Alexander and Machiela *BMC Bioinformatics* (2020) 21:14
https://doi.org/10.1186/s12859-020-3340-1

BMC Bioinformatics

SOFTWARE Open Access

LDpop: an interactive online tool to calculate and visualize geographic LD patterns

T. A. Alexander[1] and M. J. Machiela[2*]

Abstract

Background: Linkage disequilibrium (LD)—the non-random association of alleles at different loci—defines population-specific haplotypes which vary by genomic ancestry. Assessment of allelic frequencies and LD patterns from a variety of ancestral populations enables researchers to better understand population histories as well as improve genetic understanding of diseases in which risk varies by ethnicity.

Results: We created an interactive web module which allows for quick geographic visualization of linkage disequilibrium (LD) patterns between two user-specified germline variants across geographic populations included in the 1000 Genomes Project. Interactive maps and a downloadable, sortable summary table allow researchers to easily compute and compare allele frequencies and LD statistics of dbSNP catalogued variants. The geographic mapping of each SNP's allele frequencies by population as well as visualization of LD statistics allows the user to easily trace geographic allelic correlation patterns and examine population-specific differences.

Conclusions: LDpop is a free and publicly available cross-platform web tool which can be accessed online at https://ldlink.nci.nih.gov/?tab=ldpop

Keywords: Linkage disequilibrium, Genome-wide association, Geographical visualization, 1000 Genomes Project

Background

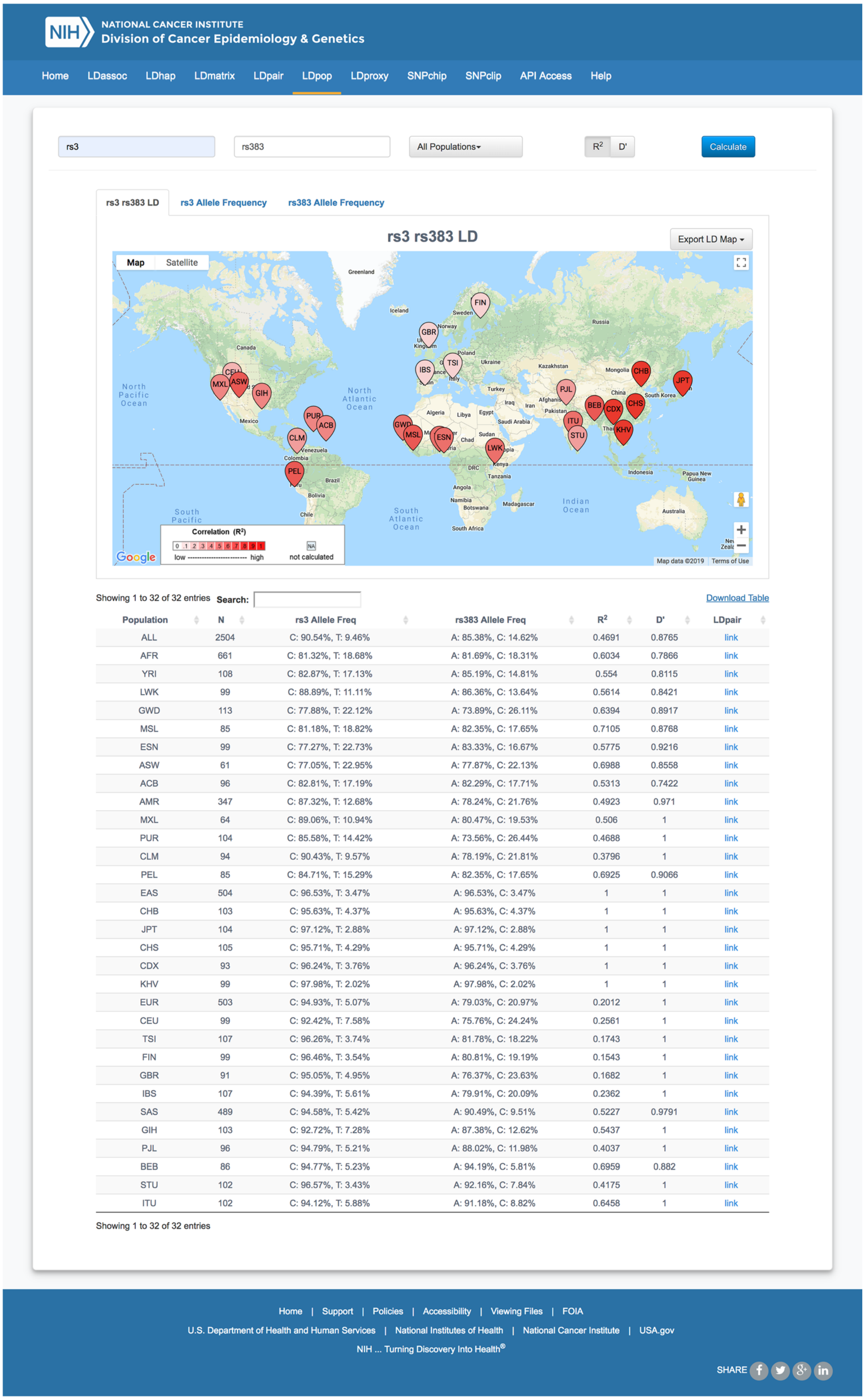
Linkage disequilibrium (LD)—the non-random association of alleles at different loci—defines population-specific haplotypes which vary by genomic ancestry [1]. Assessment of allelic frequencies and LD patterns from a variety of ancestral populations enables researchers to better understand population histories as well as improve genetic understanding of diseases in which risk varies by ethnicity. For example, genome-wide association studies (GWAS) identify germline variation associated with disease susceptibility but need to account for ancestry-specific differences in LD patterns when designing the study, analyzing markers and interpreting findings. While population geneticists have developed many datasets (e.g., 1000 Genomes Project, HapMap) [2, 3] and tools (e.g., Geography of Genetic Variants Browser) [4] to investigate differences in allelic frequencies by population group, to date no tool exists to easily explore and visualize LD patterns across 1000 Genomes population groups.

Implementation

LDpop is an online module designed to allow researchers to query LD patterns of two variants across ancestral populations of interest. LDpop estimates allele frequencies and measures of LD (D' and R^2) for each included population. The reference genetic data is from the 1000 Genomes Project Phase 3, which includes sequencing data for 2504 individuals in 26 ancestral populations which are divided into 5 super populations (e.g., African, Ad-Mixed American, East Asian, European, and South Asian) [2]. The 1000G data are available for public download in VCF format (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/).

LDpop is written in Python (2.7.15) and runs on a web-accessible virtual machine with UNIX operating system. The genomic coordinates are retrieved for

* Correspondence: mitchell.machiela@nih.gov
[2]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD 20892, USA
Full list of author information is available at the end of the article



Population	N	rs3 Allele Freq	rs383 Allele Freq	R^2	D'	LDpair
ALL	2504	C: 90.54%, T: 9.46%	A: 85.38%, C: 14.62%	0.4691	0.8765	link
AFR	661	C: 81.32%, T: 18.68%	A: 81.69%, C: 18.31%	0.6034	0.7866	link
YRI	108	C: 82.87%, T: 17.13%	A: 85.19%, C: 14.81%	0.554	0.8115	link
LWK	99	C: 88.89%, T: 11.11%	A: 86.36%, C: 13.64%	0.5614	0.8421	link
GWD	113	C: 77.88%, T: 22.12%	A: 73.89%, C: 26.11%	0.6394	0.8917	link
MSL	85	C: 81.18%, T: 18.82%	A: 82.35%, C: 17.65%	0.7105	0.8768	link
ESN	99	C: 77.27%, T: 22.73%	A: 83.33%, C: 16.67%	0.5775	0.9216	link
ASW	61	C: 77.05%, T: 22.95%	A: 77.87%, C: 22.13%	0.6988	0.8558	link
ACB	96	C: 82.81%, T: 17.19%	A: 82.29%, C: 17.71%	0.5313	0.7422	link
AMR	347	C: 87.32%, T: 12.68%	A: 78.24%, C: 21.76%	0.4923	0.971	link
MXL	64	C: 89.06%, T: 10.94%	A: 80.47%, C: 19.53%	0.506	1	link
PUR	104	C: 85.58%, T: 14.42%	A: 73.56%, C: 26.44%	0.4688	1	link
CLM	94	C: 90.43%, T: 9.57%	A: 78.19%, C: 21.81%	0.3796	1	link
PEL	85	C: 84.71%, T: 15.29%	A: 82.35%, C: 17.65%	0.6925	0.9066	link
EAS	504	C: 96.53%, T: 3.47%	A: 96.53%, C: 3.47%	1	1	link
CHB	103	C: 95.63%, T: 4.37%	A: 95.63%, C: 4.37%	1	1	link
JPT	104	C: 97.12%, T: 2.88%	A: 97.12%, C: 2.88%	1	1	link
CHS	105	C: 95.71%, T: 4.29%	A: 95.71%, C: 4.29%	1	1	link
CDX	93	C: 96.24%, T: 3.76%	A: 96.24%, C: 3.76%	1	1	link
KHV	99	C: 97.98%, T: 2.02%	A: 97.98%, C: 2.02%	1	1	link
EUR	503	C: 94.93%, T: 5.07%	A: 79.03%, C: 20.97%	0.2012	1	link
CEU	99	C: 92.42%, T: 7.58%	A: 75.76%, C: 24.24%	0.2561	1	link
TSI	107	C: 96.26%, T: 3.74%	A: 81.78%, C: 18.22%	0.1743	1	link
FIN	99	C: 96.46%, T: 3.54%	A: 80.81%, C: 19.19%	0.1543	1	link
GBR	91	C: 95.05%, T: 4.95%	A: 76.37%, C: 23.63%	0.1682	1	link
IBS	107	C: 94.39%, T: 5.61%	A: 79.91%, C: 20.09%	0.2362	1	link
SAS	489	C: 94.58%, T: 5.42%	A: 90.49%, C: 9.51%	0.5227	0.9791	link
GIH	103	C: 92.72%, T: 7.28%	A: 87.38%, C: 12.62%	0.5437	1	link
PJL	96	C: 94.79%, T: 5.21%	A: 88.02%, C: 11.98%	0.4037	1	link
BEB	86	C: 94.77%, T: 5.23%	A: 94.19%, C: 5.81%	0.6959	0.882	link
STU	102	C: 96.57%, T: 3.43%	A: 92.16%, C: 7.84%	0.4175	1	link
ITU	102	C: 94.12%, T: 5.88%	A: 91.18%, C: 8.82%	0.6458	1	link

Showing 1 to 32 of 32 entries

Fig. 1 Example of an LDpop interactive map and table. Selected tab displays a map of R^2 for rs3 and rs383 for all 1000 Genomes Project populations. Numeric data on sample size, allele frequency and LD measures are displayed in the table at the bottom of the screen capture

each query variant from an indexed MongoDb database of dbSNP version 151 and subsequently extracted from the phased 1000 Genomes Project variant call format (VCF) file using Tabix (0.2.5). LDpop uses the Google Maps API to produce the interactive geographic mapping for each population using latitude and longitudinal coordinates for each 1000 Genomes Project ancestral population. The LDpop web-accessible page is programed in HTML5 for cross-browser and cross-platform compatibility and is part of the larger LDlink collection of LD web tools [5, 6]. All code for LDpop is available from out GitHub repository: https://github.com/CBIIT/nci-webtools-dceg-linkage/.

Results

LDpop takes as input two dbSNP reference SNP numbers (rsIDs), a selection of desired populations from the 1000 Genomes Project, and a choice of which LD statistic (D′ or R^2) to report for the geographic mapping. LDpop supports queried dbSNP variants which are biallelic including both single nucleotide polymorphism (SNP) and small insertion and deletion (indel) queries. LDpop allows the user to specify any subset of populations from the subpopulations, super populations, and all populations, they are interested in examining for the analysis.

LDpop produces three geographic maps and one sortable, filterable table as output (Fig. 1). For each queried variant, the allele frequency is calculated for every population selected and the frequency percentage is plotted over the population's approximate geographic coordinates as a colored pin with deeper blue colors indicating higher allele frequencies. This allows the investigators to easily calculate and visualize changes in allele frequency across ancestral populations for each variant. A LD map is also produced displaying a computed LD statistic (D' or R^2) for the two query variants for every population selected. The mapped data point is colored in proportion to the gradient shown in the legend, with darker red signifying a higher degree of LD. All geographic mapping utilizes the Google Maps API for smooth and rapid performance. The interactive summary table at the bottom of the page has a row for each selected 1000 Genomes Project population and displays data in the number of samples in each population, allele frequencies for each variant, and calculated LD values (D′ and R^2). This table is sortable by column and has a search bar to quickly navigate through it. The table is also downloadable as a text file for local storage and future data integration and analysis.

Conclusions

LDpop is an online module designed to allow researchers to query LD patterns of two variants across ancestral populations of interest. It is designed to allow users to easily calculate and geographically visualize these LD patterns and changes in allele frequency across ancestral populations. This web tool is freely available and can be accessed at https://ldlink.nci.nih.gov/?tab=ldpop.

Abbreviations

GWAS: Genome-Wide Association Studies; Indels: Insertion/Deletion; LD: Linkage Disequilibrium; rsIDs: reference SNP numbers; SNP: Single Neucleotide Polymorphism; VCF: Variant call format

Acknowledgements

Special thanks to Kevin Jiang, Kai-Ling Chen, and Ye Wu from the Center for Biomedical Informatics and Information Technology (CBIIT) for technical assistance and web development as well as the numerous testers from DCEG that provided feedback and ensured a stable and user-friendly interface.

Availability and requirements

Project Name: LDpop.
Project Homepage: https://ldlink.nci.nih.gov/?tab=ldpop
Project Repository: https://github.com/CBIIT/nci-webtools-dceg-linkage/tree/master/LDlink
Operating Systems: Cross-platform compatible.
Programming Language: Python (2.7.15), HTML5.

Ethics approval and consent

Not Applicable.

Authors' contributions

TAA coded the Python script for LDpop and drafted the manuscript. MJM conceptually designed the project, assisted in troubleshooting the code and drafted the manuscript. All authors read and approved the final manuscript.

Funding

LDpop development was funded by the Intramural Research Program of the National Cancer Institute.

Availability of data and materials

The LDpop web tool is freely available at https://ldlink.nci.nih.gov/?tab=ldpop. The 1000G data are available for public download in VCF format (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/).

Consent for publication

Not Applicable.

Competing interests

None Declared.

Author details

[1]Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD 20740, USA. [2]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD 20892, USA.

Received: 20 September 2019 Accepted: 3 January 2020
Published online: 10 January 2020

References

1. Reich D, et al. Linkage disequilibrium in the human genome. Nat. 2001;411:199–204.
2. Genomes Project, et al. An integrated map of genetic variation from 1,092 human genomes. Nat. 2012;491:56–65.
3. International HapMap Consortium, et al. The international HapMap project. Nat. 2003;426:789–96.
4. Marcus JH, Novembre J. Visualizing the geography of genetic variants. Bioinformatics. 2016;33(4):594–5.

5. Machiela MJ, Chanock SJ. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. Bioinformatics. 2015;31(21):3555–7.
6. Machiela MJ, Chanock SJ. LDassoc: an online tool for interactively exploring genome-wide association study results and prioritizing variants for functional investigation. Bioinformatics. 2017;34(5):887–9.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.