Tractatus: an exact and subquadratic algorithm for inferring identity-by-descent multi-shared haplotype tracts

Derek Aguiar[1], Eric Morrow[2], and Sorin Istrail[1,⋆]

[1] Department of Computer Science and Center for Computational Biology,
Brown University, Providence, Rhode Island 02912, USA
Derek_Aguiar@brown.edu and Sorin_Istrail@brown.edu
[2] Departments of Molecular Biology, Cell Biology & Biochemistry and Psychiatry & Human Behavior,
Brown University, Providence, Rhode Island 02912, USA
Eric_Morrow@brown.edu

Abstract. In this work we present graph theoretic algorithms for the identification of all identical-by-descent (IBD) multi-shared haplotype tracts for an $m \times n$ haplotype matrix. We introduce Tractatus, an exact algorithm for computing all IBD haplotype tracts in time linear in the size of the input, $O(mn)$. Tractatus resolves a long standing open problem, breaking optimally the (worst-case) quadratic time barrier of $O(m^2n)$ of previous methods often cited as a bottleneck in haplotype analysis of genome-wide association study-sized data. This advance in algorithm efficiency makes an impact in a number of areas of population genomics rooted in the seminal Li-Stephens framework for modeling multi-loci linkage disequilibrium (LD) patterns with applications to the estimation of recombination rates, imputation, haplotype-based LD mapping, and haplotype phasing. We extend the Tractatus algorithm to include computation of haplotype tracts with allele mismatches and shared homozygous haplotypes in a set of genotypes. Lastly, we present a comparison of algorithmic runtime, power to infer IBD tracts, and false positive rates for simulated data and computations of homozygous haplotypes in genome-wide association study data of autism. The Tractatus algorithm is available for download at http://www.brown.edu/Research/Istrail_Lab/.

Keywords: haplotypes, haplotype tracts, graph theory, identity-by-descent

1 Introduction

1.1 Li-Stephens PAC-Likelihood Model and the $O(m^2n)$ time bound

Understanding and interpreting patters of linkage disequilibrium (LD) among multiple variants in a genome-wide population sample is a major technical challenge in population genomics. A large body of research literature is devoted to the topic including the computational framework presented in the seminal work of Li and Stephens[1]. Building on the work by Stephens *et al.* 2001[2], Hudson[3], and Fearnhead and Donnelly[4], the Li-Stephens framework led the way towards major advances in the understanding and modeling of linkage disequilibrium patterns and recombination.

The difficulties associated with modeling LD patterns at multiple loci include a number of long standing analytical obstacles. Among existing bottlenecks is the notorious (1) *curse of the pairwise*, as all the popular LD measures in the literature are pairwise measures, and the (2) *haplotype block-free* approach to avoid *ad hoc* haplotype block definitions and "fake blocks" due to recombination rate heterogeneity. Current methods for computing haplotype blocks result in the definition of *ad hoc* boundaries that sometimes present less LD within blocks than between blocks due to different patterns of recombination. This phenomenon leads to spurious block-like clusters. The Li-Stephens statistical model for LD, named the *Product of Approximate Conditionals* (PAC), is based on a generalization of coalescent theory to include recombination [3,5].

The optimization problem introduces the PAC likelihood $L_{PAC}(\rho)$

$$L_{PAC}(\rho) = \tilde{\pi}(h_1 \mid \rho)\tilde{\pi}(h_2 \mid h_1, \rho)...\tilde{\pi}(h_m \mid h_1, ..., h_{m-1}, \rho)$$

⋆ corresponding author

where $h_1, ..., h_m$ are the m sampled haplotypes, ρ denotes the recombination parameter, and $\tilde{\pi}$ represents an approximation of the corresponding conditional probabilities. Li and Stephens propose a number of such approximations for approximate likelihood functions[1]. $L_{PAC}(\rho)$ represents the unknown distribution

$$Prob(h_1, ..., h_m \mid \rho) = Prob(h_1 \mid \rho)Prob(h_2 \mid h_1, \rho)...Prob(h_m \mid h_1, ..., h_{m-1}, \rho)$$

The choice of $\tilde{\pi}$ gives the form of the likelihood objective function.

The PAC likelihood is based on expanding the modeling to capture realistic genomic structure while generalizing Ewens' sampling formula and coalescent theory. The framework iteratively samples the m haplotypes; if the first k haplotypes have been sampled $h_1, ..., h_k$, then the conditional distribution for the next sampled haplotype is $Prob(h_{k+1} \mid h_1, ..., h_k)$. $\tilde{\pi}$ approximates this distribution and is constructed to satisfy the following axioms:

1. h_{k+1} is more likely to match a haplotype from $h_1, ..., h_k$ that has been observed many times rather than a haplotype that has been observed less frequently.
2. The probability of observing a novel haplotype decreases as k increases.
3. The probability of observing a novel haplotype increases as $\theta = 4N\mu$ increases, where N is the population size and μ is the mutation rate.
4. If the next haplotype is not identical to a previously observed haplotype, it will tend to differ by a small number of mutations from an existing haplotype (as in the Ewens' sampling formula model).
5. Due to recombination, h_{k+1} will resemble haplotypes $h_1, ..., h_k$ over contiguous genomic regions; the average physical length of these regions should be larger in genomic regions where the local rate of recombination is low.

Intuitively, the next haplotype h_{k+1} should be an imperfect *mosaic* of the first k haplotypes, with the size of the mosaic fragments being smaller for higher values of the recombination rate. Although the proposed model ($\tilde{\pi_A}$ in the notation of [1]) satisfies the above axioms and has the desirable property of being efficiently computable, it has a serious disadvantage. As is stated in their article, this "unwelcome" feature of the PAC likelihoods corresponding to the choices for $\tilde{\pi}$ is *order dependence*, that is, the choices are dependent on the order of the haplotypes sampled. Other methods used in the literature, notable, Stephens *et al.* 2001[2] and Fearnhead and Donnelly[4], present the same problem of order dependence. Different haplotype sampling permutations correspond to different distributions; these probability distributions *do not satisfy the property of exchangeability* that we would expect to be satisfied by the true but unknown distribution.

1.2 Identical-by-descent haplotype tracts

Haplotype tracts, or contiguous segments of haplotypes, are identical-by-descent (IBD) if they are inherited from a common ancestor [6]. Tracts of haplotypes shared IBD are disrupted by recombination so the expected lengths of the IDB tracts depends on the pedigree structure of the sample and the number of generations till the least common ancestor at that haplotype region. The computation of IBD is fundamental to genetic mapping and can be inferred using the PAC likelihood model.

To model the effects of recombination, a hidden Markov model (HMM) is defined to achieve a mosaic construction. At every variant, it is possible to transition to any of the haplotypes generated so far with a given probability. Thus, a path through the chain starts with a segment from one haplotype and continues with a segment from another haplotype and so on. To enforce the mosaic segments to resemble haplotype tracts, the probability of continuing in the same haplotype without jumping is defined exponentially in terms of the physical distance (assumed known) between the markers; that is, if sites j and $j + 1$ are at a small genetic distance apart, then they are highly likely to exist on the same haplotype. The computation of the L_{PAC} is linear in the number of variants (n) and quadratic in the number of haplotypes (m) in the sample, hence the $O(m^2n)$ time bound.

In this work we present results that remove the pairwise quadratic dependence by computing multi-shared haplotype tracts. Multi-shared haplotype tracts are maximally shared contiguous segments of haplotypes starting and ending at the same genomic position that cannot be extended by adding more haplotypes in the sample. Because we represent the pairwise sharing in sets of haplotypes, no more than $O(mn)$ multi-shared haplotype tracts may exist.

1.3 Prior work

Building on the PAC model, the IMPUTE2 [7] and MaCH [8] algorithms employ HMMs to model a sample set of haplotypes as an imperfect mosaic of reference haplotypes. The usage of the forward-backward HMM algorithm brings these methods in the same $O(m^2n)$ time bound class. The phasing program SHAPEIT (segmented haplotype estimation and imputation tool) also builds on the PAC model by decomposing the haplotype matrix uniformly into a number of segments and creating linear time mosaics within each such *ad hoc* segmented structure[9]. The dependence on the number of segments is not considered in the time complexity.

PLINK [10], FastIBD [11], DASH [12], and IBD-Groupon [13] are algorithms based on HMMs or graph theory clustering methods that consider pairs of haplotypes to compute IBD tracts. Iterating over all such pairs takes time $O(m^2n)$ and is intractable for large samples; this intractability is best described in the recent work of Gusev *et al.* 2011.

> "Although the HMM schemes offer high resolution of detection [of IBD], the implementations require examining all pairs of samples and are intractable for GWAS-sized cohorts. ... In aggregate, these identical-by-descent segments can represent the totality of detectable recent haplotype sharing and could thus serve as refined proxies for recent variants that are generally rare and difficult to detect otherwise." Gusev *et al.* 2011 [12]

Gusev *et al.* 2009 describes the computationally efficient algorithm GERMLINE which employs a dictionary hashing approach[14]. The input haplotype matrix is divided into discrete slices or windows and haplotype words that hash to the same value are identified as shared. Due to this dependence on windows, the algorithm is inherently inexact. While the identification of small haplotype tracts within error-free windows can be performed in linear time, GERMLINE's method for handling base call errors is worst case quadratic. However, GERMLINE's runtime has been shown to be near linear time in practice [6].

In what follows, we describe the Tractatus algorithm for computing IBD multi-shared haplotype tracts from a sample of haplotypes and the Tractatus-HH algorithm for computing **h**omozygous **h**aplotypes in a sample of genotypes. Section 2 introduces the computational model and algorithms. Section 3 compares the runtime of Tractatus to a generic pairwise algorithm, compares false positive rates and power with GERMLINE, and provides an example computation of homozygous haplotype regions in genome-wide association study data of autism. Finally, sections 4 and 5 discuss implications of this algorithm, conclusions, and future directions.

2 Methods

Our work presented here addresses the lack of exchangeability in the sampling methods of the Li-Stephens model and provides a rigorous result that gives a basis for sampling with the assured exchangeability property. We also present a data structure that speeds up the HMM and the graph clustering models for the detection of identity-by-descent haplotype tracts. Informally, a *haplotype tract* or simply *tract* is a contiguous segment of a haplotype – defined by start and end variant indices – that is shared (identical) by two or more haplotypes in a given sample of haplotypes. One can then view each of the haplotypes in the set as a mosaic concatenation of tracts. Such a haplotype tract decomposition is unique and a global property of the sample. Our Tractatus algorithm computes the *Tract tree* of all the tracts of the haplotype sample in linear time in the size of the sample. The Tract tree, related to a suffix tree, represents each haplotype tract in a single root-to-internal-node path. Repeated substrings in distinct haplotypes are compressed and represented only once in the Tract tree.

2.1 The Tractatus model

Suffix trees are graph theoretic data structures for compressing the suffixes of a character string. Several algorithms exist for suffix tree construction including the notable McCreight and Ukkonen algorithms that achieve linear time and space constructions for $O(1)$ alphabets [15,16]. In 1997, Farach introduced the first

suffix-tree construction algorithm that is linear time and space for integer alphabets [17]. Extensions to suffix-trees, commonly known as generalized suffix trees, allow for suffix-tree construction of multiple strings.

The input to the problem of IBD tract inference is m haplotypes which are encoded as n-length strings of 0's and 1's corresponding to the major and minor alleles of genomic variants $v_1, ..., v_n$. Because we are interested in IBD relationships which are by definition interhaplotype, naive application of suffix-tree construction algorithms to the set of haplotypes would poorly model IBD by including intrahaplotype relationships. Let haplotype i be denoted h_i and the allele of h_i at position j be denoted $h_{i,j}$. Then, we model each haplotype $h_i = h_{i,1}, h_{i,2}, ..., h_{i,n}$ with a new string $d_i = (h_{i,1}, 1), (h_{i,2}, 2), ..., (h_{i,n}, n)$ for $1 \leq i \leq m$. Computationally, the position-allele pairs can be modeled as integers $\in [0, 1, 2, ..., 2n-1]$ where $(h_{i,j}, j)$ is $2 * j + h_{i,j}$ where $h_{i,j} \in 0, 1$. The transformed haplotype strings are termed *tractized*.

2.2 The Tractatus algorithm without errors

The Tractatus algorithm incorporates elements from integer alphabet suffix trees with auxiliary data structures and algorithms for computing IBD haplotype tracts. Firstly, a suffix tree is built from the set of m tractized haplotypes each of length n. To represent the tractized haplotypes, the alphabet size is $O(n)$, so Farach's algorithm may be used to construct a suffix tree in linear time[17]. The suffix tree built from the tractized haplotypes is termed the *Tract tree*. After the Tract tree is built, an $O(mn)$ depth first post-order traversal (DFS) is computed to label each vertex with the number of haplotype descendants. These pointers enable the computation of groups of individuals sharing a tract in linear time.

Substrings of haplotypes are compressed if they are identical and contain the same start and end positions in two or more haplotypes. We consider a path from the root to a node with k descendants as maximal if it is not contained within any other path in the Tract tree. The maximal paths can be computed using a depth first search of the Tract tree, starting with suffixes beginning at 0 and ending at suffixes beginning with $2n-1$. Of course, if a tract is shared by $k \geq 2$ haplotypes, it is represented only once in the Tract tree. Figure 1 shows the construction of the Tract tree and computation of IBD tracts.

The internal nodes of the Tract tree also have an interpretation in regards to Fisher junctions. A Fisher junction is a position in DNA between two variants such that the DNA segments that meet in this virtual point were ancestrally on different chromosomes. Fisher junctions are represented in the Tract tree where maximal tracts branch.

After maximal tracts are computed, they are quantified as IBD or IBS. Tractatus implements two methods for calling maximally shared tracts IBD or IBS. A simple tract calling method thresholds the length L (number of variants) or area (variants $\times$ haplotypes) of the tract in terms of the haplotype matrix input. A more complex method considers the probability of a region being shared IBD or identical-by-state (IBS). If two individuals are k^{th} degree cousins, the probability they share a haplotype tract IBD is 2^{-2k} due to the number of meioses between them[18]. Let the frequency of a variant at position i be f_i. Then, the probability of IBD and IBS can be combined to define the probability that a shared haplotype tract of length L and starting at position s for k^{th} degree cousins is IBD (Equation 1) [19]

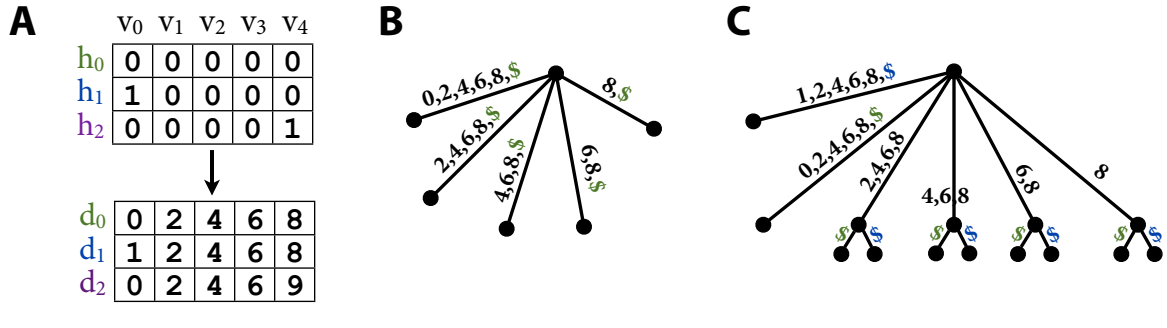
$$P(IBD|L) = \frac{2^{-2k}}{2^{-2k} + \prod_{i=s}^{s+L} (f_i^2 + (1-f_i)^2)} \tag{1}$$

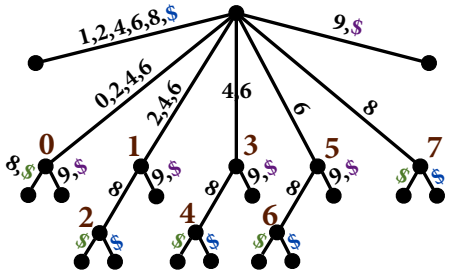
The value of k can be approximated if the population structure is known. Tractatus without errors is presented in Algorithm 1. Because the suffix tree is computable in $O(mn)$ time with $O(mn)$ nodes, the tree traversals can be computed in $O(mn)$ time thus giving Theorem 1.

Theorem 1. *Given a set of m haplotypes each of length n, Algorithm 1 computes the Tract tree and the set of IBD tracts in $O(mn)$ time and space.*

2.3 The Tractatus algorithm with errors and allele mismatches

Incorporating base call errors and additional variability gained after differentiation from the least common ancestor requires additional computations on the Tract tree and a statistical modeling of haplotype allele mismatches. The Tractatus algorithm with errors is parameterized by an estimated probability of error or





E

search order	haplotypes	tract
0	{h0,h2}	[v0,v3]
1	{h0,h1,h2}	[v1,v3]
2	{h0,h1}	[v1,v4]
3	{h0,h1,h2}	[v2,v3]
4	{h0,h1}	[v2,v4]
5	{h0,h1,h2}	[v3,v3]
6	{h0,h1}	[v3,v4]
7	{h0,h1}	[v4,v4]

Fig. 1. Construction of the Tract tree and running Tractatus on example input without errors or allele mismatches. Terminator characters \$ are colored to match tractized haplotypes and the empty string (simply the terminator character) is omitted in this example. (A) The haplotype matrix is encoded by an integer alphabet representing position-allele pairs. (B) Tractized haplotype d_0 is inserted in the Tract tree. The first tractized haplotype inserts $O(n)$ nodes into the Tract tree. (C) Tractized haplotype d_1 is inserted in the Tract tree. The suffix of d_1 starting at v_0 requires generation of a new node in the Tract tree but subsequent suffixes can be compressed along paths from the root. (D) Tractized haplotype d_2 is inserted in the Tract tree and the algorithmic search order is given in brown integers adjacent to internal nodes. Leaf nodes have exactly one terminating character (haplotype) and therefore do not have to be visited during the search. (E) The largest IBD tracts are found at search numbers 0, 1, and 2. Saving references to these tracts enables the determination that subsequent tracts are contained within already processed tracts.

input : m haplotypes each of length n, minimum length L or IBD probability p
output: set of IBD tracts

$H \leftarrow$ tractized haplotypes

$T(H) \leftarrow$ Tract tree of H

Post-order DFS of T(H) to compute descendant haplotypes from each node

DFS of T(H):
if *path in DFS is longer than L or $P(IBD) > p$ and node has at least 2 descendant haplotypes*
then
  if *tract is not contained in previously computed tract* **then**
    report as an IBD tract
  end
else
  push children nodes on stack
end

Algorithm 1: Tractatus (error free)

mismatched alleles p_t, a p-value threshold corresponding to a test for the number of errors in a tract p_h, a minimum length partial IBD tract l, and a minimum length of calling a full IBD tract L (or alternatively P(IBD) as defined in Equation 1). We will, in turn, explain the significance of each parameter.

The algorithm proceeds similarly to the error-free case. We build the tractized haplotypes, Tract tree and populate necessary data structures with a DFS. Because errors and additional variation now exist which can break up tracts (and therefore paths in the Tract tree), we compute partial tracts as evidence of IBD. We compute a DFS from the root, and a maximal partial tract is saved when the algorithm arrives at a node with path length at least l and at least 2 haplotype descendants. If we find a partial tract in a subsequent traversal, we can check in $O(1)$ time is it is contained in a maximal partial tract already computed. Figure 2 shows an example of the Tract tree construction with a single allele mismatch.

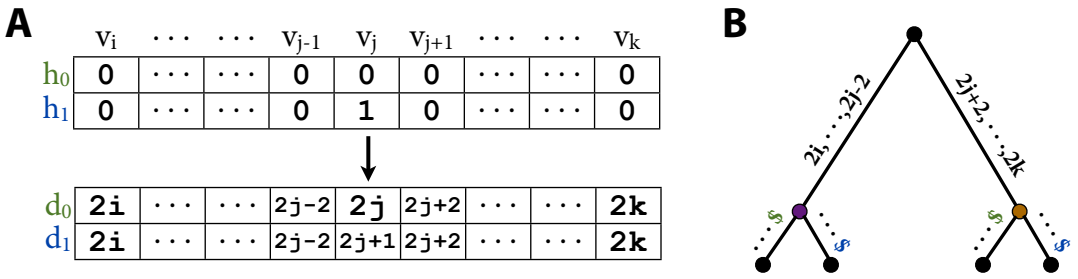


Fig. 2. Construction of the Tract tree and running Tractatus on example input with allele mismatches. (A) h_0 and h_1 share a tract IBD in the interval $[v_i, v_k]$ with a single allele mismatch at v_j. (B) By the construction of the Tract tree, there must be some path (here shown as a single edge but it may be split into a path by other haplotypes) from root to internal node that includes both $[v_i, v_k)$ and (v_k, v_i).

Because we computed the partial tracts using a DFS, the tracts are ordered by starting position. For each tract, tracts starting at a position prior and including a subset of the same haplotypes are combined if the extension is statistically probable. To determine the scan distance, we can compute a probability of observing a partially shared tract of length l given a window distance w (or this can be user defined). Assuming the generation of errors is independent and the probability of generating an error is p_e, we model the probability of generating at least k errors in an interval of l_i in t haplotypes as a Poisson process with $\lambda = p_e l_i t$. For each extension we calculate the probability of observing at least k mismatches and accept the extension if the probability is greater than p_h. The parameter p_t is used as an approximation of p_e. The haplotype consensus sequence of the tract is taken by majority rule at each variant position.

Pseudocode is given in Algorithm 2. While the algorithm is parameterized with five parameters, they are optional and default values are suitable in most cases.

Construction of the Tract-tree takes $O(mn)$ time and space. $O(mn)$ time is needed to prepare data structures and compute maximally shared partial tracts (post-order depth first search). A tract can be checked if it is contained in a previously processed tract in $O(1)$ time. It takes $O(mnw)$ to merge partial IBD tracts in the worst case when we have to extend many tracts covering a large portion of the matrix, thus yielding Theorem 2.

Theorem 2. *Given a set of m haplotypes each of length n, a scan distance w and a set of partial haplotype tracts, Algorithm 2 computes the Tract tree and set of IBD tracts in time and space $O(mn + sw)$.*

2.4 Extensions for homozygous haplotypes

A particular class of identical-by-descent relationships are long regions of extended homozygosity in genotypes. The two dominant concepts of extended regions of allelic homozygosity are the homozygous haplotype (HH) concept introduced by Miyazawa *et al.* 2007 and the well-known region or run of homozygosity

input : m haplotypes each of length n, partial tract length l, minimum length L or IBD probability p, p-value threshold p_h, estimated probability of error p_t, length of scan w
output: set of IBD tracts

$H \leftarrow$ tractized haplotypes

$T(H) \leftarrow$ Tract tree of H

Post-order DFS of T(H) to compute reachable haplotypes from each node

DFS of T(H):
if *path in DFS is longer than l, node has at least 2 descendant haplotypes, and is maximal* **then**
| add partial IBD tract to set of tracts S
else
| push children nodes on stack
end
for *tract $s \in S$* **do**
| *Check for extension in previously processed tracts within scan region w*
| *Compute probability according to number of errors in extension, p_t, the length of the extension, and the number of individuals*
| *If probability $> p_h$, merge tracts*
end
for *tract $s \in S$* **do**
| *If length of s is greater than L or $P(IBD) > p$, report as IBD tract*
end

Algorithm 2: Tractatus (with errors)

(ROH)[20]. A HH is defined as a genotype after the removal of heterozygous variants such that only homozygous variants remain. Miyazawa *et al.* 2007 compared every pair of HH in a small cohort and reported regions of consecutive matches over a threshold. ROHs are defined as extended genomic regions of homozygous variants allowing for a small number of heterozygous variants contained within. We can rigorously capture both concepts using Tractatus.

A naive model for computing HH would consider each heterozygous site as a wildcard allowing for either the 0 or 1 allele. A haplotype with k heterozygous sites would require insertion of 2^k haplotypes into the Tract tree. This immediately suggests a fixed-parameter tractable algorithm using the same machinery as Tractatus. However, we can remove the dependence on k using a key insight regarding the structure of the Tract tree and tractized haplotypes.

Errors split tracts in the Tract tree such that the shared tract fragments are on different paths from the root. Instead of encoding all 2^k possible haplotypes, we simply remove the heterozygous alleles from the tractized string. Because the position is inherently encoded in the tractized string, the removal of the heterozygous alleles would have the same effect as an error. Therefore, if we encode genotypes by simply removing heterozygous variants in the tractized string, we can run Algorithm 2 to produce all the homozygous haplotypes for a set of genotypes in linear time and space.

3 Results

The principle advantages of Tractatus over existing methods are the theoretically guaranteed subquadratic runtime and exact results in the error-free case which translate to improved results in the case with errors and allele mismatches. We evaluate the runtime of Tractatus against a generic algorithm that processes individuals in pairs using phased HapMap haplotypes from several populations. We then compare the power and false positive rates of both Tractatus and GERMLINE which is a leading method for IBD inference[14]. Finally, we show an application of Tractatus-HH by inferring homozygous haplotypes in a previously known homozygous region in the Simons Simplex Collection genome-wide association study data[21].

3.1 Tractatus vs. pairwise algorithm runtimes

To evaluate the runtime of Tractatus versus pairwise methods, we implemented the pairwise equivalent algorithm which iterates through pairs of individuals and reports tracts of variants occurring in both individuals over some threshold length of variants. The data consist of phased haplotypes from HapMap Phase III Release 2 in the ASW, CEU, CHB, CHD, GIH, JPT, LWK, MEX, MKK, TSI, and YRI populations[22]. Figure 3 left shows the independence between chromosome and computation time for the Tractatus suffix tree and the pairwise algorithm. Because the runtime of each algorithm does not depend on the chromosome, we varied the population sizes while keeping the number of variants constant for chromosome 22. Figure 3 right shows the quadratic computation time growth for the pairwise algorithm while Tractatus tree construction remains linear in the number of individuals.

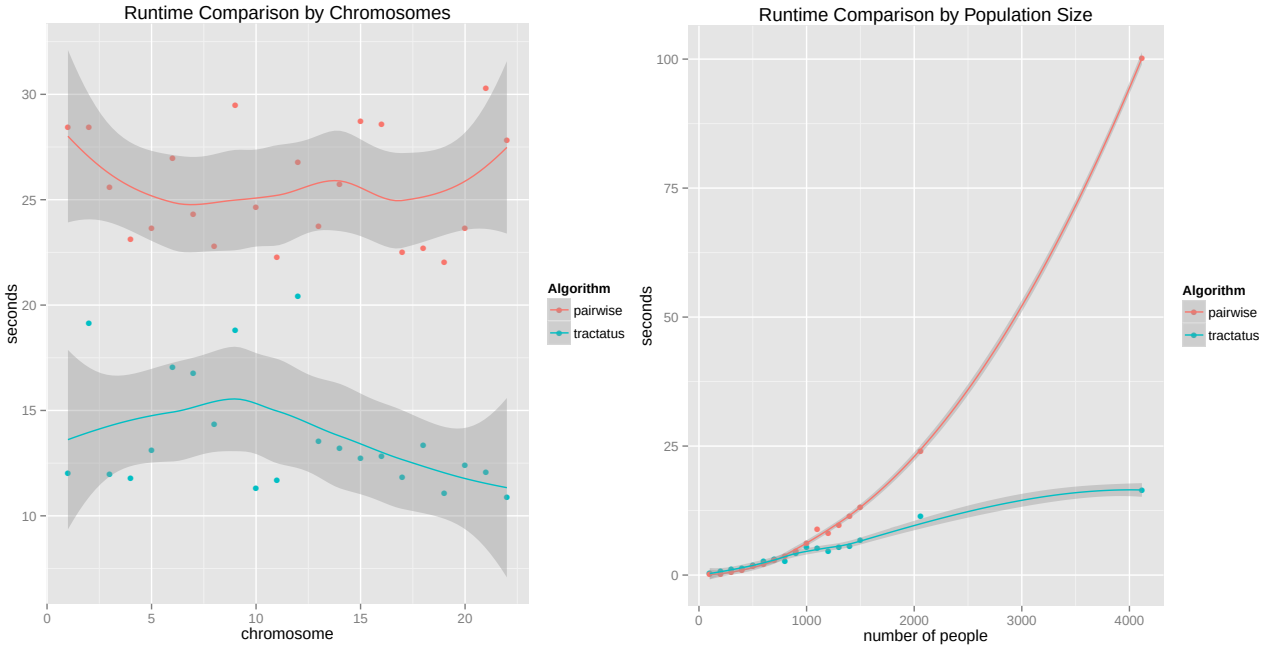


Fig. 3. Left: Tractatus and the pairwise algorithm were run on haplotypes from each chromosome of all HapMap populations for a minimum tract length of 100, and a randomly selected interval of 1000 variants. The experiment was repeated 100 times for each chromosome and elapsed time was averaged. Right: Tractatus and the pairwise algorithm were run on a randomly selected interval of 1000 variants from chromosome 22. The population size varied from 100 to twice the actual population size by resampling haplotypes with a 0.05 allele switch rate (per base).

3.2 False positive rates

Because it is difficult to construct a gold-standard baseline of true IBD regions in real data, our false positive rate and power calculations are performed on simulated data. To estimate the false positive rates for GERMLINE and Tractatus we simulated haplotypes at random and generated a single IBD region defined as having identical haplotype alleles in the region of IBD. We generated 100 haplotype matrices where $m = n = 500$ for all possible combinations of the number of individuals sharing a segment IBD $\in [3, 5, 10]$, the number of variants in the IBD region $\in [50, 60, 70, 80, 90, 100, 150, 200]$ and the single base substitution error rates $\in [0.0, 0.01, 0.05]$. In total, we generated 7200 haplotype matrices but aggregated the data across the number of individuals and variants in the IBD region because the false positive rates did not vary over these dimensions.

Table 1 shows that both algorithms have very low false positive rates in terms of the number of bases incorrectly called in an IBD region. However, Tractatus incorrectly calls less individuals in IBD regions than

GERMLINE. In this experiment, IBD regions were generated in block sizes and GERMLINE benefits from calling IBD regions in terms of blocks or windows. GERMLINE and Tractatus call a similar amount of bases IBD because Tractatus can over-estimate the ends of blocks. However, when individuals are compared, Table 1 shows that Tractatus computes a significantly smaller number of false positive IBD regions.

Table 1. False positive rates for the GERMLINE (G) and Tractatus (T) algorithms as a function of error rate. Each row corresponds to 2400 randomly generated haplotype matrices. The error rate was varied in a simulated haplotype matrix containing a single IBD region. False positive rates were calculated in terms of the number of non-IBD bases being called IBD (bases) and the number of individuals called IBD who were not in an IBD region (people) for the GERMLINE and Tractatus algorithms.

error rate	G FPR bases	T FPR bases	G FPR people	T FPR people
0.0	$1.3 \cdot 10^{-4}$	$1.16 \cdot 10^{-4}$	0.016	$2.13 \cdot 10^{-3}$
0.01	$1.2 \cdot 10^{-4}$	$1.11 \cdot 10^{-4}$	0.012	$8.72 \cdot 10^{-3}$
0.05	$6.1 \cdot 10^{-5}$	$4.18 \cdot 10^{-5}$	0.015	$7.43 \cdot 10^{-3}$

3.3 Power

We apply Tractatus and GERMLINE to the simulated data from Section 3.2 and estimate power by computing the number of times GERMLINE and Tractatus correctly call the IBD region in terms of variants and individuals. We considered an individual being called correctly if an IBD region was called and overlapped anywhere in the interval of the true IBD tract. We set the -bits and min_m options of GERMLINE to 20 and 40 respectively which sets the slice size for exact matches to 20 consecutive variants and the minimum length of a match to be 40 MB (which corresponds to 40 variants in our simulated data). For a valid comparison, we set Tractatus to accept partial tract sizes of 20 variants and a minimum length of an IBD region to 40 variants.

Figure 4 shows the power of GERMLINE and Tractatus to infer IBD as a function of IBD region length, number of haplotypes sharing the region, and the probability of base call error. Figure 4 right displays a *jagged* curve for GERMLINE which is likely due to the algorithmic dependence on window size. Both algorithms perform relatively well for shorter IBD tracts but Tractatus is clearly more powerful when the number of haplotypes sharing the tract increases or the base call error rates are low. Additionally, the minimum partial tract length for Tractatus could be lowered to increase the power to find smaller IBD tracts (at a cost of higher false positive rates). Another interesting observation is that both GERMLINE and Tractatus are able to perfectly infer all individuals sharing the IBD region in the perfect data case, but, GERMLINE is unable to compute the entire IBD interval in some data.

3.4 Homozygous haplotypes in autism GWAS data

As a proof of concept for Tractatus-HH, we extracted a $250kb$ genomic region identified as having a strong homozygosity signal in the Simons Simplex Collection[23]. The families analyzed include 1,159 simplex families each with at least one child affected with autism and genotyped on the Illumina 1Mv3 Duo microarray. Gamsiz et al. 2013 approached the problem by treating a homozygous region as a marker and testing for association or burden for the region as a whole[23]. Our analysis shows that regions of homozygosity are more complex than previously assumed and there can be multiple regions overlapping and sharing some segments of homozygous haplotypes but largely different in other segments (Table 2). We found more individuals possessing a homozygous haplotype than Gamsiz *et al.* 2013 because the probability of generating an error or heterozygous site was set to a large value (0.1) but in general this parameter can be adjusted to be more conservative.

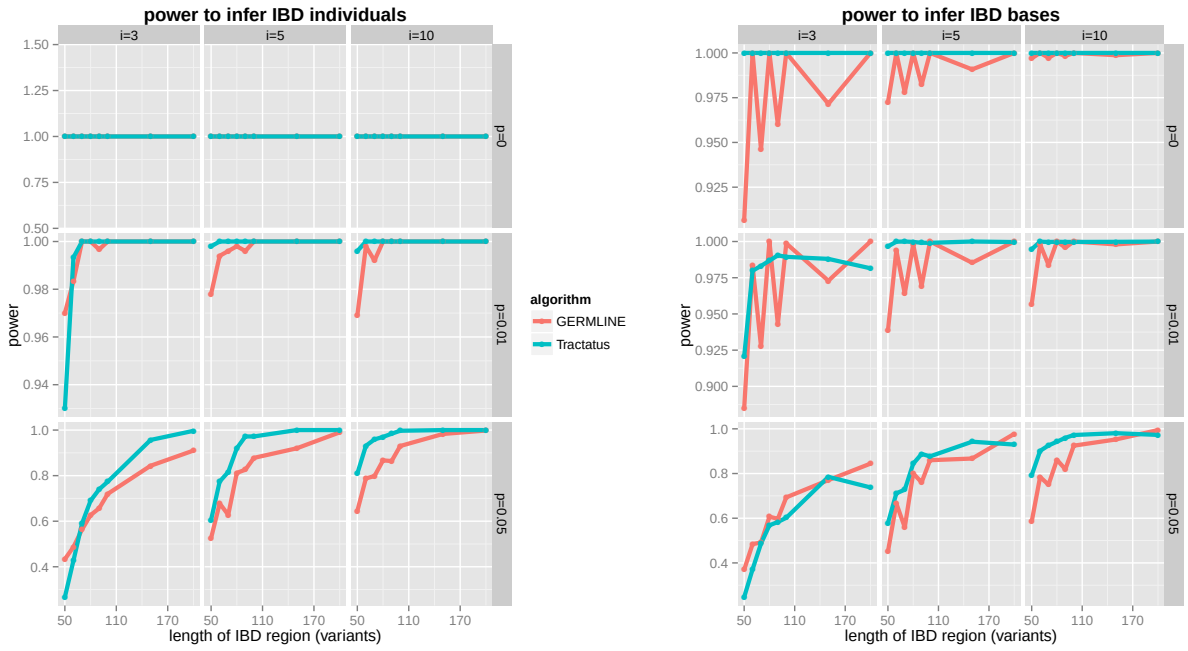


Fig. 4. The power to infer IBD by individual haplotypes (left) and variant bases (right) as a function of the length of the IBD region in variants (x-axis), the probability of base call error (p), and the number of individual haplotypes sharing the IBD segment (i).

Table 2. Analysis of a $250kb$ region of homozygosity in the Simons Simplex Collection. The homozygous interval is defined as a region start and end in terms of variants in the genomic interval, a number of individuals (size), and the number of individuals unique to the particular homozygous haplotype group (unique). One region is dominant and contains most of the individuals, but there are smaller regions with some overlap that contain unique individuals not sharing a homozygous haplotype with the larger region.

region start	region end	size	unique
0	111	20	10
0	109	20	12
0	109	252	238

4 Discussion

The importance of provable bounds and exact solutions is exemplified in Section 3 and, in particular, Figure 4. Even in the error free case, GERMLINE approximates computing IBD tracts by processing windows or vertical slices of the haplotype matrix. Tractatus is able to compute maximally shared partial tracts exactly (which are exactly the IBD tracts in the error-free case). Moreover, the inexactness of GERMLINE due to the dependence of hashing windows is exacerbated in the case of errors. If errors fall in a pattern that cause individuals sharing a segment IBD to hash to different values then GERMLINE produces false negatives. Tractatus computes all maximally shared partial tracts without dependence on windows. Lastly, in the worst case, the number of matches per word is quadratic giving GERMLINE a complexity quadratic in the number of individuals. Even though this is unrealistic in practice, Tractatus compresses individuals sharing a partial tract into a single path of the suffix tree.

The Tract tree in itself is an interesting data structure with many possible applications. Once the Tract tree is computed for a set of haplotypes, the statistics of constructing the mosaic of tract combinations can be done rigorously and completely such that sampling can be implemented in an order independent manner satisfying the exchangeability property. For the HMM constructions, the availability of the complete set of tracts would provide a rigorous basis for defining the transition probabilities and overall linear time construction. For the graph clustering methods, the Tract tree represents tracts occurring multiple times together and thus this construction will maximize the power in association studies.

Unfortunately, the issue of acquiring the haplotypes remains. Almost exclusively, algorithms for computing IBD require haplotypes due, in part, to the higher power to infer a more subtle IBD sharing than in genotype data. However, this is not a major roadblock considering haplotype phasing algorithms can be highly parallelized or made more efficient using reference panels. Additionally, haplotype assembly algorithms are very efficient and can extend genome-wide [24].

A related and important unanswered problem is to compute IBD regions in genotypes faster than the naive quadratic allele sharing algorithm. Haplotype-based IBD inference algorithms have difficulties modeling genotypes predominately because the heterozygous site introduces ambiguity in haplotype phase. We believe an approach exploiting the Tract tree may infer IBD in genotypes in subquadratic time perhaps with a direct application of the Tractatus-HH algorithm. However, the number of heterozygous variants is usually very high, so additional computation would be required to handle the large quantity of ambiguous sites.

Our analysis of the autism genome-wide association study data shows that homozygous regions cannot simply be treated as a biallelic markers. Distinct homozygous haplotypes, while having a similar signature of homozygosity, can be composed of entirely different alleles. These finding suggest that homozygous regions are complex, multi-allelic markers.

Finally, we note that a similar linear time construction could be used for constructing a Tract tree for a set of haplotypes where there is known genetic information about the distance between variants as in the Li-Stephens PAC model[1]. The genetic distance can be modeled approximately as an integer and used in a similar encoding to compress "identical" tracts.

5 Conclusions

In this work, we described the Tractatus algorithm for computing IBD tracts with and without errors and homozygous haplotypes. Tractatus represents the first provably exact algorithm for finding multi-shared IBD tracts given a set of haplotypes as input; it computes all subsets of individuals that share tracts and the corresponding shared tracts in time linear in the size of the input. By starting from an exact and rigorous algorithmic baseline, we are able to modify downstream decisions based on the global IBD tract decomposition. We compare the runtimes of Tractatus and a generic pairwise algorithm that process individuals in pairs using phased HapMap haplotypes from several populations and show decreased runtimes. Also, we exhibit superior statistical power to infer IBD tracts with less false positives than GERMLINE. Finally, with a conceptual change to the interpretation of genotypes, we showed that homozygous haplotype inference in genotypes can be modeled in the same Tractatus framework and demonstrated Tractatus-HH in a previously known homozygous region of the Simons Simplex Collection autism data.

6 Acknowledgements

This work was supported by the National Science Foundation [1048831 and 1321000 to S.I.] and NIGMS-NIH (P20GM103645). We are grateful to all of the families at the participating Simons Simplex Collection (SSC) sites, the Simons Foundation Autism Research Initiative, and the principal investigators (A.L. Beaudet, R. Bernier, J. Constantino, E.H.C., Jr., E. Fombonne, D.H.G., E. Hanson, D.E. Grice, A. Klin, R. Kochel, D. Ledbetter, C. Lord, C. Martin, D.M. Martin, R. Maxim, J. Miles, O. Ousley, K. Pelphrey, B. Peterson, J. Piggot, C. Saulnier, M.W.S., W. Stone, J.S. Sutcliffe, C.A. Walsh, Z. Warren, and E. Wijsman). We appreciate obtaining access to phenotypic data in SFARI Base.

References

1. Li, N., Stephens, M.: Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics **165**(4) (2003) 2213–2233
2. Stephens, M., Smith, N.J., Donnelly, P.: A new statistical method for haplotype reconstruction from population data. American journal of human genetics **68**(4) (April 2001) 978–989
3. Hudson, R.R.: Gene genealogies and the coalescent process. Oxford Survey in Evolutionary Biology **7** (1991) 1–44
4. Fearnhead, P., Donnelly, P.: Estimating recombination rates from population genetic data. Genetics **159**(3) (2001) 1299–1318
5. Kingman, J.F.C.: On the Genealogy of Large Populations. Journal of Applied Probability **19** (1982) 27–43
6. Browning, S.R., Browning, B.L.: Identity by descent between distant relatives: Detection and applications. Annual Review of Genetics **46**(1) (2012) 617–633
7. Howie, B.N., Donnelly, P., Marchini, J.: A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet **5**(6) (2009) e1000529
8. Li, Y., Willer, C.J., Ding, J., Scheet, P., Abecasis, G.R.: Mach: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genetic Epidemiology **34**(8) (2010) 816–834
9. Delaneau, O., Marchini, J., Zagury, J.F.: A linear complexity phasing method for thousands of genomes. Nat Meth **9**(2) (December 2011) 179–181
10. 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) (September 2007) 559–575
11. Browning, B.L., Browning, S.R.: A fast, powerful method for detecting identity by descent. American journal of human genetics **88**(2) (February 2011) 173–182
12. Gusev, A., Kenny, E.E., Lowe, J.K., Salit, J., Saxena, R., Kathiresan, S., Altshuler, D.M., Friedman, J.M., Breslow, J.L., Pe'er, I.: DASH: A Method for Identical-by-Descent Haplotype Mapping Uncovers Association with Recent Variation. Am J Hum Genet **88**(6) (June 2011) 706–717
13. He, D.: IBD-Groupon: an efficient method for detecting group-wise identity-by-descent regions simultaneously in multiple individuals based on pairwise IBD relationships. Bioinformatics **29**(13) (2013) 162–170
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**(2) (2009) 318–326
15. McCreight, E.M.: A space-economical suffix tree construction algorithm. J. ACM **23**(2) (April 1976) 262–272
16. Ukkonen, E.: On-line construction of suffix trees. Algorithmica **14**(3) (1995) 249–260
17. Farach, M.: Optimal suffix tree construction with large alphabets. In: Proceedings of the 38th Annual Symposium on Foundations of Computer Science. FOCS '97, Washington, DC, USA, IEEE Computer Society (1997) 137–143
18. Kong, A., Masson, G., Frigge, M.L., Gylfason, A., Zusmanovich, P., Thorleifsson, G., Olason, P.I., Ingason, A., Steinberg, S., Rafnar, T., et al.: Detection of sharing by descent, long-range phasing and haplotype imputation. Nature genetics **40**(9) (2008) 1068–1075
19. Halldorsson, B.V., Aguiar, D., Tarpine, R., Istrail, S.: The Clark Phaseable sample size problem: long-range phasing and loss of heterozygosity in GWAS. Journal of Computational Biology **18**(3) (2011) 323–333
20. Miyazawa, H., Kato, M., Awata, T., Kohda, M., Iwasa, H., Koyama, N., Tanaka, T., Huqu, N., Kyo, S., Okazaki, Y.: Homozygosity Haplotype Allows a Genomewide Search for the Autosomal Segments Shared among Patients. The American Journal of Human Genetics **80**(6) (June 2007) 1090–1102
21. Fischbach, G.D., Lord, C.: The Simons Simplex Collection: A Resource for Identification of Autism Genetic Risk Factors. Neuron **68**(2) (2010) 192 – 195
22. International HapMap Consortium: The International HapMap Project. Nature **426**(6968) (December 2003) 789–796

23. Gamsiz, E., Viscidi, E., Frederick, A., Nagpal, S., Sanders, S., Murtha, M., Schmidt, M., Triche, E., Geschwind, D., State, M., Istrail, S., Jr., E.C., Devlin, B., Morrow, E.: Intellectual disability is associated with increased runs of homozygosity in simplex autism. The American Journal of Human Genetics **93**(1) (2013) 103–109
24. Aguiar, D., Istrail, S.: Hapcompass: A fast cycle basis algorithm for accurate haplotype assembly of sequence data. Journal of Computational Biology (2012)