Bioinformatics, 32(4), 2016, 557–562
doi: 10.1093/bioinformatics/btv626
Advance Access Publication Date: 29 October 2015
Original Paper

OXFORD

Genetics and population analysis

SeqSIMLA2_exact: simulate multiple disease sites in large pedigrees with given disease status for diseases with low prevalence

Po-Ju Yao and Ren-Hua Chung*

Division of Biostatistics and Bioinformatics, Institute of Population Health Sciences, National Health Research Institutes, Zhunan, Miaoli County, Taiwan

*To whom correspondence should be addressed.

Associate Editor: Janet Kelso

Received on April 17, 2015; revised on September 8, 2015; accepted on October 22, 2015

Abstract

Summary: It is difficult for current simulation tools to simulate sequence data in a pre-specified pedigree structure and pre-specified affection status. Previously, we developed a flexible tool, SeqSIMLA2, for simulating sequence data in either unrelated case-control or family samples with different disease and quantitative trait models. Here we extended the tool to efficiently simulate sequences with multiple disease sites in large pedigrees with a given disease status for each pedigree member, assuming that the disease prevalence is low.

Availability and implementation: SeqSIMLA2_exact is implemented with C++ and is available at http://seqsimla.sourceforge.net.

Contact: rchung@nhri.org.tw

1 Introduction

Family studies based on next-generation sequencing (NGS) have become important in complex disease studies because a sufficient number of rare alleles that co-segregate with the disease can be observed in pedigrees (Wijsman, 2012). To evaluate the statistical power for family-based disease studies using NGS, it is necessary to simulate sequencing data in a given family structure with disease status. Moreover, as multiple variants can contribute to the disease, it is crucial to simulate the joint effects of multiple variants on the disease.

We searched the Genetic Simulation Resources website (Peng *et al.*, 2013), which provides a comprehensive catalog of simulation tools for genetics studies, and identified four tools, ForSIM (Lambert *et al.*, 2008), Nemo (Guillaume and Rougemont, 2006), simuPOP (Peng and Amos, 2008) and SeqSIMLA2 (Chung *et al.*, 2015), that can simulate sequence data in families with affection status. ForSIM, Nemo and simuPOP use forward-time algorithms, which dynamically determine pedigree structures during the simulation. Therefore, it may not be easy to obtain pre-specified pedigree structures from the simulation results. SeqSIMLA2 uses a gene-dropping algorithm to easily simulate pedigrees with pre-specified structures. However, it is very time consuming for the gene-dropping algorithm to simulate exactly the same affection status as the pre-specified status for every family member, particularly in large pedigrees and diseases with low prevalence. Another simulation tool, SimRare (Li *et al.*, 2012), which shares some properties with SeqSIMLA2, can simulate multiple disease loci in sequence data; however, SimRare only simulates unrelated case-control data.

In order to efficiently simulate genotypes conditional on the family structure and disease status, a retrospective likelihood, which is the probability of observing genotypes in each family member given the disease status, needs to be calculated. However, calculating the retrospective likelihood is typically based on computationally intensive algorithms such as Elston–Stewart (Elston and Stewart, 1971) and Markov Chain Monte Carlo (MCMC) algorithms. For example, tools such as FastSLINK (Ott, 1989), SUP (Lemire, 2006), SimWalk2 (Sobel and Lange, 1996) and MORGAN (Thompson, 2005) have implemented or extended one of the aforementioned algorithms for simulating genotypes given the trait data. Due to the computational complexity, these tools can simulate only one disease locus. Moreover, although multiple markers not associated with the disease locus can be simulated, SimWalk2 and MORGAN assume

that the markers are in linkage equilibrium, while FastSLINK and SUP can simulate few markers that are in linkage disequilibrium (LD). These tools may not be suitable to evaluate the joint effects of multiple disease loci. Moreover, these tools are not suitable to simulate NGS data, which involve millions of variants with complex LD structures.

Here we extended SeqSIMLA2 to SeqSIMLA2_exact for simulating sequence data given the pedigree structures and disease status for all family members. When the disease prevalence is low, we show that the retrospective probability can be efficiently approximated without using computationally intensive algorithms.

2 Methods

For a pedigree with n members, assume $\mathbf{x} = (x_1, ..., x_n)$ is a vector of disease phenotypes and $\mathbf{h} = (h_1, ..., h_n)$is a vector of haplotype pairs at the disease loci for the n individuals. Assume $\mathbf{c} = (c_1, ..., c_s)$ is a vector of recombination rates at s recombination hotspots within the haplotype region. Family members without parents are referred to as founders, while family members with parents in the pedigree structure are referred to as non-founders. Assume $\mathbf{h_f}$ is a subset of $\mathbf{h}$ that contains haplotype pairs of founders and $\mathbf{h_{nf}}$ is the complement of $\mathbf{h_f}$ containing haplotype pairs of non-founders. The probability of $\mathbf{h}$ can be calculated as

$$P(\mathbf{h}|\mathbf{x},\mathbf{c}) = P(\mathbf{h_f}, \mathbf{h_{nf}}|\mathbf{x},\mathbf{c}) = P(\mathbf{h_f}|\mathbf{x},\mathbf{c})P(\mathbf{h_{nf}}|\mathbf{h_f},\mathbf{x},\mathbf{c}) \quad (1)$$

First we calculate $P(\mathbf{h_f}|\mathbf{x},\mathbf{c})$. Assume $(h_1, \ldots h_{i-1})$ is a subset of haplotypes in $\mathbf{h_f}$ that have been simulated. As shown in the Appendix A, assuming no inbreeding and random mating, the probability for h_i at founder i is calculated as

$$P(h_i|h_1, ..., h_{i-1}, \mathbf{x}, \mathbf{c}) = \frac{P(h_i, x_i, \mathbf{c})P(x_{i+1}, ..., x_n|\mathbf{c}, h_1, ..., h_i, x_1, ..., x_i)}{P(x_i, \mathbf{c})P(x_{i+1}, ..., x_n|\mathbf{c}, h_1, ..., h_{i-1}, x_1, ..., x_i)} \quad (2)$$

The denominator is the sum of the numerators over all possible haplotypes of individual i. The probability $P(h_i, x_i, \mathbf{c})$ is further derived as $P(x_i \mid h_i, \mathbf{c})P(h_i, \mathbf{c})$, which can be easily calculated based on the penetrance functions and disease haplotype frequencies. Calculating the probability $P(x_{i+1}, ..., x_n|\mathbf{c}, h_1, ..., h_i, x_1, ..., x_i)$ in Eq. (2) requires the consideration of affection status for family members who have not been simulated, given the haplotypes and affection status for family members who have been simulated. This process can be time consuming for a large pedigree. To efficiently calculate the probability, we only consider the affection status for family members in the same nuclear family as individual i. It is also commonly assumed that given a person's haplotypes, the person's affection status is independent of any other person's haplotypes (Laird and Lange, 2011). Moreover, conditional on parental haplotypes, siblings' haplotypes are independent. Based on these assumptions, the probability is written as

$$P(x_{i+1}, ..., x_n|\mathbf{c}, h_1, ..., h_i, x_1, ..., x_i) = P(x_{i+1}, ..., x_k|\mathbf{c}, h_1, ..., h_i, x_1, ..., x_i)$$
$$= \sum_{r\in\Omega}\prod_{r=i+1}^{k} P(x_r|h_r, \mathbf{c}) \prod_{r=i+1}^{k} P(h_r|h_1, ..., h_i) \quad (3)$$

where individuals $i+1$ through k are in the same nuclear family as individual i, and Ω is a set of all possible haplotypes of individuals $i+1$ through k given the haplotypes for family members that have been simulated. Note that we assume that the haplotypes and affection status for the family members who have been simulated (i.e. individuals 1 to i) are independent of the affection status of other family members who are not in the same nuclear families as individuals 1 to i. Our simulation results in the Results section suggest that this assumption holds for diseases with low prevalence.

Next we calculate $P(\mathbf{h_{nf}}|\mathbf{h_f},\mathbf{x},\mathbf{c})$in Eq. (1). If h_j is the haplotype pair at non-founder j and $(h_m, \ldots h_{j-1})$ is a subset of haplotypes in $\mathbf{h_{nf}}$ that have been simulated, the probability of haplotype pair h_j at non-founder j can be calculated as

$$P(h_j|\mathbf{h_f}, h_m, ..., h_{j-1}, \mathbf{x}, \mathbf{c})$$
$$= P(h_j|h_p, h_q, \mathbf{c})P(x_{j+1}, ...x_n|\mathbf{h_f}, h_m, ..., h_j, x_1, ..., x_j, \mathbf{c})\prod_{r=1}^{j} P(x_r|h_r, \mathbf{c})/D \quad (4)$$

where $(h_p, h_q) \in \mathbf{h_f}$ are the parental haplotypes of h_j and D is the sum of the numerators over all possible haplotype pairs of a child whose parental haplotypes are(h_p, h_q). Detailed derivation of Eq. (4) can be found in the Appendix A.

Based on Eqs. (2) and (4), we developed an efficient algorithm to simulate disease haplotypes for large pedigrees:

1. Simulate $\mathbf{h_f}$ based on Eq. (2).
2. Identify one of the top founders m and set one of m's children as k.
3. Simulate h_k for k using Eq. (3).
4. 4.1. If k has a spouse, set k's spouse as k and go to step 3.
   4.2. If k has at least one sibling, set one of k's siblings as k and go to step 3. Repeat Step 4.2 until all siblings of k have been simulated.
   4.3. If k has at least one child, set one of k's children as k and go to step 3. Repeat Step 4.3 until all children of k have been simulated.

After the disease haplotypes have been simulated for all members, the sequences for the members can also be efficiently simulated using the original SeqSIMLA algorithm (Chung and Shih, 2013). To be more specific, SeqSIMLA requires a population of sequences (i.e. haplotypes) in an external file. As previously described in (Chung and Shih, 2013), these sequences can be generated by several simulation tools such as HAPGEN2 (Su *et al.*, 2011) and COSI (Schaffner *et al.*, 2005). In general, a population of 10 000 sequences is required for SeqSIMLA. Each sequence then has an index, which ranges from 1 to 10 000. After reading the external sequences, SeqSIMLA extracts the haplotypes at the disease sites selected by the user. The haplotype frequencies and haplotype pair frequencies are calculated based on the external sequences. When a haplotype pair h_i in founder i is determined based on Eq. (2), two haplotypes in the external haplotypes which are consistent with h_i are randomly selected. The indices for the two randomly selected haplotypes are assigned to individual i. When a haplotype pair h_j in non-founder j is determined based on Eq. (4), the inheritance vectors, including the crossover events, are assigned to individual j. Finally, after all members in a pedigree have been simulated, their sequences can subsequently be simulated rapidly based on the founders' sequence indices, non-founders' inheritance vectors and the external sequences.

We used simulation studies to evaluate the performance of the algorithm. Different simulation scenarios with parameters of minor allele frequencies (MAFs) and odds ratios for the disease sites, disease prevalence and family structures were used. We selected three common variants with MAFs of 0.03, 0.05 and 0.13 as the disease sites. We also selected three rare variants with MAFs of

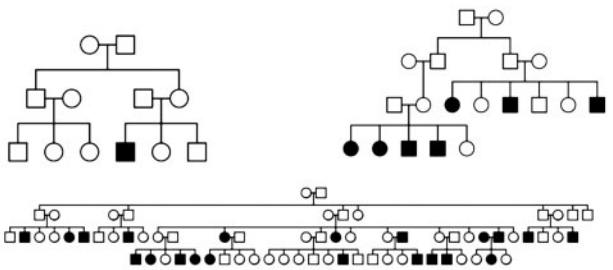

Fig. 1. Three types of pedigree structures, representing small, medium and large pedigrees, used in the simulations

0.005, 0.007 and 0.009 as another set of disease sites. A penetrance model was used based on the following logistic penetrance function in SeqSIMLA2:

$\text{logit}(P(\text{affected})) = \beta_0 + \sum_{i=1}^{u} \beta_i x_i$, where β_0 is determined by the disease prevalence, and β_i and x_i are the log odds ratio and genotype coding, respectively, for disease site i in the u disease sites. We specified the odds ratios of 1, 2 and 3 for the three common variants and higher odds ratios of 1, 5 and 8 for the three rare variants with different disease prevalences (i.e. 0.5%, 1%, 3%, 5% and 10%). Three types of family structures, shown in Figure 1, representing small, medium and large pedigrees were used.

We evaluated the performance of SeqSIMLA2_exact based on whether the 95% confidence intervals for the estimated odds ratios for the disease sites in a simulated replicate contained the true parameters specified. UNPHASED (Dudbridge, 2008) was used to estimate the odds ratios and their 95% confidence intervals for each of the simulated replicates. For each replicate, 10 000 families were simulated unless otherwise specified. Fewer pedigrees were simulated for some scenarios when analyzing 10 000 pedigrees in UNPHASED was prohibitively time-consuming. A retrospective likelihood, which was the probability of observing the genotypes in parents and their children given the affection status for the children in each nuclear family, was derived in UNPHASED. The effect sizes for the disease sites were obtained from the maximum likelihood estimates based on the likelihood function. A large pedigree is split into independent nuclear families for the inference of likelihood in UNPHASED. When multiple disease sites were simulated in our simulation studies, UNPHASED was used to estimate the odds ratios of haplotypes at the disease sites. The odds ratio for a disease site was obtained from the odds ratio for the haplotype with the disease allele at the disease site, while alleles at other disease sites were the non-disease alleles in the haplotype.

3 Results

Table 1 shows the estimates of odds ratios for the three common variants with different prevalences for the medium pedigree structure. The 95% confidence intervals generally included the true odds ratios when the disease prevalence was less than 5% or when the specified odds ratio was 1, but the estimates of odds ratios can be smaller than the true values when the disease prevalence was above 5% and the true values were not 1 (i.e. values marked in bold in Table 1). This reflects the fact that the approximation of Eq. (3) within nuclear families may not be correct when the disease prevalence is high. Table 2 shows the estimates of odds ratios for the three common variants with different pedigree structures when the prevalence is 1%. All the 95% confidence intervals included the true odds

Table 1. Estimated odds ratios for the three common variants with different prevalences

	Prevalence				
Odds ratio	0.5%	1%	3%	5%	10%
MAF = 0.13					
1[a]	0.99[b] (0.97,1.01)[c]	1.00 (0.98,1.03)	1.01 (0.98,1.03)	1.02 (0.99,1.04)	1.00 (0.96,1.03)
2	2.04 (1.98,2.09)	2.00 (1.95,2.05)	1.98 (1.92,2.03)	**1.92** (**1.88,1.97**)	**1.82** (**1.76,1.89**)
3	3.05 (2.94,3.16)	3.03 (2.93,3.14)	**2.84** (**2.75,2.94**)	**2.79** (**2.70,2.88**)	**2.51** (**2.40,2.62**)
MAF = 0.05					
1	1.00 (0.97,1.04)	0.98 (0.95,1.02)	0.99 (0.96,1.03)	0.97 (0.94,1.01)	1.01 (0.96,1.05)
2	2.00 (1.92,2.07)	2.01 (1.94,2.09)	1.98 (1.91,2.05)	**1.92** (**1.85,1.99**)	**1.82** (**1.76,1.89**)
3	3.01 (2.87,3.15)	3.04 (2.91,3.19)	2.93 (2.80,3.06)	**2.75** (**2.63,2.88**)	**2.51** (**2.40,2.62**)
MAF = 0.03					
1	1.00 (0.96,1.05)	1.00 (0.95,1.04)	1.03 (0.98,1.07)	1.01 (0.96,1.06)	1.01 (0.96,1.05)
2	1.97 (1.88, 2.07)	1.95 (1.85,2.04)	1.94 (1.85,2.04)	1.97 (1.88,2.08)	**1.82** (**1.73,1.91**)
3	3.20 (2.99,3.42)	3.08 (2.89,3.28)	3.00 (2.82,3.19)	2.94 (2.76,3.13)	**2.49** (**2.35,2.63**)

[a]The specified odds ratio.
[b]The estimated odds ratio.
[c]The 95% confidence interval.

Table 2. Estimated odds ratios for the three common variants based on the prevalence of 0.01 for different pedigree structures

	Pedigree size		
Odds ratio	Small	Medium	Large
MAF = 0.13			
1[a]	0.98[b] (0.92,1.04)[c]	0.99 (0.97,1.00)	0.98[d] (0.92,1.05)
2	1.93 (1.82,2.06)	1.98 (1.92,2.04)	1.90 (1.77,2.05)
3	2.97 (2.77,3.18)	3.00 (2.88,3.14)	3.06 (2.79,3.34)
MAF = 0.05			
1	0.97 (0.88,1.06)	1.02 (0.98,1.05)	0.96 (0.88,1.05)
2	1.87 (1.70,2.05)	2.05 (1.95,2.14)	2.14 (1.91,2.34)
3	3.12 (2.81,3.46)	3.05 (2.86,3.25)	3.01 (2.62,3.46)
MAF = 0.03			
1	0.99 (0.88,1.11)	0.96 (0.91,1.00)	1.06 (0.93,1.20)
2	2.15 (1.89,2.44)	1.96 (1.85,2.08)	2.11 (1.81,2.47)
3	3.15 (2.74,3.61)	2.99 (2.75,3.24)	3.34 (2.77,4.03)

[a]The specified odds ratio.
[b]The estimated odds ratio.
[c]The 95% confidence interval.
[d]The estimates for large pedigrees were based on 500 pedigrees.

ratios with different pedigree structures, suggesting that pedigree size does not bias the estimates. Table 3 shows the estimates of odds ratios for the three rare variants with the medium pedigree structure at the disease prevalence of 1%. Again, the 95% confidence intervals all included the true values. We also selected a mixture of five rare and common variants with disease prevalence of 1%, and found that with the exception of one case (values marked in bold), the 95% confidence intervals included the true values (Table 4).

Table 5 shows the run time of SeqSIMLA2_exact when simulating different numbers of disease sites in 2223 variants in 100 pedigrees with different structures, where *s* (i.e. the number of recombination hotspots within the haplotype region for disease sites) was fixed at 2. The run time was evaluated on a Linux server with Xeon 2.0 GHz CPUs and 96 GB of memory. The results showed that simulating 5 disease sites in 100 small or medium pedigrees can be completed in 15 s. Moreover, simulating a challenging scenario of 6900 individuals in 100 large pedigrees with exactly the same disease status for each individual as specified can be completed in 14 minutes. However, increasing the number of disease sites significantly increased the run time. Figure 2 shows the run time of SeqSIMLA2_exact when simulating 10 disease sites in 2223 variants in 100 medium pedigrees with different values of *s*. Similar to the number of disease sites, increasing *s* also significantly increased the run time.

4 Discussion

An efficient algorithm to simulate sequence data given the family structure and affection status for every family member was developed and implemented. Our simulation results suggested that the tool properly simulated multiple disease variants with the desired odds ratios when the disease prevalence was less than 5%, which is the case for many complex diseases, such as schizophrenia, autism and age-related macular degeneration that have prevalences of 1.1%, 1.4% and 1.5% in Caucasians, respectively.

As shown in Table 5 and Figure 2, the run time for SeqSIMLA2_exact increased exponentially with either the number of disease loci or the number of recombination hotspots. This can be attributed to calculating the probabilities in Eqs. (3) and (4), which requires enumeration of all possible haplotypes for individuals who have not been simulated under all possible recombination patterns, and the number of possible haplotypes and the number of recombination patterns increase exponentially with the number of disease sites and recombination hotspots, respectively. However, these calculations are only required in the first batch of the simulations. The results can subsequently be saved in memory and used in the following batches of simulations.

In our simulations, a logistic penetrance function was used to simulate multiple independent disease loci with additive effects on the disease. However, different disease models can also be simulated using SeqSIMLA2_exact. A penetrance table with penetrance functions for all possible combinations of genotypes at the disease sites can be provided to SeqSIMLA2_exact. These penetrance functions can be specified based on any type of disease model, including the incorporation of interacting loci. Moreover, different directions of effects of the disease loci can be simulated using the logistic penetrance function or the penetrance table. For example, specifying positive and negative values for the log odds ratios in the logistic penetrance function can simulate the risk and protective effects of the disease loci, respectively.

Table 3. Estimated odds ratios for the three rare variants based on the prevalence of 0.01

	MAF		
Odds ratio	0.005	0.007	0.01
1[a]	1.02[b] (0.94,1.12)[c]	1.03 (0.94,1.12)	1.00 (0.94,1.07)
5	4.75 (4.40,5.13)	4.88 (4.53,5.26)	5.01 (4.69,5.34)
8	8.07 (7.29,8.94)	8.07 (7.32,8.90)	7.96 (7.28,8.71)

[a]The specified odds ratio.
[b]The estimated odds ratio.
[c]The 95% confidence interval.

Table 4. Estimated odds ratios for a mixture of common and rare variants based on the prevalence of 0.01 using 1000 medium pedigrees

	MAF				
Odds ratio	0.374	0.286	0.056	0.020	0.011
1[a]	0.98[b] (0.91,1.05)[c]	1.03 (0.95,1.10)	1.01 (0.85,1.19)	0.77 (0.55,1.07)	1.23 (0.93,1.63)
3	3.19 (2.57,3.97)	3.31 (2.66,4.11)	**4.64 (3.28,6.55)**	3.26 (2.15,4.95)	2.66 (1.46,4.84)
5	4.08 (3.01,5.52)	4.14 (3.09,5.56)	4.38 (3.11,6.17)	5.83 (2.74,12.4)	3.05 (1.37,6.82)

[a]The specified odds ratio.
[b]The estimated odds ratio.
[c]The 95% confidence interval.

Table 5. Run time for simulating different numbers of disease sites in 100 pedigrees with different structures

	Run time		
Number of disease sites	Small	Medium	Large
1	10 s	9 s	32 s
5	12 s	15 s	13 m 38 s
10	26 s	1 m 57 s	6 h 57 m 30 s
20	5 m 22 s	21 m 8 s	58 h 54 m 17 s

s, second; m, minute; h, hour.

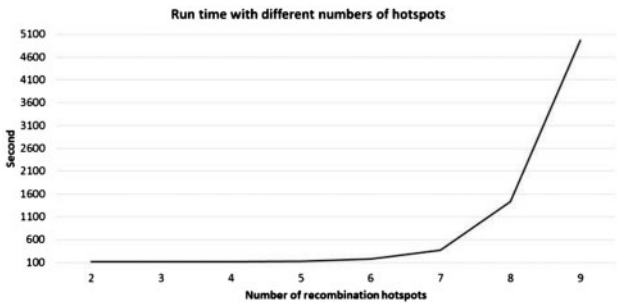

Fig. 2. Run time for simulating 100 medium pedigrees and 10 disease sites with different recombination hotspots

At present, SeqSIMLA2_exact assumes that the affection statuses for all family members are given. To account for missing phenotypes, the probability in Eq. (1) can be extended to $P(\mathbf{h}, \mathbf{x}_u|\mathbf{x}, \mathbf{c})$, where $\mathbf{x}_u$ is a vector of the affection statuses of individuals whose affection statuses were not given. We hope to implement this function in SeqSIMLA2_exact in the near future.

In conclusion, SeqSIMLA2_exact will be useful to simulate sequence data with multiple disease sites in large pedigrees when the disease prevalence is low. SeqSIMLA2_exact is implemented in SeqSIMLA2 with the -exact option, which is available from http://seqsimla.sourceforge.net.

Funding

This study was supported by grants from the National Health Research Institutes (PH-104-PP-10) and National Science Council (NSC 102-2221-E-400-001-MY2) in Taiwan.

Conflict of Interest: none declared.

References

Chung,R.H. and Shih,C.C. (2013) SeqSIMLA: a sequence and phenotype simulation tool for complex disease studies. *BMC Bioinformatics*, **14**, 199.

Chung,R.H., *et al.* (2015) SeqSIMLA2: simulating correlated quantitative traits accounting for shared environmental effects in user-specified pedigree structure. *Genet. Epidemiol.*, **39**, 20–24.

Dudbridge,F. (2008) Likelihood-based association analysis for nuclear families and unrelated subjects with missing genotype data. *Hum. Hered.*, **66**, 87–98.

Elston,R.C. and Stewart,J. (1971) A general model for the genetic analysis of pedigree data. *Hum. Hered.*, **21**, 523–542.

Guillaume,F. and Rougemont,J. (2006) Nemo: an evolutionary and population genetics programming framework. *Bioinformatics*, **22**, 2556–2557.

Laird,N.M. and Lange,C. (2011) *The Fundamentals of Modern Statistical Genetics*. New York: Springer.

Lambert,B.W. *et al.* (2008) ForSim: a tool for exploring the genetic architecture of complex traits with controlled truth. *Bioinformatics*, **24**, 1821–1822.

Lemire,M. (2006) SUP: an extension to SLINK to allow a larger number of marker loci to be simulated in pedigrees conditional on trait values. *BMC Genet.*, **7**, 40.

Li,B. *et al.* (2012) SimRare: a program to generate and analyze sequence-based data for association studies of quantitative and qualitative traits. *Bioinformatics*, **28**, 2703–2704.

Ott,J. (1989) Computer-simulation methods in human linkage analysis. *Proc. Natl. Acad. Sci. USA*, **86**, 4175–4178.

Peng,B. and Amos,C.I. (2008) Forward-time simulations of non-random mating populations using simuPOP. *Bioinformatics*, **24**, 1408–1409.

Peng,B. *et al.* (2013) Genetic Simulation Resources: a website for the registration and discovery of genetic data simulators. *Bioinformatics*, **29**, 1101–1102.

Schaffner,S.F. *et al.* (2005) Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.*, **15**, 1576–1583.

Sobel,E. and Lange,K. (1996) Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. *Am. J. Hum. Genet.*, **58**, 1323–1337.

Su,Z. *et al.* (2011) HAPGEN2: simulation of multiple disease SNPs. *Bioinformatics*, **27**, 2304–2305.

Thompson,E.A. (2005) *MCMC in the analysis of genetic data on pedigrees*. Singapore: World Scientific Co Pte Ltd.

Wijsman,E.M. (2012) The role of large pedigrees in an era of high-throughput sequencing. *Hum. Genet.*, **131**, 1555–1563.

Appendix A

A1. For simulating haplotypes in founders

We calculate the probability of a haplotype pair h_i at founder i given the known information:

$$P(h_i|h_1,...,h_{i-1},\mathbf{x},\mathbf{c}) = \frac{P(h_1,...,h_i,\mathbf{x},\mathbf{c})}{P(h_1,...,h_{i-1},\mathbf{x},\mathbf{c})}$$
$$= \frac{P(h_i,x_i,\mathbf{c})P(h_1,...,h_{i-1},x_1,...,x_{i-1}|h_i,x_i,\mathbf{c})P(x_{i+1},...,x_n|\mathbf{c},h_1,...,h_i,x_1,...,x_i)}{P(x_i,\mathbf{c})P(h_1,...,h_{i-1},x_1,...,x_{i-1}|x_i,\mathbf{c})P(x_{i+1},...,x_n|\mathbf{c},h_1,...,h_{i-1},x_1,...,x_i)} \quad \text{(A1)}$$

Assume no inbreeding and random mating so that all founder haplotypes are independent. Therefore, h_i is independent with $h_{1,...,}h_{i-1}$. Also assume that given a person u's haplotypes h_u, the person's affection status x_u is independent with any other persons' haplotypes such as h_i (Laird and Lange, 2011). Then the second term of the numerator in Eq. (A1) can be written as

$$\begin{aligned} &P(h_1,...,h_{i-1},x_1,...,x_{i-1}|h_i,x_i,\mathbf{c}) \\ &= P(h_1,...,h_{i-1}|h_i,x_i,\mathbf{c})P(x_1,...,x_{i-1}|h_1,...,h_{i-1},h_i,x_i,\mathbf{c}) \\ &= P(h_1,...,h_{i-1}|x_i,\mathbf{c})P(x_1,...,x_{i-1}|h_1,...,h_{i-1},x_i,\mathbf{c}) \\ &= P(h_1,...,h_{i-1},x_1,...,x_{i-1}|x_i,\mathbf{c}) \end{aligned} \quad \text{(A2)}$$

Given (A1) and (A2):

$$P(h_i|h_1,...,h_{i-1},\mathbf{x},\mathbf{c}) = \frac{P(h_i,x_i,\mathbf{c})P(x_{i+1},...,x_n|\mathbf{c},h_1,...,h_i,x_1,...,x_i)}{P(x_i,\mathbf{c})P(x_{i+1},...,x_n|\mathbf{c},h_1,...,h_{i-1},x_1,...,x_i)} \quad \text{(A3)}$$

Equation (A3) is Eq. (2) in the manuscript.

A2. For simulating haplotypes in non-founders

$$P(h_j|\mathbf{h_f},h_m,...,h_{j-1},\mathbf{x},\mathbf{c}) = \frac{P(\mathbf{h_f},h_m,...,h_j,\mathbf{x},\mathbf{c})}{P(\mathbf{h_f},h_m,...,h_{j-1},\mathbf{x},\mathbf{c})}$$
$$= \frac{P(\mathbf{h_f},h_m,...,h_j,\mathbf{c})P(\mathbf{x}|\mathbf{h_f},h_m,...,h_j,\mathbf{c})}{\sum_{k\in\Psi}P(\mathbf{h_f},h_m,...,h_{j-1},h_k,\mathbf{x},\mathbf{c})} \quad \text{(A4)}$$

where Ψ is a set of all possible haplotype pairs of a child whose parental haplotypes are the same as h_j's parental haplotypes.

For the first part of the numerator in Eq. (A4), conditional on individual j's parental haplotypes (i.e. h_p and h_q), individual i's haplotypes are independent from other members' haplotypes. Because $(h_p, h_q) \in \mathbf{h_f}$, the part of equation can be written as

$$\begin{aligned} &P(\mathbf{h_f},h_m,...,h_j,\mathbf{c}) = P(h_j|\mathbf{h_f},h_m,...,h_{j-1},\mathbf{c})P(\mathbf{h_f},h_m,...,h_{j-1},\mathbf{c}) \\ &= P(h_j|h_p,h_q,\mathbf{c})P(\mathbf{h_f},h_m,...,h_{j-1},\mathbf{c}) \end{aligned} \quad \text{(A5)}$$

For the second part of the numerator in Eq. (A4):

$$\begin{aligned} &P(\mathbf{x}|\mathbf{h_f},h_m,...,h_j,\mathbf{c}) \\ &= P(x_1|\mathbf{h_f},h_m,...,h_j,\mathbf{c})P(x_2|\mathbf{h_f},h_m,...,h_j,x_1,\mathbf{c})... \\ &P(x_j|\mathbf{h_f},h_m,...,h_j,x_1,...,x_{j-1},\mathbf{c})P(x_{j+1},...x_n|\mathbf{h_f},h_m,...,h_j,x_1,...,x_j,\mathbf{c}) \end{aligned} \quad \text{(A6)}$$

Again using the assumption that given a person u's haplotypes h_u, the person's affection status x_u is independent with any other

persons' haplotypes such as h_i (Laird and Lange, 2011), each term in Eq. (A6) except the last term can be written as

$$P(x_j|\mathbf{h_f}, h_m, ..., h_j, x_1, ..., x_{j-1}, \mathbf{c})$$
$$= \frac{P(x_1, ..., x_j|h_1, ..., h_j, \mathbf{c})P(h_1, ..., h_j, \mathbf{c})}{P(x_1, ..., x_{j-1}|h_1, ..., h_j, \mathbf{c})P(h_1, ..., h_j, \mathbf{c})} = \frac{\prod_{r=1}^{j} P(x_r|h_r, \mathbf{c})}{\prod_{r=1}^{j-1} P(x_r|h_r, \mathbf{c})} = P(x_j|h_j, \mathbf{c})$$

Then Eq. (A6) can be written as

$$P(\mathbf{x}|\mathbf{h_f}, h_m, ..., h_j, \mathbf{c}) = \prod_{r=1}^{j} P(x_r|h_r, \mathbf{c})P(x_{j+1}, ...x_n|\mathbf{h_f}, h_m, ..., h_j, x_1, ..., x_j, \mathbf{c}) \quad \text{(A7)}$$

Based on Eqs. (A5) and (A7), Eq. (A4) can be written as

$$P(h_j|\mathbf{h_f}, h_m, ..., h_{j-1}, \mathbf{x}, \mathbf{c}) = \frac{P(h_j|h_p, h_q, \mathbf{c})P(x_{j+1}, ...x_n|\mathbf{h_f}, h_m, ..., h_j, x_1, ..., x_j, \mathbf{c})\prod_{r=1}^{j} P(x_r|h_r, \mathbf{c})}{\sum_{k\in\Psi}\left(P(h_k|h_p, h_q, \mathbf{c})P(x_{j+1}, ...x_n|\mathbf{h_f}, h_m, ..., h_{j-1}, h_k, x_1, ..., x_j, \mathbf{c})\prod_{r=1}^{k} P(x_r|h_r, \mathbf{c})\right)}$$

which is Eq. (4) in the manuscript. The probabilities $P(x_{j+1}, ...x_n|\mathbf{h_f}, h_m, ..., h_j, x_1, ..., x_j, \mathbf{c})$can be efficiently calculated in the same manner as the calculation in Eq. (3).