Linkage analysis and the study of Mendelian disease in the era of whole exome and genome sequencing

M. Dawn Teare and Mauro F. Santibañez Koref

Advance Access publication date 14 July 2014

Abstract
Whole exome and whole genome sequencing are now routinely used in the study of inherited disease, and some of their major successes have been the identification of genes involved in disease predisposition in pedigrees where disease seems to follow Mendelian inheritance patterns. These successes include scenarios where only a single individual was sequenced and raise the question whether linkage analysis has become superfluous. Linkage analysis requires genome-wide genotyping on family-based data, and traditionally the linkage analysis was performed before the targeting sequencing stage. However, methods are emerging that seek to exploit the capability of linkage analysis to integrate data both across individuals and across pedigrees. This ability has been exploited to select samples used for sequencing studies and to identify among the variants uncovered by sequencing those mapping to regions likely to contain the gene of interest and, more generally, to improve variant detection. So, although the formal isolated linkage analysis stage is less commonly seen, when uncovering the genetic basis of Mendelian disease, methods relying heavily on genetic linkage analysis principles are being integrated directly into the whole mapping process ranging from sample selection to variant calling and filtering.

Keywords: Linkage analysis; next-generation sequencing; whole exome sequencing; whole genome sequencing; Mendelian disease

INTRODUCTION
The identification of genes that contribute to disease risk and phenotypic variation has been a key research area for many decades. Mendel was the first experimental geneticist, crossing particular types of peas and observing the phenotypic patterns. He hypothesized that these patterns could be due to the transmission of discrete physical units, subsequently called genes, each randomly transmitted from the pair of parental genetic material. Today when diseases or traits are described as ‘simple Mendelian’, this generally means the pattern of phenotypic inheritance is compatible with the transmission of alleles at a single locus, and the mode of inheritance can be clearly classified as dominant, co-dominant or recessive. Diseases or traits, which are described as ‘complex’, are believed to have multiple genes involved, and many of these genes may individually have a modest effect on risk. Many diseases, however, have been found to have both Mendelian and complex components. This review will focus only on the role of linkage analysis in mapping genes responsible for diseases with a major Mendelian component. The identification of the underlying genetic defects in Mendelian disease can be assisted by concentrating on the regions of the genome that are inherited together with disease in individuals in a pedigree, and traditionally linkage analysis was the tool used to delineate such regions [1, 2]. More recently, as a consequence of new technical developments, direct identification of the causative lesions by sequencing the whole genome of affected individuals, and apparently omitting the linkage step, is gaining popularity. This raises the question of the role of linkage analysis given the new technical possibilities. In this short review,

Corresponding author. M. Dawn Teare, School of Health and Related Research (ScHARR), University of Sheffield, Regent Court, 30 Regent Street, Sheffield S1 4DA, UK. Tel.: +44-114-222-6398; Fax: +44-114-272-4095; E-mail: m.d.teare@sheffield.ac.uk

M. Dawn Teare is a Senior Lecturer at the University of Sheffield.

Mauro F. Santibañez Koref is a Senior Lecturer at the University of Newcastle.

© The Author 2014. Published by Oxford University Press. All rights reserved. For permissions, please email: journals.permissions@oup.com
we will explore this question. First, we will recapitulate some of the features of linkage analysis, then briefly discuss next-generation sequencing (NGS) and its application in the investigation of Mendelian disease and finally explore the emerging field of the application of linkage analysis in the elucidation of Mendelian disease using NGS. This is a rapidly evolving field, and we will concentrate on emerging trends rather than details.

**LINKAGE ANALYSIS**

Originally, genetic linkage analysis was used to examine the segregation of traits in pedigrees and to assess whether the particular phenotypes tended to be co-inherited, suggesting the proximity of the genes responsible for these traits to each other in the genome. It was revolutionized when it became possible to directly assess the state of so-called polymorphic markers, i.e. sites in the genome known to be polymorphic in the human population. Today the term genetic linkage analysis is often applied to the use of methods aiming to identify the chromosomal region that is most likely to contain the gene responsible for the trait of interest. It relies on material from families that have been selected to be informative for this process, i.e. display sufficient variation in the trait of interest and where, either in individual families or in sets of families, the phenotype segregates through a number of meioses that are large enough to achieve statistically significant results. Polymorphic markers are used to track the co-segregation of genetic material with the occurrence of the phenotype in the selected families [2]. Linkage analysis has been very successful in mapping many Mendelian traits. It is at its most powerful when the phenotype is due to a single gene. Its power is unaffected by allelic heterogeneity, i.e. when there are many distinct disease alleles within the same gene in the population, but is reduced when there is incomplete penetrance, locus heterogeneity, i.e. when defects in many genes can lead to the same phenotype, and when there are non-genetic forms of the disease. To confidently identify a region of the genome, which is likely to harbour the causative gene a minimum of 10 informative meioses, is needed. An informative meiosis occurs when a parent is heterozygous for both the marker and disease predisposing allele so that it is possible to deduce which allele has been transmitted to any affected or unaffected offspring. Hence, one of the key features affecting the power of linkage analysis is the probability that specific formations of parent-offspring trios will be informative for linkage when using a specific genetic marker. The usefulness of a genetic marker in linkage analysis can be quantified by its polymorphism information content, which is a function of the marker’s allele frequencies.

Linkage analysis can also be approached from a non-parametric or allele sharing perspective, which does not require a mode of inheritance (e.g. dominant or recessive) to be known or specified in the analysis. Tracking the segregation of marker alleles and disease in a pedigree enables the comparison of the degree of allele sharing among relatives with a specific phenotype with that expected from the familial relationships alone. For example, the degree of allele sharing at a specific locus in a series of pairs of affected siblings can be compared with the degree of sharing expected for pairs of siblings. An excess of allele sharing can be interpreted as evidence for the involvement of the locus in disease predisposition. Allele sharing evidence is combined over many such sibships and genomic regions that show excessively high levels of sharing are therefore likely to contain predisposing loci.

Linkage analysis needs very few genetic markers to deduce the chromosomal region shared between affected individuals, but it cannot ‘find the gene’ alone. The more informative meioses that can be included in the linkage analysis, the narrower the shared peak region is expected to be. This peak linkage region would then be taken forward for more expensive sequencing. This led to workflows where sample collection was followed by genotyping, linkage analysis and targeted sequencing. Before the availability of a sufficiently reliable human reference sequence, targeted sequencing required identifying bacterial or yeast strains or clones carrying the sequences from the region of interest in appropriate vectors. These type of workflows were therefore often referred to as “positional cloning” strategies, and linkage analysis was a central part of them.

**WHOLE EXOME AND WHOLE GENOME SEQUENCING**

The last 8 years have seen the development and commercialization of a variety of methods allowing the sequencing of many thousands or millions of templates simultaneously (see [3] for a comparison of some currently popular platforms). Collectively,
these methods are often referred to as massively parallel sequencing (MPS) or NGS technologies [4]. Complementary developments in this period have led to methods allowing the selective enrichment of certain portions of the genome [5–8]. Initially, these techniques were used to sequence selected candidate regions in large numbers of samples [9], and NGS was seen mainly as a replacement of Sanger sequencing in the targeted sequencing step [10]. However, as the relative cost of sequencing fell, these regions became larger to include eventually all indentified regions that encode all the sequences represented in mature transcripts. These sequences are often collectively designated as the exome and the sequencing of these regions as whole exome sequencing (WES) [11]. This allows to concentrate in a relatively small fraction of the genome. Sequencing the exome was of particular interest because of the assumption that highly penetrant changes (that give rise to a Mendelian phenotype) are most likely to involve changes or abolition of gene function. This was consistent with the finding that only a small proportion of the variants implicated in Mendelian disease had been found in regulatory regions (<5% according to [12]). However, the enrichment procedures can lead to biased representation of the sequences of interest. Such biases can result, for example, in the under representation of 5’ exons, which can have a high CG content [8]. This concern and the falling sequencing cost have led to an increasing interest in sequencing the whole genome [whole genome sequencing (WGS)] when studying Mendelian disease (e.g. [13]).

In general, the analysis of NGS or WES data begins with converting the machine output, which may, for example, consist of signal intensity data, into sequences. This process is designated as base calling and is specific to the sequencing equipment used. The sequences are then mapped to a reference sequence and deviations from the reference sequence identified. This last step, often designated variant calling, leads to a set of variants. From this set, variants are prioritized according to a variety of criteria. After these filtering steps, the remaining variants are validated, i.e. their presence is confirmed using alternative techniques and their functional impact assessed. These steps are sequentially performed, and there are a rapidly growing number of software packages that can be used for each of these steps (for a recent review, see [14]) so that we omit details in favour of the discussion of general principles.

The study of Mendelian disease has been one of the earliest and most important successes of NGS [15, 16] and has allowed the identification of >300 novel disease-causing genes between 2010 and 2012 [17]. For recessive traits, in particular in consanguineous families, analysis of single individuals enabled the identification of causal alleles (see [18]). These successes were a key factor behind the renewed focus on rare diseases and are one of the factors driving the introduction of MPS into clinical diagnosis [17]. They depended to a substantial degree on the ability to predict the consequences of sequence changes. However, the failure rate is difficult to quantify given publication bias and the diversity of sequencing and analysis pipelines.

THE ROLE OF LINKAGE ANALYSIS IN THE ERA OF LARGE-SCALE SEQUENCING

In the case of Mendelian disease, where the assumption is that in an individual a single variant is responsible for disease susceptibility, one of the critical issues is filtering the large number of variants identified by variant calling [19]. This is achieved by applying a succession of steps, which can include criteria such as the predicted effect of the variant [20], whether the variant has already been observed and if it has, whether its reported frequency exceeds a certain threshold (see [21] for a discussion threshold choice). Where information from related individuals is available, criteria may include whether the variant is shared among affected individuals in the pedigree, or in an inbred recessive scenario whether affected individuals are homozygous for the variant [22] and for a simple recessive whether another variant that is predicted to be deleterious has been found in the same gene [23]. Some of these criteria, such as the prediction of the functional effect of a variant, do not yield a binary outcome, e.g. deleterious/not deleterious, but a continuous predictor, e.g. the probability that the variant is deleterious, and are therefore best seen as methods to prioritize variants for further analyses.

Linkage analysis can provide several pieces of information that may be used in this context. The first is the location of the variant of interest; the second is a measure of the strength of the evidence in for a particular region of the genome being involved in disease predisposition inheritance in a particular pedigree and the third is information on haplotypes.
In pedigrees where disease predisposition follows a dominant mode of inheritance and where only a limited (perhaps only two) affected putative gene carriers have been sequenced, one of the main difficulties can be the large number of variants shared between the sequenced individuals. However, if the region of interest has been delimited by linkage, the location with respect to the linkage interval can help to reduce the number of candidate variants to a more manageable size [24]. Bowden et al. [25] successfully identified a mutation in the ADIPOQ gene contributing to \( \sim 17\% \) of the variance of the quantitative trait plasma adiponectin levels in families selected for insulin resistance and adiposity. They had previously found very strong evidence for linkage to the 3q region and used their linkage signals to select the three most informative individuals to put through region targeted WES. More recently, two groups have used a similar single nucleotide polymorphism linkage analysis followed by WES to identify single causative mutations behind autosomal dominant forms of non-syndromic hearing loss [26, 27]. Linkage analysis can quantify the evidence for the involvement of variants at a particular chromosomal position in disease predisposition in a specific pedigree. The evidence for linkage is summarized as an logarithm of the odds score, and values above 3 are interpreted as evidence of linkage to that genomic region. This information can be used in studies involving a collection of pedigrees. In such a case, it may be desirable to prioritize individuals for sequencing from pedigrees where there is strong evidence for the involvement of a particular region of the genome, while pedigrees where linkage is unable to pinpoint a specific region are assigned a lower priority. This type of strategy has been investigated by Shi and Rao [28]. Within a pedigree, linkage analysis can also be used to select individuals for sequencing. One approach is to select individuals carrying putative at risk haplotypes in the region expected to contain the causative locus. This is of particular interest where there is incomplete penetrance, sporadic cases or genetic heterogeneity and in a wider context for familial study designs in diseases with a complex genetic component. An alternative is to choose for sequencing the smallest set of individuals that, in conjunction with genotyping data, enables the inference of the sequence in the region of interest for all individuals in the pedigree [29]. Currently, genotyping for linkage analysis is often done using genotyping microarrays, and only selected samples are sequenced. In the future, falling costs and improved variant calling will probably lead to a shift to genotyping by sequencing. Therefore, there will be no preliminary genotyping for sample or pedigree selection. This will have several consequences. The first is that sequencing can be used to estimate the expected degree of allele sharing directly, a procedure that may involve reconstructing familial relationships from the sequencing data for the whole genome and may reduce errors due to inaccurate pedigree information. The second is that linkage analysis will no longer be used to select individuals but rather to prioritize variants for further analysis.

Another emerging area of application for linkage analysis is in variant calling. The output of currently used next-generation sequencers represents a series of separate sequences. Each of them is called a read. Currently, most sequencers produce relatively short reads, 40–500 bases in length. The alignment step maps these reads to a reference sequence. Usually, this reference includes the sequences of the human nuclear and mitochondrial genomes. The alignment step allows the identification of the reads that cover any particular position of the reference sequence. Each read may show at such a position either the base corresponding to the reference or a deviation from it. Such a deviation reflects either the presence of a variant in the input material or errors that were introduced through the sequencing and alignment processes. A variant caller has to decide between these two possibilities. This means assessing the likelihood that the pattern of deviations observed at a certain position can be explained by sequencing or alignment errors or by the presence of a variant allele in the sample itself. Recent variant callers rely on models that will consider variables such as the ploidy of the starting material, the estimated error probability for each position or local sequence context [30]. In this context, ploidy refers to the number of different homologous chromosomes expected in the sample, and would be, for example, two for autosomes in a sample with material from a single human and twice the number of individuals in pooled samples. Some variant callers can also use information from comparable samples, i.e. samples that have been processed and sequenced at the same time or have a common ethnic background to improve variant calling [30, 31].

Failure to accurately call variants in the context of Mendelian disease can lead to false exclusions when
an existing variant is not detected in the sample from an obligate carrier and to false inclusions for when a variant is not called in a non-carrier. However, non-detection of a variant is often due to lack of coverage, an issue that is often not explicitly addressed by variant callers. However, using linkage information to reduce errors in calling is now incorporated into several approaches. These methods utilize the additional information obtained when related individuals are sequenced. Li et al. [32] proposed a calling procedure, which jointly considers the expected genotype sharing due to the pedigree structure and the aligned sequences. They demonstrated that this approach could significantly reduce errors even for positions where coverage was as high as 30-fold. Their implementation improved the power to detect the short insertions and deletions and also provides good power to detect de novo mutations. They extended this approach to incorporate the linkage disequilibrium between neighbouring sites and short range haplotypes when using unrelated samples or samples from parent–offspring pairs or trios [33]. Kojima et al. [34] offer a further extension to this approach using relatives from extended pedigrees in the calling procedure to rescue regions where insufficient coverage leads to apparently low levels of heterozygosity. The extra information from the segregation of alleles and short range haplotypes reduces the chance of missing a variant. As techniques emerge to allow longer read lengths, these haplotype-based approaches would be expected to further reduce variant calling errors. Peng et al. [35] systematically explored the effects of different parameters such as family structure and size, allele frequencies and coverage using a similar approach and showed that the improvement in variant calling were particularly evident for positions with low to moderate coverage (5- to 20-fold coverage).

CONCLUSION

Originally, one of the main applications of linkage analysis was to combine genotypic and phenotypic data to obtain the chromosomal location of the genetic variants affecting the phenotype of interest. Traditionally, this mapping process followed a sequence of events consisting of genotyping, linkage analysis and targeted sequencing. New, increasingly cost effective, genotyping technologies such as whole exome or genome sequencing are blurring the distinction between the genotyping and targeted sequencing stages, and methods using linkage analysis principles are more integrated into the analysis of the sequencing directly. As long as our ability to infer the functional consequences of the detected variants is limited, positional information derived from linkage analysis will help to uncover the genetic basis of human disease.

Key points
- The success of WES and WGS in identifying genes predisposing to Mendelian disease is redefining the role of linkage analysis.
- Linkage analysis has a new role in informing aspects of the study design for sequencing studies of Mendelian disease.
- These aspects include selection of pedigrees and individual samples used for sequencing, improving variant calling and filtering of candidate variants.

FUNDING

This work was supported by the University of Newcastle (M.F.S.K.) and the University of Sheffield (M.D.T.).

References


