IBDfinder and *SNPsetter*: tools for pedigree-independent identification of autozygous regions in individuals with recessive inherited disease

Ian M. Carr, Eamonn Sheridan, Bruce E. Hayward, Alexander F. Markham, David T. Bonthron

Centre for Autozygosity Mapping and Section of Genetics, Leeds Institute of Molecular Medicine, University of Leeds, Leeds, U.K.

Corresponding author:

Ian M. Carr

Leeds Institute for Molecular Medicine

Level 9, Wellcome Trust Brenner Building

St. James's University Hospital

Leeds LS9 7TF, U.K.

Abstract

Autozygosity mapping of recessive genes can be performed on a small number of affected individuals from consanguineous pedigrees. With the advent of microarray SNP analysis, acquiring genotype data has become extremely simple and quick, in comparison to gene mapping with microsatellite markers. However, the subsequent data analysis required to identify autozygous regions can still be a significant obstacle. For rapid gene identification, it may be desirable to integrate information from heterogeneous groups of affected individuals, both familial and isolated, under various assumptions of ancestry and locus heterogeneity, that are not amenable to formal linkage analysis. Unfortunately, there are few computer programs aimed specifically at facilitating this type of data sifting. Here, we demonstrate two new programs that facilitate the identification of autozygous regions within a heterogeneous SNP dataset derived from familial and sporadic affected individuals.

Keywords

autozygosity, homozygosity mapping, inbreeding, autosomal recessive, SNP, software

Introduction

The method of mapping recessive disease genes by identifying regions of identity by descent (IBD), in affected individuals who are the products of consanguineous unions, has amply rewarded its advocacy by Lander and Botstein [Lander and Botstein, 1987]. Obtaining whole-genome genetic data on which to base this approach is now much easier and cheaper than formerly, and can be readily outsourced to commercial providers. Detection of the desired "autozygous" regions linked to recessive genes should therefore be straightforward, provided that certain requirements are met in terms of (i) genotyping quality, (ii) pedigree structure, and (iii) computational approach. In practice, however, formal genetic analysis of interesting clinical phenotypes is often impaired by a failure in one of these three areas. The following common problems, corresponding to each of these three areas, are discussed in more detail below: (i) genetic datasets (usually from single nucleotide polymorphism (SNP) microarrays) may be inconsistent or marker sets may differ between experimental subjects; (ii) pedigrees may be unavailable, unreliable, or incomplete; (iii) formal computational algorithms may be incapable of dealing with complex inbred pedigrees without the introduction of unwarranted assumptions.

In addition, investigators often desire to make use of information from singleton cases where possible, not least because the recognition of small ancestral autozygous regions concordant between apparently unrelated individuals can be the key to positional identification of the disease gene. However, this desire introduces the additional difficulty of dealing with the possibility of locus heterogeneity among the cohort, a problem not usually amenable to formal analysis, given the typically small numbers of subjects.

Incompatible or inconsistent genotype data

Whole-genome linkage searches have moved from the use of low-density panels of highly informative microsatellites to much higher-density microarray analysis of biallelic SNPs.

Microarrays themselves have also quickly evolved, ~10k arrays having been successively replaced by chips offering ~50k, ~250k and ~1M SNPs. Such advances can outstrip the pace of a clinical project, so that patients ascertained over a period of years may have been genotyped with a mixture of array types. Because different arrays may not share SNPs, integrating data from these different marker sets is difficult. Furthermore, even when two datasets derive from a common array type, if their raw data were analysed at different times, their resulting SNP annotations may differ, due to changes in reference data (Fig. 1). Similarly, microarray SNP annotation data as delivered to the end user may not be concordant with the most current information presented by the genome browsers at Ensembl [Flicek et al., 2008] or UCSC [Kent et al., 2002]. Consequently, the potential exists both for misalignment of autozygous regions between different patients, and for pursuit of genomic regions incorrectly identified by interrogating the genome browsers using map positions.

Pedigree and computational limitations

While the ideal resource for autozygosity mapping comprises individual multiplex pedigrees informative enough to allow the identification of a unique autozygous locus, more typically, even in highly inbred families, only singleton cases are available. If a small cohort of unrelated singleton patients has been assembled, it may be unclear whether a common ancestral origin for the mutation in different cases is to be expected, or indeed whether locus heterogeneity exists within the cohort. Formal statistical analysis of small cohorts of this kind cannot be performed without making assumptions about the likelihood of locus and allelic heterogeneity.

Even when a promising multiplex inbred pedigree is available, the number of consanguineous loops may make it computationally difficult to perform LOD score calculations. In such cases, a recommended practice is to simplify the pedigree structure for analysis [Terwilliger and Ott, 1994]. However, this introduces doubt as to the true significance of the resulting LOD score. Furthermore, there is an implicit assumption that all familial relationships are correctly represented in the available pedigree. Empirical evidence indicates that in communities practicing customary consanguineous marriage, this assumption is usually incorrect; their complex families, with many unknown or undisclosed inbreeding loops, display observed homozygosity levels much greater than expected [Woods et al., 2006]. The unknown inbreeding loops will have an unpredictable impact on the reliability of computed linkage probabilities.

Because of these problems, we have developed ***SNPsetter*** and ***IBDfinder***, a pair of programs that respectively allow standardization of the data annotation of Affymetrix SNP data files and the detection of putative IBD regions, independently of allele frequencies and pedigree information. The IBD regions can also be compared among patients, allowing the rapid identification of overlapping regions of IBD, even if the data derive from different SNP microarray types. In common with our earlier program, ***AutoSNPa*** [Carr et al., 2006], ***IBDfinder*** does not calculate formal linkage probabilities. Rather, the graphical user interface is designed for ease of use without the need for specialist expertise in linkage methods. It allows flexible handling and inspection of large data sets, and export of detailed genotype information for regions of interest.

Materials and Methods

Subjects

The ***IBDfinder*** program is available for download (see below) together with a separate sample data file comprising 13 sets of Affymetrix 50k *Hin*dIII SNPs, derived from individuals with recurrent (familial) biparental hydatidiform mole (FBHM, OMIM 231090). In this autosomal recessive disorder, a failure to establish epigenetic imprints in the maternal germline results in failure of embryonic development and proliferation of hyperplastic trophoblastic tissue [Judson et al., 2002]. It usually results from biallelic mutations of the *NLRP7* gene in 19q13.4 [Murdoch et al., 2006]. Because the reproductive failure may confer a considerable social stigma, gathering of accurate family information and access to relatives has proven difficult, resulting in a predominance of singleton cases in our patient cohort. This small cohort is thus typical of the resource available for a rare recessive disorder; the patients also include both several unique mutations and multiple examples of an ancestral mutation, R693P. It also includes non-consanguineous cases with compound heterozygous *NLRP7* mutations, and cases with no 19q13.4 linkage.

Methods

Software development and requirements

Programming was done using Microsoft Visual Studio 2005 using the Visual Basic language. The programs requires the .NET framework 2.0 to be installed, and at the time of writing, are tested on Windows XP SP2 and Vista SP1. They are freely available for download, with documentation, from http://dna.leeds.ac.uk/ibdfinder/ and http://dna.leeds.ac.uk/snpsetter/.

SNPsetter

The ***SNPsetter*** work-flow is illustrated in Supp. Fig. S1. Its database is built around a list of SNP ***rs*** names, each linked to a single chromosome and physical position. These data can come from an Affymetrix data file, an Affymetrix annotation file or a user-created tab-delimited file. If the Affymetrix data or annotation file contains information on allele frequencies (Asian (Chinese), European or African-American) or genetic map positions (Marshfield, deCODE or slm1), these fields are also linked to the SNP ***rs*** name.

Once the core database has been created, position and frequency data can be added to or updated from a different Affymetrix file; only the link between ***rs*** name and chromosome is immutable. It is also possible to add user-defined data (*e.g.* a new set of allele frequencies), which are either appended to the existing data or used to replace the data in a specific Affymetrix-defined column. (The latter option allows new data to be used even with downstream applications that only accept files in the standard Affymetrix data file formats.)

SNPsetter can generate reduced output files containing data subsets. Options include excluding certain data fields, or selecting subsets of SNPs, filtered by chromosome, distance from other SNPs and/or allele frequency. (If SNPs are filtered by allele frequency, only the allele frequency dataset employed can be exported to the output file. Similarly, if SNPs are filtered by proximity to SNPs already added to the database, only one set of SNP positional data can be exported.)

SNPsetter can import its data either from Affymetrix tab-delimited *.xls* files or from BRLLM-format text files, but all input files must be of the same type. The BRLMM-format files do not include SNP ***rs*** identifiers. Therefore, when importing BRLMM-format data, the ***SNPsetter*** reference database must be initialized using an Affymetrix annotation file (*.csv*) that contains all the ***rs*** and Affymetrix SNP_A name data. The input files can be processed either in parallel or sequentially. The latter requires much less memory, but parallel processing of the input files enables the program’s facility to exclude SNPs that are all of the same homozygous genotype or that are all *NoCalls*.

SNPsetter also offers a facility to combine two SNP data files. This may be desirable simply in order to join two 50k or 250k files into a single 100k or 500k file. More commonly, though, a user may wish to analyze a patient cohort that includes more than one type of SNP set, *e.g.* older 10k data files and newer 50k or 500k files. In this case, one file of each type can be combined to produce a single master file. This master file can then be used as the initial ***SNPsetter*** reference file, allowing the subsequent standardization of either of the two types of input file.

Finally, ***SNPsetter*** can also generate a list of ***rs*** names from an Affymetrix file, which can then be submitted to the NCBI SNP web page (http://www.ncbi.nlm.nih.gov/projects/SNP/dbSNP.cgi?list=rslist) in order to retrieve a current list of ***rs*** names with chromosome and physical position data. This can then be used to form the initial basis of the ***SNPsetter*** reference database. This function would be particular useful to researchers using custom SNP chips, or working on organisms of which the genome has not been completely sequenced or is still subject to new builds.

More detailed descriptions of the functionality of ***SNPsetter*** are given in the illustrated on-line tutorial, at http://dna.leeds.ac.uk/snpsetter/guide/.

IBDfinder

The approach taken by ***IBDfinder*** is to generate a graphical display of a numerical score that changes as the genotype data are scanned in linear map order. Sequential scoring is also used by hidden Markov chain approaches to identifying IBD regions [Leutenegger et al., 2003] which require accurate genetic maps and allele frequencies. With SNP data, various factors could affect which regions are identified as IBD by a Markov chain approach. By definition, IBD regions will always be terminated by a heterozygous marker. However, because of the low heterozygosity of SNPs, the *start* of a “significant” IBD region may only be signalled by some notable criterion, such as homozygosity for a rare allele. The region identified as IBD will then differ according to the direction in which the chromosome is scanned. Scanning in both directions will identify two different (overlapping or non-overlapping) intervals, which if combined define a region flanked by heterozygous markers and containing at least one rare homozygous genotype. Because of the high density of current SNP arrays, even a 1cM homozygous interval (containing 100 or more SNPs) is highly likely to contain at least one rare homozygous genotype. At such high marker densities, it becomes simpler just to regard an IBD region as one that can be directly observed to be devoid of heterozygous SNPs.

The large number of SNPs in 50k and denser arrays means that even small autozygous regions typically contain >50 SNPs. As the number of SNPs in a region increases, the distribution of its allele frequencies will tend towards the genome-wide distribution. Consequently, an algorithm that utilizes allele frequencies to determine likelihood of IBD from a long sequence of markers may gain little overall, compared to one that simply counts SNPs. (Curves obtained by using allele frequencies are more variable in shape, having sharp increases in score where rare homozygous genotypes occur.) Additional to this limited extra benefit from using SNP allele frequencies is the consideration that available allele frequencies may not be valid for the population under study, particularly if small and reproductively isolated.

The lack of supplementary information needed by the ***IBDfinder*** algorithm makes it very easy to implement. However, the SNP-counting function cannot generate statistical genome-wide measures of the probability of IBD. In particular, linkage disequilibrium (LD) between SNPs, which is far from negligible at the high densities of modern arrays, is ignored. In principle, data from HapMap and similar projects could allow the effects of LD to be accounted for. However, distribution of LD blocks can vary markedly among different ethnic groups subject to differing recent selection pressures. For example, large LD blocks around the lactase gene are widely divergent among different reproductively isolated African populations [Tishkoff et al., 2007]. A general solution to this problem (which also affects earlier SNP-counting approaches such as ExcludeAR [Woods et al., 2004]) is therefore not currently available.

In practice, the graphical presentation and collation of SNP-counting results by ***IBDfinder*** allows two important steps: (a) analysis of singleton cases independently of secondary information and (b) alignment and comparison of putative IBD regions from different individuals (even if non-concordant SNP arrays were used) in order to highlight IBD regions common to a number of individuals.

Results

IBDfinder

Principle

The ***IBDfinder*** interface is illustrated in Fig. 2a-c. Typical displays of SNP data for Chr.1 (p*ter*–q*ter*, left–right) are shown for three inbred individuals. Each homozygous SNP is scored by counting (a) how many homozygous SNPs lie between it and the nearest heterozygous SNP and (b) the average number of homozygous SNPs between it and the nearest centromeric and telomeric heterozygous SNPs. These two scores (for one individual) are plotted against map coordinate in the upper panel of the user interface, as red and blue graphs, respectively. This makes it easy to see both the size of an IBD region and the number of homozygous SNPs within it. Although a heterozygous SNP should in principle reset both the (a) and (b) scores to zero, an allowance is made for genotyping error rates (see below).

The lower ***IBDfinder*** panel shows the total number and extent of IBD regions on one chromosome, summed across all individuals under analysis. The *y* coordinate in this panel represents the *number of individuals* scored as "IBD" (for detailed description see User Guide, Section 4.1) for that region. (The yellow line shows the maximum possible score.) Since each SNP data file is scored independently, the lower panel can overlay data from different types of SNP file; *e.g.* 10k *Xba*I files can be combined with 50k *Hin*dIII files. Because an individual SNP may then appear in only a subset of data files, it is necessary to group SNPs in bins of 0.125 Mb/cM when calculating the cumulative IBD score. A side-effect of this is that if a bin contains any SNP scored as "not IBD", the whole bin is scored as "not IDB". This limits the precision of the plot at the edges of IDB regions, but in practice has a small overall affect. In Fig. 2, all three individuals under study can immediately be seen to have regions of IBD that overlap to define a 11-Mb region centred on 71 Mb, over which all three are IBD (though not necessarily for the same haplotype).

Miscalls and low-density regions

Since the genotyping error rate of SNP arrays is typically ~1 in 200, any large homozygous region may contain wrongly called SNPs, which will introduce artefactual division of the IBD region. ***IBDfinder*** will therefore incorporate tolerance for the presence of a heterozygous SNP. Subject to certain user-selected constraints (see the on-line walk-through tutorial), rather than resetting the IBD score to zero, a single heterozygous SNP instead resets it to an arbitrary value above the "IBD cut-off" score. (This is actually calculated as the larger of (IBD cutoff + 1) or (selected error rate / 3)). This creates a valley in the IBD graph, but keeps the entire region above the IBD cut-off. The resulting appearance is shown in Fig. 2c, in which a >61-Mb IBD region is divided by a single heterozygous SNP (presumably a miscall). Consequently, ***IBDfinder*** can correctly identify minimal regions of IBD overlap between different affected individuals, even in the presence of the miscalled heterozygous SNP. (See Figures 10a-b in the walk-through tutorial for more detail.)

An adjustment is also made for low SNP density. The IBD score is arbitrarily reduced by 1 for each SNP short of a threshold of 5 SNPs per Mb/cM. For example, if a 2-Mb region contains only 6 SNPs, its score is reduced by 4. For a fuller explanation of score adjustment, see the ***IBDfinder*** walkthrough tutorial. The effect of the weighting can be seen by toggling the View > Show adjusted value option.

SNPsetter

We have mentioned above the difficulties that can arise from inconsistent SNP metadata (Fig. 1). To deal with this, and to address the need to integrate two or more discordant SNP datasets, we developed ***SNPsetter***, a program that creates a relational database containing all SNP metadata and patient genotypes. This database is then used to generate standardized output data files, consistently annotated with current metadata.

The effect of ***SNPsetter*** on subsequent data analysis is shown in Fig. 3. In Fig. 3a, ***IBDfinder*** is using the marker position data as supplied by Affymetrix in April 2005. In Fig. 3b, ***SNPsetter*** has first been used to exchange the SNP metadata for those later supplied by Affymetrix in September 2006. Although the number, approximate positions and scores of the IBD regions are similar in both analyses, the 2006 metadata define a smaller chromosome (due both to minor adjustments in the positions of many SNPs and to larger changes involving the centromere and 9p telomere). Another difference not apparent from the images is that the original file (11,561 SNPs) and the 2006 reference file (10,204 SNPs) contain only 8,168 SNPs in common.

Figs. 1 and 3 both illustrate a simple pitfall in autozygosity mapping; if SNP position data are not questioned and verified, IBD regions in different data files can be misaligned and any common regions consequently incorrectly placed on the current human genome sequence freeze.

SNPsetter can also be used to perform data reduction on a file. Data can be selected for removal based on allele frequency, distance from other SNPs or chromosome. Such a procedure may be desirable in order to reduce the number of SNPs analyzed by a third party application, to restrict the data shared with other researchers or simply to generate smaller files for electronic transmission. For example, we discuss below the condition familial biparental hydatidiform mole (FBHM; OMIM 231090), which usually results from *NLRP7* mutations [Murdoch et al., 2006]. The sample data files Fam2-01-N and Fam2-02-N, however, originate from FBHM patients who do not have *NLRP7* mutations. To prevent disclosure of the non-Chr.19 data, while the search for a second FBHM gene is still ongoing, we used ***SNPsetter*** to create these files,.

Usage example: identification of IBD regions among FBHM patients

To illustrate the usefulness and the limitations of ***IBDfinder***, we performed a retrospective analysis of a cohort of thirteen affected individuals with FBHM; typically for a rare recessive disorder, our cohort includes six singleton cases (Supp. Table S1). It also displays locus heterogeneity, Family L being unlinked to *NLRP7* [Judson et al., 2002; Hayward et al., 2003], while the eleven other affected individuals all have biallelic *NLRP7* mutations (data not shown).

Two patients were from a non-consanguineous English family, and were therefore excluded from autozygosity analysis. Fig. 4 shows the composite IBD plots for the best chromosomal regions identified by ***IBDfinder*** among the remaining eleven patients (all of Pakistani origin). The arrows in Fig. 4a indicate narrow peaks that result from short runs of SNPs that have low minor allele frequencies, and hence are homozygous in the majority of individuals under study. As shown in Fig. 4b, increasing the IBD cut-off eliminates these artefacts, while preserving the broader region of IBD on Chr.11p. A peak composite IBD score of 7 (of a maximum 11) is observed both here and in the region of Chr.19 containing *NLRP7* (Fig. 4c-d, black vertical line). Of the four patients not contributing to the Chr.19 IBD score, two are from Family L, as expected (Supp. Table S1). However, the other two (K2, K3) are from Family K, in which another affected relative (K1) shows a 7.7 Mb IBD region around *NLRP7*; in fact, K1, K2 and K3 were susbsequently found all to be homozygous for a previously described *NLRP7* mutation, R693P [Murdoch et al., 2006] (not shown).

The failure to detect an IBD region in K2 and K3 is due to its small size. The heterozygous *NLRP7*-flanking SNPs in both these individuals (rs400322, rs306425) define an autozygous region of at most 1.36 Mb, containing only five homozygous SNPs from the 10k set. This is consistent with the probable remoteness of the ancestor from whom the IBD mutation derived (see Discussion); Table 1 shows the highly divergent degrees of inbreeding among the three affected individuals in this pedigree, and also illustrates the point referred to earlier, that pedigree data systematically underestimate coefficients of inbreeding in complex consanguineous families.

The strength of ***IBDfinder*** in analysis of this small cohort was thus that it allowed quick summation of data across several families, including singletons with disparate mutations; this resulted in prioritization of the correct locus and narrowing to a region of overlapping homozygosity. A weakness is revealed by its failure to detect the small autozygous segments in two individuals with low coefficients of inbreeding. This limitation can be addressed to some extent by the use of present-day arrays with very high SNP densities, and adjustment of the "IBD cut-off" value to one appropriate for the value of *F*. However, even outbred individuals frequently display one or more sizeable regions of autozygosity [Broman and Weber, 1999], and in inbred individuals, many IBD segments larger than that harbouring the pathogenic mutation may be expected (Table 1). The most effective use of ***IBDfinder*** therefore comes from its ability to very rapidly identify subsets of individuals sharing the same IBD region.

Discussion

Autozygosity mapping is a powerful approach to the identification of rare recessive diseases in consanguineous pedigrees. Microarray SNP genotyping has greatly reduced the cost and time needed to obtain whole-genome genotype data, making feasible the analysis of small families or heterogeneous groups of singleton affected individuals, for which the resources would not previously have been available or justifiable. Optimal use of such clinical material may not be achievable by formal linkage methods, and flexible interrogation of the genotype sets by a variety of approaches may be the key to successful gene identification. The size of modern SNP arrays introduces considerable computational demands. ***IBDfinder*** is readily able to handle modern SNP datasets such as Affymetrix SNP 6.0 (900k SNPs) from multiple (tens of) individuals on a standard desktop computer.

SNP versus microsatellite data

Linkage analysis methods developed for use with microsatellites may not be optimal for SNP data. The following differences are relevant: (a) SNPs are less mutable than microsatellites. Large differences in allele frequencies between ethnic groups may therefore be expected to be persistent, with implications for approaches that use allele frequencies in determining an IBD statistic. (b) SNP data sets contain 10^4 to 10^6 genotypes, compared to ~400 for a typical microsatellite panel and (c) A single SNP is less informative than a typical microsatellite.

Considering the last point, for a microsatellite with *i* alleles, the chance that an individual is homozygous is Σp_i^2. For a typical dinucleotide repeat with five alleles of frequencies 0.1, 0.2, 0.4, 0.2, 0.1, this probability is 0.26 (heterozygosity = 0.74). The probability of two unrelated individuals being homozygous for the same allele is $\Sigma p_i^4 = 0.029$, and for three or four such individuals, 0.0042, 0.00066. For a maximally informative SNP, on the other hand, these probabilities for one, two, three or four concordantly homozygous individuals are 0.5, 0.125, 0.03125, 0.0078, and with only slightly suboptimal allele frequencies (0.35, 0.65) the values are 0.545, 0.19, 0.077, 0.032. In searching for concordant homozygosity among four apparently unrelated individuals (*e.g.* as an indication of a founder-effect for a recessive mutation), a typical microsatellite therefore offers ~50-fold more significance than a single typical SNP.

The informativeness of SNP data can of course be increased by considering sequences of SNP alleles as haplotypes; however the phase information missing from diploid genotypes cannot be trivially deduced from single patient or even limited family information [Hodge et al., 1999]. General approaches exist for inferring haplotypes within populations [Clark, 1990]. However, such inferred data cannot necessarily be assumed to be applicable for populations other than those from which they were derived, in particular to genetically isolated inbred groups.

In contrast, since SNP haplotypes are unambiguously resolved within IBD regions, the association of an ancestral mutation with a particular haplotype can readily be observed in the context of autozygosity mapping. This can allow very rapid narrowing of the disease gene interval (see [Uppal et al., 2008] for a recent example). Furthermore, because SNPs are less mutable, such founder haplotypes should be observable between more distantly related individuals than are microsatellite associations.

Detection of IBD segments

The formal LOD score obtainable through autozygosity mapping in an inbred family depends strongly on the allele frequency of the linked marker, since this affects the relative likelihood that the marker is identical by descent rather than by state [Mueller and Bishop, 1993].

Similar considerations determine the ability to detect autozygous genomic segments (composed of runs of homozygous genotypes) in isolated individuals. Provided the frequencies of the homozygous alleles are known, a formal LOD score can be derived, describing the likelihood that the observations result from IBD, relative to the null hypothesis that all markers within the region are IBS and in linkage equilibrium [Broman and Weber, 1999]. Unfortunately, relevant SNP allele frequencies are often not known for small reproductively isolated groups of individuals. Until a representative sample of that group has been genotyped to derive empirical allele frequencies, it will not be known whether allele frequencies drawn from outbred populations (such as those provided with Affymetrix SNP sets) are valid for the study group.

Formal probability distributions for the IBD status of genomic regions can also be obtained using a hidden Markov model [Leutenegger et al., 2003] in which the probability of a marker changing state between IBD and not-IBD is a function of the state of the preceding marker, the probability of recombination between the two markers, the marker's genotype and its allele frequencies. This approach is well-suited to highly polymorphic microsatellites at intervals of a few centimorgans. However, in the context of high-density SNP data, some of the difficulties referred to above again become relevant. At high SNP densities, the chance of recombination between any two adjacent SNPs becomes very low; consequently, a change in state is only likely to be identified when a (presumptively) very rare homozygous genotype occurs. This places undue reliance on allele frequencies of uncertain accuracy. Also, the assumptions of linkage equilibrium and lack of crossover interference break down at high marker densities.

In practice, as very high SNP densities are reached, it is reasonable (and perhaps also safer) to assume that allele frequency effects can be averaged out across a long succession of adjacent SNPs. The assignment of regions as IBD is then reduced to counting of adjacent homozygous SNPs, an approach that ***IBDfinder*** adopts in common with previous, non-graphical tools [Woods et al., 2004]. The main criticism of this empirical method is that patterns of linkage disequilibrium are ignored. We believe that this is not a practical impediment to the use of IBDfinder for its intended purpose, because most IBD segments in inbred individuals are large compared to the distances across which LD effects are significant.

Sizes of autozygous regions

The expected genetic lengths of autozygous regions, after a number of recombination events in the descendants of a founder individual, can be modelled using the so-called "broken stick" distribution [Barton and David, 1956; Macarthur, 1957]. Its predictions are consistent with the observation that even apparently outbred individuals often have autozygous segments of substantial length [Broman and Weber, 1999; Clark, 1999]. Among inbred individuals with recessive disease, the number and size of homozygous regions tend to be greater than expected from disclosed pedigree structures [Woods et al., 2006], an unsurprising observation given the incomplete description of consanguinity within such complex families.

The aim of ***IBDfinder***, in either case, is to rapidly identify and compare most of these regions, across user-selected combinations of familial and non-familial cases. ***IBDfinder*** does not determine statistical significances of individual IBD segments; this would be of little interest for gene localization, especially since in an individual patient, the pathogenic mutation is unlikely to lie within the largest IBD segment. (For example, all nine subjects homozygous for *NLRP7* mutations in our cohort have IBD segments elsewhere in the genome, substantially larger than that surrounding the pathogenic mutation.)

The more ancient is a pathogenic mutation, the smaller its surrounding ancestral founder-derived IBD region is likely to be. However, on passing through an inbred family, such a mutation is liable to become "captured" within a larger autozygous region, IBD from a more recent ancestor. In highly inbred families, this will eventually result in it residing within a very complex mosaic IBD region. In affected individuals who are less highly inbred, in contrast, the IBD segment around the mutation may not be distinguishable against the genome-wide level of "background" IBD segments. A mixed picture can present in a pedigree with multiple consanguineous loops, since the mutation may be captured in a recent large IBD region in one family branch, but not elsewhere. This principle (which is illustrated in Supp. Fig. 2) applies to FBHM Family K. The pathogenic *NLRP7* mutation R693P in Family K appears quite ancient, having now been detected in four independent families, two Indian [Murdoch et al., 2006] and two Pakistani (this report). This may explain the small size of the IBD regions in individuals K3 and K2, while in the more highly inbred individual K1, it has been captured within a larger 7.7-Mb IBD region derived from a more recent common ancestor.

These aspects are worth careful consideration, since published examples of successful autozygosity mapping may give a misleading impression of the likelihood that disease genes are to be found within large IBD segments. Since small IBD segments are harder to distinguish, they may fail to lead to successful gene identification, causing an ascertainment bias in the published literature. Certainly, a retrospective analysis of cases with known mutations (as here) is likely to reveal surprisingly small IBD segments around the causative mutation in some affected individuals.

REFERENCES

Barton D, David F. 1956. Some notes on ordered random intervals. Journal of the Royal Statistical Society 18:79-94.

Broman KW, Weber JL. 1999. Long homozygous chromosomal segments in reference families from the centre d'Etude du polymorphisme humain. Am J Hum Genet 65:1493-1500.

Carr IM, Flintoff KJ, Taylor GR, Markham AF, Bonthron DT. 2006. Interactive visual analysis of SNP data for rapid autozygosity mapping in consanguineous families. Hum Mutat 27:1041-1046.

Clark AG. 1990. Inference of haplotypes from PCR-amplified samples of diploid populations. Mol Biol Evol 7:111-122.

Clark AG. 1999. The size distribution of homozygous segments in the human genome. Am J Hum Genet 65:1489-1492.

Flicek P, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, Down T, Dyer SC, Eyre T, Fitzgerald S, Fernandez-Banet J, Gräf S, Haider S, Hammond M, Holland R, Howe KL, Howe K, Johnson N, Jenkinson A, Kähäri A, Keefe D, Kokocinski F, Kulesha E, Lawson D, Longden I, Megy K, Meidl P, Overduin B, Parker A, Pritchard B, Prlic A, Rice S, Rios D, Schuster M, Sealy I, Slater G, Smedley D, Spudich G, Trevanion S, Vilella AJ, Vogel J, White S, Wood M, Birney E, Cox T, Curwen V, Durbin R, Fernandez-Suarez XM, Herrero J, Hubbard TJP, Kasprzyk A, Proctor G, Smith J, Ureta-Vidal A, Searle S. 2008. Ensembl 2008. Nucleic Acids Res 36:D707-14.

Hayward BE, De Vos M, Judson H, Hodge D, Huntriss J, Picton HM, Sheridan E, Bonthron DT. 2003. Lack of involvement of known DNA methyltransferases in familial hydatidiform mole implies the involvement of other factors in establishment of imprinting in the human female germline. BMC Genet 4:2.

Hodge SE, Boehnke M, Spence MA. 1999. Loss of information due to ambiguous haplotyping of SNPs. Nat Genet 21:360-361.

Judson H, Hayward BE, Sheridan E, Bonthron DT. 2002. A global disorder of imprinting in the human female germ line. Nature 416:539-542.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. Genome Res 12:996-1006.

Lander ES, Botstein D. 1987. Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. Science 236:1567-1570.

Leutenegger A, Prum B, Génin E, Verny C, Lemainque A, Clerget-Darpoux F, Thompson EA. 2003. Estimation of the inbreeding coefficient through use of genomic data. Am J Hum Genet 73:516-523.

Macarthur RH. 1957. On the relative abundance of bird species. Proc Natl Acad Sci U S A 43:293-295.

Mueller RF, Bishop DT. 1993. Autozygosity mapping, complex consanguinity, and autosomal recessive disorders. J Med Genet 30:798-799.

Murdoch S, Djuric U, Mazhar B, Seoud M, Khan R, Kuick R, Bagga R, Kircheisen R, Ao A, Ratti B, Hanash S, Rouleau GA, Slim R. 2006. Mutations in NALP7 cause recurrent hydatidiform moles and reproductive wastage in humans. Nat Genet 38:300-302.

Terwilliger J, Ott J. 1994. Handbook of Human Genetic Linkage. . John Hopkins University Press, 1994.

Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, Silverman JS, Powell K, Mortensen HM, Hirbo JB, Osman M, Ibrahim M, Omar SA, Lema G, Nyambo TB, Ghori J, Bumpstead S, Pritchard JK, Wray GA, Deloukas P. 2007. Convergent adaptation of human lactase persistence in Africa and Europe. Nat Genet 39:31-40.

Uppal S, Diggle CP, Carr IM, Fishwick CWG, Ahmed M, Ibrahim GH, Helliwell PS, Latos-Bieleńska A, Phillips SEV, Markham AF, Bennett CP, Bonthron DT. 2008. Mutations in 15-hydroxyprostaglandin dehydrogenase cause primary hypertrophic osteoarthropathy. Nat Genet 40:789-793.

Woods CG, Cox J, Springell K, Hampshire DJ, Mohamed MD, McKibbin M, Stern R, Raymond FL, Sandford R, Malik Sharif S, Karbani G, Ahmed M, Bond J, Clayton D, Inglehearn CF. 2006. Quantification of homozygosity in consanguineous individuals with autosomal recessive disease. Am J Hum Genet 78:889-896.

Woods CG, Valente EM, Bond J, Roberts E. 2004. A new method for autozygosity mapping using single nucleotide polymorphisms (SNPs) and EXCLUDEAR. J Med Genet 41:e101.

Figure Legends

Figure 1

The three coloured bars indicate the regions spanned by markers rs259589 and rs611448, which in the Affymetrix 10k *Xba*I SNP set flank a region of 20 SNPs on Chr.1. In the UCSC Golden Path as of March 2006 (blue) these SNPs are located at 236.54 and 242.50 Mb (5.96 Mb interval), whereas in the Affymetrix 10k SNP data files from April 2005 (red) and September 2006 (green), the corresponding intervals are 235.51–241.38 (5.87 Mb) and 234.80–240.76 (5.96 Mb), respectively. As a result of such changes, candidate genes may be wrongly placed relative to an autozygous interval, unless the SNP array positional metadata are correctly matched (or adjusted) to the genome build.

Figure 2

Typical IBDfinder plots of SNP data (Chr.1) from three inbred individuals. Tick-marks at the top edge of the plot are 10 Mb apart. The vertical scale is adjusted to accommodate the maximum IBD score (much lower in **a**). The lower panel compares data from all three subjects, and hence is identical in **a-c**. The green line highlights low SNP density (in this case the centromere and pericentromeric heterochromatin). **c.** Tolerance of SNP miscalls in identification of IBD regions, The large autozygous region is interrupted by a single heterozygous SNP. Since this is flanked by >200 homozygous SNPs (a user-selectable tolerance), ***IBDfinder*** adjusts the displayed graph to maintain the score above the minimum cut-off (red horizontal line) chosen for assignment of a region as IBD. Consequently, the entire (>61-Mb) region scores as IBD for the purposes of assembling a composite IBD score (lower panel). However, the “twin-peak” appearance of the IBD plot remains as a strong visual cue to the presence of the dubious genotype.

Figure 3

IBDfinder display of Chromosome 9 SNP data (10k XbaI) demonstrating IBD regions. In (**a**) positional metadata from SNP output files created in April 2005 were used, while in (**b**) these have been exchanged for those available in September 2006. This has resulted, for example, in a shift of almost 5 Mb in the position of the smaller IBD segment to the right-hand side of the figure.

Figure 4

Overlapping homozygous regions in the FBHM patients, on Chr. 11 (**a, b**) and Chr. 19 (**c, d**). The homozygous SNP score cut-off is set at 10 SNPs in **a** and **c** and 15 SNPs in **b** and **d**. Arrows in (**a**) indicate narrow peaks of IBD score that reflect short runs of uninformative SNPs (see text for detail). They are eliminated by correct adjustment of the SNP score cut-off (**b**). In **c** and **d**, the position of *NLRP7* is indicated by the vertical black line.

Figure 1

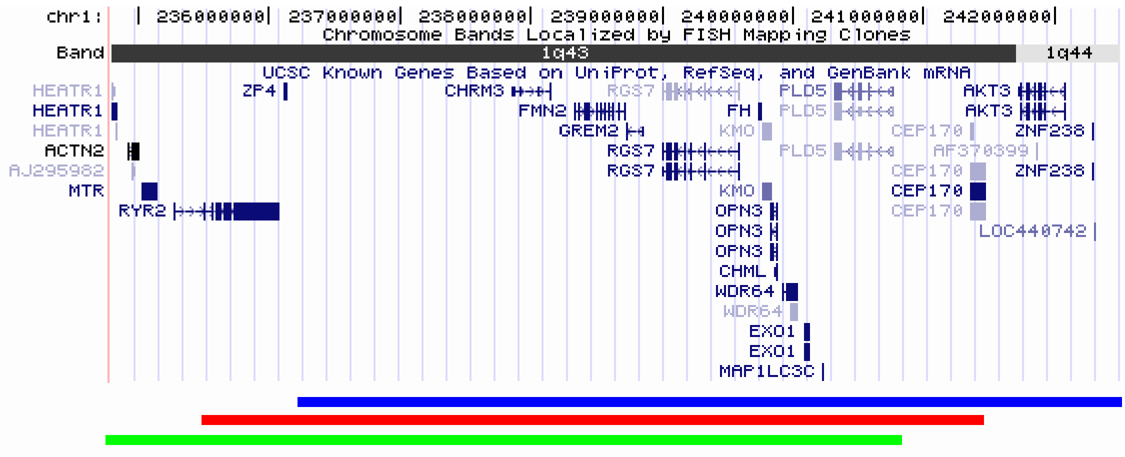
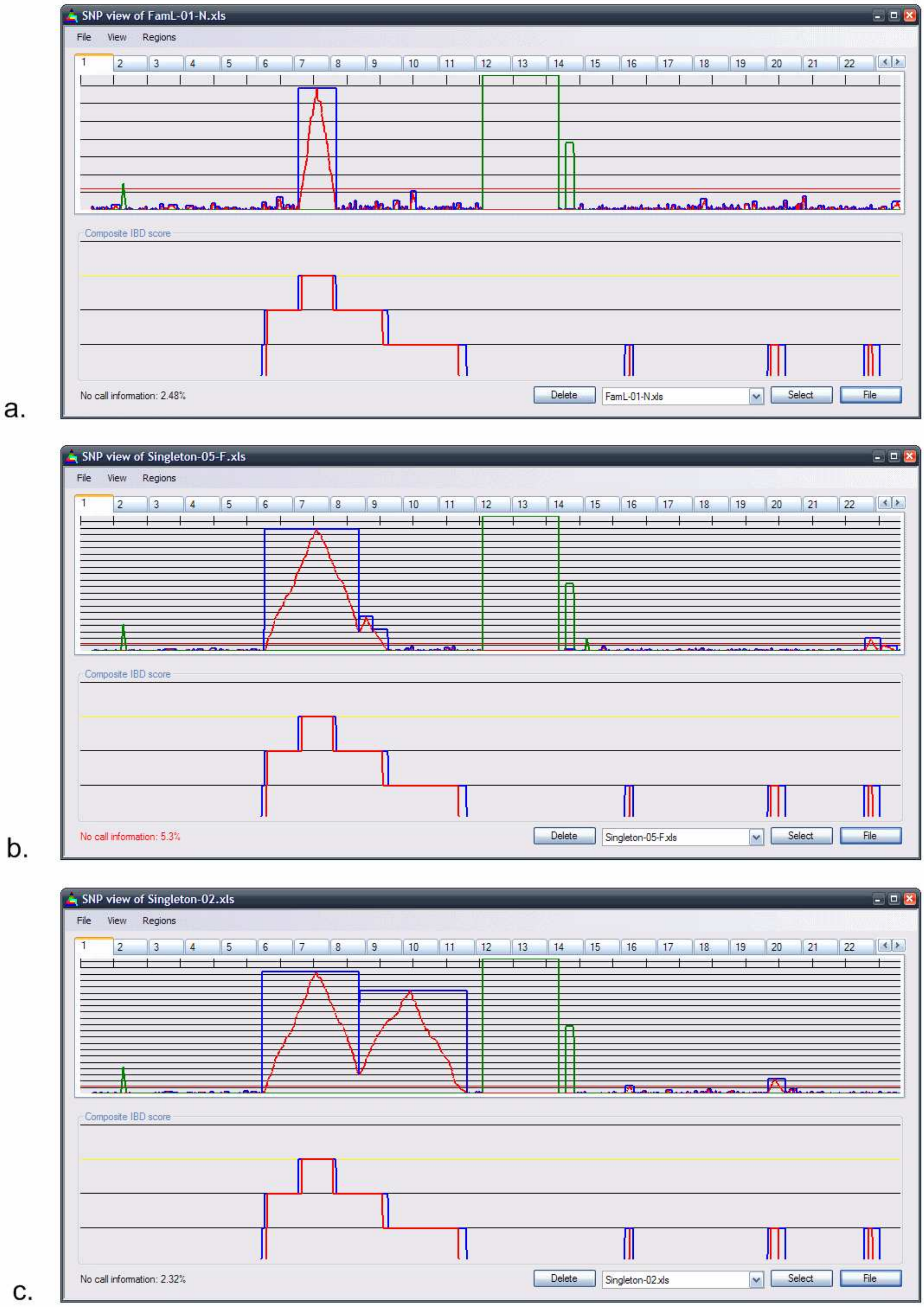

chr1:
236000000
237000000
238000000
239000000
240000000
241000000
242000000
Chromosome Bands Localized by FISH Mapping Clones
Band
1q43
1q44
UCSC Known Genes Based on UniProt, RefSeq, and GenBank mRNA
HEATR1
HEATR1
HEATR1
ACTN2
AJ295982
MTR
RYR2
ZP4
CHRM3
FMN2
GREM2
RGS7
RGS7
RGS7
FH
KMO
KMO
OPN3
OPN3
OPN3
CHML
WDR64
WDR64
EXO1
EXO1
MAP1LC3C
PLD5
PLD5
PLD5
CEP170
CEP170
CEP170
CEP170
AF370399
LOC440742
AKT3
AKT3
ZNF238
ZNF238

Figure 2



a.

SNP view of FamL-01-N.xls

No call information: 2.48%

FamL-01-N.xls

b.

SNP view of Singleton-05-F.xls

No call information: 5.3%

Singleton-05-F.xls

c.

SNP view of Singleton-02.xls

No call information: 2.32%

Singleton-02.xls

Figure 3

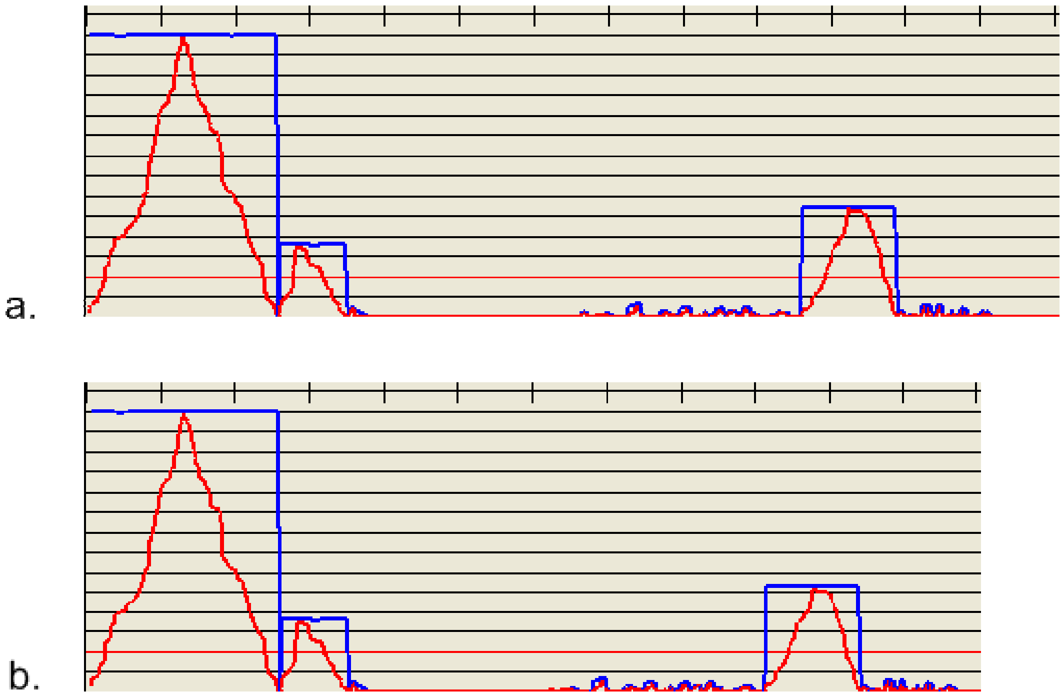


Figure 4

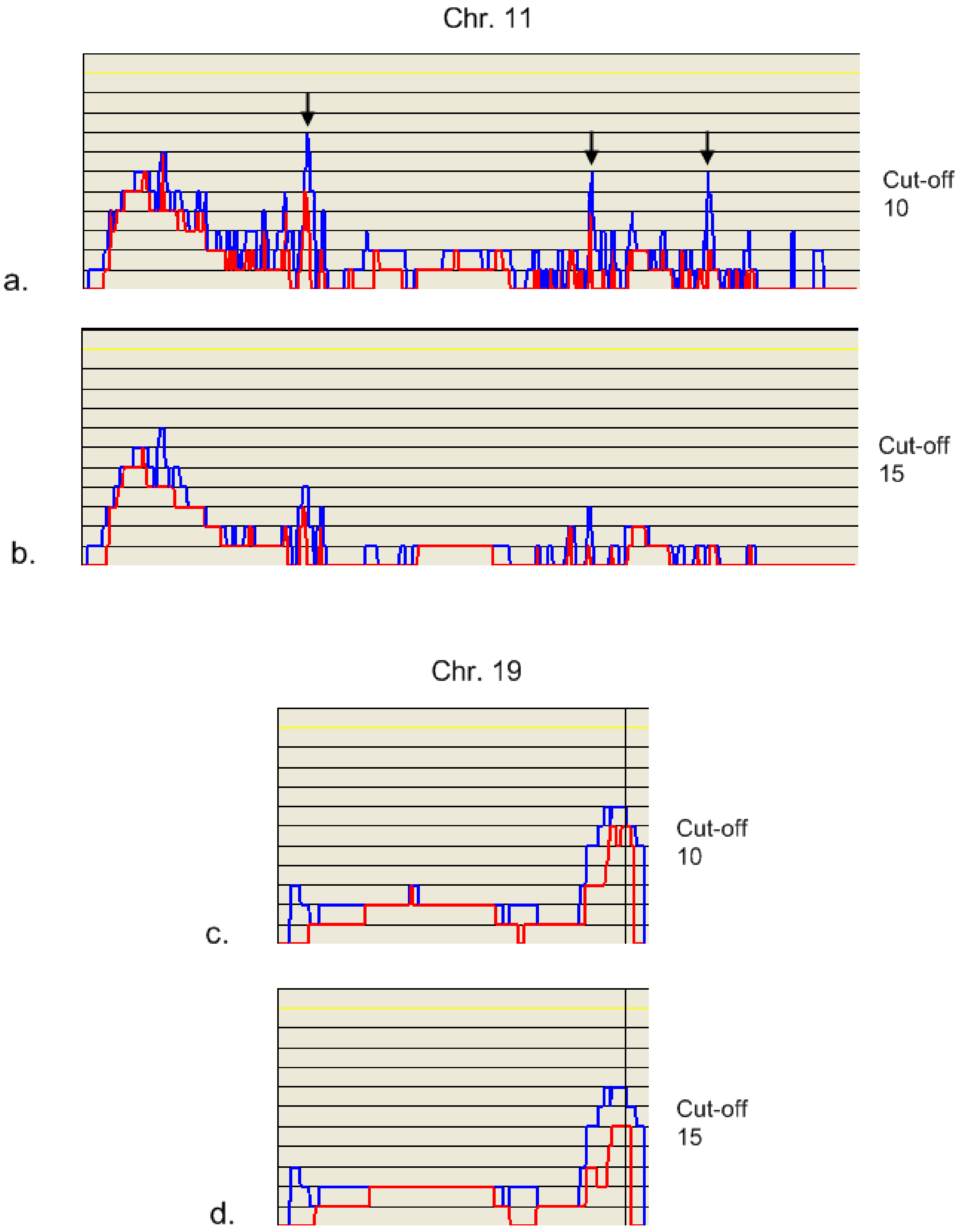


Table 1

Widely divergent degrees of autozygosity among three affected individuals from FBHM Family 1. Note also how pedigree data underestimate the coefficient of inbreeding by two- to three-fold; the observed *F* was calculated from the sum of regions >2 Mb identified by ***IBDfinder*** as IBD.

Affected individual	**Calculated *F***	**Number of IBD segments >2 Mb**	**Observed *F***	**Largest IBD segment / Mb**	***NLRP7* IBD segment / Mb**
K3	0	5	0.008	7.2	<1.4
K2	0.016	12	0.05	26	<1.4
K1	0.066	32	0.16	52	7.7

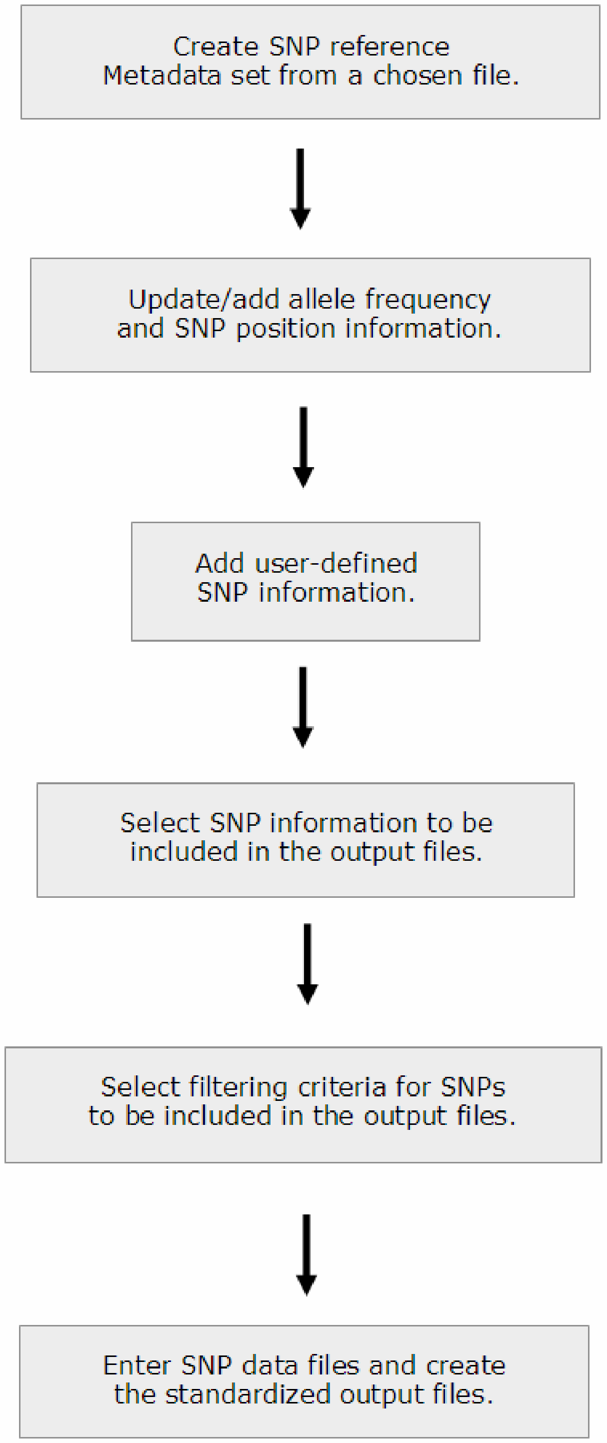
Supplementary Table S1

SNP data files supplied with IBDfinder and used in the example analysis in this article.

File name	*NLRP7* mutation status	Ethnicity, consanguinity
FamE-E1	Compound heterozygote	North European, non-consanguineous
FamE-E2	Compound heterozygote	North European, non-consanguineous
FamK-K3-F	R693P homozygous	Pakistani, consanguineous
FamK-K2-F	R693P homozygous	Pakistani, consanguineous
FamK-K1-F	R693P homozygous	Pakistani, consanguineous
FamL-01-N-Chr19	Not found	Pakistani, consanguineous
FamL-02-N-Chr19	Not found	Pakistani, consanguineous
Singleton-01	Homozygous (unique)	Pakistani, consanguineous
Singleton-02	Homozygous (unique)	Pakistani, consanguineous
Singleton-03	Homozygous (unique)	Pakistani, consanguineous
Singleton-04	Homozygous (unique)	Pakistani, consanguineous
Singleton-05-F	R693P homozygous	Pakistani, consanguineous
Singleton-06	Homozygous (unique)	Pakistani, consanguineous

Supp. Figure S1

The workflow process for standardization of Affymetrix and BRLMM data files using *SNPsetter*.


Create SNP reference Metadata set from a chosen file.
Update/add allele frequency and SNP position information.
Add user-defined SNP information.
Select SNP information to be included in the output files.
Select filtering criteria for SNPs to be included in the output files.
Enter SNP data files and create the standardized output files.

Supp. Figure S2

The size of IBD regions can differ greatly, depending on the position of the actual (as opposed to the nearest, and hence most likely) common ancestor. Individuals LI and DI could both have regions of IDB originating from individual LA (DNA flow shown for LI by red lines), but their actual common ancestor could alternatively lie outside the known pedigree (hypothetical DNA flow shown by blue lines). Both affected individuals will then share a small region of IBD (blue bar next to the ideograms) originating from the distant ancestor (DA); however, only LI will have a large IBD region (red bar) that arose from the local ancestor (LA) (and also subsumes the smaller region that LA inherited from DA). An example of this phenomenon is seen in FBHM Family 1.