Bioinformatics, 36(24), 2020, 5567–5570
doi: 10.1093/bioinformatics/btaa940
Advance Access Publication Date: 2 November 2020
Discovery Note

Genetics and population analysis

Obelisc: an identical-by-descent mapping tool based on SNP streak

Kyuto Sonehara[1] **and Yukinori Okada**[1,2,3,*]

[1]Department of Statistical Genetics, Osaka University Graduate School of Medicine, Suita 565-0871, Japan [2]Laboratory of Statistical Immunology, Immunology Frontier Research Center (WPI-IFReC), Osaka University, Suita 565-0871, Japan and [3]Integrated Frontier Research for Medical Science Division, Institute for Open and Transdisciplinary Research Initiatives, Osaka University, Suita 565-0871, Japan

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on February 1, 2020; revised on September 30, 2020; editorial decision on October 24, 2020; accepted on October 26, 2020

Abstract

Motivation: Genetic linkage analysis has made a huge contribution to the genetic mapping of Mendelian diseases. However, most previously available linkage analysis methods have limited applicability. Since parametric linkage analysis requires predefined model of inheritance with a fixed set of parameters, it is inapplicable without fully structured pedigree information. Furthermore, the analytical results are dependent on the specification of model parameters. While non-parametric linkage analysis can avoid these problems, the runs of homozygosity (ROH) mapping, a widely used non-parametric linkage analysis method, can only deal with recessive inheritance. The implementation of non-parametric linkage analyses capable of dealing with both dominant and recessive inheritance has been required.

Results: We have developed the Obelisc (**Obs**e̲rvational **li**nkage **sc**an), a flexibly applicable user-friendly non-parametric linkage analysis tool, which also provides an intuitive visualization of the analytical results. Obelisc is based on the SNP streak approach, which does not require any predefined inheritance model with parameters. In contrast to the ROH mapping, the SNP streak approach is applicable to both dominant and recessive traits. To illustrate the performance of Obelisc, we generated a pseudo-pedigree from the publicly available BioBank Japan Project genome-wide genotype dataset ($n > 180\,000$). By applying Obelisc to this pseudo-pedigree, we successfully identified the regions with inherited identical-by-descent haplotypes shared among the members of the pseudo-pedigree, which was validated by the population-based haplotype phasing approach.

Availability and implementation: Obelisc is feely available at https://github.com/qsonehara/Obelisc as a python package with example datasets.

Contact: yokada@sg.med.osaka-u.ac.jp

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Linkage analysis has played a central role in mapping susceptibility genes in research on rare diseases (Gusella *et al.*, 1983; Kamatani *et al.*, 2000). Although widespread next generation sequencing (NGS) technologies have enabled us to directly detect rare mutations involved in the etiology of rare diseases, it is still challenging to reliably detect some of those rare mutations, such as structural variations, including indels and repetitive elements (Kosugi *et al.*, 2019). In addition, NGS is still inferior to microarray technologies in terms of cost-efficiency and reliability in common variant calling (Kishikawa *et al.*, 2019). Linkage analysis utilizing genome-wide common SNP information is a powerful tool to locate disease-associated rare variants.

Linkage analysis is methodologically classified into two approaches: parametric and non-parametric (Ott et al., 2015). Parametric linkage analysis is based on a predefined model of inheritance with a fixed set of parameters, such as an allele frequency and penetrance of disease-associated variants. Based on this model, the logarithm of the odds (LOD) scores are calculated, which are statistical estimates of how likely the marker loci and the disease locus are located close together. While parametric linkage analysis has made a significant contribution to genetic mapping of Mendelian traits (Gusella *et al.*, 1983; Kamatani *et al.*, 2000), the following points need to be considered when performing such analysis: (i) fully structured pedigree information is needed, (ii) the LOD score calculation becomes complicated and computationally intensive in highly inbred pedigrees (Eerdewegh, 1989) and (iii) the analytical results are dependent on the specification of model parameters (Risch et al., 1992).

Non-parametric linkage analysis involves evaluating haplotype sharing within a case group in an observational way without predefined inheritance models containing parameters. A widely used non-parametric linkage analysis method is the runs of homozygosity (ROH) mapping (Lander *et al.*, 1987), which searches for autozygous identical-by-descent (IBD) segments inherited from a recent common ancestor by assessing stretches of consecutive homozygous SNPs. In contrast to parametric linkage analysis, ROH mapping (i) does not require any structured pedigree information, (ii) inbreeding within the pedigree is not computationally problematic owing to the simplicity of the method and (iii) it does not contain predefined parameters that affect the analytical results. While these flexibilities are an advantage of ROH mapping, it can only deal with recessive inheritance.

Here, we introduce the Obelisc (**Obs**ervational **li**nkage **sc**an), which is a non-parametric linkage analysis software, applicable to both dominant and recessive inheritance (Fig. 1a). It is based on the 'SNP streak' approach (Leibon *et al.*, 2008; Miyazawa *et al.*, 2007; Okada *et al.*, 2014; Thomas, 2010), which estimates haplotype sharing and detects candidate IBD segments shared within a case group. The SNP streak approach is a highly flexible IBD mapping method. However, there are no user-friendly tools to easily apply this method and obtain intuitively understandable mapping results. Obelisc is the first tool that enables the convenient performance of SNP streak-based IBD mapping and visualization of the results of the analysis. Obelisc also performs ROH mapping, and the user can obtain the results of both powerful non-parametric linkage analyses simultaneously.

2 Materials and methods

2.1 SNP streak-based IBD mapping

Obelisc takes PLINK (Chang *et al.*, 2015) binary PED files as input and utilizes genome-wide SNP data to detect IBD stretches. First, for each SNP site, the possibility of IBD among cases is evaluated based on the following fact: if all the cases share an IBD haplotype at the site, individuals who have homozygous reference allele and who have homozygous alternative allele are never compatible (see details in Results and Fig. 1b). After every SNP site is evaluated, Obelisc searches for regions where possible IBD sites appear continuously by a sliding window approach (Okada *et al.*, 2014). With the default settings, the scanning window is 1.5 Mbp and IBD is judged when all of the following conditions are fulfilled: (i) it contains 25 SNPs or more, (ii) it contains no nearest SNP pairs more than 1 Mbp apart and (iii) all the SNP sites in it are possible IBD sites as described above with exception of no more than 1 site. The user can change these parameters flexibly.

Obelisc scans genome-wide SNPs of the case group in the aforementioned manner to detect IBD regions shared among all of the cases. Then, separately for each control individual, Obelisc adds him or her to the case group and executes genome-wide SNP scanning again in the mixed group to distinguish whether the parts of the detected regions are also shared by the added control (i.e. the regions may not be specific to cases). The case-specificity of each part of the detected regions is calculated as $1 - n_{\text{share}}/n_{\text{all}}$, where n_{share} is the number of controls who do not share the IBD region and n_{all} is the number of all of the controls.

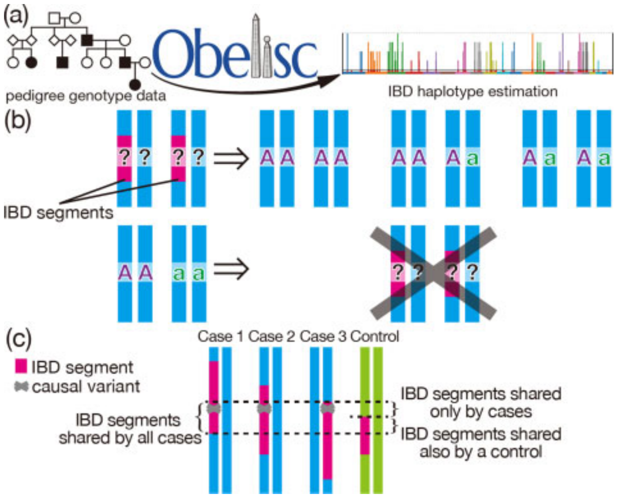

Fig. 1. Fundamental concept used in Obelisc. (**a**) We developed a new non-parametric linkage analysis tool, Obelisc. (**b**) If a pair of individuals share an IBD segment, each individual must have at least one identical allele at all of the loci within the segment. Hence if one individual has genotype AA and the other has genotype aa in the same locus, this locus cannot be within an IBD region. (**c**) Applying this concept to more than two cases narrows down candidate IBD segments shared among all of the cases. We also utilize control data to estimate the extent to which an IBD region is specifically shared among cases

2.2 Constructing a pseudo-pedigree from a publicly available BioBank Japan (BBJ) Project dataset

To demonstrate Obelisc IBD mapping with a practical example, we utilized a real genome-wide SNP genotype dataset of BBJ Project (Hirata *et al.*, 2017; Kanai *et al.*, 2018; Nagai *et al.*, 2017) ($n = 182\ 505$) accessed through the National Bioscience Database Center (Fig. 2a). We applied standard quality-control criteria to the genotype data and performed linkage disequilibrium-pruning (see Supplementary Notes).

We inferred kinships between each pair of the cohort participants and sexes of each participant only from the genotype data using PLINK (version 1.90b4.4) (Chang *et al.*, 2015). The kinships were estimated based on PI_HAT and Z1 provided by PLINK according to the following criteria: (i) if PI_HAT $\geq$ 0.38 and Z1 $\geq$ 0.8, the kinship is parent–child; (ii) if PI_HAT $\geq$ 0.38 and Z1 < 0.8, sibling; and (iii) if 0.38 > PI_HAT $\geq$ 0.17, other kinships closer than a third-degree relative. Since individuals in this publicly available biobank dataset had been de-identified and no pedigree information was provided, we constructed pseudo-pedigree charts in order to make the kinships easy to interpret (Fig. 2b; see Supplementary Notes).

We selected one of the largest pseudo-pedigrees, which consisted of nine genotyped individuals. We arbitrarily assigned artificial affection status to each individual so that the status appeared to be in accordance with autosomal dominant inheritance (Fig. 2c).

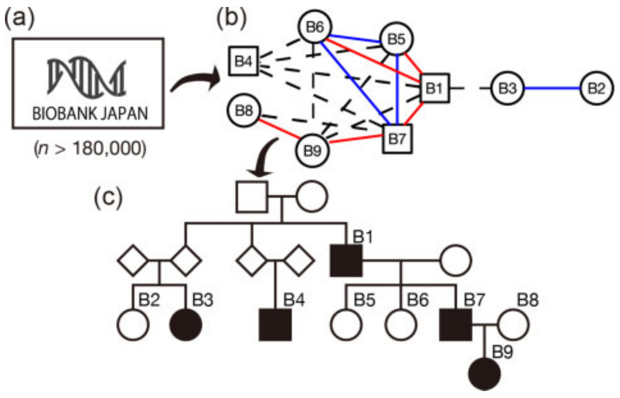

Fig. 2. Schematic overview of constructing pseudo-pedigree from public biobank data. (**a**) We utilized publicly available biobank genotype data. (**b**) A network diagram showing inferred relationships among some biobank participants. Red edges indicate parent–child relationships, blue edges indicate sibling relationships and black dashed edges indicate other relationships that are closer than third-degree relatives. (**c**) A constructed pseudo-pedigree chart based on the inferred kinships

2.3 Comparison between SNP streak-based IBD mapping results and explicitly phased haplotypes

Since the SNP streak approach finds latent identical haplotypes only using unphased genotypes, this approach essentially involves judging haplotype matches without explicit haplotype estimation. To empirically show the performance of this method, we assessed whether the IBD regions identified by Obelisc had shared haplotypes among the individuals.

We phased all of the quality-controlled BBJ genotype data with Eagle (version 2.4.1) (Loh *et al.*, 2016) and extracted inferred haplotypes of the nine members in the pseudo-pedigree. We compared each pair of inferred haplotypes. If the number of mismatched SNPs was no more than 1 per Mbp (corresponding to the default settings in the sliding window approach), we defined the pair as being IBD, as defined by Obelisc.

3 Results

3.1 Principle of the SNP streak approach

We developed an easy-to-use non-parametric linkage analysis tool, Obelisc, which also provides a graphical representation of the analytical results (Fig. 1a). The goal of the SNP streak approach is to detect IBD haplotypes shared within a case group from genome-wide SNP genotype data. The SNP streak approach is fundamentally based on the fact that, if a pair of individuals share an IBD segment harboring the disease-causing variant, they must carry at least one identical allele at all of the loci within the segment. In other words, thinking of a pair of individuals and a biallelic locus A/a, if one has a homozygous genotype AA and the other has the other homozygous genotype aa, this locus cannot be within the region IBD in this pair (Fig. 1b). It is worth noting that this concept not only permits the exact match of the genotypes within the pair, but also allows the combination of heterozygote and only one of the two homozygotes. The condition described here is the necessary but not sufficient condition for concluding that a segment is IBD, and a sufficient length of contiguous SNP stretches in this condition is required to make the inference reliable.

By extending this concept to multiple affected individuals, we can narrow down the number and the size of IBD candidate segments in all of the cases. Moreover, by adding each one of the unaffected to the case group and estimating IBD regions, we can further filter out candidate IBD regions also shared by a control (Fig. 1c). We defined IBD case-specificity per detected region as the proportion of the number of the controls not sharing the region relative to all of the controls. The IBD case-specificity indicates how rarely the detected IBD region is also shared among the controls. The implementation of this indicator is one of the distinctive points of our IBD mapping method compared with other existing SNP streak-based approaches (Leibon *et al.*, 2008; Miyazawa *et al.*, 2007; Okada *et al.*, 2014; Thomas, 2010). This indicator allows users to estimate the penetrance of the causal mutation within the detected region.

3.2 IBD mapping demonstration with a pseudo-pedigree derived from public biobank data

To illustrate SNP streak-based IBD mapping implemented in Obelisc, we generated pseudo-pedigree data from a publicly available genotype dataset of the BBJ Project (Fig. 2; 492 861 autosomal variants and 11 342 X-chromosomal variants of 182 108 individuals, mainly of Japanese ancestry) (Hirata *et al.*, 2017; Kanai *et al.*, 2018; Nagai *et al.*, 2017).

First, we estimated the pairwise kinships of all of the pairs of individuals in the dataset using PLINK (Chang *et al.*, 2015). Of all of the 16 581 570 778 pairs of individuals, 12 226 pairs (7.37×10^{-5}%) were estimated to be closer than third-degree relatives (Supplementary Fig. S1a; see Section 2). Of these 12 226 relative pairs, 6911 pairs (56.5%) were estimated to be first-degree relatives (parent–child or sibling kinship; Supplementary Fig. S1b). These observations demonstrate that a non-negligible fraction of the biobank participants is not independent but closely related (Yengo et al., 2019). To construct pseudo-pedigrees based on the inferred kinships, we drew kinship network diagrams, whose nodes indicate individuals and whose edges indicate the kinships of the pairs of individuals. While the first-degree relatives can be clearly discriminated between parent–child and sibling kinship by genotypes (see Section 2), discrimination of the second-degree relatives was unclear. Therefore, to construct pseudo-pedigrees consistent with the estimated kinships, we first drew kinship network diagrams only composed of first-degree relative edges. Then, we appended second-degree relative edges to them.

The largest diagram composed of first-degree relatives comprised ten individuals (Supplementary Figs S2 and S3a). We appended the second-degree relative edges to this largest diagram and constructed a pseudo-pedigree in accordance with the estimated kinships (Supplementary Fig. S3b). The constructed pseudo-pedigree appeared to be composed of two separate pedigrees connected by one marriage with one child. Since unrelated individuals' genotype data do not have much information for linkage analysis, we adopted only one of the two connected pedigrees for SNP streak-based IBD mapping. The adopted pseudo-pedigree was composed of nine genotyped individuals (B1–B9). Applying our IBD mapping method to all the pairs in the pseudo-pedigree confirmed the validity of the method (Supplementary Fig. S4).

To demonstrate our IBD mapping method for identifying disease associated regions, we arbitrarily assigned affection status as if it is autosomal dominant: five (B1, B3, B4, B7 and B9) are affected, and four (B2, B5, B6 and B8) are unaffected (Fig. 2c). We applied IBD mapping implemented in Obelisc to this pseudo-pedigree based on the real genotype data. The IBD mapping results are shown in Figure 3a. A total of 65 candidate IBD regions were identified as harboring the causal mutation responsible for the artificially assigned affection status (mean length=3.38 Mbp). Furthermore, five of the identified regions were estimated to be exclusively shared in the case group (chromosome 1p34-33, 3q33, 12q24, 15q24 and 17q22; mean length=1.55 Mbp; Supplementary Table). We confirmed that decreasing the number of cases resulted in larger IBD regions, and that decreasing the number of samples did in larger case-specific IBD regions, as expected (Supplementary Fig. S5).

While the affection status of each member was arbitrarily assigned, it was, in principle, assumed that the case-assigned members truly shared an identical haplotype within a detected region even if this region would not harbor causal variants. To examine

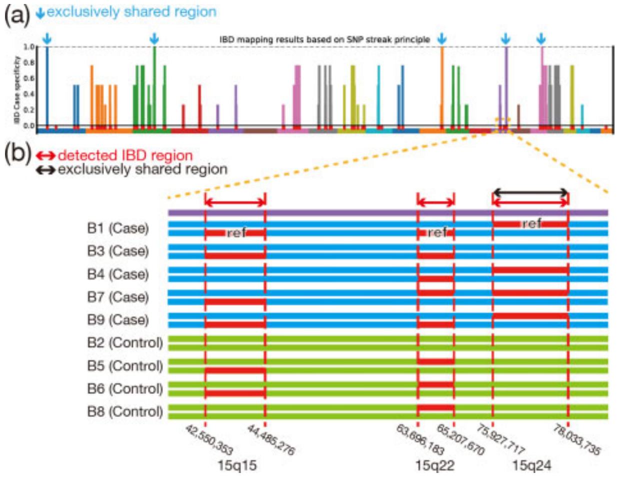

Fig. 3. IBD mapping results and comparison to haplotypes estimated by population-based haplotype phasing approach. (**a**) IBD mapping results by SNP streak approach. The horizontal axis indicates the genomic position, and the vertical axis indicates IBD case-specificity. (**b**) We compared all three IBD regions within chromosome 15 detected by the SNP streak approach with inferred haplotypes. Haplotypes of each individual in the pseudo-pedigree were compared with those of B1. Red segments indicate that phased haplotypes of the regions are identical in alleles

whether each case's haplotype of a detected region is identical to another case's, we compared inferred haplotypes of the pseudo-pedigree members within the regions. Although the true haplotype phases of the detected regions were unknown, thanks to the large sample size of the BBJ cohort, we were able to accurately estimate the haplotype phases of individuals in the cohort with reasonable certainty by Eagle2 (Loh *et al.*, 2016), a population-based phasing tool.

As an example, we selected an exclusively shared IBD region located within chromosome 15q24 and two nearby IBD regions (within 15q15 and 15q22). We inferred the haplotype phases of these regions of the pseudo-pedigree members and compared phased haplotypes of each region (Fig. 3b). We applied a strict criterion for judging a pair of haplotypes as identical, namely, requiring exact matching throughout the region except for 1 locus per Mbp. Based on this rigorous criterion, 4, 5 and 4 of the 5 cases were estimated to have IBD haplotypes in the detected region within chromosome 15q15, 15q22 and 15q24. In addition, we compared each of the inferred haplotypes of both cases and controls to verify the estimated case-specificity of the regions. The estimated case-specificity of the region within chromosome 15q15, 15q22 and 15q24 was 0.5, 0.0 and 1.0, respectively. In the explicit phasing analysis, 2, 3 and 0 of the total of 4 controls shared the IBD haplotype with the cases (corresponding to real case-specificity 0.5, 0.25 and 1.0), showing high concordance with the estimation. Despite the strict criterion applied for defining IBD, our SNP streak-based IBD mapping successfully identified IBD regions shared among the cases and estimated the case-specificity of each region.

4 Discussion

We developed the Obelisc, which is a flexibly applicable, user-friendly non-parametric IBD mapping software. Obelisc implements both the above-mentioned SNP streak-based IBD mapping and ROH mapping simultaneously, and can visualize genome-wide mapping results. This tool enables users planning to investigate the genetics of rare diseases to easily conduct two effective non-parametric linkage analyses simultaneously.

We introduced IBD case specificity to indicate whether the detected regions are merely common among the background population. Obtaining the null distribution of IBD case specificity and making a statistical test feasible is left for a future improvement. As attempted in ROH mapping (Narasimhan *et al.*, 2016), the use of hidden Markov model may improve the performance of IBD mapping, and it is also a possible future work.

By utilizing public biobank genotype data, we successfully constructed a pseudo-pedigree based on the genetic kinships and sexes, which could be used as an illustrative example for SNP streak-based IBD mapping. We drew the pseudo-pedigree chart (Fig. 2c) to obtain an intuitive understanding of the kinships between the related individuals and the assignment of the affection status. However, we note that our IBD mapping only needed the affection status of each individual and did not utilize the constructed pseudo-pedigree structure. This is one of the most useful characteristics of non-parametric linkage analysis. With the large-scale sample size, biobank genotype datasets are likely to contain multiple sets of unexpected related samples, as shown in this study. Non-parametric linkage analysis does not require pedigree structure information, and the kinships can be inferred using the genotype data. Therefore, if only phenotype information is available, non-parametric linkage analysis can also be applied to the hidden pedigrees in biobank-scale datasets. Our study highlights this novel possibility of a biobank dataset investigation.

In conclusion, Obelisc facilitates disease-associated gene mapping in research on rare diseases, especially when recessive inheritance is not assumed or unstructured pedigree data is provided. Our tool complements current pitfalls in linkage analysis where existing tools are unsuitable and contributes to elucidating the etiology of unexplained rare diseases.

Acknowledgements

The authors thank the members of the BioBank Japan Project for supporting the study.

Funding

This study was supported by JSPS KAKENHI [19H01021, 20K21834], AMED [JP20km0405211, JP20ek0109413, JP20ek0410075, JP20gm4010006 and JP20km0405217] and Bioinformatics Initiative of Osaka University Graduate School of Medicine, Osaka University. Y.O. and K.S. were supported by the Takeda Science Foundation.

Conflict of Interest: none declared.

References

Chang,C.C. *et al.* (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, **4**, 7.

Eerdewegh,P.V. (1989) Linkage analysis with inbreeding. *Genet. Epidemiol.*, **6**, 277–279.

Gusella,J.F. *et al.* (1983) A polymorphic DNA marker genetically linked to Huntington's disease. *Nature*, **306**, 234–238.

Hirata,M. *et al.* (2017) Cross-sectional analysis of BioBank Japan clinical data: a large cohort of 200,000 patients with 47 common diseases. *J. Epidemiol.*, **27**, S9–S21.

Kamatani,N. *et al.* (2000) Localization of a gene for familial juvenile hyperuricemic nephropathy causing underexcretion-type gout to 16p12 by genome-wide linkage analysis of a large family. *Arthritis Rheum.*, **43**, 925–929.

Kanai,M. *et al.* (2018) Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. *Nat. Genet.*, **50**, 390–400.

Kishikawa,T. *et al.* (2019) Empirical evaluation of variant calling accuracy using ultra-deep whole-genome sequencing data. *Sci. Rep.*, **9**, 1–10.

Kosugi,S. *et al.* (2019) Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biol.*, **20**, 117.

Lander,E.S. *et al.* (1987) Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. *Science*, **236**, 1567–1570.

Leibon,G. *et al.* (2008) A SNP streak model for the identification of genetic regions identical-by-descent. *Stat. Appl. Genet. Mol. Biol.*, **7**, Article 16.

Loh,P.-R. *et al.* (2016) Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet.*, **48**, 1443–1448.

Miyazawa,H. *et al.* (2007) Homozygosity haplotype allows a genomewide search for the autosomal segments shared among patients. *Am. J. Hum. Genet.*, **80**, 1090–1102.

Nagai,A. *et al.* (2017) Overview of the BioBank Japan Project: study design and profile. *J. Epidemiol.*, **27**, S2–S8.

Narasimhan,V. *et al.* (2016) BCFtools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data. *Bioinformatics*, **32**, 1749–1751.

Okada,Y. *et al.* (2014) Integration of sequence data from a consanguineous family with genetic data from an outbred population identifies PLB1 as a candidate rheumatoid arthritis risk gene. *PLoS One*, **9**, e87645.

Ott,J. *et al.* (2015) Genetic linkage analysis in the age of whole-genome sequencing. *Nat. Rev. Genet.*, **16**, 275–284.

Risch,N. *et al.* (1992) Model misspecification and multipoint linkage analysis. *Hum. Hered.*, **42**, 77–92.

Thomas,A. (2010) Assessment of SNP streak statistics using gene drop simulation with linkage disequilibrium. *Genet. Epidemiol.*, **34**, 119–124.

Yengo,L. *et al.* (2019) Extreme inbreeding in a European ancestry sample from the contemporary UK population. *Nat. Commun.*, **10**, 1–11.