

# UCLA

## UCLA Previously Published Works

### Title

TOP-LD: A tool to explore linkage disequilibrium with TOPMed whole-genome sequence data

### Permalink

<https://escholarship.org/uc/item/0dg9r8n3>

### Journal

American Journal of Human Genetics, 109(6)

### ISSN

0002-9297

### Authors

Huang, Le  
Rosen, Jonathan D  
Sun, Quan  
et al.

### Publication Date

2022-06-01

### DOI

10.1016/j.ajhg.2022.04.006

Peer reviewed

# TOP-LD: A tool to explore linkage disequilibrium with TOPMed whole-genome sequence data

## Authors

Le Huang, Jonathan D. Rosen, Quan Sun, ...,  
Alex P. Reiner, Paul L. Auer, Yun Li

## Correspondence

[pauer@mcw.edu](mailto:pauer@mcw.edu) (P.L.A.),  
[yunli@med.unc.edu](mailto:yunli@med.unc.edu) (Y.L.)



Huang et al., 2022, *The American Journal of Human Genetics* 109, 1175–1181

June 2, 2022 © 2022 The Authors.

<https://doi.org/10.1016/j.ajhg.2022.04.006>

# TOP-LD: A tool to explore linkage disequilibrium with TOPMed whole-genome sequence data

Le Huang,<sup>1,19</sup> Jonathan D. Rosen,<sup>2,19</sup> Quan Sun,<sup>2,19</sup> Jiawen Chen,<sup>2</sup> Marsha M. Wheeler,<sup>3</sup> Ying Zhou,<sup>4</sup> Yuan-I Min,<sup>5</sup> Charles Kooperberg,<sup>4</sup> Matthew P. Conomos,<sup>6</sup> Adrienne M. Stilp,<sup>6</sup> Stephen S. Rich,<sup>7</sup> Jerome I. Rotter,<sup>8</sup> Ani Manichaikul,<sup>7</sup> Ruth J.F. Loos,<sup>9,10</sup> Eimear E. Kenny,<sup>9</sup> Thomas W. Blackwell,<sup>11</sup> Albert V. Smith,<sup>11</sup> Goo Jun,<sup>12</sup> Fritz J. Sedlazeck,<sup>13</sup> Ginger Metcalf,<sup>13</sup> Eric Boerwinkle,<sup>14</sup> NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium, Laura M. Raffield,<sup>15</sup> Alex P. Reiner,<sup>16,4</sup> Paul L. Auer,<sup>17,\*</sup> and Yun Li<sup>2,15,18,\*</sup>

## Summary

Current publicly available tools that allow rapid exploration of linkage disequilibrium (LD) between markers (e.g., HaploReg and LDlink) are based on whole-genome sequence (WGS) data from 2,504 individuals in the 1000 Genomes Project. Here, we present TOP-LD, an online tool to explore LD inferred with high-coverage (~30×) WGS data from 15,578 individuals in the NHLBI Trans-Omics for Precision Medicine (TOPMed) program. TOP-LD provides a significant upgrade compared to current LD tools, as the TOPMed WGS data provide a more comprehensive representation of genetic variation than the 1000 Genomes data, particularly for rare variants and in the specific populations that we analyzed. For example, TOP-LD encompasses LD information for 150.3, 62.2, and 36.7 million variants for European, African, and East Asian ancestral samples, respectively, offering 2.6- to 9.1-fold increase in variant coverage compared to HaploReg 4.0 or LDlink. In addition, TOP-LD includes tens of thousands of structural variants (SVs). We demonstrate the value of TOP-LD in fine-mapping at the *GGT1* locus associated with gamma glutamyltransferase in the African ancestry participants in UK Biobank. Beyond fine-mapping, TOP-LD can facilitate a wide range of applications that are based on summary statistics and estimates of LD. TOP-LD is freely available online.

Linkage disequilibrium (LD), i.e., the non-random association of alleles at different variant sites in a given population, is an important genetic phenomenon. Patterns of LD between genetic markers can be leveraged to gain insights in a variety of different applications, from population genetic research to disease association studies.<sup>1,2</sup> With the growth of whole-genome sequencing (WGS) and high-throughput array and genotype imputation technologies, resources for calculating LD across populations have expanded to encompass multiple populations at variant sites with increasingly rare frequencies.<sup>3-6</sup> Due to the centrality of LD in a host of applications, multiple tools exist for querying LD between genetic markers in different populations. The current most widely used LD lookup tools, HaploReg<sup>7</sup> and LDlink,<sup>8</sup> base their LD estimates on the 1000 Genomes data. Specifically, HaploReg uses phase

1 and LDlink uses phase 3 1000 Genomes data. Although the 1000 Genomes data contains LD information on >99% of genetic markers with minor allele frequency (MAF) > 1% in a variety of populations,<sup>4</sup> there remains a dearth of publicly available information on LD between markers with MAF < 1%. We have created a new LD lookup tool (called “TOP-LD”), in the spirit of HaploReg and LDlink, that is based on deep (30×) WGS data from the NHLBI Trans-Omics for Precision Medicine (TOPMed) Program. Because the TOPMed data contain much larger sample sizes with greater depth of sequencing than the 1000 Genomes project, TOP-LD provides a significant upgrade in LD information availability, specifically by including single-nucleotide variants and small indels (referred to hereafter simply as “SNVs”) with MAF < 1% as well as structural variants (SVs). Here, we describe the data and

<sup>1</sup>Curriculum in Bioinformatics and Computational Biology, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA; <sup>2</sup>Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA; <sup>3</sup>Department of Genome Sciences, University of Washington, Seattle, WA 98105, USA; <sup>4</sup>Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA; <sup>5</sup>Department of Medicine, University of Mississippi Medical Center, Jackson, MS 39216, USA; <sup>6</sup>Department of Biostatistics, University of Washington, Seattle, WA 98105, USA; <sup>7</sup>Center for Public Health Genomics, Department of Public Health Sciences, University of Virginia School of Medicine, Charlottesville, VA 22908, USA; <sup>8</sup>The Institute for Translational Genomics and Population Sciences, Department of Pediatrics, The Lundquist Institute for Biomedical Innovation at Harbor-UCLA Medical Center, Torrance, CA 90502, USA; <sup>9</sup>The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York City, NY 10029, USA; <sup>10</sup>Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, 2200 Copenhagen, Denmark; <sup>11</sup>TOPMed Informatics Research Center, University of Michigan, Department of Biostatistics, Ann Arbor, MI 48109, USA; <sup>12</sup>Human Genetics Center, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA; <sup>13</sup>Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX 77030, USA; <sup>14</sup>Human Genetics Center, Department of Epidemiology, Human Genetics, and Environmental Sciences, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA; <sup>15</sup>Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA; <sup>16</sup>Department of Epidemiology, University of Washington, Seattle, WA 98195, USA; <sup>17</sup>Division of Biostatistics, Institute for Health and Equity, and Cancer Center, Medical College of Wisconsin, Milwaukee, WI 53226, USA; <sup>18</sup>Department of Computer Science, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

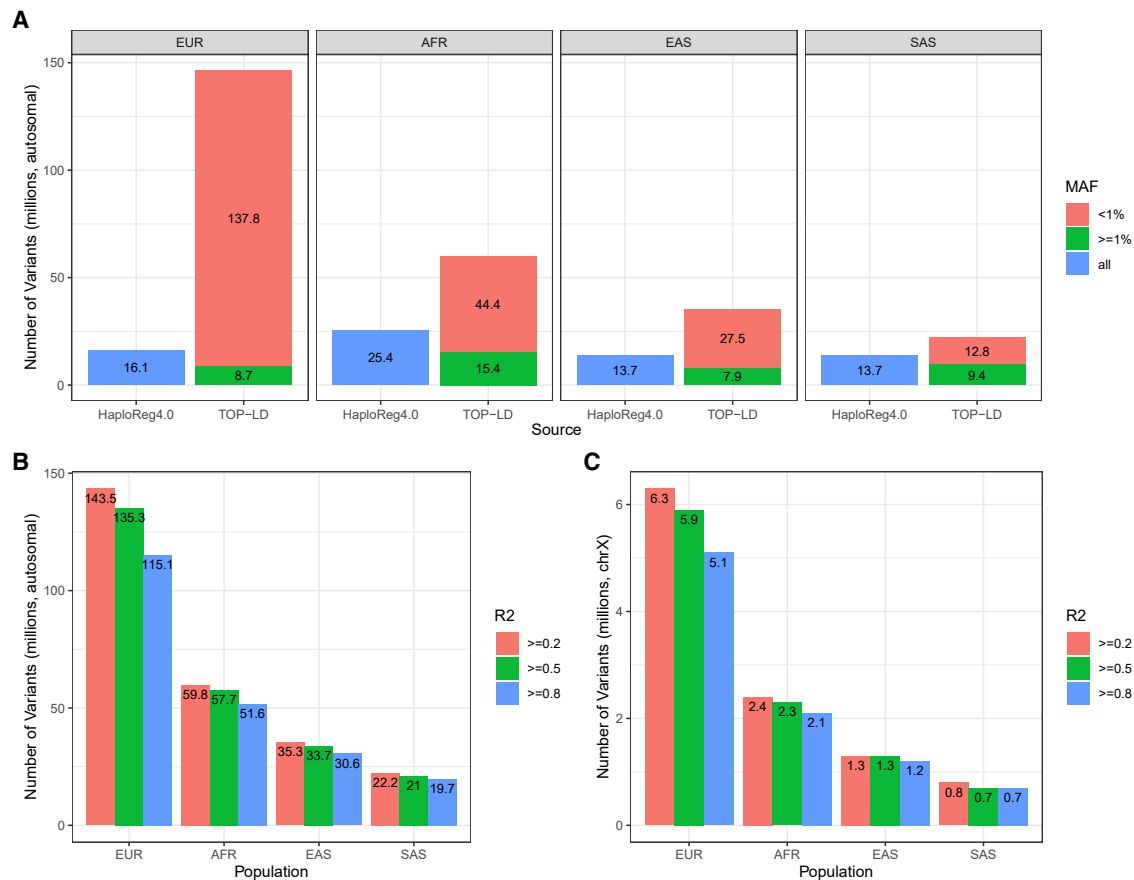
<sup>19</sup>These authors contributed equally

\*Correspondence: [pauer@mcw.edu](mailto:pauer@mcw.edu) (P.L.A.), [yunli@med.unc.edu](mailto:yunli@med.unc.edu) (Y.L.)

<https://doi.org/10.1016/j.ajhg.2022.04.006>

© 2022 The Authors. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).





**Figure 1. Number of variants included in TOP-LD**

(A) Comparison of autosomal variants with HaploReg 4.0 by population. Blue bars on the left show total number of autosomal variants in HaploReg4.0. Green and red indicate common (MAF  $\geq 1\%$ ) and uncommon (MAF  $< 1\%$ ) autosomal variants in TOP-LD. Note that HaploReg4.0 provides LD for ASN (Asian) with no separate information for EAS and SAS. Therefore, we used the same 13.7 million ASN variants for comparison in both EAS and SAS.

(B) Number of autosomal variants in TOP-LD breaking down by LD  $R^2$  threshold. The majority of the variants have at least one LD proxy with  $R^2 \geq 0.8$ .

(C) Number of chrX variants in TOP-LD breaking down by LD  $R^2$  threshold.

(Note: LD information downloaded from HaploReg4.0 does not contain chromosome X. Therefore, we compared TOP-LD with HaploReg4.0 only for autosomal variants).

methods that went into creating TOP-LD along with specific examples of how TOP-LD can provide essential information that is missed by HaploReg and LDlink.

We used TOPMed WGS data<sup>6</sup> from the following four cohorts: BioMe Biobank (BioMe), the Multi-Ethnic Study of Atherosclerosis (MESA), the Jackson Heart Study (JHS), and the Women's Health Initiative (WHI). We aimed to provide LD estimates for genetically homogeneous groups of individuals from one of the following four ancestral populations: European (EUR), African (AFR), East Asian (EAS), and South Asian (SAS). To select appropriate samples, we first inferred local and global ancestry for all participants in these four cohorts by using RFMix,<sup>9</sup> with reference populations including five ancestral groups, namely African, Native American, East Asian, European, and South Asian. After local ancestry inference, we then retained only TOPMed samples with  $>90\%$  estimated ancestry from a single population, as estimated via RFMix. We further removed related individuals by using a stringent kinship

coefficient threshold of  $2^{-5.5}$  obtained via PC-Relate.<sup>10</sup> This threshold of  $2^{-5.5}$  removes pairs within as far as fifth degree relationship. The final dataset included 1,335 unrelated individuals of African, 844 of East Asian, 13,160 of European, and 239 of South Asian ancestry for pairwise LD inference. Regarding variants, we started with all TOPMed freeze 8 polymorphic variants that passed quality control and retained multi-allelic variants or multiple entries at the same position, resulting in a total of 23.0–153.0 million SNVs in each of the ancestral groups (Figure 1A, Table S1).

We inferred LD separately within each of the four ancestral groups, for all pairs of variants within 1 Mb of each other, and retained LD pairs meeting a minimum  $R^2$  threshold of 0.2. The reported  $R^2$  between two variants is the squared Pearson correlation coefficient between their phased haplotypes, where phasing was performed with Eagle 2.4 for all polymorphic variants, similar to phasing of the freeze 5 data.<sup>6</sup> No minimum minor allele count thresholding was used, that

**Table 1. Summary of SVs by population**

Population	Number of SVs	Number of SVs in LD w/SNVs <sup>a</sup>	Number of SVs with MAF < 0.01
EUR	79,004	16,301	69,011
AFR	44,859	15,151	27,978
SAS	16,511	10,392	7,292
EAS	20,789	7,498	12,902

<sup>a</sup>Number of SVs having at least one SNV LD tag with  $R^2 \geq 0.8$ .

is, even singletons in our sample were included in LD calculations. We also report the direction of each association as either positive (+) or negative (–) on the basis of the sign of the Pearson correlation coefficient between the corresponding pair of reference (REF) alleles. In addition to  $R^2$ , we also report D-prime statistics for each pair of variants meeting the  $R^2$  of 0.2.

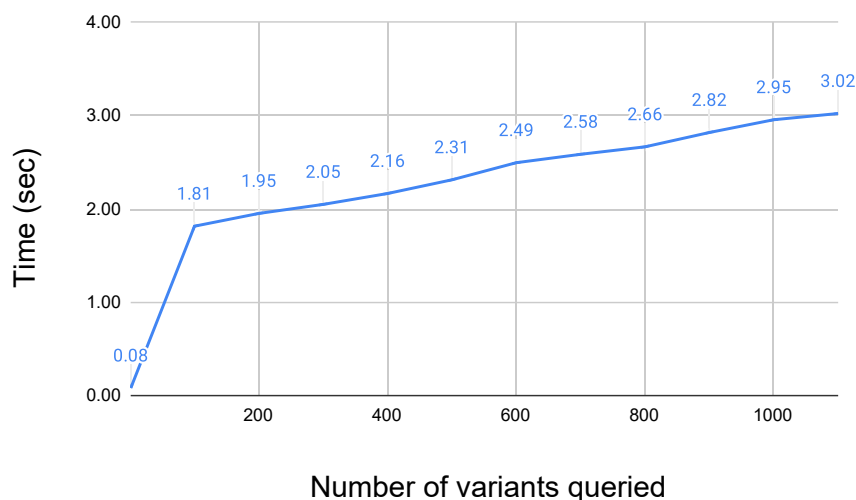
We filtered chromosome X to exclude the pseudo-autosomal regions: PAR1 (bp 10,001–2,781,479, GRCh38) and PAR2 (bp 155,701,383–156,030,895, GRCh38). Variants that were not coded as homozygous in the males were excluded from the LD calculations. We inferred LD for the remaining variants by using a total of  $2F + M$  haplotypes, where F and M are the numbers of females and males, respectively.

The TOPMed structural variant (SV) call-set freeze 1 was merged with a reduced TOPMed SNV call-set where SNVs with  $MAF < 0.1\%$  were filtered out before merging, and then the merged SV-SNV dataset was phased with Eagle2.<sup>11</sup> SVs with  $>10\%$  missingness were removed prior to phasing. For each ancestry group, we included 16.5–79K SVs (deletions, duplications, and inversion) with the majority being lower frequency (e.g., 7–69K with  $MAF < 1\%$ ) (Table 1). LD values were subsequently estimated as the squared Pearson correlation coefficient between the corresponding pair of phased alleles.

TOPMed LD information was then loaded into the TOP-LD website, which is powered by a combination of MySQL,

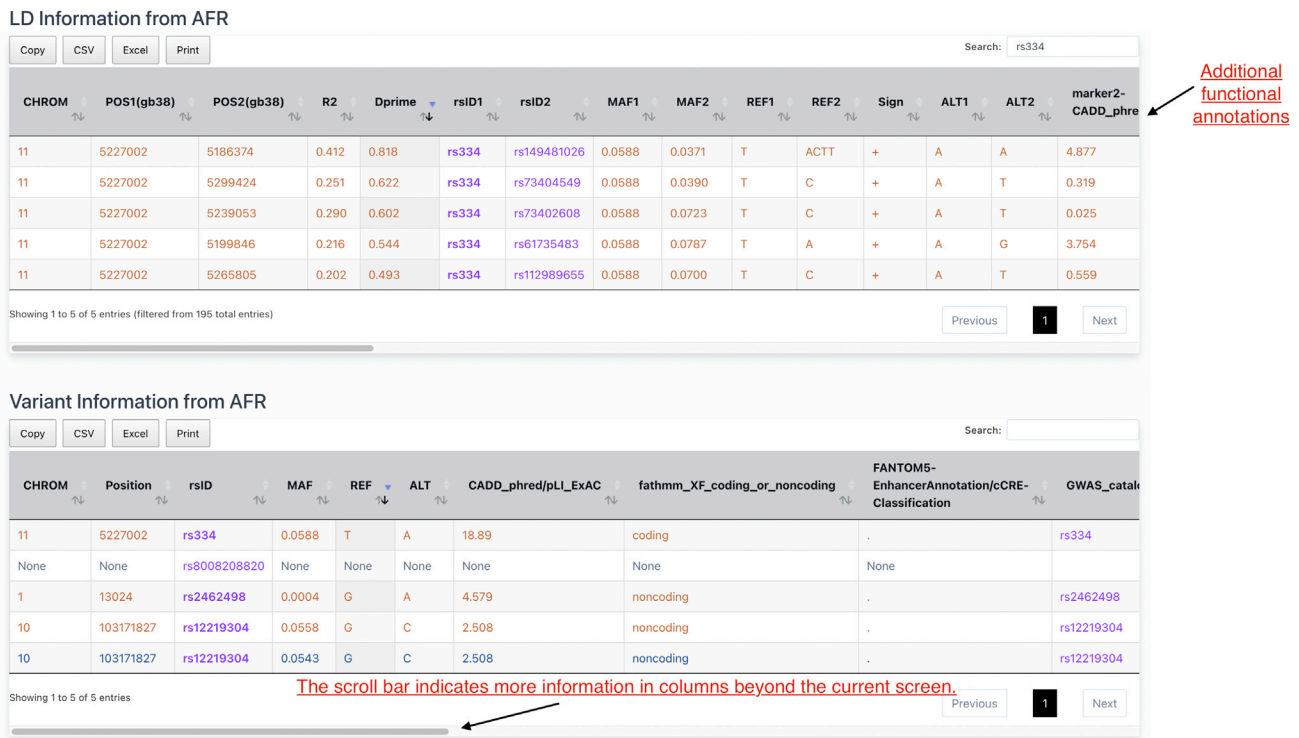
PHP, Javascript, and Apache2 under the CloudSQL and Compute Engine of Google Cloud Platform. The web interface provides access to all precomputed LD estimates. Users have the option to either paste or upload a file containing variant(s) of interest. Users can specify the population (East Asian, European, African, or South Asian) in which LD was estimated. In TOP-LD, markers are identified by rsID, or chr:position, or chr:position:REF:ALT for SNVs, or TOPMed variant names for SVs (in the format of DEL/DUP/INV\_chr:startPosition-endPosition, for example, DEL\_10:85001–97300). TOP-LD returns all variants within a pre-specified LD threshold (ranging from  $R^2$  values of 0.2 to 1.0) with the query variant. TOP-LD supports fast batch queries (Figure 2); querying a single variant takes  $\sim 0.5$  s, while a batch query of 500 variants takes  $\sim 2.3$  seconds. TOP-LD currently allows a maximum of 500 variants in one query.

After submitting the query, the website auto-directs to a result page that contains two parts: LD information on the top panel and variant information on the bottom panel. The latter provides basic information for the queried variants, including position, marker name, alleles (REF and ALT), and minor allele frequency (MAF). Markers not in the database will have “none” for all fields except marker names. The LD panel displays related LD metrics, one pair of variants on each line, including both  $R^2$ ,  $D'$ , and the sign of LD (measured between REF alleles of the two variants), along with marker name, marker position,



**Figure 2. Elapsed time (in seconds) for queries**

The x axis represents the number of variants queried, and the y axis represents the elapsed time.



**Figure 3. An example query result**

The result contains two parts. The top part “LD information from AFR” shows the LD information where each line provides information between a query variant (rsID1) and one of its corresponding LD proxies (rsID2). The bottom part “variant information from AFR” provides variant information, which shows basic information for each query variant. From the bottom part, we know that the user’s query includes four variants: rs334, rs8008208820, rs2462498, and rs12219304. Variants not included in LD calculation will have “none” records. For instance, rs8008208820 in this example query is not involved in LD inference and therefore will not have any LD proxies in the top part simply because of no data. Records from SV inference are in blue and those from SNV data are in orange. Some variants may appear twice because they are included in both SNV LD calculation and SV calculation. For example, in this example, rs12219304 appeared twice with MAF 0.0558 from the SNV source (second last record in orange) and MAF 0.0543 from the SV source (last record in blue).

alleles, and frequency for both variants in the pair (Figure 3). In addition, we provide the following pieces of information for SNVs from WGS annotation<sup>12</sup>: CADD score (phred-scaled), fathmm\_XF\_coding\_or\_noncoding classification, FANTOM5 enhancer annotations, gene name, and relative location to gene as well as a link to GWAS catalog query results.<sup>13</sup> For SVs, we provide a variety of annotations including gene(s) overlapping the SV, the SV’s location relative to gene, the gene’s pLI score, overlapping candidate *cis*-regulatory regions (cCREs) from ENCODE SCREEN.<sup>14,15</sup> The query results can be sorted, searched, copied, exported, and printed for further analyses.

The TOP-LD tool leverages TOPMed WGS data, whose much larger sample size and high depth sequencing lead to LD information for a much larger number of variants compared to the 1000 Genomes Project. As shown in Figure 1A and Table S1, TOP-LD offers 2.6- to 9.1-fold increase in variant coverage compared to the other state-of-the-art resources such as HaploReg 4.0 or LDlink. For example, for the European population, TOP-LD includes 146.5 million autosomal SNVs, while HaploReg 4.0 or LDlink contains 16.1 million variants. Not surprisingly,

the vast majority of the variants in TOP-LD that are not in 1000 Genomes, contributing to the up to  $9.1\times$  increase, are low frequency or rare. For example, out of the 146.5 million autosomal SNVs cataloged in the TOP-LD European population, 137.8 million have  $MAF < 0.01$  (Figure 1A, Table S1). Most of the variants have LD proxies. For example, 115.1 out of the 146.5 (78.6%) million autosomal variants have at least one LD tag with  $R^2 \geq 0.8$  and if we further relax the  $R^2$  threshold to 0.5 and 0.2, the number increases to 135.3 (92.4%) and 143.5 (98.0%), respectively (Figure 1B).

For chromosome X, we have included 6.5 million, 2.4 million, 1.3 million, and 760,000 variants for the European, African, East Asian, and South Asian populations, respectively (Table S1). Similar to the autosomal variants, the majority of these variants have at least one LD proxy with  $R^2 \geq 0.8$ : 5.1 million, European; 2.1 million, African; 1.2 million East Asian; 690,000, South Asian (Figure 1C, Table S2).

To evaluate the consistency between TOP-LD estimates and those from Haploreg v4.1, we collected the set of overlapping variants based on rsID with  $MAF \geq 0.05$  for Europeans and Africans. This set of variants was further filtered

**Table 2. Summary statistics of distinct working truth at GGT1 locus associated with gamma glutamyltransferase**

Signal	Variant	Position (hg38)	Effect allele	Unconditional p value	p value conditional on previous signals <sup>a</sup>	Effect allele frequency
1	rs4049904	24609759	G	2.82e-61	N/A	10.27%
2	rs73404962	24598530	G	4.46e-29	2.00e-36	5.63%
3	rs743369	24588099	A	9.94e-36	7.51e-27	11.94%
4	rs6004193	24598329	C	4.23e-41	3.25e-19	18.27%
5	rs57719575	24609020	C	3.97e-38	1.98e-24	14.86%
6	rs3876101	24607291	A	2.66e-15	1.17e-13	35.45%
7	rs116161010	24585912	T	5.69e-17	7.70e-9	7.13%

<sup>a</sup>The p values are reported from the sequential conditional analysis. For example, we report the p value for rs73404962 conditional on rs4049904, the p value of rs743369 conditional on both rs4049904 and rs73404962, and so forth.

such that the MAF values were within 10% of each other because large MAF differences would induce large LD differences. Figures S1 and S2 show high level of agreement between TOP-LD and Haploreg v4.1 LD estimates (e.g., Pearson correlation = 0.972 and 0.962 for European and African chromosome 1, respectively). Similarly, comparison of the chromosome X TOP-LD estimates for females and males again show high level of consistency (Pearson correlation = 0.992 and 0.975 for European and African population, respectively) (Figures S3 and S4).

To demonstrate the utility of TOP-LD, we performed fine-mapping at the GGT1 locus on chromosome 22, which is known to be associated with gamma glutamyltransferase.<sup>16</sup> We performed sequential conditional analysis with EPACTS<sup>17</sup> by using individual-level data among 8,768 UK Biobank participants of African ancestry following the same strategy in our previous work<sup>18</sup> adjusting for the same covariates as in Sun et al.<sup>19</sup> The sequential conditional analyses with individual-level data identified seven distinct signals at the GGT1 locus associated with gamma glutamyltransferase (Table 2). Because we used individual-level data for this conditional analysis, we considered these seven distinct signals to be the “working truth.”

We then carried out fine-mapping analysis with the FINEMAP method<sup>20</sup> by using only GWAS summary statistics from Sun et al.<sup>19</sup> We applied FINEMAP with an LD reference either from TOP-LD or from the 1000 Genomes Project and assessed the performance by comparing the results with “working truth” established from the sequential conditional analysis of the individual-level data.

FINEMAP produced 95% credible sets containing five variants when using either the 1000 Genomes (1000G) Project LD panel or the TOP-LD panel (see Table 3). However, the 1000G-based credible set contained only one of the seven signals from the “working truth” set. In contrast, the TOP-LD-based credible set contained three of the seven signals from the “working truth” set. In addition, because the lead variant from each conditional analysis (corresponding to each distinct signal) is selected somewhat arbitrarily, we also considered their LD proxies. When we considered any LD proxy (using a lenient R<sup>2</sup> threshold of 0.2) of a variant in the working truth set, the 1000G-based results still only identified a single signal from the working truth, whereas the TOP-LD-based results identified four of the seven signals (Table 3).

We also used TOP-LD to aid in the identification and prioritization of potentially causal structural variants at GWAS loci. For example, our recent association analysis<sup>21</sup> with TOPMed data identified an African-specific (MAF = 0.129) variant rs28450540 associated with lower monocyte count (p = 3.65 × 10<sup>-17</sup>). Query for LD tags via TOP-LD revealed a ~600 bp deletion near SIPR3 in perfect LD (R<sup>2</sup> = 1) with rs28450540 in the African population. We performed genome editing in monocytic and primary human HSPCs followed by xenotransplantation, which provides evidence that the deletion disrupts an SIPR3 monocyte enhancer leading to decreased SIPR3 expression. These preliminary data from functional experiments suggest that the 600 bp deletion is most likely causal but would have been missed in standard association analysis with

**Table 3. FINEMAP credible-set variants**

		Variant 1	Variant 2	Variant 3	Variant 4	Variant 5
1000G reference	credible-set variant	rs4049904	rs147866692	rs570263050	rs115231893	22:24649848:G:A (hg38)
	LD with working truth	1 (w/rs4049904 itself)	0.464 (w/rs4049904)	0.606 (w/rs4049904)	0.275 (w/rs4049904)	0.434 (w/rs4049904)
TOP-LD reference	credible-set variant	rs4049904	rs743369	rs57719575	rs2073397	rs5751902
	LD with working truth	1 (w/rs4049904 itself)	1 (w/rs743369 itself)	1 (w/rs57719575 itself)	0.83 (w/rs6004193)	0.51 (w/rs6004193)

The two five-variant credible sets provided by FINEMAP with either 1000G or TOP-LD as reference. For each credible-set variant, we list the corresponding variant (and the LD Rsq) from the working truth that has the highest LD.

only SNVs.<sup>22</sup> TOP-LD offers a simple and efficient approach to rescue such putative causal structural variants.

LD information, reflecting recombination, natural selection, and demographic history, has always been of intense interest in population genetics and complex trait association studies. LD information is also indispensable for a wide range of other applications, including GWAS follow-up and many summary-statistics-based inferences including fine-mapping, imputation of association summary statistics, construction of polygenic risk scores (PRSs), and interpretation and prioritization of GWAS results for further functional and clinical studies. TOP-LD significantly boosts the coverage of lower frequency variants by harnessing the power of high-coverage (~30×) WGS data of over 15,000 individuals primarily of a single continental ancestry. We demonstrate the utility of TOP-LD in fine-mapping at the *GGT1* locus and variant prioritization at the *S1PR3* locus. The LD information provided by TOP-LD will facilitate a range of essential inferences for common and rare variation across a diverse range of populations.

#### Data and code availability

Data generated for this study can be accessed via the TOP-LD web portal: <http://topld.genetics.unc.edu>.

#### Supplemental information

Supplemental information can be found online at <https://doi.org/10.1016/j.ajhg.2022.04.006>.

#### Acknowledgments

We gratefully acknowledge the studies and participants who provided biological samples and data for TOPMed. The TOPMed Banner Authorship list can be found at: <https://www.nhlbiwgs.org/topmed-banner-authorship>. The project described is supported by funding from the National Institutes of Health through R01HL129132 and U01HG011720 (Y.L.), R01HL146500 (A.P.R.), and KL2TR002490 (L.M.R.). Y.L., J.D.R., and L.H. are also partially supported by R01HL146500, U24 AR076730, and U01DA052713. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

#### Declaration of interests

L.M.R. is a consultant for the TOPMed Administrative Coordinating Center (through Westat).

Received: December 20, 2021

Accepted: April 8, 2022

Published: May 2, 2022

#### Web resources

EPACTS, <https://genome.sph.umich.edu/wiki/EPACTS>  
FINEMAP, <http://www.christianbenner.com/>

HaploReg, <https://pubs.broadinstitute.org/mammals/haploreg/>

LDlink, <https://ldlink.nci.nih.gov/>

TOP-LD, <http://topld.genetics.unc.edu/>

TOPMed, <https://topmed.nhlbi.nih.gov/>

#### References

1. Slatkin, M. (2008). Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. *Nat. Rev. Genet.* 9, 477–485. <https://doi.org/10.1038/nrg2361>.
2. Bush, W.S., and Moore, J.H. (2012). Chapter 11: Genome-wide association studies. *PLoS Comput. Biol.* 8, e1002822. <https://doi.org/10.1371/journal.pcbi.1002822>.
3. Choudhury, A., Aron, S., Botigué, L.R., Sengupta, D., Botha, G., Bensekell, T., Wells, G., Kumuthini, J., Shriner, D., Fakim, Y.J., et al. (2020). High-depth African genomes inform human migration and health. *Nature* 586, 741–748. <https://doi.org/10.1038/s41586-020-2859-7>.
4. The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. *Nature* 526, 68–74.
5. Consortium, I.H., Gibbs, R.A., Peltonen, L., Altshuler, D.M., Gibbs, R.A., Peltonen, L., Dermitzakis, E., Schaffner, S.F., Yu, F., Peltonen, L., et al. (2010). Integrating common and rare genetic variation in diverse human populations. *Nature* 467, 52–58. <https://doi.org/10.1038/nature09298>.
6. Taliun, D., Harris, D.N., Kessler, M.D., Carlson, J., Szpiech, Z.A., Torres, R., Taliun, S.A.G., Corvelo, A., Gogarten, S.M., Kang, H.M., et al. (2021). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* 590, 290–299. <https://doi.org/10.1038/s41586-021-03205-y>.
7. Ward, L.D., and Kellis, M. (2016). HaploReg v4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. *Nucleic Acids Res.* 44, D877–D881. <https://doi.org/10.1093/nar/gkv1340>.
8. Machiela, M.J., and Chanock, S.J. (2015). LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics (Oxford, England)* 31, 3555–3557. <https://doi.org/10.1093/bioinformatics/btv402>.
9. Maples, B.K., Gravel, S., Kenny, E.E., and Bustamante, C.D. (2013). RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am. J. Hum. Genet.* 93, 278–288. <https://doi.org/10.1016/j.ajhg.2013.06.020>.
10. Conomos, M.P., Reiner, A.P., Weir, B.S., and Thornton, T.A. (2016). Model-free estimation of recent genetic relatedness. *Am. J. Hum. Genet.* 98, 127–148. <https://doi.org/10.1016/j.ajhg.2015.11.022>.
11. Loh, P.R., Danecek, P., Palamara, P.F., Fuchsberger, C., A Reshef, Y., K Finucane, H., Schoenherr, S., Forer, L., McCarthy, S., Abecasis, G.R., et al. (2016). Reference-based phasing using the haplotype reference consortium panel. *Nat. Genet.* 48, 1443–1448. <https://doi.org/10.1038/ng.3679>.
12. Liu, X., White, S., Peng, B., Johnson, A.D., Brody, J.A., Li, A.H., Huang, Z., Carroll, A., Wei, P., Gibbs, R., et al. (2016). WGS: an annotation pipeline for human genome sequencing studies. *J. Med. Genet.* 53, 111–112. <https://doi.org/10.1136/jmedgenet-2015-103423>.
13. MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., et al.



- (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* 45, D896–d901. <https://doi.org/10.1093/nar/gkw1133>.
14. Geoffroy, V., Herenger, Y., Kress, A., Stoetzel, C., Piton, A., Dollfus, H., and Muller, J. (2018). AnnotSV: an integrated tool for structural variations annotation. *Bioinformatics (Oxford, England)* 34, 3572–3574. <https://doi.org/10.1093/bioinformatics/bty304>.
  15. Moore, J.E., Moore, J.E., Purcaro, M.J., Pratt, H.E., Epstein, C.B., Shores, N., Adrian, J., Kawli, T., Davis, C.A., Dobin, A., et al. (2020). Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* 583, 699–710. <https://doi.org/10.1038/s41586-020-2493-4>.
  16. Pazoki, R., Vujkovic, M., Elliott, J., Evangelou, E., Gill, D., Ghanbari, M., van der Most, P.J., Pinto, R.C., Wielscher, M., Farlik, M., et al. (2021). Genetic analysis in European ancestry individuals identifies 517 loci associated with liver enzymes. *Nat. Commun.* 12, 2579. <https://doi.org/10.1038/s41467-021-22338-2>.
  17. Kang, H.M., Sul, J.H., Service, S.K., Zaitlen, N.A., Kong, S.Y., Freimer, N.B., Sabatti, C., and Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* 42, 348–354. <https://doi.org/10.1038/ng.548>.
  18. Raffield, L.M., Iyengar, A.K., Wang, B., Gaynor, S.M., Spracklen, C.N., Zhong, X., Kowalski, M.H., Salimi, S., Polfus, L.M., Benjamin, E.J., et al. (2020). Allelic heterogeneity at the CRP locus identified by whole-genome sequencing in multi-ancestry cohorts. *Am. J. Hum. Genet.* 106, 112–120. <https://doi.org/10.1016/j.ajhg.2019.12.002>.
  19. Sun, Q., Graff, M., Rowland, B., Wen, J., Huang, L., Miller-Fleming, T.W., Haessler, J., Preuss, M.H., Chai, J.F., Lee, M.P., et al. (2021). Analyses of biomarker traits in diverse UK biobank participants identify associations missed by European-centric analysis strategies. *J. Hum. Genet.* 67, 87–93. <https://doi.org/10.1038/s10038-021-00968-0>.
  20. Benner, C., Spencer, C.C.A., Havulinna, A.S., Salomaa, V., Ripatti, S., and Pirinen, M. (2016). FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics (Oxford, England)* 32, 1493–1501. <https://doi.org/10.1093/bioinformatics/btw018>.
  21. Mikhaylova, A.V., McHugh, C.P., Polfus, L.M., Raffield, L.M., Boorgula, M.P., Blackwell, T.W., Brody, J.A., Broome, J., Chami, N., Chen, M.H., et al. (2021). Whole-genome sequencing in diverse subjects identifies genetic correlates of leukocyte traits: the NHLBI TOPMed program. *Am. J. Hum. Genet.* 108, 1836–1851. <https://doi.org/10.1016/j.ajhg.2021.08.007>.
  22. Wheeler, M.M., Stilp, A.M., Rao, S., Halldórsson, B.V., Beyter, D., Wen, J., Mikhaylova, A.V., McHugh, C.P., Lane, J., Jiang, M.-Z., et al. (2021). Whole Genome sequencing identifies common and rare structural variants contributing to hematologic traits in the NHLBI TOPMed program. Preprint at medRxiv. <https://doi.org/10.1101/2021.12.16.21267871>.