DOI 10.1038/s41439-018-0030-x

Human Genome Variation

SOFTWARE REPORT Open Access

eLD: entropy-based linkage disequilibrium index between multiallelic sites

Yukinori Okada[1,2]

Abstract

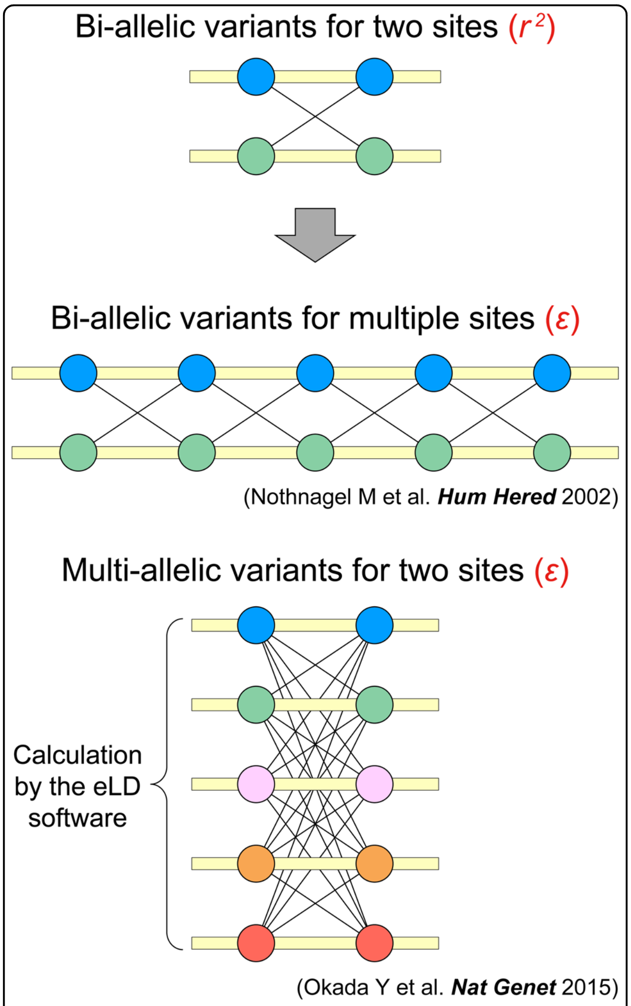
Quantification of linkage disequilibrium (LD) is a critical step in studies investigating human genome variations. Commonly used LD indices such as r^2 handle LD of biallelic variants for two sites. As shown in a previously introduced LD index of ε, normalized entropy difference of the haplotype frequency between LD and linkage equilibrium (LE) could be utilized to estimate LD of biallelic variants for multiple sites. Here, we developed eLD (**e**ntropy-based **L**inkage **D**isequilibrium index between multiallelic sites) as publicly available software to calculate ε of multiallelic variants for two sites. Application of eLD could dissect complex LD structures among multiple HLA genes (e.g., strong LD among *HLA-DRB1*, *HLA-DQA1*, and *HLA-DQB1* in East Asians). Use of eLD is not restricted to haplotype-based LD; it is also applicable to genotype-based LD. Therefore, eLD enables estimation of *trans*-regional LD of SNP genotypes at two unlinked loci, such as the nonlinear LD between functional missense variants of *ADH1B* (rs1229984 [Arg47His]) and *ALDH2* (rs671 [Glu504Lys]).

Linkage disequilibrium (LD) is defined as the nonrandom distribution of alleles at different loci[1]. Quantitative assessment of LD in a population of interest is an important procedure to conduct fine-mapping of causal variants embedded in the disease risk loci identified by genome-wide association studies (GWAS)[2]. Population-specific features of LD are related to ethnically heterogeneous distributions of single-nucleotide polymorphisms (SNPs)[3]. The most widely used measurements of LD are r^2 and D'; both values quantify LD between biallelic variants (i.e., SNPs) for two sites, reflecting nonrandom distributions of four haplotypes consisting of pairwise combinations of the alleles. Specifically, r^2 can be interpreted as Pearson's correlation measurement (R^2) of allele distributions and is known to be proportional to χ^2 values of genotype–phenotype association statistics between two sites[1]. LD values can easily be calculated using publicly available software (e.g., PLINK and vcftools), or using downloaded pre-calculated values from websites (e.g., HaploReg and LocusZoom).

Nothnagel et al.[4] previously demonstrated that r^2 can also be interpreted as normalized entropy in haplotype frequencies, and introduced a novel LD index named ε (see definition in Supplementary Information). ε represents the normalized entropy difference of the haplotype frequencies between LD and those expected under the null hypothesis of no LD (i.e., linkage equilibrium [LE]). The value of ε ranges between 0 and 1, with larger values indicating stronger LD. Application of ε enabled LD quantification of biallelic variants for multiple sites (Fig. 1)[4], which was effective in selecting tag SNPs free from ambiguous definitions of LD blocks in an unbiased manner[5].

We have recently extended ε to further quantify LD of multiallelic variants for two sites as described elsewhere (Fig. 1)[6]. Here, we developed eLD (**e**ntropy-based **L**inkage **D**isequilibrium index between multiallelic sites) as publicly available software to calculate the ε of multiallelic variants for two sites (see the software URL). Various multiallelic variants exist with important clinical impacts in terms of genotype–phenotype associations. Of these, polymorphisms of human leukocyte antigen (HLA) genes

Correspondence: Yukinori Okada (yokada@sg.med.osaka-u.ac.jp)
[1]Department of Statistical Genetics, Osaka University Graduate School of Medicine, Suita 565-0871, Japan
[2]Laboratory of Statistical Immunology, Immunology Frontier Research Center (WPI-IFReC), Osaka University, Suita 565-0871, Japan


Bi-allelic variants for two sites (r^2)
Bi-allelic variants for multiple sites (ε)
Multi-allelic variants for two sites (ε)
Calculation by the eLD software

(Nothnagel M et al. ***Hum Hered*** 2002)

(Okada Y et al. ***Nat Genet*** 2015)

Fig. 1 eLD: entropy-based linkage disequilibrium index between multiallelic sites. While r^2 quantifies LD of biallelic variants for two sites, ε quantifies LD of biallelic variants for multiple sites[4] or multiallelic variants for two sites[6]. We developed eLD (**e**ntropy-based **L**inkage **D**isequilibrium index between multiallelic sites) as publicly available software to calculate ε of multiallelic variants for two sites (see the software URL)

in the major histocompatibility (MHC) locus have a wide spectrum of risk for a variety of human diseases. While elucidation of the complex LD structure of HLA genes has been challenging, application of ε clearly identified hidden LD relationships among the HLA genes[6]. For example, we observed relatively strong LD between *HLA-C* and *HLA-B*, among *HLA-DRB1*, *HLA-DQA1*, and *HLA-DQB1*, and between *HLA-DPA1* and *HLA-DPB1* ($\varepsilon > 0.15$; calculated using 4-digit classical alleles of a subset of the East Asian subjects [$n = 300$] enrolled in the original studies[6,7]; Fig. 2). Since estimation of the haplotype frequency could be biased when its distribution is sparse, an option to combine the alleles with frequencies lower than the defined threshold (0.05 in default settings) into a single dummy allele is implemented in eLD.

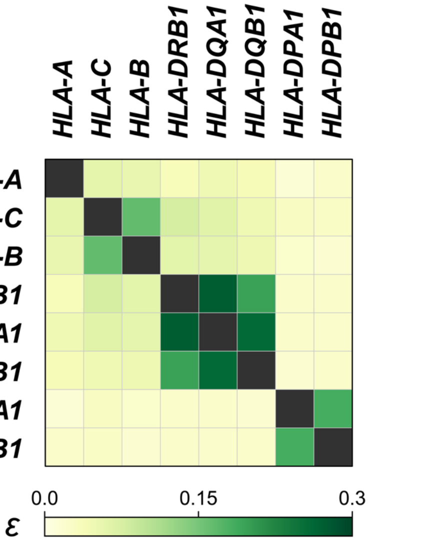


Fig. 2 Pairwise quantification of LD among the classical HLA gene variants. Pairwise LD index of ε among the 4-digit alleles of the classical HLA genes were evaluated using eLD. Phased HLA alleles were obtained from a subset of the East Asian subjects ($n = 300$) of the original studies[6, 7]. Strong LD between *HLA-C* and *HLA-B*, among *HLA-DRB1*, *HLA-DQA1*, and *HLA-DQB1*, and between *HLA-DPA1* and *HLA-DPB1* was observed ($\varepsilon > 0.15$)

One of the novel features of eLD is to empirically estimate a value of ε in a null hypothesis of LE ($=\varepsilon_{\_\mathrm{NULL}}$). Additionally, it also calculates the ε actually observed in a given data set ($=\varepsilon_{\_\mathrm{Observed}}$). eLD calculates $\varepsilon_{\_\mathrm{NULL}}$ based on a permutation approach. By randomly shuffling connections of the alleles between the two sites, $\varepsilon_{\_\mathrm{NULL}}$ is estimated as the mean value of ε obtained in each iteration step (×1000 iterations in default settings). Since the baseline value of $\varepsilon_{\_\mathrm{NULL}}$ depends on the number of alleles in each site, calculation of $\varepsilon_{\_\mathrm{NULL}}$ as well as $\varepsilon_{\_\mathrm{Observed}}$ would help to evaluate the relative strength of LD relationships at the observed sites.

Another feature of the software is that application of eLD is not restricted to haplotype-based LD; it is also applicable to genotype-based LD. Using eLD, one can estimate LD between loci where phasing of the haplotypes is theoretically difficult. As an illustrative example, we estimated *trans*-regional LD in two unlinked loci: *ADH1B* at 4q23 and *ALDH2* at 12q24. *ADH1B* and *ALDH2* harbor well known functional missense variants at rs1229984 (Arg47His) and rs671 (Glu504Lys), respectively. Both of these SNPs have pleiotropic effects on a number of human complex traits, including dietary habits. Studies investigating natural selection pressure identified strong significant positive selection on these missense variants in Japanese or other East Asian populations, which was closely linked to geographical heterogeneity in allele frequency spectra of these SNPs even within a single population[8]. Here, using eLD, we calculated ε to estimate *trans*-regional

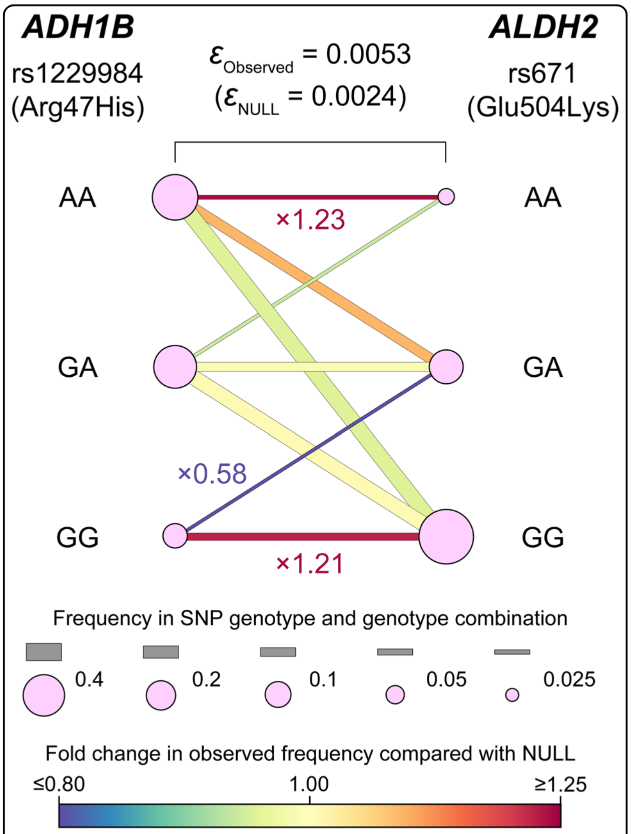


Fig. 3 ***Trans*-regional LD between functional missense variants of *ADH1B* and *ALDH2*.** eLD can quantify genotype-based LD as well as haplotype-based LD and thus can estimate *trans*-regional LD between two unlinked loci without haplotype phasing. LD between genotypes of the two functional missense variants of *ADH1B* (rs1229984 [Arg47His]) and *ALDH2* (rs671 [Glu504Lys]) was assessed by using eLD. Observed ε (= $\varepsilon_{\_\mathrm{Observed}}$) and ε expected under the null hypothesis of LE (= $\varepsilon_{\_\mathrm{NULL}}$) are indicated. Frequencies of the genotypes and genotype combinations are visualized as in the legend. We observed nonrandom and nonlinear distribution of the genotypes between the variants (i.e., reduced frequency in rs1229984GG-rs671GA)

LD between rs1229984 and rs671 (Fig. 3). We obtained genotypes for these SNPs from East Asian subjects within the 1000 Genomes Projects ($n = 504$, phase 3 version 5), and found a high $\varepsilon_{\_\mathrm{Observed}}$ value (=0.0053) when compared to $\varepsilon_{\_\mathrm{NULL}}$ (= 0.0024). As expected from natural selection pressure on these variants[8], rs1229984AA-rs671AA genotypes and rs1229984GG-rs671GG genotypes had increased frequencies compared to those variants in LE (≥1.21-fold), while rs1229984GG-rs671GA genotypes had decreased frequencies (0.58-fold) compared to those variants in LE. While Pearson's correlation between genotypes can also evaluate *trans*-regional LD, nonlinear relationships of genotypes (such as the reduced frequency of rs1229984GG-rs671GA) would not have been reflected with this measurement.

In summary, we developed software, which we named eLD, that quantifies the entropy-based LD index of ε in multiallelic variants for two sites, such as LD between highly polymorphic HLA genes. eLD also enables estimation of *trans*-regional LD of SNP genotypes, such as functional variants of *ADH1B* and *ALDH2*. We note that normalized entropy has increased the potential to dissect complex dependencies among human genome variations (e.g., Y-chromosomal short tandem repeat [STR] marker selection[9]), and development of additional methodology should be warranted.

Software availability

eLD is freely available at http://www.sg.med.osaka-u.ac.jp/tools.html with example data sets.

Acknowledgements

We thank Prof. Michael Nothnagel for his thoughtful suggestions. This study was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI (15H05911) and the Japan Agency for Medical Research and Development (AMED; 18gm6010001h0003 and 18ek0410041h0002).

Conflict of interest

The author declares that he has no conflict of interest.

Supplementary information is available for this paper at https://doi.org/10.1038/s41439-018-0030-x.

Received: 9 September 2018 Revised: 18 September 2018 Accepted: 18 September 2018.
Published online: 22 October 2018

References

1. Slatkin, M. Linkage disequilibrium - understanding the evolutionary past and mapping the medical future. *Nat. Rev. Genet.* **9**, 477–485 (2008).
2. Schaid, D. J. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat. Rev. Genet.* **19**, 91–504 (2018).
3. Kanai, M. et al. Empirical estimation of genome-wide significance thresholds based on the 1000 Genomes Project data set. *J. Hum. Genet.* **61**, 861–866 (2016).
4. Nothnagel, M. et al. Entropy as a measure for linkage disequilibrium over multilocus haplotype blocks. *Hum. Hered.* **54**, 186–198 (2005).
5. Nothnagel, M. & Rohde, K. The effect of single-nucleotide polymorphism marker selection on patterns of haplotype blocks and haplotype frequency estimates. *Am. J. Hum. Genet.* **77**, 988–998 (2005).
6. Okada, Y. et al. Construction of a population-specific HLA imputation reference panel and its application to Graves' disease risk in Japanese. *Nat. Genet.* **47**, 798–802 (2015).
7. Okada, Y. et al. Risk for ACPA-positive rheumatoid arthritis is driven by shared HLA amino acid polymorphisms in Asian and European populations. *Hum. Mol. Genet.* **23**, 6916–6926 (2014).
8. Okada, Y. et al. Deep whole-genome sequencing reveals recent selection signatures linked to evolution and disease risk of Japanese. *Nat. Commun.* **9**, 1631 (2018).
9. Siegert, S. et al. Shannon's equivocation for forensic Y-STR marker selection. *Forensic Sci. Int. Genet.* **16**, 216–225 (2015).