# Haplotype reconstruction for genetically complex regions with ambiguous genotype calls: illustration by the KIR gene region

Burg, L.L.J. van der; Wreede, L.C. de; Baldauf, H.; Sauter, J.; Schetelig, J.; Putter, H.; Böhringer, S.

Received: 2 February 2023 | Revised: 6 September 2023 | Accepted: 25 September 2023

DOI: 10.1002/gepi.22538

RESEARCH ARTICLE

Genetic Epidemiology

OFFICIAL JOURNAL INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY www.geneticepi.org

WILEY

# Haplotype reconstruction for genetically complex regions with ambiguous genotype calls: Illustration by the *KIR* gene region

**Lars L. J. van der Burg**[1] | **Liesbeth C. de Wreede**[1,2] | **Henning Baldauf**[2] | **Jürgen Sauter**[2] | **Johannes Schetelig**[2,3] | **Hein Putter**[1] | **Stefan Böhringer**[1]

[1]Biomedical Data Sciences, LUMC, Leiden, The Netherlands

[2]DKMS, Dresden/Tübingen, Germany

[3]Department of Internal Medicine I, University Hospital Carl Gustav Carus, Dresden, Germany

**Correspondence**
Lars L. J. van der Burg, Biomedical Data Sciences, LUMC, Leiden, The Netherlands.
Email: l.l.j.van_der_burg@lumc.nl

**Abstract**

Advances in DNA sequencing technologies have enabled genotyping of complex genetic regions exhibiting copy number variation and high allelic diversity, yet it is impossible to derive exact genotypes in all cases, often resulting in ambiguous genotype calls, that is, partially missing data. An example of such a gene region is the killer-cell immunoglobulin-like receptor (*KIR*) genes. These genes are of special interest in the context of allogeneic hematopoietic stem cell transplantation. For such complex gene regions, current haplotype reconstruction methods are not feasible as they cannot cope with the complexity of the data. We present an expectation–maximization (EM)-algorithm to estimate haplotype frequencies (HTFs) which deals with the missing data components, and takes into account linkage disequilibrium (LD) between genes. To cope with the exponential increase in the number of haplotypes as genes are added, we add three components to a standard EM-algorithm implementation. First, reconstruction is performed iteratively, adding one gene at a time. Second, after each step, haplotypes with frequencies below a threshold are collapsed in a rare haplotype group. Third, the HTF of the rare haplotype group is profiled in subsequent iterations to improve estimates. A simulation study evaluates the effect of combining information of multiple genes on the estimates of these frequencies. We show that estimated HTFs are approximately unbiased. Our simulation study shows that the EM-algorithm is able to combine information from multiple genes when LD is high, whereas increased ambiguity levels increase bias. Linear regression models based on this EM, show that a large number of haplotypes can be problematic for unbiased effect size estimation and that models need to be sparse. In a real data analysis of *KIR* genotypes, we compare HTFs to those obtained in an independent study. Our new EM-algorithm-based method is the first to account for the full genetic architecture of complex gene regions,

such as the *KIR* gene region. This algorithm can handle the numerous observed ambiguities, and allows for the collapsing of haplotypes to perform implicit dimension reduction. Combining information from multiple genes improves haplotype reconstruction.

**KEYWORDS**
EM-algorithm, haplotype reconstruction, *KIR* genes

## 1 | INTRODUCTION

The advance of DNA sequencing technologies has resulted in the generation of an enormous amount of genotype data which can be used to detect associations between genes and diseases or normal traits. Most often, this data only contains unphased genotypes, that is, not indicating which alleles lie on the same chromosome (Browning & Browning, 2011). Arguably, the analysis of genotype data is best exploited through phased haplotypes, because an allelic variation in a gene can alter upstream gene expression, potentially affecting gene functioning (Browning & Browning, 2011; Tewhey et al., 2011). Haplotypes can be obtained at considerable cost experimentally, by long-range sequencing or via genotyping of family members (Becker & Knapp, 2004; Böhringer & Pfeiffer, 2009; Dudbridge, 2003, 2008; Osoegawa et al., 2019). Alternatively, haplotypes can be reconstructed statistically (Excoffier & Slatkin, 1995).

Many of the haplotype reconstruction methods are based on an expectation–maximization (EM)-algorithm, which is able to handle diploid, unphased genotypes (Excoffier & Slatkin, 1995). The unobserved phase of haplotypes is treated as missing data and maximum likelihood estimates (MLEs) can be derived when Hardy–Weinberg equilibrium (HWE) is assumed (Browning & Browning, 2011; Dudbridge, 2003, 2008; Excoffier & Slatkin, 1995; Schäfer et al., 2017). Important for the EM-algorithm is to limit the number of possible haplotypes, to keep the algorithm computationally feasible. Iterative strategies have been devised, for example, so-called partition–ligation where a divide-and-conquer strategy is used (Qin et al., 2002). Another common approach is collapsing of rare haplotypes (Sinnwell & Schaid, 2018). It is also possible to exploit information from first-degree relatives to limit the number of possible haplotypes (Becker & Knapp, 2004; Solloch et al., 2020).

In many cases, haplotype reconstruction is an intermediate step in a disease association analysis, yielding posterior probabilities as input for outcome analyses. In such analyses, either the effects of alleles separately or of haplotypes can be estimated in regression models. Additionally, haplotype reconstruction can be helpful in resolving ambiguous genotype calls.

Also Bayesian approaches are possible, which allow one to incorporate additional assumptions, such as sequence similarity, or additional, external information (Browning & Browning, 2011; Stephens et al., 2001). Incorporating such information can improve haplotype reconstruction under certain circumstances (Browning & Browning, 2009).

There is a general trade-off within and across all methods with regard to the computation costs and phasing accuracy. As a consequence, haplotype reconstructions might differ for a given data set, posing an additional analysis challenge (Al Bkhetan et al., 2019). Additionally, the stochastic nature of some algorithms requires care in choosing various parameters to ensure reproducible analyses (Becker & Knapp, 2004). Nevertheless, low-coverage next-generation sequencing (NGS) data combined with reference data might be informative enough to challenge conventional single-nucleotide polymorphism arrays in genotyping efficiency (Rubinacci et al., 2021).

Recently, more genetically complex gene regions, such as the killer-cell immunoglobulin-like receptor (*KIR*) gene region have been investigated. KIR receptors play an important role in regulating natural killer cell functioning. In the context of allogeneic hematopoietic stem cell transplantation (alloHCT), the KIR genotype of the stem cell donor might hence determine immune reactions against residual leukemic cells and thus influence the risk of relapse after transplantation. So far, research groups failed to validate predictive models for patient outcomes after alloHCT based on the absence or presence of single KIRs and their cognate ligands (Boudreau et al., 2017; Schetelig et al., 2020, 2021; Venstrom et al., 2012). Full haplotype information, which describes the *KIR* gene repertoire more comprehensively, might be more informative to predict outcomes (Cooley et al., 2010; Guethlein et al., 2021).

There are 17 *KIR* genes described in the human genome. A haplotype usually contains between 7 and 12 genes and due to copy number variation (CNV), multiple or no gene copies per haplotype are possible, with up to four copies per gene having been reported (Jiang et al., 2012; Wagner et al., 2018). Additionally, extensive

allelic polymorphism occurs, with diversity ranging between 16 and 158 alleles for a single gene (Béziat et al., 2016; Wagner et al., 2018). Due to sequence similarities, it is difficult for current typing approaches to genotype a *KIR* gene correctly on its allelic level, resulting in ambiguities. However, sequence differences between alleles can greatly influence receptor functioning (Béziat et al., 2016). The high diversity of the *KIR* gene region has been attributed to its role in the immune system in analogy with other gene regions, like, human leukocyte antigen (*HLA*) (Wagner et al., 2018).

Due to this complexity, a given genotype call results in a high number of potential diplotypes for each individual, which cannot be handled by conventional EM-algorithms. Additionally, due to limitations in the genotyping process, several forms of ambiguity remain in marginal genotype calls. Therefore, apart from haplotype phase, ambiguities have to be dealt with. To our knowledge current methods are infeasible for such type of data (Natsuhiko, 2012; Sakaue et al., 2022; Sinnwell & Schaid, 2018; Vukcevic et al., 2015). We therefore propose a method to reconstruct haplotypes for complex gene regions, such as the *KIR* loci (Juhos et al., 2016) that can accommodate various forms of ambiguity. Inferring the haplotype structure, on the one hand, allows one to help resolving observed ambiguities, on the other hand allows one to infer true, biological effects. Such biological implications are usually tackled in a second step after reconstruction, when haplotypes are associated with an outcome by means of regression analyses. It is therefore important to develop practical strategies to model haplotype associations when a high number of haplotypes can be used as predictors.

Our paper is structured as follows. In Section 2, we introduce the data and present an EM-based algorithm for genetically complex regions which can handle CNVs, ambiguity in genotype calls, high allelic polymorphism, and computationally scales to many genes. In Section 2, we also present simulation studies, which are used in Section 3 to evaluate the properties of the method under different scenarios. We analyze an extensive real KIR data set. We close with a discussion and give an outlook on future directions.

## 2 | MATERIAL AND METHODS

### 2.1 | Data

#### 2.1.1 | Structure of genotype data

In their most general form, genotype calls contain the number of gene copies and the allelic variant of each gene copy. Different allelic variants of the same gene are indicated by an asterisk, for example, *KIR2DL1*001*. Further explanation of the KIR allele nomenclature can be found in Appendix A. Due to limitations in the genotyping process, different forms of ambiguity can occur (Wagner et al., 2018). First, allelic ambiguity denotes the case of several possible alleles for a single gene copy. Second, genotypic ambiguity corresponds to multiple possibilities for the complete genotype. Third, CNV implies that there are possibly none or multiple copies of a gene on a single haplotype. As there is no location information for multiple copies, we consider multiple copies arranged in tandem when present on the same chromosome. Notationally, these copies are then separated by a hat ("^"). Illustrations for these three forms are shown in Supporting Information Table S1. Fourth, allelic variants reported as "NEW" have not been described before. Fifth, a genotype call reported as "POS" implies that the number of gene copies and the possible allelic variants are unknown but that the gene is present. Sixth, a negative genotype call ("NEG") indicates that there are no copies of that gene in the donor's genome. Technically, "NEG" does not represent an ambiguity, but it is a nonstandard genotype call. Additionally, the well-known phasing ambiguities occur when combining two heterozygous loci.

#### 2.1.2 | KIR data

The *KIR* genotyping data were collected from 4077 hematopoietic stem cell donors who were registered and provided written informed consent at the Collaborative Biobank (CoBi). Approval for reuse of the data in the current study was obtained from CoBi. Each of the donors had supplied a graft for an unrelated *HLA* compatible patient with Myelodysplastic Syndromes or Acute Myeloid Leukemia patient. More detailed information about these cohorts is described elsewhere (Schetelig et al., 2020, 2021). For each of the 17 *KIR* genes, a separate genotype call was obtained via high-resolution short-amplicon-based NGS of exons 3, 4, 5, 7, 8, and 9. Data preprocessing was performed by combining genotyping results of the *KIR2DS4* and *KIR2DS4N* into a single locus due to high sequence similarities, and by collapsing all allelic ambiguities that originate from nonsequenced exons into an artificial allele (indicated in results by a trailing "c") (Solloch et al., 2020). Subsequent data processing steps were performed to call genotypes as described elsewhere (Wagner et al., 2018).

In these data, all genotyped individuals have at least one genotype call that has one of the above-mentioned ambiguities, with on average ambiguities for 9.68 *KIR* genes (out of the 16 genes) on the allelic level. An overview of the frequencies of ambiguities is shown in

**TABLE 1** Genotype information and ambiguities for the seven *KIR* genes, given separately.

| | 2DL1 | 2DL2 | 2DL3 | 3DL1 | 2DS1 | 2DS4 | 3DS1 |
|---|---|---|---|---|---|---|---|
| # Genotypes | 586 | 117 | 482 | 2881 | 31 | 185 | 65 |
| # Alleles | 20 | 10 | 16 | 66 | 7 | 10 | 15 |
| *Genotype calls* | | | | | | | |
| Unamb | 396 | 3559 | 1922 | 1300 | 3816 | 1119 | 2606 |
| Amb | 3681 | 518 | 2155 | 2777 | 261 | 2958 | 1471 |
| *Type of ambiguities* | | | | | | | |
| Allelic | 3206 | 98 | 229 | 822 | 4 | 1319 | 1458 |
| Genotypic | 438 | 91 | 73 | 55 | 690 | 482 | 5 |
| CNV | 2641 | 356 | 1838 | 2551 | 173 | 2368 | 200 |
| NEW | 10 | 3 | 5 | 5 | 32 | 24 | 0 |
| POS | 140 | 100 | 178 | 27 | 70 | 257 | 11 |
| NEG | 131 | 1936 | 395 | 170 | 2541 | 171 | 2554 |

*Note*: Number of unique genotypes (genotypes)/alleles include potential alleles due to copy number variation (CNV) (first two rows). The last six rows list forms of observed ambiguity (see text). Each genotype call can have multiple ambiguity forms, except for the genotypes with a "POS" denotation. The multiple CNV (mCNV) indicates that there is a genotype possibility with at least two gene copies.

Abbreviations: Amb, ambiguously; KIR, killer-cell immunoglobulin-like receptor; NEG, negative genotype call; NEW, allelic variants not been described before; POS, possible allelic variants; Unamb, unambiguously.

Table 1 and Supporting Information Table S2. For the analysis, a subset of seven genes was chosen based on their potential effects reported earlier (Boudreau et al., 2017; Cooley et al., 2010; Guethlein et al., 2021; Schetelig et al., 2020, 2021; Venstrom et al., 2012). Donors that had a "POS" call for any of these seven genes were excluded to reduce complexity, leaving a total of 3547 donors in the data set.

A simple illustrative example of these genotype calls and how the information of genes is combined is shown in Supporting Information Table S3. For each gene, a set of diplotypes is compatible with the genotype call, which are combined to form a multilocus diplotype. In our diplotype notation, the two haplotypes of a diplotype are separated by a plus ("+") and the alleles of the different loci on a haplotype by a minus ("−"). If there is no gene copy for a gene, it is indicated with "NEG." In the example of Supporting Information Table S3, ambiguities are omitted in the genotype calls which would otherwise make the diplotype lists of each gene more extensive.

## 2.2 | Statistical methods

For this study, the information of the independently obtained genotype calls are combined to form multilocus haplotypes and diplotypes. The DNA sequence variations observed at a single locus are called alleles, although a tuple of alleles spanning multiple loci on the same chromosome is called a haplotype. When considering a variable number of loci, we also include single alleles as a special case of a haplotype to simplify the notation. We number the distinct haplotypes $j = 1, ..., M$. Each diplotype is a pair of haplotypes, that is, $H_i = (H_{i1}, H_{i2})$. We number the distinct diplotypes $k = 1, ..., D$. We stress, that we treat CNV occurring within a locus as alleles, that is, the complete variation within a locus including tandem copies is enumerated each being treated as a separate allele.

Haplotype frequencies (HTFs) can be estimated by optimizing the likelihood (see below). Under the hypothetical complete data scenario, no ambiguities are present and the true diplotype of each individual is known. The maximum likelihood estimator is then given as the sample frequencies of the haplotypes. However, due to the genotype and phase ambiguities in the data, for each individual, a multitude of haplotypes and diplotypes is compatible with their observed genotype. In this observed-data scenario, all these configurations have to be considered and, due to the nonexistence of a closed-form maximum likelihood estimator, require numerical optimization.

### 2.2.1 | Complete data likelihood

Complete data are given by diplotypes $H = (H_1, ..., H_N)$ for individuals $i = 1, ..., N$. Assuming that individuals are

independently, identically distributed and HWE holds, the likelihood is given by

$$L_{\text{compl}}(\pi) = P(H = h; \pi) = \prod_{i=1}^{N} \pi_{h_{i1}} \pi_{h_{i2}} c_{h_{i1},h_{i2}}. \quad (1)$$

Here, $\pi = (\pi_1, ..., \pi_M)$ is the parameter vector of HTFs and $c_{h_{i1},h_{i2}}$ is a constant representing heterozygosity (i.e., $c_{h_{i1},h_{i2}} = 2$ for a heterozygote diplotype and 1 otherwise). After rearranging the terms by haplotypes, the log-likelihood becomes

$$\ell_{\text{compl}}(\pi) = \sum_{j=1}^{M} \log(\pi_j) \sum_{i=1}^{N} \underbrace{(I\{h_{i1} = j\} + I\{h_{i2} = j\})}_{x_{ij}} + C, \quad (2)$$

where $I$ is an indicator function and $x_{ij}$ is the number of times haplotype $j$ occurs in the diplotype of individual $i$, and $C = \sum_{i=1}^{N} c_{h_{i1},h_{i2}}$ is a constant that does not depend on $\pi$.

The maximum likelihood estimator of $\pi_j$ is given by the population frequency:

$$\hat{\pi}_j = \frac{N_j}{2N}, \quad N_j = \sum_{i=1}^{N} x_{ij}. \quad (3)$$

### 2.2.2 | Observed-data likelihood

Observed data are given by the (multilocus) (ambiguous) genotype calls of individuals, denoted with $G_i$ for individual $i$: $G_i = (G_{i1}, ..., G_{iL})$, where $L$ is the number of loci. In general, the genotypes are ambiguous, that is, they represent a set of possible haplotype pairs. We denote with $\{h_1|h_2\} \sim g_i$ the set of possible diplotypes which are compatible with genotype $g_i$. Constructing this set requires taking into account both genotype ambiguities and phase uncertainty. The observed-data likelihood is then given by

$$L_{\text{obs}}(\pi) = \prod_{i=1}^{N} \left( \sum_{\{h_1|h_2\} \sim g_i} \pi_{h_1} \pi_{h_2} c_{h_1,h_2} \right). \quad (4)$$

### 2.2.3 | The EM-algorithm

Algorithmically, we derive MLEs using an EM-algorithm and start with the form as introduced by Excoffier and Slatkin (1995). The algorithm iterates by computing expectations in the E-step and updating $\pi^{(l)}$ in the M-step (Dempster et al., 1977).

*E-step*: Denote the current estimate of $\pi$ in iteration $l$ by $\pi^{(l)}$. Formally, we compute the expected complete data log-likelihood as $E_{\pi^{(l)}}[\ell_{\text{compl}}(\pi)|G]$, which gives

$$l_E(\pi; H = h|G = g) = \sum_{j=1}^{M} \log(\pi_j) \sum_{i=1}^{N} \underbrace{\sum_{\{h_1|h_2\} \sim g_i} E_{\pi^{(l)}}[I\{H_1 = j\} + I\{H_2 = j\}|g_i]}_{\hat{x}_{ij}}. \quad (5)$$

Now, $\hat{x}_{ij} = \sum_{\{h_1|h_2\} \sim g_i} [I\{h_1 = j\} + I\{h_2 = j\}](P_{\pi^{(l)}}(H_i = (h_1|h_2)|g_i))$ and with Bayes' rule, we get

$$P_{\pi^{(l)}}(H_i = (h_1|h_2)|g_i) = \underbrace{P(G_i = g_i|(h_1|h_2))}_{I_{(h_1|h_2) \sim g_i}} \frac{P(H_i = (h_1|h_2))}{P(G_i = g_i)} = I_{(h_1|h_2) \sim g_i} \frac{\pi_{h_1} \pi_{h_2} c_{h_1,h_2}}{P(G_i = g_i)}. \quad (6)$$

Here, $I_{(h_1|h_2) \sim g_i}$ indicates whether genotype $g_i$ is compatible with diplotype $H_i$.

*M-step*: The M-step is the same as the ML-estimation, plugging in $\hat{x}_{ij}$ for $x_{ij}$, leading to $\hat{\pi}_j^{(l+1)} = \frac{1}{2N} \sum_{i=1}^{N} \hat{x}_{ij}$.

For the algorithm to start, initial HTFs are based on potential haplotype occurrences in the compatible diplotypes assuming a uniform distribution. Iterations are continued until the largest change in estimated HTFs of two subsequent iterations is smaller than $1 \times 10^{-5}$ or until a maximum number of 200 iterations is reached. On the basis of simulations (see below), this limit seems sufficient to guarantee convergence. The EM-algorithm guarantees to improve the observed-data log-likelihood with each iteration of the algorithm. It may, however, only converge to a local maximum. After convergence, individual-specific frequency vectors are obtained from Equation (6), quantifying how likely each diplotype is for each individual.

A main limitation of the EM-algorithm is the limited number of haplotypes that can computationally be handled. Especially, for complex gene regions, the number of haplotypes can quickly become too large, making the current EM-algorithm computationally not feasible. We address this problem by first, collapsing rare haplotypes into a special category and second, by building up HTF estimation iteratively, by adding one locus a time into the reconstruction. Collapsing of the rare haplotypes occurs after each addition of a new locus. Due to ambiguities, and unlike the standard EM-algorithm, HTFs estimated for a smaller set of loci previously estimated cannot be considered fixed when

adding a locus. This can lead to problems with the frequency estimation of the collapsed group as explained below. Therefore, a third modification of the EM-algorithm is required in that the frequency of the collapsed group is fixed during estimation, leading to a profile EM-algorithm.

### 2.2.4 | Dimension reduction

To limit the number of haplotypes in the EM-algorithm, haplotypes with a frequency below a predefined threshold $\epsilon$ will be collapsed. Collapsing is performed after the EM-algorithm converged with the loci currently under reconstruction. To this end, all haplotypes to be collapsed are regrouped into a new, collapsed haplotype. Next, when a new locus is added into the reconstruction and HTFs were estimated, the next round of collapsing is executed. The HTFs estimated in previous steps thus cannot be assumed fixed, but are changing in general due to additional information from the added locus that might help resolve ambiguities observed in the previous reconstruction. A full optimization of all HTFs is therefore required for every step.

A newly occurring problem is that, when collapsing haplotypes into a group, full information on ambiguities cannot be retained in all cases. This happens, for example, when two individuals have ambiguous genotypes and both carry a different low-frequency haplotype in some reconstructions. If these rare haplotypes are collapsed, the ambiguities are no longer uniquely defined, because now both haplotypes are considered to be the same. As all HTFs are reoptimized in every step, this can lead to bias in estimates and lead to changes in marginal HTFs as compared with previous steps (spill-back). By absorbing more ambiguities, the rare haplotype group tends to increase in frequency in subsequent steps and this problem has to be dealt with by profiling.

### 2.2.5 | Profiling collapsed haplotypes

As the collapsed group can contain hundreds of haplotypes, the problem of spill-back mentioned above can have substantial influence on HTF estimates. To avoid this, the frequency of the collapsed group is fixed in each iteration at the sum of the HTFs in this collapsed group as estimated in the previous iteration of the EM-algorithm. We call this the profile EM-algorithm.

As ambiguities cannot be carried over for the collapsed haplotypes in all cases, also the linkage disequilibrium (LD; see Section 2.2.8) is affected by collapsing. To be able to reconstruct marginal HTFs from the final reconstruction, we additionally require marginal HTFs to be unaffected by the collapsing. To this end, after adding a new gene (gene $K$) into the reconstruction, the newly estimated HTFs are reparametrized into the marginal HTFs of the previously considered loci ($K-1$ genes), the allele frequencies (AFs) of the newly added locus ($K$th gene) and the LD values between the previous loci and the new locus (Appendix B).

Subsequently, LD values for haplotypes falling into the collapsed group with other haplotypes are set to 0. Because the collapsed group consists of a large number of haplotypes it is assumed to average out. Because these haplotypes are rare this adjustment only has a small impact.

### 2.2.6 | Reconstruction algorithm

The algorithm described so far provides a reconstruction for a single, fixed order of loci. As the collapsing depends on the order in which loci are added, in principle, all orderings should be considered, leading to exponential complexity. We use the following heuristic approach to achieve quadratic complexity. For a given reconstruction, we add each of the remaining loci in turn. Each time, HTFs are marginalized back to the starting reconstruction. We then compare how much the HTFs of the starting reconstruction have changed with the new locus, by evaluating

$$T_{\text{new}} := \sum_{j=1}^{M} \left(\pi_j^{(1)} - \pi_j^{(\text{new})}\right)^2, \tag{7}$$

where $\pi_j^{(1)}$ and $\pi_j^{(\text{new})}$ are HTFs of the starting gene with the starting reconstruction and with a newly added locus, respectively. The reconstruction is continued with the locus with maximal $T_{\text{new}}$. The added locus is therefore the most informative for the reconstruction and should help resolve ambiguities that have not yet been resolved. To start the algorithm, all loci are considered as starting locus. In Figure 1, we show all steps of the haplotype reconstruction for a fixed starting locus.

### 2.2.7 | Kullback–Leibler divergence (KLD)

To evaluate estimation performance, we use the KLD. KLD is a nonsymmetric measure indicating how well two probability distributions, $\pi$ and $\pi^R$, are alike. For HTFs, the KLD is given by

loci_in_reconstruction $\leftarrow (l_{start\_gene})$
$\hat{\pi} \leftarrow$ EM(loci_in_reconstruction)
$\hat{\pi}_{\text{coll}} \leftarrow$ collapse_frequencies($\hat{\pi}$, threshold)
**for i = 2, ..., #loci do**:
  $T_l \leftarrow ()$ ▷ Start with empty vector
  **for** l ∈ loci \ loci_in_reconstruction **do**:
    $\hat{\pi}_l \leftarrow \text{EM}_{\text{prof}}((\text{loci\_in\_reconstruction}, l), \hat{\pi}_{\text{coll}})$
    $T_l \leftarrow (T_l, T_{\text{new}}(\hat{\pi}, \text{marginal}(\hat{\pi}_l)))$
  $l_{\text{next\_gene}} \leftarrow \arg\max T_l$ ▷ Pick locus with highest $T_{\text{new}}$
  loci_in_reconstruction $\leftarrow (\text{loci\_in\_reconstruction}, l_{\text{next\_gene}})$
  $\hat{\pi}_{\text{coll}} \leftarrow$ collapse_frequencies($\hat{\pi}$, threshold)
**return** $\hat{\pi}_{\text{coll}}$

**FIGURE 1** Profile EM-algorithm for full reconstruction. $\hat{\pi}$ contains MLEs for the starting gene. Collapsed versions are stored in $\hat{\pi}_{\text{coll}}$. *loci_in_reconstruction* contains the currently reconstructed set of loci. $i$ iterates the number of loci and the iteration indexed by $l$ performs the search for the next locus to choose. $T_l$ stores statistic (7) for the locus candidates. $l_{next\_gene}$ is the number of the locus to continue the reconstruction with. Collapsing haplotype frequencies (HTFs) imply collapsing of haplotypes in the data. EM, expectation–maximization; MLE, maximum likelihood estimate.

$$KLD(\pi||\pi^R) = \sum_{j=1}^{M} \pi_j \log\left(\frac{\pi_j}{\pi_j^R}\right). \tag{8}$$

### 2.2.8 | Linkage disequilibrium

LD is the covariance between pairs of alleles of different loci: $\delta_{(i,a),(j,b)}$, where $i, j$ denote the two loci, $a, b$ are alleles at these loci. Fixing loci and alleles, we use $\delta$ for brevity. To compare LD values of different combinations, the LD is standardized, $\delta' = \delta/\delta_{\max}$, where $\delta$ is the unstandardized LD and $\delta_{\max}$ the theoretical maximum of $\delta$ for fixed AFs. When $\delta'$ is zero, there is no association, although a positive (negative) $\delta'$ means the alleles occur more (less) frequently in phase than expected under independence.

### 2.2.9 | Information $R^2$

One major goal is to evaluate whether adding loci to the reconstruction can improve the resolution of ambiguities observed for a given locus, that is, can ambiguities be resolved using information from additional loci? To this end, we fix a given locus and marginalize the estimated frequency vector per individual to get AFs per individual for the given gene. This marginalization is performed after each addition of a locus and allows one to evaluate accuracy gain (or loss) as loci are added to the reconstruction. We quantify this change by calculating an information $R^2$ in analogy to an imputation accuracy, that is, $R^2 = I_Y / I_X$, where $I_Y$ is the observed Fisher information and $I_X$ is the expected complete Fisher information of all probabilities estimated for each individual in the EM-algorithm (Appendix C). Informally, the information $R^2$ can be interpreted as the proportion of information retained in the observed (i.e., incomplete) data compared with data with known haplotypes, that is, the effective sample size.

### 2.2.10 | Majority vote (MV) analysis

To determine the benefit of applying the profile EM-algorithm in association analyses, we could not compare it to other methods mentioned in the introduction as they cannot cope with ambiguities and CNVs simultaneously. Instead, we compare with a simpler method. In this so-called MV method each individual gets assigned a single diplotype thus creating a complete data set which is simple to analyze. The choice of diplotype is based on the haplotype reconstruction of the profile EM-algorithm (with all rare haplotypes collapsed). To this end, we define the compatibility count of diplotypes as the count of how often a diplotype was compatible with observed genotypes across all individuals. The diplotype with the highest compatibility count in the reconstruction is then considered to be the individual's actual diplotype. When an individual has two or more diplotypes with the same highest frequency, a uniform, random draw is performed.

## 2.3 | Simulations

Our simulation studies are designed based on the aims, data-generating mechanisms, estimands, methods, and

performance measures (ADEMP)-structure discussed in Morris et al. (2018), which provides a structured approach for the planning and reporting of a simulation study.

### 2.3.1 | Aims

The aim of the simulations is to evaluate the profile EM-algorithm and to determine the effect of varying LD values and/or the degree of genotype ambiguities on the precision of haplotype reconstruction. As a secondary goal, we want to assess the added value of haplotype reconstruction through the profile EM-algorithm introduced in this paper in estimating regression coefficients in an outcome model for allele effects, where the simulated outcome represents a continuous outcome of a patient after alloHCT, such as immune reconstitution of patient cells.

### 2.3.2 | Data-generation mechanism

Two types of simulations were performed, each with a different underlying data-generating mechanism. For both studies, diplotypes consisting of three loci were simulated, based on AFs for each locus and certain LD values between the loci.

For diplotype simulation, loci were added iteratively, where after each addition, the LD for the matching haplotypes was altered by changing it with fractions $\theta$ to $\delta'_{\text{sim}} = \delta' + \theta(\delta_{\max} - \delta')$. The LD patterns applied to the data differ between the two simulation studies. Simulated diplotypes were converted to genotypes. After data generation, ambiguities were added to these genotype calls based on varying ambiguity scenarios. An outcome was simulated by generating a normally distributed outcome based on the truly simulated diplotypes and allele effects.

In simulation study I, the added ambiguities and LD fractions were chosen to explore the full possible spectrum, varying between being absent and very high. Due to this variation, the effect of more extreme conditions on diplotype reconstruction can be analyzed. For the second simulation study, AFs, ambiguities, and LD values were directly obtained from the analysis of *KIR2DS1*, *KIR2DL1*, and *KIR2DL3* in the real KIR data set. These genes were selected because of their different numbers of alleles, ambiguities, and AF patterns. To determine the effect of the LD between genes, LD fractions were varied around the point estimate of the real data.

The two simulation studies and the simulation of the outcome are further described in Appendix D. In short, each scenario consists of three genes, each gene with 4–7 alleles. A design matrix with intercept and all potential alleles, pairwise and three-locus haplotypes is generated, which is based on true, simulated diplotypes. Alleles are indicated by a single digit or number (e.g., "A," "001," or "NEG1"), where haplotypes are combinations of such alleles. Because all haplotypes have a fixed structure, their first allele is always from the first gene, their second allele from the second gene, and so forth. Together with regression coefficients the design matrix is used to simulate a linear outcome. For the analysis of the outcome, the EM-algorithm or MV are used to derive a design matrix which will be used for the regression. Because of the added ambiguities in the data, the design matrices of the outcome simulation and analysis are not necessarily the same.

### 2.3.3 | Estimands

The main estimand of the simulation studies are HTFs. By comparing estimated with true HTFs, the potential of the profile EM-algorithm to resolve observed ambiguities can be judged. A second estimand is effect sizes in the linear regression models. Often, association analyses based on genetic information are of interest. Our simulations give insight into such applications.

### 2.3.4 | Methods

For each scenario, 100 independent replications ($n_{\text{sim}}$) were run, each with a sample size of 2000 ($n_{\text{obs}}$). Each simulated data set was analyzed with the profile EM-algorithm and the MV analysis method. For both, the threshold for collapsing rare haplotypes was set at $1 \times 10^{-7}$, as derived in the real data analysis (Section 4). Each data set consisted of three simulated genes. To facilitate valid comparisons between all replications, the order in which the genes were added into the reconstruction was fixed as given in Supporting Information Tables S4 and S5. Additionally, to ensure that in each replication the effect sizes can be compared, all haplotypes with a theoretical probability >0.05 (the candidate list haplotypes) were never collapsed into the collapsed haplotype, even if their frequency was estimated to be under the threshold.

The output of the EM-algorithm contains expected counts of diplotypes; these are used as predictors in

linear regression with the simulation outcome per individual. More information about modeling choices is given in Appendix D.

#### 2.3.5 | Performance measures

Impact of the effect of the different scenarios and methods on the estimation of the HTFs is assessed via the KLD and the root mean square error (RMSE)

$$RMSE = \sum_{j=1}^{M'} \sqrt{\frac{1}{n} \sum_{i=1}^{n_{\text{sim}}} (\hat{Z}_{ij} - Z_{ij})^2}, \tag{9}$$

where $M'$ is the number of haplotypes for which estimates are available, and the matrix $Z$ contains these estimates. The RMSE is estimated over all 100 replications and incorporates both the bias and the variance of the estimated HTFs. Standardized RMSE (sRMSE) values were calculated as

$$sRMSE = \frac{RMSE}{Z_{ij}}. \tag{10}$$

For the comparison of the estimated HTFs with their theoretical value, KLD was estimated with the haplotypes that occur in all 100 independent replications (Section 2.2.7). Additionally, to determine whether the addition of genes into one analysis enhances the haplotype reconstruction, the information $R^2$ (Section 2.2.9) is estimated for all alleles of the starting gene that occurred in at least 5% of the replications.

To quantify simulation uncertainty due to using $n_{\text{sim}} = 100$, Monte Carlo Standard Errors (MCSEs) are estimated for the Mean Square Error (MSE) of the HTFs of each scenario. We did not choose the number of replications according to expected MCSE (Morris et al., 2018), due to computational complexity. Instead, we justify $n_{\text{sim}}$ empirically. Replicating high-ambiguity scenarios 10 times, the MCSE did not exceed $5 \times 10^{-4}$, which we deem acceptable.

For the linear regression models, RMSE and bias were estimated for the effect size estimates. Additionally, the estimated prediction error (EPE) is calculated to evaluate the accuracy of the model estimates:

$$EPE = \frac{1}{n_{\text{sim}}} \sum_{i=1}^{n_{\text{sim}}} \frac{1}{n_{\text{obs}}} \sum_{j=1}^{n_{\text{obs}}} \left( y_{ij}^c - \hat{y}_{ij} \right)^2, \tag{11}$$

with $y_i^c$ the true outcome and $\hat{y}_i$ the predicted outcome.

## 3 | RESULT OF SIMULATIONS

### 3.1 | Simulation study I

Figure 2 illustrates that the EM-algorithm estimates are consistent throughout the different scenarios: with an increase in ambiguities the sRMSE increases, which is partly compensated by increasing the LD. In contrast, the MV analysis shows less consistency, because increasing ambiguities result in both higher and lower sRMSEs, depending on the allele or haplotypes. It is clear that the HTFs are estimated more accurately in the profile EM-algorithm than that in the MV analysis, although this is mainly prominent in the two- and three-gene haplotypes, because the sRMSE values for a single gene are consistently low and not influenced by LD in the data. The sRMSE values are also plotted for all 16 scenarios in a nested-loop plot (Supporting Information Figure S1), where the lines below the plot indicate the parameter combinations of the different scenarios (Rücker & Schwarzer, 2014). Here, KLD values are also shown which quantify the benefit of the profile EM-algorithm compared with the MV analysis. How diplotype frequencies estimated with the EM-algorithm change with the addition of genes are illustrated in Supporting Information Figure S2.

For the information $R^2$ calculations, the differences after each addition of a new gene are estimated and shown in Table 2. To compare the information $R^2$ for a given allele of the starting gene, the larger reconstruction is marginalized back to the AFs of gene 1. In principle, these differences should always be positive as new information is added, however, collapsing and lack of LD might also lead to loss of information. Lack of LD induces additional variability, potentially making HTF estimation less precise. HTFs are different between scenarios due to varying LD values, and due to sampling variations not all haplotypes occur in each replication.

Table 2a shows the information $R^2$ without any added LD or ambiguities. The truly simulated alleles—the alleles included in the diplotype simulation (i.e., the green-colored alleles in the table)—all already have a high amount of available information in the first gene analysis, which does not increase with additional genes. In contrast, most nonsimulated alleles—which should be estimated with frequency zero but the occurrence of which could not be excluded during reconstruction due to ambiguities—benefit from the second and third gene, that is, their absence becomes more apparent. These observations become more pronounced in the scenario with high LD (Table 2b).

With high ambiguity added to the data, the results are different (Table 2c). As expected, the amount of information in the one-gene analysis is much lower, and the impact of

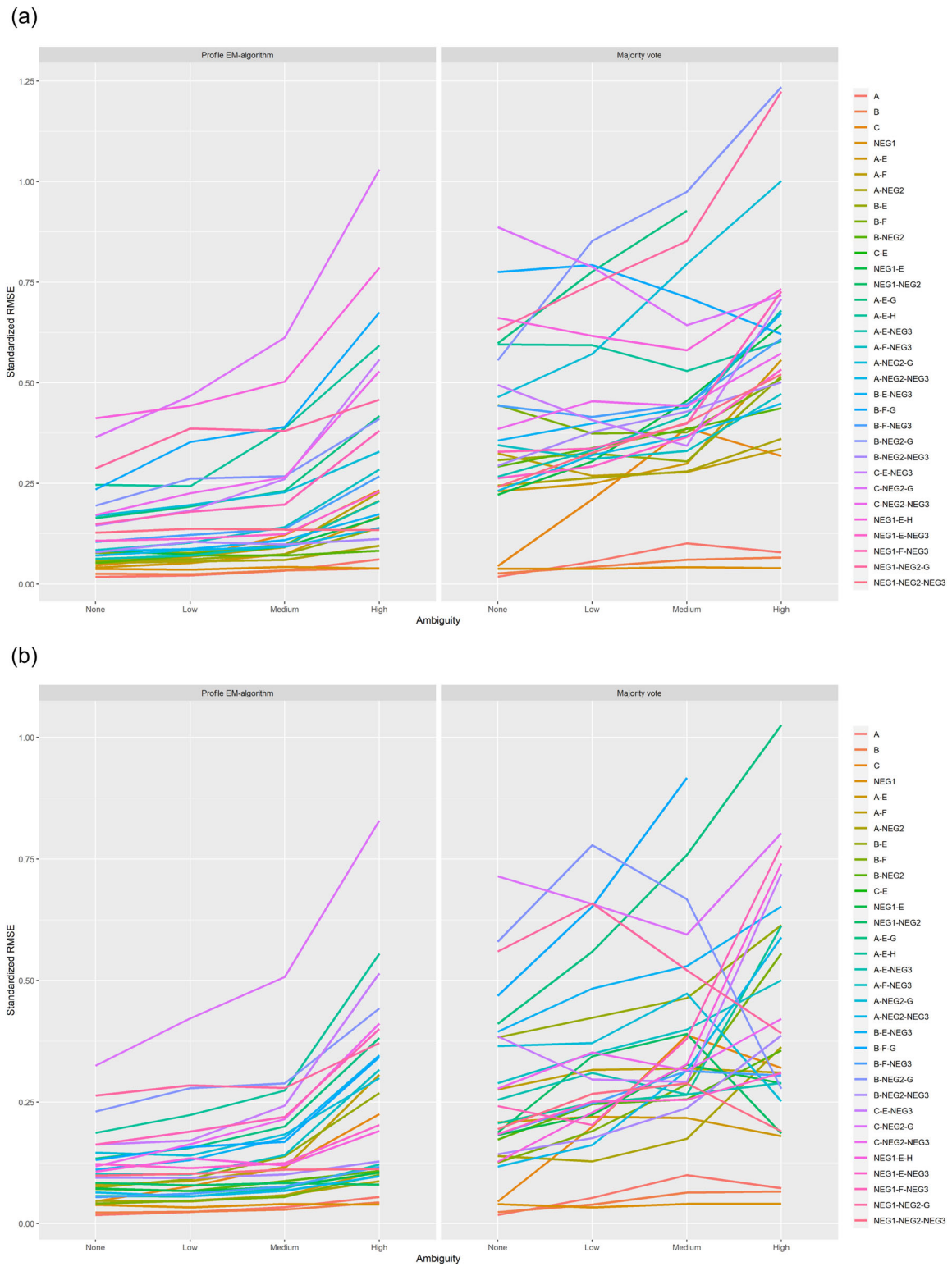


**FIGURE 2** Comparison of HTF estimation. Line graphs of sRMSEs for HTFs for the four scenarios of simulation study I with (a) no LD and (b) medium LD. Left are the four ambiguity scenarios estimated with the profile EM-algorithm, right the four ambiguity scenarios estimated with the majority vote analysis. Each line indicates the sRMSE progression of a specific allele or HT over the four ambiguity scenarios. The four alleles plotted ("A," "B," "C," and "NEG1") are all from gene 1, where the haplotypes are combinations of genes 1 and 2 (when there are two alleles separated by a "−," e.g., "A–E") or a combination of genes 1–3 (when there are three alleles, e.g., "A-E-NEG3"). These eight scenarios are together with the scenarios with low and high LD plotted together in Supporting Information Figure S1. EM, expectation–maximization; HT, haplotype; HTFs, haplotype frequencies; LD, linkage disequilibrium; sRMSE, standardized root mean square error.

**TABLE 2** Information $R^2$ simulation study I.


*Note*: For the scenarios with (a) no LD and no ambiguity is added; (b) high LD (0.8) and no ambiguity; (c) no LD and high ambiguity; (d) high LD and high ambiguity. The information $R^2$ is estimated for all haplotypes of the starting gene; average values are shown over the retained haplotypes in the independent replications. The first column of each heatmap shows the percentage of replications where the haplotype is retained, only haplotypes that are retained in at least 5% of the one-gene analyses are included. Haplotypes are ordered based on the AFs in the one-gene analysis, which are denoted in the second column. Haplotypes used for the diplotype simulation are colored green, haplotypes not used for the simulation are denoted in black. The average information $R^2$ of the one-gene analysis is shown in the third column, with the difference in the two- and three-gene combinations to these values shown in the fourth and fifth columns, respectively. The sixth column shows the amount of information for each haplotype in the three gene combinations. The color spectrum indicates whether the amount of information increases (green) or decreases (red) in the combination with respect to the one gene analysis. Completely white boxes indicate that there is no difference. The information $R^2$ heatmaps of the other 12 scenarios are displayed in Supporting Information Table S6.

Abbreviations: AFs, allele frequencies; LD, linkage disequilibrium.

adding genes into the reconstruction on the increase in information is attenuated. For most truly simulated alleles now a relative small, but positive, effect is observed with additional genes being added. Supporting Information Tables S6A and S6B show the information progression with high LD/low ambiguity and low LD/high ambiguity, respectively. The low level of ambiguity added to the data has a big effect on the benefit of adding genes, where the addition of a low level of LD only has a small impact. Overall, the combination of LD and ambiguity is beneficial for the simulated haplotypes, with throughout beneficial, albeit small, effects.

This LD/ambiguity effect is best observed when both are high (Table 2d). Haplotypes D, A^B, and C^D show a clear increase in the amount of information with multiple genes, with the amount of information for haplotype D doubling in the three-gene combination. The truly simulated alleles show a clear increase in the amount of information with multiple genes, especially compared with the scenario without any added LD. In contrast, the nonsimulated haplotypes do not seem to benefit from the added LD, compared with their values with only high levels of ambiguity.

Linear models were run based on expected allele/haplotype counts (Supporting Information Figure S3). Two observations stand out. First, most RMSE values vary between 0 and 0.6, without any big differences between the profile EM-algorithm and the MV analysis. Second, for some haplotypes in most EM-algorithm scenarios the RMSE values are very high. These outliers are caused by some extreme estimated effect sizes in some replications, originating from the relation between the main- and the two-gene combination haplotypes. For example, the four extreme haplotypes in the first column (illustrated by the triangles on top) are "B" and all two-gene combinations of "B" ("B-E," "B-F," and "B-NEG2"). In one replication, "B" gets attributed an effect size of 30, although its two-gene combinations all have effects of −30. Because a donor with "B" must also have any of the three other haplotypes, the effects balance out.

This balancing out is also apparent when examining the EPE, which is comparable for the two methods (Supporting Information Figure S3). For both, higher EPE values are observed with more ambiguity in the data, although extra LD decreases the error. Interestingly, with the highest added level of LD, hardly any extreme RMSE values occur. When comparing the median values of the absolute bias of each replication, the impact of the extreme effect sizes is decreased (Supporting Information Figure S4). With low ambiguity added, the median values are comparable for both methods. However, with higher values, the bias with the profile EM-algorithm is considerably lower. Increasing the LD between genes again lowers the bias.

To determine whether the findings of the three-gene simulation study can be extrapolated to analyses with more genes (like the real data analysis with seven *KIR* genes), the scenario with medium LD and medium ambiguity was extended with four additional genes. Details are given in Appendix E. The resulting information $R^2$ plot with seven genes (Supporting Information Figure S5) shows that the trend of increasing information by adding genes continues for these additional genes. For almost all truly simulated alleles there is a clear increase in the amount of information, which is much higher than observed in the three-gene simulation studies.

## 3.2 | Simulation study II

Except for two special scenarios where the LD between all observed haplotype combinations equals −0.2 or −0.5 (Figure 3b), the effect of the LD adjustments has little impact on performance measures (Figure 3). These two scenarios perform worse than the rest. Possibly, this is because of lower "NEG" AFs due to the altered LD values for all three genes. In general, fewer "NEG" alleles mean more ambiguity in the data and thus more uncertainty in the estimates.

The profile EM-algorithm performs better for all scenarios, particularly because of higher RMSE values for certain haplotypes (e.g., "001-003-002") in the MV analysis. This is probably because of their low frequency, which is just above the candidate list threshold. This shows a fundamental difference between the two methods. The MV method chooses one diplotype for each individual, thereby having a bias towards frequently occurring haplotypes. As a consequence, rare haplotypes, which are more abundant with higher levels of ambiguities, are lost. In contrast, the EM-algorithm does not ignore rare haplotypes, resulting in more moderate errors. The sRMSE values are also plotted in a nested-loop plot (Supporting Information Figure S6). Again, KLD values are given to quantify the benefit of the profile EM-algorithm over the MV analysis.

Table 3b shows the information $R^2$ differences for one of the special scenarios with LD equal to −0.5. In contrast to the latter scenarios, most alleles of the special scenario are lost with the addition of a second gene. The heatmaps of the standard scenarios are similar to those of simulation study I (Table 3). Truly simulated haplotypes already have a high amount of information in the one-gene analysis, although the nonsimulated haplotypes benefit from additional genes. More results are shown in the supplement (Supporting Information Table S7).

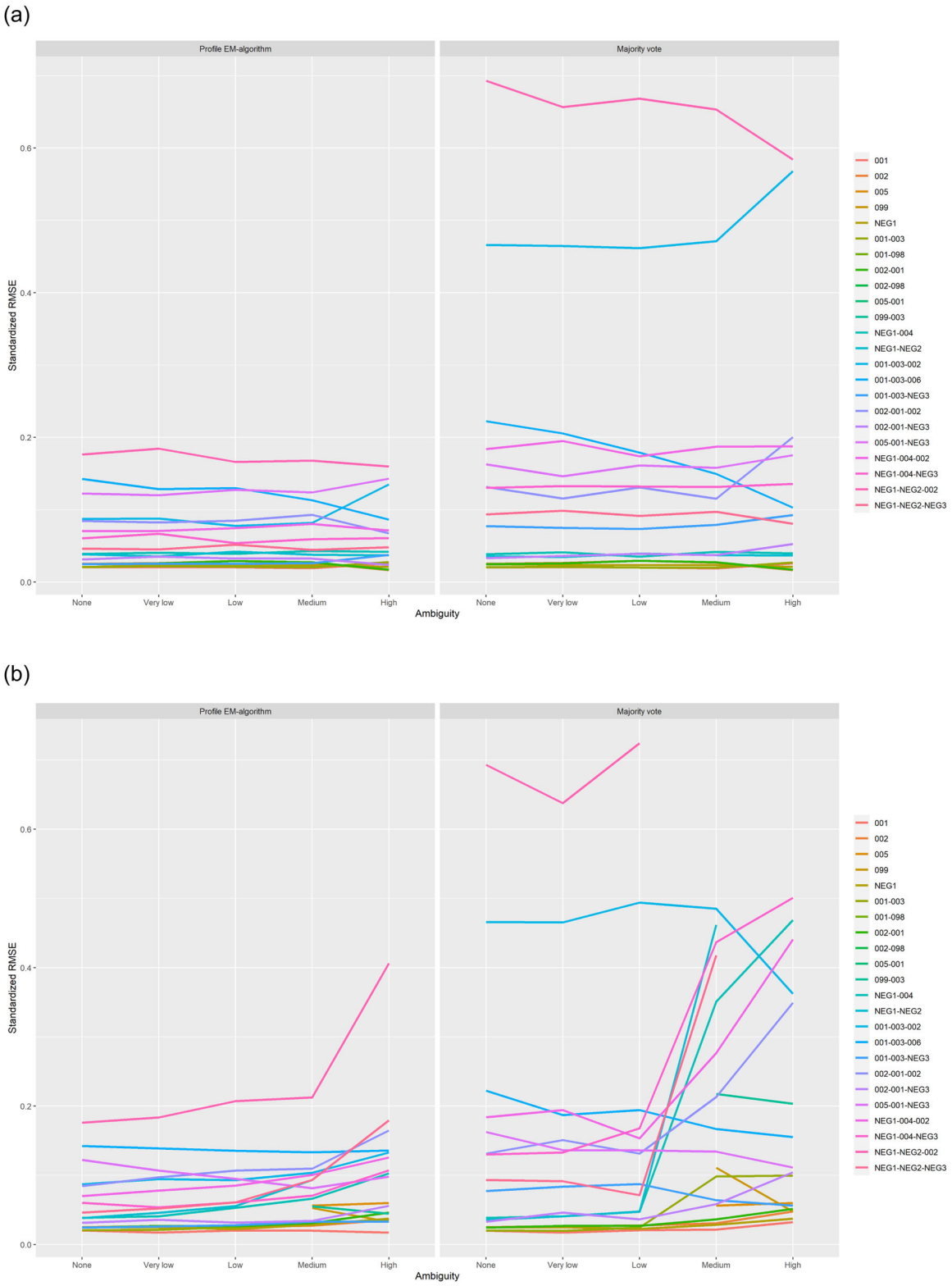


**FIGURE 3** Comparison of HTF estimation. Line graphs of sRMSEs for HTFs for the five scenarios of simulation study II with (a) positive LD or (b) negative LD added to all haplotypes. The scenario with no alterations is plotted in both (a) and (b). Left are the five scenarios estimated with the profile EM-algorithm, right the five scenarios estimated with the majority vote analysis. Each line indicates the sRMSE progression of a specific allele or HT over the five LD-altered scenarios. The four alleles plotted ("001," "002," "005," "099," and "NEG1") are all from gene 1, where the haplotypes are combinations of genes 1 and 2 (when there are two alleles separated by a "–," e.g., "001-003") or a combination of genes 1–3 (when there are three alleles, e.g., "001-003-NEG3"). These nine scenarios are together with the scenarios where LD is only altered for one haplotype combination plotted together in Supporting Information Figure S6. LD, linkage disequilibrium; HTFs, haplotype frequencies; sRMSE, standardized root mean square error; EM, expectation–maximization; HT, haplotype.

**TABLE 3** Information $R^2$ simulation study II.


*Note*: For scenario (a) the LD is not altered, the LD of all haplotypes is decreased by (b) high LD (0.5), (c) low LD (0.1), or increased by (d) high LD (0.5). The information $R^2$ is estimated for all haplotypes of the starting gene; average values are shown over the retained haplotypes in the independent replications. The first column of each heatmap shows the percentage of replications where the haplotype is retained; only haplotypes that in at least 5% of the one-gene analyses are retained are included. Haplotypes are ordered based on the AFs in the one-gene analysis, which are denoted in the second column. Haplotypes used for the diplotype simulation are colored green, haplotypes not used for the simulation are denoted in black. The average information $R^2$ of the one-gene analysis is shown in the third column, with the difference in the two- and three-gene combinations to these values shown in the fourth and fifth columns, respectively. The sixth column shows the amount of information for each haplotype in the three gene combinations. The color spectrum indicates whether the amount of information increases (green) or decreases (red) in the combination with respect to the one gene analysis. Completely white boxes indicate that there is no difference, whereas gray boxes indicate that the corresponding haplotype is completely collapsed. The information $R^2$ heatmaps of the other 13 scenarios are displayed in Supporting Information Table S7.

Abbreviations: AFs, allele frequencies; LD, linkage disequilibrium.

For all 17 different scenarios, linear models were run to estimate the haplotype effect sizes (Supporting Information Figure S7). Interestingly, the main effects of both models are biased. These biases are observed in all scenarios and not limited to haplotypes with a nonzero effect size (Supporting Information Figure S8). This is also the reason why none of the haplotypes has a RMSE close to zero, as is the case for the first simulation

study. Moreover, the extreme effect sizes for some haplotypes in the EM-algorithm are again noticeable, which are especially prevalent for haplotypes containing the "004" allele of the second gene. These haplotypes have relatively low frequencies, and therefore their effect sizes have a high variance. Due to the biases in the main effects, it is not surprising that the extreme effect sizes are also observed when looking at the median effects (Supporting Information Figure S9). As before, these extreme values are not observed in the EPEs of these linear models, which for most scenarios are comparable, with ratios slightly above one.

# 4 | KIR DATA ANALYSIS

## 4.1 | Analysis design

Each of the seven selected genes (*KIR2DS1*, *KIR2DS4*, *KIR3DS1*, *KIR2DL1*, *KIR2DL2*, *KIR2DL3*, and *KIR3DL1*) of the 3547 selected donors was used as a starting gene in the profile EM-algorithm and for the MV analysis. Both analyses were run with varying threshold values ($1 \times 10^{-i}$ for $i \in \{04, 05, 06, 07, 08, 10, 12, 15\}$) for the collapsing of rare haplotypes. The choice of a threshold value is a trade-off between computational speed, bias, and the amount of information retained.

After each analysis, the final gene sequence was converted into one common gene order to facilitate comparisons between the analyses with the different starting genes. The optimal threshold value is determined by estimating the dissimilarity between the HTFs of all subsequent analyses with the KLD for the haplotypes that occur in both analyses.

HTFs are compared with those obtained in the MV analysis, and to the estimates from another study (Solloch et al., 2020), taken as a rough comparison. In this study, HTFs are estimated in a different sample with a comparable ethnical background with a lower number of participants. This study limited the number of compatible diplotypes by including the parent-offspring genetic relationship, and the complexity is further reduced by only allowing diplotypes where the two haplotypes both met copy number patterns of 92 previously described *KIR* haplotypes. Participants that did not have any remaining options were discarded from the study. Differences with the HTFs in our study are therefore expected but HTFs should be similar.

## 4.2 | Results

Of high importance for the analysis is the chosen threshold value because this has a big impact on the number of haplotypes retained and the estimated frequencies, for example, with a threshold value of $1 \times 10^{-4}$ the analysis with starting gene *KIR2DS4* ends up with 19 uncollapsed haplotypes, although with $1 \times 10^{-15}$ 471 haplotypes are uniquely retained. In general this also results in higher running time and requires more main memory (Supporting Information Figure S10), potentially exceeding available resources.

Conceptually, the optimal threshold value implies correct reconstruction of "important" haplotypes and collapsing of others. We determine this value via a sensitivity analysis for the profile EM-algorithm. Figure 4 shows the HTF dissimilarity between the subsequent analyses. Especially the differences between the higher threshold values are evident, with a high dissimilarity between the analyses with $1 \times 10^{-4}$, $1 \times 10^{-5}$, and $1 \times 10^{-6}$ as threshold value. For two of the seven genes, dissimilarities still decrease when running the analysis with threshold $1 \times 10^{-7}$, where even lower threshold values do not change HTFs any more. On the basis of the HTF differences, $1 \times 10^{-7}$ seems to be the optimal threshold value for the real KIR data set and is therefore used for all ensuing analyses.

Table 4 shows the average values and standard deviations over the analyses with different starting genes for both analysis methods of the 20 haplotypes with the highest frequency in the profile EM-algorithm analysis. Although there are some apparent frequency differences between the two methods, that is, the frequency of the top two haplotypes differ by 0.009 and 0.013, respectively, the frequency order of the haplotypes closely matches. The standard deviation of the estimates in the MV method are somewhat higher than those from the profile EM-algorithm. However, for both methods the SDs are still low, indicating that the estimated HTFs are robust against the choice of starting locus. Additionally, the HTFs of a family-based study in a similar population are included (Solloch et al., 2020), which provides a rough comparison with some notable differences. For example, certain allelic variants which were unambiguously determined to be present in our data set did not occur in the rough comparison study. For some haplotypes, their frequencies are closer to those estimated in the profile EM-algorithm, whereas others coincide more with the frequencies from the MV analysis. Overall, the HTFs of the comparison study are comparable to those observed here, suggesting that with the collapsing of rare haplotypes, realistic HTFs can be obtained.

The information $R^2$ is estimated for all alleles of the starting gene after each reconstruction step to

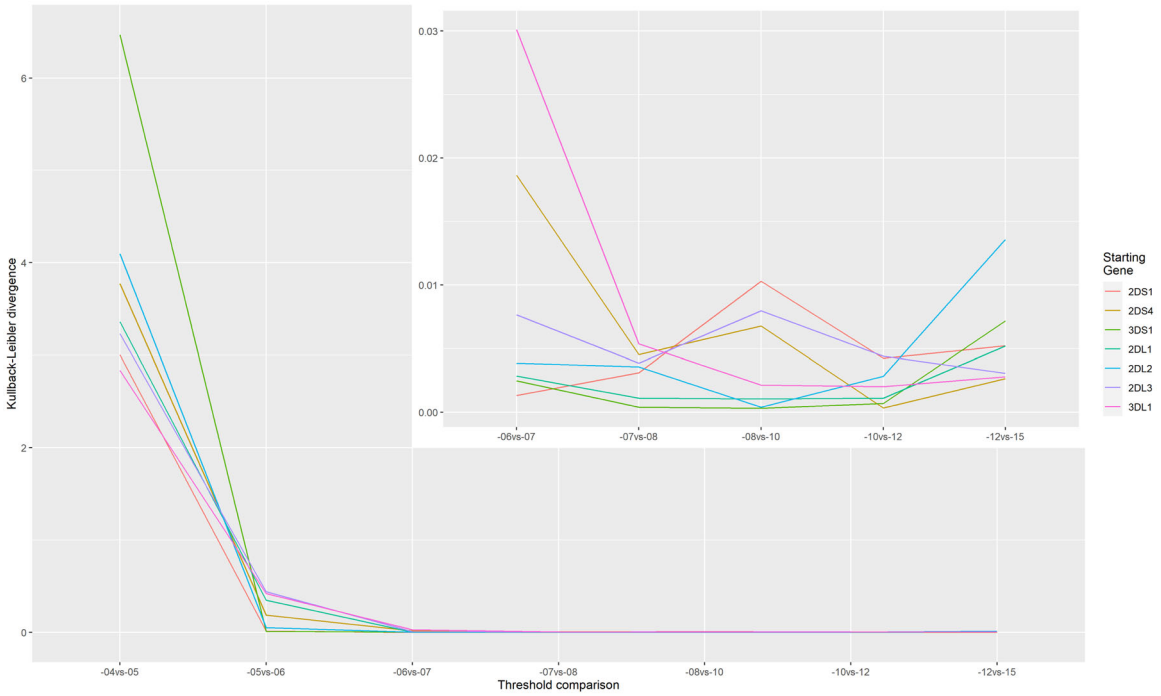


**FIGURE 4** Threshold value comparison. The dissimilarity between the HTFs of analyses run with all subsequent threshold values ($1 \times 10^{-i}$ for $i \in \{04, 05, 06, 07, 08, 10, 12, 15\}$) is estimated with the KLD. When HTFs are identical, KLD equals 0. HTFs, haplotype frequencies; KLD, Kullback–Leibler divergence.

determine whether adding genes into the reconstruction enhances HTF estimation. Table 5 shows the information $R^2$ for the analysis with *KIR2DL1* as the starting gene, the information $R^2$ progressions for the other six starting genes are given in Supporting Information Table S8. As observed above, haplotypes with a moderate or high HTF in general already have a high amount of information in the one-gene analysis, although haplotypes with low frequencies start with little available information, and benefit most from adding genes. However, also alleles with already a sufficient amount of information present can benefit from adding genes, that is, *KIR2DL1*032N* and *KIR2DS1*005* (Supporting Information Table S8), where the differences clearly increase after the inclusion of the first and sixth genes, respectively.

The addition of genes can also have a negative effect on the amount of information in the analysis. For certain haplotypes, that is, *KIR2DL1*003c^004* and *KIR2DL1*020*, the amount of information retained in the analysis is reduced, leading to lower information $R^2$ differences than observed in the single starting gene analysis. This reduction can be compensated for at later steps, as can be seen for *KIR2DL1*003c^004*. Moreover, certain haplotypes are lost (indicated by NA) due to the collapsing of the rare haplotypes.

## 5 | DISCUSSION

In this study, we developed an EM-algorithm to estimate HTFs for *KIR* genotype data when genotype calling is ambiguous. We allow for a wide range of ambiguities, including CNVs, thereby making this algorithm applicable to all human genetic regions. To our knowledge, other haplotype reconstruction methods do not allow for this wide range of ambiguities in the data. Our algorithm is important even if there is no intention to conduct haplotype-based analyses as the reconstruction of a single locus can be improved by taking into account additional loci. This is a major difference from the situation when genotypes are measured unambiguously (without error) where haplotype reconstruction can only contribute to impute missing genotypes. This benefit, however, can only be realized when sufficient LD is present in the data.

To deal with the high complexity of the data, our algorithm contains both algorithmic optimizations and heuristic modifications. First, we take an iterative approach, starting the reconstruction with a single locus and adding further loci one at a time. This approach is also taken by standard HTF estimation algorithms (Balliu et al., 2019; Sinnwell & Schaid, 2018). Unlike the standard approaches, our algorithm is potentially sensitive to the order of reconstruction, which is determined

**TABLE 4** Haplotype frequency comparison.

| Haplotype | Profile EM-algorithm | MV | Solloch et al. |
|---|---|---|---|
| NEG-010-003c-NEG-001-NEG-005 | 0.086 (0.0003) | 0.095 (0.0010) | 0.099 |
| NEG-006-001c-NEG-002-NEG-004 | 0.075 (0.0002) | 0.088 (0.0003) | 0.082 |
| NEG-003-003c-NEG-001-NEG-001c | 0.067 (0.0002) | 0.065 (0.0020) | 0.069 |
| 002c-NEG-003c-NEG-001-013c-NEG | 0.051 (0.0003) | 0.045 (0.0020) | 0.029 |
| NEG-001c-NEG-003-NEG-NEG-002 | 0.047 (0.0001) | 0.052 (0.0004) | 0.043 |
| 002c-NEG-004-001-NEG-013c-NEG | 0.047 (0.0001) | 0.052 (0.0002) | 0.025 |
| NEG-006-003c-NEG-001-NEG-004 | 0.046 (0.0002) | 0.038 (0.0002) | 0.046 |
| NEG-010-001c-NEG-002-NEG-005 | 0.035 (0.0001) | 0.028 (0.0010) | 0.029 |
| NEG-001c-003c-NEG-001-NEG-015c | 0.034 (0.0001) | 0.038 (0.0003) | 0.043 |
| NEG-003-001c-NEG-002-NEG-001c | 0.034 (0.0002) | 0.032 (0.0020) | 0.042 |
| NEG-001c-001c-NEG-002-NEG-002 | 0.032 (0.0003) | 0.031 (0.0010) | 0.033 |
| 002c-NEG-001c-NEG-002-013c-NEG | 0.030 (0.0001) | 0.030 (0.0020) | 0.012 |
| NEG-003-NEG-003-NEG-NEG-001c | 0.029 (0.0003) | 0.031 (0.0002) | 0.015 |
| NEG-003-003c-NEG-001-NEG-008 | 0.029 (0.0001) | 0.032 (0.0010) | 0.003 |
| 002c-NEG-NEG-003-NEG-013c-NEG | 0.019 (0.0002) | 0.018 (0.0001) | 0.013 |
| NEG-001c-003c-NEG-001-NEG-002 | 0.017 (0.0000) | 0.016 (0.0003) | 0.012 |
| NEG-006-004-001-NEG-NEG-004 | 0.017 (0.0001) | 0.015 (0.0004) | 0.010 |
| NEG-004-001c-NEG-002-NEG-007c | 0.015 (0.0001) | 0.016 (0.0003) | 0.018 |
| 006-NEG-003c-NEG-001-013c-NEG | 0.014 (0.0001) | 0.015 (0.0002) | 0.009 |
| NEG-003-NEG-001-NEG-NEG-001c | 0.012 (0.0002) | 0.011 (0.0003) | 0.011 |

*Note*: Average HTFs (SD) in the profile EM-algorithm and the majority vote (MV) analysis, together with the HTFs observed in a study, here used as a rough comparison (Solloch et al., 2020). Analyses were run for each of the seven starting genes with $1 \times 10^{-7}$ as a threshold value and haplotype names were recoded into a common order (*KIR2DS1-KIR2DS4-KIR2DL1-KIR2DL2-KIR2DL3-KIR3DS1-KIR3DL1*). The top 20 haplotypes are ordered numerically based on the estimates of the profile EM-algorithm.

Abbreviations: EM, expectation–maximization; HTFs, haplotype frequencies.

in a greedy way. However, our analyses suggest that already with moderate threshold values for collapsing rare haplotypes this is not the case in practice. Second, we introduce collapsing of rare haplotypes similar to other algorithms (Sinnwell & Schaid, 2018). In our case, collapsing destroys information about ambiguities, so that potential bias is introduced. We show in simulations that bias can be minimized when the collapsing threshold is kept low, that is, smaller than the frequency associated with a single occurrence of a haplotype in the present data ($\frac{1}{2n}$). Running the reconstruction with such a low threshold allows for faithful reconstruction, keeps the complexity low, and avoids inaccuracies when computing marginal frequencies, that is, computing AFs based on HTFs spanning several loci. The collapsing step makes a third modification necessary, which we call the profile EM. When ambiguous genotype calls are compatible with a rare, collapsed haplotype, full information about ambiguities cannot be retained. As the full vector of HTFs is updated in every iterative step, marginal HTFs of previously estimated loci can change, leading to accumulation of error in the EM-algorithm (spill-back, Section 2.2.4). We avoid this by fixing the frequency of haplotypes collapsed so-far in the EM-algorithm. After convergence, both the set of collapsed haplotypes and its frequencies are updated. We justify this step by using the knowledge that the collapsed haplotypes are indeed very rare (probably not present in the sample), so that this knowledge can be safely used in a subsequent step.

In summary, we can reconstruct haplotypes in large data sets containing millions of potential haplotypes such as in the analysis of seven *KIR* genes. Estimates of HTFs are unbiased and are more accurate than those obtained with the MV approach. One stated goal was to "borrow" information across loci, that is, use loci in LD to reduce ambiguities at a given locus. Our simulations show that benefits are greatest when both the level of ambiguities and LD are high. As a caveat, we point out that we parameterized LD in terms of $\delta'$ which is not directly comparable across pairs of loci. For the real KIR

**TABLE 5** Information $R^2$ **for *KIR2DL1***.


*Note*: Column (1) shows the HTFs estimated in the one-gene analysis where the amount of information in this analysis is denoted in column (2). The differences in the amount of information with an increasing number of genes in the combination, compared with this one-gene analysis, are denoted in columns (3)–(8). The color spectrum indicates whether the amount of information increases (green) or decreases (red) in the combination with respect to the one-gene analysis. Completely white boxes indicate that there is no difference, whereas gray boxes indicate that the corresponding haplotype is completely collapsed.

Abbreviation: HTFs, haplotype frequencies.

data application, we could see improvements in reconstruction for less frequent haplotypes. The gain in overall accuracy therefore clearly depends on the data set and the intended downstream analysis.

The HTFs in the first simulation study are estimated with greater precision with the profile EM-algorithm than with the MV analysis. Increasing the ambiguity in the data reduces this precision, which can only partly be

compensated by high LD. In simulation study II, the amount of ambiguity was fixed and data-derived levels of LD were already present. Small LD alterations had no noticeable effect, although large adjustments can lead to biased HTF estimation. Both methods have a bias towards frequently occurring haplotypes. However, because the EM-algorithm gives a frequency vector of HTFs for each subject, the rare haplotypes are better represented than in the all-or-nothing approach of the MV method. The downside of this is observed in the information $R^2$ heatmaps, where the nonsimulated haplotypes benefit more from adding loci than the truly simulated haplotypes. This is especially the case with high LD and smaller amounts of ambiguities, although the impact of the nonsimulated haplotypes remains small. Higher ambiguity levels partly attenuate this effect.

We have also investigated the behavior of linear regression models with the reconstructed haplotypes as predictors and investigated KLD and RMSEs of parameter estimates. A motivation for this simulation component is that an ideal EM-algorithm should also take into account outcome data. Such an EM-algorithm should not be subject to elaborate regression modeling to allow algorithmically driven analysis, justifying a saturated model. We show that using reconstruction based on the EM-algorithm in certain cases leads to less accurate regression coefficient estimates as compared with the MV reconstruction. This surprising result can be mainly explained by the fact that the MV reconstruction method tends to ignore rare haplotypes, thereby performing an implicit penalization. Also, keeping relatively rare haplotypes in the model can induce collinearity by inducing negative correlations. This interpretation is supported by the fact that ill-conditioned design matrices most frequently appear for simulation scenarios with low LD and/or high ambiguity, both being cases where haplotype diversity is high. A major conclusion concerning model building is that models should be kept sparse or that a penalized model should be used, although for the latter statistical testing becomes challenging (Böhringer & Lohmann, 2022).

## 5.1 | Computational aspects and method performance

The presented profile EM-algorithm can deal with a high diversity of scenarios. It only requires a list containing all diplotypes that are compatible with an individual's genotype which can be prespecified or automatically constructed. Some computational cost has to be expected; Supporting Information Figure S10 displays the time and required memory to run the analyses with varying number of genes, sample size, and threshold values. As a proof-of-concept we have also run simulations with seven genes which took roughly 110 GiB memory and 24 h of running time per analysis. This proof-of-concept simulation study shows that the borrowing of information of the profile EM-algorithm continues with a higher number of genes in the analysis. Currently, the EM-algorithms are run starting with a uniform distribution on compatible diplotypes. Other values can be supplied, for example, by using estimated HTFs from previous reconstructions, which might speed up convergence.

## 5.2 | KIR analysis

One major motivation for this study is to better understand the role of *KIR* genes in allogeneic stem cell transplantation. Reconstruction of haplotypes with our EM-algorithm creates new opportunities to test for the impact of donor genotype information on patient outcomes. A comparison with HTF estimates from a previous, family-based study shows good agreement (Solloch et al., 2020). Discrepancies can be explained by sampling fluctuation, but also by differences in methodology, as our algorithm is the most comprehensive in handling ambiguities so far. Albeit more validation is desirable, we are not aware of public data sets which would contain all the ambiguities considered here.

## 5.3 | Future work

Several extensions of our current algorithm seem interesting. First, incorporating data with a known phase would be straightforward but has not yet been implemented. Second, we have assumed that missingness is random, that is, the chance of an ambiguity only depends on the genotype, but not on the underlying diplotype. In practice, this might not be the case. For example, a given diplotype might be more difficult to call as compared with another diplotype that is also compatible with the given genotype. Although conceptually this scenario is likely, we are not aware of experimental quantification of such calling biases with respect to diplotypes. Including such knowledge would lead to Bayesian models that could include the a priori calling biases for which previous knowledge would be required.

Third, the inclusion of outcome data into the reconstruction algorithm is interesting. This might further improve accuracy in the presence of an association signal. Such an extension would be challenging for at least two reasons. First, as our outcome simulations

show, outcome models might be difficult to fit when larger models are considered, requiring some regularization. Second, as HTFs are updated for all loci in every iteration of the algorithm, convergence issues might arise. We see this as a challenging but promising line of future research.

In conclusion, we have developed an EM-based algorithm to reconstruct haplotypes in genetically complex regions and demonstrated robust, unbiased estimation of HTFs in all relevant scenarios.

## AUTHOR SUMMARY

For certain patients with hematological malignancies, stem cell transplantation is the only curative treatment. Important for this immunological procedure is a good match between patient and donor, thus a lot of research focuses on improving this match. However, the genotyping process of promising genes can be subject to uncertainties and ambiguities (e.g., that of the *KIR* gene region), making it difficult to evaluate their role in patient survival. In this article we introduce an algorithm that imputes these uncertainties and ambiguities by incorporating the information of neighboring genes. Because of the dimensionality of the problem three alterations had to be made to the classical algorithm to ensure valid probability estimation. In this introduction paper, we show that the probabilities estimated in our algorithm are comparable to those observed in an independent study. With reliable probabilities, the effect of gene regions that are difficult to genotype on patient survival can be mapped in a better and more comprehensive way, which can then potentially be included in the patient-donor selection criteria.

## SOFTWARE

Software is available as an R package to conduct all analyses performed in this study (https://github.com/lljvanderburg/haploemcnv). We briefly outline the main functions:

- *all_options*: to recode the per gene genotype calls into diplotype lists which are required as input for the analysis.
- *EM_algorithm*: to run a single EM-algorithm round for the supplied data.
- *HT_reconstruction*: to run the profile EM-algorithm for the supplied number of diplotype lists. Multiple parameters control the reconstruction, for example, the number of starting genes used and the threshold value below which rare haplotypes are collapsed.

At the moment, installation requires the *devtools* packages and is performed via:

```
remotes::install_github("lljvanderburg/haploemcnv)
```

Input has to be provided as character strings as multilocus genotypes as follows (cf. GLSC, 2022):

+ is the separation of gene copies (in genotype) and haplotypes (in diplotype).
− is the separation of alleles that are in the haplotype phase.
^ is the separation of multiple alleles of the same gene on a single haplotype; the so-called CNVs.
/ denotes allelic ambiguity in the genotype call.
| denotes genotype ambiguity in the genotype call.

## ACKNOWLEDGMENTS

The authors would like to thank Bose Falk and Ute Solloch for sharing and explaining the data and for discussing the results. This research has been facilitated by the Collaborative Biobank (https://www.cobi-biobank.com) and we would also like to thank the patients for contributing their information.

## CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

## DATA AVAILABILITY STATEMENT

The real KIR data set consists of the genotyping data of 17 *KIR* genes for 4077 individuals. The following *KIR* genes are included in the data: *KIR2DL1*, *KIR2DL2*, *KIR2DL3*, *KIR2DL4*, *KIR2DL5*, *KIR2DP1*, *KIR2DS1*, *KIR2DS2*, *KIR2DS3*, *KIR2DS4*, *KIR2DS4N*, *KIR2DS5*, *KIR3DL1*, *KIR3DL2*, *KIR3DL3*, *KIR3DP1*, and *KIR3DS1*. The data set is managed by the Collaborative Biobank (CoBi), where any bona fide scientist can apply to get access to its data sets. More information about the access can be found under access policy at https://www.cobi-biobank.com/scientists/welcome/. The data that support the findings of this study are available from Collaborative Biobank (CoBi). Restrictions apply to the availability of these data, which were used under license for this study. Data are available from https://www.cobi-biobank.com/scientists/welcome/with the permission of Collaborative Biobank (CoBi).

## ORCID

*Lars L. J. van der Burg* https://orcid.org/0000-0003-2359-828X

## REFERENCES

Al Bkhetan, Z., Zobel, J., Kowalczyk, A., Verspoor, K., & Goudey, B. (2019). Exploring effective approaches for haplotype block phasing. *BMC Bioinformatics*, *20*(540). https://doi.org/10.1186/s12859-019-3095-8

Balliu, B., Houwing-Duistermaat, J. J., & Böhringer, S. (2019). Powerful testing via hierarchical linkage disequilibrium in haplotype association studies. *Biometrical Journal*, *61*(3), 747–768. https://doi.org/10.1002/bimj.201800053

Becker, T., & Knapp, M. (2004). Maximum-likelihood estimation of haplotype frequencies in nuclear families. *Genetic Epidemiology*, *27*(1), 21–32. https://doi.org/10.1002/gepi.10323

Béziat, V., Hilton, H. G., Norman, P. J., & Traherne, J. A. (2016). Deciphering the killer-cell immunoglobulin-like receptor system at super-resolution for natural killer and T-cell biology. *Immunology*, *150*(3), 248–264. https://doi.org/10.1111/imm.12684

Böhringer, S., & Lohmann, D. (2022). Exact model comparisons in the plausibility framework. *Journal of Statistical Planning and Inference*, *217*, 224–240. https://doi.org/10.1016/j.jspi.2021.07.013

Böhringer, S., & Pfeiffer, R. M. (2009). A model for fine mapping in family based association studies. *Human Heredity*, *67*, 226–236. https://doi.org/10.1159/000194976

Boudreau, J. E., Giglio, F., Luduec, J. B. L., Shaffer, B. C., Hsu, K. C., Gooley, T. A., Stevenson, P. A., Malkki, M., Petersdorf, E. W., Rajalingam, R., Reed, E. F., Hou, L., Hurley, C. K., Noreen, H., Vierra-Green, C., Haagenson, M., Spellman, S., & Yu, N. (2017). KIR3DL1/HLA-B subtypes govern acute myelogenous leukemia relapse after hematopoietic cell transplantation. *Journal of Clinical Oncology*, *35*(20), 2268–2278. https://doi.org/10.1200/JCO.2016.70.7059

Browning, B. L., & Browning, S. R. (2009). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *The American Journal of Human Genetics*, *84*(2), 210–223. https://doi.org/10.1016/j.ajhg.2009.01.005

Browning, S. R., & Browning, B. L. (2011). Haplotype phasing: Existing methods and new developments. *Nature Reviews Genetics*, *12*, 703–714. https://doi.org/10.1038/nrg3054

Cooley, S., Weisdorf, D. J., Guethlein, L. A., Klein, J. P., Wang, T., Le, C. T., Marsh, S. G. E., Geraghty, D., Spellman, S., Haagenson, M. D., Ladner, M., Trachtenberg, E., Parham, P., & Miller, J. S. (2010). Donor selection for natural killer cell receptor genes leads to superior survival after unrelated transplantation for acute myelogenous leukemia. *Blood*, *116*(14), 2411–2420. https://doi.org/10.1182/blood-2010-05-283051

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood for incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1–22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

Dudbridge, F. (2003). Pedigree disequilibrium tests for multilocus haplotypes. *Genetic Epidemiology*, *25*(2), 115–121. https://doi.org/10.1002/gepi.10252

Dudbridge, F. (2008). Likelihood-based association analysis for nuclear families and unrelated subjects with missing genotype data. *Human Heredity*, *66*(2), 87–98. https://doi.org/10.1159/000119108

EMBL-EBI. (2022). *KIR allele nomenclature*. https://www.ebi.ac.uk/ipd/kir/about/nomenclature/allele/

Excoffier, L., & Slatkin, M. (1995). Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Molecular Biology and Evolution*, *12*(5), 921–927. https://doi.org/10.1093/oxfordjournals.molbev.a040269

GLSC. (2022). *Glsc 1.0 syntax*. https://glstring.org/syntax.html

Guethlein, L. A., Beyzaie, N., Nemat-Gorgani, N., Wang, T., Ramesh, V., Marin, W. M., Hollenbach, J. A., Schetelig, J., Spellman, S. R., Marsh, S. G., Cooley, S., Weisdorf, D. J., Norman, P. J., Miller, J. S., & Parham, P. (2021). Following transplantation for acute myelogenous leukemia donor, *KIR Cen B02* better protects against relapse than *KIR Cen B01*. *The Journal of Immunology*, *206*(12), 3064–3071. https://doi.org/10.4049/jimmunol.2100119

Handgretinger, R., Lang, P., & André, M. C. (2016). Exploitation of natural killer cells for the treatment of acute leukemia. *Blood*, *127*(26), 3341–3349. https://doi.org/10.1182/blood-2015-12-629055

Jiang, W., Johnson, C., Jayaraman, J., Simecek, N., Noble, J., Moffatt, M. F., Cookson, W. O., Trowsdale, J., & Traherne, J. A. (2012). Copy number variation leads to considerable diversity for B but not a haplotypes of the human KIR genes encoding NK cell receptors. *Genome Research*, *22*(10), 1845–1854. https://doi.org/10.1101/gr.137976.112

Juhos, S., Krisztina, R., & György, H. (2016). On genotyping polymorphic HLA genes—Ambiguities and quality measures using NGS. In J. K. Kulski (Ed.), *Next Generation Sequencing—Advances, Applications and Challenges*. IntechOpen. https://doi.org/10.5772/61592

Morris, T. P., White, I. R., & Crowther, M. J. (2018). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, *38*(11), 2074–2012. https://doi.org/10.1002/sim.8086

Natsuhiko, K. (2012). *Disentangler*. https://kumasakanatsuhiko.jp/projects/disentangler/

Osoegawa, K., Mallempati, K. C., Gangavarapu, S., Oki, A., Gendzekhadze, K., Marino, S. R., Brown, N. K., Bettinotti, M. P., Weimer, E. T., Montero-Martín, G., Creary, L. E., Vayntrub, T. A., Chang, C.-J., Askar, M., Mack, S. J., & Fernández-Viña, M. A. (2019). HLA alleles and haplotypes observed in 263 US families. *Human Immunology*, *80*(9), 644–660. https://doi.org/10.1016/j.humimm.2019.05.018

Qin, Z. S., Tianhua, N., & Liu, J. S. (2002). Partition–ligation–expectation–maximization algorithm for haplotype inference with single-nucleotide polymorphisms. *The American Journal of Human Genetics*, *71*(5), 1242–1247. https://doi.org/10.1086/344207

Rubinacci, S., Ribiero, D. M., Hofmeister, R. J., & Delaneau, O. (2021). Efficient phasing and imputation of low-coverage sequencing data using large reference panels. *Nature Genetics*, *53*, 120–126. https://doi.org/10.1038/s41588-020-00756-0

Rücker, G., & Schwarzer, G. (2014). Presenting simulation results in a nested loop plot. *BMC Medical Research Methodology*, *14*(129). https://doi.org/10.1186/1471-2288-14-129

Sakaue, S., Hosomichi, K., Hirata, J., Nakaoka, H., Yamazaki, K., Yawata, M., Yawata, N., Naito, T., Umeno, J., Kawaguchi, T., Matsui, T., Motoya, S., Suzuki, Y., Inoko, H., Tajima, A., Morisaki, T., Matsuda, K., Kamatani, Y., Yamamoto, K., ...

Okada, Y. (2022). Decoding the diversity of killer immunoglobulin-like receptors by deep sequencing and a high-resolution imputation method. *Cell Genomics*, *2*(3), 100101. https://doi.org/10.1016/j.xgen.2022.100101

Schäfer, C., Schmidt, A. H., & Sauter, J. (2017). Hapl-o-Mat: Open-source software for HLA haplotype frequency estimation from ambiguous and heterogeneous data. *BMC Bioinformatics*, *18*(284). https://doi.org/10.1186/s12859-017-1692-y

Schetelig, J., Baldauf, H., Heidenreich, F., Massalski, C., Frank, S., Sauter, J., Stelljes, M., AyuketanAyuk, F., Bethge, W., Bug, G., Klein, S., Wendler, S., Lange, V., de Wreede, L. C., Fürst, D., Kobbe, G., Ottinger, H., Beelen, D., Mytilineos, J., … Bornhäuser, M. (2020). External validation of models for KIR2DS1/KIR3DL1 informed selection of hematopoietic cell donors fails. *Blood*, *135*(16), 1386–1395. https://doi.org/10.1182/blood.2019002887

Schetelig, J., Baldauf, H., Koster, L., Kuxhausen, M., Heidenreich, F., de Wreede, L. C., Spellman, S., van Gelder, M., Bruno, B., Onida, F., Lange, V., Massalski, C., Potter, V., Ljungman, P., Schaap, N., Hayden, P., Lee, S. J., Kröger, N., Hsu, K., … Robin, M. (2021). Haplotype motif-based models for KIR-genotype informed selection of hematopoietic cell donors fail to predict outcome of patients with myelodysplastic syndromes or secondary acute myeloid leukemia. *Frontiers in Immunology*, *11*. https://doi.org/10.3389/fimmu.2020.584520

Sinnwell, J. P., & Schaid, D. J. (2018). *haplo.stats: Statistical analysis of haplotypes with traits and covariates when linkage phase is ambiguous*. R package version 1.7.9. https://CRAN.R-project.org/package=haplo.stats

Solloch, U. V., Schefzyk, D., Schäfer, G., Massalaski, C., Kohler, M., Pruschke, J., Heidl, A., Schetelig, J., Schmidt, A. H., Lange, V., & Sauter, J. (2020). Estimation of German KIR allele group haplotype frequencies. *Frontiers in Immunology*, *11*. https://doi.org/10.3389/fimmu.2020.00429

Stephens, M., Smith, N. J., & Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. *The American Journal of Human Genetics*, *68*(4), 978–989. https://doi.org/10.1086/319501

Tewhey, R., Bansal, V., Torkamani, A., Topol, E. J., & Schrok, N. J. (2011). The importance of phase information for human genomics. *Nature Reviews Genetics*, *12*, 215–223. https://doi.org/10.1038/nrg2950

Venstrom, J. M., Pittari, G., Gooley, T. A., Chewning, J. H., Spellman, S., Haagenson, M., Gallagher, M. M., Malkki, M., Petersdorf, E., Dupont, B., & Hsu, K. C. (2012). HLA-C-dependent prevention of leukemia relapse by donor activating KIR2DS1. *New England Journal of Medicine*, *367*(9), 805–816. https://doi.org/10.1056/NEJMoa1200503

Vukcevic, D., Traherne, J. A., Naess, S., Ellinghaus, E., Kamatani, Y., Dilthey, A., Lathrop, M., Karlsen, T. H., Franke, A., Moffatt, M., Cookson, W., Trowsdale, J., McVean, G., Sawcer, S., & Leslie, S. (2015). Imputation of KIR types from SNP variation data. *The American Journal of Human Genetics*, *97*(4), 593–607. https://doi.org/10.1016/j.ajhg.2015.09.005

Wagner, I., Schefzyk, D., Pruschke, J., Schöfl, G., Schöne, B., Gruber, N., Lang, K., Hofmann, J., Gnahm, C., Heyn, B., Marin, W. M., Dandekar, R., Hollenbach, J. A., Schetelig, J., Pingel, J., Norman, P. J., Sauter, J., Schmidt, A. H., & Lange, V. (2018). Allele-level KIR genotyping of more than a million samples: Workflow, algorithm, and observations. *Frontiers in Immunology*, *9*. https://doi.org/10.3389/fimmu.2018.02843

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** van der Burg, L. L. J., de Wreede, L. C., Baldauf, H., Sauter, J., Schetelig, J., Putter, H., & Böhringer, S. (2024). Haplotype reconstruction for genetically complex regions with ambiguous genotype calls: Illustration by the *KIR* gene region. *Genetic Epidemiology*, *48*, 3–26. https://doi.org/10.1002/gepi.22538

## APPENDIX A: *KIR* GENE REGION

There are 17 genes described in the *KIR* gene region. However, the KIR diversity is much bigger due to allelic polymorphism. To distinguish between these variants, a nomenclature for the *KIR* genes is developed (EMBL-EBI, 2022), here explained with KIR2DL1*0030201 as example:

- KIR: The acronym of the gene. For all 17 *KIR* genes this is the same.
- 2D: The number of immunoglobulin domains in the gene. There are two variants: two domains (2D) or three domains (3D).
- L: Representing the length of the cytoplasmic tail. There are three options: “L” genes result in long-tailed receptors; “S” genes result in short-tailed receptors and “P” indicates to the two pseudogenes—which are not capable of coding for proteins (Handgretinger et al., 2016).
- 1: Distinguishes between genes with the same number of immunoglobulin domains and cytoplasmic tail.
- 003: Distinguish between allelic variants that have nonsynonymous amino-acid substitutions in the coding part of the gene.
- 02: Distinguishes between allelic variants that have synonymous amino-acid substitutions in the coding part of the gene.
- 01: Distinguishes between allelic variants that have amino-acid substitutions in the noncoding part of the gene.

Where the digits of the last three points are increasing in the order in which the allelic variant is described. In this paper, the interest is on the allelic variants with nonsynonymous substitutions. Thus, alleles with only differences in the last two points are considered to be the same allelic variant.

The genomic organization of KIR haplotypes is structured relative to the fixed position of the so-called framework genes. *KIR3DL3* and *KIR3DL2* mark the boundaries of the *KIR* gene cluster and *KIR3DP1* and *KIR2DL4* are placed centrally, between which a recombination hotspot is located (Jiang et al., 2012). This divides the gene cluster into a centromeric and a telomeric motif, with each motif expressing a specific set of *KIR* genes (Cooley et al., 2010).

## APPENDIX B: REPARAMETRIZATION

Assume that the set of loci is partitioned into $K-1$ loci, the haplotypes of which are indexed by $i$ and a single, remaining locus, the alleles of which are indexed by $j$. HTFs at all loci are therefore given by the vector $(\pi_{(i,j)})_{(i,j)}$. We can now derive marginal HTFs and LD parameters as follows:

$$\pi_i := \sum_j \pi_{(i,j)}, \quad \pi_j := \sum_i \pi_{(i,j)}, \quad \delta_{i,j} := \pi_{(i,j)} - \pi_i \pi_j.$$

The mapping $(\pi_{(i,j)})_{(i,j)} \mapsto ((\pi_i)_i, (\pi_j)_j), (\delta_{i,j})_{(i,j)})$ is bijective and thereby defines a reparametrization.

## APPENDIX C: INFORMATION $R^2$

To derive the $R^2$, we focus on a single locus and dichotomize the problem by analyzing one allele and considering all other alleles to be the alternative allele. This implies that for each allele a separate $R^2$ is calculated. Let $X = (X_1, \ldots X_N)$ a dichotomized genotype vector for individuals $i = 1, \ldots, N$, where $X_i = (X_{i0}, X_{i1}, X_{i2})$ are the expected genotype counts $(0, 1, 2)$ for individual $i$. Under HWE assumptions and with AF $\theta$, the likelihood is then given by

$$L(\theta, X) = \prod_i \binom{2}{2X_{i2} + X_{i1}} \theta^{2X_{i2}+X_{i1}} (1-\theta)^{2-(2X_{i2}+X_{i1})}.$$

The score function $S$ is then given by

$$S(\theta, X) = \sum_i S(\theta, X_i) = \sum_i \frac{2X_{i2} + X_{i1} - 2\theta}{(1-\theta)\theta}.$$

For the second derivative of the log-likelihood we have

$$B(\theta, X) = \frac{\partial^2 \log L(\theta, X)}{\partial \theta^2} = \sum_i \frac{2\theta^2 + (1-2\theta)(2X_{i2} + X_{i1})}{((1-\theta)\theta)^2}.$$

For the expected information $I_X$, we get

$$I_X = \sum_i (X_{i0} B(\theta, (1,0,0)) + X_{i1} B(\theta, (0,1,0)) + X_{i2} B(\theta, (0,0,1))).$$

For the observed information, $I_Y$ calculates as follows:

$$I_Y = \sum_i S(\theta, X_i)^2.$$

## APPENDIX D: SIMULATION SETUP

### D.1 | Simulation study I

In the real KIR data set, roughly, two different allele distributions can be observed. Some genes have one allele with a very high frequency (>0.7; often the "NEG" allele). In contrast, the other group of genes exhibits a more uniform distribution (i.e., all AFs < 0.5). The AFs used for the three genes in this simulation study are given in Supporting Information Table S4, also indicating for which combinations LD was varied. Genotype ambiguity was added by specifying a set of alleles which can be ambiguous and replacing these alleles with an ambiguous genotype comprising this full allele set. These ambiguities were applied to all ambiguity alleles (when adding the ambiguity *A*/*B*, thus both to allele *A* and *B*) with given probabilities. Both for this probability and the added LD we define four levels: none (0), low (0.2), medium (0.4), or high (0.8). In total, this leads to 16 simulation scenarios, one for each LD and ambiguity combination. The same probability was used for each locus and allele when assigning ambiguities (Supporting Information Table S4). Note that, in each scenario, there is always some base level of ambiguity, even when none is explicitly added, based on haplotype phasing and CNV alleles.

### D.2 | Simulation study II

In the base scenario, AFs and LD values are derived from analysis of the three selected *KIR* genes To limit the total number of haplotypes, alleles of each gene with a frequency lower than 0.01 were collapsed in a gene-specific collapsed allele. This collapsed allele was considered as a new allele for each gene (Supporting Information Table S5). This collapsing was reflected in the real KIR data set thereby limiting the number of ambiguities in these data. The frequencies of the remaining ambiguities in the data were used as probabilities for adding ambiguities to simulated data. In the simulations, we consider two types of ambiguity: *allelic* and *genotypic* ambiguity.

To determine the effect of varying LD values, the LD values of a single haplotype combination, or that of all haplotypes with a nonzero frequency were set to the following fractions of $\delta'$: very low ($\pm 0.05$), low ($\pm 0.1$), medium ($\pm 0.2$), or high ($\pm 0.5$) value. In total, this results in 17 different scenarios, including the analysis without any LD alterations. The single haplotype that was altered was the combination of the collapsed alleles of *KIR2DL1* and *KIR2DL3*. This combination was assumed to have a $\delta'$ of zero, and was both increased and decreased.

### D.3 | Simulation of outcome

To determine the effect of ambiguities on estimating allele/haplotype effects in regression models, we use a linear model with a simulated, normally distributed outcome. The design matrix contains columns for intercept, alleles, pairwise, and three-locus haplotypes (the predictor set). We use additive coding, that is, the number of times a haplotype is present in a diplotype. The design matrix therefore only contains values {0, 1, 2}. Reference alleles were chosen to be the collapsed group of alleles per locus. Reference alleles or haplotypes containing a reference allele are excluded from the matrix to ensure identifiability. This matrix $X$ has dimensions $(n_{\text{obs}} \times A)$, where $A$ is the cardinality of the predictor set without reference alleles/haplotypes plus one (for the intercept). Outcomes are drawn according to the following model:

$$Y = X\beta + \epsilon, \tag{D1}$$

where $\beta = (\beta_1, ..., \beta_A)$, intercept $\beta_1 = 3$ and $\epsilon_i \sim N(0, 0.5)$. Because there is no clear linear outcome for the *KIR* genes, these values are determined randomly and have no biological interpretation. Supporting Information Tables S4 and S5 list the effect sizes in parentheses for the different scenarios. In all simulations, we only assign nonzero coefficients to alleles, all other (higher-order) haplotype effects were set to zero.

### D.4 | Analysis of outcome

The output of the EM-algorithm contains expected counts of diplotypes. These can be transformed into expected numbers for alleles, pairwise and three-loci haplotypes through marginalization. These expected counts are used as entries in the design matrix which has a similar form to the one used to generate the outcome, after adding an intercept and removing reference columns. We call this matrix $\tilde{X}$ $(n_{\text{obs}} \times \tilde{A})$, where $\tilde{A}$ is again the number of predictors. Due to the stochastic nature of the reconstruction algorithm, $\tilde{A}$ might differ from $A$, the number of predictors used for data generation. To fix the dimension for comparability, all haplotypes not in the candidate list are collapsed and used as references. This ensures that all replications get effect size estimates for the same haplotypes. The analysis model is given as follows:

$$Y = \tilde{X}\beta + \epsilon, \tag{D2}$$

where $\tilde{X}$ and $\beta$ are interpreted in the same way as for the data-generation model. The $\beta$s will be estimated via standard linear least square regression. In our analysis, we have chosen a saturated model which potentially exhibits high collinearity if LD is high.

## APPENDIX E: SEVEN-GENE SIMULATIONS

Where the real data analysis is executed with seven *KIR* genes, both simulation studies only include three genes. To investigate consistency between real data analysis and the simulation studies, simulation study I was extended to also include seven genes. The setup of this seven-gene simulation study is similar to that of simulation study I, with some adjustments mentioned below.

### E.1 | Data-generation mechanism

For this study, four additional genes were simulated (Supporting Information Table S9), which complement the three earlier presented genes (Supporting Information Table S4). LD and ambiguity are added to the simulated genes as described before, however, because of much higher computational costs—roughly 110 GiB memory and 24 h of running time per replication—this simulation study is only executed for the scenario with medium LD and medium ambiguities.

### E.2 | Estimands, methods, and performance measures

The estimand of this simulation study is the HTFs. For each scenario, 100 independent replications ($n_{\text{sim}}$) were run, each with a sample size of 2000 ($n_{\text{obs}}$). Each simulated data set was only analyzed with the profile EM-algorithm. Whether the addition of four extra genes into these analyses enhances the haplotype reconstruction, in a similar way as for simulation study I, is determined via the Information $R^2$ (Section 2.2.9).

### E.3 | Results

The heatmap of the seven-gene simulation study (Supporting Information Figure S5) shows a similar pattern as observed in simulation study I. With the three genes simulation studies, the amount of information with one gene is already quite high, does not increase with two genes but shows a small increase with the addition of the third gene. The seven-gene simulation study shows that this increase becomes only more profound with more genes in the analysis. Although the increase is especially visible for the nonsimulated alleles, the truly simulated alleles also clearly benefit from having a larger number of genes in the simulation.