps were successfully produced from WGS data, and correlate well with D' based upon HapMap data (e.g. Fig. 1).

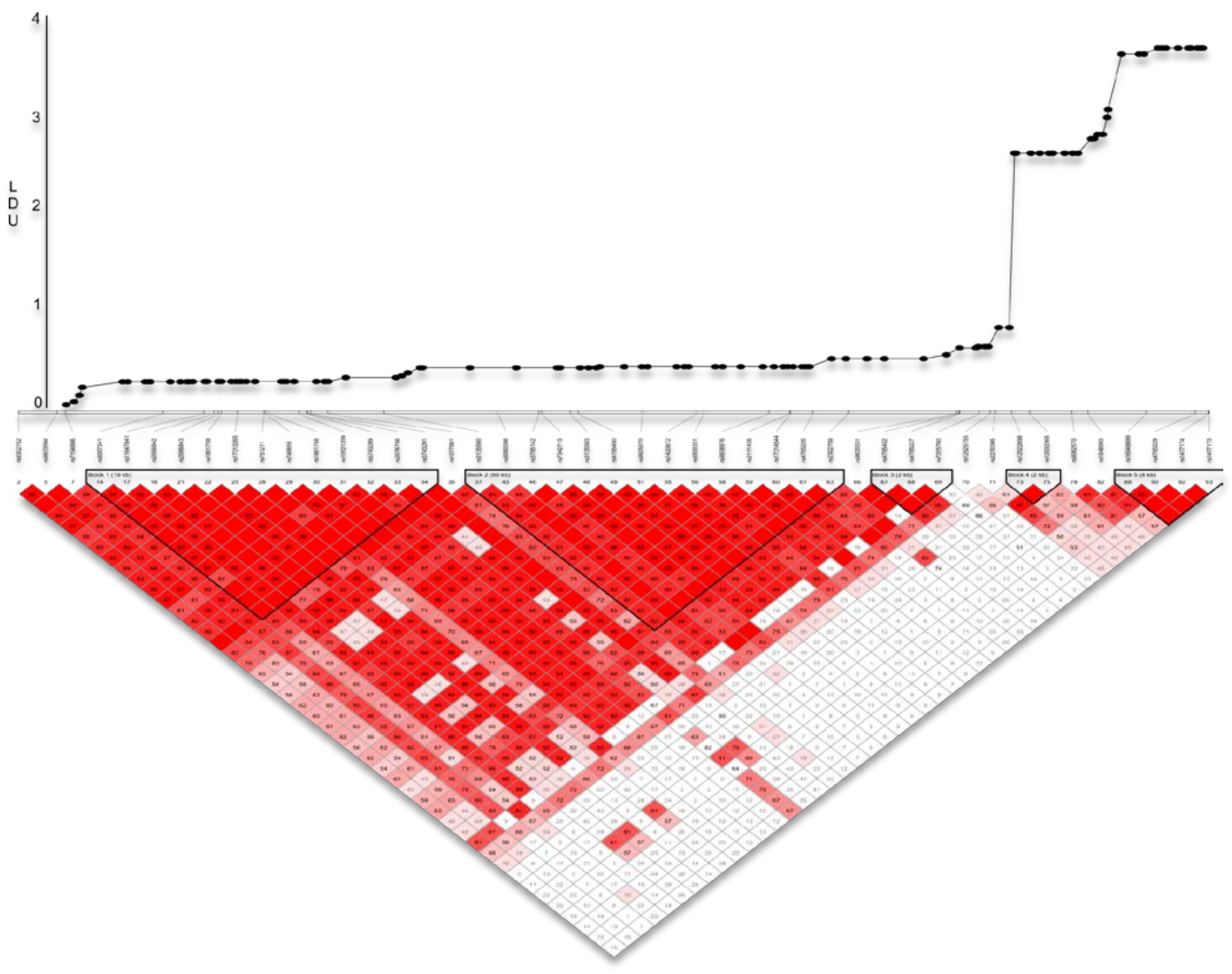

**Fig. 1** | Representative comparison of CEU LDU map with Haploview LD Plot (red squares indicating high D' values) for a small genomic region. Regions of breakdown in LD can be seen to closely correlate.

The chromosomal average LDU rate is 0.050 LDU/kb. As we have previously shown [4,5], transcribed DNA (based upon the RefGene database) has a significantly shorter LDU/kb ratio, as do modified histone marked regions. DNase hypersensitive regions have inflated ratios, potentially signifying diversifying selection.

| Feature | LDU/kb | *p* |
|---|---|---|
| Chromosome mean | 0.050 | - |
| Transcribed | 0.044 | 0.005 |
| DNase hypersensitive | 0.057 | $4.3\times10^{-13}$ |
| CpG island | 0.048 | 0.573 |
| H3K4m1 | 0.045 | 0.067 |
| H3K9ac | 0.040 | 0.0001 |

**Table 1** | LDU/kb ratios for functional marks in the genome, and significance for deviation from the chromosome mean (1 sample t-test).

## Discussion

LD patterns provide a potential method for the identification of functional elements in the genome by the identification of regions under selection. This is a potential avenue for prioritisation of regions harbouring variants of uncertain significance identified in patient sequencing.

Further work should be undertaken in order to identify alternative methods with better sensitivity and resolution in order to accurately identify functional loci in population data.

In addition to this human work, LD mapping is a highly valuable tool in other species such as *Gallus gallus*, allowing for the identification of regions under selection within breeding programmes.

## Conclusions

LD is an informative tool in the identification of functional regions of the genome, and may be beneficial in medical and agricultural genetics.

### References

1. Morton *et al.* (2001), *Proc Natl Acad Sci U S A* **98**:5217-21
2. Kuo *et al.* (2007), *Methods in Molecular Biology* **376**:47-57
3. Ennis (2007), *Methods in Molecular Biology* **376**:59-70
4. Pengelly *et al.* (2015), *BMC Genomics* **16**:666
5. Pengelly *et al.* (2016), *Heredity* **117**:375-82

### Acknowledgements

We thank the 1000 Genomes Project for public-access to data, and Rick Tearle for advice on Complete Genomics data. Analyses were performed on Iridis 4. Banner image created by Christoph Bock.