Published in final edited form as:

Genet Epidemiol. 2018 September ; 42(6): 528–538. doi:10.1002/gepi.22130.

Efficient Computation of the Joint Probability of Multiple Inherited Risk Alleles from Pedigree Data

Thomas Madsen^{1,2}, Danielle Braun^{1,2}, Gang Peng³, Giovanni Parmigiani^{1,2}, and Lorenzo Trippa

¹Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute

²Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA

³Department of Biostatistics, Yale School of Public Health, New Haven CT 06520, USA

Abstract

The Elston-Stewart peeling algorithm enables estimation of an individual's probability of harboring germline risk alleles based on pedigree data, and serves as the computational backbone of important genetic counseling tools. However, it remains limited to the analysis of risk alleles at a small number of genetic loci because its computing time grows exponentially with the number of loci considered. We propose a novel, approximate version of this algorithm, dubbed the *peeling and paring* algorithm, which scales polynomially in the number of loci. This allows extending peeling-based models to include many genetic loci. The algorithm creates a trade-off between accuracy and speed, and allows the user to control this trade-off. We provide exact bounds on the approximation error and evaluate it in realistic simulations. Results show that the loss of accuracy due to the approximation is negligible in important applications. This algorithm will improve genetic counseling tools by increasing the number of pathogenic risk alleles that can be addressed. To illustrate we create an extended 5-genes version of BRCAPRO, a widely-used model for estimating the carrier probabilities of *BRCA1* and *BRCA2* risk alleles and assess its computational properties.

Keywords

Family History; Germline Risk Alleles; Mendelian Risk Prediction; Peeling Algorithm

1 Introduction

Until recently, patients referred to genetic testing to assess cancer susceptibility were tested only for few well-known and well-studied susceptibility genes. The rapid drop in the cost of DNA sequencing has led to the availability of multi-genes cancer panels, which currently test 25 to 123 cancer susceptibility genes at a relatively low cost (Plichta et al., 2016). This is challenging traditional computational approaches used to support decision making about genetic testing. In this paper we tackle this challenge.

To illustrate, women with a family history of breast or ovarian cancer were until recently tested only for deleterious germline variants in the *BRCA1* and *BRCA2* genes, which are associated with an increased risk of several diseases (Lynch et al., 2014). These hereditary genetic variants are rare in the general population, but have considerably higher prevalence in individuals with a family history of breast and ovarian cancers (Ellisen & Haber, 1998). More recently, deleterious variants of *ATM*, *CHEK2*, *PALB2* and several other genes have also been associated with an increased risk of several cancer types including breast cancer (Marabelli et al., 2016; Schmidt et al., 2016; A. C. Antoniou et al., 2014). Before genetic tests are ordered, it is often of interest to support the testing decision by assessing pre-test estimates of the probability of carrying a deleterious variant (Halbert, 2004).

The importance of accurate estimates of carrier probabilities is well established (Nelson et al., 2014). Mendelian models, such as BRCAPRO (Berry et al., 1997; Mazzola et al., 2015) or BOADICEA (A. C. Antoniou et al., 2008), estimate pre-test carrier probabilities of *BRCA1* and *BRCA2* germline deleterious variants based on family history of relevant cancers, as well as other risk factors. These models are grounded in principles of Mendelian genetics, Bayesian probability theory, and knowledge of deleterious variants, including genotypes' prevalences and genotype-specific penetrance (age dependent probability of developing cancer conditional on genotype), both of which are estimated in the general population. Mendelian modeling can be applied to estimate carrier probabilities for risk alleles associated with any hereditary syndrome. These models are included in widely used clinical software packages (Euhus, 2001) and recommended by professional societies (Riley et al., 2012).

Most existing Mendelian risk prediction models focus on either one cancer, or a single syndrome. Recently, strong evidence is emerging that syndromes once thought to be distinct are determined by genetic variants that increase the risk of multiple cancers (Kastrinos et al., 2009; Kastrinos & Syngal, 2011; Hruban et al., 2010; Moran et al., 2012; Kote-Jarai et al., 2011). Many of these associations are still unknown and a recent study by Schrader et al. (2015) found that only 81 of 198 participants with presumed pathogenic germline variants in cancer-susceptibility genes had diagnoses considered to be associated with the mutated gene. For example, only 28 out of 52 mutations in *BRCA1* and *BRCA2* were observed in participants with breast, ovary, prostate and pancreas cancer, while the remaining 24 had diagnoses of other cancer types. This indicates important opportunities for better probabilistic modeling of cancer pedigrees, and in turn refinements in risk stratification, screening and management of risk. For example, ovarian cancer patients who until recently were candidates for only *BRCA* testing, may instead carry risk alleles in mismatch repair (*MMR*) genes, which might impact genetic testing decisions and colonoscopy frequency for family members.

The input to a Mendelian model is a detailed pedigree, including age, gender, ethnicity, race, cancer diagnosis status, age at diagnosis, and risk factors. Also included are the exact relations between the individuals, typically captured by linking each individual ID with his or her maternal and paternal ID, as well as the risk allele prevalences and penetrances. The output is the proband's (individual who comes into counseling) posterior genotypic distribution, i.e. the probability of each genotype, based on the information in the pedigree.

The genotypic distribution assigns probabilities to each genotype considered by a Mendelian model. For example, BRCAPRO provides probabilities, conditional on pedigree information, for four possible genotypes of the proband: no deleterious variants, deleterious variant only at *BRCA1*, deleterious variant only at *BRCA2*, or deleterious variants at both *BRCA1* and *BRCA2*. In this model a binary variable per gene captures, at the individual level, deleterious variants for *BRCA1* and *BRCA2*. Similarly, in the remainder of this article we will consider models with more than 2 genes, and one binary variable per gene to indicate individual deleterious variants.

In Mendelian models, pedigree information is translated into individual genotype probabilities via three key components: first, specification of an inheritance mechanism for risk alleles; second, specification of the prevalence and penetrance parameters; and third, application of Bayes' rule (E. Murphy & Mutalik, 1969). Bayesian inference permits calculation, conditional on the family history, of genotype probabilities for the proband.

Calculation of the posterior genotypic distribution directly via the law of total probability requires the computationally intensive consideration of every possible configuration of genotypes in the pedigree. We use the term *genotype configuration* to refer to the collection of the genotypes of all members of the pedigree. Figure 1 (left panel) illustrates a single configuration. Consider N family members and K genes. In our analyses we make the simplifying assumption that each gene has two possible states, it either harbors a pathogenic variant or not. Then the number of configurations is 2^{NK} . In other contexts, or more complex models, the number of possible states per gene may need to be larger. For example when homozygous carriers of deleterious mutations are viable, one can define three states corresponding to homozygous and heterozygous deleterious variants, and wild type. Note that, if we add this third state to each gene, then the number of configurations increases to 3^{NK} . The Elston-Stewart peeling algorithm (Elston & Stewart, 1971) facilitates burdensome integration over these large sets of configurations with a series of recursive computations. The algorithm enables integration over all possible genotypic configurations of the pedigree consistent with Mendelian inheritance, to compute the proband posterior genotypic distribution. This computational procedure exploits the independence of siblings' genotypes conditional on their parents' genotypes. It is similar to Bayesian computation methods for graphs of random variables, such as message-passing algorithms (K. P. Murphy et al., 1999). Within pedigrees each node is a random variable representative of an individual's genotype, and the graph encodes parent-sibling relationships.

Nonetheless, the time complexity of the peeling algorithm is polynomial in the number of possible genotypes, and exponential in the number of genes with risk alleles. Even for relatively small numbers of genes (fewer than ten) and small pedigrees (fewer than twenty members), the computation is far too time-consuming for a clinical counseling setting, where speed is paramount. Additionally, the time requirements of the peeling algorithm grow linearly in the number of individuals in the pedigree, and do not depend on the number of outcomes (e.g. cancer types) or the complexity of the penetrance model.

In the context of breast and ovarian cancer risk prediction, to guide genotype testing and to improve risk management and prevention, increasing evidence is available that risk alleles in

many other genes, in addition to *BRCA1* and *BRCA2*, contribute significantly to both breast and ovarian cancer risk. For instance, in breast cancer, risk alleles in *PALB1*, *CHEK2*, *ATM*, *TP53*, and *PTEN*, as well as dozens of single-nucleotide polymorphisms, are all known to be risk-heightening, though many with considerably lower penetrance than *BRCA1/2* risk alleles (Kurian et al., 2017).

The purpose of this work is to introduce a new algorithm, the *peeling and paring* algorithm, for approximate computations of the posterior genotypic distribution in Mendelian models with a large number of genes with risk heightening variants. This algorithm can replace the standard peeling algorithm in Mendelian models, allowing them to include additional genetic components. Although our illustration of this algorithm will be in breast cancer, the algorithm can be applied to any Mendelian model.

2 Methods

2.1 Notation and Overview of Mendelian Models

This section outlines the general formalism of Mendelian risk prediction models. Let N be the family size, index i identifies relatives in the family, and let U_i , $i = 0, \dots, N$, indicate the corresponding genotypes. The proband is indexed by $i = 0$. In what follows, N and the family relations, for example parent-sibling relations, are fixed, while genotypes and phenotypes (disease status; in our setting cancer diagnoses) will be part of a joint probability model. Phenotypes are denoted by $\mathbf{Y} = (\mathbf{Y}_0, \dots, \mathbf{Y}_N)$. In the simplest case, we consider a single cancer type; $\mathbf{Y}_i = (c_i, a_i)$, where a_i denotes a possibly censored age of diagnosis and c_i denotes the censoring variable indicating if cancer occurred or not. Age of diagnosis can be censored by death or current age of individuals at time of interview. Models such as BRCAPRO and MMRPRO also include more specific information, including preventive surgeries and molecular characteristics.

The unknown vector U_0 represents the genotype of the proband, with each entry corresponding to one of the genes considered. For example, the current version of MMRPRO considers pathogenic variants in the genes *MLH1*, *MSH2*, and *MSH6*, and consequently U_0 has three binary entries in this case. We will denote the number of genes in the model by K . Let $\mathcal{U}$ represent the set of all possible genotypes. This has cardinality g^K , where g is the number of possible states for a given locus. In our illustrations $g = 2$. In genetic counseling, we are interested in estimating the conditional probabilities $P(U_0 = u / Y_0, Y_1, \dots, Y_N)$ for all genotypes $u \in \mathcal{U}$.

Following E. Murphy & Mutalik (1969) and Lange (2003) we can apply Bayes' rule to find the desired posterior distribution as follows:

$$P(U_0 = u | Y_0 = y_0, \dots, Y_N = y_N) \propto P(U_0 = u)P(Y_0 = y_0, \dots, Y_N = y_N | U_0 = u).$$

The term $P(Y_0, Y_1, \dots, Y_N | U_0 = u)$ is the conditional probability of the observed phenotypes for the whole pedigree given the genotype u of the proband. To compute the

posterior distribution starting from a penetrance model at the individual level, it is necessary to integrate over the genotype configurations $(U_1, \dots, U_N)$ through

$$P(Y_0 = y_0, \dots, Y_N = y_N | U_0 = u) = \sum_{u_1, \dots, u_N \in \mathcal{U}} P(Y_0 = y_0, \dots, Y_N = y_N | U_0 = u, U_1 = u_1, \dots, U_N = u_N) P(U_1 = u_1, \dots, U_N = u_N | U_0 = u).$$

We start by modeling $P(U_0, U_1, \dots, U_N)$ and $P(U_1, \dots, U_N | U_0)$, the probability distribution of all family members' genotypes conditional on the proband's genotype, in the absence of any information about phenotypes. These quantities can be derived using properties of Mendelian inheritance and knowledge of the prevalence of risk alleles at each genetic locus. For a trio, with $K = 1$, it is sufficient to specify the marginal probability of maternal and paternal genotypes, which are assumed *a priori* independent and identically distributed, and the conditional distribution of the sibling genotype when the risk allele can be inherited from one or both parents. Marginal and conditional probabilities allow one to straightforwardly extend the model from trios to larger pedigrees ($N \geq 3$). The extension is defined by the assumption of genotype independence across first-generation family members, followed by conditionally independent genotypes for second-generation family members, with conditional distributions identical as in the trio case, which in turn are followed by conditionally independent genotypes for the subsequent generation (Berry et al., 1997). In Bayesian terms, this modular mechanism forms our prior model $P(U_0, U_1, \dots, U_N)$. With $K > 1$ we assume that risk alleles at different loci are independent.

Next, we model $P(Y_0, \dots, Y_N | U_0, \dots, U_N)$ by assuming that individual's phenotypes are conditionally independent. That is, conditional on U_i , Y_i is independent of other genotypes U_j and phenotypes Y_j , $j \neq i$. Under this assumption,

$$P(Y_0, \dots, Y_N | U_0, \dots, U_N) = \prod_{i=1}^N P(Y_i | U_i).$$

$P(Y_i | U_i)$ is the probability of the phenotype given the genotype, also referred to as the penetrance, which can be found in the epidemiological literature. For example in BRCAPRO $P(Y_i | U_i)$ coincides with penetrance functions for breast and ovarian cancers for carriers of *BRCA1* and *BRCA2* risk alleles, as well as for non-carriers, which have been estimated in several studies (Chen & Parmigiani, 2007). The summation in the above equation is over all possible genotypic configurations of the pedigree; the cardinality of this set grows exponentially both with the size of the pedigree and the number of genes considered.

The peeling algorithm reduces the computational burden of this integration by making its complexity and computational time linear with respect to the size of the pedigree N . In the peeling algorithm, computing time still grows exponentially with the number of genes K . Currently peeling computations for genetic counseling become unfeasible with a modest number of genes, say 10. We briefly review the basic principle of the peeling algorithm and subsequently propose an effective approximation.

2.2 Peeling and Paring

The peeling algorithm applies iterative local integration to compute the exact genotype posterior probabilities. The basic principle of local integration is similar to other successful computational methods including message-passing procedures (Pearl, 1982) and the Viterbi algorithm (Forney, 1973). These procedures are used to compute posterior estimates of latent variables, for example in Hidden Markov models, when their dependence and conditional independence relations can be described by graphs. Similarly, the peeling algorithm formalizes the fact that to compute the posterior probability of an individual's genotype U_i , $P(U_i/Y_0 = y_0, \dots, Y_N = y_N)$, we do not need the genotype conditional distribution of the entire family, but rather only need the genotype conditional distribution of first degree relatives (local integrations), penetrance estimates $P(Y_i/U_i)$, and elementary transformations of these quantities. Since the genotype conditional distribution of first degree relatives are not available from the start, iterative computations are used to progressively go from preliminary estimates or partial knowledge of $P(U_0, \dots, U_N/Y_0, \dots, Y_N)$ to the exact posterior distribution.

The peeling algorithm is the standard approach for conducting posterior computations in Mendelian models. Nonetheless it is only applicable under settings in which the model contains only a modest number of genes, say less than $K = 10$, which are jointly modeled with the phenotypes. The algorithm requires consideration of all 2^K genotypic configurations of all individuals. Recall that we consider only two states per gene: wild type and risk-increasing variants. The cardinality of genotypes rapidly increases with K , and computing times increase exponentially. The same holds with more than two possible states per gene. Therefore, computations become unfeasible with a large number of genes. We propose a simple and effective approximation procedure, to make the algorithm applicable in such settings.

Our proposed procedure, peeling and paring, restricts the set of possible genotypes under consideration for each individual. In particular, the key idea is to disregard all genotypes which contain more than T risk alleles, and apply local integration, as in peeling, on the restricted set of genotypes $\mathcal{U}_T$. For this reason, we dub our algorithm the 'peeling and paring' algorithm, since it performs peeling on a pared set of genotypes. T is an integer approximation parameter chosen by the user. The cardinality of $\mathcal{U}_T$ when $T < K$ can be considerably lower than 2^K , reducing the number of possible individual genotypes, and substantially mitigating the computational bottleneck. When $T = K$, the peeling and paring algorithm coincides with the original peeling algorithm.

After the space of genotypes $\mathcal{U}$ is restricted to $\mathcal{U}_T$ the *a priori* probabilities of its components are renormalized. The Mendelian model remains identical in every other aspect. Normalization is used to specify a probability model, our *a priori* approximation, in which we enforce all genotypes U_i to be included in a restricted set $\mathcal{U}_T$.

The resulting integration step has the following form:

$$P(Y_0, Y_1, \dots, Y_N | U_0 = u) = \sum_{u_1, \dots, u_N \in \mathcal{U}_T} [P(Y_0, Y_1, \dots, Y_N | U_0 = u, U_1 = u_1, \dots, U_N = u_N) \times P(U_1 = u_1, \dots, U_N = u_N | U_0 = u)]$$

where $\mathcal{U}_T$ is the set of all genotypes with T or fewer risk alleles, i.e. binary vectors with K entries in which the digit 1, indicating pathogenic variants, appears at most T times. In our approximation, normalization guarantees that

$$\sum_{u_1, \dots, u_N \in \mathcal{U}_T} P(U_1 = u_1, \dots, U_N = u_N | U_0 = u) = 1.$$

The output of the exact peeling algorithm and our approximation have nearly identical interpretations. They are both exact posterior distributions from two Bayesian models. The only difference between the two models is that in our approximation we assume that the probabilities of genotypes U_i harboring more than T risk alleles are equal to zero.

Figure 1 depicts the posterior genotypic distributions (right panel) for the example pedigree in the left panel with various levels of T . As expected, all three posteriors assign similar probabilities to the most likely genotypes. Note that no probability is assigned to genotypes containing two or more risk alleles when $T = 1$, and no probability is assigned to genotypes containing three or more risk alleles when $T = 2$. The 'missing' probability in each of these distributions is reallocated to the genotypes with fewer risk alleles. All three posterior distributions sum to one, though this is difficult to perceive due to the logarithmic scale of the y-axis.

To illustrate the gain in computational efficiency provided by our method, we consider a simple example. Using the notation above, the number of possible genotypes is 2^K . If $K = 25$, a realistic number for current genetic testing panels, then the number of possible genotypes is $2^{25} = 33,554,432$. However, if we restrict the set of possible genotypes to those containing $T = 2$ or fewer risk alleles, then the number of possible genotypes is reduced to 326. Since the runtime of the peeling algorithm depends roughly linearly on this quantity, the approximation reduces the required number of computations by approximately fifteen orders of magnitude. Figure 2 illustrates the relationship between the cardinality of the set of individual genotypes and the number K of genetic alterations considered, with different constraints $T < K$ and without any constraint ($T = K$). One goal of our work has been to identify sensible values of T , which provide sufficiently precise approximations of the posterior genotypic distributions in practice while remaining computationally feasible.

Our approximation approach, based on restrictions of the genotypes' set $\mathcal{U}$, can be easily extended to models where the assumption of two possible states per gene or locus is relaxed, for example including wild type, heterozygous and homozygous pathogenic variants. In the next section we illustrate results under the assumption of two states per genes, as in several existing Mendelian models including BRCAPro.

The improvement in speed achieved by our algorithm comes at a cost because some genotype configurations are not being considered. The algorithm provides an approximation to the posterior genotypic distribution. Selecting a value for T involves a trade-off between precision and efficiency; higher values of T will tend to produce a more precise approximation of the posterior genotypic distribution, but at the price of an increased computational burden. The goal of the following section is to quantify under realistic scenarios this loss of accuracy.

3 Simulation Study

3.1 Simulation Scenarios

We use simulations to evaluate the peeling and paring algorithm, and how it approximates the exact posterior genotypic distribution. To this end we developed a generalization of the BRCAPRO model which incorporates five genes associated with breast cancer: *BRCA1*, *BRCA2*, *ATM*, *CHEK2*, and *PALB2*. We used estimates from the literature for the model parameters. The model has not been empirically validated yet, and in what follows, we use it as an illustration. We use simulated pedigrees to evaluate peeling and paring approximations and compare them to the exact ($T = K = 5$) posterior genotype distributions, based on this realistic model.

We generated pedigrees with the same structure as the one in Figure 1. More specifically, each pedigree has a single set of grandparents, their two offspring and spouses, and four grandchildren. One of the female grandchildren was chosen as the proband. We first simulate genotypes assuming independent Mendelian inheritance of the five genes. We then simulate phenotypes conditional on the individual genomic profile; $P(Y_i|U_i)$. Prevalence and age-dependent penetrance estimates for *BRCA1* and *BRCA2* carriers and non-carriers came from the BayesMendel R package (Chen & Parmigiani, 2007; Chen et al., 2004). Penetrance estimates for the other three genes were obtained from a comprehensive literature review as part of a clinical decision support tool, All Syndromes Known to Man Evaluator, or ASK2ME (Braun et al., 2018), which selected a study for each of these three genes (Marabelli et al., 2016; Schmidt et al., 2016; A. C. Antoniou et al., 2014). Penetrance and prevalence estimates for the five genes are shown in Figure 3. We assumed no interaction effects between the risk alleles; penetrance functions for genotype with deleterious variants at multiple genes were calculated assuming additivity of the risk increase associated to the deleterious status of each gene. For simplicity, we did not consider risk modifiers such as race and preventive surgeries.

3.2 Accuracy Metric

To evaluate our approximation, we computed for each simulated family the total variation, contrasting the exact and approximate genotype posterior distribution:

$$\text{Total Variation}(p, p^T) = \sum_{u \in \mathcal{U}} |p(u|\text{data}) - p^T(u|\text{data})|.$$

Here $\mathcal{U}$ is the set of all possible genotypes (without restrictions), u is an individual genotype, $p(u|\text{data})$ is the exact posterior probability of u obtained with the standard peeling algorithm, and $p^T(u|\text{data})$ is the approximate posterior probability of u obtained with the peeling and paring algorithm. Higher values of the approximation parameter T are expected to lead to smaller total variation discrepancies.

We chose to simulate from a probability model with only five genes so that exact posterior computation remains feasible and provides a benchmark of comparison to evaluate the approximate posterior distribution. For each simulated family we calculated the approximate posterior distribution at various levels of T as well as the exact posterior distribution.

3.3 Simulation Results

We generated 1000 simulated families and computed the total variation for each proband under each level of the approximation parameter $T = 1, 2, 3$. Figure 4a shows that the precision of the approximation increases by orders of magnitude with each increment of T . Even small values of $T = 1, 2$, provide approximations to the posterior genotypic distribution that we think could be potentially valuable in clinical applications, with total variation summaries below 1% (dotted line in Figure 4a). This suggests that it is possible to reduce the number of genotypes considered by the peeling algorithm while still retaining accuracy and the resulting estimates are consistent with the exact posterior distributions. In our simulations, $T = 2$ guarantees total variation discrepancies that would alter only minimally individual risk management and prevention recommendations.

To investigate whether the accuracy of the approximation depends on the size of the pedigree, we also generated 1000 larger pedigrees (22 members) accordingly to the same breast cancer model. Figure 4b shows the total variation when using the approximation level $T = 2$ and the pedigree size is equal either to 7 or 22. This panel shows that the accuracy of the approximation remains similar when the pedigree size changes.

3.4 Larger Number of Risk Alleles

Next we explored peeling and pairing in scenarios with relevant risk alleles in more than 20 genes, where the number of possible individual genotypes is above 10^6 . We combine the five breast cancer genes previously described with risk-increasing variants for 20 additional genes. We assume that risk alleles in each of these additional loci have a prevalence of 1% and increase the risk of breast cancer by 10% compared to non-carriers. These values are in line with many well-known breast cancer-associated SNPs. We generated 10,000 pedigrees, each with the same structure as in Figure 1, using this 25-genes model, and another 10,000 pedigrees using the previous 5-genes model.

In this case, exact posterior computation becomes unfeasible. We focus on discrimination, and evaluate the peeling and paring algorithm's ability to discriminate between carriers and non-carriers by calculating the area under the receiver operating characteristic curve (AUC). In addition to peeling and paring, we also estimated the posterior genotypic distribution using a Gibbs sampler. Gibbs sampling procedures are a popular class of Monte Carlo Markov Chain algorithms for sampling from multivariate distributions. In our application the

computational times of the Gibbs sampler are larger than for peeling and paring. The accuracy of the Gibbs algorithm improves as it runs longer. The Gibbs sampler estimates converges to the exact posterior as the Markov Chain length goes to infinity. For meaningful comparisons, we evaluated the AUC with the Gibbs sampler truncated at one and ten minutes of computations (using R on a 2.2 GHz Intel Core machine). All the peeling-and-paring calculations that we performed took less than thirty seconds. We compared the Gibbs sampler to the peeling and paring algorithm at various values of the approximation parameter T . We evaluate these two algorithms based on their ability to discriminate i) whether the proband carries a risk allele in *BRCA1*, and ii) whether the proband carries a risk allele in any gene included in the model. The results of these simulations are summarized in Figure 5.

In the 5-genes model, the peeling and paring algorithm only suffers a negligible loss in discrimination, even for low levels of T , compared to exact posterior computations. In fact, in our simulated population of 10,000 pedigrees, when we evaluate discrimination of probands with and without deleterious variants, the empirical AUC for the peeling and paring algorithm with $T = 1$ (0.726) is nearly identical to the AUC for $T = 2$ (0.723) or $T = 5$ (exact peeling) (0.723). In this simulation study peeling and paring outperformed the MCMC approach, even when the Gibbs sampler runs for ten minutes per pedigree (Figure 5). AUCs values across algorithms tend to be higher when we focus on discrimination of probands with and without *BRCA1* deleterious variants, compared to carriers versus non carriers of any deleterious variant, plausibly because *BRCA1* is more penetrant than the other genes in the simulation model.

In the 25-genes model, it is prohibitive to compute the exact posterior distributions. Also in these simulations, when we evaluate discrimination of probands with and without deleterious variants, the peeling and paring algorithm even with $T = 1$ outperforms the MCMC algorithm: the AUC with $T = 1$ is 0.717, while for the MCMC algorithm with 10 minutes computing time it is 0.608. As for the 5-genes simulation model, AUCs to summarize discrimination of probands with and without any deleterious variant, when compared to discrimination of probands with and without *BRCA1* deleterious variants, tend to be lower. Overall, the peeling and paring in our simulations outperforms the MCMC approach in terms of discrimination accuracy.

4 Exact Error Bounds

One advantage of our approach compared to other posterior approximation methods such as Markov Chain Monte Carlo or variational inference, is that we can bound the error of our approximations. For subsets of genotypes we can compute upper and lower bounds of the exact posterior probability, together with the approximated posterior probability ($T < K$). This bounding is useful because it can inform the choice of the approximation parameter T . For instance, we can increase T until a desired level of precision is achieved at the individual level. The derivation of these bounds can be found in the Appendix.

To demonstrate the utility of error bounds, we consider an example with $K = 2$. Each risk allele has prevalence equal to 0.02. Consider a family with three members where parents are

indexed by 1 and 2 and the offspring is indexed by 0. Let 00 denote the wild-type genotype, 01 denote wild-type variant for gene 1 and a risk allele at gene 2, and so forth. Furthermore, assume the following risk function with binary phenotype: $P(y_i = 1 | u_i = 00) = 0.1$, $P(y_i = 1 | u_i = 01) = 0.2$, $P(y_i = 1 | u_i = 10) = 0.3$, $P(y_i = 1 | u_i = 11) = 0.4$, with identical probabilities across genders. We restrict $\mathcal{U}$ to genotypes with deleterious variants only for one or none of the genes ($T = 1$). Table I shows the bounds on genotype probabilities, for each of the six possible phenotype configurations.

We illustrate how bounds can facilitate testing decisions in this example. Assume first that the threshold to recommend genomic testing for the proband is a 5% probability that he/she is a carrier of a risk allele. Then approximate computations ($T = 1$) together with the upper and lower bounds (first and last row of Table I) are sufficient, because intervals from lower bounds to upper bounds never include the threshold 0.05. Importantly the computation of the bounds is associated to a minimal increase of the ($T = 1$) computational burden. Here the bounds indicate that the decision between genetic testing and not remains the same whether approximate ($T = 1$) or exact ($T = 2$) Mendelian computations are used. We can now increase the threshold to 8%, to illustrate that the bounds in some cases indicate the need of more precise computations. Indeed, when both parents present the phenotype of interest (4th column of Table I), the upper and lower bounds suggest the use of exact computations ($T = 2$). The interval between the upper and lower bounds include 8%; therefore testing decisions with approximate ($T = 1$) and exact ($T = 2$) computations could differ.

5 Discussion

We have presented a generalization of the peeling algorithm that allows for efficient and precise approximation of the posterior genotypic distribution in the multi-gene Mendelian models. The method can be adjusted to a desired level of accuracy by the choice of the hyper-parameter T . We showed that even small values of T can provide accurate approximations to the exact posterior genotypic distribution, with essentially no loss in discrimination. Our method will allow for the development of Mendelian risk prediction models accounting for potential deleterious variants in a larger number of genes compared to models currently in use. This has the potential to better capture the underlying biology of disease than present models. Furthermore, we presented a method for computing exact bounds for the errors of the proposed approximation, a feature which enables the identification of sensible hyper-parameter choices that strike a balance between precision and efficiency.

Since many susceptibility risk alleles are risk-heightening for multiple cancer types, multi-loci models will be able to leverage more family history information by including more cancer types. Furthermore, this innovation allows for Mendelian risk prediction models to be developed for diseases whose heritable component is thinly spread across many different loci. This aspect is relevant in polygenic models such as BOADICEA (A. Antoniou et al., 2004). Moreover, Mendelian models such as the one used for illustration in this work have the advantage of allowing one to borrow information from multiple cancer types.

The proposed approximation procedure is motivated by the context of low-prevalence, high-penetrance variants. Current genetic testing panels for inherited cancer susceptibility (Plichta et al., 2016) can include over one hundred genes, but it is reasonable to assume that the probability that an individual—even in a high risk population—is a carrier of more than a modest number of deleterious variants, say $T=5$, is negligible. In contrast, for diseases caused by a large number of more common, low-penetrance variants, where each individual is likely to be carrier of an extensive set of risk-associated variants, our approximation may not be ideal. Alternative approximation methods – perhaps based on central limit theorem arguments – remain necessary in this context.

Our illustration included an extension of the BRCAPRO model to evaluate the performance of the proposed peeling-paring algorithm. It is important to note that this model has not yet been validated.

Our simulations did not take into account the possibility of missing data, which is likely to be present in clinical settings. Recent articles (Biswas et al., 2013, 2016) have explored the use of Mendelian models with incomplete data (for example missing information about a family member, such as age or cancer diagnosis), or when only summaries, such as the number of first degree relatives with cancer diagnoses, are available. These studies suggest that Mendelian models can be adapted to incomplete pedigrees and that the models' performances (in terms of AUC) are robust to modest missingness. The methods identified in these articles, including data imputation, could be extended to to relatively complex Mendelian models, as those considered in our work.

An important difference between our approach and algorithms such as Lander and Green (1987), is the assumption of independent inheritance across distinct loci. This assumption is appropriate for Mendelian models of commonly tested cancer gene mutations, though it would not be tenable in a linkage analysis, the primary focus in Lander and Green. In our work we did not evaluate peeling and paring beyond this assumption, nor did we investigate connections with models and algorithms that account for linkage disequilibrium.

The application of the peeling and paring algorithm is not limited to cancer syndromes; several complex diseases have a heritable risk component that involves multiple genetic loci, including diabetes (Prasad & Groop, 2015), coronary artery disease (Roberts, 2015), Alzheimer's disease (Lambert et al., 2013), multiple sclerosis (Bashinskaya et al., 2015) and chronic obstructive pulmonary disease (Cho et al., 2014). The method that we discussed could support the use of Mendelian models for these diseases. The algorithm enables the development and use of more powerful risk prediction models, which leverage a proband's entire family history and include multiple disease types and susceptibility loci, to estimate carrier probabilities and future risks.

Acknowledgments

We gratefully acknowledge support from the National Cancer Institute at the National Institutes of Health [4P30CA006516-51] which supported Dr. Parmigiani.

Appendix

Derivation of Error Bounds

We derive a set of exact error bounds. Initially consider a pedigree with three individuals: two parents and an offspring, the proband. We will index the parents by 1 and 2 and the offspring by 0. Let U_i represent the latent genotype for the i th individual and y_i be the phenotype. Let $\mathcal{U}$ be the set of all possible genotypes, and let $\mathcal{U}^*$ be the restricted set of genotypes.

We first consider the distribution of U_0 , conditional on y_1 and y_2 . We can define upper P^u and lower P^l bounds for $u \in \mathcal{U}^*$

$$P(U_i = u_i | Y_i = y_i) \leq P^u(U_i = u_i | Y_i = y_i) = \frac{P(U_i = u)P(y_i | u)}{\sum_{u \in \mathcal{U}^*} P(u)P(y_i | u)}$$

and

$$P(U_i = u_i | Y_i = y_i) \geq P^l(U_i = u_i | Y_i = y_i) = \frac{P(U_i = u)P(y_i | u)}{\max_u (P(y_i | u)) \left(1 - \sum_{v \in \mathcal{U}^*} P(v)\right) + \sum_{u \in \mathcal{U}^*} P(u)P(y_i | v)}$$

for $i = 1, 2$. Computation of these bounds is linear with the cardinality of $\mathcal{U}^*$. We can proceed to the subsequent generation and, using the above expressions we can bound

$$P(U_0 = u_0 | Y_1 = y_1, Y_2 = y_2) \leq P^u(U_0 = u_0 | Y_1 = y_1, Y_2 = y_2)$$

and

$$P(U_0 = u_0 | Y_1 = y_1, Y_2 = y_2) \geq P^l(U_0 = u_0 | Y_1 = y_1, Y_2 = y_2)$$

for $u_0 \in \mathcal{U}^*$. For example,

$$P^l(U_0 = u_0 | Y_1 = y_1, Y_2 = y_2) = \sum_{u_1, u_2} P(U_0 = u_0 | U_1 = u_1, U_2 = u_2) \prod_{i=1}^2 P^l(U_i = u_i | Y_i = y_i).$$

The computational time in this case is polynomial with respect to the cardinality of $\mathcal{U}^*$.

These last equations allow one to directly obtain the final bounds of interest

$$P(U_0 = u_0 | Y_0 = y_0, Y_1 = y_1, Y_2 = y_2) \leq P^u(U_0 = u_0 | Y_0 = y_0, Y_1 = y_1, Y_2 = y_2)$$

and

$$P(U_0 = u_0 | Y_0 = y_0, Y_1 = y_1, Y_2 = y_2) \geq P^{\ell}(U_0 = u_0 | Y_0 = y_0, Y_1 = y_1, Y_2 = y_2).$$

For example,

$$P^u(U_0 = u_0 | Y_0 = y_0, Y_1 = y_1, Y_2 = y_2) = \frac{P^u(U_0 = u_0 | Y_1 = y_1, Y_2 = y_2)P(Y_0 = y_0 | U_0 = u_0)}{\sum_u P^{\ell}(U = u | Y_1 = y_1, Y_2 = y_2)P(Y_0 = y_0 | U = u)}.$$

For larger pedigrees we can iterate the use of the illustrated arguments and obtain bounds

$$P^u(U_0 = u_0 | Y = y),$$

where Y denotes the phenotypes in the pedigree, and

$$P^{\ell}(U_0 = u_0 | Y = y).$$

References

- Antoniou A, Pharoah P, Smith P, Easton D. The boadicea model of genetic susceptibility to breast and ovarian cancer. *British journal of cancer*. 2004; 91(8):1580–1590. [PubMed: 15381934]
- Antoniou AC, Casadei S, Heikkinen T, Barrowdale D, Pylkäs K, Roberts J, et al. Breast-cancer risk in families with mutations in *palb2*. *New England Journal of Medicine*. 2014; 371(6):497–506. [PubMed: 25099575]
- Antoniou AC, Cunningham AP, Peto J, Evans DG, Lalloo F, Narod SA, Easton DF. The BOADICEA model of genetic susceptibility to breast and ovarian cancers: updates and extensions. *British Journal of Cancer*. 2008 Apr; 98(8):1457–1466. [PubMed: 18349832]
- Bashinskaya V, Kulakova O, Boyko A, Favorov A, Favorova O. A review of genome-wide association studies for multiple sclerosis: classical and hypothesis-driven approaches. *Human genetics*. 2015; 134(11–12):1143–1162. [PubMed: 26407970]
- Berry D, Parmigiani G, Sanchez J, Schildkraut J, Winer E. Probability of carrying a mutation of breast-ovarian cancer gene *brca1* based on family history. *Journal of the National Cancer Institute*. 1997; 89(3):227–237. [PubMed: 9017003]
- Biswas S, Atienza P, Chipman J, Blackford AL, Arun B, Hughes K, Parmigiani G. A two-stage approach to genetic risk assessment in primary care. *Breast cancer research and treatment*. 2016; 155(2):375–383. [PubMed: 26786860]
- Biswas S, Atienza P, Chipman J, Hughes K, Barrera AMG, Amos CI, Parmigiani G. Simplifying clinical use of the genetic risk prediction model *brcapro*. *Breast cancer research and treatment*. 2013; 139(2):571–579. [PubMed: 23690142]
- Braun D, Yang J, Griffin M, Parmigiani G, Hughes KS. A Clinical Decision Support Tool to Predict Cancer Risk for Commonly Tested Cancer-Related Germline Mutations. *Journal of Genetic Counseling*. 2018 Mar; 371(6):1–13. Retrieved from . DOI: 10.1007/s10897-018-0238-4
- Chen S, Parmigiani G. Meta-analysis of *BRCA1* and *BRCA2* penetrance. *Journal of clinical oncology*. 2007; 25(11):1329–1333. [PubMed: 17416853]
- Chen S, Wang W, Broman K, Katki HA, Parmigiani G. Bayesmendel: an r environment for mendelian risk prediction. 2004
- Cho MH, McDonald MLN, Zhou X, Mattheisen M, Castaldi PJ, Hersh CP, et al. Risk loci for chronic obstructive pulmonary disease: a genome-wide association study and meta-analysis. *The lancet Respiratory medicine*. 2014; 2(3):214–225. [PubMed: 24621683]

- Ellisen LW, Haber DA. Hereditary breast cancer. *Annual review of medicine*. 1998; 49(1):425–436.
- Elston RC, Stewart J. A general model for the genetic analysis of pedigree data. *Human heredity*. 1971; 21(6):523–542. [PubMed: 5149961]
- Euhus DM. Understanding mathematical models for breast cancer risk assessment and counseling. *The breast journal*. 2001; 7(4):224–232. [PubMed: 11678799]
- Forney GD. The viterbi algorithm. *Proceedings of the IEEE*. 1973; 61(3):268–278.
- Halbert C. Decisions and outcomes of genetic testing for inherited breast cancer risk. *Annals of oncology*. 2004; 15(suppl 1):i35–i39. [PubMed: 15280185]
- Hruban RH, Canto MI, Goggins M, Schulick R, Klein AP. Update on familial pancreatic cancer. *Advances in surgery*. 2010; 44(1):293–311. [PubMed: 20919528]
- Kastrinos F, Mukherjee B, Tayob N, Wang F, Sparr J, Raymond VM, Syngal S. Risk of pancreatic cancer in families with lynch syndrome. *Jama*. 2009; 302(16):1790–1795. [PubMed: 19861671]
- Kastrinos F, Syngal S. Inherited colorectal cancer syndromes. *Cancer journal (Sudbury, Mass)*. 2011; 17(6):405.
- Kote-Jarai Z, Leongamornlert D, Saunders E, Tymrakiewicz M, Castro E, Mahmud N, et al. Brca2 is a moderate penetrance gene contributing to young-onset prostate cancer: implications for genetic testing in prostate cancer patients. *British journal of cancer*. 2011; 105(8):1230–1234. [PubMed: 21952622]
- Kurian AW, Hughes E, Handorf EA, Gutin A, Allen B, Hartman AR, Hall MJ. Breast and ovarian cancer penetrance estimates derived from germline multiple-gene sequencing results in women. *JCO Precision Oncology*. 2017; 1:1–12.
- Lambert JC, Ibrahim-Verbaas CA, Harold D, Naj AC, Sims R, Bellenguez C, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for alzheimer's disease. *Nature genetics*. 2013; 45(12):1452–1458. [PubMed: 24162737]
- Lander E, Green P. Construction of multilocus genetic linkage maps in humans. *Proceedings of the National Academy of Sciences*. 1987; 84(8):2363–2367.
- Lange K. *Mathematical and statistical methods for genetic analysis*. Springer Science & Business Media; 2003.
- Lynch HT, Drescher K, Knezetic J, Lanspa S. Genetics, biomarkers, hereditary cancer syndrome diagnosis, heterogeneity and treatment: a review. *Current treatment options in oncology*. 2014; 15(3):429–442. [PubMed: 24827900]
- Marabelli M, Cheng SC, Parmigiani G. Penetrance of atm gene mutations in breast cancer: A meta-analysis of different measures of risk. *Genetic epidemiology*. 2016
- Mazzola E, Blackford A, Parmigiani G, Biswas S. Recent Enhancements to the Genetic Risk Prediction Model BRCAPRO. *Cancer Inform*. 2015; 14(Suppl 2):147–157. [PubMed: 25983549]
- Moran A, O'Hara C, Khan S, Shack L, Woodward E, Maher E, Evans D. Risk of cancer other than breast or ovarian in individuals with brca1 and brca2 mutations. *Familial cancer*. 2012; 11(2):235–242. [PubMed: 22187320]
- Murphy E, Mutalik G. The application of bayesian methods in genetic counselling. *Human Heredity*. 1969; 19(2):126–151.
- Murphy KP, Weiss Y, Jordan MI. Loopy belief propagation for approximate inference: An empirical study. *Proceedings of the fifteenth conference on uncertainty in artificial intelligence*. 1999:467–475.
- Nelson H, Pappas M, Zakher B, Mitchell J, Okinaka-Hu L, Fu R. Risk assessment, genetic counseling, and genetic testing for BRCA-related cancer in women: A systematic review to update the U.S. preventive services task force recommendation. *Annals of Internal Medicine*. 2014; 160(4):255–266. Retrieved from + DOI: 10.7326/M13-1684 DOI: 10.7326/M13-1684 [PubMed: 24366442]
- Pearl J. *Reverend bayes on inference engines: A distributed hierarchical approach*. Cognitive Systems Laboratory, School of Engineering and Applied Science, University of California; Los Angeles: 1982.
- Plichta JK, Griffin M, Thakuria J, Hughes KS. What's New in Genetic Testing for Cancer Susceptibility? *Oncology (Williston Park, N.Y)*. 2016 Sep; 30(9):787–799.

- Prasad RB, Groop L. Genetics of type 2 diabetes?pitfalls and possibilities. *Genes*. 2015; 6(1):87–123. [PubMed: 25774817]
- Riley BD, Culver JO, Skrzynia C, Senter LA, Peters JA, Costalas JW, Trepanier AM. Essential elements of genetic cancer risk assessment, counseling, and testing: updated recommendations of the National Society of Genetic Counselors. *J Genet Couns*. 2012 Apr; 21(2):151–161. [PubMed: 22134580]
- Roberts R. A genetic basis for coronary artery disease. *Trends in cardiovascular medicine*. 2015; 25(3): 171–178. [PubMed: 25453988]
- Schmidt MK, Hogervorst F, Van Hien R, Cornelissen S, Broeks A, Adank MA, et al. Age-and tumor subtype-specific breast cancer risk estimates for CHEK2* 1100delc carriers. *Journal of clinical oncology*. 2016; 34(23):2750–2760. [PubMed: 27269948]
- Schrader KA, Cheng DT, Prasad M, Walsh M, Zehir A, Thomas T, Robson M. Germline variants in targeted tumor sequencing using matched normal dna. *JAMA Oncology*. 2015; doi: 10.1001/jamaoncol.2015.5208

Figure 1.

Left. Pedigree with individual genotypes consistent with Mendelian inheritance. An individual's genotype is represented as series of +/– characters; each character represents risk heightening variants at a particular genetic locus. In this example we consider five genes. Crosses (+) wild type, dashes (–) indicate deleterious variants. Circles indicate females, squares men. Individuals with a previous breast cancer diagnosis are shaded, the question mark indicates the proband. **Right.** An example of two approximate (peeling and paring with $T = 1, 2$) and one exact (peeling with $T = 5$) posterior genotypic distributions (probabilities in logarithmic scale) for the pedigree on the left. In this example exact and approximate calculations generate nearly identical genotype distributions for the proband. The details of these computations are described in the Methods section.

Figure 2.

The relationship between cardinality of genotypes' set and the number K of genes with potential risk heightening variants. P&P indicates the proposed peeling-pairing algorithm, with different $T < K$ constraints, as indicated by the legend. Conventional peeling ($T = K$) is included.

Gene	<i>BRCA1</i>	<i>BRCA2</i>	<i>ATM</i>	<i>PALB2</i>	<i>CHEK2</i>
Prevalence	0.0061	0.0068	0.0010	0.0080	0.0010

Figure 3.

Prevalence and cumulative penetrance estimates with deleterious variants of the five genes used in the modified BRCAPRO simulation model. Prevalence and age-dependent cumulative penetrance estimates for *BRCA1* and *BRCA2* carriers and non-carriers came from the BayesMendel R package (Chen & Parmigiani, 2007; Chen et al., 2004). Penetrance and cumulative prevalence estimates for *ATM*, *PALB2*, and *CHEK2* were obtained from (Marabelli et al., 2016; A. C. Antoniou et al., 2014; Schmidt et al., 2016).

Figure 4.

Approximation accuracy across 1000 simulated breast cancer pedigrees. **Left.** Shows the distribution of the total variation across 1000 simulated breast cancer pedigrees for three different choices of the approximation parameter T . For each of the simulated families we computed four different genotype distributions: the exact posterior ($T = 5$) and three approximations ($T = 1, 2, 3$). The accuracy of the approximation increases with T . The dotted black line indicates 1% total variation. Even low values of $T = 1, 2$, consistently achieve this level of accuracy. Standard density estimation has been used for the graphical illustration of the 1000 simulation replicates shown in the figure. **Right.** Shows the distribution of total variation for different sizes of the pedigree ($N = 22, 7$). This panel corresponds to a choice of the approximation parameter $T = 2$.

	5 Gene Model		25 Gene Model	
	BRCA1 AUC	Any Mut. AUC	BRCA1 AUC	Any Mut. AUC
P & P, T=1	0.740	0.726	0.794	0.717
P & P, T=2	0.742	0.723	0.791	0.693
P & P, T=5	0.742	0.723	-	-
MCMC, 1 min	0.687	0.697	0.584	0.556
MCMC, 10 min	0.719	0.701	0.716	0.608

Figure 5.

Model discrimination results in a simulated cohort of 10,000 pedigrees, comparing the peeling and paring algorithm (with $T = 1, 2$), to a Gibbs MCMC approach. These algorithms were compared on the basis of two discrimination metrics. These are AUCs that evaluate posterior approximations ability to discriminate (i) whether or not the proband carries a *BRCA1* risk allele (top panels), and whether or not the proband carries a risk allele at any gene in the model (bottom panels). In the 5-genes model (left panels), computations by exact peeling are feasible ($T = 5$), but in the 25-genes setting (right panels), computations become infeasible both for $T = 5$ and $T = 25$. In the five gene-model, ROC curves for $T = 2$ and $T = 5$ were visually indistinguishable from the $T = 1$ and were omitted from the panels.

Error bounding. Trio families (parents and one offspring) with one binary phenotype and conditional phenotype distribution described in the previous paragraph. Probabilities that the offspring is a carrier of one or two germline risk alleles in all six possible family configurations.

Table 1

Affected Family Members						
$y = 1$	no one	one parent	offspring	both parents	offspring & parent	all
lower bound	0.0272	0.0510	0.0761	0.0725	0.1424	0.1947
approximate ($T=1$)	0.0272	0.0512	0.0762	0.0739	0.1434	0.2016
exact ($T=2$)	0.0274	0.0518	0.0768	0.0750	0.1451	0.2040
upper bound	0.0280	0.0557	0.0783	0.0970	0.1550	0.2557