Genetic linkage analysis in the age of whole-genome sequencing

Jurg Ott[1,2], Jing Wang[1] and Suzanne M. Leal[3]

Abstract | For many years, linkage analysis was the primary tool used for the genetic mapping of Mendelian and complex traits with familial aggregation. Linkage analysis was largely supplanted by the wide adoption of genome-wide association studies (GWASs). However, with the recent increased use of whole-genome sequencing (WGS), linkage analysis is again emerging as an important and powerful analysis method for the identification of genes involved in disease aetiology, often in conjunction with WGS filtering approaches. Here, we review the principles of linkage analysis and provide practical guidelines for carrying out linkage studies using WGS data.

Genetic mapping
The ordering of loci on a chromosome and the determination of the distances between two adjacent loci. For short distances, the recombination fraction can serve as a measure of genetic distance, with the unit of measurement being the centimorgan (cM); 1 cM = 1% recombination fraction.

[1]*Key Laboratory of Mental Health, Institute of Psychology, Chinese Academy of Sciences, 16 Lincui Road, Beijing 100101, China.*
[2]*Laboratory of Statistical Genetics, Rockefeller University, 1230 York Avenue, New York, New York 10065, USA.*
[3]*Center for Statistical Genetics, Department of Human and Molecular Genetics, Baylor College of Medicine, One Baylor Plaza, Houston, Texas 77030, USA.*
Correspondence to J.O. and S.M.L.
e-mails: ottjurg@psych.ac.cn; sleal@bcm.edu
doi:10.1038/nrg3908
Published online
31 March 2015

Linkage analysis was the predominant statistical genetic mapping approach used in the latter half of the twentieth century. More recently, the focus shifted to association studies of complex traits that analyse common variants, which have a modest effect. For such variants, association analyses are more powerful than linkage analyses, and genome-wide association studies (GWASs) using single-nucleotide polymorphism (SNP) marker loci became the preferred association mapping tool. However, an emerging view is that rare variants, which are not well interrogated by GWASs, could be responsible for a substantial proportion of complex human disease[1]. Importantly, the increased availability of exome and whole-genome sequence data has brought linkage analysis once again to the forefront owing to the development of powerful methods to detect rare variants involved in disease aetiology using family-based data; such an approach has many advantages over simply using filter methods to identify causal variants. Several reviews[2–5] and books[6–8] have been written on genetic linkage analysis, but none, to our knowledge, covers linkage analysis coupled with whole-genome sequencing (WGS).

Several recent studies have generated genome-wide association data for families. For example, the T2D-GENES (Type 2 Diabetes Genetic Exploration by Next-generation sequencing in Ethnic Samples) consortium has generated WGS data on 1,043 individuals from 20 Mexican families and reported analysis of risk variants for type 2 diabetes. However, for cost reasons, most studies currently only obtain WGS data for a small number of family members.

To date, most family-based WGS studies have therefore been analysed using filtering approaches, and only a few family members are prioritized for sequencing (FIG. 1). However, filtering approaches do not offer statistical evidence of a variant's involvement in disease susceptibility, whereas linkage analysis does provide this statistical support. With the decreasing cost of sequencing, it will become more commonplace to have WGS data available for every informative pedigree member.

This Review provides the reader with a practical guide for performing linkage analysis to identify variants that are responsible for Mendelian[9] trait aetiology. After briefly mentioning the relative merits of linkage and association analysis, we discuss linkage algorithms and their implementations in computer programs, with a special emphasis on the use of sequence data. We then outline a step-by-step approach to successful linkage analysis using WGS data.

Genome-wide linkage analysis

For all informative family members, genotypes can be generated using SNP arrays and analysed using genome-wide linkage analysis. This approach is beneficial in that it evaluates DNA sample quality; elucidates whether specified familial relationships are correct; allows the detection of mis-specification of affection status and locus heterogeneity; aids the selection of an individual (or individuals) to undergo WGS; and facilitates the mapping of the disease locus to a region (or regions) of the genome, thus reducing the number of variants that need to be followed up. Linkage analysis can also provide statistical evidence of the involvement of a variant or gene in disease aetiology and can be performed either directly using WGS data or after filtering using data on variants

© 2015 Macmillan Publishers Limited. All rights reserved

that have been followed up by sequencing[10] across entire families. However, it should be noted that although linkage analysis provides statistical evidence that a variant is involved in disease aetiology, false positives can occur when the variant that is tested is only in linkage disequilibrium with the causal variant. When filter approaches are used, phenocopies[11,12] and reduced penetrance can inhibit the ability to elucidate the causal variant but, because parametric linkage analysis incorporates a penetrance model, even under these circumstances the causal variant can usually be mapped.

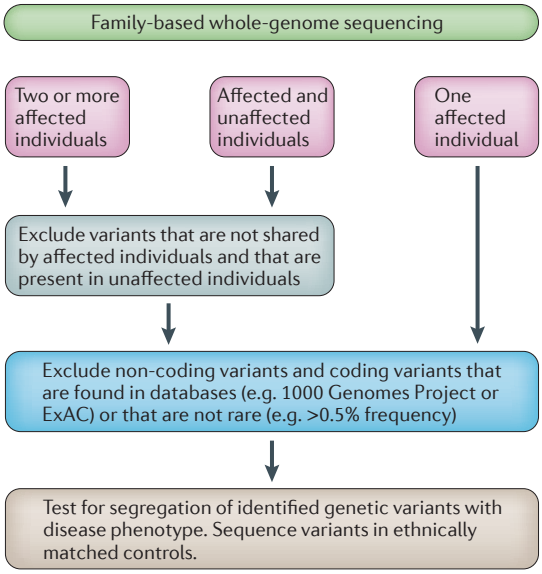


Figure 1 | **Workflow for the whole-genome sequencing filtering approach in human family data.** Usually, one, two or more affected individuals, or affected and unaffected individuals, in a family have their genomes or exomes sequenced. Variants that are not predicted to be nonsense, missense or splice-site variants are usually excluded from further analyses because it is unlikely that they are causal. When the mode of inheritance of a disease is known, this information can be used to aid the selection of variants. For example, for an autosomal dominant disease, the affected pedigree member's sequence data should display a heterozygous causal variant. Sequence data on additional pedigree members can help to reduce the number of variants that could potentially be disease causing. A final filtering step is performed in which those variants that are present in the databases dbSNP, 1000 Genomes, ExAC and Exome Variant Server are excluded. Additionally, bioinformatic tools, such as Polyphen-2 (REF. 102), and measures of conservation, for example, PhyloP[103], are often used to predict whether a variant is deleterious and therefore likely to be disease causing. Even after filtering steps, there may be many variants that need to be followed up in the remaining family members to elucidate whether the variant (or variants) segregate with the disease phenotype. If the family is from a population that is not represented in databases, then ethnically matched controls need to be sequenced to evaluate the frequency of the variant (or variants).

Association analysis versus linkage analysis

Pertinent reviews of family-based association analysis have previously been published[13–15], and only highlights are therefore presented here. Genetic linkage and association between two loci are both related to recombination — in the former, recombination events are scored over a limited number of observed generations, whereas the latter relies on large numbers of unobserved recombination events in past generations. As generations go by after an initial disease mutation has occurred, recombination events (crossing over) with surrounding markers tend to occur closer and closer to the disease locus so that measurable association between disease and marker loci extends only over short distances of up to 100 kb[16,17], corresponding approximately to a recombination fraction (represented by θ) of 0.001, given 1 Mb ≈ 1 cM. Most differences between association and linkage analysis are due to this difference in the number of generations.

Association analysis using common variants generally allows for finer mapping than linkage analysis using SNP loci, but one potentially problematic aspect of association analysis is population stratification, which can lead to an increased number of false-positive results if not properly accounted for[18]. This is not a problem in linkage analysis because children's genotypes depend on those of their parents and not on population genotype frequencies. However, if some parental genotype data are missing, using incorrect marker allele frequencies can increase type I and II errors. It has thus been tempting to combine positive aspects of linkage and association analysis, which may be achieved by using family-based rather than population-based control individuals. Consider an affected individual and his or her parents. At a given marker locus, the alleles inherited by the child may be contrasted with the alleles that are not inherited[19,20], where the latter can be shown to be representative of the alleles in the population[21]. The most well-known use of such family-based controls is probably the transmission disequilibrium test (TDT)[22]. For this to apply to multiple offspring, the null hypothesis of the TDT must include absence of linkage ($\theta = 0.5$), so the TDT is a test for linkage that is only powerful when there is both linkage and association[21]. The TDT has been extended (the rare variant-TDT (RV-TDT))[23] for use with WGS data incorporating several rare variant association tests and has been implemented in the Family-Based Association Test Toolkit (FBAT) suite of programs[24]. Some rare variant association tests[25] analyse variants in aggregate (usually across a genomic region such as a gene) instead of analysing individual rare variants. It has been shown that analysing rare variants in aggregate is much more powerful than the individual analysis of rare variants[25,26].

Approaches for linkage analysis

LOD scores. Linkage analysis can be carried out between a putative disease locus and a single marker locus (two-point linkage) or across a set of markers (multipoint analysis) consisting of a small number of markers or

Genetic linkage
A phenomenon whereby two alleles, one each at two different loci, are transmitted together from parents to offspring more often than expected by chance. It leads to a recombination fraction smaller than 0.5.

Phenocopies
Individuals that exhibit the phenotype of a Mendelian trait but that are not carriers of a susceptible genotype. Phenocopies were thought to result from non-genetic factors, but genes at locations other than those under current consideration can also lead to (genetic) phenocopies.

Penetrance
The conditional probability of being affected given one of the genotypes at the disease locus, '++', '+d' or 'dd', where 'd' is the disease allele and '+' the non-disease (wild-type) allele. More generally, penetrance is the conditional probability of a phenotype given a genotype.

Recombination
Two alleles, one from each of two loci, can be inherited from one parent but originate from two different grandparents. If the two marker loci are on the same chromosome, a recombination is the result of an odd number of crossovers between the markers.

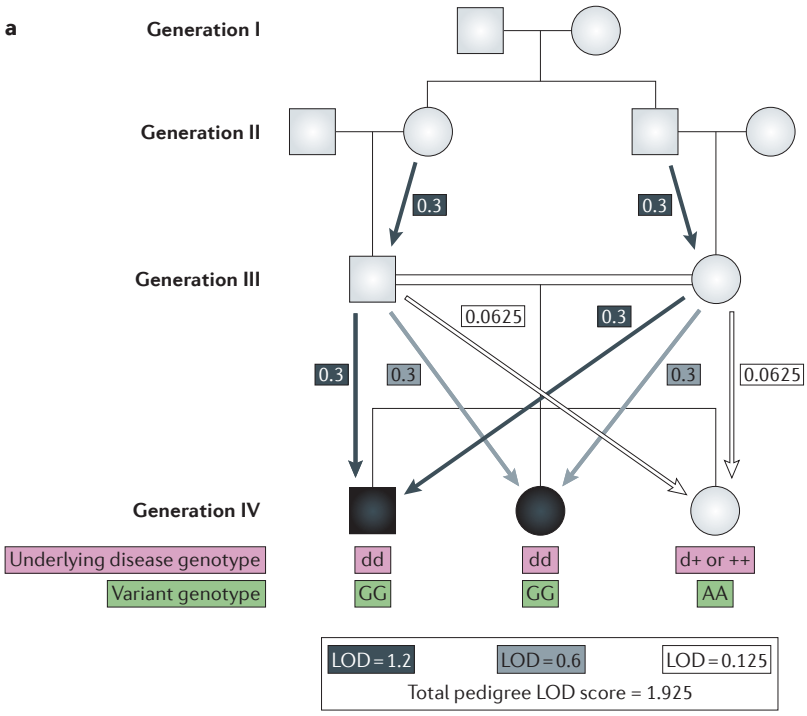
Crossing over
A cytogenetic phenomenon that occurs during the formation of human gametes (egg or sperm cells). The salient feature of crossing over is that it occurs semi-randomly along chromosomes, with at least one crossover occurring on each chromosome in meiosis.

Recombination fraction
(θ). The expected proportion of recombinant children divided by the total number of recombinant and non-recombinant children. For two loci in close proximity to each other, θ is small owing to the randomness of crossing over, but it increases to 0.5 for loci that are far apart.

© 2015 Macmillan Publishers Limited. All rights reserved

LOD score
$Z(x) = \log_{10}[L(x)/L(\infty)]$ is the logarithm of the likelihood ratio, with the numerator being calculated under the assumption of linkage and the denominator under no linkage. A LOD score of 3.3 or higher has been shown to correspond to a genome-wide significance level of 0.05.

even all markers on a given chromosome. For multipoint analysis, the LOD score, $Z(x) = \log_{10}[L(x)/L(\infty)]$, is computed as the logarithm of the likelihood ratio, with the numerator specifying a position, x, of the putative disease locus on the marker map. For the denominator, one assumes the disease locus to be off the map — that is, infinitely far away from the markers (FIG. 2). The multipoint LOD score can furnish a curve over all markers on a chromosome (FIG. 3); the maximum of this curve, over all chromosomes, then represents the estimated position of the disease locus on the human gene map provided that the maximum LOD score is at least equal to 3.3 (REF. 101). Evidence for linkage can be obtained from a single pedigree or multiple pedigrees with LOD scores summed at the same θ or map position. When linkage analysis was previously performed with marker loci and the individual genes within a region had to be sequenced using, for example, Sanger sequencing, false-positive regions would not be followed up owing to reasons of time and cost, so it was important for a pedigree or a group of pedigrees to meet the genome-wide significance level. There is less concern now with meeting this criterion because it is quick and relatively inexpensive to follow by WGS of associated regions. Smaller pedigrees with suggestive LOD scores can still be followed up with WGS, although there may be multiple linkage regions that could potentially harbour the causative variant. If a putative causal variant is identified in a small pedigree, it is imperative that additional families are identified that segregate either the same variant or another putatively causal variant within the same gene. If a variant is identified that segregates with a



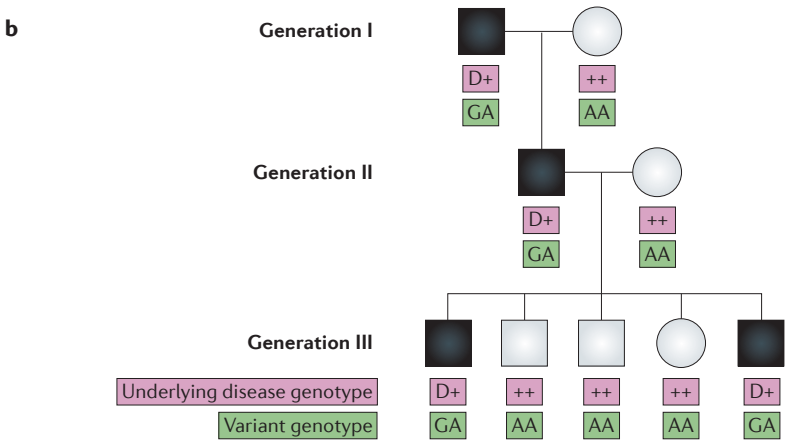


Figure 2 | **Linkage information for a first-cousin mating for an autosomal recessive trait and a phase-known autosomal dominant trait.** The disease is fully penetrant without phenocopies and has a minor allele frequency of 0.0001. Circles represent females and squares males. Individuals represented by solid black symbols are affected, and individuals represented by white symbols are unaffected. Shown below each individual in generation IV are the possible underlying disease genotypes. **a** | An autosomal recessive trait pedigree in which the affected children are offspring of first-cousin parents is shown. Consanguinity is indicated by the double horizontal line. The affected individuals are homozygous for a variant that is either causal or in perfect linkage disequilibrium with the causal variant. The unaffected sibling is homozygous wild type. The arrows show each informative meiosis and the contribution to the LOD score. For this pedigree configuration, the rare variant must have entered the pedigree through one of the great-grandparents. The meiosis events from the great-grandparents to their children do not contribute to the LOD score; however, the meiosis events from the affected children's grandparents to their parents and from the parents to the first affected child each contribute 0.3 to the LOD score, yielding a total LOD score of 1.2. The second affected child only adds 0.6 to the LOD score for the family because only the meioses from her parents yield new linkage information. Each additional unaffected child only yields an additional LOD score of 0.125 because for unaffected children it is not possible to elucidate whether they are homozygous wild type or causal-variant carriers; each of these possibilities have a probability of 1/3 and 2/3, respectively. These two probabilities are incorporated into the calculation of the LOD score, and linkage information is therefore lost. **b** | A phase-known autosomal dominant pedigree with five children is shown. This pedigree with five offspring for which there are no recombination events will lead to a maximum LOD score of 1.5 at $\theta = 0$, where $Z(\theta) = \log_{10}[(1-\theta)^5/(\frac{1}{2})^5]$. However, if no genotype information is available for the grandparents (shown in generation I), making the pedigree phase-unknown, the pedigree will yield a maximum LOD score of 1.2 at $\theta = 0$, where $Z(\theta) = \log_{10}[((1-\theta)^5 + \theta^5)/((\frac{1}{2})^5 + (\frac{1}{2})^5)]$.

© 2015 Macmillan Publishers Limited. All rights reserved

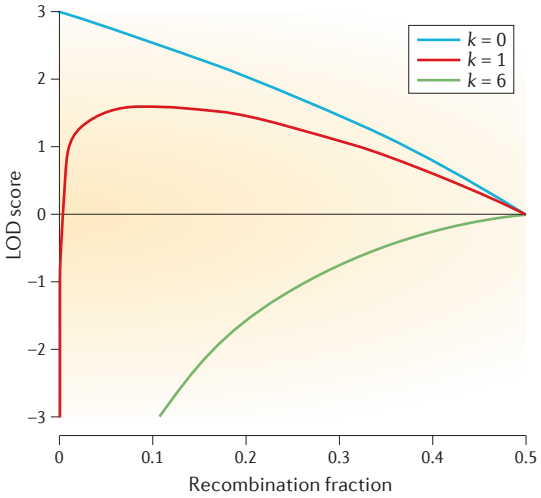


Figure 3 | **LOD score curves for a phase-known autosomal dominant pedigree with ten children in the third generation.** The LOD score curve is displayed for *k* recombination events (*k* = 0, 1 and 6) out of 10 meioses. The disease phenotype segregating in this pedigree is fully penetrant and has no phenocopies. Phenotype and genotype information are available for all pedigree members. The marker locus that is analysed is fully informative. The maximum LOD scores are 3.0 at a recombination fraction of 0 (θ = 0), 1.6 at θ = 0.1 and 0 at θ = 0.5, respectively, for *k* = 0, 1 and 6. When multiple pedigrees are analysed, the resulting LOD scores can be summed across families at either the same θ value or the same map position.

phenotype in a large pedigree and produces a LOD score >3.3, it is also desirable to have additional pedigrees that segregate the same variant or different variants within the same gene. Even with a significant LOD score, the finding could be a false positive or the variant that was identified may only be in linkage disequilibrium with the causal variant, which may not have been observed in the sequence data; for example, the variant might not have been captured or there might have been insufficient read depth. Additionally, performing functional studies can be important on two levels: to provide additional evidence for gene causality and to better understand the role of the gene in disease aetiology[27].

A parent must be heterozygous at each of two loci to be 'informative for linkage'; otherwise, there is insufficient information to distinguish recombinant from non-recombinant events in offspring. When grandparents are unavailable (that is, in 'phase-unknown pedigrees'), there must be at least two children in the third generation for linkage to be potentially informative. In some instances, such as for autosomal recessive and X-linked recessive traits, grandparents do not help to set the phase because they are usually unaffected and disease allele carriers cannot be distinguished from non-carriers. Pedigrees in which the grandparents provide phase information are known as 'phase-known pedigrees'. In suitable situations, the number of recombinant events (k) and of non-recombinant events ($n-k$) can be counted directly. The estimate of the recombination fraction is then simply $\theta = k/n$. Generally, however, the recombination fraction is estimated by the maximum likelihood (LOD score) method.

The recombination fraction tends to be different in males and females. It may also depend on age[28], but human studies have provided varied results[29–32].

Penetrance. For many traits, penetrance is incomplete. For example, in torsion dystonia, penetrance has been estimated as 29%[33]; that is, fewer than one-third of disease-gene carriers express the trait. Penetrance can be age and sex dependent. For example, in Huntington disease, penetrance is zero at birth and gradually increases to 100% later in life[34,35]. Multiple penetrance classes in linkage analysis can have functions similar to those of predictor variables in logistic regression for case–control association studies[6,36]. If the penetrance for a disease is unknown or not well established, an 'affected-only' analysis can be performed, in which individuals who are unaffected are given an unknown affection status.

We distinguish between two penetrances: g for genetic cases and f for phenocopies, with $g > f$. In many linkage studies, f is taken to be a small number, such as 0.01 or smaller. The penetrance ratio, g/f, is analogous to the risk ratio in epidemiology[37] and indicates how well the disease phenotype (or any phenotype, for that matter) can discriminate between underlying genotypes.

Initial SNP genotyping. Performing linkage analysis with a SNP genotyping array can be beneficial in the identification of concerns about the data set. SNP genotyping can elucidate potential problems with the quality of the DNA samples, detect whether samples have been swapped and indicate instances in which a relationship has not been correctly specified. Additionally, if the pedigree produces a much lower LOD score than that expected for the number of informative meioses, this might indicate problems with phenotypic information that need to be rectified before additional analysis can be performed, or it might indicate that locus heterogeneity is present in the pedigree. The resulting linkage results and haplotype reconstruction can also aid in the selection of individuals for WGS: selection could be based on the smallest haplotype or on a haplotype that overlaps across affected individuals. Additionally, haplotype information can elucidate whether there are individuals within a pedigree who are phenocopies and therefore should not be selected for WGS. After performing WGS, fewer variants need to be followed up than when performing filtering alone because the causal variant is likely to be in the linked regions. For a family that can establish linkage, this strategy usually only yields 1–3 variants that have to be followed up in additional pedigree members and ethnically matched controls.

Linkage algorithms. With few exceptions[38], the calculation of pedigree likelihoods is done recursively by starting with a portion of the data and then working

© 2015 Macmillan Publishers Limited. All rights reserved

through the rest of the data. Two main algorithms are in general use. The Elston–Stewart algorithm[39–42] recursion takes place over individuals in a pedigree so that computing effort is linear with pedigree size but increases exponentially with the number of loci considered simultaneously. Conversely, the Lander–Green algorithm[43] recursion takes place over loci so that computing effort increases linearly with the number of loci but exponentially with family size. For multipoint analysis, the marker map is generally limited to 6–8 markers in the Elston–Stewart algorithm, whereas thousands of markers on any chromosome can be accommodated by the Lander–Green algorithm. However, the Lander–Green algorithm can only handle small- to medium-sized families, whereas the Elston–Stewart algorithm is applicable to very large pedigrees. Many programs have been developed that implement the Elston–Stewart algorithm (for example, LINKAGE[44] and FASTLINK[45]) and the Lander–Green algorithm (for example, GeneHunter[46] and MERLIN[47]). Although these programs were not developed for analysing WGS data, analysis of sequence data can easily be performed by converting Variant Cell Format (VCF) files into the linkage file format (also known as the PLINK[48] file format). These conversions can readily be made with PLINK version 1.9, VCFtools[49], Variant Association Tools (VAT)[50] or SEQLinkage[51]. If the conversion is performed using either VCFtools or VAT, the user will have to create a file that contains parameter information. However, SEQLinkage will create both the pedigree and parameter files for direct use in linkage analysis. Additionally, SEQLinkage can be used to directly perform linkage analysis using VCF files. However, the analysis of individual rare variants can be poorly powered when using SEQLinkage, so the software authors, motivated by rare-variant association tests, developed the collapsed haplotype pattern (CHP) method, which aggregates rare variants within regions (usually a gene) to create a 'super locus' (REF. 51). The CHP method is more powerful than analysing rare variants individually, particularly in the presence of allelic heterogeneity.

There is no exact algorithm that can realistically accommodate large families and large numbers of loci, but computer-based methods have been developed to approximate linkage likelihoods for these situations. They are generally based on Markov-chain Monte Carlo (MCMC) methods[52] and can allow for multiple disease loci, large family pedigrees and large numbers of marker loci. Two examples of linkage analysis programs that use MCMC approaches are Loki[53] and SimWalk2 (REF. 54).

Multipoint analysis is useful when analysing SNP genotyping arrays for which the genotypes for the causal variant are unavailable because analysing multiple markers is usually more informative than analysing an individual SNP marker locus. However, for sequence data, there is no advantage in performing multipoint analysis if genotype data are available for the causal variant because no additional linkage information will be obtained.

Mendelian inheritance model
The Mendelian laws of inheritance, when applied to variants, stipulate that an individual carries two copies (alleles) of a given nucleotide and passes one of them at random to each of their offspring. Disease may be the result of a single copy of the allele (dominant inheritance) or of two copies (recessive inheritance) in an individual.

Parameter-free methods. So-called parameter-based ('parametric') methods require specification of an inheritance model for the trait locus, unlike allele sharing (parameter-free) methods, which do not require specification of a disease model. Parameter-free methods are sometimes referred to as 'non-parametric', but this term should be avoided because it means, in the statistics literature, that analysis is carried out on ranks rather than the original data, which is not the case for these methods. The simplest type of allele-sharing analysis is based on affected sibpairs (ASPs)[55,56], but more-sophisticated approaches have been developed[57,58]. However, many of these methods do imply a specific Mendelian inheritance model. For example, analysis of identity-by-descent (IBD) sharing in affected siblings has been shown to be equivalent to an analysis under a fully penetrant recessive mode of inheritance[59]. Newer parameter-free approaches make use of large pedigrees and both affected and unaffected individuals[58]. Allele-sharing linkage methods are based on the principle that if two relatives with a similar phenotype (for example, both affected) inherit the same marker allele from a common ancestor more often than expected by chance, then this indicates that a disease locus is linked with the marker locus. Various sharing statistics have been developed[57], but the subject of parameter-free linkage analysis is beyond the scope of this Review.

Extended approaches. Methods have also been developed to allow for two trait loci[56,60,61], often referred to as digenic inheritance[62], but it is not entirely clear what LOD score threshold for significance should be applied to such bivariate analyses[63], and their power gain over single-locus analyses has been questioned[64].

For children who are affected with an autosomal recessive trait and whose parents are cousins or similarly close relations, marker loci linked with the trait locus tend to be homozygous[65]. These runs of homozygosity can be quickly detected for either SNP genotype array or WGS data using, for example, Homozygosity Mapper[66]. Linkage analysis can be used to obtain LOD scores for the variants within the region (or regions) of homozygosity through multipoint analysis (for SNP genotyping array data) or two-point linkage (for WGS data). It should be noted that in the very rare circumstance that the disease trait in a consanguineous pedigree is due to compound heterozygous variants[67], homozygosity mapping will not lead to detection of the region harbouring the causal variants, although linkage analysis results will not be influenced by the variants being compound heterozygotes instead of being homozygous.

Steps for a successful linkage study

Phenotyping. Two classes of phenotypes can be distinguished: qualitative and quantitative traits. Qualitative traits consist of a discrete number of classes, such as 'affected' and 'unaffected', whereas quantitative traits occur with a continuous distribution. In this Review, we focus on qualitative (disease) traits.

For many traits there is little question as to who is affected and who is not. Even when disease definition might be ambiguious, there are usually medical rules to

© 2015 Macmillan Publishers Limited. All rights reserved

determine disease status — for example, the conditions that need to be satisfied for someone to be diagnosed as schizophrenic. Whether these rules are genetically relevant is generally unclear, and researchers sometimes choose to rely on 'endophenotypes'; that is, phenotypes correlated with disease that might be closer to gene action than the overall disease definition. It can also be more powerful to analyse separate underlying quantitative phenotypes instead of an overall clinical phenotype (for example, hypertension) that might be based on several quantitative traits. For example, rather than applying the medical diagnosis of hypertension, researchers working with the Lyon hypertensive rat carried out linkage analysis with each of three components of blood pressure (systolic, diastolic and pulse pressure, with pulse pressure being the difference between systolic and diastolic blood pressure); they found significant results for two different loci that each control a different blood pressure component[68]. Such clear results might have been difficult to obtain if hypertension was considered as a single phenotype.

Various approaches can be taken to accommodate multiple phenotypes involved in a disease. For example, two different lipid levels have been analysed jointly in a bivariate analysis relating to diabetes[69]. Most often, however, multiple phenotypes are suitably combined as a weighted sum[70–73], which is then used as a single quantitative trait in linkage analysis, or tests on single phenotypes are combined to show their joint effect[74]. It is best to avoid dichotomizing a quantitative trait because substantial linkage information can be lost. A number of programs, including FASTLINK and MERLIN, can perform quantitative trait linkage analysis.

Selecting family members for sequencing. If only a fraction of all family members can be sequenced owing to reasons of cost, scientists are faced with the dilemma of which pedigree members to select for WGS. SNP genotyping data can aid selection, but such data are not always available. Some general guidelines are given below, but more advanced approaches rely on computer simulation, which emulates a linkage study by generating marker data and analysing it with the same parameter that will be used in the linkage study. For example, SLINK could be used to generate marker data, and MSIM could be used to perform the analysis. A sophisticated statistical framework for prioritizing individuals for sequencing has recently been developed and implemented in a computer program called GIGI-Pick[42] (TABLE 1).

Consider an autosomal recessive trait that is carried by two unaffected parents, who are cousins, and by their two children, who are affected with the trait. At least one of the affected offspring should be sequenced because, in this family, this child can yield a LOD score of 1.20 when both the trait and the linked variant alleles are rare. If an additional individual is to be sequenced, should this be a parent or the other affected child? A parent with an unknown genotype is likely to be heterozygous for a rare variant, and the affected child will have one variant in common with the parent, so each additional affected child can produce an LOD score increment of 0.60. Thus, it is less important to sequence the parents than to sequence the affected siblings in this situation. However, sequencing parents is necessary for identifying compound heterozygotes and *de novo* events.

For dominant traits, it is generally best to sequence distantly related individuals, a principle established some 20 years ago[75]. The same rare allele occurring in two relatives is likely to represent two copies of an ancestral allele rather than two alleles independently acquired by the two individuals, which translates into an LOD score that increases with increasing distance of relationship. For rare variants for a disease without phenocopies, even only two distantly related individuals can yield sufficient linkage information. For example, consider two second cousins affected with an autosomal dominant trait for which the causal variant has a minor allele frequency of 0.0001; all other relatives are of unknown disease and marker status. The resulting LOD score is equal to 1.20, and more-distant relationships can yield even higher LOD scores.

For WGS studies, one or two unaffected individuals in a family should also be sequenced as controls, but in linkage analysis a negative LOD score is a sufficient indication that a given variant is not linked with the trait gene. For traits with reduced penetrance, unaffected pedigree members can be carriers of causal variants and therefore do not make ideal controls. Usually, there is no need to obtain variant frequencies in unaffected controls because this same information can readily be obtained from databases such as dbSNP[76], Exome Variant Server[77], ExAC and 1000 Genomes[78]. Variants that occur at higher frequencies in these databases — for example, >0.5% — are unlikely to be causal[1]. It should be noted that even fully penetrant disease variants may be present in variant databases for several reasons: these are not databases of disease-free individuals, and for autosomal recessive traits disease-free carriers may be included.

SNPs from sequence data. After individuals in a family have been sequenced, variants are extracted from the sequence data[79]. If a given variant is not observed in available databases, it can be assumed to be rare and given a low allele frequency: for example, 0.0001. When performing linkage analysis, it is necessary to have VCF files that contain genotype information for every family member for which there is a variant site in at least one of the pedigree members. If this information is not available, it is impossible to distinguish between missing data and an individual who is a homozygous non-carrier. Additionally, if a sufficient number of family members are being subjected to WGS, a family-aware variant caller[80], such as that implemented in the Genome Analysis Toolkit (GATK)[81], should be used to increase the accuracy of the variant calls.

Quality control. For WGS data, quality control can be performed as previously described[50]; however, these procedures will not completely eliminate genotyping errors from WGS data. In contrast to association studies[82], in families genotyping errors have traditionally been

© 2015 Macmillan Publishers Limited. All rights reserved

Table 1 | **Computer implementations**

Software	Description and purpose	URL	Refs
easyLINKAGE	Integrated suite that generates the necessary files and performs analysis using several programs, including GeneHunter	http://sourceforge.net/projects/easylinkage/	92
Genome Analysis Toolkit (GATK)	Toolkit for call variants from WGS data	https://www.broadinstitute.org/gatk/	81
GeneHunter	• Lander–Green algorithm • Can be applied to small- to medium-sized families and large numbers of marker loci • Features parametric and non-parametric multipoint linkage analysis for qualitative and quantitative traits	http://www.broadinstitute.org/ftp/distribution/software/genehunter/	46
GIGI-Check	• MCMC algorithm • Detects Mendelian-consistent genotyping errors	https://faculty.washington.edu/wijsman/progdists/gigi/software/GIGI-Check/GIGI-Check.html	86
GIGI-Pick	Can be used to prioritize individuals for WGS	https://faculty.washington.edu/wijsman/progdists/gigi/software/GIGI-Pick/GIGI-Pick.html	42
LINKAGE and FASTLINK	• Elston–Stewart algorithm • Can be applied to large families but only to a limited number of marker loci for multipoint, parametric two-point and parametric multipoint linkage analysis for qualitative and quantitative traits • Can be used to detect Mendelian-consistent genotyping errors	http://www.jurgott.org/linkage/LinkagePC.html; http://www.ncbi.nlm.nih.gov/CBBresearch/Schaffer/fastlink.html	44, 93
Loki	• MCMC algorithm • Can be used to perform multipoint linkage and segregation analysis of quantitative traits on large pedigrees	http://www.stat.washington.edu/thompson/Genepi/Loki.shtml	94
Mendel	• Elston–Stewart and Lander–Green algorithms • Can be applied to the analysis of qualitative or quantitative traits in pedigree- or population-based data • Can combine multiple (rare) variants into superloci	http://www.genetics.ucla.edu/software/mendel	95, 96
MERLIN	• Lander–Green algorithm • Can be applied to small- to medium-sized pedigrees • Handles closely spaced SNPs by combining them into superloci	http://www.sph.umich.edu/csg/abecasis/Merlin/	47
MSIM	• Elston–Stewart algorithm • Can be used to analyse simulated pedigree data to evaluate power, maximum LOD scores and expected LOD scores • Useful for simulation studies to evaluate the most-informative pedigree members to select for WGS	http://watson.hgen.pitt.edu/docs/SLink.html	97
PedCheck	Detects genotype incompatibilities in pedigree data	http://watson.hgen.pitt.edu/register/docs/pedcheck.html	98
PLINK	• Whole-genome association tool set for genotype data • Can be used to estimate IBD sharing between two individuals	http://pngu.mgh.harvard.edu/~purcell/plink/	48
Pseudomarker	• Family-based association testing (joint linkage and linkage analysis) for qualitative traits using cases and controls, trios, sibpairs, sibships and extended families	http://www.helsinki.fi/~tsjuntun/pseudomarker/	99
SEQLinkage	• Elston–Stewart algorithm • Can be used for parametric linkage analysis of WGS data using the collapsed haplotype pattern method • Generates linkage files from VCF files for use with any linkage program that performs parametric linkage analysis	http://bioinformatics.org/seqlink	51
SimWalk2	• MCMC algorithm • Can handle large pedigrees and an intermediate number of marker loci for parametric and non-parametric multipoint linkage analysis of qualitative and quantitative traits	http://www.genetics.ucla.edu/software/	54
SLINK and FastSLINK	• Simulates pedigree data that are conditional and unconditional on qualitative and quantitative phenotypes • Limited in size of pedigree and number of marker loci • MSIM can be used to analyse the simulated pedigrees	http://watson.hgen.pitt.edu/docs/SLink.html	97
Superlink	• Bayesian networks • Can handle large, complex pedigrees with multiple inbreeding loops segregating dichotomous traits • Performs two-point and multipoint (with a limited number of markers) linkage analysis	http://bioinfo.cs.technion.ac.il/superlink/	100

© 2015 Macmillan Publishers Limited. All rights reserved

Table 1 (cont.) | **Computer implementations**

Software	Description and purpose	URL	Refs
TLINKAGE	• Elston–Stewart algorithm • Can handle large pedigrees • Performs parametric two-locus linkage analysis	http://www.jurgott.org/linkage/tlinkage.htm	63
Variant Association Tools (VAT)	• Pipeline for quality control and analysis of WGS and genotype data • Can be used to generate linkage files for VCF files and to estimate IBD sharing between a pair of individuals	http://varianttools.sourceforge.net/VAT	50
VCFtools	Program to manipulate VCF files	http://vcftools.sourceforge.net/	49

IBD, identity-by-descent; MCMC, Markov-chain Monte Carlo; SNP, single-nucleotide polymorphism; VCF, Variant Cell Format; WGS, whole-genome sequencing.

detected as Mendelian inconsistencies[83–85]. However, particularly in small families, and given the biallelic nature of most variants, all sequencing errors will not be detected as Mendelian inconsistencies, and the fraction of such undetected errors can be high. MERLIN (double recombination events over short distances) and GIGI-Check (MCMC approach) are able to detect Mendelian-consistent errors[86].

Mendelian inconsistencies may be due to sequencing error or pedigree inconsistency (adoption, non-paternity[87] or swapped samples). In the case of pedigree inconsistency, large numbers of variants are expected to exhibit inconsistencies. To identify a specific individual causing these errors, it is useful to estimate the proportion of alleles shared IBD 0, 1 and 2 for pairs of individuals (implemented, for example, in the VAT or the PLINK programs). For example, for siblings, these proportions are expected to be 0.25, 0.5 and 0.25, respectively; if IBD proportions deviate from these values, then the two individuals are unlikely to be full siblings.

Computing LOD scores. To compute parameter-based LOD scores for linkage between a hypothesized disease locus and a given DNA variant, one needs to specify Mendelian model parameters such as allele frequencies and penetrances. Some handy rules are as follows. For example, consider a recessive trait so that the (homozygous) susceptibility genotype has frequency p^2, where p is the disease allele frequency under Hardy–Weinberg equilibrium. Trait population frequency (K) is then predicted to be $K = gp^2 + f(1 - p^2)$, where g and f are the respective penetrances for genetic cases and phenocopies. Fixing, for example, $f = 0.01$ and $g = 0.90$, allows the disease allele frequency to be determined as $p = \sqrt{[(K - f)/(g - f)]}$. In large families, penetrances may be estimated by maximum likelihood in suitable computer programs, but this is rarely done. Rather, one may determine the fraction of obligate disease-gene carriers who are unaffected, which should be approximately equal to $1 - g$. For age-dependent penetrance, it is generally sufficient to find two time points, a_1 and a_2, where a_1 is the youngest age at which anyone has been diagnosed with the disease and a_2 is the oldest age at which the disease has manifested. Then, in a coordinate system with age as the x axis and penetrance as the y axis, the age-of-onset curve is approximated by a straight line rising from a penetrance of 0 at a_1 to the maximum penetrance at a_2 (REF. 7).

Heterogeneity. Two types of heterogeneity may be distinguished: locus heterogeneity and allelic heterogeneity. Allelic heterogeneity refers to different alleles at the same locus (gene) conferring disease risk on different families or individuals, whereas in locus heterogeneity, different genes, possibly on different chromosomes, are disease causing. In linkage analysis, in contrast to association analysis performed with SNP marker loci, allelic heterogeneity does not generally represent a problem because linkage refers to a relationship between loci, not alleles. With WGS data, variants at different sites in the same gene may lead to disease; when such rare variants are being analysed using rare-variant association methods, allelic heterogeneity does not present a problem because rare variants within a gene region are analysed in aggregate. When allelic heterogeneity is present within a causal gene and individual variants are analysed, there can be a great loss of power because different pedigrees will not be informative for the same variant for pedigrees in which disease aetiology is due to the same gene but not the same variant; therefore, when LOD scores are summed across pedigrees, most pedigrees will not be informative and the power to detect linkage will be low. However, this problem can be avoided by using the CHP method described above, which analyses rare variants within a gene region in aggregate.

When analysing SNP marker loci, locus heterogeneity generally leads to a mixture of families that do and do not exhibit linkage to a given variant. Thus, in addition to estimating the recombination fraction θ in families with linkage, at the same time one also estimates the proportion (α) of linked families. The likelihood is maximized over α and θ, and the resulting LOD scores are known as heterogeneity LOD scores (HLODs)[88]. When analysing rare variants, locus heterogeneity does not usually have a great impact because families that are not linked to the causal gene generally do not have an informative variant within the causal gene region; therefore, instead of producing negative LOD scores, they are uninformative for linkage.

Conclusion

Linkage analysis is again emerging as an extremely useful method in genomic analysis, particularly for the identification of rare variants associated with a complex trait with high penetrance. Linkage analysis has many advantages over filtering approaches in terms of limiting the number of genes that have to be analysed;

© 2015 Macmillan Publishers Limited. All rights reserved

namely, it takes account of phenocopies and reduced penetrance, which are often features of Mendelian traits, and in addition it provides statistical evidence of the involvement of a variant in disease aetiology. Many new disease susceptibility genes have been successfully identified using linkage analysis coupled with WGS, and this strategy has been successfully used to identify the association of rare variants to phenotypic traits such as hearing impairment[10,89], familial goitres[90] and familial hypertension[91]. In the future, with the reduction in cost of WGS, linkage analysis of WGS data will be widely used.

1. McClellan, J. & King, M. C. Genetic heterogeneity in human disease. *Cell* **141**, 210–217 (2010).
2. Lander, E. S. & Schork, N. J. Genetic dissection of complex traits. *Science* **265**, 2037–2048 (1994).
3. Pulst, S. M. Genetic linkage analysis. *Arch. Neurol.* **56**, 667–672 (1999).
4. Altshuler, D., Daly, M. J. & Lander, E. S. Genetic mapping in human disease. *Science* **322**, 881–888 (2008).
5. Bailey-Wilson, J. E. & Wilson, A. F. Linkage analysis in the next-generation sequencing era. *Hum. Hered.* **72**, 228–236 (2011).
6. Terwilliger, J. D. & Ott, J. *Handbook of Human Genetic Linkage* (Johns Hopkins Univ. Press, 1994).
7. Ott, J. *Analysis of Human Genetic Linkage* (Johns Hopkins Univ. Press, 1999).
8. Lange, K. *Mathematical and Statistical Methods for Genetic Analysis* (Springer, 2002).
9. Mendel, G. J. Versuche über Pflanzen-Hybriden. *Verh. Naturforsch. Ver. Brünn* **4**, 3–47 (in German) (1866).
10. Santos-Cortez, R. L. *et al.* Mutations in *KARS*, encoding lysyl-tRNA synthetase, cause autosomal-recessive nonsyndromic hearing impairment DFNB89. *Am. J. Hum. Genet.* **93**, 132–140 (2013).
11. Goldschmidt, R. Gen und Ausseneigenschaft (Untersuchungen an *Drosophila*) I. *Z. Indukt Abstamm Vererbungsl* **69**, 38–69 (in German) (1935).
12. Goldschmidt, R. B. Phenocopies. *Sci. Am.* **181**, 46–49 (1949).
13. Laird, N. M. & Lange, C. Family-based designs in the age of large-scale gene-association studies. *Nature Rev. Genet.* **7**, 385–394 (2006).
14. Laird, N. M. & Lange, C. Family-based methods for linkage and association analysis. *Adv. Genet.* **60**, 219–252 (2008).
15. Ott, J., Kamatani, Y. & Lathrop, M. Family-based designs for genome-wide association studies. *Nature Rev. Genet.* **12**, 465–474 (2011).
16. Weiss, K. M. & Clark, A. G. Linkage disequilibrium and the mapping of complex human traits. *Trends Genet.* **18**, 19–24 (2002).
17. Ott, J. & Wang, J. Multiple phenotypes in genome-wide genetic mapping studies. *Protein Cell* **2**, 519–522 (2011).
18. Sasieni, P. D. From genotypes to genes: doubling the sample size. *Biometrics* **53**, 1253–1261 (1997).
A clear description of how population substructure leads to deviation from Hardy–Weinberg equilibrium and, consequently, to false-positive evidence of allelic association.
19. Falk, C. T. & Rubinstein, P. Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations. *Ann. Hum. Genet.* **51**, 227–233 (1987).
20. Terwilliger, J. D. & Ott, J. A haplotype-based 'haplotype relative risk' approach to detecting allelic associations. *Hum. Hered.* **42**, 337–346 (1992).
21. Ott, J. Statistical properties of the haplotype relative risk. *Genet. Epidemiol.* **6**, 127–130 (1989).
22. Spielman, R. S., McGinnis, R. E. & Ewens, W. J. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am. J. Hum. Genet.* **52**, 506–516 (1993).
The derivation of the highly successful TDT as a test for linkage and association.
23. He, Z. *et al.* Rare-variant extensions of the transmission disequilibrium test: application to autism exome sequence data. *Am. J. Hum. Genet.* **94**, 33–46 (2014).
24. De, G., Yip, W. K., Ionita-Laza, I. & Laird, N. Rare variant analysis for family-based design. *PLoS ONE* **8**, e48495 (2013).
25. Li, B. & Leal, S. M. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet.* **83**, 311–321 (2008).
The first derivation of collapsing methods for rare variants, leading to what is now known as burden tests.
26. Lee, S., Abecasis, G. R., Boehnke, M. & Lin, X. Rare-variant association analysis: study designs and statistical tests. *Am. J. Hum. Genet.* **95**, 5–23 (2014).
27. MacArthur, D. G. *et al.* Guidelines for investigating causality of sequence variants in human disease. *Nature* **508**, 469–476 (2014).
28. Haldane, J. B. S. & Crew, F. A. E. Change of linkage in poultry with age. *Nature* **115**, 641 (1925).
29. Renwick, J. H. & Schulze, J. Male and female recombination fractions for the nail-patella:ABO linkage in man. *Ann. Hum. Genet.* **28**, 37992 (1965).
30. Elston, R. C., Lange, E. & Namboodiri, K. K. Age trends in human chiasma frequencies and recombination fractions. II. Method for analyzing recombination fractions and applications to the ABO:nail-patella linkage. *Am. J. Hum. Genet.* **28**, 69–76 (1976).
31. Tanzi, R. E. *et al.* A genetic linkage map of human chromosome 21: analysis of recombination as a function of sex and age. *Am. J. Hum. Genet.* **50**, 551–558 (1992).
32. Shi, Q. *et al.* Absence of age effect on meiotic recombination between human X and Y chromosomes. *Am. J. Hum. Genet.* **71**, 254–261 (2002).
33. Kostic, V. S. *et al.* Intrafamilial phenotypic and genetic heterogeneity of dystonia. *J. Neurol. Sci.* **250**, 92–96 (2006).
34. Gusella, J. F. *et al.* A polymorphic DNA marker genetically linked to Huntington's disease. *Nature* **306**, 234–238 (1983).
35. Lee, J. M. *et al.* CAG repeat expansion in Huntington disease determines age at onset in a fully dominant fashion. *Neurology* **78**, 690–695 (2012).
36. Ott, J. & Falk, C. T. Epistatic association and linkage analysis in human families. *Hum. Genet.* **62**, 296–300 (1982).
37. Ott, J. in *Genetic Approaches to Mental Disorders* (eds Gershon, E. S. & Cloninger, C. R.) 63–75 (American Psychiatric Press, 1994).
38. Renwick, J. H. & Schulze, J. A computer program for the processing of linkage data from large pedigrees. *Excerpta Med. Int. Congr Ser.* **32**, E145 (1961).
39. Elston, R. C. & Stewart, J. A general model for the genetic analysis of pedigree data. *Hum. Hered.* **21**, 523–542 (1971).
A recursive method of likelihood calculation in large pedigrees, now known as the Elston–Stewart algorithm. It formed the basis for modern linkage analysis.
40. Elston, R. C., George, V. T. & Severtson, F. The Elston–Stewart algorithm for continuous genotypes and environmental factors. *Hum. Hered.* **42**, 16–27 (1992).
41. Ott, J. Estimation of the recombination fraction in human pedigrees: efficient computation of the likelihood for human linkage studies. *Am. J. Hum. Genet.* **26**, 588–597 (1974).
The first generally available linkage program for large pedigrees, LIPED.
42. Cheung, C. Y., Marchani Blue, E. & Wijsman, E. M. A statistical framework to guide sequencing choices in pedigrees. *Am. J. Hum. Genet.* **94**, 257–267 (2014).
43. Lander, E. S. & Green, P. Construction of multilocus genetic linkage maps in humans. *Proc. Natl Acad. Sci. USA* **84**, 2363–2367 (1987).
44. Lathrop, G. M., Lalouel, J. M., Julier, C. & Ott, J. Strategies for multilocus linkage analysis in humans. *Proc. Natl Acad. Sci. USA* **81**, 3443–3446 (1984).
45. Cottingham, R. W. Jr., Idury, R. M. & Schaffer, A. A. Faster sequential genetic linkage computations. *Am. J. Hum. Genet.* **53**, 252–263 (1993).
46. Kruglyak, L., Daly, M. J., Reeve-Daly, M. P. & Lander, E. S. Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am. J. Hum. Genet.* **58**, 1347–1363 (1996).
47. Abecasis, G. R., Cherny, S. S., Cookson, W. O. & Cardon, L. R. MERLIN — rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genet.* **30**, 97–101 (2002).
48. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
49. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
50. Wang, G. T., Peng, B. & Leal, S. M. Variant association tools for quality control and analysis of large-scale sequence and genotyping array data. *Am. J. Hum. Genet.* **94**, 770–783 (2014).
51. Wang, G. T., Zhang, D., Li, B., Dai, H. & Leal, S. M. Collapsed haplotype pattern method for linkage analysis of next generation sequence data. *Eur. J. Hum. Genet.* (in the press).
52. Thomas, D. C. & Cortessis, V. A. Gibbs sampling approach to linkage analysis. *Hum. Hered.* **42**, 63–76 (1992).
53. Heath, S. C. Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. *Am. J. Hum. Genet.* **61**, 748–760 (1997).
54. Sobel, E., Sengul, H. & Weeks, D. E. Multipoint estimation of identity-by-descent probabilities at arbitrary positions among marker loci on general pedigrees. *Hum. Hered.* **52**, 121–131 (2001).
55. Penrose, L. S. The detection of autosomal linkage in data which consist of pairs of brothers and sisters of unspecified parentage. *Ann. Eugen.* **6**, 133–138 (1935).
56. Knapp, M., Seuchter, S. A. & Baur, M. P. Two-locus disease models with two marker loci: the power of affected-sib-pair tests. *Am. J. Hum. Genet.* **55**, 1030–1041 (1994).
57. Whittemore, A. S. & Halpern, J. A class of tests for linkage using affected pedigree members. *Biometrics* **50**, 118–127 (1994).
58. Basu, S., Stephens, M., Pankow, J. S. & Thompson, E. A. A likelihood-based trait-model-free approach for linkage detection of binary trait. *Biometrics* **66**, 205–213 (2010).
59. Knapp, M., Seuchter, S. A. & Baur, M. P. Linkage analysis in nuclear families. 2: relationship between affected sib-pair tests and lod score analysis. *Hum. Hered.* **44**, 44–51 (1994).
60. Su, M. & Thompson, E. A. Computationally efficient multipoint linkage analysis on extended pedigrees for trait models with two contributing major loci. *Genet. Epidemiol.* **36**, 602–611 (2012).
61. Dietter, J. *et al.* Efficient two-trait-locus linkage analysis through program optimization and parallelization: application to hypercholesterolemia. *Eur. J. Hum. Genet.* **12**, 542–550 (2004).
62. Schaffer, A. A. Digenic inheritance in medical genetics. *J. Med. Genet.* **50**, 641–652 (2013).
63. Schork, N. J., Boehnke, M., Terwilliger, J. D. & Ott, J. Two-trait-locus linkage analysis: a powerful strategy for mapping complex genetic traits. *Am. J. Hum. Genet.* **53**, 1127–1136 (1993).
64. Sham, P. C., MacLean, C. J. & Kendler, K. S. Two-locus versus one-locus LODs for complex traits. *Am. J. Hum. Genet.* **55**, 855–858 (1994).
65. Smith, C. A. B. The detection of linkage in human genetics. *J. R. Statist. Soc. Series B (Methodol.)* **15**, 153–192 (1953).
66. Powell, J. E., Visscher, P. M. & Goddard, M. E. Reconciling the analysis of IBD and IBS in complex trait studies. *Nature Rev. Genet.* **11**, 800–805 (2010).
67. Kamphans, T. *et al.* Filtering for compound heterozygous sequence variants in non-consanguineous pedigrees. *PLoS ONE* **8**, e70151 (2013).
68. Dubay, C. *et al.* Genetic determinants of diastolic and pulse pressure map to different loci in Lyon hypertensive rats. *Nature Genet.* **3**, 354–357 (1993).
69. Hasstedt, S. J., Hanis, C. L. & Elbein, S. C. Univariate and bivariate linkage analysis identifies pleiotropic loci underlying lipid levels and type 2 diabetes risk. *Ann. Hum. Genet.* **74**, 308–315 (2010).

© 2015 Macmillan Publishers Limited. All rights reserved

70. Amos, C. I. *et al.* An approach to the multivariate analysis of high-density-lipoprotein cholesterol in a large kindred: the Bogalusa Heart Study. *Genet. Epidemiol.* **3**, 255–267 (1986).
71. Allison, D. B. *et al.* Multiple phenotype modeling in gene-mapping studies of quantitative traits: power advantages. *Am. J. Hum. Genet.* **63**, 1190–1201 (1998).
72. Ott, J. & Rabinowitz, D. A principal-components approach based on heritability for combining phenotype information. *Hum. Hered.* **49**, 106–111 (1999).
73. Suo, C. *et al.* Analysis of multiple phenotypes in genome-wide genetic mapping studies. *BMC Bioinformatics* **14**, 151 (2013).
74. Doyle, A. E. *et al.* Multivariate genomewide linkage scan of neurocognitive traits and ADHD symptoms: suggestive linkage to 3q13. *Am. J. Med. Genet. B. Neuropsychiatr. Genet.* **147B**, 1399–1411 (2008).
75. Houwen, R. H. *et al.* Genome screening by searching for shared segments: mapping a gene for benign recurrent intrahepatic cholestasis. *Nature Genet.* **8**, 380–386 (1994).
76. Smigielski, E. M., Sirotkin, K., Ward, M. & Sherry, S. T. dbSNP: a database of single nucleotide polymorphisms. *Nucleic Acids Res.* **28**, 352–355 (2000).
77. Tennessen, J. A. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64–69 (2012).
78. Genomes Project, C. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
79. Smith, K. R. *et al.* Reducing the exome search space for Mendelian diseases using genetic linkage analysis of exome genotypes. *Genome Biol.* **12**, R85 (2011).
80. Li, B. *et al.* A likelihood-based framework for variant calling and *de novo* mutation detection in families. *PLoS Genet.* **8**, e1002944 (2012).
81. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
A description of the widely used GATK tool for analysis of WGS data.
82. Bentley, D. *et al.* Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
83. Brzustowicz, L. M. *et al.* Molecular and statistical approaches to the detection and correction of errors in genotype databases. *Am. J. Hum. Genet.* **53**, 1137–1145 (1993).
84. Ott, J. Detecting marker inconsistencies in human gene mapping. *Hum. Hered.* **43**, 25–30 (1993).
85. Gordon, D., Leal, S. M., Heath, S. C. & Ott, J. An analytic solution to single nucleotide polymorphism error-detection rates in nuclear families: implications for study design. *Pac. Symp. Biocomput.* **2**, 663–674 (2000).
86. Cheung, C. Y., Thompson, E. A. & Wijsman, E. M. Detection of Mendelian consistent genotyping errors in pedigrees. *Genet. Epidemiol.* **38**, 291–299 (2014).
87. Neale, M. C., Neale, B. M. & Sullivan, P. F. Nonpaternity in linkage studies of extremely discordant sib pairs. *Am. J. Hum. Genet.* **70**, 526–529 (2002).
88. Hodge, S. E., Vieland, V. J. & Greenberg, D. A. HLODs remain powerful tools for detection of linkage in the presence of genetic heterogeneity. *Am. J. Hum. Genet.* **70**, 556–559 (2002).
89. Santos-Cortez, R. L. *et al.* Adenylate cyclase 1 (ADCY1) mutations cause recessive hearing impairment in humans and defects in hair cell function and hearing in zebrafish. *Hum. Mol. Genet.* **23**, 3289–3298 (2014).
90. Yan, J. *et al.* Combined linkage analysis and exome sequencing identifies novel genes for familial goiter. *J. Hum. Genet.* **58**, 366–377 (2013).
91. Louis-Dit-Picard, H. *et al. KLHL3* mutations cause familial hyperkalemic hypertension by impairing ion transport in the distal nephron. *Nature Genet.* **44**, 456–460 (2012).
92. Hoffmann, K. & Lindner, T. H. easyLINKAGE-Plus — automated linkage analyses using large-scale SNP data. *Bioinformatics* **21**, 3565–3567 (2005).
93. Lathrop, G. M., Lalouel, J. M., Julier, C. & Ott, J. Multilocus linkage analysis in humans: detection of linkage and estimation of recombination. *Am. J. Hum. Genet.* **37**, 482–498 (1985).
94. Heath, S. C., Snow, G. L., Thompson, E. A., Tseng, C. & Wijsman, E. M. MCMC segregation and linkage analysis. *Genet. Epidemiol.* **14**, 1011–1016 (1997).
95. Lange, K. *et al.* Mendel: the Swiss army knife of genetic analysis programs. *Bioinformatics* **29**, 1568–1570 (2013).
96. Lange, K., Weeks, D. & Boehnke, M. Programs for pedigree analysis: MENDEL, FISHER, and dGENE. *Genet. Epidemiol.* **5**, 471–472 (1988).
97. Schaffer, A. A., Lemire, M., Ott, J., Lathrop, G. M. & Weeks, D. E. Coordinated conditional simulation with SLINK and SUP of many markers linked or associated to a trait in large pedigrees. *Hum. Hered.* **71**, 126–134 (2011).
98. O'Connell, J. R. & Weeks, D. E. PedCheck: a program for identification of genotype incompatibilities in linkage analysis. *Am. J. Hum. Genet.* **63**, 259–266 (1998).
99. Gertz, E. M. *et al.* PSEUDOMARKER 2.0: efficient computation of likelihoods using NOMAD. *BMC Bioinformatics* **15**, 47 (2014).
100. Fishelson, M. & Geiger, D. Exact genetic linkage computations for general pedigrees. *Bioinformatics* **18**, S189–S198 (2002).
101. Lander, E. & Kruglyak, L. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nature Genet.* **11**, 241–247 (1995).
The derivation of the critical LOD score of 3.3 for a significance level of 0.05 in genome-scan linkage analysis.
102. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. in *Current Protocols in Human Genetics* (eds Haines, J. L. *et al.*) Ch. 7.20 (Wiley, 2013).
103. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20**, 110–121 (2010).

Acknowledgements
This work was supported by the Natural Science Foundation of China grant 31470070 (to J.O.) and the US National Institutes of Health grants R01 DC003594, R01 DC011651 and U54 HG006493 (to S.M.L.).

Competing interests statement
The authors declare no competing interests.

DATABASES
1000 Genomes: http://www.1000genomes.org/
dbSNP: http://www.ncbi.nlm.nih.gov/projects/SNP
ExAC: http://exac.broadinstitute.org/
Exome Variant Server: http://evs.gs.washington.edu/EVS/
Online Mendelian Inheritance in Man: http://omim.org/

FURTHER INFORMATION
easyLINKAGE: http://sourceforge.net/projects/easylinkage/
Family-Based Association Test Toolkit: http://www.hsph.harvard.edu/fbat/default.html
Genome Analysis Toolkit: https://www.broadinstitute.org/gatk/
GeneHunter: http://www.broadinstitute.org/ftp/distribution/software/genehunter/
GIGI-Check: https://faculty.washington.edu/wijsman/progdists/gigi/software/GIGI-Check/GIGI-Check.html
GIGI-Pick: https://faculty.washington.edu/wijsman/progdists/gigi/software/GIGI-Pick/GIGI-Pick.html
Handbook of Human Genetic Linkage: http://www.jurgott.org/linkage/LinkageHandbook.pdf
Homozygosity Mapper: http://www.homozygositymapper.org/
LINKAGE: http://www.jurgott.org/linkage/LinkagePC.html
FASTLINK: http://www.ncbi.nlm.nih.gov/CBBresearch/Schaffer/fastlink.html
Loki: http://www.stat.washington.edu/thompson/Genepi/Loki.shtml
Mendel: http://www.genetics.ucla.edu/software/mendel
MERLIN: http://www.sph.umich.edu/csg/abecasis/Merlin/
MSIM: http://watson.hgen.pitt.edu/docs/SLink.html
PedCheck: http://watson.hgen.pitt.edu/register/docs/pedcheck.html
PLINK: http://pngu.mgh.harvard.edu/~purcell/plink/
PolyPhen-2: http://genetics.bwh.harvard.edu/pph2/
PhyloP: http://ccg.vital-it.ch/mga/hg19/phylop/phylop.html
Pseudomarker: http://www.helsinki.fi/~tsjuntun/pseudomarker/
RV-TDT: http://bioinformatics.org/rv-tdt/
SEQLinkage: http://bioinformatics.org/seqlink
SimWalk2: http://www.genetics.ucla.edu/software/
SLINK: http://watson.hgen.pitt.edu/docs/SLink.html
Superlink: http://bioinfo.cs.technion.ac.il/superlink/
TLINKAGE: http://www.jurgott.org/linkage/tlinkage.htm
Variant Association Tools: http://varianttools.sourceforge.net/VAT
VCFtools: http://vcftools.sourceforge.net/

ALL LINKS ARE ACTIVE IN THE ONLINE PDF

© 2015 Macmillan Publishers Limited. All rights reserved