Detection of short identity by descent segments using low-frequency variants

Chonghao Wang[1], Werner Pieter Veldsman[1], Lu Zhang[1,2*]

[1]Department of Computer Science, Hong Kong Baptist University, Hong Kong SAR, China

[2]Institute for Research and Continuing Education, Hong Kong Baptist University, Shenzhen, China

*Correspondence and requests for materials should be addressed to Lu Zhang (email: ericluzhang@hkbu.edu.hk)

Abstract

Identity by descent (IBD) segments are inherited genomic sequences from the same common ancestors that serve as biomarkers during lineage analysis, disease mutation mapping, and within broader fields of population genetics. Current computational tools used for the detection of IBD segments focus on identifying long IBD segments, with short IBD segments (i.e. shorter than 2 cM) often going undetected. Here, we present SILO, a method with a remarkably improved ability to detect short IBD segments as compared to state-of-the-art IBD detection tools. SILO detects IBD segments by considering both common and low-frequency variants (LFVs), where the probability of the presence of a LFV minor allele is assumed to follow a Bernoulli distribution with a Beta prior. SILO was benchmarked against GERMLINE2, hap-IBD, HapFABIA, Parente, and TRUFFLE on simulated data and on a real pedigree from the 1,000 Genomes Project. The results show that SILO has an unrivaled ability to detect short IBD segments, while its ability to detect long IBD segments is similar to that of contemporary tools in the field.

Introduction

Each individual inherits a haploid copy of their full chromosomal complement from each of their parents. If two individuals happen to share a genomic segment by random chance, then the two segments are said to be identical by state (IBS). Furthermore, if these two segments were inherited from the same recent common ancestor, then the two segments are said to be identical by descent (IBD). IBD segments can be used in many different applications. For

example, IBD segments among individuals in a population could be used to explore the demographic history of the population [1, 2], be used to locate disease susceptible variants or genes [3, 4, 5, 6] and be used to phase genomes [7, 8].

Many computational tools (**Table 1**) have been developed to detect IBD segments [9], either using single nucleotide polymorphism (SNP) microarray or next generation sequencing data. In general, there are two types of tools for identifying IBD segments. The first type searches for identical haplotypes or non-conflicting genotypes exceeding a certain segment length threshold. This type of tools is usually computationally efficient and can be applied to datasets consisting of thousands of individuals or even an entire biobank dataset [10, 11]. RaPID [12], for example, uses the Positional Burrows-Wheeler Transform to identify potential IBD segments and applies random projections to allow for some genotype errors. GERMLINE2 [13] employs hash functions to detect candidate IBD segments rapidly and searches the regions flanking each candidate segment to determine whether the segment could be extended. iLASH [14] utilizes a locality-sensitive hashing technique to find IBD segments instead of the conventional hashing methods used by GERMLINE [15]. Hap-IBD [10] also uses the Positional Burrows-Wheeler Transform to locate candidate IBD segments, and it uses a seed-and-extend algorithm to extend candidate IBD segments in order to obtain longer IBD segments. IBIS [16] divides a chromosome into many small windows. It reports a set of contiguous windows as an IBD segment if the windows are long enough with limited mismatches. TRUFFLE [17] searches for long segments with non-conflicting SNP genotypes along a chromosome while allowing some genotype errors, and it reports these segments as IBD.

The second type of IBD detection tools uses probabilistic models to infer the probability of a segment being IBD. BEAGLE IBD [18] utilizes a hidden Markov model to infer IBD segments. Parente [19] uses a naive Bayes model to calculate the posterior probabilities of genotypes in a window under the assumption that it is either IBD or non-IBD. Parente2 [20] offers two major improvements on Parente. First, Parente2 considers linkage disequilibrium (LD) among the markers in each window by incorporating haplotype frequencies, while Parente assumes independence on such markers. Second, Parente2 adds two additional structures, the augmented windows and window subsets, into the model framework, which makes the model more nuanced.

Most IBD detection tools based on haplotype information use common variants (e.g. SNPs with minor allele frequencies (MAFs) above 5%) to detect IBD segments, as it is difficult to phase low-frequency variants (LFVs) and rare variants by population LD information. In contrast, genotype-based IBD detection tools generally do not have a strict restriction on MAFs, e.g. TRUFFLE. However, they either treat LFVs and rare variants in a similar manner to how they treat common variants, or they retain only common variants during quality control because LFVs and rare variants are believed to be less informative in long IBD segment detection.

As Zhou *et al.* [10] stated, the existing tools can correctly detect IBD segments with lengths longer than 2 cM, but they may not be able to show sufficient power at a low false positive rate to detect short IBD segments. Short IBD segments are, however, important since they can be used in various applications, such as imputing missing genotypes and inferring haplotypes, especially where there are no long IBD segments. Short IBD segments can also be useful in identifying private or low-frequency disease causal variants in a pedigree, where linkage analysis has insufficient power to prioritize the inherited disease causal mutations for nuclear families and incomplete large pedigrees. Moreover, existing IBD mapping strategies rely only on long IBD segments shared by close relatives of patients due to a limited ability to identify short IBD segments, but this may result in a huge number of candidate mutations required to be further validated. In such situations, short IBD segments shared by the distant relatives of

patients are more informative in narrowing down mutation searching and validation space, such as the situations in autosomal dominant Alzheimer's disease [21] and hereditary deafness [22].

Previous studies suggested that the use of LFVs might not be suitable for detecting IBD segments because what what appeared to be LFVs might in actual fact be artifacts that arose from genotype errors [10]. However, significant advancements in sequencing technology have led to substantial reductions in genotype error rates. For example, as Ma *et al.* [23] highlighted, the genotype error rates could be suppressed to 10^{-5} to 10^{-4}. This magnitude of the reduction of errors significantly reduces both the false positive rates and false negatives rates when utilizing LFVs. In our opinion, there are specific scenarios during IBD segment detection in which LFVs would not only be suitable but also advantageous. First, if two individuals share minor alleles of multiple LFVs within a short IBS segment, it is highly probable that this segment is IBD, as the probability of all variants encountering genotype errors is low. Second, if the alleles of surrounding common variants are concordant, this indicates that the segment is more likely to be IBD. HapFABIA [24] is an example of a computational tool that uses rare variants in a biclustering technique to detect short IBD segments. It assumes short IBS regions shared by more individuals are more likely to be IBD, but this limits its ability to be applied to a dataset of a small population or a pedigree with few numbers. Additionally, common variants could also be used in conjunction with LFVs to improve the sensitivity and reduce false positives in short IBD segment detection.

To address these shortcomings, we implemented SILO (detection of **S**hort **I**dentity by descent segments using **LO**w-frequency variants), a novel algorithm to detect long and short IBD segments by considering both common variants and LFVs. SILO employs an approach similar to that used by Parente2 to identify long IBD segments (e.g. lengths >= 2 cM) using common variants, but adapts the approach to detect short IBD segments using LFVs (**Methods**). SILO assumes that the probability of the presense of a LFV minor allele follows a Bernoulli distribution, with its parameter modeled by a Beta distribution. SILO implements a seed-and-extend algorithm to identify short IBD segments and implements an cleaning procedure to remove false positive IBD segments. We compared SILO to five state-of-the-art IBD detection tools (hap-IBD, GERMLINE2, TRUFFLE, Parente and HapFABIA) on simulated data and on a real pedigree from the 1000 Genomes Project. SILO demonstrated a significantly superior ability in detecting short IBD segments in simulated data with an average F_1 score that was 13 times higher than that obtained by the nearest competitor for those IBD segments between 0.1 to 0.2 cM. In the exercise on a real three-generation pedigree from the 1000 Genomes Project, SILO also identified many validated short IBD segments which were missed by the other tools. These results show that SILO has a ground-breaking ability to detect short IBD segments.

Results

An overview of SILO

SILO employs a two-stage procedure to identify both long and short IBD segments between a pair of individuals. In the first stage, SILO identifies long IBD segments based on common variants. The common variants in one chromosome are divided into contiguous windows (without overlap) with the same number of common variants. SILO uses inner log-likelihood ratios (inner-LLRs) to determine the likelihood of a window being IBD versus non-IBD. Due to high variances of inner-LLRs within certain windows, SILO converts inner-LLRs to outer log-likelihood ratios (outer-LLRs) using empirical distributions. A sliding block (one window for each step) is defined to include several contiguous windows until their total length equals

or slightly exceeds the pre-defined block length threshold (3 cM by default). The block outer-LLR is obtained by combining the outer-LLRs from all the included windows. When the block outer-LLR surpasses a predefined threshold value (T_{LLR}), it signifies that this specific block is identified as an IBD segment between the two individuals.

In the second stage, SILO seeks to detect short IBD segments using a seed-and-extend algorithm (**Figure 1C**). A seed is defined as a window (i) with shared minor alleles of LFVs between two individuals, and (ii) with a combined-LLR (the window outer-LLR and LFV LLRs) no less than T_{LLR}. SILO extends from the seed to its surrounding windows and terminates until certain conditions are met (**Methods**). Moreover, SILO employs a cleaning procedure to remove potentially false positive IBD regions based on discordant alleles of LFVs. The workflow of SILO is shown in **Figure 1**.

The performance of SILO on simulated data

We used Fastsimcoal2 [25, 26] to simulate 1,000 individuals as the reference population and 150 individuals as the test set. We removed latent IBD segments of the tested pairs in the test set and then implanted short IBD segments into them for evaluating the prediction performance of IBD detection tools. For the short IBD segments, to investigate how each IBD detection tool performed on different IBD length scales, we selected four different ranges: 0.1 cM-0.2 cM, 0.2 cM-0.5 cM, 0.5 cM-1 cM and 1 cM-2 cM. For each range, one, three, five, and seven LFVs were inserted into these IBD segments respectively. This was done in order to explore their impacts on the performance of SILO (see section **The cumulative impact of LFVs**). In total, there were 16 different experiments, and each experiment had 100 replicates. From **Figure 2**, it was observed that SILO showed significant improvements over the other tools in the detection of very short IBD segments. In the experiments on the shortest (0.1-0.2 cM) IBD segments (**Figure 2A**), all tools underperformed compared to SILO. Hap-IBD, which performed the best among the other tools, could merely get a mean F_1 score of 0.06. In contrast, SILO obtained a mean F_1 score of 0.8. As the length of IBD segments became larger, the other tools displayed higher F_1 scores. Although there were notable improvements compared to the IBD length of the 0.1 cM-0.2 cM range, the other tools still had difficulties in reliably finding 0.2 cM-0.5 cM IBD segments, where the F_1 score of the best performing method was 0.39, while SILO achieved a F_1 score in excess of 0.8 (**Figure 2B**). When the IBD segment lengths exceed 0.5 cM, some of the other tools show F_1 scores that are comparable to those that are obtained by SILO. Specifically, the mean F_1 score of hap-IBD in the experiment on IBD segments with a length in the range 0.5 cM-1 cM was 0.86, and the mean F_1 scores of hap-IBD, GERMLINE2 and Parente in the experiment of IBD segments with a length in the range 1 cM-2 cM were 0.92, 0.84 and 0.91, respectively (**Figure 2C** and **Figure 2D**). This trend reflects that the detection power of LFVs becomes milder as the IBD segments become longer.

Apart from F_1 score, we also investigated differences in IBD segment length between inferred and true IBD segments in the simulated data for each tool. As shown in **Figure 3**, we observed results that are consistent with those we had obtained during F_1 scoring. When the difference between inferred and true segment length for a method was closer to 0, the method had a higher mean F_1 score. For example, the length differences of SILO in the case of 0.1 cM-0.2 cM were closer to 0 than the other tools (**Figure 3A**), and the accompanying mean F_1 score of SILO was also higher than others. The observation that a decrease in length difference has an inverse relationship with the F_1 score validated our choice of scoring metric since a decrease in length difference implies higher accuracy.

Unlike hap-IBD, GERMLINE2 and Parente that achieved at least some inference accuracy,

TRUFFLE and HapFABIA both got F_1 scores of 0 in all experiments. It was to be expected that TRUFFLE would obtain such results as it utilizes genotype but not haplotype information and was thus limited to the detection of long IBD segments. In contrast, HapFABIA, that could utilize genotypes or haplotypes, was supposed to be able to detect short IBD segments. We investigated the cause of its inability to detect short IBD segments and found that it could not detect IBD segments when a test set contained a relatively small number of individuals. This can be due to HapFABIA having difficulties in classifying rare tagSNPs defined by its model when there are only a few individuals. On the other hand, as pointed out by Hochreiter [24], HapFABIA reported short IBD segments on a majority basis whereby it regarded a segment as IBD with a measure of confidence that was determined by using the number of individuals sharing it in a dataset. This approach may become a limitation to HapFABIA when there are only a few individuals in a dataset.

The cumulative impact of LFVs

The usage of LFVs to detect short IBD segments is a primary feature of SILO. We therefore investigated the influence of different numbers of LFVs on short IBD segment detection in the simulated data. **Figure S1** showed that multiple LFVs in a candidate short IBD segment improved prediction performance, but the improvements plateaued as the number of LFVs increased. For example, the mean F_1 scores when inserting one, three, five, and seven LFVs in 0.1 cM-0.2 cM IBD segments were 0.804, 0.853, 0.855 and 0.857, respectively. The implication is that when the number of LFVs reaches a saturation value (3 in this example), there is sufficient information to detect a short IBD segment, and additional LFVs add less value to the prediction. We furthermore found that the cumulative influence of LFVs was eroded as the lengths of IBD segments increased. For instance, the mean F_1 scores when inserting one, three, five, and seven LFVs in 1 cM-2 cM IBD segments were nearly identical instead of reaching a saturation value. This demonstrated that LFVs were less informative in longer IBD regions. A similar conclusion could be drawn when examining length differences between inferred and simulated IBD segments. As shown in **Figure S2**, the length difference between inferred IBD segments and simulated IBD segments was closer to 0 when there were more LFVs within the segments. The presence of LFVs therefore has a positive impact on segment length inference accuracy.

The performance of SILO on real data

Because it was difficult to determine the ground truth for IBD segments in real data, we selected six members in a pedigree from the 1000 Genomes Project [27, 28] as the test set and evaluated the performance of IBD detection tools based on the known relationships of individuals (**Figure S3**). We addressed two experimental concerns during the evaluation of tools' performance on real data. The first concern was related to the proportion of IBD segments in the autosomes given the relationship between two individuals. Because the relationships of the six individuals in the pedigree were known, the expected proportion of IBD segments could be derived [29]. **Figure 4A** showed that the proportions obtained by SILO generally matched the expected proportions. There was a noticeable difference between the expected and inferred proportions of IBD segments of NA19660-NA19686 and NA19661-NA19686. However, as shown in **Table S1**, the other tools showed similar trends with SILO on proportionality. For example, the proportions of GERMLINE2 for the two pairs just mentioned were 0.331 and 0.647, respectively, and the proportions of Parente for them were 0.325 and 0.678, respectively.

This suggested that either the true proportions for these two pairs differed substantially from their expected proportions or that all tools (including SILO) had difficulties in inferring IBD segments for these two pairs.

The second experimental concern was related to short IBD segment detection performance. In contrast to the first concern that mostly dealt with long IBD segments, here we focused on studying the ability of SILO over the other tools to detect short IBD segments. In order to confirm whether the inferred short IBD segments were truly IBD or not, we chose the pair NA19662-NA19686 for analysis. Our aim was to determine where NA19685 carried minor alleles of LFVs in the inferred short IBD segments, and where NA19684 did not. By carrying out the preceding analysis on inferred short IBD segments, we were able to determine whether these segments were inherited from a common ancestor (i.e. NA19660 or NA19661). There were 150 inferred short IBD segments by SILO shared between NA19662 and NA19686 across all autosomes where NA19685 carried minor alleles while NA19684 did not. We extracted 93 short IBD segments that were most probable to be true from the 150 segments. We furthermore sought to determine how many of the 93 segments could be detected or covered by the other tools (**Figure 4B**). Among the other tools, GERMLINE2 could only cover 14% of the short IBD segments detected by SILO ($>= 80\%$ overlap was required to call a segment as covered). Hap-IBD, Parente and TRUFFLE all covered only 1% of the short IBD segments detected by SILO. HapFABIA could not cover any short IBD segments detected by SILO, likely due to the previously mentioned sample size issue. To ensure that our ground truth assumption was unbiased and did not favor SILO over the other tools, we extracted IBD segments inferred by the other tools, applied the same selection procedures, and examined how many segments could reciprocally be found by SILO. The corresponding results are shown in **Figure S4**-**S7**. In contrast to the limited ability of the other tools to detect short IBD segments inferred by SILO, SILO could detect most IBD segments inferred by the other tools.

In addition to the benchmark study, we also evaluated SILO's ability to detect real short IBD segments in a practical context using chromosome 22. SILO inferred that there was a short IBD segment of around 0.005 cM in length shared between NA19662 and NA19686 (**Figure 5**). This segment was near the end of chromosome 22, and there were seven SNVs with discordant alleles upstream of the segment in a 0.15 cM region shown in **Figure 5**. The presence of these variants prohibited the extension of the short IBD segments to a long IBD segment. In **Figure 5**, the haplotypes of each pedigree member in the inferred IBD segment were shown (**Methods**). We observed that NA19660, NA19662, NA19685 and NA19686 all carried a minor allele of a LFV in one of their haplotypes, while NA19661 and NA19684 did not carry the minor alleles. Furthermore, the four haplotypes that carried minor alleles were identical. Therefore, we could confirm that NA19662 and NA19686 both inherited this haplotype from NA19660, which meant the inferred short IBD segment was a true IBD segment.

The influence of hyperparameters

There are several hyperparameters that can influence the performance of SILO. For the *window size*, it was shown in **Figure S8** that incrementing *window size* improved results towards an optimal *window size* (8-15 common variants) beyond which a decrease in performance was seen again. The hyperparameters T_N^{neg} and T_P^{neg} (thresholds of the numbers (N) and proportions (P) of windows with negative outer-LLRs in the seed-and-extend algorithm) affected the determination of the boundary points of inferred short IBD segments. The inferred short IBD segments of NA19662 and NA19686 were extracted, and classified as true or false as described in the section **SILO's performance on real data**. The precision was defined as the number of

true inferred short IBD segments divided by the number of all inferred short IBD segments. As presented in **Table S2**, we observed that when T_N^{neg} and T_P^{neg} became higher, the mean lengths of inferred short IBD segments became longer. However, an increase in T_N^{neg} and T_P^{neg} was accompanied by a decrease in precision. In essence, SILO tends to infer partial short IBD segments when the two thresholds are low, while it tends to expand inferred IBD segments, which potentially leads to more false positives, when the two thresholds are high. Therefore, a balance should be struck when choosing T_N^{neg} and T_P^{neg}.

Discussion

The identification of short IBD segments has long been a challenge in the field of population genetics. Given that a coincidental sharing of minor alleles of LFVs is less probable compared to common variants, LFVs enable a more reliable identification of short IBD segments. However, the limitations of high genotype error rates in the past [30] hindered the effective utilization of LFVs. Fortunately, with the advances in sequencing technology, the genotype error rates have reduced significantly [23]. As a result, the probability of sharing minor alleles of LFVs becomes much higher than that of encountering genotype errors. In light of these developments, we present SILO to address this problem by incorporating LFVs. We demonstrated that SILO significantly improved short IBD segment detection over existing tools in both simulated and real datasets.

SILO intrinsically considers the impacts of genotype errors. If the genotype error rates are high, SILO will put smaller weights on the observed genotypes and increase weights on other possible genotypes. We investigated the impacts of different genotype error rates on SILO. With a genotype error rate of 1%, SILO did not fully utilize the information from LFVs because there was a higher chance that the sharing of minor alleles of LFVs was attributed to genotype errors. Furthermore, it was observed that a genotype error rate of 0.1% could generally make SILO extract sufficient information from LFVs rather than regarding them as genotype errors, and a genotype error rate of 0.001% was enough for SILO to fully utilize the information from LFVs. Thus, SILO can be effectively applied to datasets generated by contemporary sequencing technology platforms.

In addition to the quality of LFVs, it is also important to consider the influence of MAF intervals of these variants. A longer MAF interval is supposed to contain more LFVs, and SILO will thus detect more short IBD candidates. An important caveat is that variants with higher MAFs have a higher possibility of occurring by chance, which can potentially increase the false positive rate. To assess the impact of different MAF intervals, we tested 2 different intervals: 0.5% to 1% and 0.5% to 5%. The precision of the 2 intervals were 0.75 and 0.61, respectively, in the context of short IBD segments detection between NA19662 and NA19686. Furthermore, the numbers of corresponding inferred short IBD segments were 132 and 919, respectively. We could observe that a longer interval with larger MAFs introduced more false positives. Based on these findings, we recommend setting the interval of LFVs to be between 0.5% and 1%.

A challenging problem in IBD segment detection is the determination of IBD segment boundary points. When SILO detects long IBD segments, the minimum length of an IBD segment is the block length. As a block consists of several windows, it is possible that the windows near the block boundary are non-IBD (e.g. IBS segments). However, if the summation of outer-LLRs of all windows in the block exceeds the threshold, the entire block is considered an IBD segment. For short IBD segments detection, the seed-and-extend algorithm may also

include some non-IBD regions. To mitigate this issue, SILO employs a cleaning procedure. This procedure utilizes LFVs with discordant alleles to identify non-IBD regions surrounding them. If inferred IBD segments overlap with these non-IBD regions, the overlapping parts are considered false positives and are removed from the inferred IBD segments.

SILO makes a significant contribution to the field by improving the detection on short IBD segments. Our results demonstrated that SILO outperformed state-of-the-art tools in the identification of short IBD segments and it achieved comparable performance with these tools in detecting longer IBD segments. In practice, SILO is recommended to be employed in small populations or pedigrees. Additionally, it can serve as a follow-up IBD detection tool after faster tools such as hap-IBD, to identify IBD segments in a small population of interest within a larger population.

Methods

Calculation of the outer log-likelihood ratios of common variants

SILO detects long IBD segments based on common variants (MAF>=5%). As illustrated in **Figure 1**, a chromosome is split into contiguous windows (without overlap) and each window contains n common variants. A sliding block (one window per step) is built on top of the windows, and its length is measured in centimorgan (3 cM by default). The windows are incorporated in a block until the block length equals or slightly exceeds 3 cM. SILO refers to the similar framework of Parente2 to identify long IBD segments, while it modifies the way to split bins in empirical distributions and removes the augmented windows and window subsets in Parente2 to allow the extension of windows with LFVs for short IBD segment detection.

The probabilities of observing $g(w)$ and $g'(w)$ from IBD and non-IBD segments are defined as

$$p(g(w),g'(w)|IBD) = \sum_{h_1,h_2,h_3} p(g(w)|h_1,h_2)p(g'(w)|h_2,h_3)f_w(h_1)f_w(h_2)f_w(h_3), \quad (1)$$

$$p(g(w),g'(w)|non-IBD) = \sum_{h_1,h_2,h_3,h_4} p(g(w)|h_1,h_2)p(g'(w)|h_3,h_4)f_w(h_1)f_w(h_2)f_w(h_3)f_w(h_4), \quad (2)$$

where $p(g(w)|h_1,h_2)$ is defined as

$$p(g(w)|h_1,h_2) = \prod_{i\in m(w)} p(g(i)|a_i^1,a_i^2), \quad (3)$$

$$p(g(i)|a_i^1,a_i^2) = \begin{cases} 1-\varepsilon & g(i) = a_i^1 + a_i^2 \\ \frac{\varepsilon}{2} & \text{otherwise.} \end{cases} \quad (4)$$

We denote $g(w)$ and $g'(w)$ as the genotypes of common variants in window w of two individuals. The genotypes are represented as 0, 1 and 2 based on the number of major alleles. The haplotypes of two individuals can be denoted as (h_1,h_2) and (h_2,h_3), if they inherited h_2 from the same ancestor as IBD segments; otherwise, they are represented as (h_1,h_2) and (h_3,h_4). $m(w)$ is a set of all common variants included in window w, a_i^1 and a_i^2 represent the alleles of the common variant i from h_1 and h_2. $f_w(h_i)$ ($i \in \{1,2,3,4\}$) is the frequency of h_i in window w, which is estimated in the available haplotypes from the same population (taken as the reference population). ε is the genotype error rate of common variants (default $\varepsilon = 0.5\%$). $p(g'(w)|h_2,h_3)$ and $p(g'(w)|h_3,h_4)$ can be defined in the same way. The inner-LLR

($\gamma_w(g(w), g'(w))$) of the window w is used to compare the magnitude of $p(g(w), g'(w)|IBD)$ and $p(g(w), g'(w)|non-IBD)$ and show the degree of differences between them, and it is defined as

$$\gamma_w(g(w), g'(w)) = log(\frac{p(g(w), g'(w)|IBD)}{p(g(w), g'(w)|non-IBD)}). \tag{5}$$

Because there can be some windows in a block that have high variances for $\gamma_w(g(w), g'(w))$ and thus make the inner-LLR of the block sensitive to such factor, we generate two empirical distributions (Q_I and $Q_{\bar{I}}$) for $\gamma_w(g(w), g'(w))$ under the assumption of IBD and non-IBD segments, respectively. Q_I and $Q_{\bar{I}}$ are generated by two groups of $\gamma_w(g(w), g'(w))$ values for IBD and non-IBD segments. We split these $\gamma_w(g(w), g'(w))$ values as non-overlapping bins to approximate the probability density function of Q_I and $Q_{\bar{I}}$. We consider the values in ($-$inf, -18) as bin_L and in (18, inf) as bin_H for extreme values, and the remaining bins (bin size=2) span over [-18, 18]. We define outer-LLR ($\lambda_w(g(w), g'(w))$) as

$$\lambda_w(g(w), g'(w)) = log(\frac{Q_I(\gamma_w(g(w), g'(w)))}{Q_{\bar{I}}(\gamma_w(g(w), g'(w)))}). \tag{6}$$

The inner-LLR (γ_b) and outer-LLR (λ_b) of a block are calculated by summing up the corresponding value for each involved window due to the independence assumption among these windows.

Calculation of the log-likelihood ratios of LFVs

We assume LFVs are in trivial LD with the other variants. Therefore, we calculate probabilities and LLR for individual LFVs instead of windows. The probabilities of observing $g(m)$ and $g'(m)$ from IBD and non-IBD segments are

$$p(g(m), g'(m)|IBD) = \sum_{h_1,h_2,h_3} p(g(m)|h_1,h_2)p(g'(m)|h_2,h_3)p(h_1,h_2,h_3|D), \tag{7}$$

$$p(g(m), g'(m)|non-IBD) = \sum_{h_1,h_2,h_3,h_4} p(g(m)|h_1,h_2)p(g'(m)|h_3,h_4)p(h_1,h_2,h_3,h_4|D), \tag{8}$$

where $g(m)$ and $g'(m)$ as the genotypes of LFV m of two individuals, $p(h_1,h_2,h_3|D)$ and $p(h_1,h_2,h_3,h_4|D)$ stand for the joint probabilities of haplotypes from two individuals given D. D represents the available haplotypes from the same population, such as CEU in the 1000 Genomes Project. For common variants, the haplotype joint probabilities could be approximated by the product of haplotype frequencies [20]. This strategy is not effective for LFVs due to their MAFs being low and inaccurately estimated in the population. We assume the probability of the presence of a LFV minor allele follows a Bernoulli Distribution,

$$\begin{aligned} p(h_1,h_2,h_3|D) &= \int_0^1 p(h_1,h_2,h_3|\theta,D)p(\theta|D)d\theta \\ &= \int_0^1 p(h_1|\theta,D)p(h_2|\theta,D)p(h_3|\theta,D)p(\theta|D)d\theta \\ &= \int_0^1 \theta^{h_1}(1-\theta)^{1-h_1}\theta^{h_2}(1-\theta)^{1-h_2}\theta^{h_3}(1-\theta)^{1-h_3}p(\theta|D)d\theta. \end{aligned} \tag{9}$$

θ is a latent variable that estimates the parameter of the Bernoulli distribution. $p(\theta|D)$ follows a beta distribution and it can be expanded as

$$p(\theta|D) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}, \tag{10}$$

where α and β are the numbers of major and minor alleles of the reference population and test sets for the LFV m. The LLR τ_m of the LFV m is defined as

$$\tau_m(g(m), g'(m)) = log(\frac{p(g(m), g'(m)|IBD)}{p(g(m), g'(m)|non-IBD)}) \tag{11}$$

Determination of IBD segments

IBD segments can be identified by considering the LLRs of common variants (λ_b) and LFVs (τ_m). If λ_b is larger than T_{LLR} (by default 3), the block is considered to be a long IBD segment. SILO applies a seed-and-extend algorithm (**Figure 1C**) to identify short IBD segments using both common variants and LFVs. In the seed-and-extend algorithm, each LFV is assigned to a window that covers it or nearest to it. We define the combined-LLR ϕ_w of a window w as

$$\phi_w = \begin{cases} \lambda_w + \Sigma_{i \in v(w)} \tau_i & \text{if } v(w) \text{ is not an empty set,} \\ \lambda_w & \text{otherwise.} \end{cases} \tag{12}$$

$v(w)$ is the set of all LFVs in the window w. A seed is defined as a window w if $\phi_w >= T_{LLR}$ and w contains LFVs with $\Sigma_{i \in v(w)} \tau_i > 1$. During the extension, SILO will first consider the left flanking window and then the right flanking window (**Figure 1C**). The extension procedure terminates if one of the following criteria is satisfied: (i) the summation of combined-LLRs of the involved windows becomes smaller than T_{LLR}; (ii) the number of windows with negative λ_w exceeds the threshold T_N^{neg}; (iii) the proportion of windows with negative λ_w exceeds the threshold T_P^{neg}; (iv) the ϕ_s of the newly added window s is smaller than -3; (v) the extended region is longer than 2 cM (since long IBD segments are supposed to be detected by common variants). If the extension is stopped in one direction, the extension in the opposite direction continues until it also meets one of the above criteria. The final post-extension region is regarded as a short IBD segment.

Removal of false positive IBD segments

SILO includes a cleaning procedure to remove potential false positive IBD segments using a seed-and-extend algorithm. A seed is defined as a window w if $\phi_w < T_{LLR}$ and w contains LFVs with $\Sigma_{i \in v(w)} \tau_i < 0$. The extension procedure is similar to the one described above with different termination conditions, and it is terminated if one of the following criteria is satisfied: (i) the summation of combined-LLRs of the involved windows becomes larger than or equal to T_{LLR}; (ii) the λ_s of the newly added window s is larger than -1; (iii) the extended region is longer than 1 cM. The overlaps of the extended regions and IBD segments are the false positive regions. SILO removes these false positive regions from the inferred IBD segments.

Generation of simulated data

Fastsimcoal2 [25, 26] was used to generate raw SNVs following the similar strategy used in Refined IBD [31]. In this study, we simulated a genomic region of 30 Mb. The effective population size in the current generation was 24.3 million. We sampled 1,000 individuals (2,000 haplotypes) in the current generation as the reference population, and sampled another 150 individuals (300 haplotypes) in the current generation as a test set. The mutation rate was set to 2.5×10^{-8} and the recombination rate was 10^{-8} (1 Mb = 1 cM). There was no transition bias in simulation. We only kept common variants (MAF>= 5%) in the reference population.

Since latent IBD segments might exist in the test set, which could introduce unexpected noises in performance evaluations, we adjusted the strategy used by Parente2 [20] to break up those IBD segments by dividing the genomic sequences into many small contiguous windows, each with a length of 0.05 cM. Briefly, we shuffled the haplotypes in each window of the individuals in the test set, thereby generating 150 composite individuals, where the two haplotypes of each may come from different individuals. We assumed no IBD segments existed in these composite individuals, and they replaced the original individuals in the test set.

We selected a pair of individuals from the test set and transplanted a haplotype segment (0.1 cM to 2 cM) at a random position from one individual to the other. LFVs were inserted into these IBD segments, with MAFs randomly selected between 0.5% to 1%. For each LFV, the shared haplotype between two individuals was assigned a minor allele, while the other allele for each individual was determined based on its MAF. We also manually introduced genotype errors with a rate of 0.5% for common variants and 0.001% for LFVs. We simulated 100 times for each combination of the different length intervals of IBD segments (0.1 cM-0.2 cM, 0.2 cM-0.5 cM, 0.5 cM-1 cM, and 1 cM-2 cM) and the number of LFVs (1, 3, 5 and 7).

IBD inference using a real pedigree from the 1000 Genomes Project

We downloaded the haplotypes (GRCh38) of the individuals in the 1000 Genomes Project from The International Genome Sample Resource (IGSR) [27, 28]. A pedigree of Mexican descent was selected, which consisted of six members. Their relationship to one another are represented in **Figure S3** based on the pedigree information from `https://www.internationalgenome.org`. We also selected individuals that were unrelated and who had annotated population codes of either GBR, CEU, TSI, FIN, IBS, MXL or PUR to constitute the reference population for SILO. All individuals belonging to the Mexican pedigree were removed from the reference population. As a result, we retained 674 individuals in the reference population. For quality control, we removed SNVs with missing value rates exceeding 10% and kept only bi-allelic SNVs. In addition, we only kept common variants (MAF$\geq$5%) and LFVs (0.5%$\leq$MAF$\leq$ 1%).

We compared the inferred IBD segments shared by NA19662 and NA19686 from different tools. Specifically, we considered the inferred long IBD segments (>2 cM) to be true if the following requirements were all satisfied: (i) there existed a nearly identical haplotype (fewer than 0.5% common variants and LFVs with discordant alleles in total) between NA19662 and NA19686; (ii) the shared haplotype was observed in NA19685 but did not exist in NA19684; (iii) the shared haplotype was observed in either NA19660 or NA19661. Besides the three requirements, we added an additional requirement for short IBD segment validation: (iv) the shared haplotype between NA19662 and NA19686 should contain the minor allele of at least one LFV, while the haplotypes of NA19684 did not.

Benchmark framework

We benchmarked the performance of SILO against five IBD detection tools: hap-IBD [10], GERMLINE2 [13], TRUFFLE [17], Parente [19] and HapFABIA [24]. The relevant software versions are listed in **Table S3**. We conducted a hyperparameter analysis for these tools and selected the optimal combination of hyperparameters for each tool as indicated by the highest F_1 score in the simulated data (**Table S4-S7**). We then comparatively evaluated the performance based on F_1 score and length differences between inferred and simulated IBD segments. We used the recommended or default hyperparameters of each tool to identify IBD segments of the Mexican pedigree (**Table S8**), and evaluated the performance with the following metrics:

(i) the comparison of inferred and expected IBD segment proportions; (ii) the proportion of validated inferred IBD segments of each tool that were also detected by each of the other tools.

Data availability

Simulated data were generated by fastsimcoal2 (`http://cmpg.unibe.ch/software/fastsimcoal27/`). The haplotypes from the 1000 Genomes Project are available at `http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/working/20220422_3202_phased_SNV_INDEL_SV/`.

Code availability

SILO's source code and documentation are available at `https://github.com/ericcombiolab/SILO`.

References

[1] Pier Francesco Palamara et al. "Length distributions of identity by descent reveal fine-scale demographic history". en. In: *Am. J. Hum. Genet.* 91.5 (Nov. 2012), pp. 809–822.

[2] Edmund Gilbert, Ashwini Shanmugam, and Gianpiero L. Cavalleri. "Revealing the recent demographic history of Europe via haplotype sharing in the UK Biobank". In: *Proceedings of the National Academy of Sciences* 119.25 (2022), e2119281119. DOI: 10.1073/pnas.2119281119. URL: https://www.pnas.org/doi/abs/10.1073/pnas.2119281119.

[3] Sharon R Browning and Elizabeth A Thompson. "Detecting rare variant associations by identity-by-descent mapping in case-control studies". en. In: *Genetics* 190.4 (Apr. 2012), pp. 1521–1531.

[4] Eric Letouzé et al. "Identity by descent mapping of founder mutations in cancer using high-resolution tumor SNP data". en. In: *PLoS One* 7.5 (May 2012), e35897.

[5] Dingge Ying et al. "HaploShare: identification of extended haplotypes shared by cases and evaluation against controls". en. In: *Genome Biol.* 16.1 (May 2015), p. 92.

[6] Wen-Chi Hsueh et al. "Identity-by-descent mapping identifies major locus for serum triglycerides in Amerindians largely explained by an APOC3 founder mutation". en. In: *Circ. Cardiovasc. Genet.* 10.6 (Dec. 2017).

[7] Kimmo Palin et al. "Identity-by-descent-based phasing and imputation in founder populations using graphical models". en. In: *Genet. Epidemiol.* 35.8 (Dec. 2011), pp. 853–860.

[8] Keith Noto and Luong Ruiz. "Accurate genome-wide phasing from IBD data". en. In: *BMC Bioinformatics* 23.1 (Nov. 2022), p. 502.

[9] Evan L. Sticca, Gillian M. Belbin, and Christopher R. Gignoux. "Current Developments in Detection of Identity-by-Descent Methods and Applications". In: *Frontiers in Genetics* 12 (2021). ISSN: 1664-8021. DOI: 10.3389/fgene.2021.722602.

[10] Y. Zhou, S. R. Browning, and B. L. Browning. "A Fast and Simple Method for Detecting Identity-by-Descent Segments in Large-Scale Data". In: *Am J Hum Genet* 106.4 (Apr. 2020), pp. 426–437.

[11] S. R. Browning and B. L. Browning. "Identity by descent between distant relatives: detection and applications". In: *Annu Rev Genet* 46 (2012), pp. 617–633.

[12] Ardalan Naseri et al. "RaPID: ultra-fast, powerful, and accurate detection of segments identical by descent (IBD) in biobank-scale cohorts". en. In: *Genome Biol.* 20.1 (July 2019), p. 143.

[13] J. Nait Saada et al. "Identity-by-descent detection across 487,409 British samples reveals fine scale population structure and ultra-rare variant associations". In: *Nat Commun* 11.1 (Nov. 2020), p. 6130.

[14] R. Shemirani et al. "Rapid detection of identity-by-descent tracts for mega-scale datasets". In: *Nat Commun* 12.1 (June 2021), p. 3546.

[15] Alexander Gusev et al. "Whole population, genome-wide mapping of hidden relatedness". en. In: *Genome Res.* 19.2 (Feb. 2009), pp. 318–326.

[16] D. N. Seidman et al. "Rapid, Phase-free Detection of Long Identity-by-Descent Segments Enables Effective Relationship Classification". In: *Am J Hum Genet* 106.4 (Apr. 2020), pp. 453–466.

[17] A. Dimitromanolakis, A. D. Paterson, and L. Sun. "Fast and Accurate Shared Segment Detection and Relatedness Estimation in Un-phased Genetic Data via TRUFFLE". In: *Am J Hum Genet* 105.1 (July 2019), pp. 78–88.

[18] Sharon R Browning and Brian L Browning. "High-resolution detection of identity by descent in unrelated individuals". en. In: *Am. J. Hum. Genet.* 86.4 (Apr. 2010), pp. 526–539.

[19] Jesse Rodriguez, Serafim Batzoglou, and Sivan Bercovici. "An Accurate Method for Inferring Relatedness in Large Datasets of Unphased Genotypes via an Embedded Likelihood-Ratio Test". In: Springer, Apr. 2013, pp. 212–229. ISBN: 978-3-642-37194-3. DOI: 10.1007/978-3-642-37195-0_18.

[20] J. M. Rodriguez et al. "Parente2: a fast and accurate method for detecting identity by descent". In: *Genome Res* 25.2 (Feb. 2015), pp. 280–289.

[21] R. J. Bateman et al. "Autosomal-dominant Alzheimer's disease: a review and proposal for the prevention of Alzheimer's disease". In: *Alzheimers Res Ther* 3.1 (Jan. 2011), p. 1.

[22] Victoria A. S. Nyst. "Hereditary Deafness in a Former Fishing Village on the Dutch Coast". In: *The Journal of Deaf Studies and Deaf Education* 21.1 (Sept. 2015), pp. 94–103. ISSN: 1081-4159. DOI: 10.1093/deafed/env045. URL: https://doi.org/10.1093/deafed/env045.

[23] X. Ma et al. "Analysis of error profiles in deep next-generation sequencing data". In: *Genome Biol* 20.1 (Mar. 2019), p. 50.

[24] S. Hochreiter. "HapFABIA: identification of very short segments of identity by descent characterized by rare variants in large sequencing data". In: *Nucleic Acids Res* 41.22 (Dec. 2013), e202.

[25] L. Excofffier et al. "fastsimcoal2: demographic inference under complex evolutionary scenarios". In: *Bioinformatics* (June 2021).

[26] L. Excoffier et al. "Robust demographic inference from genomic and SNP data". In: *PLoS Genet* 9.10 (Oct. 2013), e1003905.

[27] Marta Byrska-Bishop et al. "High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios". en. In: *Cell* 185.18 (Sept. 2022), 3426–3440.e19.

[28] Susan Fairley et al. "The International Genome Sample Resource (IGSR) collection of open human genomic variation resources". en. In: *Nucleic Acids Res.* 48.D1 (Jan. 2020), pp. D941–D947.

[29] Lei Sun and Apostolos Dimitromanolakis. "PREST-plus identifies pedigree errors and cryptic relatedness in the GAW18 sample using genome-wide SNP data". en. In: *BMC Proc.* 8.Suppl 1 Genetic Analysis Workshop 18Vanessa Olmo (June 2014), S23.

[30] T. C. Glenn. "Field guide to next-generation DNA sequencers". In: *Mol Ecol Resour* 11.5 (Sept. 2011), pp. 759–769.

[31] B. L. Browning and S. R. Browning. "Improving the accuracy and efficiency of identity-by-descent detection in population data". In: *Genetics* 194.2 (June 2013), pp. 459–471.

Acknowledgments

We would like to thank Dr. Sivan Bercovici and Prof. Serafim Batzoglou for their useful comments and supervision in the early stage of this project. We thank the Research Grants Council, Hong Kong, the Research Committee of Hong Kong Baptist University and the Interdisciplinary Research Clusters Matching Scheme for their kind support of this project.

Author contribution

LZ conceived the study, CW collected data, developed the software, and conducted experiments. CW and LZ wrote the manuscript. W.P.V and L.Z. edited the manuscript. All authors reviewed the manuscript.

Funding

This research was partially supported by the Hong Kong Research Grant Council Early Career Scheme (HKBU 22201419), Guangdong-Hong Kong Technology Cooperation Funding Scheme (GHX/133/20SZ), HKBU Start-up Grant Tier 2 (RC-SGT2/19-20/SCI/007), HKBU IRCMS (No. IRCMS/19-20/D02) and the Guangdong Basic and Applied Basic Research Foundation (No. 2021A1515012226).

Competing interests

The authors declare no competing interests.

Materials & Correspondence

Correspondence and requests for materials should be addressed to Lu Zhang.

Method	Main algorithm	Haplotypes/Genotypes	Probabilistic model
BEAGLE IBD [18]	LHCM and HMM	Haplotypes	Yes
GERMLINE [15]	Hash and extension	Haplotypes	No
GERMLINE2 [13]	Hash and extension	Haplotypes	No
Hap-IBD [10]	PBWT and extension	Haplotypes	No
HapFABIA [24]	Biclustering	Haplotypes/Genotypes	No
IBIS [16]	Search long shared regions	Genotypes	No
iLASH [14]	Locality sensitive hashing	Haplotypes	No
Parente [19]	Naive Bayes	Genotypes	Yes
Parente2 [20]	Naive Bayes	Genotypes	Yes
RaPID [12]	PBWT	haplotypes	No
TRUFFLE [17]	Search long shared regions	Genotypes	No

Table 1: An overview of IBD detection tools. LHCM: localized haplotype cluster model; HMM: hidden Markov model; PBWT: positional Burrows-Wheeler transform.

Figure 1: SILO framework. (**A**) The SILO algorithm. (**B**) SILO's block-and-window structure. A chromosome is divided into many windows. Each window contains a fixed number of common variants (eight in this figure). A block is built on top of windows, and its length is

measured in centimorgan (3 cM in this figure). (**C**) Seed-and-extend diagram. During the extension, SILO first extends to the left flanking window and then to the right flanking window. If the extension to a window stops, the extension in that direction stops.

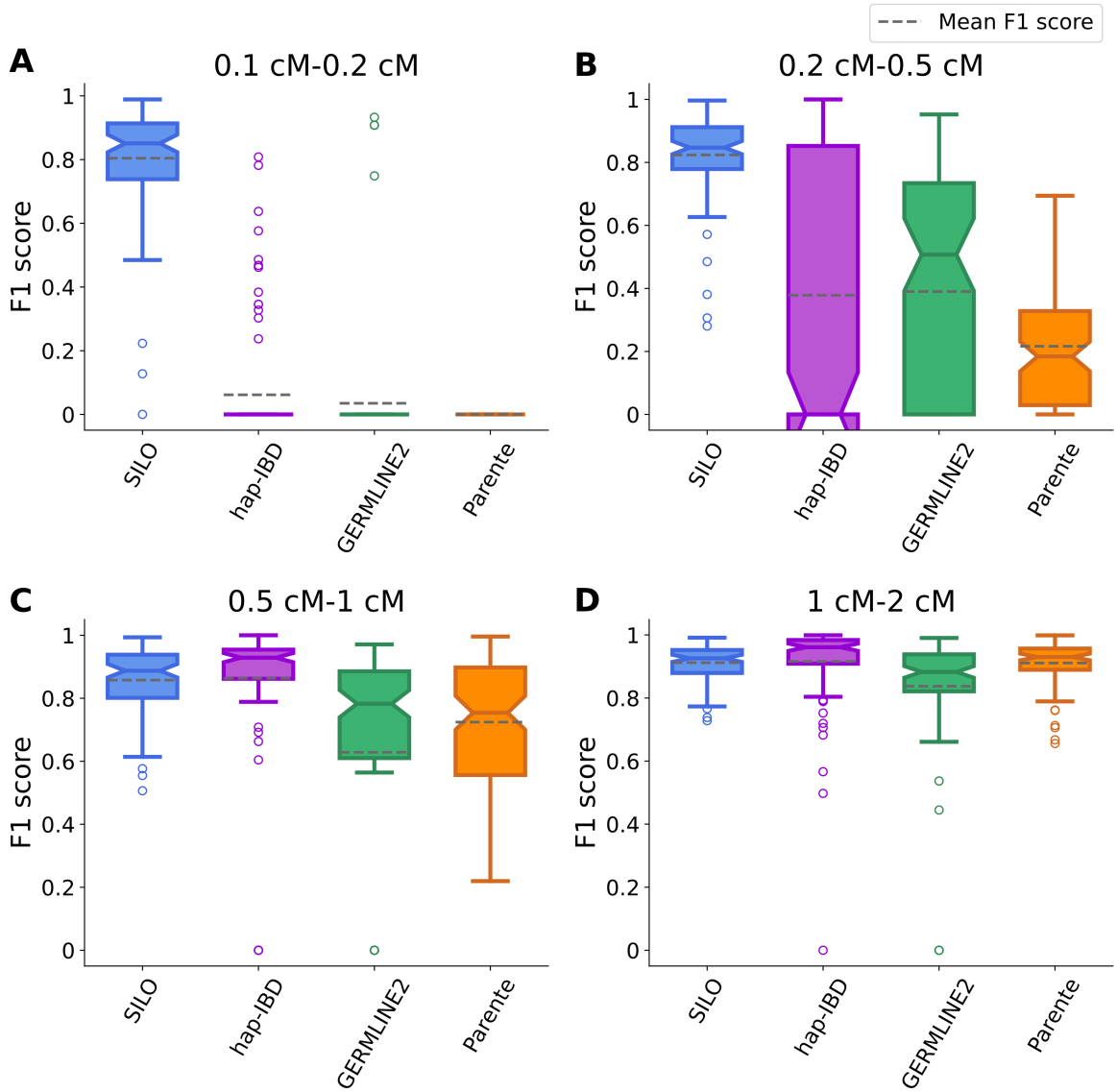

Figure 2: F_1 scores of each tool in the simulated data. Single LFV insertion results are demonstrated in this figure for different length scales (**A**-**D**). Refer to **Figure S1** for multiple LFVs insertion results.

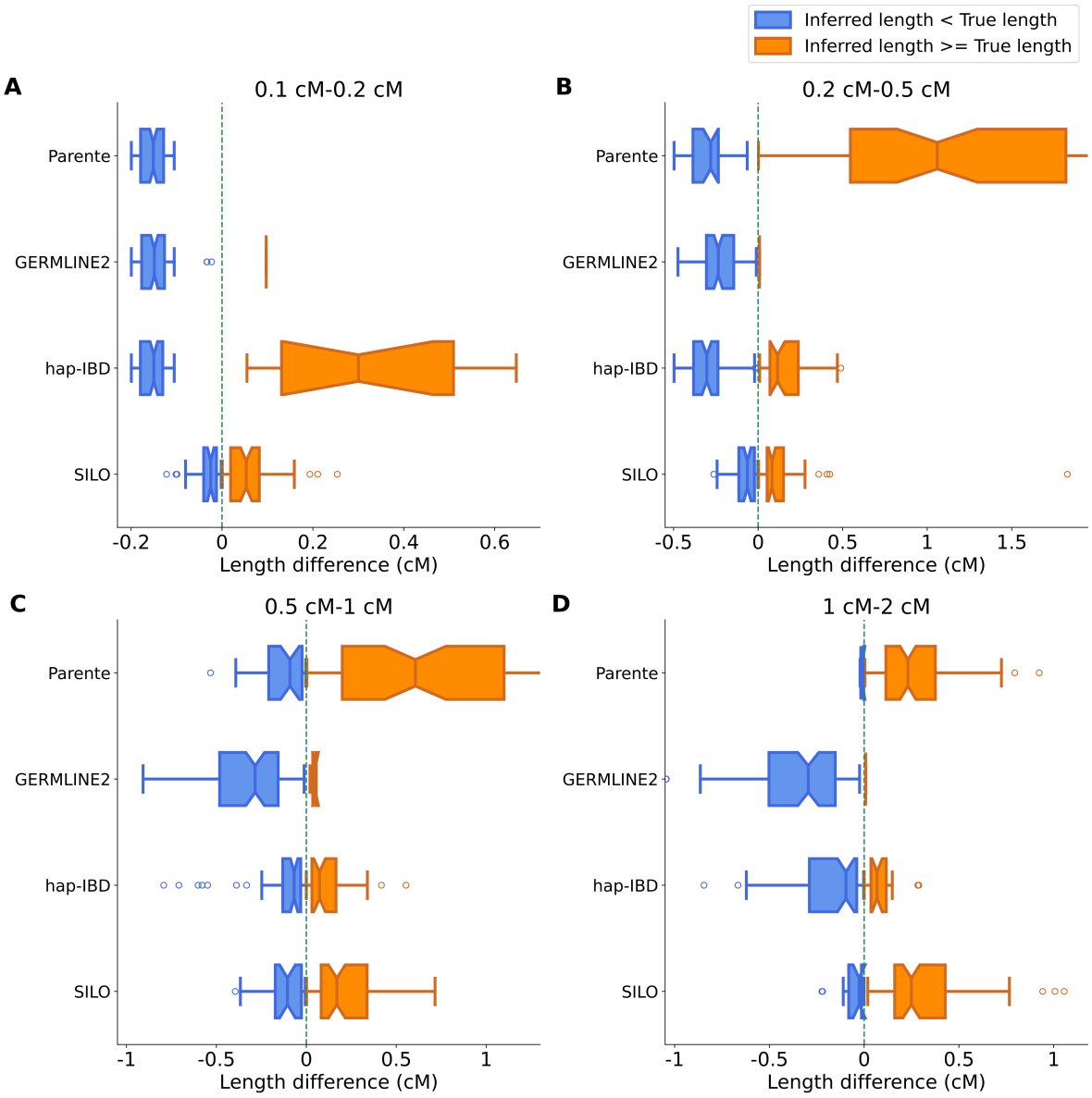

Figure 3: Length differences between inferred IBD segments and true IBD segments in the simulated data. Single LFV insertion results are demonstrated in this figure for different length scales (**A**-**D**). Refer to **Figure S2** for multiple LFVs insertion results.

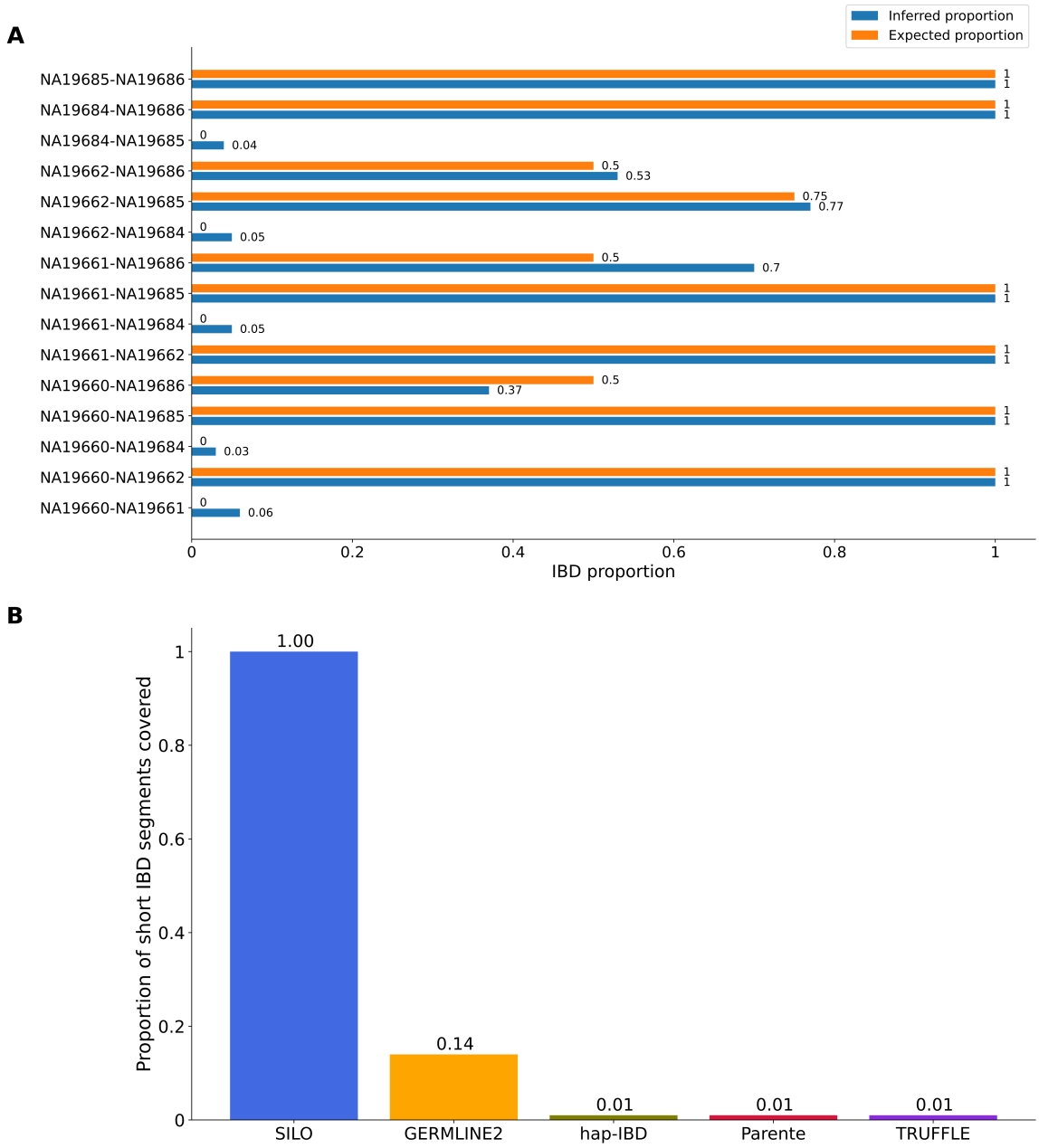

Figure 4: (**A**) The proportions of inferred autosomal IBD segments by SILO compared to the expected proportions of IBD segments in a real Mexican pedigree from the 1000 Genomes Project (**B**) The proportions of short IBD segments covered (or found) by each method relative to the validated short IBD segments found by SILO. See **Figure S4-S7** for the results of reciprocal detection of true IBD segments found by each of the other four tools.

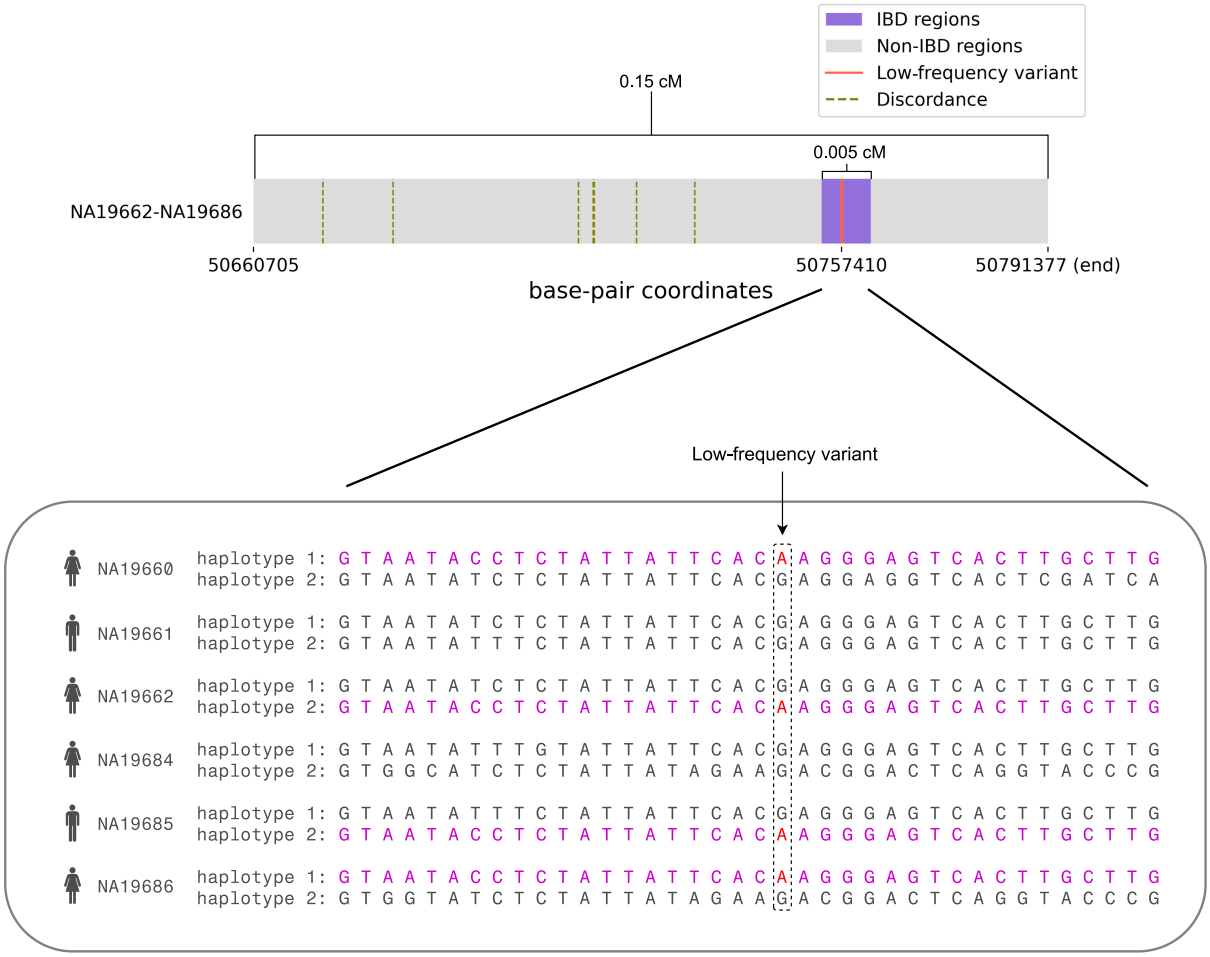

Figure 5: An example of a short IBD segment shared by NA19662 and NA19686 on the 22nd chromosome. This short IBD segment is inherently unextendible due to the presence of seven SNVs with discordant alleles upstream.