Proc. Assoc. Advmt. Anim. Breed. Genet. 22:465-468

EPINETR: A FORWARD-TIME SIMULATOR FOR EPISTATIC NETWORK MODELLING IN R

Dion C. Detterer[1,2], Paul Kwan[1] and Cedric Gondro[1,2]

[1] School of Science and Technology, University of New England, Armidale NSW 2351, Australia
[2] The Centre for Genetic Analysis and Applications, University of New England, Armidale NSW 2351, Australia

SUMMARY

The problem of the missing heritability hinders our understanding of the relationship between genetic markers and complex quantitative traits, in turn limiting informed selection of mates for animal breeding purposes. To this end, we have developed epinetr, a software package for R designed to facilitate the investigation of the possible contribution of gene interaction networks to the missing heritability.

INTRODUCTION

Since the advent of the genome-wide association study (GWAS) in 2005 (Haines *et al.* 2005; Vissler *et al.* 2012), thousands of genetic variants have been identified which contribute to complex traits in either livestock (Tenghe *et. al.* 2016) or humans (Li *et al.* 2016), with an application for livestock being a genetically-informed artificial selection for desirable traits. However, a gap emerged between current heritability estimates for these traits and the contribution of the identified variants: the so-called "missing heritability" problem (Manolio *et al.* 2009; Zuk *et al.* 2014). Several explanations were put forth to explain this disparity (Manolio *et al.* 2009; Eichler *et al.* 2010); among these, the effect of epistasis (i.e. gene-gene interaction) on heritability estimates is an explanation that has attracted considerable attention (Huang 2012; Zuk *et al.* 2012; Bloom *et al.* 2013). Simulations are currently the most viable approach to test epistatic models and how they affect our estimates of additive genetic variance (Hoban *et al.* 2012).

There is thus a need in animal breeding for flexible simulators that can accommodate a wide variety of randomly-generated and user-generated epistatic models while still providing parameters to control other factors. As an aid to further research on the genetic architecture of epistasis, a need also exists for a network-based approach to epistatic modelling in simulators. To this end, we have developed epinetr, a package for the statistical environment R, soon to be submitted to CRAN: epinetr is a forward-time simulator designed specifically for the study of high-order epistatic networks and how they impact estimates of genetic parameters and selection decisions of complex quantitative traits.

This paper first gives an overview of the design decisions behind epinetr, it then discusses the epinetr simulator itself, the features and parameters within the simulator and its ability to handle complex epistatic networks.

DESIGN CONSIDERATIONS

The two broad categories of population genetics simulators form a simple dichotomy: simulators that work forwards-in-time and those that work backwards-in-time (Hoban *et al.* 2012). As can be inferred from the nomenclature, forwards-in-time (or forward-time) simulators start with a population and work forwards to track individuals and pedigrees via selection, recombination and mutation across generations; on the other hand, backwards-in-time (or coalescent) simulators work backwards to infer genetic histories. Forwards-in-time simulators demand more computational resources than backwards-in-time simulators simply due to the level of granularity required (i.e. per-individual simulation); at present, forwards-in-time simulators include Easypop (Balloux 2001),

GenomePop (Carvajal-Rodríguez 2008) and FREGENE (Chadeau-Hyam 2008), none of which include mention of epistatic modelling capabilities in the associated literature. Both simuPOP (Peng and Kimmel 2005) and quantiNemo (Neuenschwander 2008) are forwards-in-time simulators that do allow for statistical epistatic modelling; the same is true for the more recent simulator SELAM (Corbett-Detig and Jones 2016).

Backwards-in-time simulators such as SNPsim (Posada and Wiuf 2003), SIMCOAL2 (Laval and Excoffier 2004), GENOME (Liang *et al.* 2007) and MaCS (Chen *et al.* 2009) are typically more computationally efficient than forwards-in-time simulators, but there is a trade-off: they are not as suited to modelling complexity or natural or artificial selection (Hoban *et al.* 2012). This limits their application to the study of epistatic impact on selection for complex traits.

Existing outside this dichotomy is EpiSIM (Shang 2013), which allows for the simulation of simple 2-way interactions.

The choice was made to build a forward-time simulator, as this allowed for the use of complex selection scenarios. As a further consideration, there is evidence to suggest that epistatic networks exhibit a small world or scale-free structure (Tyler *et al.* 2009; Mackay 2014). While this appears to be a fruitful avenue to pursue, a more general point emerges: the actual network structure may be the key to understanding the underlying mechanics of epistasis, including the relationship between genes and phenotypes. For this reason, epinetr includes the ability to both automatically generate random and scale-free epistatic networks or alternatively input user-defined epistatic networks that can be generated by an external model based on previous knowledge (or a hypothesis) of the underlying architecture of a trait.

In a nutshell, the epinetr package is designed as a tool to investigate potential epistatic sources of missing heritability using network models.

PACKAGE FEATURES

The epinetr package is written for the R statistical software environment, allowing for complex analysis to take place in the same environment as the actual simulation. It includes a set of classes that enable users to perform common operations both before and after the simulation with simple commands, as well as provisions for specifying a large set of population parameters.

Typically, there are 5 broad steps in the workflow:

1. Define population parameters and construct the initial population
2. Attach additive effects to the population
3. Attach an epistatic network to the population and visualise the network
4. Run a forward-time simulation of the population and plot the simulation run

Parameters are specified using a simple parameter file. Below we give an overview of the parameter options available.

Population size, given at initialisation, is fixed throughout the simulation run. However, because litter size is specified by a user-defined probability mass function, some generations may be smaller than the fixed population size. For this reason, another pair of parameters controlling the maximum lifespans of sires and dams may be violated.

Allele frequencies can be inferred from a haplotype file or specified directly, thus allowing for "sideways simulation" (by first using a coalescent simulator to arrive at the allele frequencies); alternatively, haplotypes can be used directly as the initial population.

Both broad- and narrow-sense heritability can be specified, controlling the contributions of additive, epistatic and environmental effects to the overall variance of the trait being studied.

Selection is performed either randomly or via linear ranking; the mutation rate is a single number while recombination probabilities can be optionally specified, thus allowing for the simulation of hotspots. Separate truncation rates for sires and dams can also be specified, as can an initial burn-in period of random selection.

A chromosomal map for the single nucleotide polymorphisms (SNP) is required, with the user determining which SNP are used for quantitative trait loci (QTL) in the epistatic network; alternatively, the user can specify the number of QTL which are then selected from the SNP at random.

The number of times a sire can mate during a single generation can be specified.

Once a population is generated using the above parameters, additive effects across all SNP can then be attached. Effect sizes (i.e. the absolute value of the coefficients) are determined by the restrictions of the population parameters; however, they can be sampled from any distribution specified by the user, including user-defined functions.

Epistatic modelling. By specifying an incidence matrix (with each row representing a QTL and each column representing an interaction between QTL), the user can determine the structure of the epistatic network; alternatively, the system can generate a random or scale-free network for the population with a single command. In either case, the orders of interaction included in the network are specified by the user and limited only by the number of QTL in the population; in addition, scale-free networks can be given a minimum number of interactions per QTL.

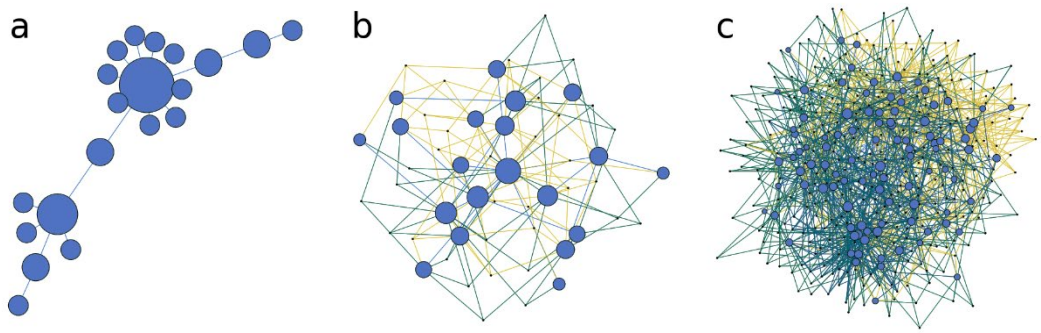

Figure 1. Three unique scale-free epistatic networks generated automatically from within epinetr: a) a 20-QTL network comprised of 2-way interactions; b) a 20-QTL network comprised of 2-, 3- and 4-way interactions; and c) a 100-QTL network comprised of 2-, 3-, 4- and 5-way interactions

The network structure can be easily visualised using a simple plot command. Figure 1 depicts three potential epistatic scale-free networks generated automatically and visualised from within epinetr.

The result of a simulation run is a set of files giving allele frequencies and pedigrees for each individual in each generation, as well as haplotypes for each individual in the final generation (or, optionally, each generation). Most importantly, the additive, epistatic and environmental contribution to each individual's phenotype is given as an output. Finally, the mean, maximum and minimum phenotypic values within the population across generations can also be easily plotted using a single command.

CONCLUSION

epinetr is an R package designed to facilitate the modelling and analysis of epistatic networks and their effects on estimates of genetic parameters and selection decisions within populations, filling an important niche in population genetics simulation. It is hoped that it will be a valuable tool to better understand how different models of genetic architecture, particularly epistasis, relate to the problem of missing heritability.

ACKNOWLEDGEMENT
This project was supported by a grant from the Next-Generation BioGreen 21 Program (No. PJ01134906), Rural Development Administration, Republic of Korea and Australian Research Council (DP130100542).

REFERENCES
Balloux F. (2001). *J. Hered.* **92**: 301.
Bloom J.S., Ehrenreich I.M., Loo W.T., Lite T.L.V. and Kruglyak L. (2013) *Nature* **494**: 234.
Carvajal-Rodríguez A. (2008) *BMC Bioinformatics* **9**: 223.
Chadeau-Hyam M., Hoggart C.J., O'Reilly P.F., Whittaker J.C., De Iorio M. and Balding D.J. (2008) *BMC Bioinformatics* **9**: 364.
Chen G.K., Marjoram P. and Wall J.D. (2009). *Genome Res.* **19**: 136.
Corbett-Detig R. and Jones, M. (2016) *Bioinformatics* **32**: 3025.
Eichler E.E., Flint J., Gibson G., Kong A., Leal S.M., Moore J.H. and Nadeau J.H. (2010) *Nature Rev. Genet.* **11**: 446.
Haines J.L., Hauser M.A., Schmidt S., Scott W.K., Olson L.M, Gallins P., Spencer K.L., Kwan S.Y., Noureddine M., Gilbert J.R., Schnetz-Boutaud M., Agarwal A., Postel E.A. and Pericak-Vance M.A. (2005) *Science* **308**: 419.
Hoban S., Bertorelle G. and Gaggiotti O.E. (2012) *Nature Rev. Genet.* **13**: 110.
Huang W., Richards S., Carbone M.A., Zhu D., Anholt R.R., Ayroles J.F., Duncana L., Jordana K.W., Lawrencea F., Magwirea M.M., Warner C.B., Blankenburg K., Han Y., Javaid M., Jayaseelan J., Jhangiani S.N., Muzny D., Ongeri F., Perales L., Wu Y., Zhang Y., Zou X., Stone E.A., Gibbs R.A. and Mackay T.F.C. (2012) *Proc. Natl. Sci. USA* **109**: 15553.
Laval G. and Excoffier L. (2004) *Bioinformatics* **20**: 2485.
Li M.J., Liu Z., Wang P., Wong M.P., Nelson M.R., Kocher J.P.A., Yeager M., Sham P.C., Chanock S.J., Xia Z. and Wang, J. (2016) *Nucleic Acids Res.* **44**: D869.
Liang L., Zöllner S. and Abecasis G.R. (2007) *Bioinformatics*, **23**: 1565.
Mackay, T.F. (2014) *Nature Rev. Genet.* **15**: 22.
Manolio T.A., Collins F.S., Cox N.J., Goldstein D.B., Hindorff, L.A., Hunter D.J., McCarthy M.I., Ramos E.M., Cardon L.R., Chakravarti A., Cho J.H., Guttmacher A.H., Kong A., Kruglyak L., Mardis E., Rotimi C.N., Slatkin M., Valle D., Whittemore A.S., Boehnke M., Clark A.G., Eichler E.E., Gibson G., Haines J.L., Mackay T.F.C., McCarroll S.A. and Visscher P.M. (2009) *Nature* **461**: 747.
Neuenschwander S., Guillaume F. and Goudet, J. (2008) *Bioinformatics* **24**: 1552.
Peng B. and Kimmel M. (2005) *Bioinformatics* **21**: 3686.
Posada D. and Wiuf, C. (2003) *Bioinformatics* **19**: 289.
Shang J., Zhang J., Lei X., Zhao W. and Dong Y. (2013) *Genes Genomics* **35**: 305.
Tenghe A.M.M., Bouwman A.C., Berglund B., Strandberg E., de Koning D.J. and Veerkamp R. F. (2016) *J. Dairy Sci.* **99**: 5470.
Tyler A.L., Asselbergs F.W., Williams S.M. and Moore, J. H. (2009) *Bioessays* **31**: 220.
Visscher P.M., Brown M.A., McCarthy M.I. and Yang J. (2012) *Am. J. Hum. Genet.* **90**: 7.
Zuk O., Hechter E., Sunyaev S.R. and Lander E.S. (2012) *Proc. Natl. Sci. USA* **109**: 1193.
Zuk O., Schaffner S.F., Samocha K., Do R., Hechter E., Kathiresan S., Daly M.J., Neale B.M., Sunyaev S.R. and Lander, E.S. (2014*) Proc. Natl. Sci. USA* **111**: E455.