European Journal of Human Genetics (2008) 16, 854–860

www.nature.com/ejhg

ARTICLE

An approach for cutting large and complex pedigrees for linkage analysis

Fan Liu[1], Anatoliy Kirichenko[2], Tatiana I Axenovich[2], Cornelia M van Duijn[1] and Yurii S Aulchenko*[,1]

[1]*Department of Epidemiology & Biostatistics, Genetic Epidemiology unit, Erasmus MC, Rotterdam, The Netherlands;* [2]*Siberian Division, Institute of Cytology and Genetics, Russian Academy of Sciences, Novosibirsk, Russia*

Utilizing large pedigrees in linkage analysis is a computationally challenging task. The pedigree size limits applicability of the Lander–Green–Kruglyak algorithm for linkage analysis. A common solution is to split large pedigrees into smaller computable subunits. We present a pedigree-splitting method that, within a user supplied bit-size limit, identifies subpedigrees having the maximal number of subjects of interest (eg patients) who share a common ancestor. We compare our method with the maximum clique partitioning method using a large and complex human pedigree consisting of 50 patients with Alzheimer's disease ascertained from genetically isolated Dutch population. We show that under a bit-size limit our method can assign more patients to subpedigrees than the clique partitioning method, particularly when splitting deep pedigrees where the subjects of interest are scattered in recent generations and are relatively distantly related via multiple genealogic connections. Our pedigree-splitting algorithm and associated software can facilitate genome-wide linkage scans searching for rare mutations in large pedigrees coming from genetically isolated populations. The software package PedCut implementing our approach is available at http://mga.bionet.nsc.ru/soft/index.html.
European Journal of Human Genetics (2008) **16,** 854–860; doi:10.1038/ejhg.2008.24; published online 27 February 2008

Keywords: large complex pedigrees; splitting; Lander–Green–Kruglyak algorithm; linkage analysis; genetically isolated population

Introduction

Parametric linkage analysis searching for a unique founder mutation is a powerful tool to identify rare genetic variants with large effects. Genetically isolated populations provide extended pedigrees for linkage analysis as evidenced by numerous successes for complex traits, including type 2 diabetes[1] and Alzheimer's disease.[2] In such populations, pedigrees can be reconstructed based on genealogical records resulting in deep pedigrees that include a large number of multiple lines of descent. However, utilizing such large pedigrees in linkage analysis is computationally challenging.

For exact multipoint linkage analysis, several software packages implementing Lander–Green–Kruglyak algorithm,[3,4] such as Genehunter,[3] Merlin,[5] and Allegro,[6] are frequently used. The computational complexity of this algorithm increases linearly with the number of markers, but exponentially with the bit-size of the pedigree. The bit-size is defined as two times the number of individuals with parents presented in the pedigree minus the number of pedigree founders.[3] As long as the pedigree bit-size is small, the Lander–Green–Kruglyak algorithm can analyse a large number of markers. In modern implementations,[5] the time to compute multipoint LOD score using the Lander–Green–Kruglyak algorithm also depends on the fraction

*Correspondence: Dr YS Aulchenko, Department of Epidemiology & Biostatistics, Erasmus MC Rotterdam, Postbus 2040, 3000 CA Rotterdam, The Netherlands.
Tel: +31 10 408 7362; Fax: +31 10 408 9382;
E-mail: i.aoultchenko@erasmusmc.nl
Received 9 July 2007; revised 11 December 2007; accepted 17 January 2008; published online 27 February 2008

of pedigree members with missing genotypic information, which may be large for deep pedigrees because phenotypes and genotypes are usually unknown from the upper generations. Therefore, exact multipoint calculations are limited to pedigrees of several dozens of bits. Programs using Markov-chain Monte Carlo (MCMC) algorithms, such as Simwalk2,[7] Loki,[8] and Morgan[9] can analyse larger pedigrees but still fall short in the context of large pedigrees with hundreds of loops, especially for genome screens with large number of densely spaced markers.[10,11]

A common solution to reduce the computational burden is to split a large pedigree into smaller, and thus computable, subunits. Analyses of Hutterite pedigrees have revealed that a substantial amount of linkage information may be lost when truncating the pedigree for linkage analysis based on recent generations.[12] Manual splitting by an expert is only possible for relatively small pedigrees. For cutting large pedigrees, a semi-automatic method has been proposed based on factor analysis.[13] This method relies partly on the expert decisions and often yields subpedigrees that are still too complex for the Lander–Green–Kruglyak algorithm-based linkage analysis. Falchi *et al*[14] suggested a pedigree-splitting method based on graph theory maximum-cliques partitioning algorithm. This algorithm, although automatic, requires a number of parameters to be prespecified, for example a maximum number of generations, range for the measure of relatedness used to group individuals, and the range of the number of subjects of interest in a subpedigree. The optimization of these parameters largely depends on examining different sets of resultant subpedigrees. Furthermore, the available software implementing this algorithm does not guarantee that all of the resultant subpedigrees fall within specific bit-size limit and thus can be efficiently analysed by parametric linkage analysis using the Lander–Green–Kruglyak algorithm.

In this work, we develop a fast automatic algorithm for splitting large pedigrees based on user-specified maximum bit-size restriction. The algorithm specifically aims to split deep pedigrees where patients are relatively remotely related through genealogic connections and to produce an optimal set of subpedigrees for parametric linkage analysis of binary traits under rare dominant mutation model.

Pedigree splitting algorithm

In pedigree splitting we focus on the family members who have known genotypes and/or phenotypes. These people are denoted as 'subjects of interest' (SOI). We assume that some SOI share the genetic variant explaining their phenotype identical by descent from their common ancestor(s). The aim of our heuristic algorithm is to split a large pedigree into subpedigrees containing a maximal number of SOI who are related to a common ancestor and the bit sizes of the resultant pedigrees should be smaller than or equal to that specified by the user. This aim can be theoretically achieved by evaluating all SOI subgroups in terms of bit-sizes of pedigrees relating them to a common ancestor. The number of subgroups to evaluate, however, grows exponentially with the number of SOI studied, and becomes prohibitively large with more than 20–30 SOI. The number of subgroups to evaluate, therefore need to be reduced.

In our algorithm the kinship coefficient, ϕ_{ij}, is used to measure the degree of relatedness between SOI. The kinship coefficient is defined as the expected probability that two alleles randomly sampled from a pair of relatives i and j are copies of the same ancestral allele (identical by descent). For example the kinship coefficient is 1/4 if i and j are first-degree relatives (eg siblings) and 1/8 if they are second-degree relatives (eg uncle–niece). In our work, the coefficients were calculated using a modified version of the PEDIG software developed by Didier Boichard.[15]

In the first step, the algorithm constructs a matrix of subgroups that are sorted based on kinship. Given a group of SOI of size N, consider individual $m \in (1, N)$. The set of relatives of m is sorted in decreasing order according to their kinship to m, so as m is the first element of this set. Let us call this set as R_m. Let $R_m(f)$ be a set of the first f elements of R_m, $f \in (2, N)$. When $f=2$, $R_m(2)$ includes m and his closest relative. When $f=3$, the next closest relative of m *is* included into $R_m(2)$, so that $R_m(1) \subset R_m(2) \subset \ldots R_m(N)$.

Iterating the central individual m over all SOI gives an N by N-*1* matrix,

$$\begin{bmatrix} R_1(2) & R_1(3) & \cdots & R_1(f) & \cdots & R_1(N) \\ R_2(2) & R_2(3) & \cdots & R_2(f) & \cdots & R_2(N) \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ R_m(2) & R_m(3) & \cdots & R_m(f) & \cdots & R_m(N) \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ R_N(2) & R_N(3) & \cdots & R_N(f) & \cdots & R_N(N) \end{bmatrix}$$

where each row represents the subgroups of SOI derived from the same central individual but having different sizes (2 to N). Each column represents the subgroups of the same size, which are derived from each individual, who is considered as the center of the corresponding subgroup. Identify unique groups of relatives by iterating over matrix R and give $R_m(f)$ a binary index of 1 when the content of $R_m(f)$ is seen for the first time, otherwise assign zero.

At the second step of our algorithm, the genealogies connecting the members of identified unique subgroups (as identified by index of 1) are constructed using the PedHunter program developed by Agarwala *et al*.[16] For each subgroup, a subpedigree linking the maximal number of subgroup members to the most recent common ancestor is reconstructed and the bit-size of every subpedigree is computed. For each row R_m, the construction of subpedigrees starts at $R_m(2)$ and stops when the bit-size of $R_m(f)$ violates the maximal bit-size limit. There is no need to

construct subpedigrees for $R_m(f+1)$ to $R_m(N)$ because the bit-size of $R_m(f+1)$ is always greater than that of $R_m(f)$. At this stage, all constructed subpedigrees satisfy the bit-size limit and contain a unique configuration of SOI. The subpedigree connecting the largest number of SOI, as heuristically the most interesting pedigree, is selected as the first subpedigree. When several non-overlapping subpedigrees are eligible all of them are selected. When several partly overlapping subpedigrees are eligible, the one with the smallest bit-size is selected. If still several subpedigrees are eligible, the one with the highest average kinship among SOI is selected. When, additionally, the average kinships are the same, a random subpedigree is selected and the alternative selections are saved in a log file. From our experience, the latter scenario is very rare for reasonably large bit-sizes and complex pedigrees with multiple lines of descent. Within the search space, this exhaustive algorithm guarantees that the selected subpedigree has the maximal number of SOI in respect to the bit-size restriction B.

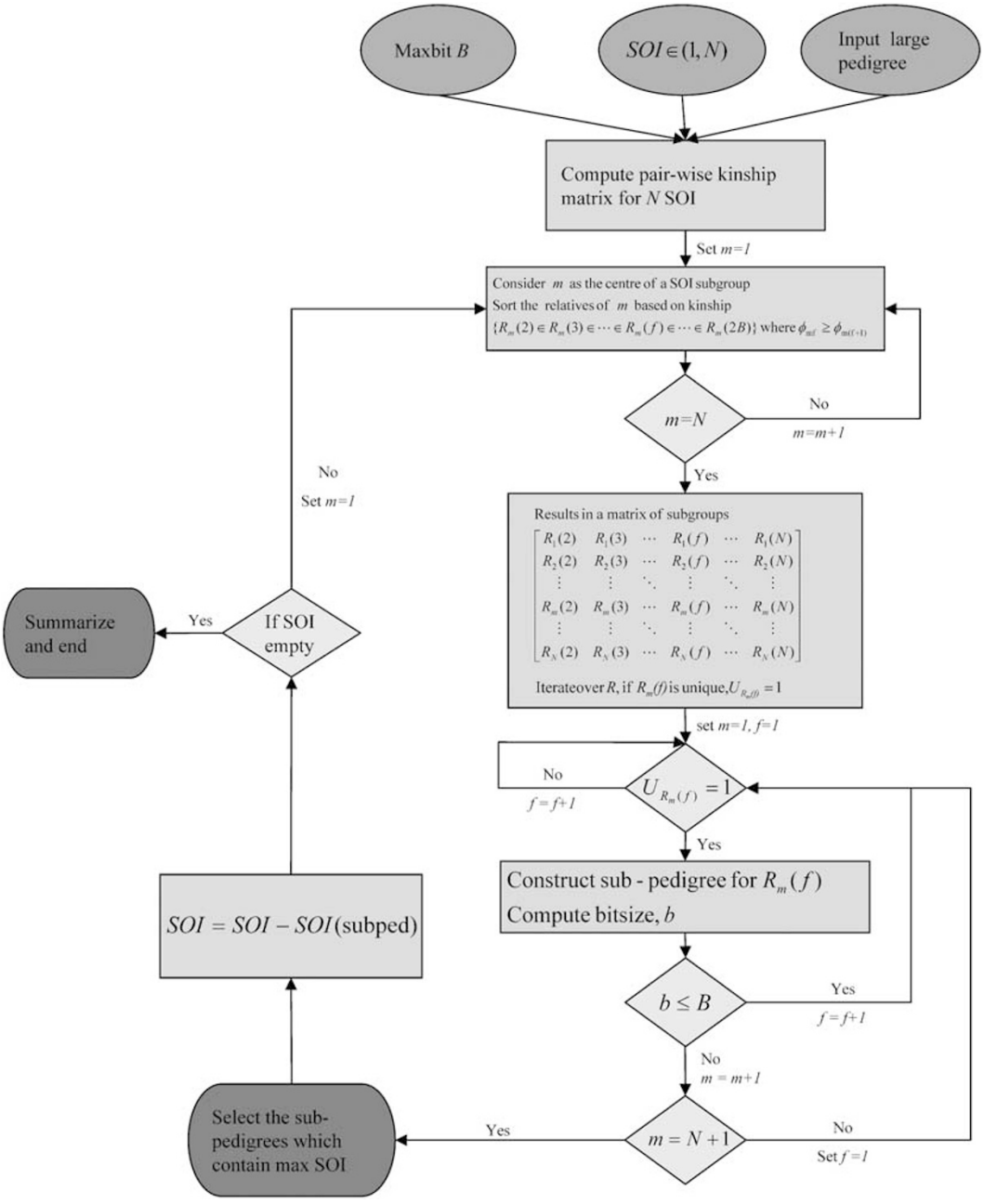

Figure 1 Flowchart of the pedigree splitting algorithm.

At the third step, the algorithm removes the SOI belonging to the identified pedigree(s) from further consideration and is recursively applied, starting with the Step 1, to the remaining SOI, until no further SOI can be assigned to a subpedigree.

The algorithm is presented in Figure 1. The described algorithm is implemented in a software package, PedCut, which is available at http://mga.bionet.nsc.ru/soft/index.html.

Splitting a large pedigree with PedCut and Greffa

We tested properties of our program PedCut using a large and complex pedigree and compared it with the Greffa program developed by Falchi *et al*.[14] This program implements the maximum clique-partitioning algorithm. The pedigree comprises a part of the genealogy used in a genome-wide screen for late onset Alzheimer's disease (AD) in genetically isolated Dutch population.[2] The original pedigree contains 103 AD patients and 4645 family members. For demonstration purpose, we used a fraction of the original pedigree that contains 50 randomly selected AD patients and their 2460 ancestors spanning over 18 generations (Table 1 and Figure 2). These 50 AD patients, who are considered as SOI, are scattered in the most recent generations. Except five sibling–pairs and two aunt–niece pairs, all SOI are related distantly via multiple genealogic connections (Table 1). Here a genealogic connection is defined as a unique genealogic pathway via which an allele identical by descent can be transmitted. For example a pair of siblings have two unique lines connecting them, one from the mother and another from the father whereas a pair of half sibs have only one line connecting them via the shared parent.

We used maximum bit-size restriction of 18, 24, and 30 bits to test PedCut and Greffa. To make the results from Greffa comparable with our results, we tried a number of combinations of the parameters required by Greffa and reported the results from the optimal set of parameters, which give subpedigrees with the maximum number of SOI within 18, 24 or 30 bits.

Table 2 shows that PedCut can assign more patients into subpedigrees compared to Greffa in splitting this pedigree. Under any bit-size limit investigated, PedCut could successfully assign most SOI to subpedigrees whereas Greffa left a considerable number of SOI unassigned. For example, Greffa left 12 (24%) to 28 (56%) SOI unassigned, and thus completely lost for linkage analysis. On the other hand, PedCut left at maximum one (2%) SOI unassigned. The only person who could not be assigned to a pedigree by PedCut at bit-size of 18 and 24 is connected to a closest SOI through 10 meioses; this person was assigned when bit-size limit was set to 30.

Also, the average number of SOI per subpedigree from PedCut was larger (2.9, 3.5, and 4.2 for 18, 24, and 30 bits) than that from Greffa (2.9, 2.4, and 2.5). Furthermore, the maximum number of SOI from Pedcut (6, 6, and 7 for 18, 24, and 30 bits) is larger than that from Greffa (5, 4, and 3). Finally, compared to Greffa, PedCut produced more uniformly sized subpedigrees, as evidenced by a narrower range of bit and pedigree-sizes (Table 2).

With an Intel Pentium 4 2.4 GHz CPU, PedCut could split the pedigree in about 3 min. The resultant subpedigrees from PedCut using 24 bits as the pedigree limit are depicted in Supplementary Figure 1, where the 14 subpedigrees are ordered in the same sequence as PedCut identified them. All five sib-pairs were captured by the first two subpedigrees and the two aunt–niece pairs were captured by subpedigrees 3 and 4. The pedigrees 5 and 6 also contain multiple SOI who are relatively closely related to each other. The subpedigree seven contains two SOI who are double-first cousins and the subpedigree 10 contains two SOI who are second cousins. The subpedigrees 8, 9, 11 and 12 contain multiple distantly related SOI. The pedigree 13 and 14 contains only two distantly related SOI each.

Power comparison

We compared the power to detect linkage to a rare fully penetrant variant using subpedigrees derived in the previous section. Using these pedigrees, we simulated genetic data conditional on the phenotypes observed, using method of Boehnke,[17] implemented in software package SIMLINK version 4.12. We assumed a genetically homogenous binary trait, determined by two underlying alleles (D being causal and d being normal allele). The penetrance of DD was fixed at 1.0 and penetrance of dd was

Table 1 Genealogic characteristics of the pedigree including 50 Alzheimer's disease patients

Characteristics	*Value ± SD (min–max)*
Number of Alzheimer's disease patients	50
Number of generations	18
Number of individuals	2510
Pedigree bit-size	2839
Average number of genealogic connections between a pair of patients	378.5 (0–2673)
Average number of meioses separating a pair of patients	18.9 ± 3.1 (2–22)
Number of SOI pairs with non-zero kinship	1214
Sum of pair-wise kinship coefficient	8.74598
Mean kinship coefficient	0.0072 ± 0.0198

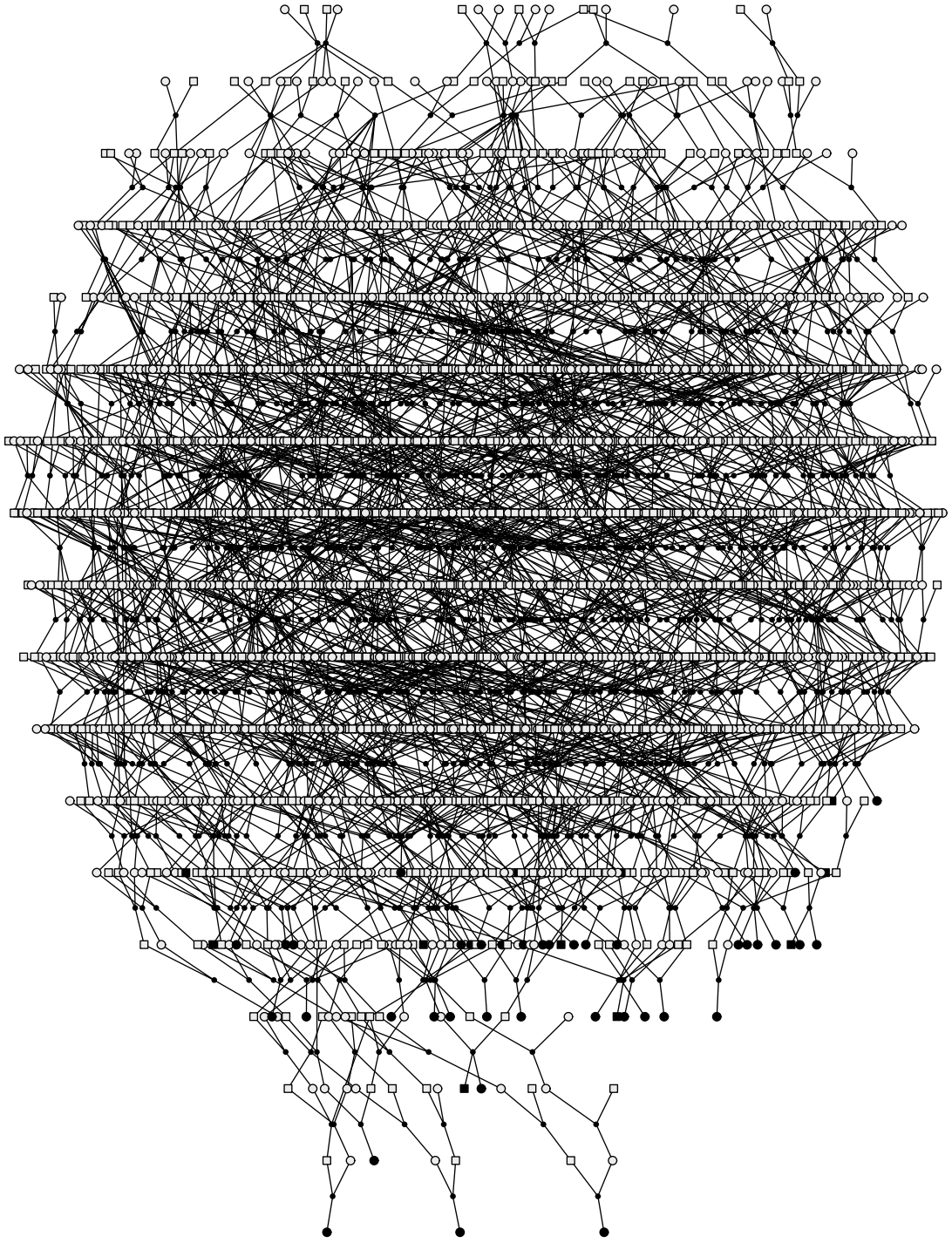

Figure 2 The initial pedigree consisting of 50 Alzheimer patients (blue: patients). This pedigree was drawn using software package Pedfiddler version 0.5 (http://www.medicine.mcgill.ca/statgene/software.html).

fixed at 0.0. The penetrance for Dd was set at 1.0 (dominant model), 0.75, 0.5, 0.25, or 0.0 (recessive model). The frequency of D allele was set in such manner, that locus-specific population attributable fraction was 0.0001. One marker locus having 5 alleles with equal frequency of 0.2 was modelled. One thousand replicas were simulated and analysed under each scenario concerning Dd penetrance. The distance between the trait and marker loci was set to zero. Consequent linkage analysis was performed using correct model.

Table 2 Result of a pedigree of 50 Alzheimer's disease patients using PedCut and Greffa under various bit-size restrictions

	Greffa			*PedCut*		
	18 bits	*24 bits*	*30 bits*	*18 bits*	*24 bits*	*30 bits*
Number of resultant subpedigrees	13	9	15	17	14	12
Pedigree bits (min–max)	9.4 (2–18)	9.2 (4–23)	15.1 (6–30)	12.0 (6–18)	16.1 (6–24)	23.2 (15–29)
Pedigree size (min–max)	16.2 (4–27)	15.1 (8–28)	23.5 (12–45)	20.1 (11–33)	25.2 (12–41)	37.3 (24–51)
Average number of SOI per subpedigree (min–max)	2.9 (2–5)	2.4 (2–4)	2.5 (2–3)	2.9 (2–6)	3.5 (2–6)	4.2 (3–7)
Average number of generations of subpedigrees (min–max)	4.4 (2–5)	4.2 (3–5)	5.2 (4–6)	4.6 (2–5)	5.7 (4–9)	6.1 (4–9)
Number of SOI who could not be assigned to any subpedigree	13	28	12	1	1	0
Number of SOI pairs with non-zero kinship	35	16	36	57	74	95
Sum of kinship derived from subpedigrees	1.836	0.777	1.030	2.173	2.133	2.236
Mean kinship derived from subpedigrees	0.052	0.049	0.029	0.038	0.029	0.024
Parameters used for the Greffa program[a]						
Minimum clique size	2	2	2			
Maximum clique size	5	6	3			
Maximum number of generations	7	7	9			
Minimum pairwise kinship	0.01	0.001	0.0067			

[a]An optimal set of parameters was chosen for Greffa that gives subpedigrees with the maximum number of SOI within 18, 24, or 30 bits.

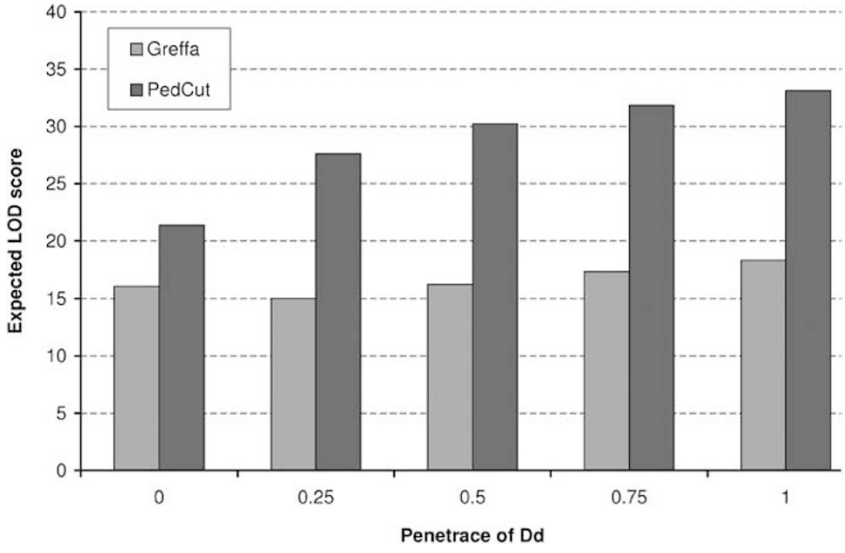

Figure 3 Expected LOD score for pedigrees derived using PedCut and Greffa with a bit-size limit of 18.

For 18-bits pedigrees, results of power calculation are shown in Figure 3. The subpedigrees derived from PedCut consistently showed higher power compared to the ones from Greffa, across all models analysed. This is most likely due to the fact that Greffa left a considerable amount of patients unassigned to any subpedigree. Of interest, the subpedigrees derived from PedCut showed a clear trend of increase in power when the underlying genetic model was becoming more dominant. The same trend exists for the pedigrees from Greffa but in much less degree. Similar results were obtained for 24- and 30-bit pedigrees as well (not shown).

Discussion

In this work we have developed an algorithm that recursively groups the subjects of interest (SOI, eg genotyped patients) into subgroups that fall within a certain bit-size limit and include the maximum number of SOI who share a common ancestor. With the algorithm we exploit the basic rationale of linkage analysis, that is that affected relatives are likely to share the disease-causal allele identical by descent from a common ancestor. Fast grouping is achieved by prioritizing relatives using kinship. Our algorithm guarantees that the derived subpedigrees can be directly and efficiently analysed by software implementing

the Lander–Green–Kruglyak algorithm. Further, it is fully automated.

Any method for subpedigree identification involves breaking relationships between SOI. Breaking relationships may increase false-positive linkage signals[18,19] or reduce the power of linkage analysis.[12] Bias is especially pronounced when close relationships are broken. In our program we include a default option to preserve close relationships between SOI, keeping all second cousin or closer relationships between SOI. In general, type 1 error for split pedigree should be worked out by gene-dropping using complete pedigree, and re-analysis using split pedigrees.[2]

Exhaustive search algorithm guarantees (within the search space) that the selected subpedigree has the maximal number of SOI in respect to the bit-size restriction, if no equally eligible subpedigrees are available. The latter situation, though possible, is in our experience unlikely in deep pedigrees coming from genetically isolated populations, where mating is mostly random, and multiples genealogic connections are observed between SOI.

We have previously applied PedCut to split a pedigree including 4645 people of which 112 were Alzheimer's disease patients.[2] In that study, we could assign 103 patients to 35 pedigrees using bit-size limit of 35 in about 11 min. Empirical threshold for 5% genome-wide significance, established using simulations in complete pedigree, was estimated to be 3.64. Regions of significant linkage were genotyped using dense single-nucleotide polymorphisms (SNPs) marker map in an independent cohort. Significant associations were observed between some of these regions and cognitive function. This proves applicability of our approach in real studies.

In summary, we developed a heuristic algorithm that is suitable to split large and complex pedigrees coming from genetically isolated populations. Our algorithm and associated software (PedCut, http://mga.bionet.nsc.ru/soft/index.html) can facilitate fast genome-wide linkage search for rare mutations.

Acknowledgements

We thank three anonymous reviewers for valuable comments. This work was supported by Netherlands Organization for Scientific Research (NWO, 91203014), the joint grant from the Netherlands Organization for Scientific Research and the Russian Foundation for Basic Research (NWO-RFBR), the Centre of Medical Systems Biology (CMSB), Hersenstichting Nederland, Internationale Stichting Alzheimer Onderzoek (ISAO), Alzheimer Association project number 04516, Hersenstichting Nederland project number 12F04(2).76, Interuniversity Attraction Poles (IUAP) program. YSA was supported by 'Stichting MS'.

References

1 Grant SF, Thorleifsson G, Reynisdottir I *et al*: Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. *Nat Genet* 2006; **38**: 320–323.
2 Liu F, Arias-Vásquez A, Sleegers K *et al*: A genome-wide screen for late onset Alzheimer's disease in a genetically isolated Dutch population. *Am J Hum Genet* 2007; **81**: 17–31.
3 Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES: Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am J Hum Genet* 1996; **58**: 1347–1363.
4 Lander ES, Green P: Construction of multilocus genetic linkage maps in humans. *Proc Natl Acad Sci USA* 1987; **84**: 2363–2367.
5 Abecasis GR, Cherny SS, Cookson WO, Cardon LR: Merlin – rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet* 2002; **30**: 97–101.
6 Gudbjartsson DF, Thorvaldsson T, Kong A, Gunnarsson G, Ingolfsdottir A: Allegro version 2. *Nat Genet* 2005; **37**: 1015–1016.
7 Sobel E, Lange K: Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. *Am J Hum Genet* 1996; **58**: 1323–1337.
8 Heath SC: Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. *Am J Hum Genet* 1997; **61**: 748–760.
9 Sung YJ, Thompson EA, Wijsman EM: MCMC-based linkage analysis for complex traits on general pedigrees: multipoint analysis with a two-locus model and a polygenic component. *Genet Epidemiol* 2007; **31**: 103–114.
10 Service S, Molina J, Deyoung J *et al*: Results of a SNP genome screen in a large Costa Rican pedigree segregating for severe bipolar disorder. *Am J Med Genet B Neuropsychiatr Genet* 2006; **141**: 367–373.
11 Sieh W, Basu S, Fu AQ *et al*: Comparison of marker types and map assumptions using Markov chain Monte Carlo-based linkage analysis of COGA data. *BMC Genet* 2005; **6** (Suppl 1): S11.
12 Dyer TD, Blangero J, Williams JT, Goring HH, Mahaney MC: The effect of pedigree complexity on quantitative trait linkage analysis. *Genet Epidemiol* 2001; **21** (Suppl 1): S236–S243.
13 Pankratz VS, Iturria SJ: A pedigree partitioning approach to quantitative trait loci mapping of IgE serum level in the GAW12 Hutterite data. *Genet Epidemiol* 2001; **21** (Suppl 1): S258–S263.
14 Falchi M, Forabosco P, Mocci E *et al*: A genomewide search using an original pairwise sampling approach for large genealogies identifies a new locus for total and low-density lipoprotein cholesterol in two genetically differentiated isolates of Sardinia. *Am J Hum Genet* 2004; **75**: 1015–1031.
15 Boichard D: PEDIG: a fortran package for pedigree analysis suited for large populations. *In Proceedings of the 7th World Congress on Genetics Applied to Livestock Production, Montpellier, 2002-08-19/23* 2002; **32**: 525–528.
16 Agarwala R, Biesecker LG, Hopkins KA, Francomano CA, Schaffer AA: Software for constructing and verifying pedigrees within large genealogies and an application to the Old Order Amish of Lancaster County. *Genome Res* 1998; **8**: 211–221.
17 Boehnke M: Estimating the power of a proposed linkage study: a practical computer simulation approach. *Am J Hum Genet* 1986; **39**: 513–527.
18 Liu F, Elefante S, van Duijn CM, Aulchenko YS: Ignoring distant genealogic loops leads to false-positives in homozygosity mapping. *Ann Hum Genet* 2006; **70**: 965–970.
19 Miano MG, Jacobson SG, Carothers A *et al*: Pitfalls in homozygosity mapping. *Am J Hum Genet* 2000; **67**: 1348–1351.

Supplementary Information accompanies the paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)