An Accurate Method for Inferring Relatedness in Large Datasets of Unphased Genotypes via an Embedded Likelihood-Ratio Test

Jesse M. Rodriguez[⋆,1,2], Serafim Batzoglou[1], and Sivan Bercovici[1]

[1] Department of Computer Science, Stanford University
[2] Biomedical Informatics Program, Stanford University
jesserod@cs.stanford.edu

Abstract. Studies that map disease genes rely on accurate annotations that indicate whether individuals in the studied cohorts are related to each other or not. For example, in genome-wide association studies, the cohort members are assumed to be unrelated to one another. Investigators can correct for individuals in a cohort with previously-unknown shared familial descent by detecting genomic segments that are shared between them, which are considered to be identical by descent (IBD). Alternatively, elevated frequencies of IBD segments near a particular locus among affected individuals can be indicative of a disease-associated gene. As genotyping studies grow to use increasingly large sample sizes and meta-analyses begin to include many data sets, accurate and efficient detection of hidden relatedness becomes a challenge. To enable disease-mapping studies of increasingly large cohorts, a fast and accurate method to detect IBD segments is required.

We present PARENTE, a novel method for detecting related pairs of individuals and shared haplotypic segments within these pairs. PARENTE is a computationally-efficient method based on an embedded likelihood ratio test. As demonstrated by the results of our simulations, our method exhibits better accuracy than the current state of the art, and can be used for the analysis of large genotyped cohorts. PARENTE's higher accuracy becomes even more significant in more challenging scenarios, such as detecting shorter IBD segments or when an extremely low false-positive rate is required. PARENTE is publicly and freely available at *http://parente.stanford.edu/*.

Keywords: Population genetics, IBD, relatedness.

1 Introduction

Genomic sequence variants such as single-nucleotide variants, insertions, and deletions, are being constantly introduced to populations with each generation. As mutation rates are considered to be relatively low, [10] and as genetic drift drives allele frequencies to become fixed, it is reasonable to assume that two

⋆ Corresponding author.

M. Deng et al. (Eds.): RECOMB 2013, LNBI 7821, pp. 212–229, 2013.

individuals carrying the same allele have actually inherited it from a common ancestor; in such a case, the alleles can be said to be identical-by-descent (IBD). This strict definition of IBD holds for the majority of evident human germline mutations, and with high probability. Many biological applications, however, are driven by the study of longer shared stretches that cover multiple mutations. Using knowledge of such longer shared segments, inferences can be made regarding ancestry [27], population demographics [15, 19, 23], and perhaps more important, the location of disease susceptibility genes [2, 22, 4]. For such applications, the alleles of two individuals that were inherited from a *recent* common ancestor are called IBD, whereas the alleles that simply have the same allelic state but did not originate from a recent common ancestor are called identical-in-state (IIS). Note that alleles that are IBD are also IIS, but multiple independent mutation events can cause two alleles to be IIS but not IBD. It follows that in the case of a recent common ancestor, IBD alleles are harbored within longer segments containing additional IBD alleles; the more recent the common ancestor, the fewer meiosis occurred, and the longer the shared segment. In this work, we describe two individuals as being *related* to one another if they share an IBD segment from a recent common ancestor.

Identity-by-descent (IBD) inference is defined as the process of detecting genomic segments that were inherited from recent common ancestors in a given set of genotyped individuals. In the problem's simplest form, a pedigree describing the connection between sampled individuals is provided with the genotypes in order to identify the segments. Given the pedigree, a model can be derived to explicitly capture these relationships when the genotypes are examined. The most common model used is based on a factorial hidden Markov model (factorial-HMM) [26, 12] with a hidden state space defined by selector variables that determine the inheritance pattern in the pedigree [18, 1, 13, 16, 20]. More recently, such methods were extended to model linkage disequilibrium (LD) between neighboring markers, enabling the detection of shorter IBD segments [4]. The main use of these models is in the application of genetic linkage analysis. When a hereditary disease is studied in a family of healthy and affected individuals, linkage analysis is applied to identify loci that are associated with the hereditary disease; these loci may contain genes or regulatory elements that increase the probability of having the disease. The premise of linkage analysis is that affected individuals will share an IBD segment around the disease locus, and that this segment is not shared (or less likely to be shared) by healthy individuals [11, 18, 9, 24].

In the large majority of hereditary disease studies, however, the relationship between sampled individuals is unknown. In genome-wide association studies (GWAS), sampled individuals are assumed to be *unrelated*. However, it is common to have hidden relationships (also known as cryptic relationships) within large sampled cohorts [5, 15, 17].

The accurate detection of IBD segments within these samples enables the correction for the cryptic relationships, for example, by removing related individuals from analysis. Conversely, instead of discarding related individuals, IBD

mapping [7, 22, 4] can be applied, directly associating the levels of IBD with phenotype in the process of mapping disease susceptibility genes.

Extensive previous work has focused on developing methods for the accurate detection of IBD segments without using pedigree information. Most commonly, an HMM or a factorial-HMM is applied to infer the IBD segments. Purcell et al. presented PLINK [25], which uses a simple three-state model, counting the occurrences of IBD per position given the observed genotypes of two individuals. In BEAGLE, by Browning and Browning [8], a factorial HMM was developed to phase and simultaneously detect the specific haplotypes that are shared between examined individuals. To improve accuracy, the BEAGLE model captured complex linkage-disequilibrium patterns by extending the state space to accommodate the haplotypic structure found in the data and measuring the patterns' frequencies. In the work by Bercovici et al. the inheritance vector capturing the relationship between two individuals was explicitly modeled, and LD was incorporated via a first-order Markov model at the level of the founders [4]. The explicit modeling of both relationship and LD was shown to significantly improve performance. Similar to others, the work further demonstrated that these accurate inference methods could be used to detect the IBD enrichment evident around disease-gene loci, highlighting the value of IBD detection in the mapping of disease susceptibility genes. Moltke et al. presented a Markov Chain Monte Carlo (MCMC) approach for the detection of IBD regions where segments of chromosomes are it iteratively partitioned into sets of identical descent [22]. In the above methods, there exists a tradeoff between accuracy and running time. Nonetheless, in most of the above methods, the complexity of the analysis in all these methods is quadratic in the number of individuals. Simply, every pair of individuals must be examined for relatedness. GERMLINE, by Gusev et al. aimed to reduce the time complexity of IBD inference at the cost of lower accuracy [14]. The GERMLINE method performs the IBD analysis on phased data. By populating hash tables with segments taken from the phased data, the method efficiently determines potential seeds of segments that are shared between individuals. These segments are then extended to determine if sufficient evidence exists to support IBD between specific pairs of individuals. As GERMLINE requires phased data in order to operate, the individuals are first phased using BEAGLE [6]. In a later work by Browning and Browning, fastIBD [5] was developed to efficiently determine IBD segments between pairs of individuals in large cohorts of thousands of samples in a feasible timeframe. Similar to GERMLINE, fastIBD employed a sliding window approach to allow efficient computation. Pairs of individuals sharing the same state in fastIBD's factorial HMM are considered in the evaluation of subsequent windows; shared segments are extended for pairs of individuals with a high probability of IBD. While GERMLINE provides a more time-efficient solution, previous work has shown the method to have a reduced ability to detect more ancient IBD segments in comparison to more accurate methods such as fastIBD. As phasing can be prohibitive when analyzing extremely large datasets, Henn et al. developed a method aimed at detecting larger IBD segments based on reverse-homozygous

positions that does not require phasing [15]. While providing an efficient approach for IBD detection, the method is tuned to detect larger IBD segments, in order to achieve required specificity.

While advances in IBD detection have been made in recent years, accurately detecting IBD in large cohorts remains a challenge. As the cost of genotyping decreases, the number of genotyped individuals is increasing rapidly, and the genotyping density is growing to include millions of markers per sample. Since many of the accurate methods investigate all pairs of individuals for relatedness, the analysis complexity grows quadratically with the number of individuals in a studied sample. Such challenges require that IBD detection methods have high computational efficiency. More importantly, since the vast majority of examined pairs of individuals are unlikely to be related, an IBD detection method must exhibit extremely high specificity in order to avoid reporting an overwhelming number of false positives.

In this paper we present PARENTE, a novel method for the detection of IBD that exhibits high accuracy, and can be efficiently used for the analysis of large genotyped cohorts. PARENTE employs a variant of a likelihood-ratio test along with local thresholding to achieve significantly higher accuracy than the current state of the art. Our method can be applied directly on genotype data, without needing to first phase the genotypes, a step that can be computationally-intensive. The primary goal of our method is to efficiently detect which pairs of individuals in large corhorts are *related* to one another, in feasible time. This is done by finding pairs of individuals that share at least one IBD segment greater than x cM in size. Once these related pairs are identified, one can determine specific IBD segment boundaries as a post-processing step using a more complex IBD detection method of higher computational cost. We further show that PARENTE can also be directly used for the localization of the IBD segments within the related pairs, providing highly accurate results. PARENTE was able to successfully detect pairs of related individuals sharing a 6 cM IBD segment (the expected average IBD segment size for 7th cousins) with 90% sensitivity at a 5×10^{-5} false positive rate. In the more challenging case of a 4 cM shared segment, it detects related pairs with 86% sensitivity at a 8×10^{-3} false positive rate, which represents a 28% relative increase in sensitivity compared to fastIBD, a state-of-the-art method. Finally, we observed that PARENTE is an order of magnitude faster than fastIBD, as well. These results highlight the relevance of our method for the accurate and efficient analysis of large cohorts.

2 Methods

The PARENTE model employs a window-based approach, whereby multiple consecutive markers are grouped together and their joint probability is estimated given a hypothesized IBD state. Subsequently, the probabilities of multiple non-overlapping windows are merged via a naive Bayes model, producing the probability for the assumed IBD state in a given block of pre-defined length. The block lengths are derived from a target timespan covering common ancestors of interest, and the required accuracy as driven by the application.

Given N individuals sampled over M biallelic markers, let G be defined as the genotype matrix. We use $g_{i,j} \in \{0, 1, 2\}$ to denote the major allele count observed in the j^{th} marker of the i^{th} individual, and g_i as the vector corresponding to all M genotyped markers sampled for individual i. The measured genotypes G are assumed to have originated from a set of $2N$ underlying hidden haplotypes, denoted by the matrix H. The maternal and paternal alleles of the j^{th} marker in the i^{th} individual are marked as $h^m_{i,j} \in \{0, 1\}$ and $h^p_{i,j} \in \{0, 1\}$, respectively, corresponding to the major allele count in each. More broadly, however, we use h^*_j as a symbol to signify one of the alleles at the j^{th} marker, corresponding to one of the population haplotypes comprising an individual's genotype. We use f_j to denote the major allele frequency of the j^{th} marker in the sampled population. The M markers covering the genome are partitioned into a set of consecutive windows $W = \{w_1, ..., w_{\frac{M}{k}}\}$, each of size k. We use $m(w)$ to denote the indices of the k consecutive markers within the w^{th} window, and $g_{i,m(w)}$ as the partial genotyping vector for individual i corresponding to these k markers. Finally, we define a block $B = \{w_t, ..., w_{t+k-1}\}$ as a set of consecutive windows.

For a target IBD block length l (in cM), the PARENTE method is defined as follows. All $\binom{N}{2}$ pairs of individuals are enumerated. For each pair of individuals, the genome is scanned by sliding a block B across each chromosome, where each block B starts from one of the $\frac{M}{k}$ possible window positions. The examined block B includes all successive windows that contain markers that are at most l cM away from the first marker of the first window in that block. For each such block B and pair of individuals i, i', an aggregated block score $\Lambda_B(g_i, g_{i'})$ is defined as follows:

$$\Lambda_B(g_i, g_{i'}) = \sum_{w \in B} \log s_w(g_{i,m(w)}, g_{i',m(w)}) \tag{1}$$

where $s_w(g_{i,m(w)}, g_{i',m(w)})$ is a window-specific score, computed using the genotypes of the two examined individuals i, i' within an examined window w. We call a pair of individuals i and i' to be IBD in block B whenever $\Lambda_B(g_i, g_{i'}) > T_B$, where T_B is a pre-defined threshold associated with block B. We compute this score for each block in the genome and call a pair of individuals to be *related* if any block in the genome is called to be IBD. The threshold T_B is defined such that the false-positive rate is controlled to a desired level. The block score $\Lambda_B(g_i, g_{i'})$ can be efficiently computed along the genome of two individuals. As blocks are scanned, window-scores corresponding to windows that are no longer part of the newly examined block B' are subtracted from the current block score $\Lambda_{B'}(g_i, g_{i'})$, and the window-scores corresponding to newly joining windows are simply added.

In the remainder of this section we derive two instantiations for the score function $s_w(g_{i,m(w)}, g_{i',m(w)})$. We first derive a score function s_w using a likelihood-ratio approach. We continue by deriving an embedded likelihood-ratio score which corrects for the reduced performance stemming from windows exhibiting high variance in the likelihood-ratio score. Finally, we will describe how the block-specific score threshold T_B is defined. In the Results section, we show

that higher variance is associated with windows that have reduced ability to distinguish between genotypes originating from related individuals from those originating from unrelated individuals.

2.1 Likelihood Ratio Test

To efficiently detect IBD, we first develop a likelihood ratio-test (LRT) variant of our method. Within a sliding block comparing two individuals' genotypes, we contrast the probability that the they are IBD in the block against the probability that they are are not IBD. The LRT score is computed by estimating the likelihood of the individuals' genotypes within each block under two models, namely a model M_{IBD} corresponding to the hypothesis the two examined individuals are related, and a model $M_{\overline{\text{IBD}}}$ corresponding to the hypothesis the two individuals are unrelated.

As suggested by Equation 1, for both M_{IBD} and $M_{\overline{\text{IBD}}}$, we model the genotypes within a block B using a naive Bayes approach whereby all windows are independent given the IBD status of the two examined individuals within B. The probabilities of the genotypes within each window $w \in B$ comprising an examined block B are considered separately, and the product of these probabilities defines the probability of the observed genotypes within the examined block (or as a sum, under our log formulation). Namely, given a block of interest B, and the genotype of two examined individuals g_i and $g_{i'}$, the window-specific score in Equation 1 is defined as:

$$s_w^{\text{LR}}(g_{i,m(w)}, g_{i',m(w)}) = \frac{p_{M_{\text{IBD}}}(g_{i,m(w)}, g_{i',m(w)})}{p_{M_{\overline{\text{IBD}}}}(g_{i,m(w)}, g_{i',m(w)})} \tag{2}$$

Under the assumption that the sampled markers are in linkage equilibrium, meaning that the alleles within a window are not associated, the genotype probabilities under the two models are given by:

$$\begin{aligned} p_{M_{\text{IBD}}}(g_{i,m(w)}, g_{i',m(w)}) &= \prod_{j \in m(w)} p_{M_{\text{IBD}}}(g_{i,j}, g_{i',j}) \\ p_{M_{\overline{\text{IBD}}}}(g_{i,m(w)}, g_{i',m(w)}) &= \prod_{j \in m(w)} p_{M_{\overline{\text{IBD}}}}(g_{i,j}, g_{i',j}). \end{aligned} \tag{3}$$

The probability of the genotype pair $g_{i,j}, g_{i',j}$ under our two models is then defined as:

$$\begin{aligned} p_{M_{\text{IBD}}}(g_{i,j}, g_{i',j}) &= \sum_{h_j^1, h_j^2, h_j^3} p(g_{i,j}|h_j^1, h_j^2) \cdot p(g_{i',j}|h_j^1, h_j^3) \cdot p(h_j^1) \cdot p(h_j^2) \cdot p(h_j^3) \\ p_{M_{\overline{\text{IBD}}}}(g_{i,j}, g_{i',j}) &= \sum_{h_j^1, h_j^2, h_j^3, h_j^4} p(g_{i,j}|h_j^1, h_j^2) \cdot p(g_{i',j}|h_j^3, h_j^4) \cdot p(h_j^1) \cdot p(h_j^2) \cdot p(h_j^3) \cdot p(h_j^4) \end{aligned} \tag{4}$$

where $p(h_j^*) = f_j^{h_j^*} \cdot (1-f_j)^{(1-h_j^*)}$ as determined by the allele frequency at marker f_j. The probability $p(g_{i,j}|h_j^1, h_j^2)$ that the genotype $g_{i,j}$ was sampled given the underlying haplotypes h_j^1 and h_j^2, must accommodate for genotyping errors. We define $p(g_{i,j}|h_j^1, h_j^2)$ as follows:

$$p(g_{i,j}|h_j^1, h_j^2) = \begin{cases} 1-\epsilon & g_{i,j} = h_j^1 + h_j^2 \\ \frac{\epsilon}{2} & \text{otherwise} \end{cases} \tag{5}$$

where the parameter ϵ is tuned to capture the amount of expected genotyping error. Finally, to accommodate for missing data, we set the likelihood ratio at a marker to 0.5 if either genotype is missing.

We note that in the above model, the individuals can share at most a single haplotype. We further note that under the assumption of linkage equilibrium, the equivalent of a block LRT score $\Lambda_B(g_i, g_{i'})$ can be directly computed without windows by using the sums of log of the genotype probabilities, as defined by Equation 4. We utilize the window-based s_w formulation described in Equation 2 to facilitate our description of an extension that accounts for local score variability, which we now derive.

2.2 Embedded Likelihood Ratio Test

The model described thus far provides an efficient approach to identifying pairs of individuals that share a common ancestor, and in particular to detecting specific regions that are IBD. While alleviating some of the performance-related challenges that are evident when examining large cohorts by providing a computationally feasible approach, the model is sensitive to windows exhibiting highly variable scores. Namely, for each block, the window-score of a small sub-set of windows plays a critical role in the determination of the final block score. It is the high variability of such windows that limits the performance of the likelihood-ratio based test.

One approach that corrects for the detrimental impact of high-variance windows is based on the direct examination of window-level performance. The distribution of window-score can be examined given the genotypes from unrelated individuals, and contrasted against the distribution of the window-score given genotypes from related individuals. By contrasting these distributions, it is possible to detect and control for the impact of highly-variable windows. Specifically, to apply such a correction, we treat the LR described by Equation 2 as a random variable $S_w^{LR} = s_w^{\text{LR}}(g_{i,m(w)}, g_{i',m(w)})$. We then define two Gaussian models for the distribution of S_w^{LR}, one corresponding to the distribution of the score under related individuals, and a second corresponding to the distribution of the score given unrelated individuals:

$$S_w^{LR}|IBD \sim N(\mu_{w,\text{IBD}}, \sigma_{w,\text{IBD}}), \quad S_w^{LR}|\overline{\text{IBD}} \sim N(\mu_{w,\overline{\text{IBD}}}, \sigma_{w,\overline{\text{IBD}}}). \tag{6}$$

Our modified score, which we term *embedded likelihood-ratio* (ELR), is finally defined as:

$$s_w^{ELR}(g_{i,m(w)}, g_{i',m(w)}) = \frac{P(S_w^{LR} = s_w^{\mathrm{LR}}(g_{i,m(w)}, g_{i',m(w)})|\mathrm{IBD})}{P(S_w^{LR} = s_w^{\mathrm{LR}}(g_{i,m(w)}, g_{i',m(w)})|\overline{\mathrm{IBD}})}. \quad (7)$$

In total, 4 additional parameters define our new model. Namely, the mean μ and standard deviation σ of the normal distributions used to approximate the behavior of our initial score s_B^{LR} under observations originating from related and unrelated individuals. In order to estimate these parameters, phased data is used to simulate related and unrelated individuals, yielding the means to compute empirical estimates for the score distributions. The phased haplotypes can be either generated from datasets containing trios, or via computationally-phased individuals. It is important to note that current phasing methods offer a sufficiently low switch-error rate such that their performance should have a negligible effect when considering haplotypes within a window of moderate size.

2.3 Genotyping-Error Function

In Equation 5 we describe the probability of genotypes given the hidden underlying haplotype. The conditional probability $p(g_{i,j}|h_j^1, h_j^2)$ derived accounts for genotyping error. While providing a more realistic model, it can in fact reduce the statistical power when failing to reject unrelated individuals. The lower power stems from the fact the impact of reverse-homozygous genotypes is reduced; such observations can be attributed to sampling errors rather than indication of unrelatedness under the realistic model. One can increase the penalty under such scenarios by controlling the genotyping error parameter ϵ. Our method strives to reduce the amount of false-positive pairs detected. Thus, we extend our method by introducing a genotyping-error function that increases the contrast between IBD and non-IBD segments. Specifically, when estimating the model parameters, we use ϵ as the genotyping error rate, whereas during inference, we replace ϵ in Equation 5 with a function $\phi(\epsilon) = v \cdot \epsilon$, where v is a scaling factor. In the Result section, we used $v = \frac{1}{100}$.

2.4 Likelihood-Ratio Test Threshold

When applying likelihood-ratio tests, thresholds are selected so as to control the false-positive rate. Specifically, the distribution of the test is examined under examples originating from the null distribution, and a threshold is selected to guarantee an expected performance in terms of false-positives. It is common to select a single, global threshold to control for the global proportion of type I errors. However, as each block in our method contains windows of different score distribution, a local, block-specific threshold T_B can be applied to improve the performance. In our method, we explore the distribution of $\Lambda_B(g_i, g_{i'})$ given the genotypes of unrelated individuals for each block, thus accommodating to the

local behavior of our score. Given a training set of unrelated pairs and their corresponding block scores $D_{b,\overline{\text{IBD}}}$, we define the block threshold as:

$$T_B = \max(D_{b,\overline{\text{IBD}}}) + c\sigma_{D_{b,\overline{\text{IBD}}}} \tag{8}$$

where $\sigma_{D_{b,\overline{\text{IBD}}}}$ is the standard deviation observed in the block-scores, and c scales the margin defined by the standard deviation. In our experiments, we use values between -1.5 and 2.5 for the scaling-factor c.

In the Results section, we demonstrate that the combination of ELR and a block-specific threshold T_B provides superior performance in comparison to current state-of-the-art methods.

3 Results

The performance of PARENTE was evaluated using simulated data. We show that PARENTE has a superior accuracy performance when compared against fastIBD, which is considered state-of-the-art method for the accurate and efficient detection of IBD. We further explore the relative contribution to performance stemming from the use of the likelihood-ratio approach (LRT), the embedded LRT (ELRT) approach, and finally the use of a local threshold versus a global threshold. As a note on notation, for the remainder of this paper, we present the *window score* as $\log s_w(g_{i,m(w)}, g_{i',m(w)})$ instead of $s_w(g_{i,m(w)}, g_{i',m(w)})$.

Constructing Training and Testing Datasets. To train and evaluate the performance of PARENTE, we used the phased data from three Asian populations of the the HapMap Phase III panel [3]: Han Chinese in Beijing, China (CHB); Japanese in Tokyo, Japan (JPT); and Chinese in Metropolitan Denver, Colorado (CHB). Our experiments used polymorphic SNPs from the long arm of human chromosome 1. We randomly partitioned the unrelated individuals from these populations into a set of 154 training haplotypes and a set of 366 testing haplotypes. To create a larger dataset of unrelated individuals, we used the original haplotypes to generate *composite haplotypes* by simulating mosaics of the original haplotypes using an approach similar to [8]. Briefly, to generate a composite haplotype, we considered every 0.2 cM segment across the chromosome; for each segment, we copied the corresponding segment from one of the original haplotypes chosen uniformly at random. Due to the random process, some longer segments of two composite haplotypes were copied from the same original haplotype. Therefore, we removed 36 composite haplotypes that had more than 0.8 cM of contiguous sequence that was generated from the same original haplotype as another composite haplotype. A total of 500 composite training haplotypes and $1,000$ composite testing haplotypes were generated. In all of our experiments we use these composite haplotypes for training and testing. Thus, henceforth, we will refer to these composite haplotypes as simply training and testing haplotypes.

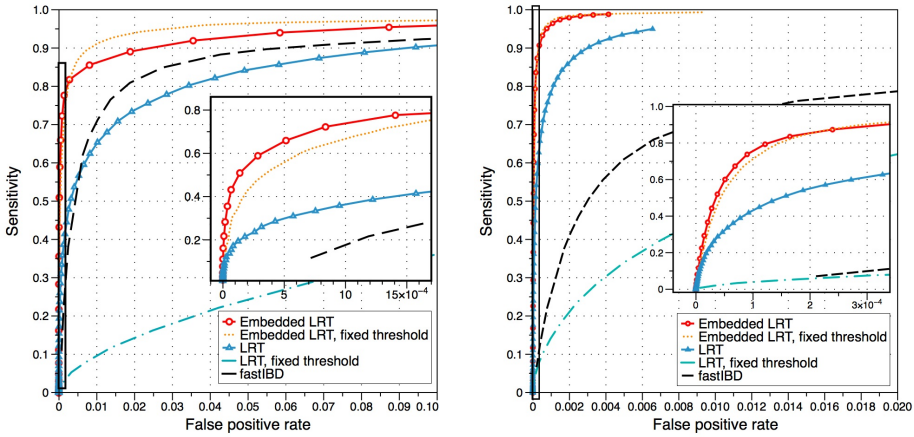

Fig. 1. (a) Performance of PARENTE for detecting related pairs of individuals sharing 4 cM IBD segments in comparison to fastIBD. PARENTE was applied using three different strategies: LRT, LRT with local thresholding, and ELRT. The magnified inset highlights PARENTE's superior performance when considering the high-specificity range. (b) Performance of PARENTE for detecting IBD segments compared to fastIBD. The same experiments from (a) were used, but the sensitivity and false positive rate were calculated based on the number of SNPs in IBD and non-IBD segments. Similarly, the magnified inset highlights PARENTE's superior performance in the high-specificity range.

Simulations to Evaluate Performance. To evaluate and characterize the performance of PARENTE, we created simulated pairs of related individuals that shared a single IBD segment of a specific size, ranging between 3 and 8 cM. We used a bootstrap approach to measure accuracy, using 100 trials per experiment, averaging the results of all trials within an experiment. For each trial, we simulated 80 pairs of related individuals by generating 80 pairs of composite individuals and inserting one shared IBD segment of a given size at a random position along the chromosome. After genotypes were copied and IBD was injected, a genotypic error rate of $\epsilon = 0.005$ was applied, changing the genotype call to one of the other two genotypes with equal probability. We designated the first simulated individual of each pair to be a *query* individual and the second individual as the *database* individual. Then we used PARENTE to predict whether IBD existed between each query individual and all database individuals by labeling a pair as IBD if at least one block had a score passing the block-specific threshold. We calculated sensitivity as the number of IBD pairs correctly predicted out of 8,000 true IBD pairs per experiment, and false positive rates as the number of non-IBD pairs incorrectly predicted as IBD out of the 632,000 non-IBD pairs per experiment.

When aiming to detect IBD segments of a particular length L (in cM), we defined the blocks to have the largest size possible l such that $L - 0.5 \leq l \leq L - 0.1$. We used block sizes slightly smaller than the target IBD segment size

to account issues related to block-boundary, stemming from the varying density of the SNP array and the fact that blocks start at window boundaries (and not at arbitrary SNPs). This was done to increase the likelihood that at least one block fit completely within the any arbitrary IBD segment of length L.

In all our experiments, we used a window size of $k = 20$ SNPs per window, and simulated a single 4 cM segment for each related pair of individuals, except where stated otherwise.

PARENTE's Accuracy and Comparison to Fastibd. Our goal was to produce a fast, accurate method to predict IBD. We thus compared the performance of PARENTE to fastIBD [5], an efficient IBD detection method. fastIBD was previously shown to have higher accuracy than GERMLINE [14], a scalable IBD detection platform, and comparable accuracy to BEAGLE's slower, high-accuracy IBD inference method [8]. We evaluated the performance of fastIBD on our simulated dataset using the default parameters and IBD detection thresholds ranging from 1×10^{-6} to 1×10^{-30}. Following fastIBD's authors recommendations, we ran fastIBD ten times with ten different seeds and aggregated the results by taking the minimum score observed at each position in any of the runs. We applied a size filter to the fastIBD predictions, only considering called segments longer than 1 cM, a value selected for yielding the best performance for fastIBD. fastIBD further recommends providing additional genotypes to aid in training fastIBD's internal haplotype model. Our experiments indicate that the use of additional haplotypes did not increase the performance (results not shown). As fastIBD infers IBD segments from all pairs in a given cohort, all the query and database individuals was provided simultaneously, while only considering calls that were made between query and database individuals, following PARENTE's mode of operation.

To compare the accuracy of PARENTE and fastIBD, we performed the simulations described above, measuring accuracy on detecting which pairs of individuals shared a simulated 4 cM IBD segment. The results shown in Figure 1a demonstrate that PARENTE has a significantly higher accuracy in comparison to fastIBD when detecting pairs of related individuals. This difference in sensitivity further grows at high-specificity levels, which is a crucial parameter when analyzing large cohorts. Note that the use of a local threshold for the ELRT provides superior high-specificity performance over a global threshold strategy. In the case of the LRT, the local threshold provides a large increase in sensitivity at all specificity levels. We further compared the performance of PARENTE and fastIBD in the task of accurately determining the location and boundaries of IBD segments (see Figure 1b). Our experiments demonstrate that PARENTE achieves higher per-SNP, per-pair accuracy when compared to fastIBD. We note that when running fastIBD for this analysis we did not enforce the called segment size filter, as fastIBD performed better when the filter was not applied. The sensitivity for each related pair of individuals was measured as the fraction SNPs in the simulated IBD segment successfully detected to be IBD. For all pairs in the experiment, we measured the false positive rate as the fraction of

SNPs not in IBD segments that were incorrectly called as IBD. Since blocks can overlap in PARENTE, we labeled a SNP as IBD if it belonged to any block that had a score above the threshold.

We characterized PARENTE performance on a range of simulated IBD segment sizes from 3 cM to 8 cM, as depicted in Figure 2. These results show that PARENTE excels at high-specificity detection of IBD segments. For instance, PARENTE was able to successfully detect 8 cM IBD segments with 94% sensitivity and nearly zero false positive rate, and 6 cM IBD segments with 90% sensitivity and a 5×10^{-5} false positive rate.

As efficiency is key in the analysis of large cohorts, we measure execution time. In our experiments, the running time for PARENTE was approximately 10 times less than that of fastIBD. Specifically, PARENTE was able to process ∼15 individual pairs per second on our trials of 6,400 pairs. Note that we measured running time in pairs per second as fastIBD analyzes all pairs within a cohort, whereas PARENTE was run on all pairings between query and database individuals.

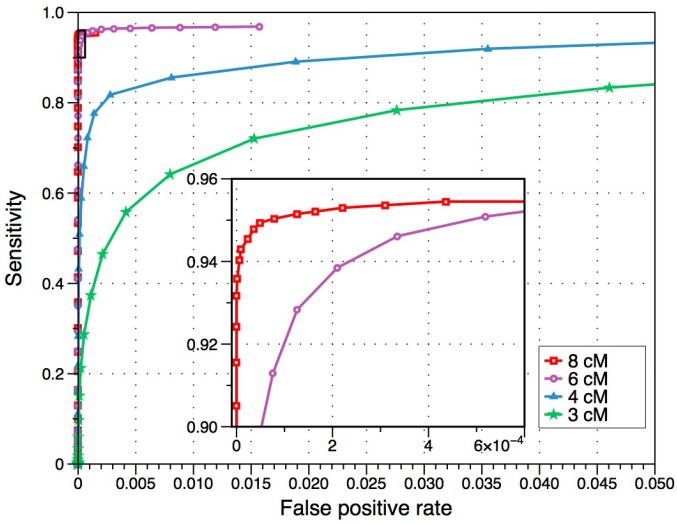

Fig. 2. Performance of PARENTE for detecting related pairs of individuals sharing IBD segments of various sizes. The magnified inset shows PARENTE's high sensitivity achieved at near-zero false positive rates for larger IBD segments.

Training PARENTE's Model and Thresholds. In order to compute our embedded LRT score, P_{IBD} and $P_{\overline{\text{IBD}}}$ first need to be evaluated for every window w. Simulated pairs of related and unrelated individuals was used for this process (see Equations 6,7). Simulated pairs of related individuals' genotypes were simulated so that each pair shared one entire haplotype along the chromosome. Specificially, each pair of related genotypes was generated by randomly selecting

one haplotype from the training data to be shared by both genotypes as well as a unique haplotype for each genotype so that three distinct haplotypes were sampled. Pair of unrelated genotypes were simulated by randomly choosing four distinct training haplotypes, using two of the haplotypes for one genotype and the remaining two haplotypes for the second genotype. A total of $2,000$ pairs of related genotypes and $2,000$ pairs of unrelated genotypes were generated. For each window w and each pair of related and unrelated genotypes, we computed the LRT score assuming a genotyping error rate of $\epsilon = 0.005$; we then fit window-specific normal distributions to the scores of related and unrelated pairs resulting in $(\mu_{w,\text{IBD}}, \sigma_{w,\text{IBD}})$ and $(\mu_{w,\overline{\text{IBD}}}, \sigma_{w,\overline{\text{IBD}}})$, respectively.

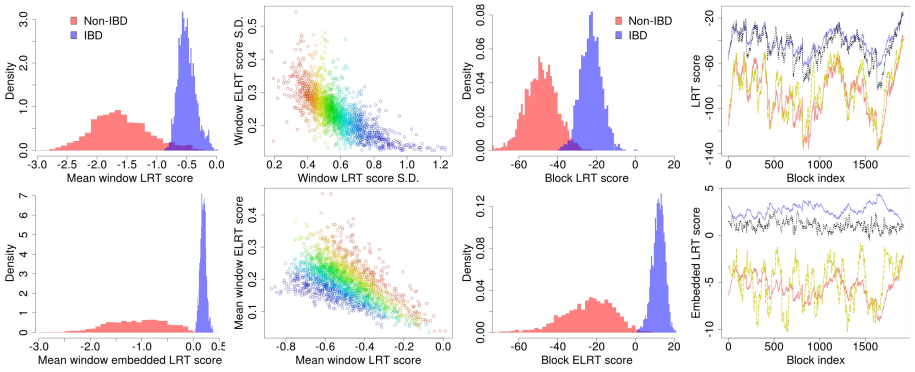

Fig. 3. (a) For each window, the mean window score of the IBD and non-IBD training data was computed; the histogram of these means is shown for the LRT and ELRT scores. When compared to the LRT score, the ELRT score has more separation between the IBD and non-IBD distributions, the boundary between them becomes centered at zero, and the IBD score variance is reduced. (b) Mean and standard deviation of window LRT scores and ELRT scores for IBD training data was computed. Each point represents a specific window, with the same color used to denote the same window in both plots. This illustrates the extent to which the ELRT reduces the variance of windows with high-variance, low-negative-mean LRT scores. (c) For a particular block, a histogram of the scores observed in the training data are shown. As with windows, the ELRT block scores feature better separation between IBD and non-IBD individuals, with a boundary close to zero. (d) Scores and thresholds across a chromosomal segment based on training data. The red line represents the mean score for non-IBD training data and the dark blue line represents the mean score for IBD training data. The yellow dashed line is the score for a single unrelated pair at each block. The dotted black line shows a local, block-specific threshold. This figure illustrates the consistent and improved separation between IBD and non-IBD score distributions at blocks across the chromosome for the ELRT over the LRT.

Embedded LRT and Local Thresholds. We computed the LRT and ELRT scores for windows and blocks for the unrelated and related training data and examined their properties in order to explore the differences between the ELRT

and LRT strategies. Figure 3a shows the distribution of the average window-score for IBD and non-IBD segments. The figure demonstrates three notable properties of the ELRT, when compared to the LRT. First and foremost, there is greater separation between the scores of IBD and non-IBD segments; second, the boundary between the scores of IBD and non-IBD segments is very close to zero, suggesting well calibrated scores; third, the variance of the scores of IBD segments is controlled. To understand the role of the ELRT's reduction variance of the IBD window scores, we plotted the mean and standard deviation of the window scores for ELRT versus LRT (see Figure 3b). Note that the ideal score distribution for IBD segments would have a high mean and low variance in order to serve as a reliable predictor for the IBD state. Therefore, these plots clearly demonstrate that ELRT controls for windows that are unreliable predictors of IBD. Specifically, the windows with high variance and low negative mean LRT scores (the blue and violet points in the figure) are mapped to lower variance ELRT scores. We note that even though there is a negative trend between the average LRT scores and average ELRT scores, the ELRT scores stay above zero, the apparent boundary between IBD and non-IBD scores. The ELRT advantages at the window level translate to the block level, as seen in Figure 3c. This greater block score separation consequently allows PARENTE to achieve higher accuracy when using the embedded LRT score. In Figure 3d, the mean of these distributions can be seen for many blocks along chromosome, demonstrating the stability of the increased separation of the ELRT across the chromosome. This figure also shows the high variation in the block thresholds in for the LRT, which explains why the LRT's performance increases significantly when using block-specific thresholds compared to a global threshold.

Accuracy Performance Characteristics. Finally, we conducted additional experiments aimed at characterizing the performance of PARENTE. Specifically, we examined the effect of genotyping errors, the use of the genotyping-error function $\phi(\epsilon)$, and the effect of varying the window size k. First, we explored PARENTE's performance with and without $\phi(\epsilon)$, assessing differences in accuracy. When using $\phi(\epsilon)$ with the scaling factor $v = \frac{1}{100}$, PARENTE's sensitivity increased from 75% to 86% at the 1% FPR level. The improvement in sensitivity further increased at the 0.1% FPR level, from 45% when using ϵ to 73%, when $\phi(\epsilon)$ was applied. Next, we demonstrated that PARENTE is robust to changes in the window size parameter. IBD pairs were inferred on simulations with 4 cM injected IBD segments for a window size of 10, 20, and 30 SNPs per window. When using the LRT score, PARENTE's sensitivity changed less than 0.5% at the 0.1% FPR level. The differences were due to the fact that block boundaries were generated to begin and end at window boundaries, resulting in block definitions that were slightly different given the window size. As noted earlier, the varying windows size does not effect the LRT score, as the window-based model is equivalent to the direct computation of the score at the block level. Simply, the LRT score of a block can be equivalently computed by summing the individual SNP LRT scores or the window LRT scores. When using the embedded LRT score,

PARENTE's sensitivity varied by less than 2% at the 0.01% FPR level across the different window sizes. These differences can be attributed to differences in the window models as well as block boundary differences. Finally, we explored the extent to which genotyping errors affected PARENTE's performance. To this end, we repeated the simulations but introduced genotyping errors at different rates: 1%, 0.5%, and 0%. The model parameters ϵ and $\phi(\epsilon)$ were unchanged from previously described experiments, being set to $\epsilon = 0.005$ and $\frac{\epsilon}{100}$, respectively. We found that at the 0.1% FPR level, the sensitivity increased from 66% to 74% to 76% for the 1%, 0.5%, and 0% error rates, respectively. These results illustrate that PARENTE is robust to a realistic range of error rates of less than 0.5%.

4 Discussion

To improve computational efficiency when applying the described scoring functions, the log window score $\log s_w(g_{i,m(w)}, g_{i',m(w)})$ can be pre-computed for all possible pairs of genotypes for every window. For instance, with a window size of 5 SNPs, each window requires only $\frac{(3^5)(3^5+1)}{2} = 29,646$ values per window. The block score $\Lambda_b(g_i, g_{i'})$ can then be computed efficiently by retrieving and summing these values.

The model presented here assumes markers within each window are in linkage equilibrium. One approach to satisfy this assumption is via marker pruning using tools such as PLINK [25]. Alternatively, our model can be extended so as to incorporate the LD evident between neighboring markers. Previous work has shown that modeling LD can improve the performance of IBD methods [4].

In our work, the applied block-specific threshold strategy was based on the observed scores of unrelated pairs in the training data. The rationale behind this approach was to extremely control for false positions, since we aim to identify IBD in extremely large cohorts. Therefore, we calculated the threshold based on the maximum and variance of the observed training scores and a provided constant, c (see Equation 8). The default value $c = 0$ yielded a threshold with good performance (82% sensitivity at a 3×10^{-3} FPR for the embedded LRT); c can be adjusted to achieve the preferred tradeoff between specificity and sensitivity. We have observed that the margin between the related and unrelated distributions varies between blocks (see Figure 3). One may be able to increase sensitivity without loss of specificity by increasing the thresholds at blocks where the margin is large. In future work, we aim to explore additional stronger thresholding schemes in order to increase PARENTE's accuracy.

PARENTE makes the assumption that IBD segments along the genome are independent of one another, which holds true for distant relatives with relatively small IBD segments (eg 5 cM) that are expected to have at most one shared IBD segment. The assumption may not hold true for closely-related individuals, which are expected to share several IBD segments. However, due to the close relationships in these scenarios, these IBD segments also tend to be very large. Because PARENTE can accurately detect individual small IBD segments, it can

Table 1. As window size increases, a Gaussian distribution fits window LRT scores better. Given a window size (SNPs per window), the Kolmogorov−Smirnov test was performed on the scores of the training data for each window along the chromosomal segment. The mean p-value of all the windows is reported here.

	SNPs per window				
	3	**5**	**10**	**15**	**20**
Mean non-IBD KS p-value	7e-12	4e-11	8e-7	5e-6	2e-5
Mean IBD KS p-value	1e-9	4e-5	0.003	0.008	0.017

also detect each individual larger IBD segment, without needing to take into account that several large IBD segments may appear across the genome.

Our model uses a normal approximation of the LRT score distribution in order to compute the ELRT scores. With a window size of 20 SNPs per window, as used in our experiments, the LRT score distributions of most windows reasonably follow a Gaussian distribution. Naturally, however, for smaller window sizes (such as 3 SNPs window), most windows had score distributions that does not fit a Gaussian distribution. The poor approximation of the LRT score via a Gaussian distribution resulted in reduced performance (results not shown). We quantified window LRT score normality across various window sizes by using a Kolmogorov−Smirnov (KS) test on the related and unrelated training LRT scores for each window. The mean p-value of all the windows along the chromosome was computed. Table 1 shows these results, illustrating that the approximation using a Gaussian distribution provides a better fit as the window size increases. These observations indicate that it may be worthwhile to explore alternative parametric and empirical distributions for LRT, evaluating their impact on PARENTE's accuracy, especially when using small window sizes.

In this paper we presented PARENTE, a novel method for the accurate and efficient detection of IBD. Our results demonstrate that PARENTE has a superior accuracy in comparison to previous state-of-the-art methods, especially when set to control for extremely low false-positive rates. Furthermore, the methods efficiency enables the analysis of large-cohorts sampled over dense marker sets. As larger dataset are collected and sampled at an increasingly higher resolution via next-generation sequencing [21, 28], efficient methods such as PARENTE that can operate on non-phased genotype data become vital for their analysis. PARENTE is publicly and freely available at *http://parente.stanford.edu/*.

Acknowledgments. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1147470. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. This work is also supported by a grant from the Stanford-KAUST alliance for academic excellence. We would like to thank Kelly Gilbert for helpful feedback in preparing the manuscript and two anonymous reviewers for many helpful comments.

References

[1] Abecasis, G.R., Cherny, S.S., Cookson, W.O., Cardon, L.R.: Merlin–rapid analysis of dense genetic maps using sparse gene flow trees. Nat. Genet. 30(1), 97–101 (2002)
[2] Alkuraya, F.S.: Homozygosity mapping: one more tool in the clinical geneticist's toolbox. Genet. Med. 12(4), 236–239 (2010)
[3] Altshuler, D.M., Gibbs, R.A., Peltonen, L., Dermitzakis, E., Schaffner, S.F., Yu, F., Bonnen, P.E., De Bakker, P.I.W., Deloukas, P., Gabriel, S.B., et al.: Integrating common and rare genetic variation in diverse human populations. Nature 467(7311), 52–58 (2010)
[4] Bercovici, S., Meek, C., Wexler, Y., Geiger, D.: Estimating genome-wide ibd sharing from snp data via an efficient hidden markov model of ld with application to gene mapping. Bioinformatics 26(12), i175–i182 (2010)
[5] Browning, B.L., Browning, S.R.: A fast, powerful method for detecting identity by descent. American Journal of Human Genetics 88(2), 173–182 (2011)
[6] Browning, S., Browning, B.: Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am. J. Hum. Genet. 81(5), 1084–1097 (2007)
[7] Browning, S., Thompson, E.: Detecting Rare Variant Associations by Identity by Descent Mapping in Case-control Studies. Genetics 190, 1521–1531 (2012)
[8] Browning, S.R., Browning, B.L.: High-Resolution Detection of Identity by Descent in Unrelated Individuals. American Journal of Human Genetics 86(4), 526–539 (2010)
[9] Carey, V.J.: Mathematical and statistical methods for genetic analysis (2nd ed.). kenneth lange. Journal of the American Statistical Association 100, 712 (2005)
[10] Conrad, D.F., Keebler, J.E.M., DePristo, M.A., Lindsay, S.J., Zhang, Y., Casals, F., Idaghdour, Y., Hartl, C.L., Torroja, C., Garimella, K.V., Zilversmit, M., Cartwright, R., Rouleau, G.A., Daly, M., Stone, E.A., Hurles, M.E., Awadalla, P., for the 1000 Genomes Project: Variation in genome-wide mutation rates within and between human families. Nature Genetics (2011)
[11] Elston, R., Stewart, J.: A general model for the analysis of pedigree data. Hum. Hered. 21, 523–542 (1971)
[12] Ghahramani, Z., Jordan, M.I., Smyth, P.: Factorial hidden markov models. In: Machine Learning. MIT Press (1997)
[13] Gudbjartsson, D.F., Thorvaldsson, T., Kong, A., Gunnarsson, G., Ingolfsdottir, A.: Allegro version 2. Nature Genetics 37(10), 1015–1016 (2005)
[14] Gusev, A., Lowe, J.K., Stoffel, M., Daly, M.J., Altshuler, D., Breslow, J.L., Friedman, J.M., Pe'er, I.: Whole population, genome-wide mapping of hidden relatedness. Genome Research 19, 318–326 (2009), doi:10.1101/gr.081398.108
[15] Henn, B.M., Hon, L., Macpherson, J.M., Eriksson, N., Saxonov, S., Pe'er, I., Mountain, J.L.: Cryptic distant relatives are common in both isolated and cosmopolitan genetic samples. PLoS ONE 7(4), e34267 (2012)
[16] Ingólfsdóttir, A., Gudbjartsson, D.: Genetic Linkage Analysis Algorithms and Their Implementation. In: Priami, C., Merelli, E., Gonzalez, P., Omicini, A. (eds.) Transactions on Computational Systems Biology III. LNCS (LNBI), vol. 3737, pp. 123–144. Springer, Heidelberg (2005)
[17] Kyriazopoulou-Panagiotopoulou, S., Kashef Haghighi, D., Aerni, S.J., Sundquist, A., Bercovici, S., Batzoglou, S.: Reconstruction of genealogical relationships with applications to phase iii of hapmap. Bioinformatics 27(13), i333–i341 (2011)

[18] Lander, E.S., Green, P.: Construction of multilocus genetic maps in humans. Proceedings of the National Academy of Sciences 84, 2363–2367 (1987)
[19] Li, M.-H., Strandén, I., Tiirikka, T., Sevón-Aimonen, M.-L., Kantanen, J.: A comparison of approaches to estimate the inbreeding coefficient and pairwise relatedness using genomic and pedigree data in a sheep population. PLoS ONE 6(11), e26256 (2011)
[20] Markianos, K., Daly, M.J., Kruglyak, L.: Efficient multipoint linkage analysis through reduction of inheritance space. Am. J. Hum. Genet. 68(4), 963–977 (2001)
[21] 1000 Genomes Project. A map of human genome variation from population-scale sequencing. Nature 467(7319),1061–1073 (2010)
[22] Moltke, I., Albrechtsen, A., Thomas, Nielsen, F.C., Nielsen, R.: A method for detecting IBD regions simultaneously in multiple individuals with applications to disease genetics. Genome Research 21(7), 1168–1180 (2011)
[23] Nalls, M.A., Simon-Sanchez, J., Gibbs, J.R., Paisan-Ruiz, C., Bras, J.T., Tanaka, T., Matarin, M., Scholz, S., Weitz, C., Harris, T.B., Ferrucci, L., Hardy, J., Singleton, A.B.: Measures of autozygosity in decline: Globalization, urbanization, and its implications for medical genetics. PLoS Genet 5(3), e1000415 (2009)
[24] Ott, J.: Analysis of Human Genetic Linkage. The Johns Hopkins series in contemporary medicine and public health. Johns Hopkins University Press (1999)
[25] Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., Sham, P.C.: PLINK: a tool set for whole-genome association and population-based linkage analyses. American Journal of Human Genetics 81(3), 559–575 (2007)
[26] Rabiner, L.R.: A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE, 257–286 (1989)
[27] Ralph, P., Coop, G.: The geography of recent genetic ancestry across Europe (July 2012)
[28] WTCCC. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447(7145), 661–678 (2007)