Medicine

UNIVERSITY OF Southampton

High Resolution Linkage Disequilibrium Maps Derived from Whole-genome Sequencing Data

Reuben J. Pengelly, William Tapper, Andrew Collins & Sarah Ennis

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 ρ is the correlation between 2 markers, L is the component of ρ not due to physical LD, M is linkage at 0 distance, ε 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 εd), broadly analogous to centimorgan maps derived from linkage data.

Aims

Our group have previously reported LD maps based upon HapMap [3] and whole-exome data [4]. Now that whole-genome sequencing (WGS) data are available for several populations, we aimed to generate LD maps based upon these data, bypassing some of the limitations associated with the previous data.

Methods

High depth of coverage Complete Genomics sequenced WGS data was retrieved in VCF from the 1000 Genomes Project FTP for the populations denoted CEU (n=96), CHS (n=93) and YRI (n=80), and analysed using a custom pipeline (Fig. 1).

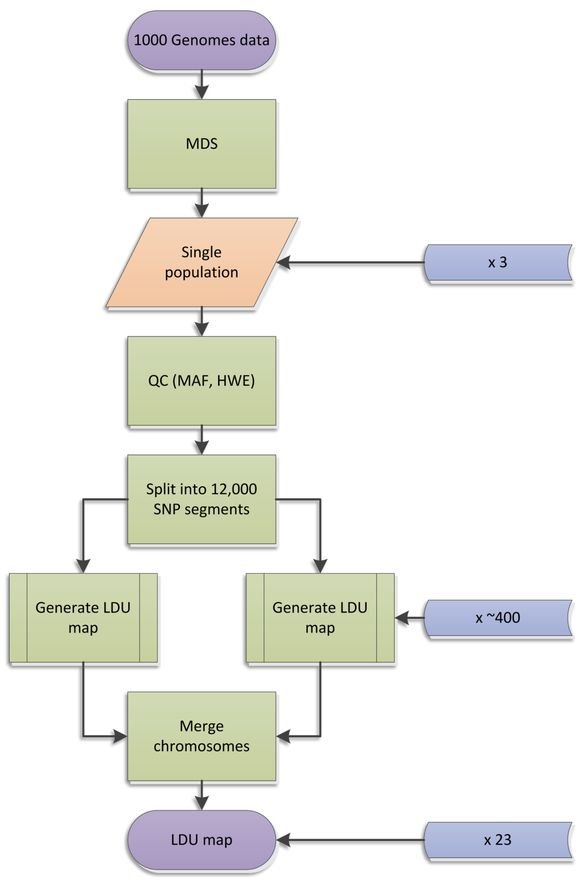

Fig. 1 | Overview of analysis workflow for LD map generation. Populations were confirmed with multi-dimensional scaling, followed by quality pruning of SNPs. Sub-sections of chromosome were then analysed in parallel, and map regions merged to give the final map. Analysis requires ~10 CPU-weeks/population.

Results

LD maps were successfully produced from WGS data, and correlate well with D' based upon HapMap data (e.g. Fig. 2).

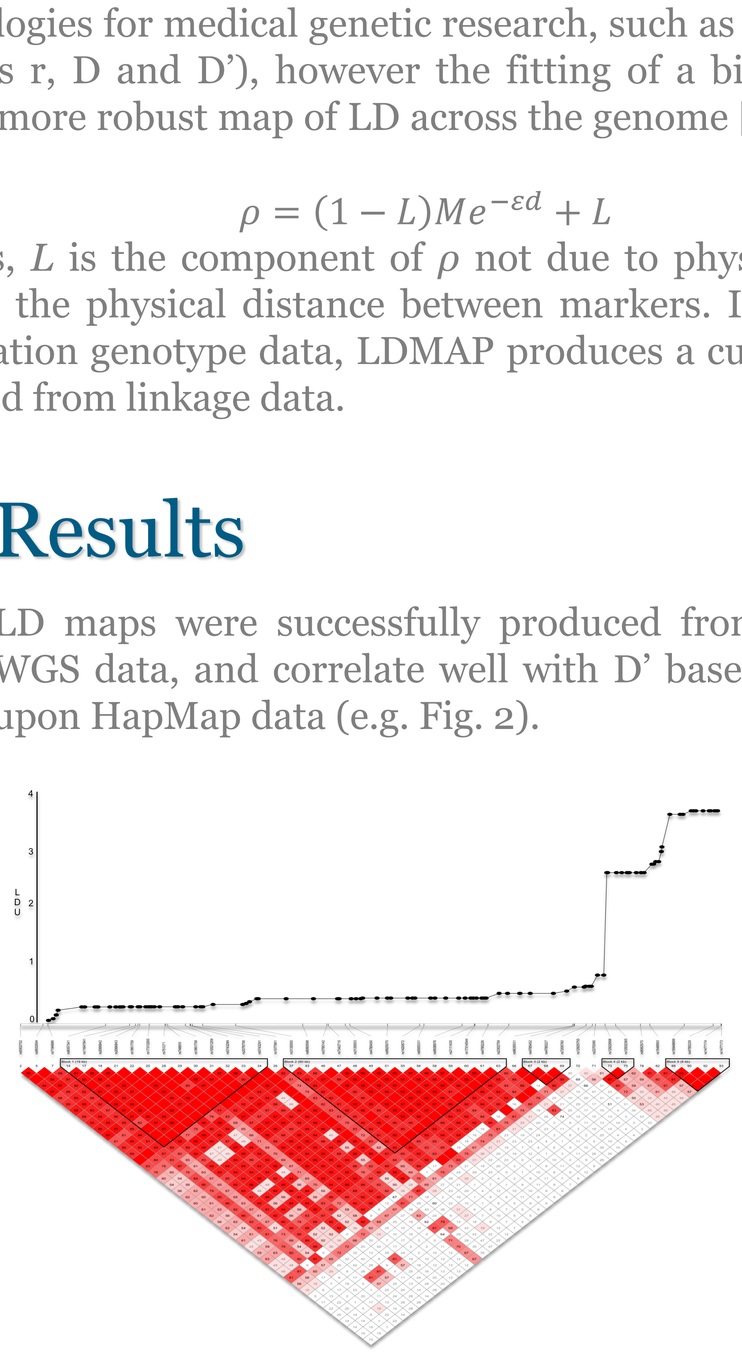

Fig. 2 | Representative comparison of CEU LDU map with Haploview LD Plot (red squares indicating high D' values) for a small genomic region. Regions of breakdown in LD can be seen to closely correlate.

These new LD maps also correlate well with previous HapMap-derived maps, although the greater resolution of the new maps is readily apparent in many areas, filling in previous holes in the map (e.g. Fig 3).

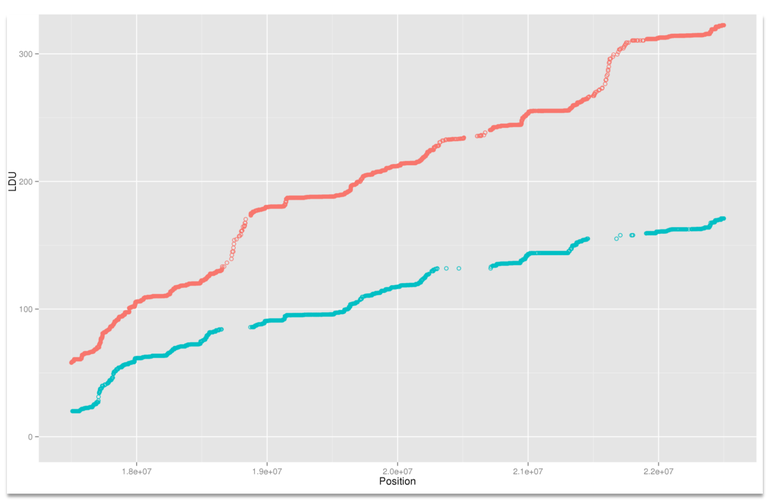

Fig. 3 | Comparison of LD maps created using CEU HapMap (blue) and WGS data (red), across a region of Chr22. The 3 large 'holes' (regions in which there is no information on LD, often due to insufficient marker density) apparent in the HapMap data are largely infilled in the new map due to the higher resolution.

We investigated the relative data-saturation achieved for each population by generating LD maps for random subsets of the data. CEU and CHS could be seen to be approaching saturation (Fig. 4). For YRI however, it appears that more than 80 individuals will be required to approach saturation.

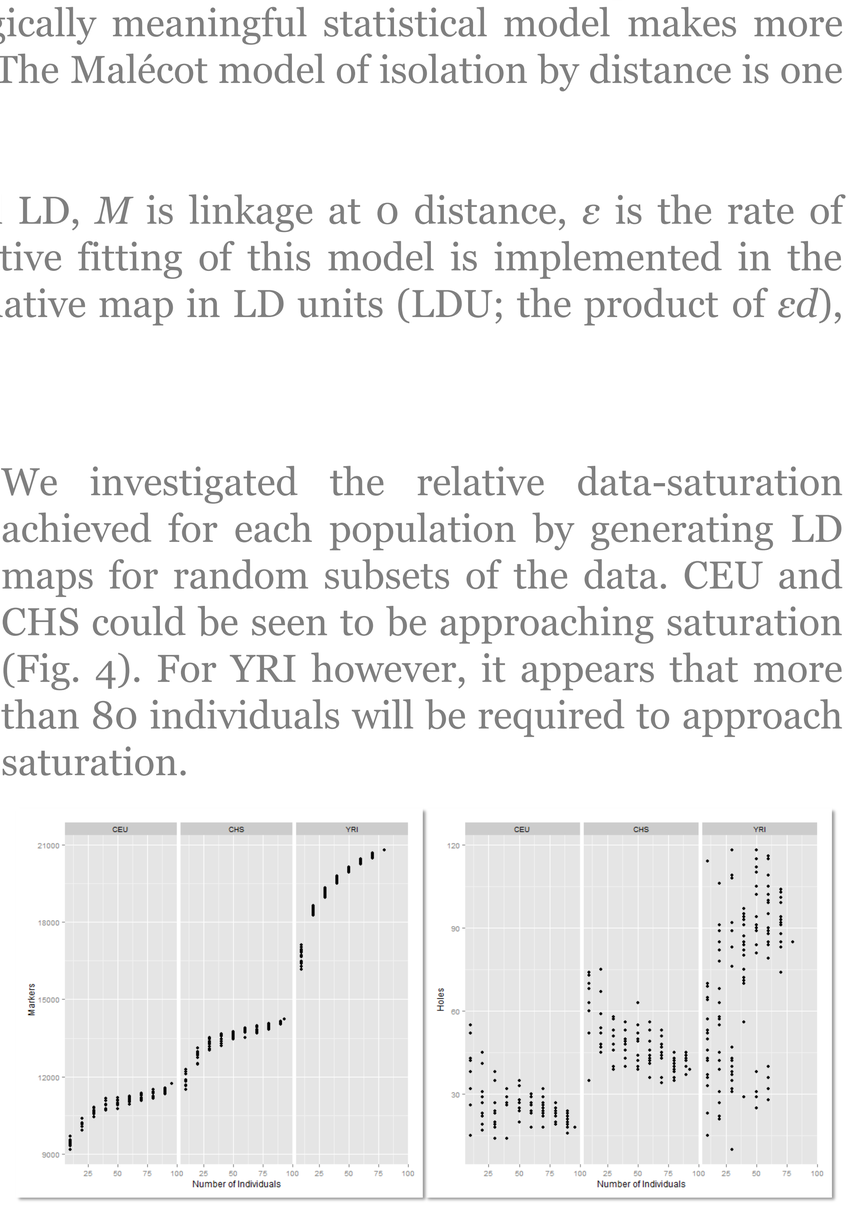

Fig. 4 | Number of markers available (left), and holes in final maps (right) for LD maps generated for a region of Chr22 on random subsets of individuals for genotype data. Available markers increase with sample-size and a decrease in the number of holes is seen, giving a more complete map, achieving a partial plateau in CEU & CHS, though not in YRI. 10 pseudoreplicates were performed for all conditions.

Conclusions

It is clear that WGS allows for a far greater resolution to be achieved for an LD map than previous data. In populations with a less recent bottleneck, greater numbers of individuals will be required to saturate the maps, increasing robustness. The final WGS maps will allow for refinement of LD based genetic investigations, hopefully aiding the identification of the causes of some GWAS association signals, as well as further investigation of biological mechanisms underlying the LD patterns in the genome.

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. Tapper *et al.* (2005), *Proc Natl Acad Sci U S A* **102**:11835-9
4. Gibson *et al.* (2013), *Hum Genet* **132**:233-43

Population Details

CEU - CEPH (Utah residents with ancestry from northern and western Europe)
CHS – Southern Han Chinese
YRI – Yoruba in Ibidan, Nigeria

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.