---

**Article:**
Fox, E.A., Wright, A.E., Fumagalli, M. et al. (1 more author) (2019) ngsLD: evaluating linkage disequilibrium using genotype likelihoods. Bioinformatics. ISSN 1367-4803

https://doi.org/10.1093/bioinformatics/btz200

---

*Bioinformatics*
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Applications Note



Genetics and population analysis

# *ngsLD*: evaluating linkage disequilibrium using genotype likelihoods

**Emma A. Fox [1], Alison E. Wright [2], Matteo Fumagalli [1], Filipe G. Vieira [3,*]**

[1]Department of Life Sciences, Silwood Park Campus, Imperial College London, Ascot, United Kingdom
[2]Department of Animal and Plant Sciences, University of Sheffield, United Kingdom
[3]Center for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Denmark

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

## Abstract

**Motivation:** Linkage disequilibrium measures the correlation between genetic loci and is highly informative for association mapping and population genetics. As many studies rely on called genotypes for estimating linkage disequilibrium, their results can be affected by data uncertainty, especially when employing a low read depth sequencing strategy. Furthermore, there is a manifest lack of tools for the analysis of large-scale, low-depth and short-read sequencing data from non-model organisms with limited sample sizes.
**Results:** *ngsLD* addresses these issues by estimating linkage disequilibrium directly from genotype likelihoods in a fast, reliable and user-friendly implementation. This method makes use of the full information available from sequencing data and provides accurate estimates of linkage disequilibrium patterns compared to approaches based on genotype calling. We conducted a case study to investigate how linkage disequilibrium decays over physical distance in two avian species.
**Availability:** The methods presented in this work were implemented in C/C++ and are freely available for non-commercial use from https://github.com/fgvieira/ngsLD
**Contact:** fgvieira@snm.ku.dk
**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

## 1 Introduction

Linkage disequilibrium (LD) measures the non random association of alleles at different loci. There are several definitions of LD, each one differing in the information they rely on to measure the statistical association between loci (Slatkin, 2008). One of the most widely used is $D$, which quantifies the difference between the frequency of haplotypes carrying two pairs of alleles at different loci and the product of the individual allele frequencies. A standardized version of $D$ called $D'$ is also often used, since it takes into account the full range of possible $D$ values. Another common definition of LD is $r^2$, which is widely interpreted as the squared correlation coefficient for the occurrence of pairs of alleles at different loci. Both $D'$ and $r^2$ are defined between 0 and 1.

Information on LD patterns across the genome is useful in both medical and evolutionary genetics. In the latter case, LD information is used to infer past historical events and demographic history of the species, or population, under investigation. Indeed, several factors affect the extent of LD, including mating system, recombination rate, mutation rate, genetic drift, and population structure. As a consequence, patterns of LD in a genome have been used to infer several population genetic parameters (Tenesa *et al.*, 2007).

Accuracy in the calculation of LD is therefore vital to make sensible inferences about the population of interest. As more traditional methods to measure LD rely on resolving individual haplotypes from genotype data, they are not applicable on low depth sequencing data, where only few reads cover each position on average. Furthermore, in the study of non-model organisms, the lack of reference information and large sample sizes prevent the use of imputation of missing data and haplotype phasing. Recent studies attempted to integrate data uncertainty into the estimation of LD and obtained promising results for high frequency alleles and moderate sample sizes (Maruki and Lynch, 2014; Bilton *et al.*, 2018).

Here we present `ngsLD`, a comprehensive package designed to calculate several measures of LD directly from genotype likelihoods. Using simulations, we show that this method is particularly suitable for low depth sequencing. Finally, we apply `ngsLD` on a mRNA sequencing

 

dataset from two avian species and confirm a significant difference in their effective population sizes.

## 2 Methods

We implemented two algorithms to estimate LD levels from genotype likelihoods. The main one is a maximum likelihood approach to estimate haplotype frequencies between pairs of sites using an expectation maximization (EM) algorithm (Excoffier and Slatkin, 1995) and, from these, calculate $D$, $D'$ and $r^2$. The EM functions were adapted from `bcftools v0.1.18` (Li, 2011). A second approach is based on the squared Pearson correlation ($r^2$) between expected genotypes, calculated from the genotype posterior probabilities.

`ngsLD` accepts both genotype likelihoods and genotype data as input, and outputs the pairwise LD between all pairs of valid SNPS, with running time compatible with genome-wide datasets 1. It supports several filtering options, such as distance between SNPs, minor allele frequency, and random subset pairs of SNPs. Apart from `ngsLD`, we also provide several auxiliary scripts to perform some of the most common LD-related analyses, such as LD decay curve fitting, LD blocks plotting and SNP pruning.

## 3 Results

At high sequencing depth ($>= \ 20\times$) we observed no significant differences using either called genotypes (CG) or genotype likelihoods (GL). At lower depths, results show a rise in both Root Mean Square Deviation (RMSD) and Mean Standard Bias (MSB), but markedly lower for the GL method (Fig. S1, Tables S2-S3). The largest difference in performance was observed when estimating $r^2$ and $D'$ at low depth, with the EM method based on GL implemented outperforming estimates based on CG. As an example, $D'$ estimates from GL at $2\times$ have similar RMSD to those based on CG at $5\times$, and $r^2$ estimates from GL at $5\times$ with similar RMSD to those based on CG at $10\times$. Estimate of $r^2$ from expected genotypes tend to display low accuracy at depths $< 10\times$ (Tables S2-S3), regardless of the method used (CG or GL).

We then assessed the accuracy of fitting LD decay curves from $r^2$ (Fig. 1, S2) and $D'$ (Fig. S3). At higher depth ($> \ 5\times$), we do not observe significant differences between CG and GL, probably because the amount of SNPs available can partially compensate for the lower coverage. However, at $<= 5\times$, we observe a loss of fitting power when using CG but not when using GL. For very low sequencing depths ($<= \ 1\times$), it is difficult to obtain reliable fittings, although this might be mitigated with larger sample sizes (Fig. S4). The script for pruning of SNPs based on their LD levels (`prune_graph.pl`) removed $3,810$ SNPs in one simulation, from a total of $67,726$, and drastically reduced LD levels as expected (Fig. S5). We also illustrate the use of the script to plot LD blocks (`LD_blocks.sh`), and plot the region between $35kb$ and $50kb$ from the simulated data at $50\times$ (Fig. S6).

## 4 Application

We analyzed mRNA sequencing data from gonad and spleen of 10 mallard ducks (*Anas platyrhynchos*) and 11 turkeys (*Meleagris gallopavo*) (Harrison *et al.*, 2015), that has highly variable coverage distributions (Fig. S7 and S8). Both populations were captive reared and have previously been shown to have different effective population sizes (Wright *et al.*, 2015) but with reasonably conserved recombination rates. We processed each data set separately and fitted an LD decay curve for each species. We observe a significantly greater intercept and slope of LD decay against physical distance in the turkey (Fig. S9), consistent with lower effective population size in the turkey population, as previously shown (Wright *et al.*, 2015).

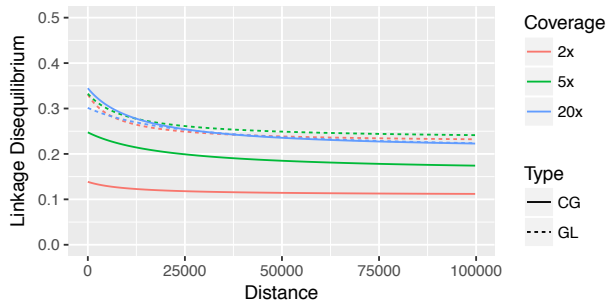


**Fig. 1.** Fitting of $r^2$ decay from simulated data from both called genotypes (CG) and genotype likelihoods (GL), across different coverages. The best-fitted curves were based on $250bps$ bins.

When comparing `ngsLD` against `GUS-LD` (Bilton *et al.*, 2018) on a subset of this real dataset, we observe that `ngsLD` is less affected by artificially downsampling data to lower depths (Fig. S10), although both methods tend to overestimate LD values.

## 5 Conclusion

`ngsLD` provides a valid and robust solution for estimating LD values from low depth sequencing data and limited sample sizes. We show that we can successfully infer LD values for depths as low as $2\times$, while still keeping acceptable error rates. We also provide several companion scripts designed to perform common LD-related analyses, such as SNP pruning and LD-decay fitting, which are the baseline for population genetic inferences. Finally, we show that `ngsLD` outperforms an existing method for LD estimation from low depth data in accuracy and memory-usage.

## Acknowledgements

We would like to thank Shyam Gopalakrishnan, Tin-Yu Hui and Ryan Waples for helpful discussions.

*Funding*: EAF was supported by funding from the MRes Computational Methods in Ecology and Evolution course at Imperial College London, and AEW by a NERC Independent Research Fellowship (NE/N013948/1).

## References

Bilton, T. P., McEwan, J. C., Clarke, S. M., Brauning, R., van Stijn, T. C., Rowe, S. J., and Dodds, K. G. (2018). Linkage disequilibrium estimation in low coverage high-throughput sequencing data. *Genetics*.

Excoffier, L. and Slatkin, M. (1995). Maximum-likelihood estimation of molecular haplotypefrequencies in a diploid population. *Mol. Biol. Evol.*, **12**(5), 921–927.

Harrison, P. W., Wright, A. E., Zimmer, F., Dean, R., Montgomery, S. H., Pointer, M. A., and Mank, J. E. (2015). Sexual selection drives evolution and rapid turnover of male gene expression. *Proc. Natl. Acad. Sci.*, **112**(14), 4393–4398.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, **27**(21), 2987–93.

Maruki, T. and Lynch, M. (2014). Genome-wide estimation of linkage disequilibrium from population-level high-throughput sequencing data. *Genetics*.

Slatkin, M. (2008). Linkage disequilibrium - Understanding the evolutionary past and mapping the medical future.

Tenesa, A., Navarro, P., Hayes, B. J., Duffy, D. L., Clarke, G. M., Goddard, M. E., and Visscher, P. M. (2007). Recent human effective population size estimated from linkage disequilibrium. *Genome Res.*, **17**(4), 520–526.

Wright, A. E., Harrison, P. W., Zimmer, F., Montgomery, S. H., Pointer, M. A., and Mank, J. E. (2015). Variation in promiscuity and sexual selection drives avian rate of Faster-Z evolution. *Mol. Ecol.*, **24**(6), 1218–1235.