Whole Population, Genomewide Mapping of Hidden Relatedness

Alexander Gusev[1], Jennifer K Lowe[2,6], Markus Stoffel[3], Mark Daly[4,5], David Altshuler[4,6,7], Jeffrey M. Friedman[2,8], Jan L. Breslow[2], Itsik Pe'er[1]

[1]Columbia University; [2]Rockefeller University; [3]ETH Zurich; [4]Broad Institute of Harvard and MIT; [5]Center for Human Genetic Research, Massachusetts General Hospital; [6]Department of Molecular Biology, Massachusetts General Hospital; [7]Department of Genetics HMS and Department of Medicine, HMS; [8]Howard Hughes Medical Institute

Correspondence should be addressed to:
Alexander Gusev (gusev@cs.columbia.edu) or Itsik Pe'er (itsik@cs.columbia.edu)
TEL: +1-212-939-7135
1214 Amsterdam Ave. Mailcode 0401
New York, NY 10027

Abstract

The ability to identify and quantify genealogical relationships between individuals within a population is an important step in accurately using such data for disease analysis and improving our understanding of demography. However, exhaustive pair-wise analysis which has been successful in small cohorts cannot keep up with the current torrent of genotype data. We present GERMLINE, a robust algorithm for identifying pairwise segmental sharing which scales linearly with the number of input individuals. Our approach is based on a dictionary of haplotypes, used to efficiently discover short exact matches between individuals and then expand these matches to identify long nearly-identical segmental sharing that is indicative of relatedness. We use GERMLINE to comprehensively survey hidden relatedness both in the HapMap as well as in a densely typed island population of 3,000 individuals. We verify that GERMLINE is in concordance with other methods when they can process the data, and also facilitates analysis of larger scale studies. We also demonstrate novel applications of precise analysis of hidden relatedness to detection of haplotype phasing errors and structural variation. We show that shared segment discovery can help identifying phasing errors and potentially resolve them. Finally, we use detected identity of genomic segments for exposing polymorphic deletions that are otherwise challenging to detect, with 8/14 deletions in the HapMap samples and 153/200 deletions in the island data having independent experimental validation.

1 Introduction

The new era of genome-wide examination of human variation has brought about major advances in mapping complex diseases and traits. With the rapid success of large scale analysis, a wide range of studies have been initiated which plan to offer public access to unprecedented amounts of genetic data. These analyses fall into two categories, association and linkage, each having complementary drawbacks. While linkage analysis hinges on difficult to find directly related individuals with rare variants of significant phenotypic penetrance, association studies can be hampered by unaccounted population substructure and a focus on common alleles. In the foreseeable future, extensive genetic information may be available for considerable fractions of the population, further blurring the distinction between examinations of “related” and “unrelated” individuals, linkage and association. These schools of genetic investigation would ideally converge to best analyze such data. Efficient and effective imputation of genealogical relationships from genetic data is essential for such convergence.

A recently published study of relatedness in the Human Haplotype Map detailed the surprising discovery of an abundance of long haplotype segments shared between individuals purported to be unrelated (Consortium 2005). These matches, too long and exact to have occurred by chance, are likely to be representatives of so-called “hidden” relatedness. When a haploid copy is co-inherited by individuals from a common ancestor, the locus is referred to as being identical by descent (IBD); entire regions of such loci, or IBD shared segments, from an unreported ancestor can be indicative of hidden relatedness. Identifiable hidden relatedness, highlighted even by the relatively small sample HapMap, is expected to be nearly commonplace as the number of sampled individuals increases.

Reviewing classical theory (Donnelly 1983), we recall that a shared ancestor may leave a pair of contemporary descendents with no genetic trace in common, but when a variant is indeed co-inherited, it is likely to be very long, providing unequivocal evidence for IBD. Formally, if we consider a pair of diploid individuals that share a single ancestor k generations back, presence of IBD at a particular locus would require co-inheritance across $2k$ meioses. Each meiosis having a 0.5 probability of transmitting a copy, the total probability such a copy being inherited down both lineages is 2^{1-2k} – less than 1% for any k≥4 (third half-cousins or less related). Despite being a low probability event, when it occurs such sharing would imply a very long segment to be nearly identical across the two samples (Thompson 2007). In the same example, segment length is expected to be $d/2k$ where d is the genetic distance, defined as 100 centimorgans (cM). This rate of change in IBD status along a pair of genomes facilitates computing the expected number of IBD segments genomewide. Table 1 illustrates these statistics for a pair of individuals k generations apart.

While IBD segments are long enough to be distinguished when they are present, the identification of these relatively rare events becomes increasingly difficult in the presence of many samples or markers. Much focus has been dedicated to algorithms for estimating IBD between pairs of individuals and a small number of loci. Essentially, these methods attempt to quantify the likelihood of having two, one, or none alleles IBD in a pair of individuals; introduced as Wright's inbreeding coefficient (Wright 1921) and later quantified for IBD (Malécot 1948).

Subsequent work has been split between genome-wide estimations of IBD and segment-wide analysis. Genome-wide estimation, such as the construction of pairwise IBD matrices (Mao and Xu 2005), is of such low resolution as to be undiscerning for individuals more than several generations apart. As we have shown in Table 1, genome-wide presence of IBD is relatively insignificant in comparison to arbitrary allele sharing; individual segmental occurrences are long enough to be significant, but are unaccounted for in a genome-wide estimate. On the other hand, segmental analysis has focused on exact multi-point estimators, which are generally unstable in the presence of genotype error and mutation (Hill and Hernandez-Sanchez 2007); or approximate calculations for IBD between pairs of individuals, which have been computationally intractable for more than a handful of representative loci (Hill and Weir 2007). However, with the development of SNP genotyping, comparisons between IBD estimation in SNP maps and microsatellite panels show that SNP analysis results in more uniform and informative results (Hinrichs, Bertelsen, Bierut, Dunn, Jin, Kauwe and Suarez 2005). To this end, genome-wide estimation as well as a whole-genome scan for segmental sharing has been implemented in the PLINK analysis tool-set (Purcell, Neale, Todd-Brown, Thomas, Ferreira, Bender, Maller, Sklar, de Bakker, Daly et al. 2007) which uses a method of moments approach. This algorithm is executed upon all pairs in a population and dependent on pairwise estimates of overall IBD. It is clear that genome-wide segmental sharing discovery offers a level of precision unmatched by whole-genome estimates. However, as available sample-size increases, strict pairwise analysis in quadratic time will rapidly become impractical. This is further confounded by the fact that IBD identification is prompted by a large searchable population. An application for segmental identification in time proportional to the number of samples is at the core of IBD discovery and imperative to keep up with the booming pace of data availability.

This paper introduces an algorithm for linear-time discovery of segmental sharing and the corresponding implementation - Genetic Error-tolerant Regional Matching with LINear-time Extension (GERMLINE). GERMLINE precisely identifies shared segments between pairs of individuals from an input of phased or unphased genotype data, and uses these segments for data inference when possible. We are motivated by the assumption that most portions of IBD sharing are error-free fragments, and will therefore be completely identical. GERMLINE therefore divides the input sequences into shorter fragments or "slices" and creates a dictionary of allele-combination "words" for each haplotype sample. The dictionary is used to identify exact matches between individuals on a slice-by-slice basis. GERMLINE then uses a dynamic programming algorithm to extend the pairwise matches along neighboring slices while allowing for a small number of mismatches; identifying shared segments of a desired minimum length. The goal of this study is to demonstrate the accuracy and efficiency of GERMLINE in large samples, and establish the usefulness of the identified segments in understanding the landscape of genetics – specifically, identifying structural variation and recovering haplotype phase.

The paper is structured as follows. We first present the GERMLINE analysis results in several varied populations; we analyze GERMLINE's efficiency and accuracy as compared to state-of-the-art segmental sharing applications; we then explore the performance of GERMLINE in a large and unphased population; finally, we detail a novel application of IBD segments to identifying phasing error and structural variation. The algorithm is presented in detail in the Materials & Methods section; we first introduce the hashing approach to identifying whole haplotype segment sharing; we then extend this algorithm to analyze smaller slices and merge partially matching contiguous slices into long segments; next we present an approach for inferring phase from genotype data in the presence of a partial pedigree; and we describe an iterative model under which GERMLINE can use discovered IBD segments to infer missing data.

2 Results

2.1 Comparison with other methods

To determine the effectiveness of the word matching algorithm in phase-known data, GERMLINE was used to identify IBD in the HapMap Phase II phased release. We compare the results with those of the PLINK whole genome data analysis toolset (Purcell; (Purcell, Neale, Todd-Brown, Thomas, Ferreira, Bender, Maller, Sklar, de Bakker, Daly et al. 2007), which can detect extended IBD with the "segmental sharing" runtime option. PLINK uses a hidden Markov Model (HMM) approach to estimate multipoint probabilities of IBD in pairs of individuals based on IBS sharing. Like GERMLINE, PLINK can report all identified shared segments as well as the samples and markers they are involved in. We calculated two sets of results for the segmental sharing option; a default under which every available SNP was processed, and a pruned set which excluded highly linked SNPs in accordance with a strategy suggested in PLINK documentation (Purcell).

Table 2 shows runtime for GERMLINE as well as the two comparison implementations on the HapMap cohorts. All results represent the sum of independent chromosome runs, with minimum segment length set at 5-megabase (Mb). Running on all SNPs, GERMLINE significantly outperforms PLINK – requiring approximately 40 minutes for the entire analysis compared to just under 35 hours with PLINK. When PLINK is run on the pruned datasets, consisting of approximately 3.2% of the total SNPs, the genome-wide GERMLINE computation still requires approximately 3-fold less time (pruning time not included).

In the absence of ground truth data, it is difficult to fully measure the accuracy and completeness of identified segments. We did, however, compare our results to a relatedness study by the HapMap Project (Frazer Ballinger Cox Hinds Stuve Gibbs Belmont Boudreau Hardenbol Leal et al. 2007) to gauge overall effectiveness. The HapMap Project used a thoroughly pruned set of SNPs with the PLINK HMM implementation to identify segmental sharing and presented their findings as population-wide statistics. Replicating the HapMap experimental set-up, we searched for segments over 1Mb in length and at least 50 SNPs. Table 3 gives the distribution statistics for segments discovered by GERMLINE and PLINK in each of the four HapMap populations, as well as identity by state statistics available only for GERMLINE. In all sets, GERMLINE identified significantly more segments and maintained a higher total distance spanned at a near perfect IBS rate. The increased sensitivity of GERMLINE to detect shorter shared segments is evident by the lower mean length of a detected segment. Furthermore, GERMLINE is sensitive enough to detect breaks in IBD matches - breaks that have real biological meaning, as explained below. These effects further reduce the mean and maximum length of detected segments while increasing accuracy.

2.2 Performance on unphased or partially phased data

We assessed the accuracy of GERMLINE in unphased genotypes on the Kosraean population data. For each individual, GERMLINE accepted as input the genotype sequence and parent identifiers, where available. Based on performance assessments, slice size was fixed at 32 SNPs with words considered "ambiguous" and unprocessed if they contained more than 6 unphased SNPs (see Methods – *4.2.4*). Because of the tight relatedness of the sample set, GERMLINE searched for shared segments over 10-centimorgan (cM) in length according to a genetic database (Duffy 2006) which compiled the results of several standard genetic maps (Kong, Murphy, Raj, He, White and Matise 2004; (Lien, Szyda, Schechinger, Rappold and Arnheim 2000).

Table 4 shows genome-wide segment discovery and data resolution statistics as a percentage of all slices, with individuals grouped by number of parents and children. We observe that the pedigree is representative of typical heterozygosity rates, with 56% of all slices being ambiguous, and can gauge how effective GERMLINE is on genotype data based on how much of this ambiguity it can resolve (for genotype resolution, see Methods – *4.2*). Three measures were used to evaluate unphased data resolution: the percentage of ambiguous slices in the raw data set, the percentage remaining after Mendel-based phasing, and the percentage remaining after GERMLINE identified and imputed from all shared segments (post Mendel phasing); we also present the percentage of slices which contained at least one shared segment. Overall, GERMLINE was able to resolve 77% of the ambiguous slices while finding shared segments in 63% of the entire sample. In individuals with at least one known parent (partially phased data), GERMLINE resolved 96.5% of the ambiguous slices, leaving only 2% of the entire sample unprocessed. However, in those samples that had no known relatives in the pedigree (unphased data), GERMLINE was able to resolve very little of the ambiguous slices, leaving 56.3% of all slices ambiguous after matching and finding shared segments in only 1.3% of the entire sample. Nevertheless, of the 567 individuals with no parents or children, all but 6 were identified in at least one shared segment and thereby connected into the pedigree.

Although we attempted to confirm these results with the PLINK segmental sharing algorithm, it was only able to identify whole-chromosome sharing. In its current form, the PLINK implementation may still need specific tuning in order to resolve relatedness in the Kosraean data set. Computationally, PLINK required 556 hours to complete analysis of the 8 shortest chromosomes while GERMLINE processed the same data in 130 hours.

2.3 Segmental gaps

In identifying IBD on the HapMap data, we discovered a number of long shared segments which were broken up by short regions (generally less than 100 SNPs) that contained unusually low IBS, referred to as "gaps". We suspected these gaps to be indicative of phasing errors or structural variation. We demonstrate that, indeed, IBD gaps come in these two flavors, as manifest by their allelic makeup.

Each individual in the HapMap phased data is considered to have one Transmitted and one Untransmitted haplotype relative to its child, referred to henceforth as the T and U haplotypes respectively. In regions of high heterozygosity, we identified gaps in which putative IBD switches in an individual from one haplotype to the other and then back at the end of the gap. This pattern may be explained by phasing inconsistency which results in incorrectly oriented haplotypes at heterozygous sites, commonly referred to as "switch error" (Lin, Cutler, Zwick and Chakravarti 2002; (Marchini, Cutler, Patterson, Stephens, Eskin, Halperin, Lin, Qin, Munro, Abecasis et al. 2006). Table 5 provides an example of three contiguous segments on chromosome 18 between individuals NA06993 and NA07056 (CEU population) with IBS measurements taken for the two pairs of notable haplotypes. In the first region, the two individuals are in nearly complete IBS along their respective Transmitted haplotypes; in the subsequent 42 SNP gap, none of the 16 heterozygous positions continue the shared segment, but rather they match the Transmitted haplotype of NA06993 with the Untransmitted haplotype of NA07056; in the remainder of the shared segment, the IBS switches back to the two Transmitted chromosomes. This IBS switch back and forth is consistent with two rapid recombination sites during the NA07056 meiosis. However, looking at the genotype data reveals that all samples in the trio involving NA06993 are heterozygous at the 16 positions, implying that the phasing is completely computational and not forced by Mendelian relationships; the lack of direct information makes such regions particularly ripe for short phasing errors. Searching in the two HapMap cohorts with known trio data for gaps in which IBD mismatches were contained to heterozygous sites and phasing was not based on familial information, we identified 58 such regions.

Another class of gaps is characterized by regions of unusually low heterozygosity, suggestive of structural variation. A region which exhibits loss of heterozygosity may, in fact, be representative of incorrectly typed hemizygosity resulting from a segmental deletion along the otherwise shared haplotype. Such regions would also feature relatively low identity rates because SNP matching is effectively being counted on the alternate haplotypes to those in actual IBD. We searched through the HapMap samples for gaps in long segmental sharing which exhibited this characteristic of loss of heterozygosity as well as a high rate of IBS mismatches. Table 6 documents an example of two such regions between pairs of individuals in the CEU population. In the first region (NA12264 and NA12155), two shared segments over 6,000 SNPs in length straddle a 44 SNP gap that exhibits loss of heterozygosity and a decrease in IBS. Similarly, the second region (NA12717 & NA11840) contains two shared segments of over 5,000 SNPs in length straddling a 14 SNP gap that exhibits loss of heterozygosity and nearly complete lack of identity. The large size of shared segments essentially guarantees these regions to be IBD. This assumption, coupled with significant loss of heterozygosity in only one individual of the pair suggests that such gaps are de novo segmental deletions. To validate these supposed deletions, we attempted to find overlapping deletion regions in the Database of Genomic Variants (Iafrate, Feuk, Rivera, Listewnik, Donahoe, Qi, Scherer and Lee 2004), which is a summary of structural variation identified in a number of studies, including experimental examination of structural variants in the HapMap cell lines (Redon, Ishikawa, Fitch, Feuk, Perry, Andrews, Fiegler, Shapero, Carson, Chen et al. 2006). Overall, we identified 14 gaps in IBD which formed runs of consecutive homozygous SNPs and were able to confirm 8 in such a manner.

We further explored the discovery of segmental deletions in the Kosraean population data where IBD is more prevalent and we observe a larger number of gaps. We used a binomial score to rank potential deletions in homozygous gaps based on the number of mismatching SNPs and the rate of mismatch in the flanking shared segments, measured across all shared segments with the suspected gap (see Appendix A2). Figure 1 shows an example of such a gap, plotting the normalized fluorescence intensity measures across a 2Mb region containing the putative deletion. This region clearly coincides with a significant decrease in intensity values, supporting the hypothesis. We validated the 200 most statistically significant gaps with three means of verification: (1) the Affymetrix Copy Number Analysis Tool (Huang, Wei, Zhang, Liu, Bignell, Stratton, Futreal, Wooster, Jones and Shapero 2004), which processes fluorescence intensity in an HMM-based algorithm to identify blocks of structural variation common to many individuals, (2) examining deviations from the average in normalized fluorescence intensity values, which can help identify deletions that are too short or uncommon for CNAT, and (3), overlap with deletions reported in the Database of Genomic Variants. Figure 2 reports validation by these criteria: 120 segments were verified in the Database of Genomic Variants (dbGV), 65 showed significant deviations in intensity (Intensity), and 17 were identified by the Affymetrix Copy Number Analysis Tool (CNAT). These results suggest that IBD-based search for deletions is more sensitive than other existing tools for SNP array data.

3 Discussion

We present GERMLINE, a program for genome-wide discovery of IBD shared segments in large populations. We introduced a linear-time time algorithm for identifying short identical genomic "slices" between pairs of individuals and then extending the boundaries of these slices to identify long shared segments representative of IBD. We also implemented strategies for error-robust haplotype inference from genotype data in the presence of pedigree information – utilizing discoverable IBD to fill in the gaps of missing or unphased markers. The program is specifically intended for analyzing large and complex data-sets.

We established the stability and accuracy of GERMLINE on the four diverse populations of phased haplotype data from the HapMap. GERMLINE proved to be several orders of magnitude faster than the state-of-the-art segmental sharing detection of PLINK in genome-wide analysis, with linear scalability that enables analysis of thousands of samples. Likewise, GERMLINE maintained higher efficiency in comparison to PLINK analysis of a significantly pruned set of SNPs. On the same data, we have also shown that GERMLINE maintains a high level of accuracy as compared to the results of segmental sharing analysis completed by the International HapMap Consortium. In particular, GERMLINE identified a larger number of short shared segments – a self-described weak-point of the segmental sharing identification by the HapMap Consortium.

The accuracy and efficiency of GERMLINE on phase data motivated IBD analysis of genotype data from the significantly larger and more densely related population of Kosrae, Micronesia. Running GERMLINE on a pedigree of nearly 3,000 individuals, we were particularly successful in identifying segmental sharing for individuals with at least one direct relative. We did find that GERMLINE underperformed in individuals with no known relatives, chiefly because their genotypes have a significant number of data points with unresolved phase, making it difficult to identify initial IBD segments used for further inference. Nevertheless, GERMLINE was able to identify some amount of sharing in all but six of the total sample set. Overall, we found that GERMLINE inferred 87% of the unphased genotypes and identified shared segments in 67% of the total genomic data.

A novel result of our IBD analysis in various populations was the identification of short "gaps" in long IBD shared segments. We hypothesized that these gaps were indicative of phasing error or structural variation, and confirmed this in many of the identified cases. A number of these gaps were found to contain IBD loss only in heterozygous sites that could not be phased according to Mendelian rules, suggesting that incorrect statistical phasing generated by the PHASE algorithm was the cause. In such instances, GERMLINE could potentially provide the correct phasing – one which would maintain identity between the pair of individuals. The remaining gaps showed a propensity for loss of heterozygosity, suggesting that the presence of deletions resulted in loss of IBD. We were successful in verifying these deletions by examining probe intensity values, CNV discovery algorithm results, and previous experimental findings. It is important to note that many of the putative deletions we discovered were not identified by the CNV detection algorithm, presumably because they were present in a minority of the population and deletion intensity does not deviate as significantly from the norm as other multiple copy variation. The IBD-based approach can fill this shortcoming by identifying de novo deletions that are not widely evident. This original strategy of using IBD sharing to detect polymorphic deletions is a promising tool for association analysis of such microdeletions, that are being recognized as important targets for disease association study (Kumar, KaraMohamed, Sudi, Conrad, Brune, Badner, Gilliam, Nowak, Cook, Dobyns et al. 2008; (Weiss, Shen, Korn, Arking, Miller, Fossdal, Saemundsen, Stefansson, Ferreira, Green et al. 2008). Moreover, unlike traditional computational methods for CNV detection our analysis does not depend on highly variable and noisy probe intensity data and can be used to complement newer array technologies that include CNV probes. Lastly, the potential for discovering the short regions of IBD-loss which underlie putative deletions is explicitly driven by the precision and accuracy offered by GERMLINE.

Looking ahead, as genotyping data volume continues to increase, the presence of such hidden relatedness will become ubiquitous. With GERMLINE, we have overcome the computational burden of pairwise sample analysis by developing an algorithm which scales linearly with the number of input individuals, and facilitates scanning large cohorts for IBD. Resolving such knowledge of hidden relatedness has been shown to add statistical power to heritable trait association mapping (Almasy and Blangero 1998; (Dodds, Amer and Auvray 2007; (Meuwissen and Goddard 2007). Furthermore, we project that mapping of IBD segments would enable linkage analysis across distantly related samples (Albers, Stankovich, Thomson, Bahlo and Kappen 2008), and extend this paradigm, currently very powerful in pedigrees (Lander and Kruglyak 1995), to study tens and hundreds of thousands of unrelated individuals. Population-based linkage analysis would integrate the strength of association studies in collection of unrelated samples with the powers of linkage to detect rare variants. The data made available from GERMLINE analysis is thus a step towards comprehensive disease mapping in large populations.

4 Materials

Haplotypes are the fundamental components of DNA inherited through generations; therefore, we devise a search for IBD that is based on directly matching portions of haplotypes between individual samples. Such a search is naturally simpler if shared segments are identical throughout the entirety of the haplotype considered, allele calls are error-free, and the phase of input sequences is known. This simple case facilitates direct matching of haplotypes to one another, and we first present GERMLINE, our methodology for efficient IBD detection, in such a demonstrative scenario. We will then introduce realistic complexities of segment-limited matching, data errors and unphased genotype data one by one. We present extensions to the fundamental GERMLINE algorithm to handle such realistic information and still detect IBD and hidden relatedness.

4.1 Haplotype IBD matching

Our methodology for discovering identity by descent within haplotype data is based on an associative structure or *dictionary* in which we store, in time proportional to the size of the data, a list of duplicate words representing haplotype portions. Input is provided as phase-known haplotype data and output is a list of shared segments and the corresponding chromosome pairs that match along them.

4.1.1 *Problem formulation*

Formally, given the input set of haplotypes for n individuals and s SNPs along a genomic segment of interest, we accept as input a $2n \times s$ matrix $\mathbf{H}_{\text{real}}$ with rows corresponding to haplotypes and columns to SNPs. Haplotype calls are represented by a binary alphabet corresponding to the alleles and a third character for missing data. Due to errors and noise that are likely to be present in the input, we will denote **H** to be the clean binary matrix of true haplotypes underlying $\mathbf{H}_{\text{real}}$. A matrix entry $\mathbf{H}[i,j]$ is 1 if haplotype i carries the minor allele of SNP j, and 0 otherwise. Each row of **H** can therefore be regarded as a binary vector. When two haplotypes (i,i') are IBD , the corresponding rows (i,i') will be identical in **H**. The goal of the algorithm is therefore to accept a matrix **H** and output a set **L** of IBD shared segments (i,i''). In this section we will define the algorithm using **H** and in subsequent sections present a reduction from $\mathbf{H}_{\text{real}}$ to **H**.

We begin by defining an algorithm for identifying matches across the s SNPs in matrix **H**. We rely on a dictionary of haplotypes: the set **D** of size no larger than $2n$ consisting of non-redundant rows from **H**. **D** is implemented as a hash-table data structure with constant-time insertion and lookup: the key is a binary vector of length s, and the value is a set of individual haplotypes having identical rows. Once **D** is constructed, each pair of rows indexed by the same key in **D** is a match. The set **M**(**H**) of all such matches can be obtained by traversing all keys in **D**, and all pairs of rows per key as in MATCH (Algorithm 1).

4.1.2 *Identical matching across subsets of **H***

When considering a large fraction of a chromosome, true IBD may not span all of the available SNPs. In a whole-genome or whole-chromosome context we are therefore interested in detecting partial matches - pairs of individuals that share a common recent ancestor only along a segment of a chromosome. As such, we establish a defined threshold on the minimum length for an IBD shared segment. The choice of L_{IBD} corresponds to the expected segment length for the most distantly related individuals we still aim at detecting (see Table 1). Formally, we define the length $\mathrm{L}(j,j')$ of an interval between columns j and j' as the genetic distance between the corresponding genes. A pair of samples (i,i') is assumed to form a shared segment in a SNP region $[j,j']$ if their included SNPs are identical and $\mathrm{L}(j,j')$ supersedes L_{IBD}. We now propose a divide & conquer strategy for using MATCH to discover such shared segments. Our goal is to identify pairs of long identical segments that are shorter than the length of **H**. We can approximate this by dividing the columns of **H** into equal vertical intervals, or *slices*, and finding pairs of individuals that are matching along contiguous slices. We thus distinguish between a *haplotype*, which is an entire row of H and a *word* that represents the part of a haplotype that intersects a slice of H. A match between two individuals along several slices can be considered *extended* if the words of these individuals also match in the succeeding slice, otherwise the match *terminates.* A shared segment can thus be redefined as belonging to a pair of individuals that extend across several word pairs, and will represent an IBD segment rounded to the nearest slice break.

Formally, our algorithm accepts as input **H** and iteratively uses MATCH to generate a set **M′** of all shared segments in **H**. We vertically slice **H** into non-overlapping, equal width sub matrices of δ columns, with each slice denoted as $\mathbf{H}_k$. The algorithm is a dynamic program that scans slices along the chromosome and maintains sets of the terminated and extendible matches within the current slice. To avoid redundant conversion from SNPs to slices, we will henceforth refer to a match $(i,i'; j,j')$ as (i,i', m,m') such that $j=m\cdot\delta$ and $j'=(m'+1)\cdot\delta$-1. At each slice $k=0\rightarrow(s/\delta)$-1, we compute independent $\mathbf{M}_k$=MATCH($\mathbf{H}_k$), the set of identical matches at k. As detailed in EXTEND (Algorithm 2), we extend complementary matches in neighboring slices by examining $\mathbf{M}_k$ and generating a corresponding set $\mathbf{M}_k'$ that contains all extended matches; naturally, each extended match contains m, the start of the match, in the range 0 to k-1 and m', the end of the match, equal to k. In the first slice, define $\mathbf{M}_0'=\mathbf{M}_0$; subsequently, initiate $\mathbf{M}_k' = \mathbf{M}_k$, and search through all matches $\mathbf{M}_k'$ for extendable matches from $\mathbf{M}_{k-1}'$. Where they exist, we updated the starting position for matches in slice k to be that of the match in slice k-1. Similarly, terminated matches present in $\mathbf{M}_{k-1}'$ but not $\mathbf{M}_k$ are either discarded or added as IBD to **M′** depending on their length. Upon completing the final iteration, all matches in $\mathbf{M}_{(s/\delta)-1}$ are discarded or added to **M′** in the same manner.

4.1.3 Genotyping error

Up until this point, we have ignored the effect genotyping errors may have on identifying matches. While modern genotyping platforms achieve accuracy levels >99% (Paynter, Skibola, Skibola, Buffler, Wiemels and Smith 2006), across many slices the chance of an error becomes non-negligible even in a single sample. Across thousands of samples and along densely typed complete chromosomes the presence of errors approaches certainty. For IBD matching, random errors are unlikely to produce false-positives. We are, however concerned about error-induced false negatives. Intuitively, a true IBD match would be present in several consecutive slices and may be detected by matching in any of them. Assuming random, uniform, and independent error rate ε per SNP, the number of mismatching allele calls between a pair of IBD haplotype intervals of length δ SNPs is Poisson distributed with parameter λ= 2δε. Across an IBD segment of length s_{IBD} SNPs (where L(s_{IBD})=L_{IBD}), the expected number of matching slices can be modeled as

$$E(N_{\mathrm{IBD}}) = \left\lfloor \frac{s_{\mathrm{IBD}}}{\delta} \right\rfloor e^{-2\varepsilon\delta}$$

and typically being flanked by nearly identical slices. Specifically, the chance of these words to be identical is (1-ε)$^{2\delta}$, with the length of a complete observed match in an IBD region thus geometrically distributed.

4.1.4 Nearly identical matching

In practice, if δ is set to such a value that E(N_{IBD}) is much larger than 1, potentially long IBD segments can be discovered by searching for identical matches contained therein. The entire segment can then be identified by merging with the nearly identical flanking matches. We use ε to determine a threshold for allowed mismatches in otherwise identical intervals that is low enough to eliminate false positive matches. Specifically, if error rate ε is conservatively assumed to be 0.01; minimum length for IBD is 2,000 SNPs; and slice length is 100 SNPs; the expected number of matching intervals is 2.7 and 1 mismatching bit is allowed in each slice. The haplotypes corresponding to such slices are defined as being nearly identical. In MERGE-PARTIAL (Algorithm 3), we modify EXTEND to detect nearly identical matches in the input data by implementing a post-processing step that adds nearly identical matches to **L**′. Algorithm 3 details how each terminal match from the previous slice may additionally be extended by a nearly identical match in the current slice. Implemented at each slice, this procedure effectively maintains contiguous partially matching segments as long as they were initially added to some $\mathbf{L}_k'$ as a result of an identical match.

We can now integrate all of the modules into a complete procedure which accepts as input matrix **H** and returns the set **L** of all contiguous identical or nearly identical matches in **H**, based on predefined length and mismatch thresholds. As described in Algorithm 4, **H** is divided into slices that are analyzed sequentially: at each slice, use MATCH to compute a list of identical matches within that interval, use MERGE to extend these matches with those of the preceding interval, and use MERGE-PARTIAL to find terminal matches in the previous slice that can be extended with a partial matches in the current. Matches from previous slices that are not extendable are discarded or stored in **L′**, which is the set of shared segments returned after all slices have been processed.

4.1.5 Algorithm complexity

Computationally, the algorithm has a significant gap between worst-case and typical scenarios. Specifically, the time complexity is highly dependent on expected number of matches, which is determined by the underlying population structure. In general, if the average number of matches per word is *m*, the complete computation requires O(*sn*) to build the dictionary and O(*sm*) to attempt extension on all matches. In the worst case, where **L** is formed by the Cartesian square of the 2*n* haplotypes, m=4n^2 and GERMLINE is comparable to pairwise exhaustive search. In practice, if we make a naïve assumption of independence of sites, the expected number of matches occurring at a slice at random is

$$\binom{n}{2}(p_s{}^2 + q_s{}^2)^{h_{\text{len}}}$$

where p_s and q_s are the allele frequencies. A more realistic assumption would acknowledge local LD within each segment. If we denote a set of population haplotype frequencies **f** of size f_{n}, the expected number of matches occurring at a word is

$$\binom{\frac{nf_n}{2}}{}\sum_{i=0}^{f_n}\mathbf{f}(i)^2$$

Even in large samples sizes this factor is low enough where overall complexity approaches O(*sn*).

4.2 Genotype IBD matching

We now consider a more difficult version of IBD matching where completely phased data is unavailable. Instead of haplotypes, the input consists of genotype calls which are either homozygous to a particular allele, or heterozygous. In the presence of genotype data only, GERMLINE utilizes several approaches to infer as much of the input haplotype as possible by leveraging knowledge of the pedigree structure in conjunction with Mendel's rules of inheritance. This is a solution that is particularly effective in densely related data with a resolved pedigree structure (Baruch, Weller, Cohen-Zinder, Ron and Seroussi 2006; (Marchini, Cutler, Patterson, Stephens, Eskin, Halperin, Lin, Qin, Munro, Abecasis et al. 2006; (Qian and Beckmann 2002; (Sobel and Lange 1996). We first describe extraction of such partial phasing information from a given pedigree, and then present GERMLINE as extended to partially phased data.

4.2.1 *Single locus phasing*

We begin by examining families of two parents and one child, known as *trios*. Initially, we will use the pedigree structure to infer haplotype information at a single locus. We will then extend the algorithm to inference on multiple unlinked markers. We introduce a sample trio of genotypes $\mathbf{p_1}$,$\mathbf{p_2}$, and $\mathbf{c}$; with haplotypes $\mathbf{p_{11}}/\mathbf{p_{12}}$, $\mathbf{p_{21}}/\mathbf{p_{22}}$, and $\mathbf{c_1}/\mathbf{c_2}$. In accordance with Mendel's rules, we establish the convention that $\mathbf{c_1}=\mathbf{p_{11}}$, $\mathbf{c_2}=\mathbf{p_{21}}$, and that a heterozygous genotype represents two non-identical haplotypes. With these restrictions, the presence of any homozygous individual provides enough information to resolve all family members at that marker. Let us consider the three potential cases where a single family member is homozygous. If $\mathbf{c}$ is the only homozygous individual with allele a (where a' denotes the alternate allele), we can immediately set $\mathbf{c_1}=\mathbf{c_2}=a$; by convention, we can also set $\mathbf{p_{11}}=\mathbf{p_{21}}=a$; and $\mathbf{p_{12}}=\mathbf{p_{22}}=a'$. Likewise, if only $\mathbf{p_1}$ is homozygous with allele a, we set $\mathbf{p_{11}}=\mathbf{p_{12}}=a$; as well as $\mathbf{c_1}=a$, $\mathbf{c_2}=a'$; and finally $\mathbf{p_{21}}=a'$ and $\mathbf{p_{22}}=a$. An identical solution holds for the third case, where $\mathbf{p_2}$ is homozygous. In this way, the entire trio can be phased unambiguously from one homozygous individual. Intuitively, these rules are also true for trios with more than one homozygous member.

4.2.2 *Multi-locus phasing with recombination*

Such an approach is effective on a per-locus basis; however, we cannot extend the identified inheritance pattern across multiple loci in the presence of recombination. We accommodate for this fault by loosening our restrictions so that they can be reassigned on the fly. We again begin with the principle that each parent has two haplotypes. For each parent-child transmission, and at each site, we define a bit that selects which of the parents' haplotypes is being transmitted. In the event of recombination which results in inconsistency between the parent and child haplotypes, this transmission bit is flipped between the two unlinked loci. With one degree of freedom in each parent, we can assume without loss of generality that the first transmitted haplotype is '0' from each. In an isolated trio, both parents can be considered "founders," which results in inherent ambiguity when defining parental haplotypes as Transmitted or Untransmitted as they can generally be classified as both. Using a variable transmission bit, which essentially identifies the grand-paternal Transmitted haplotype, is a way to resolve this.

To accomplish this, GERMLINE constructs a directed acyclical graph from the pedigree, with nodes corresponding to individuals and edges representing direct relation/inheritance. By placing the individuals within a connected graph rather than as isolated trios, phase can be propagated across multiple generations. Every individual is initiated with a default transmission bit setting representing the grand-paternal inheritance pattern. GERMLINE then traverses each node in lexicographical order from parents to children and examines the genotype of the corresponding individual. If a node has known parents and is heterozygous at a marker, GERMLINE attempts to phase the node according to the transmission bit as described previously. If such a phasing results in Mendelian inconsistencies, the marker is either indicative of a recombination event or genotype error. Because IBD matching is especially susceptible to single SNP errors, a recombination event is preferred and the transmission bit is flipped if doing so eliminates the violations. Otherwise, the genotype error is reported and left untreated. Although this dependence on Mendel errors in order to identify a recombination may be inexact for general phasing, it does not interfere with IBD identification which is primarily conducted on the scale of centimorgans.

4.2.3 Phasing without parents

In cases where the family pedigree is incomplete, an individual with a large number of children also holds the potential to resolve phase. In such an individual, we attempt to resolve instances when an inconsistency occurs between the transmission pattern of preceding loci and the genotypes of the current locus. By using a simple voting scheme across the offspring, we can determine the more likely phasing for the family. Each child votes on the phasing for the individual that is consistent with its inheritance pattern; a difference in vote indicates either a recombination event or genotype failure which resulted in a minority of offspring being called for the deviant haplotype. Because a recombination to multiple offspring is extremely unlikely, multiple votes for both potential phases are likely to be the result of a genotype error in the individual or at that marker in general. In this case, GERMLINE reports and ignores the site. Conversely, all but one vote for a specific phasing implies that the alternative was a result of recombination to the minority offspring and its inheritance pattern is flipped.

4.2.4 Partially phased IBD matching

Hereafter, we assume that the input data has been processed according to the aforementioned inference methods, resulting in the phasing of some heterozygous sites – we will refer to these, as well as the phase-known homozygous sites as *unambiguous*. We now extend the haplotype IBD matching algorithm to allow for such partially phased data. Let us return to the matrix $\mathbf{H}^{\text{real}}$, which is structurally identical to $\mathbf{H}$ but represents haplotypes with an alphabet of three characters – one each for the major and minor alleles, and a third for missing or unphased loci (referred to as the *ambiguous* character). Specifically, an individual with a missing or unphased locus will have the ambiguous character in both haplotypes at that column. We will accept the matrix $\mathbf{H}^{\text{real}}$ as input and reduce it to an $\mathbf{H}$ such that it can be used with Algorithm 4 to generate an output list of shared segments.

Conceptually, this reduction is done by exhaustively adding to $\mathbf{H}$ all phases that could have arisen from $\mathbf{H}^{\text{real}}$ without contradicting the unambiguous sites. As this will naturally result in many fictitious haplotypes being added to the dictionary, we follow-up by pruning based on detected IBS segments that further resolve the haplotype content of the data. Formally, we reconstruct the matrix of haplotypes $\mathbf{H}$ with rows that correspond to the rows of $\mathbf{H}^{\text{real}}$ permuted at the ambiguous sites. For one row in $\mathbf{H}^{\text{real}}$, each slice k contains $a(k)$ ambiguous bits and $2^{a(k)-1}$ corresponding potential haplotype phasings. Consequently, when $a(k)$ is very small, it is practical to insert each of these permutations as rows in $\mathbf{H}_k$. As detailed in MATCH-GENOTYPE (Algorithm 5), we fix a threshold α for maximum ambiguous bits, then for each row in $\mathbf{H}^{\text{real}}$ where $a(k) \leq \alpha$: generate $2^{a(k)-1}$ unambiguous potential haplotypes by permuting the value at each ambiguous bit and add each haplotype as a row to $\mathbf{H}_k$. We continue in this manner across each slice, generating $\mathbf{H}_k$ which are then used in further subroutines in exactly as with completely phased data.

Individual slices that have prohibitively high $a(k)$ must be ignored in the exact matching step, but we can make a modification to the DISTANCE function in Algorithm 3 which allows them to be used in nearly identical matches. Rather than just compute a bitwise XOR between the two haplotypes, as is done with unambiguous haplotypes, we mask out the ambiguous bits and compare only those that have been resolved. This procedure essentially ignores the ambiguous bits in nearly identical matching. As such, we must also ensure that a statistically significant number of bits are actually being considered in the distance comparison. To do so, we maintain a threshold for allowed masked bits, on the combined $a(k)$ of the compared haplotypes. The threshold value is calculated from the allele frequencies for all individuals at slice k to represent the minimum number of bits required for a non-random match at that slice. By loosening the distance algorithm, we allow some ambiguous individual slices to be used in nearly identical matching.

The remaining subroutines remain identical in implementation to those used for phased data matching. Overall, this process for genotype IBD matching is conceptually identical to haplotype matching but conducted over a sample space of $O(2^{\alpha}n)$.

4.2.5 Haplotype inference from IBD matches

We now detail the complete algorithm which integrates partial phasing with IBD matching. Because we are identifying IBD that is within relatively few generations, it is highly likely that any difference between haplotypes is due to genotyping error or incomplete phasing. Therefore, GERMLINE attempts to further infer phase data directly from the identified shared segments based on the assumption that long, nearly phased segments indicative of IBD in fact imply fully identical haplotypes. In practice, for each individual with long matches, the algorithm imputes ambiguous bits from the corresponding matching haplotype where available. Let us return to the set **M** of pairs of matching haplotypes detailed in Algorithm 5. After GERMLINE has computed and extended all matches, each haplotype i has a corresponding set of matches $\mathbf{M}(i)$ (subset of **M**) where every member of $\mathbf{M}(i)$ is a match which includes i. GERMLINE traverses each $\mathbf{M}(i)$, examining the matches in the order they appear within the genome; here, let every match at slice j be contained within the set $\mathbf{M}_j(i)$. Each member of $\mathbf{M}_j(i)$ is used for imputation by directly copying all bits which are phased in that member but not in haplotype i_j. The elements are chosen in increasing order of match likelihood $L(m)$ (see Appendix A1). After all matches have been inferred from, the newly resolved information is propagated throughout the pedigree according to the Mendelian rules described previously. With newly resolved information resulting in fewer ambiguous haplotypes, GERMLINE restarts the matching process to discover IBD shared segments more precisely and across a wider set of usable haplotype slices. The complete algorithm integrates phasing and IBD matching in an iterative process to infer the underlying haplotype data.

4.3 Implementation

GERMLINE was implemented in the C++ language as described. All experiments were conducted on a cluster computer with each node running two 2.4 GHz Intel Pentium 4 Xeon CPUs and 2 GB of memory, operating on Red Hat Enterprise Linux 3. No parallelization was used and all runtimes are reported for independent runs. The program source and documentation is maintained at: http://www.cs.columbia.edu/~gusev/germline/

We used the following datasets in our experiments:

- *Pacific islands from Kosrae, Federated States of Micronesia* (Bonnen, Pe'er, Plenge, Salit, Lowe, Shapero, Lifton, Breslow, Daly, Reich et al. 2006). Data on 3,000 individuals genotyped for 600,000 SNP's at Affymetrix and Rockefeller University has been made available to our group through collaboration. Our test set contained 2906 of these individuals and a total of 429,925 markers across 22 chromosomes. The individuals are densely related and known pedigree data for these individuals taken in trio form to aid in partial phasing. Table 3 contains a breakdown of sample size by number of parents and children. The data is made available in (Burkhardt, Kenny, Birkeland, Josowitz, Lowe, Salit, Noel, Pe'er, Daly, Altshuler et al. 2008).

- *International HapMap Project* (Consortium 2005). We used the haplotypes from four panels of individuals in the HapMap Phase II release 21 phased data: 2 parents each from 30 trios in the CEU and YRI populations, as well as 45 unrelated individuals in the JPT and CHB populations. The haplotypes were generated using the PHASE phasing algorithm (Stephens, Smith and Donnelly 2001) on Phase I+II genotyping data files. Table 1 shows the number of SNPs in each panel. Estimates by the HapMap Project show that any two individuals share approximately 0.5% of their genome through recent IBD (Frazer Ballinger Cox Hinds Stuve Gibbs Belmont Boudreau Hardenbol Leal et al. 2007).

Appendix

A1 Match likelihood scoring function

For a partially phased individual that has multiple differing matching haplotypes at a specific slice, it may be necessary to prioritize the matches most indicative of IBD. To this end, we have defined a scoring function, L which accepts a potential match *m* as parameter and calculates the likelihood of the match having occurred at random. The function works based on the assumption that matches which are less likely to be random are more likely to represent true IBD. Although this score is used on a per-slice basis, the calculation inquires along the length of the extended match between the pair involved in *m*. At each slice *k* involved in the extended match, we calculate a ratio between the number of haplotypes matching to *m*, defined as $\text{count}_k(m)$, and the overall number of haplotypes (calculated as the sum of all count_k). The multi-point score is then expressed as follows:

$$\mathrm{L}(m) = \prod_k \left[\frac{\text{count}_k(m)}{\sum_i \text{count}_k(m_i)} \right]$$

where *k* spans the extended slices, and *i* iterates over all non-redundant matches at *k*. All matches to an individual are then ordered by their score in ascending order to determine those most representative of IBD. Practically, L(*m*) favors extended matches that are longer as well as haplotypes that are rarer.

A2 Gap likelihood scoring function

To prioritize the segmental gaps which were most likely to be representative of a deletion, we developed a scoring function based on the number of mismatching SNPs and levels of homozygosity in the gap. Because many gaps occurred in the same position across several individuals, we used all those involved in the calculation to identify gaps which had a higher number of mismatches than expected from the surrounding regions. At each gap, we determined the rate of mismatch in the flanking shared segments and used this as the expected probability *p* for an independent SNP in the region to have a mismatch. After identifying all pairs of individuals that had IBD segments which were broken up by the gap, we counted the number of mismatches *n* occurring in each respective pair within the gap, as well as the total number of SNPs *k* in the gap across each pair. The values *p*, *k*, *n* were then used in the probability mass function for the binomial distribution as follows:

$$\left(\frac{n}{k} \right) p^k (1-p)^{n-k}$$

to score each gap according to how likely the involved mismatches were to have the same rate as the surrounding IBD regions. Although a deletion resulting in a high number of mismatches could occur on either or both individuals in a matching pair, it would be expected to be typed as completely homozygous where it was present. Therefore, to identify the specific individuals that contained the putative deletion, we searched for only those that had nearly complete loss of heterozygosity in the gap region. The final scoring function returns both the likelihood of a gap as well as the suspected individuals involved.

Figure Legends

Figure 1. Candidate Deletion Flourescence Intensity

Population average is shown in red (filled circle).

Individuals identified as having the deletion are shown in blue (open circles).

Dark blue bar represents deletion identified by GERMLINE.

Figure 2. Verified Candidate Deletion Regions (Top 200)

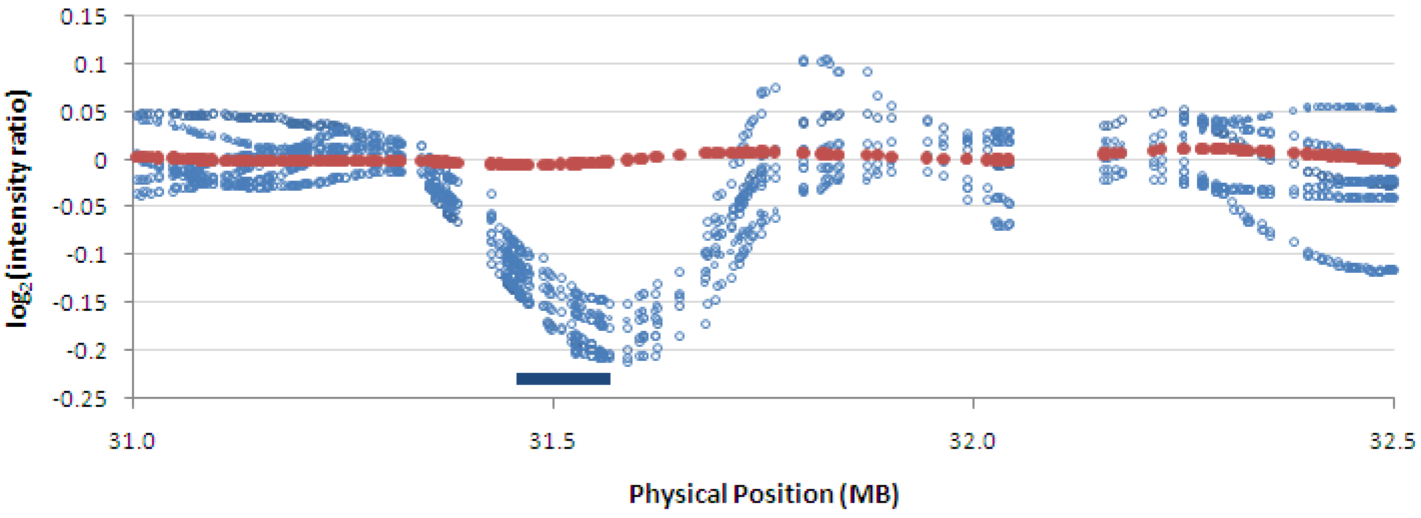
Identified in Database of Genomic Variants shown in blue.

Verified by Affymetrix Copy Number Analysis Tool shown in green.

Verified as deviation from population average intensity shown in red.

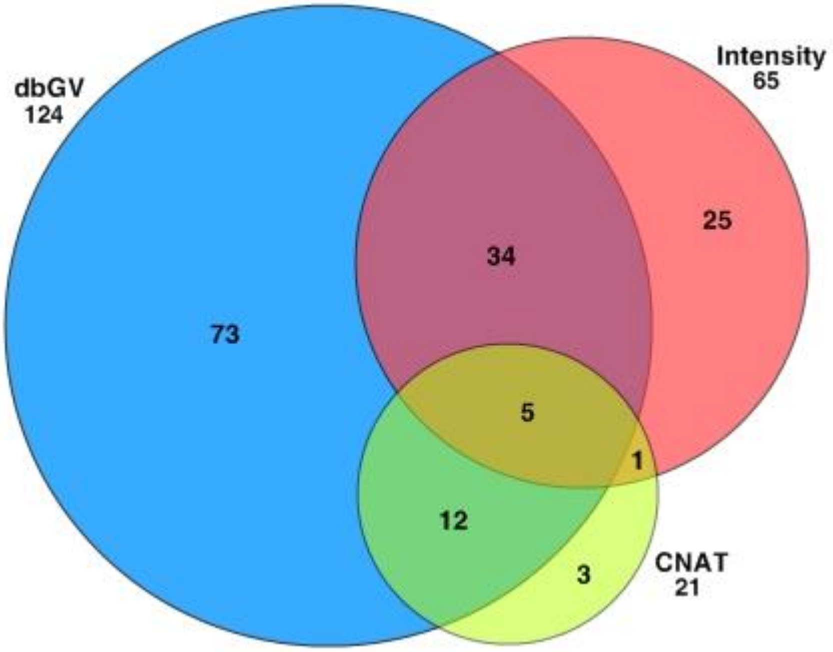
Figures

Figure 1. Candidate Deletion Flourescence Intensity



Population average is shown in red (filled circle).
Individuals identified as having the deletion are shown in blue (open circles).
Dark blue bar represents deletion identified by GERMLINE.

Figure 2. Verified Candidate Deletion Regions (Top 200)



Tables

Table 1. IBD Between Individuals of Shared Ancestry from a Single Source

Meiosis to common ancestor (k)	Likelihood	Expected Segment Length (cM)	Expected # of Potential Segments Genomewide[a]	Expected Total Segment Length (cM)
1 *(half sibs)*	50.00%	50.00	75	1,875.00
2 *(half cousins)*	12.50%	25.00	150	468.75
3	3.13%	16.67	225	117.36
4	0.78%	12.50	300	29.50
5	0.20%	10.00	375	7.40
6	0.05%	8.33	450	1.83
7	0.01%	7.14	525	0.43

[a] Human genome length taken from (Kong, Murphy, Raj, He, White and Matise 2004)

Algorithm 1

```
MATCH(H):
    define set M
    for $h_i$ in H do D.INSERT($h_i \rightarrow i$)
    for $h$ in D do
        for $i$ in D($h$) do
            for $i' \neq i$ in  D($h$) do
                M.ADD($i, i'$)
    return M
```

Algorithm 2

```
EXTEND($\mathbf{M}_{k-1}'$,$\mathbf{M}_k$ ):
    let $\mathbf{M}_k' := \mathbf{M}_k$
    for $m_k$ in $\mathbf{M}_k'$ do
        $i$ := $m_k$.INDIVIDUAL[1]
        $i'$ := $m_k$.INDIVIDUAL[2]
        if $\mathbf{M}_{k-1}'$.CONTAINS($i,i'$)
        then
            $m_{k-1}$ := $\mathbf{M}_{k-1}'$[i,i′]
            $m$.MATCH-START = $m_{k-1}$.MATCH-START
            $\mathbf{M}_{k-1}'$.REMOVE($m_{k-1}$)
    return $\mathbf{M}_w'$
```

Algorithm 3

```
EXTEND-PARTIAL($\mathbf{M}_{k\text{-}1}'$,$\mathbf{M}_k'$):
    given $d$
    for $m_{\text{k-1}}$ in $\mathbf{M}_{k\text{-}1}'$do
        $i$ := $m_{\text{k-1}}$.INDIVIDUAL[1]
        $i'$ := $m_{\text{k-1}}$.INDIVIDUAL[2]
        if DISTANCE($H_{ki}$,$H_{\underline{\text{k}}\text{i}'}$) ≤ $d$ then
            $\text{m}_{\text{k-1}}$.MATCH-END = $k$
            $\mathbf{M}_{k\text{-}1}'$.REMOVE($m_{\text{k-1}}$)
            $\mathbf{M}_{\text{w}}'$.ADD($m_{\text{k-1}}$)
    return $\mathbf{M}_k'$
```

Algorithm 4

```
HAPLOTYPE-IBD($\mathbf{H}$):
    given $s$, $s_{IBD}$ , $h_{len}$
    let $\mathbf{M}_0$ = MATCH($\mathbf{H}_0$)
    let $\mathbf{M}_0'$ = $\mathbf{M}_0$
    define $\mathbf{M}'$
    for $k$ in $1 \rightarrow (s/h_{len})-1$ do
        let $\mathbf{M}_k$ := MATCH($\mathbf{H}_k$)
        let $\mathbf{M}_k'$ := EXTEND($\mathbf{M}_{k-1}'$,$\mathbf{M}_k$)
        $\mathbf{M}_k'$ := EXTEND-PARTIAL($\mathbf{M}_{k-1}'$,$\mathbf{M}_k'$)
        for $m_{k-1}$ in $\mathbf{M}_{k-1}'$ do
            if LENGTH($m_{k-1}$) $\geq h_{IBD}$
            then $\mathbf{M}'$.ADD($m_{k-1}$)
    return $\mathbf{M}'$
```

Algorithm 5

```
MATCH-GENOTYPE($\mathbf{H}^{real}$):
    given $a_k$
    define set L
    for $h_i$ in H do
        if COUNT-AMBIGUOUS($h_i$) $\leq a_k$ then
            for $h_i'$ in PERMUTE-AMBIGUOUS($h_i$) do
                D.INSERT($h_i' \rightarrow i$)
    for $h$ in D do
        for $i$ in D($h$) do
            for $i' \neq i$ in  D($h$) do
                M.ADD($i,i'$)
    return L
```

```
GENOTYPE-IBD($\mathbf{H}^{real}$):
    given $a$, $s$, $s_{IBD}$ , $h_{len}$
    let $\mathbf{M}_0$ = MATCH-GENOTYPE($\mathbf{H}_0^{real}$)
    let $\mathbf{M}_0'$ = $\mathbf{M}_0$
    define M′
    for $k$ in $1 \rightarrow (s/h_{len})-1$ do
        let $\mathbf{M}_k$ := MATCH-GENOTYPE($\mathbf{H}_k^{real}$)
        let $\mathbf{M}_k'$ := EXTEND($\mathbf{M}_{k-1}'$,$\mathbf{M}_k$)
        $\mathbf{M}_k'$ := EXTEND-PARTIAL($\mathbf{M}_{k-1}'$,$\mathbf{M}_k'$)
        for $m_{k-1}$ in $\mathbf{M}_{k-1}'$ do
            if LENGTH($m_{k-1}$) $\geq h_{IBD}$
            then M′.ADD($m_{k-1}$)
    return M′
```

Table 2. Runtime Comparison with PLINK

Population	All SNPs	GERMLINE All	PLINK All	Pruned SNPs	PLINK Pruned
CEU	2,557,252	0:14:07	09:34:26	72,778	0:18:35
YRI	2,856,346	0:14:19	11:04:55	140,919	0:32:20
JPT	2,419,983	0:06:19	06:49:45	56,190	0:08:16
CHB	2,419,983	0:06:19	06:53:26	60,726	0:08:13

Table 3. Shared Segment Discovery in HapMap

Population	CEU		YRI		JPT		CHB	
	GERMLINE	PLINK	GERMLINE	PLINK	GERMLINE	PLINK	GERMLINE	PLINK
Total number of segments	7,120	427	7.842	250	913	273	540	146
Total distance spanned (Mb)	12,744	2,336	15,658	1,416	1,679	1,301	1,108	704
Mean segment length (Mb)	1.8	5.5	2.0	5.7	1.8	4.8	2.1	4.8
Maximum segment length (Mb)	25.9	56.2	29.9	51.7	22.8	25.3	22.8	15.0
Mean identity by state (IBS)[a]	99.8%	-	99.9%	-	99.8%	-	99.8%	-

[a] IBS statistics not available for HapMap relatedness study

Table 4. Resolution of Ambiguity[a] in Sample (% of Slices)

Parents	Children	Samples	Ambiguous initially	Ambiguous after Mendel phasing	Ambiguous after matching	Shared segment found
0	0	567	56.4%	56.4%	56.3%	1.3%
0	1+	582	56.4%	12.2%	4.8%	63.6%
1	0	715	56.5%	20.2%	4.0%	70.0%
1	1+	250	56.1%	3.1%	0.3%	91.6%
2	0	677	56.2%	4.3%	0.7%	90.1%
2	1+	115	56.6%	0.9%	0.1%	96.4%
all	all	2,906	56.4%	19.7%	13.1%	62.9%

[a] An ambiguous slice is defined as having greater than 6 of 32 unresolved heterozygous SNPs

Table 5. Potential Phasing Irregularity in HapMap

Region[a]	SNPs	Heterozygous SNPs	IBS [T/T]	IBS [T/U]
chr18:44042249-48624309	4912	2069	99.5%	36.6%
chr18:48624918-48669808	42	16	0% [b]	100%
chr18:48670200-59773196	11659	5211	99.6%	38.5%

[a] Samples NA06993 & NA07056 – CEU Population

[b] Identity by state measured over heterozygous sites

Table 6. Potential Deletion Regions in HapMap

Region	Homozygous SNPs	IBS	SNPs	length cM
chr11:63543965-74961563[a;b]	53.7%[e]	99.9%	6,950	10.2
chr11:74964989-75019488	100%[a]	40.9%	44	-
chr11:75020500-80863996	58.4%[e]	99.9%	6,004	7.3
chr1: 104546035-110007685[c;d]	50.2%[e]	99.9%	5,180	5.9
chr1: 110007814-110015547	100%[d]	57.1%	14	-
chr1: 110015973-115590284	56.3%[e]	99.7%	5,059	7.7

CEU samples: [a]NA12264, [b]NA12155, [c]NA12717, [d]NA11840

[e]Average over both samples

References

Albers, C.A., J. Stankovich, R. Thomson, M. Bahlo, and H.J. Kappen. 2008. Multipoint approximations of identity-by-descent probabilities for accurate linkage analysis of distantly related individuals. *Am J Hum Genet* **82:** 607-622.

Almasy, L. and J. Blangero. 1998. Multipoint quantitative-trait linkage analysis in general pedigrees. *Am J Hum Genet* **62:** 1198-1211.

Baruch, E., J.I. Weller, M. Cohen-Zinder, M. Ron, and E. Seroussi. 2006. Efficient inference of haplotypes from genotypes on a large animal pedigree. *Genetics* **172:** 1757-1765.

Bonnen, P.E., I. Pe'er, R.M. Plenge, J. Salit, J.K. Lowe, M.H. Shapero, R.P. Lifton, J.L. Breslow, M.J. Daly, D.E. Reich et al. 2006. Evaluating potential for whole-genome studies in Kosrae, an isolated population in Micronesia. *Nat Genet* **38:** 214-217.

Burkhardt, R., E. Kenny, A. Birkeland, R. Josowitz, J. Lowe, J. Salit, M. Noel, I. Pe’er, M. Daly, D. Altshuler et al. 2008. Common SNPs in HMGCR in Micronesians and Caucasians associated with LDLcholesterol

levels affect alternative splicing of exon13. *Proceedings of the National Academy of Sciences (Submitted for Publication)*.

Consortium, I.H. 2005. A haplotype map of the human genome. *Nature* **437:** 1299-1320.

Dodds, K.G., P.R. Amer, and B. Auvray. 2007. Using genetic markers in unpedigreed populations to detect a heritable trait. *J Zhejiang Univ Sci B* **8:** 782-786.

Donnelly, K.P. 1983. The probability that related individuals share some section of genome identical by descent. *Theor Popul Biol* **23:** 34-63.

Duffy, D.L. 2006. An integrated genetic map for linkage analysis. *Behav Genet* **36:** 4-6.

Frazer, K.A. D.G. Ballinger D.R. Cox D.A. Hinds L.L. Stuve R.A. Gibbs J.W. Belmont A. Boudreau P. Hardenbol S.M. Leal et al. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449:** 851-861.

Hill, W.G. and J. Hernandez-Sanchez. 2007. Prediction of multilocus identity-by-descent. *Genetics* **176:** 2307-2315.

Hill, W.G. and B.S. Weir. 2007. Prediction of multi-locus inbreeding coefficients and relation to linkage disequilibrium in random mating populations. *Theor Popul Biol* **72:** 179-185.

Hinrichs, A.L., S. Bertelsen, L.J. Bierut, G. Dunn, C.H. Jin, J.S. Kauwe, and B.K. Suarez. 2005. Multipoint identity-by-descent computations for single-point polymorphism and microsatellite maps. *BMC Genet* **6 Suppl 1:** S34.

Huang, J., W. Wei, J. Zhang, G. Liu, G.R. Bignell, M.R. Stratton, P.A. Futreal, R. Wooster, K.W. Jones, and M.H. Shapero. 2004. Whole genome DNA copy number changes identified by high density oligonucleotide arrays. *Hum Genomics* **1:** 287-299.

Iafrate, A.J., L. Feuk, M.N. Rivera, M.L. Listewnik, P.K. Donahoe, Y. Qi, S.W. Scherer, and C. Lee. 2004. Detection of large-scale variation in the human genome. *Nat Genet* **36:** 949-951.

Kong, X., K. Murphy, T. Raj, C. He, P.S. White, and T.C. Matise. 2004. A combined linkage-physical map of the human genome. *Am J Hum Genet* **75:** 1143-1148.

Kumar, R.A., S. KaraMohamed, J. Sudi, D.F. Conrad, C. Brune, J.A. Badner, T.C. Gilliam, N.J. Nowak, E.H. Cook, Jr., W.B. Dobyns et al. 2008. Recurrent 16p11.2 microdeletions in autism. *Hum Mol Genet* **17:** 628-638.

Lander, E. and L. Kruglyak. 1995. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat Genet* **11:** 241-247.

Lien, S., J. Szyda, B. Schechinger, G. Rappold, and N. Arnheim. 2000. Evidence for heterogeneity in recombination in the human pseudoautosomal region: high resolution analysis by sperm typing and radiation-hybrid mapping. *Am J Hum Genet* **66:** 557-566.

Lin, S., D.J. Cutler, M.E. Zwick, and A. Chakravarti. 2002. Haplotype inference in random population samples. *Am J Hum Genet* **71:** 1129-1137.

Malécot, G. 1948. *Les mathématiques de l'hérédité*. Masson, Paris.

Mao, Y. and S. Xu. 2005. A Monte Carlo algorithm for computing the IBD matrices using incomplete marker information. *Heredity* **94:** 305-315.

Marchini, J., D. Cutler, N. Patterson, M. Stephens, E. Eskin, E. Halperin, S. Lin, Z.S. Qin, H.M. Munro, G.R. Abecasis et al. 2006. A comparison of phasing algorithms for trios and unrelated individuals. *Am J Hum Genet* **78:** 437-450.

Meuwissen, T.H. and M.E. Goddard. 2007. Multipoint identity-by-descent prediction using dense markers to map quantitative trait loci and estimate effective population size. *Genetics* **176:** 2551-2560.

Paynter, R.A., D.R. Skibola, C.F. Skibola, P.A. Buffler, J.L. Wiemels, and M.T. Smith. 2006. Accuracy of multiplexed Illumina platform-based single-nucleotide polymorphism genotyping compared between genomic and whole genome amplified DNA collected from multiple sources. *Cancer Epidemiol Biomarkers Prev* **15:** 2533-2536.

Purcell, S. PLINK 1.00 http://pngu.mgh.harvard.edu/purcell/plink/.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M.A. Ferreira, D. Bender, J. Maller, P. Sklar, P.I. de Bakker, M.J. Daly et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81:** 559-575.

Qian, D. and L. Beckmann. 2002. Minimum-recombinant haplotyping in pedigrees. *Am J Hum Genet* **70:** 1434-1445.

Redon, R., S. Ishikawa, K.R. Fitch, L. Feuk, G.H. Perry, T.D. Andrews, H. Fiegler, M.H. Shapero, A.R. Carson, W. Chen et al. 2006. Global variation in copy number in the human genome. *Nature* **444:** 444-454.

Sobel, E. and K. Lange. 1996. Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. *Am J Hum Genet* **58:** 1323-1337.

Stephens, M., N.J. Smith, and P. Donnelly. 2001. A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet* **68:** 978-989.

Thompson, E.A. 2007. The IBD process along four chromosomes. *Theor Popul Biol*.

Weiss, L.A., Y. Shen, J.M. Korn, D.E. Arking, D.T. Miller, R. Fossdal, E. Saemundsen, H. Stefansson, M.A. Ferreira, T. Green et al. 2008. Association between Microdeletion and Microduplication at 16p11.2 and Autism. *N Engl J Med*.

Wright, S. 1921. Systems of Mating. I. the Biometric Relations between Parent and Offspring. *Genetics* **6:** 111-123.