An updated map of GRCh38 linkage disequilibrium blocks based on European ancestry data

James W. MacDonald[1*], Tabitha Harrison[2], Theo K. Bammler[1], Nicholas Mancuso[3,4] and Sara Lindström[2,5]

[1]Department of Environmental and Occupational Health Sciences, University of Washington, Seattle, WA., [2] Department of Epidemiology, University of Washington, Seattle, WA.,[3] Department of Population and Public Health Sciences, Keck School of Medicine, University of Southern California, Los Angeles, CA, [4]Center for Genetic Epidemiology, Keck School of Medicine, University of Southern California, Los Angeles, CA, [5] Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA.

*To whom correspondence should be addressed.

Abstract

A map of approximately independent linkage disequilibrium (LD) blocks has many uses in statistical genetics. Current LD block maps are based on sparse recombination maps and only available for GRCh37 (hg19) and prior genome assemblies. We generated LD blocks for European (EUR) ancestry populations using a new large recombination map in GRCh38.

1 Introduction

With an increasing number of large-scale genome-wide association studies (GWAS) relying on meta-analysis, many newly developed statistical methods circumvent the need for individual-level data, and instead require GWAS summary statistics only. These methods often rely on an external reference panel such as the 1,000 Genomes Project (The 1000 Genomes, 2015) to model population-specific linkage disequilibrium (LD) patterns. Further, approaches to study the local genetic architecture utilize population-specific maps of approximately independent LD blocks (Shi *et al.*, 2019), and methods to build these blocks have been described previously (Berisa and Pickrell, 2016). However, current blocks were generated based on GRCh37 coordinates, and as more data are mapped to the GRCh38 genome assembly, updated block coordinates are needed. One straightforward solution is to convert existing GRCh37 LD blocks to GRCh38 positions using reources such as liftOver (Kent *et al.,* 2002). However, this approach produces large unmappable regions, as liftOver aims to map short genomic sequences between genome builds, with longer genomic regions becoming fragmented and scattered across multiple chromosomes. To overcome this issue, we estimated new approximately independent LD blocks in European ancestry populations using a recently generated genome-wide recombination map based on parent-child pairs from Iceland (Halldorsson *et al.,* 2019).

2 Methods

We applied LDetect (Berisa and Pickrell, 2016) (https://bitbucket.org/nygcresearch/ldetect/src/master/), which utilizes population-specific variants and a recombination map file to generate LD blocks. We used a recent high-resolution recombination map based on more than 115,000 Icelandic individuals (Halldorsson *et al.,* 2019). This genetic map has a higher resolution than the previously used recombination map from HapMap, with a sex-averaged resolution of 683 bp as compared to 1,324 for HapMap. Further, the Icelandic recombination map is natively based on GRCh38. We downloaded VCF files from the 1000 Genomes GRCh38 December 2018 biallelic single nucleotide variant (SNV) data

(http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000_genomes_project/release/20181203_biallelic_SNV/) and removed variants with a minor allele frequency < 0.01. We interpolated genetic distances for the remaining variants using the recombination map and a custom R script, which is included in the scripts directory of our GitHub repository. We then partitioned each chromosome by first generating naïve blocks containing 5,000 SNVs each. To ensure that these naïve blocks did not terminate within regions of high LD, we extended the end of each partition until a shrunken covariance estimator (Wen and Stephens, 2010) between the first and last SNV was negligible (S_{ij}<1.5x10^{-8}).

The resulting partitions had a median length of 1.4 Mb with a median 300 Kb overlap and allowed us to efficiently compute the partition specific SNV covariance in parallel. When computing the covariance, we used only European ancestry sub-populations (TSI, IBS, CEU, GBR) to improve consistency with the recombination map based on European ancestry individuals. We computed the covariance minima across each partition and then selected block-specific breakpoints using a low-pass filter with local search algorithm (the fourier-ls algorithm). A more detailed description of our methods, including all our code, as well as coordinates for approximately independent LD blocks in BED format, can be found in our GitHub repository (https://github.com/jmacdon/LDblocks_GRCh38).

3 Results

Overall block statistics are presented in Table 1. We generated a total of 1,361 LD blocks, which is fewer than the existing 1,703 GRCh37 European ancestry LD blocks. The GRCh38 blocks were also longer and more variable in length (Figure 1). As compared to GRCh37, the block lengths for GRCh38 have a median 20% increase in both block length and median absolute deviation (MAD). While these updated LD blocks are useful for European-ancestry populations, a limitation of this work is the lack of GRCh38 LD blocks for other ancestries. Unfortunately, no high-resolution GRCh38 genetic maps currently exist for other populations. LD blocks for African and Asian populations, based on GRCh37, are available on the LDetect bitbucket data repository in BED format (https://bitbucket.org/nygcresearch/ldetect-data/src/master/).

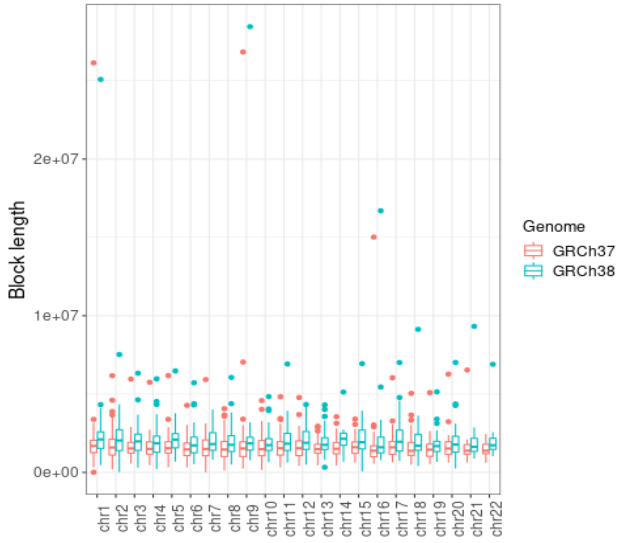

Figure 1: Chromosome block length ($\log_{10}$(bp)) distribution for European genetic ancestry population, by genomic build.

Table 1. Length of LD blocks per chromosome in European ancestry populations. The long maximum block lengths for chromosomes 1, 9, 12, and 18 are due to inclusion of centromeric regions.

		Block lengths			
Chr	Number of independent blocks	Minimum	Median	Mean	Maximum
1	105	462,196	2,107,324	2,370,759	25,079,459
2	112	16,711	2,035,590	2,162,153	7,523,344
3	97	304,389	1,979,216	2,043,080	6,328,687
4	99	213,473	1,864,556	1,920,946	5,973,432
5	86	687,633	2,092,825	2,108,730	6,475,954
6	91	528,636	1,726,746	1,874,950	5,719,137
7	79	820,310	1,810,865	2,016,499	4,018,159
8	75	502,057	1,764,896	1,933,396	6,059,508
9	60	768,908	1,853,721	2,303,518	28,445,823
10	71	651,869	1,741,079	1,883,871	4,838,136
11	67	634,461	1,827,928	2,014,146	6,922,454
12	65	462,196	2,107,324	2,370,759	25,079,459
13	50	16,711	2,035,590	2,162,153	7,523,344
14	44	496,123	1,886,408	2,050,063	4,333,981
15	39	326,640	1,758,701	1,923,637	4,308,493
16	42	678,246	2,143,594	2,064,639	5,125,362
17	37	56,827	1,929,831	2,178,751	6,940,834
18	39	785,996	1,604,375	2,147,863	16,695,683
19	32	725,130	1,954,139	2,246,547	7,011,010
20	31	413,976	1,705,935	2,057,086	9,131,209
21	20	666,967	1,680,202	1,828,644	5,140,729
22	20	254,307	1,779,760	2,073,227	7,014,334

Funding

This work was supported by the National Institute of Health (CA194393), as well as the National Institute of Environmental Health Services (P30-ES007033).

Conflict of Interest: none declared.

References

The 1000 Genomes Project Consortium. A global reference for human genet-ic variation. Nature 526, 68–74 (2015). PMID: 26432245

Berisa, T. and Pickrell, J.K., 2016. Approximately independent linkage disequilibrium blocks in human populations. Bioinformatics, 32(2), p.283. PMID: 26395773

Halldorsson, B.V., Palsson, G., Stefansson, O.A., Jonsson, H., Hardarson, M.T., Eggertsson, H.P., Gunnarsson, B., Oddsson, A., Halldorsson, G.H., Zink, F. and Gudjonsson, S.A., 2019. Characterizing mutagenic effects of recombination through a sequence-level genetic map. Science, 363(6425). PMID: 30679340.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M. and Haussler, D., 2002. The human genome browser at UCSC. Genome research, 12(6), pp.996-1006. PMID: 12045153

Shi H, Mancuso N, Spendlove S, Pasaniuc B. Local Genetic Correlation Gives Insights into the Shared Genetic Architecture of Complex Traits. Am J Hum Genet. 2017 Nov 2;101(5):737-751. doi: 10.1016/j.ajhg.2017.09.022. PMID: 29100087

Wen, X. and Stephens, M., 2010. Using linear predictors to impute allele frequencies from summary or pooled genotype data. The annals of applied statistics, 4(3), p.1158. PMID: 21479081