Bioinformatics, 35(1), 2019, 164–166
doi: 10.1093/bioinformatics/bty547
Advance Access Publication Date: 2 July 2018
Applications Note

OXFORD

Genetics and population analysis

emeraLD: rapid linkage disequilibrium estimation with massive datasets

Corbin Quick[1,*], Christian Fuchsberger[1,2,3], Daniel Taliun[1], Gonçalo Abecasis[1], Michael Boehnke[1] and Hyun Min Kang[1]

[1]Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, MI 48109, USA, [2]Institute for Biomedicine, Eurac Research, Affiliated Institute of the University of Lübeck, Bolzano, Italy and [3]Division of Genetic Epidemiology, Department of Medical Genetics, Molecular and Clinical Pharmacology, Medical University of Innsbruck, Innsbruck, Austria

*To whom correspondence should be addressed.

Associate Editor: Ioannis Xenarios

Received on April 16, 2018; revised on April 16, 2018; editorial decision on June 27, 2018; accepted on June 28, 2018

Abstract

Summary: Estimating linkage disequilibrium (LD) is essential for a wide range of summary statistics-based association methods for genome-wide association studies. Large genetic datasets, e.g. the TOPMed WGS project and UK Biobank, enable more accurate and comprehensive LD estimates, but increase the computational burden of LD estimation. Here, we describe emeraLD (Efficient Methods for Estimation and Random Access of LD), a computational tool that leverages sparsity and haplotype structure to estimate LD up to 2 orders of magnitude faster than current tools.

Availability and implementation: emeraLD is implemented in C++, and is open source under GPLv3. Source code and documentation are freely available at http://github.com/statgen/emeraLD.

Contact: corbinq@umich.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Linkage disequilibrium (LD) is of fundamental interest in population genetics as a vestige of natural selection and demographic history, and is essential for a wide range of analyses from summary statistics in genome-wide association studies (GWAS). Motivated by data-sharing and logistical constraints, a variety of tools have been developed for analysis of GWAS summary statistics rather than individual-level data—for example, fine-mapping (Benner *et al.*, 2017) and conditional analysis (Yang *et al.*, 2012). These methods often rely on LD estimates from an external dataset, which are ideally calculated on-the-fly rather than pre-computed and stored due to prohibitive storage costs. For example, the 1000 Genomes Project Phase 3 panel includes over 35 M shared variants (1000 Genomes Project Consortium, 2015), corresponding to $> 4 \times 10^{11}$ pairwise LD coefficients within 1 Mbp windows genome-wide. Analyzing GWAS summary statistics from increasingly large and diverse studies will require estimating LD with correspondingly large and diverse cohorts, prompting a need for efficient and scalable methods to estimate LD with massive datasets.

2 Materials and methods

2.1 LD statistics

Three common measures of LD are the LD coefficient D (the covariance of genotypes), the standardized LD coefficient D' (D divided by its maximum value given allele frequencies) and the Pearson correlation r or its square. Each of these statistics can be written as a function of allele frequencies, sample size and the dot product of genotype vectors. The dot product must be calculated for each pair of variants, whereas allele frequencies can be can be computed once for each variant and stored.

2.2 Computational approach

2.2.1 Sparse representation of phased genotypes

Given phased genotypes, we keep a $\{0, 1\}^{2n}$ vector of genotypes (where 1 indicates the minor allele) and sparse vector containing the indexes of non-zero entries for each variant. If the minor allele is non-reference in the input file, we reverse the sign of its LD statistics for consistency. The dot product $m_{jk} = \mathrm{G}_j \cdot \mathrm{G}_k$ between variants j

Table 1. Benchmarking: CPU time and memory usage

Tool:	m3vcftools	PLINK 1.9	LDstore	emeraLD[a]	Absolute[a]
Format:	M3VCF.gz	BED	BGEN	M3VCF.gz	
CPU Time relative to emeraLD					
1 KGP	18.8	1.3	4.4	1.0	8.5 m
HRC	44.7	6.8	16.8	1.0	2.6 m
UKB	473.7	128.4	250.6	1.0	19.9 m
Memory usage relative to emeraLD					
1 KGP	0.7	137.6	372.4	1.0	43.8 MiB
HRC	0.6	10.7	26.1	1.0	156.9 MiB
UKB	0.4	4.7	4.8	1.0	4.8 GiB

Note: CPU time and memory to calculate LD in a 1 Mbp region of chr20 (28 126 variants in 1 KGP; 13 174 in HRC and 32 783 in UKB). All experiments were run on a 2.8 GHz Intel Xeon CPU.

[a] Absolute time or memory for emeraLD as reference.

and k can be calculated in min (m_j, m_k) operations, where m_j is the minor allele count at variant j, using the sparse-by-dense product formula $m_{jk} = \sum_{i\in C_j} G_{ik}$, where $C_j = \{i|G_{ij} = 1\}$ is the set of minor-allele carriers for variant j.

2.2.2 Sparse representation of unphased genotypes

For unphased genotypes, we store a $\{0, 1, 2\}^n$ genotype vector and sparse vectors indexing heterozygotes and minor-allele homozygotes for each variant. In this case, the dot product can be calculated in $\min(N_{j1} + N_{j2}, N_{k1} + N_{k2})$ operations, where N_{ji} is the count of genotype i at variant j.

2.2.3 Haplotype block representation

Due to the limited diversity of human haplotypes (Wall and Pritchard, 2003), the number of distinct haplotypes in a haplotype block with J biallelic variants is typically small relative to the sample size $2n$ or to 2^J. M3VCF format (Das *et al*., 2016) stores genotypes using a compact haplotype representation, which requires far less storage than VCF (Danecek *et al*., 2011). Given M3VCF input, we pre-compute the counts N^b_h of each haplotype h in each block b, and index the haplotypes H^b_j containing the minor allele at variant j in block b. For variants in the same block, the dot product can be calculated in $\min(c^b_j, c^b_k)$ operations, where $c^b_k = \#H^b_k$ is the number of haplotypes that carry the minor allele at variant k, using the formula $m_{jk} = \sum_{h\in H^b_j} 1_{H^b_k}(h)N^b_h$. For variants in different blocks, we use sparse genotype representation to efficiently estimate LD.

2.2.4 Approximation by sub-sampling for large sample sizes

When both variants j and k have large MAC (e.g. common variants and/or large sample sizes), calculating sparse-by-dense products becomes expensive. In this case, we use an informed sub-sampling approach to efficiently estimate LD while maintaining a user-specified bound on the precision of LD estimates. In Supplementary Materials, we derive an optimal approximate estimator $\tilde{r}_\ell$, which can be calculated in at most ℓ operations for any pair of variants while increasing the MSE by no more than $1/\ell$ relative to exact LD estimates (or $2/\ell$ for unphased genotypes), where ℓ is user-specified. In very large datasets ($n > 50$ K), this approach decreased computation time for common variants (MAF $> 5\%$) by an order of magnitude or more.

3 Results

3.1 Implementation and usage

We implemented our methods in an open source C++ tool, emeraLD (efficient methods for estimation and random access of LD), which accepts VCF.gz and M3VCF.gz formats and leverages Tabix (Li, 2011) and HTSlib for rapid querying and random access of genomic regions. emeraLD includes options to facilitate a variety of common analyses involving LD, and can also be used via an R interface included with source files.

3.2 Performance

We used WGS genotype data from the 1000 Genomes Project Phase 3 (1 KGP; $n = 2504$), Haplotype Reference Consortium (HRC; $n = 32\,470$; Haplotype Reference Consortium, 2016), and imputed genotype data from the UK Biobank (UKBB; $n = 487\,409$) to compare performance between emeraLD and PLINK v1.9 (Chang *et al*., 2015; Purcell and Chang, 2016), LDstore (Benner *et al*., 2017) and m3vcftools (Das *et al*., 2016). In large datasets, emeraLD is up to two orders of magnitude faster than existing tools (Table 1). Times reported for emeraLD used $\ell = 1000$ (MSE of approximation ≤ 0.001); for for UKB and HRC, this reduced overall computation time by ~50%. Using M3VCF.gz files reduced computation time for emeraLD by ~30–50% relative to VCF.gz.

3.3 Applications

Our approach will be implemented in a forthcoming web-based service capable of providing LD from panels with >60 K samples in real time. This enables use of improved LD information in web-based interactive analysis and visualization tools such as LocusZoom (Pruim *et al*., 2010).

We have also used emeraLD to estimate LD on-the-fly for gene-based association and functional enrichment analysis of GWAS summary statistics. This approach avoids pre-computing and storing LD without compromising speed.

4 Conclusions

Here, we described methods to efficiently estimate LD with large datasets by leveraging the natural sparsity and redundancy of genetic data. We also developed an approximation approach to improve computational efficiency while maintaining a user-specified bound on statistical precision. Finally, we described an open-source software implementation of our methods.

Acknowledgements

The authors acknowledge the developers of M3VCF and VCF formats, and the cohorts and staffs of the HRC and 1KGP. This research has been conducted using the UKBB Resource under Application Number 24460.

Funding

The authors acknowledge support from NIH grants HG000376 (M.B.), HG009976 (MB), HG007022 (G.R.A.), HG006513 (G.R.A.) and U01HL137182 (H.M.K.).

Conflict of Interest: none declared.

References

1000 Genomes Project Consortium (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.

Benner,C. *et al*. (2017) Prospects of fine-mapping trait-associated genomic regions by using summary statistics from genome-wide association studies. *Am. J. Hum. Genet*., **101**, 539–551.

Chang,C.C. *et al*. (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, **4**.

Danecek,P. *et al*. (2011) The variant call format and vcftools. *Bioinformatics*, **27**, 2156–2158.

Das,S. *et al*. (2016) Next-generation genotype imputation service and methods. *Nat. Genet*., **48**, 1284–1287.

Haplotype Reference Consortium (2016) A reference panel of 64, 976 haplotypes for genotype imputation. *Nat. Genet*., **48**, 1279–1283.

Li,H. (2011) Tabix: fast retrieval of sequence features from generic tab-delimited files. *Bioinformatics*, **27**, 718–719.

Pruim,R.J. *et al*. (2010) Locuszoom: regional visualization of genome-wide association scan results. *Bioinformatics*, **26**, 2336–2337.

Purcell,S. and Chang,C. (2016) Plink 1.9 package. www.cog-genomics.org/plink/1.9/.

Wall,J.D. and Pritchard,J.K. (2003) Haplotype blocks and linkage disequilibrium in the human genome. *Nat. Rev. Genet*., **4**, 587–597.

Yang,J. *et al*. (2012) Conditional and joint multiple-snp analysis of gwas summary statistics identifies additional variants influencing complex traits. *Nat. Genet*., **44**, 369–375.