Bioinformatics, 33(13), 2017, 2059–2062
doi: 10.1093/bioinformatics/btx102
Advance Access Publication Date: 16 February 2017
Applications Note

OXFORD

Genetics and population analysis

GARLIC: Genomic Autozygosity Regions Likelihood-based Inference and Classification

Zachary A. Szpiech[1,*], Alexandra Blant[2] and Trevor J. Pemberton[2]

[1]Department of Bioengineering and Therapeutic Sciences, University of California — San Francisco, San Francisco, CA 94158, USA and [2]Department of Biochemistry and Medical Genetics, University of Manitoba, Winnipeg, MB R3E 0J9, Canada

*To whom correspondence should be addressed.

Associate Editor: Bonnie Berger

Received on August 9, 2016; revised on January 24, 2017; editorial decision on February 14, 2017; accepted on February 15, 2017

Abstract

Summary: Runs of homozygosity (ROH) are important genomic features that manifest when identical-by-descent haplotypes are inherited from parents. Their length distributions and genomic locations are informative about population history and they are useful for mapping recessive loci contributing to both Mendelian and complex disease risk. Here, we present software implementing a model-based method (Pemberton *et al.*, 2012) for inferring ROH in genome-wide SNP datasets that incorporates population-specific parameters and a genotyping error rate as well as provides a length-based classification module to identify biologically interesting classes of ROH. Using simulations, we evaluate the performance of this method.

Availability and Implementation: *GARLIC* is written in C++. Source code and pre-compiled binaries (Windows, OSX and Linux) are hosted on GitHub (https://github.com/szpiech/garlic) under the GNU General Public License version 3.

Contact: zachary.szpiech@ucsf.edu

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

1 Introduction

Runs of homozygous genotypes (ROH) are a commonly used indicator of genomic autozygosity, regions of the genome where the exact same haplotype has been inherited identical by descent (IBD) from an ancestor common to both parents. Originally conceived as an approach to improve the accuracy of studies performing homozygosity mapping (Lander and Botstein, 1987) of recessive Mendelian diseases (Broman and Weber, 1999), ROH have subsequently formed the foundation of studies investigating the contributions of recessive deleterious variants to the genetic risk for complex diseases (Christofidou *et al.*, 2015; Keller *et al.*, 2012; McLaughlin *et al.*, 2015) and the genetic determination of complex traits (Campbell *et al.*, 2007; Fareed and Afzal, 2014; Howrigan *et al.*, 2015; Joshi *et al.*, 2015; McQuillan *et al.*, 2012; Power *et al.*, 2014; Rudan *et al.*, 2003) as well as the identification of novel genes underlying numerous complex diseases (Ghani *et al.*, 2015; Lencz *et al.*, 2007; McLaughlin *et al.*, 2015; Sud *et al.*, 2015; Yang *et al.*, 2012) and human standing height (Yang *et al.*, 2010). Moreover, they have provided unique insights into the demographic and sociocultural processes that have shaped genomic variation patterns in contemporary worldwide human populations (Gibson *et al.*, 2006; McQuillan *et al.*, 2008; Nalls *et al.*, 2009; Pemberton *et al.*, 2012; Szpiech *et al.*, 2013), ancient hominins (Meyer *et al.*, 2012; Prufer *et al.*, 2014), non-human primates (Prado-Martinez *et al.*, 2013; Xue *et al.*, 2015) and livestock (Curik *et al.*, 2014; Zhang *et al.*, 2015).

There are several methods for identifying ROH in genomes, which fall into two general categories: genotype-counting and model-based. Genotype-counting methods search for long tracts of consecutive homozygous genotypes with parameters that define the maximum numbers of heterozygotes and missing genotypes allowable. Software implementing such an approach include *PLINK* (Purcell *et al.*, 2007), *GERMLINE* (Gusev *et al.*, 2009) and *cgaTOH* (Zhang *et al.*, 2013). Model-based methods assert

probability models that discriminate between autozygous and non-autozygous regions and incorporate parameters such as allele frequency and recombination rate that can be estimated from the data. Software such as *BEAGLE* (Browning and Browning, 2013), *H3M2* (Magi *et al.*, 2014), *FILTUS* (Vigeland *et al.*, 2016) and *BCFtools/RoH* (Narasimhan *et al.*, 2016) utilize model-based methods.

Here we implement the model-based method for calling ROH originally published by Pemberton *et al.* (2012) as well as their population-based ROH length-based classification approach that partitions ROH into classes that broadly represent ROH generated by different population processes, a functionality that is unique to our software. We compare this method against those implemented in the popular *PLINK* (Purcell *et al.*, 2007) and the well-performing *BCFtools/RoH* (Narasimhan *et al.*, 2016) software, which are representative of their categories and well-suited to analyzing the same types of data as the Pemberton *et al.* (2012) method (e.g. *H3M2* (Magi *et al.*, 2014) and *FILTUS* (Vigeland *et al.*, 2016) are explicitly designed for analyzing whole exome sequencing data).

2 Materials and methods

Pemberton *et al.* (2012) introduced an ROH calling pipeline that uses a logarithm of the odds (*LOD*) score measure of autozygosity applied in a sliding-window framework to infer ROH in high-density genome-wide SNP genotype data. This method incorporates a genotype error rate and population-specific allele frequencies (Broman and Weber, 1999; Wang *et al.*, 2009) in contrast to popular genotype counting methods such as the one implemented in *PLINK* (Purcell *et al.*, 2007). The *LOD* score of window w in individual i is calculated as the sum of the log-likelihood ratios of the K SNPs in the window:

$$LOD(w,i) = \sum_{k=1}^{K} log_{10}\left(\frac{\Pr[G_{i,k}|X_k = 1]}{\Pr[G_{i,k}|X_k = 0]}\right).$$

Here, $\Pr[G_{i,k}|X_k = 1]$ is the probability of observing genotype $G_{i,k}$ under the hypothesis of autozygosity ($X_k = 1$), and $\Pr[G_{i,k}|X_k = 0]$ is the probability of observing genotype $G_{i,k}$ under the hypothesis of non-autozygosity ($X_k = 0$). For a biallelic locus with alleles A and B that have population frequencies p_A and p_B and a genotype error rate ε, the genotype probabilities under the autozygosity and non-autozygosity hypotheses are given in Supplementary Table S1. Genotypes with missing data are assigned a *LOD* score of 0.

Calculating $LOD(w,i)$ for all windows in all individuals in a given sample set, examination of the distribution of scores shows clear bimodality (see Fig. 1 from Pemberton *et al.*, 2012). Windows in the left-hand mode support the hypothesis of non-autozygosity and those in the right-hand mode support the hypothesis of autozygosity. As window size increases the area under the autozygous mode decreases until it disappears, likely reflecting the size above which window lengths are frequently longer than those of most common true autozygous regions and therefore encompass non-autozygous regions that mask the presence of autozygosity. A reasonable window size for ROH detection is thus the largest window size where the distribution of $LOD(w,i)$ is bimodal, with windows defined as autozygous if their $LOD(w,i)$ is greater than the local minimum between the two modes (Pemberton *et al.*, 2012).

To facilitate window size selection, *GARLIC* can begin at a user-defined window size and then increase it in increments of a user-defined step size. At each window size the distribution of $LOD(w,i)$ scores is estimated via Gaussian kernel density estimation (KDE) with the *FIGtree* package (Morariu *et al.*, 2009) until an *ad hoc* heuristic (essentially, the residual sum of squared errors from a linear spline fit) designed to summarize the magnitude of the oscillations in the KDE distribution of $LOD(w,i)$ scores (see Fig. 1 from Pemberton *et al.* (2012)) is sufficiently small; this approach broadly identifies the largest window size that is bimodal with the provided sample set. Once a window size has been chosen, windows are called autozygous if their $LOD(w,i)$ score is greater than the local minimum between the two modes in the *LOD* score KDE distribution at that window size.

ROH are then constructed in the following way. Each variable site in the data is included in multiple windows (i.e. a site is included in 100 different windows given a window size of 100), and near the edges of a true autozygous region some sites will be included in both high-scoring and low-scoring windows as the window enters or leaves a region with high support for autozygosity. Therefore we require at least a certain (user-definable with --overlap-frac flag) number of high-scoring windows to cover a site before it is included in an ROH region. Finally, we do not construct an ROH across a (user-definable with --max-gap flag) maximum gap (default 200 kb). ROH are thus comprised of a concatenation of consecutive sites that meet these criteria.

GARLIC will also classify ROH into the three length groups (Pemberton *et al.*, 2012) that broadly correspond to ROH arising from distinct processes: (i) short ROH reflecting homozygosity for ancient haplotypes contributing to local linkage disequilibrium (LD) patterns, (ii) medium ROH reflecting homozygosity arising from recent population demographic changes (e.g. bottlenecks) and (iii) long ROH reflecting homozygosity due to recent parental relatedness. This classification is performed using an inbuilt Gaussian mixture function that fits a 3-component model to the ROH length distribution or based on user-provided length thresholds (set with the --size-bounds flag).

ROH are output along with length class information in UCSC's plain-text BED format to enable easy visualization in the UCSC Genome Browser (Speir *et al.*, 2016) or further downstream analysis by the user.

3 Results

In order to facilitate the uptake of this likelihood-based ROH inference method, we implement the approach of Pemberton *et al.* (2012) in the software *GARLIC* (Genomic Autozygosity Regions Likelihood Inference and Classification). A schematic of the analysis pipeline implemented in *GARLIC* is provided in Supplementary Figure S1. Genotype data and its associated individual information are accepted in the popular transposed-pedigree (TPED) and family (TFAM) file formats. Allele frequencies used in the *LOD* score calculation can either be estimated by *GARLIC* from the genotype data or provided in a separate file. We recommend that a minimum of 25 individuals per analysis group be provided if allele frequencies are to be calculated from the provided genotype data, and if multiple groups of different sample sizes are to be analyzed jointly, the standardized size resampling flag should be used to mitigate biases introduced into allele frequency estimates by the different sizes. Note, however, that this may introduce some stochasticity in the identification of very short ROH. The genotype error rate ε to be used in the calculations is provided by the user on the command line, and could be calculated based upon the observed rate of genotype discordance between duplicate samples or based on industry-reported genotyping error rates.

We evaluate *GARLIC*'s performance using forward simulations that report true autozygous regions (Kardos *et al.*, 2015). We generate 100 replicates of 30 individuals simulated with 150k biallelic sites across 250 Mbps. We further introduce genotyping errors at a rate of 0.001. To introduce varying sizes of ROH, we vary population size over the last 50 generations of the simulation. We keep the population size constant for 20 generations, followed by an 80% reduction in population size for 10 generations, followed by a recovery for 15 generations, and then a 94% reduction in population size for the final 5 generations.

We allowed *GARLIC* to automatically choose window size (--auto-winsize), which was consistently chosen to be either 100 or 110, and we specified an error rate of 0.001. Finally, we evaluated a range of possible 'overlap fractions' to determine a default value to be used for constructing ROH. Supplementary Figure S2 suggests that this parameter be set to 0.25, which is the default setting in *GARLIC*. All other parameters were set to default values. We also compare *GARLIC* results with those of the popular genotype counting method implemented in *PLINK* (using a matched window size; Purcell *et al.*, 2007) and *BCFtools/RoH* (Narasimhan *et al.*, 2016). *GARLIC* achieves better power (Supplementary Fig. S3A) than both *PLINK* and *BCFtools/RoH*, although it has a marginally worse false positive rate (Supplementary Fig. S3B) due solely to differences in ROH boundary placement.

Overall, *GARLIC* performs comparably to existing methods, while offering the advantage of in-built ROH length classification. A limitation of *GARLIC* is that it requires population allele frequencies in order to identify ROH. While *PLINK* can easily analyze single genomes without extra information, *GARLIC* would require a separate file of allele frequencies for the individual's population of origin; *BCFtools/RoH* is similarly limited. This is likewise the case for small datasets comprised only of individuals known to be highly inbred compared to their source population. Additionally, our simulations did not consider variable recombination rate, which may adversely affect *GARLIC* performance in favor of methods that explicitly handle it. However, planned future updates to the model-based *GARLIC* method will address this.

GARLIC is user friendly and open source, offering a simple implementation of the population-specific ROH calling pipeline of Pemberton *et al.* (2012). The runtime of *GARLIC* depends on the parameters set. As a guide, we re-analyzed the Pemberton *et al.* (2012) data comprised of 1839 individuals from 64 worldwide human populations typed at 577 489 SNPs, and *GARLIC* took under 16 minutes to complete ROH calling and classification. Source code and pre-compiled binaries for *GARLIC* under the Windows, OSX and Linux environments are hosted on GitHub (https://github.com/szpiech/garlic).

Acknowledgements

The authors thank Ryan D. Hernandez and Noah A. Rosenberg for helpful comments on the manuscript. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Funding

This research was supported by an institutional start-up fund from the University of Manitoba (T.J.P.) and by the University of Manitoba Graduate Enhancement of Tri-Council Stipends (GETS) program (A.B.). Partial support to Z.A.S was provided by the National Human Genome Research Institute of the National Institutes of Health under award number R01HG007644 (awarded to Ryan D. Hernandez, University of California - San Francisco).

Conflict of Interest: none declared.

References

Broman,K.W. and Weber,J.L. (1999) Long homozygous chromosomal segments in reference families from the centre d'Etude du polymorphisme humain. *Am. J. Hum. Genet.*, **65**, 1493–1500.

Browning,B.L. and Browning,S.R. (2013) Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics*, **194**, 459–471.

Campbell,H. *et al.* (2007) Effects of genome-wide heterozygosity on a range of biomedically relevant human quantitative traits. *Hum. Mol. Genet.*, **16**, 233–241.

Christofidou,P. *et al.* (2015) Runs of homozygosity: association with coronary artery disease and gene expression in monocytes and macrophages. *Am. J. Hum. Genet.*, **97**, 228–237.

Curik,I. *et al.* (2014) Inbreeding and runs of homozygosity: a possible solution to an old problem. *Livest. Sci.*, **166**, 26–34.

Fareed,M. and Afzal,M. (2014) Evidence of inbreeding depression on height, weight, and body mass index: a population-based child cohort study. *Am. J. Hum. Biol.*, **26**, 784–795.

Ghani,M. *et al.* (2015) Association of long runs of homozygosity with Alzheimer disease among African American individuals. *JAMA Neurol.*, **72**, 1313–1323.

Gibson,J. *et al.* (2006) Extended tracts of homozygosity in outbred human populations. *Hum. Mol. Genet.*, **15**, 789–795.

Gusev,A. *et al.* (2009) Whole population, genome-wide mapping of hidden relatedness. *Genome Res.*, **19**, 318–326.

Howrigan,D.P. *et al.* (2015) Genome-wide autozygosity is associated with lower general cognitive ability. *Mol. Psychiatry*, **21**, 837–843.

Joshi,P.K. *et al.* (2015) Directional dominance on stature and cognition in diverse human populations. *Nature*, **523**, 459–462.

Kardos,M. *et al.* (2015) Measuring individual inbreeding in the age of genomics: marker-based measures are better than pedigrees. *Heredity (Edinb)*, **115**, 63–72.

Keller,M.C. *et al.* (2012) Runs of homozygosity implicate autozygosity as a schizophrenia risk factor. *PLoS Genet.*, **8**, e1002656.

Lander,E.S. and Botstein,D. (1987) Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. *Science*, **236**, 1567–1570.

Lencz,T. *et al.* (2007) Runs of homozygosity reveal highly penetrant recessive loci in schizophrenia. *Proc. Natl. Acad. Sci. U. S. A.*, **104**, 19942–19947.

Magi,A. *et al.* (2014) H3M2: detection of runs of homozygosity from whole-exome sequencing data. *Bioinformatics*, **30**, 2852–2859.

McLaughlin,R.L. *et al.* (2015) Homozygosity mapping in an Irish ALS case-control cohort describes local demographic phenomena and points towards potential recessive risk loci. *Genomics*, **105**, 237–241.

McQuillan,R. *et al.* (2012) Evidence of inbreeding depression on human height. *PLoS Genet.*, **8**, e1002655.

McQuillan,R. *et al.* (2008) Runs of homozygosity in European populations. *Am. J. Hum. Genet.*, **83**, 359–372.

Meyer,M. *et al.* (2012) A high-coverage genome sequence from an archaic Denisovan individual. *Science*, **338**, 222–226.

Morariu,V.I. *et al.* (2009) Automatic online tuning for fast Gaussian summation. In: Koller,D. *et al.* (eds.) *Advances in Neural Information Processing Systems 21*, pp. 1113–1120. Curran Associates, Inc., Red Hook, NY.

Nalls,M.A. *et al.* (2009) Measures of autozygosity in decline: globalization, urbanization, and its implications for medical genetics. *PLoS Genet.*, **5**, e1000415.

Narasimhan,V. *et al.* (2016) BCFtools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data. *Bioinformatics*, **32**, 1749–1751.

Pemberton,T.J. *et al.* (2012) Genomic patterns of homozygosity in worldwide human populations. *Am. J. Hum. Genet.*, **91**, 275–292.

Power,R.A. *et al*. (2014) Genome-wide estimates of inbreeding in unrelated individuals and their association with cognitive ability. *Eur. J. Hum. Genet*., **22**, 386–390.

Prado-Martinez,J. *et al*. (2013) Great ape genetic diversity and population history. *Nature*, **499**, 471–475.

Prufer,K. *et al*. (2014) The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*, **505**, 43–49.

Purcell,S. *et al*. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet*., **81**, 559–575.

Rudan,I. *et al*. (2003) Inbreeding and the genetic complexity of human hypertension. *Genetics*, **163**, 1011–1021.

Speir,M.L. *et al*. (2016) The UCSC Genome Browser database: 2016 update. *Nucleic Acids Res*., **44**, D717–D725.

Sud,A. *et al*. (2015) Genome-wide homozygosity signature and risk of Hodgkin lymphoma. *Sci. Rep*., **5**, 14315.

Szpiech,Z.A. *et al*. (2013) Long runs of homozygosity are enriched for deleterious variation. *Am. J. Hum. Genet*., **93**, 90–102.

Vigeland,M.D. *et al*. (2016) FILTUS: a desktop GUI for fast and efficient detection of disease-causing variants, including a novel autozygosity detector. *Bioinformatics*, **32**, 1592–1594.

Wang,S. *et al*. (2009) Genome-wide autozygosity mapping in human populations. *Genet. Epidemiol*., **33**, 172–180.

Xue,Y. *et al*. (2015) Mountain gorilla genomes reveal the impact of long-term population decline and inbreeding. *Science*, **348**, 242–245.

Yang,H.C. *et al*. (2012) A genome-wide homozygosity association study identifies runs of homozygosity associated with rheumatoid arthritis in the human major histocompatibility complex. *PLoS One*, **7**, e34840.

Yang,T.L. *et al*. (2010) Runs of homozygosity identify a recessive locus 12q21.31 for human adult height. *J. Clin. Endocrinol. Metab*., **95**, 3777–3782.

Zhang,L. *et al*. (2013) cgaTOH: extended approach for identifying tracts of homozygosity. *PLoS One*, **8**, e57772.

Zhang,Q. *et al*. (2015) Runs of homozygosity and distribution of functional variants in the cattle genome. *BMC Genomics*, **16**, 542.