Linkage Disequilibrium Maps for European and African Populations Constructed from Whole Genome Sequence Data

Alejandra Vergara-Lope*, M. Reza Jabalameli*,
Clare Horscroft, Sarah Ennis, Andrew Collins
& Reuben J. Pengelly†

May 31, 2019

Human Genetics & Genomic Medicine, Faculty of Medicine, University of Southampton, Southampton, UK.
*These authors contributed equally.
†Corresponding author: Reuben Pengelly (R.J.Pengelly@soton.ac.uk)

Abstract

Quantification of linkage disequilibrium (LD) patterns in the human genome is essential for genome-wide association studies, selection signature mapping and studies of recombination. Whole genome sequence (WGS) data provides optimal source data for this quantification as it is free from biases introduced by the design of array genotyping platforms. The Malécot-Morton model of LD allows the creation of a cumulative map for each choromosome, analogous to an LD form of a linkage map. Here we report LD maps generated from WGS data for a large population of European ancestry, as well as populations of Baganda, Ethiopian and Zulu ancestry. We achieve high average genetic marker densities of 2.3–4.6/kb. These maps show good agreement with prior, low resolution maps and are consistent between populations. Files are provided in BED format to allow researchers to readily utilise this resource.

Background & Summary

Mapping of linkage disequilibrium (LD) is invaluable for many endeavours including identifying signatures of selection, refinement of signals in genome-wide association studies and studies into recombination [1, 2, 3].

One approach to the quantification of LD is the generation of LD maps applying the Malécot-Morton model [4, 5]. The product generated utilising the Malécot-Morton model are maps in cumulative linkage disequilibrium units

(LDU), which are broadly analogous to an LD-based form of centimorgans. Previous studies have reported maps generated from array based genotyping data in multiple populations (e.g. [6]), allowing for cross-population comparisons.

The mathematical basis of *LDMAP* has been previously described [4, 5]. In brief, *LDMAP* generates a cumulative map of LD distances between markers, based upon the Malécot-Morton model of association by distance:

$$\hat{\rho} = (1 - L)\, M e^{-\epsilon d} + L \tag{1}$$

where $\hat{\rho}$ is the association between two markers in a population, L is the component of $\hat{\rho}$ not due to LD, but due to confounding factors such as recent founder effects, M is the association at 0 distance (approximately 1 for monophyletic haplotypes), ϵ is the rate of decline in the association between the markers and d is the physical distance between the markers [5]. The final LDU map is built by cumulative addition of ϵd for each inter-marker span.

The increasing availability of whole genome sequencing (WGS) data allows the investigation of LD patterns at the highest level, without the impact of issues such as ascertainment bias in the selection of single nucleotide polymorphism (SNP) markers. We have previously shown that WGS-based maps provide tangible benefits in their practical application. Arrays have been designed to give a reasonable coverage of LD information for a reduced set of SNPs, as such they have limited resolution and population-specific biases are introduced during SNP selection. Given that WGS variant identification is 'hypothesis free' (i.e. SNPs are not required to be pre-defined as in array genotyping), these data, and thus these maps, represent a maximally informative resource [7].

The lack of ascertainment bias for SNP data collection is particularly important for African populations, as they have the greatest population diversity and are often under-represented in genomic studies. Though they are often under-represented, these populations are particularly informative for many studies, given the extended time since a population bottleneck [7, 8]. Higher resolution maps allow for analyses on a finer scale of the patterns of LD, such as structure within genes [9].

Here, we report our generation of WGS based LD maps for four populations, one of European and three of African descent. These maps provide a valuable population genetic resource, providing a maximal resolution, selection bias free, dataset for studies which require the incorporation of LD statistics.

Methods

Autosomal WGS data from two cohort sequencing studies was utilised. African populations were sequenced within the African Genome Diversity Project [8],

utilising Illumina short read sequencing to an average depth of 4x. European ancestry individuals were sequenced by the Wellderly Study [10], utilising Complete Genomics high depth sequencing. Multidimensional scaling as implemented in *PLINK* [11] was applied to ensure genetic homogeneity within the sub-cohorts.

SNPs were subject to quality control prior to map generation. Specifically, they were required to have a minor allele frequency $\geq 1\%$, $< 5\%$ genotype missingness and not to significantly deviate from Hardy-Weinberg equilibrium (at $\alpha = 10^{-3}$). All analyses were undertaken using the reference genome GRCh37 (hg19).

LD maps were made using *LDMAP* with default parameters. Owing to the computational intensity of LD map generation, this was performed for 12,000 marker overlapping segments, which were then concatenated into full chromosome maps, removing the 25 terminal markers of each segment to avoid end effects.

Code availability

The core *LDMAP* software is written in C, and made available at `www.soton.ac.uk/genomicinformatics/research/ld.page`.

Data Records

LD maps reported here are freely available at `https://doi.org/10.6084/m9.figshare.7850882` (data citation 1). These data are in Browser Extensible Data (BED) format, including the cumulative LDU position of every SNP marker within the generated maps. Additionally, these data are also made available as the kb/LDU ratio for each inter-SNP span providing a view of the regional 'intensity' of LD.

For the African populations [8], 95–100 individuals were utilised for each subpopulation, yielding approximately 14 million SNP markers (Table 1). The European map utilised 454 individuals [10], yielding approximately 7.5 million markers. The increased population diversity for the African compared to European population can be seen in the increase common SNP density, as well as the longer LDU length which corresponds to the longer total haplotypic diversity within a population.

Table 1: Key statistics for generated LD maps

Population	Individuals	Marker count	Density[a]	LDU
Baganda	100	13,439,201	4.35	129,640
Ethiopian	95	13,892,209	4.48	107,001
European	454	7,062,420	2.28	63,427
Zulu	100	14,205,839	4.59	130,156

[a]Average SNP markers per kb.

Technical Validation

For these data, we can determine that they are robust as they are consistent with prior, lower resolution maps, and that they are consistent between populations assessed (Figure 1). As we know that patterns of recombination and thus LD are broadly consistent between populations, this meets our prior expectations; furthermore the total map lengths are proportional to time since an effective population bottleneck (being longer in African populations reflecting the additional diversity present) [6, 12, 7].

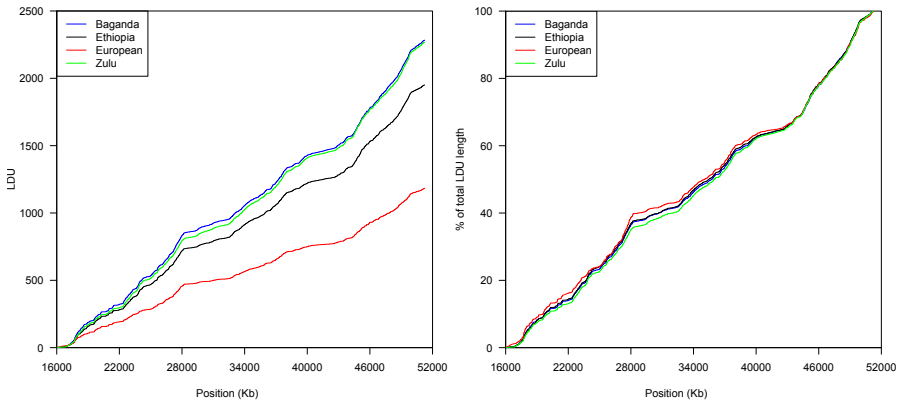

Figure 1: Comparison of the four maps for chromosome 22. The raw cumulative maps are shown (left), as well as maps normalised to have the same total length (right). It can be seen that the contour profiles of the maps are highly similar, though there is variation in the total map length.

Usage Notes

Maps can be readily incorporated into genomic analyses using tools such as BEDTools [13], allowing annotation of regions with LD information for subsequent analysis such as determining whether a genomic feature has higher LD than background on average.

Genome wide association studies using a composite likelihood model can be undertaken with LD information as provided here, allowing for additional power for signal detection and refinement [14, 2].

Acknowledgements

The authors acknowledge the use of the IRIDIS High Performance Computing Facility, and associated support services at the University of Southampton, in the completion of this work.

Author contributions

MRJ undertook data analysis.
AV-L undertook data analysis.
SE contributed to study design and supervision.
AC contributed to study design and supervision.
RJP contributed to study design, data analysis, supervision and wrote the manuscript.

Competing financial interests

The authors declare no competing financial interests.

References

[1] C. Horscroft, S. Ennis, R. J. Pengelly, T. J. Sluckin, and A. Collins, "Sequencing era methods for identifying signatures of selection in the genome," *Briefings in Bioinformatics*, vol. bby064, 2018.

[2] H. Elding, W. Lau, D. M. Swallow, and N. Maniatis, "Refinement in localization and identification of gene regions associated with crohn disease," *American Journal of Human Genetics*, vol. 92, no. 1, pp. 107–113, 2013.

[3] A. Auton and G. McVean, "Recombination rate estimation in the presence of hotspots," *Genome Research*, vol. 17, no. 8, pp. 1219–1227, 2007.

[4] T.-Y. Kuo, W. Lau, and A. R. Collins, "LDMAP: the construction of high-resolution linkage disequilibrium maps of the human genome," in *Linkage Disequilibrium and Association Mapping* (A. R. Collins, ed.), vol. 376 of *Methods in Molecular Biology*, pp. 47–57, Humana Press, 2007.

[5] W. Tapper, A. Collins, J. Gibson, N. Maniatis, S. Ennis, and N. E. Morton, "A map of the human genome in linkage disequilibrium units," *Proc Natl Acad Sci U S A*, vol. 102, no. 33, pp. 11835–9, 2005.

[6] S. Service, J. DeYoung, M. Karayiorgou, J. L. Roos, H. Pretorious, G. Bedoya, J. Ospina, A. Ruiz-Linares, A. Macedo, J. A. Palha, *et al.*, "Magnitude and distribution of linkage disequilibrium in population isolates and implications for genome-wide association studies," *Nature Genetics*, vol. 38, no. 5, p. 556, 2006.

[7] R. J. Pengelly, W. Tapper, J. Gibson, M. Knut, R. Tearle, A. Collins, and S. Ennis, "Whole genome sequences are required to fully resolve the linkage disequilibrium structure of human populations," *BMC Genomics*, vol. 16, p. 666, 2015.

[8] D. Gurdasani, T. Carstensen, F. Tekola-Ayele, L. Pagani, I. Tachmazidou, K. Hatzikotoulas, S. Karthikeyan, L. Iles, M. O. Pollard, A. Choudhury, *et al.*, "The african genome variation project shapes medical genetics in africa," *Nature*, vol. 517, no. 7534, p. 327, 2015.

[9] A. Vergara-Lope, S. Ennis, I. Vorechovsky, R. J. Pengelly, and A. Collins, "Heterogeneity in the extent of linkage disequilibrium among exonic, intronic, non-coding rna and intergenic chromosome regions," *European Journal of Human Genetics*, p. 1, 2019.

[10] G. A. Erikson, D. L. Bodian, M. Rueda, B. Molparia, E. R. Scott, A. A. Scott-Van Zeeland, S. E. Topol, N. E. Wineinger, J. E. Niederhuber, E. J. Topol, *et al.*, "Whole-genome sequencing of a healthy aging cohort," *Cell*, vol. 165, no. 4, pp. 1002–1011, 2016.

[11] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. de Bakker, M. J. Daly, and P. C. Sham, "Plink: a tool set for whole-genome association and population-based linkage analyses," *American Journal of Human Genetics*, vol. 81, no. 3, pp. 559–75, 2007.

[12] C. Bhérer, C. L. Campbell, and A. Auton, "Refined genetic maps reveal sexual dimorphism in human meiotic recombination at multiple scales," *Nature Communications*, vol. 8, p. 14994, 2017.

[13] A. R. Quinlan and I. M. Hall, "Bedtools: a flexible suite of utilities for comparing genomic features," *Bioinformatics*, vol. 26, no. 6, pp. 841–842, 2010.

[14] A. Collins and W. Lau, "Chromscan: genome-wide association using a linkage disequilibrium map," *Journal of Human Genetics*, vol. 53, no. 2, pp. 121–126, 2008.

Data Citations

Bibliographic information for the data records described in the manuscript.

1. Vergara-Lope A, Jabal MR, Horscroft C, Ennis S, Collins A & Pengelly RJ. *Figshare* 10.6084/m9.figshare.7850882 (2019).