European Journal of Human Genetics (2004) 12, 542–550

www.nature.com/ejhg

ARTICLE

Efficient two-trait-locus linkage analysis through program optimization and parallelization: application to hypercholesterolemia

Johannes Dietter*[,1], Alexander Spiegel[2], Dieter an Mey[2], Hans-Joachim Pflug[2], Hussam Al-Kateb[3], Katrin Hoffmann[4], Thomas F Wienker[1] and Konstantin Strauch[1]

[1]*Institute for Medical Biometry, Informatics and Epidemiology, University of Bonn, Bonn, Germany;* [2]*Center for Computing and Communication, High Performance Computing, Aachen University, Aachen, Germany;* [3]*Institute of Molecular Genetics, Heidelberg University, Heidelberg, Germany;* [4]*Department of Nephrology, Medical University Clinic, Würzburg, Germany, Gene Mapping Center, Max Delbrück Center for Molecular Medicine, Berlin, Germany*

We have optimized and parallelized the GENEHUNTER-TWOLOCUS program that allows to perform linkage analysis with two trait loci in the multimarker context. The optimization of the serial program, before parallelization, results in a speedup of a factor of more than 10. The parallelization affects the two-locus-score calculation, which is predominant in terms of computation time. We obtain perfect speedup, that is, the computation time decreases exactly by a factor of the number of processors. In addition, two-locus LOD and NPL scores are now calculated for varying genetic positions of both disease loci, not just one locus varied and the position of the other disease locus fixed, as before. This results in easily interpretable 3-D plots. We have reanalyzed a pedigree with hypercholesterolemia using our new version of GENEHUNTER-TWOLOCUS. Whereas originally, two individuals had to be discarded due to excessive computation-time demands, the entire 17-bit pedigree could now be analyzed as a whole. We obtain a two-trait-locus LOD score of 5.49 under a multiplicative model, compared to LOD scores of 3.08 and 2.87 under a heterogeneity and additive model, respectively. This further increases evidence for linkage to both 1p36.1–p35 and 13q22–q32 regions, and corroborates the hypothesis that the two genes act in a multiplicative way on LDL cholesterol level. Furthermore, we compare the computation times for two-trait-locus analysis needed by the programs GENEHUNTER-TWOLOCUS, TLINKAGE, and SUPERLINK. Altogether, our algorithmic improvements of GENEHUNTER-TWOLOCUS allow researchers to analyze complex diseases under realistic two-trait-locus models with pedigrees of reasonable size and using many markers.

European Journal of Human Genetics (2004) **12,** 542–550. doi:10.1038/sj.ejhg.5201196
Published online 21 April 2004

Keywords: complex traits; two locus models; locus heterogeneity; epistasis; interaction

Introduction

Many and probably most human diseases are caused or substantially influenced by genetic variants. The concept of multifactorial causation implies the cooperation of variants of more than one gene and exogenous factors, such as lifestyle, nutrition, habits, exposition towards toxic substances, etc, in the etiology of frequent diseases. Among these diseases are mental disorders, diabetes,

***Correspondence: Dr J Dietter, Institute for Medical Biometry, Informatics and Epidemiology, University of Bonn, Sigmund-Freud-Str. 25, 53105 Bonn, Germany. Tel: +49 228 287 5243; Fax: +49 228 287 5854; E-mail: dietter@imsdd.meb.uni-bonn.de**
Received 3 September 2003; revised 9 February 2004; accepted 2 March 2004

multiple sclerosis, asthma, metabolic disorders, and atopy. A first attempt to model these diseases more realistically in a linkage analysis is to analyze them with two trait loci.

There has already been extensive research and methodological development for two-trait-locus linkage analysis.[1] The methods can be roughly divided into parametric (LOD-score) and nonparametric (NPL) analysis. NPL analysis delivers results without relying on specific model assumptions, while with the LOD-score method a disease model has to be specified. Even though researchers may find it convenient not to worry about the disease model when using NPL analysis, the LOD-score method offers the possibility to gain information about the underlying biological mechanism by testing different disease models. This is especially helpful in the context of two trait loci, where it is of interest how the two loci act or interact. Throughout this paper, we will assume that the trait of interest is dichotomous.

Previously, our group has extended the GENEHUNTER software,[2,3] which performs parametric and nonparametric multipoint linkage analysis, to GENEHUNTER-TWOLOCUS.[1] This program is capable to calculate LOD and NPL scores with two trait loci. It has already been successfully applied in several projects, for example, house dust mite allergy,[1] autosomal recessive familial hypercholesterolemia,[4] and asthma.[5] However, due to otherwise excessive computation-time demands, the two-locus calculations as implemented in GENEHUNTER-TWOLOCUS are only feasible for moderate pedigree sizes. The number of effective meioses ($2 \times$ nonfounders$-$founders) should be restricted to 12–13. This may put a severe limit to some studies. In order to be able to calculate LOD scores for larger pedigrees, we adopted a two-fold strategy. At first, those parts of the GENEHUNTER-TWOLOCUS program have been thoroughly optimized, which, in consequence to the two-locus extension, are responsible for most of the computation time. This will be beneficial already on a single processor workstation. In addition, we have parallelized the optimized program parts, because even after optimization, these program parts still consume by far most of the computation time. The two-locus extension is ideally suited for parallelization in the context of the Lander–Green algorithm.[6,7] This allows the program to run in parallel on local Linux clusters at research institutes, but it is also possible to make use of massively parallel supercomputers that offer outstanding computational power.

The possibility to parallelize the code of linkage programs to make use of workstation clusters or parallel computers has been exploited before. Principally, there are two algorithmical approaches tackling the computational demands associated with linkage analysis.

The Elston–Stewart algorithm[8] allows to analyze large pedigrees. But due to the exponential increase in computation time and memory requirements with the number of loci, it can handle only three or four multiallelic markers. The Lander–Green algorithm,[6,7] on the other hand, is able to cope with a multitude of markers but only with pedigrees of moderate size, because computation time and memory requirements increase exponentially with the number of meioses in the pedigree. Programs based on both of these algorithms have been parallelized before. Kothari *et al*[9] parallelized LINKMAP of the LINKAGE package[10,11] as an example for an Elston–Stewart-based program. The CRI-MAP program,[6] which utilizes the Lander–Green algorithm,[6,7] has been parallelized by Matise *et al*.[12] Please see also the references in these papers. GENEHUNTER, which is also based on the Lander–Green algorithm, has been parallelized by Conant *et al*.[13]

The work mentioned so far is somewhat complementary to our work described here, as the authors of the aforementioned studies parallelized the 'standard' versions of the corresponding programs that solely deal with the single trait locus. Our goal for this work was to reduce the additional computation time of the GENEHUNTER-TWOLOCUS program that is caused by the two-trait-locus extension. In order to do so, we have optimized and parallelized the corresponding program parts.

Methods

Algorithmic procedure within GENEHUNTER-TWOLOCUS

GENEHUNTER, which is written in the programming language C, splits the calculation for linkage analysis in two parts:[2]

- Extraction of information about the inheritance pattern in a pedigree that depends only on the markers.
- Definition of a statistic or score to assess linkage, for a given inheritance pattern, which depends only on the trait information on all pedigree members.

According to this idea, one defines a scoring function $S(w,\phi)$ that quantifies (ie, scores) to what degree the inheritance vector w indicates the presence of a disease gene at a given position, in consideration of the trait phenotypes ϕ. The inheritance vector w specifies for every meiosis whether the paternally or maternally inherited allele has been transmitted (bit 0 or 1, respectively). In general, several inheritance vectors are compatible with the information supplied by the marker data. Therefore, the probability distribution $P(v(x)) = w)$ has to be evaluated over the set V of all possible inheritance vectors. Here, $v(x)$ denotes the inheritance vector at genetic position x of the putative disease locus relative to the marker group used. These ideas can be formalized by introducing the averaged scoring function $\bar{S}(x, \phi)$ (see Kruglyak *et al*[2]):

$$\bar{S}(x,\phi) = \sum_{w \in V} S(w,\phi) P(v(x) = w) \quad (1)$$

This formulation is valid for one disease locus. To be able to handle two disease loci with genetic positions x_1 and x_2, respectively, $\bar{S}$ can be written as follows:[1]

$$\bar{S}(x_1,x_2,\phi) = \sum_{w_1,w_2 \in V} S(w_1,w_2,\phi) P(v(x_1) = w_1, v(x_2) = w_2)$$

$P(v(x_1) = w_1, v(x_2) = w_2)$ is the probability that the inheritance vector equals w_1 at disease gene location x_1 and w_2 at disease gene location x_2. $S(w_1,w_2,\phi)$ is the two-locus extension of $S(w,\phi)$ in (1). It rates the compatibility of the inheritance vector tuple w_1,w_2 and trait phenotypes ϕ. If one assumes two unlinked marker groups (eg on two different chromosomes) and two trait loci, one linked to each marker group, then $P(v(x_1) = w_1, v(x_2) = w_2)$ factorizes into $P(v(x_1) = w_1)P(v(x_2) = w_2)$, since the inheritance vectors at the two locations are independent. In this case, when taking the expectation of $S(w_1,w_2,\phi)$ over the inheritance distributions at both loci, we obtain

$$\bar{S}(x_1,x_2,\phi) = \sum_{w_1,w_2 \in V} S(w_1,w_2,\phi) P(v(x_1) = w_1) P(v(x_2) = w_2) \quad (2)$$

Like the single-trait-locus formulation, the two-trait-locus formulation given in formula (2) allows to integrate both parametric and NPL analysis within the same framework of the Lander–Green algorithm.[1] With parametric (LOD-score) analysis, $S(w_1,w_2,\phi)$ equals the likelihood ratio for the two trait loci. With NPL analysis, two-locus extensions of the scoring functions S_{pairs} and S_{all} are used for $S(w_1,w_2,\phi)$. They evaluate sharing of alleles identical-by-descent for affected individuals, simultaneously at both disease loci. By looking at formula (2), one can tell as to which part of the calculation will cause most of the additional computational effort. The cost of computing the inheritance distributions $P(v(x_i) = w_i)$ is only doubled, since it now has to be determined for the first and the second disease locus. But $S(w_1,w_2,\phi)$ becomes a matrix with two-trait-locus analysis and has to be calculated for all w_1 and $w_2 \in V$. Thus, while in the single-locus version the scoring function has to be calculated only N times (with N being the number of possible inheritance vectors), it now has to be evaluated N^2 times. It can be shown that $N = 2^{2n-f}$, where n is the number of nonfounders and f is the number of founders in the pedigree. To be more specific, with a single-trait-locus analysis, if one nonfounder is added, the computational cost is quadrupled (2×2), since this person adds two meioses to the pedigree. This effect is much more dramatic with a two-trait-locus analysis: if one person is added, both dimensions of the $S(w_1,w_2,\phi)$ matrix are quadrupled, and thus, computing time increases by a factor of $4^2 = 16$.

Generally, in the context of the Lander–Green algorithm, $2n-f$ is the number of bits of the inheritance vector (ie, effective meioses); it determines the computation time and memory requirements of a pedigree. With the original GENEHUNTER-TWOLOCUS version, the number of bits of a pedigree should not exceed 12 or 13. Evidently this rather small pedigree size puts a severe limit on two-trait-locus studies. Therefore, in order to reduce the computational effort, one has to optimize those program parts that contribute to $S(w_1,w_2,\phi)$. This will be described in the next section.

Optimization of time-critical program parts

In order to improve the performance of the serial code – and as a consequence also the performance of the parallel version as well – we at first selected a test data set, which allowed executing many program runs in a limited time frame. This test data set only dealt with nine effective meioses and thus lead to a very short run time, but it was expected to reveal run time behavior similar to our target data set with 17 meioses. Repeatedly, during these test runs, performance information was collected and evaluated, and the program code was then modified to cut down the run time. For this purpose, we used the performance analyzer toolset, an important part of Sun Microsystem's programming environment (Sun Microsystems Forte Developer 7 Program Performance Analysis Tools http://docs.sun.com/db/doc/816-2458). By this means those portions of the code that are very expensive by not utilizing the hardware in a favorable manner can easily be identified. With this toolset it was easy to find out that the original program was spending more than 99% of its run time in two functions (peel and brute_force_analyze) which are consuming less than 1% of the program code. This is, of course, a very lucky case, because changing only a limited code portion can have a major effect on performance. There are some well-known techniques of program optimization that could be applied. In many cases, this will be done automatically by an optimizing compiler. But then there are cases where the code is too complicated or where not enough information is available at compile time. Here, we did the following manual code changes to GENEHUNTER-TWOLOCUS in order to improve the performance:

- Extraction of loop-invariant code.
- Replacement of case constructs by bit manipulations.
- Loop interchange to improve loop unrolling.
- Subroutine inlining.

These modifications of only a small part of the serial code of GENEHUNTER-TWOLOCUS lead to a considerable reduction of the runtime of the whole program without the cost of additional hardware. More details on these optimization steps can be found in the supplementary electronic information.

Parallelization of the two-locus extension

The computation time can be further reduced by parallelization of those parts, which still remain time critical. It turns out that the two-locus extension is particularly well suited for that. This becomes apparent if one rewrites (2) as

$$\overline{S}(x_1, x_2, \phi) = \sum_{w_2 \in V} \left[\sum_{w_1 \in V} S(w_1, w_2, \phi) P(v(x_1) = w_1) \right] \cdot P(v(x_2) = w_2) = \sum_{w_2 \in V} S'(w_2, \phi)_{x_1} . P(v(x_2) = w_2) \quad (3)$$

Obviously, the sums in the squared brackets, that is, $S'(w_2,\phi)_{x1}$, can be calculated independently from one another for each w_2, that is, they can be computed in parallel by different processors. The more technical details concerning the parallelization can be found in the supplementary electronic information.

Even after optimization and parallelization, GENEHUNTER-TWOLOCUS may still run for hours or even days with larger pedigrees. To avoid the annoyance of losing the results in case that a lengthy program run breaks down shortly before regular program termination, we have implemented a restart mechanism into GENEHUNTER-TWOLOCUS. This enables the user to continue the analysis at a point before the system crash.

Results

Optimization and parallelization results

Figure 1 shows the distribution of the run time between program parts in the original and the optimized program versions running with the test data sets. We measured a speedup factor of about 14 in this relatively small case. In larger cases the speedup is most likely to be even higher, because the most time-consuming program parts are getting even more dominant.

The parallelization efforts have been similarly successful. In order to measure the effect of the parallelization, one defines the 'speedup' $S(p)$ of a program due to parallelization as $S(p) = T(1)/T(p)$, where p is the number of processors used and $T(p)$ is the time spent for program execution with p processors. As already mentioned, there is no interprocessor communication during the parallel execution, and thus, the program shows ideal speed up. For example, with eight processors one observes a speedup of eight (Figure 2). However, speedup decreases when the total execution time is in the range of some minutes, as input/output operations then become noticeable. These results were obtained for an analysis of a pedigree with 11 bits. However, even with a 17-bit pedigree, we observed linear scaling up to 272 processors (data not shown). The combined effect of serial optimization and parallelization is shown in Figure 3 for the 17-bit pedigree.

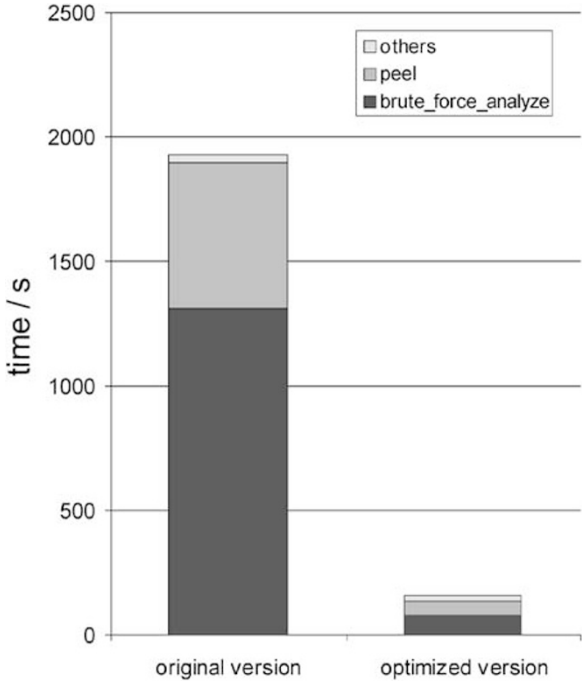

Figure 1 Runtime distribution, separated by the different program parts, in the original *versus* the optimized program version for a small test data set.

Application to a linkage study of autosomal recessive familial hypercholesterolemia

GENEHUNTER-TWOLOCUS was successfully applied in a study of familial hypercholesterolemia.[4] With standard single-disease-locus linkage analysis under a recessive model, two regions were identified; 1p35 (LOD = 3.07) and 13q (LOD = 3.08). When both disease loci were jointly taken into account, a significantly higher LOD score was obtained, compared to an analysis with only one disease locus. In addition, by comparing the results obtained with different two-locus disease models, information on the biological mechanism leading to the disease was gained. In the original study as described by Al-Kateb *et al*,[4] due to the large size of the pedigree, the family could not be analyzed as a whole with the previous version of GENEHUNTER-TWOLOCUS. Two informative individuals therefore had to be discarded. However, the combination of serial optimization and parallelization described here enabled us to analyze the complete pedigree, which is shown in Figure 4. Therefore, we have recalculated the LOD score for all three disease models discussed by Al-Kateb *et al*[4] (see Table 3 therein), that is, for the multiplicative, additive, and heterogeneity model, each assuming a recessive mode of inheritance at both loci. Each of these three jobs, with a pedigree size corresponding to 17 bits, would have required about 3 years with the former GENEHUNTER-TWOLOCUS version (standard PentiumIII 1 GHz PC). With the optimized and parallelized version, a run with 64 processors

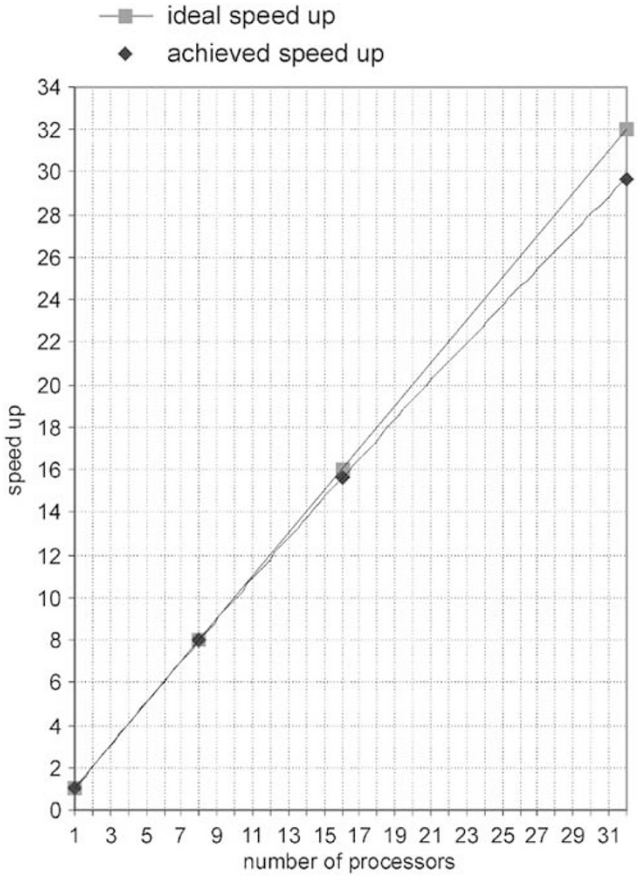

Figure 2 Speedup of program execution, $S(p) = T(1)/T(p)$, as a function of the number of processors p. $T(1)$ and $T(p)$ denote the time spent for program execution with one and p processors, respectively. The pedigree used has been truncated to 11 effective meioses. The test runs have been performed with 1, 8, 16, and 32 processors. The execution time is 3120 s with one processor and 105 seconds with 32 processors. The difference between the curves for ideal and observed speedup is due to the time GENEHUNTER-TWOLOCUS spends with input/output operations. This effect vanishes with larger pedigrees.

takes 30 h, and a run with 272 nodes takes only 7 h. These calculations have been made on the 4 Sun Fire 15 K compute nodes, which are part of the Sun Fire SMP-Cluster of Aachen University (Germany). The machines of this cluster are equipped with 672 UltraSPARC III (Cu) 900 MHz processors altogether.

Table 1 shows the maximum LOD score for the three disease models mentioned above. The first column displays the results as stated by Al-Kateb *et al*[4] for the truncated 13-bit pedigree; the new results for the complete 17-bit pedigree are shown in column 2. Obviously, if all individuals are taken into account, the two-locus LOD scores under all three models increase when compared to the 13-bit analysis, although the gain is not paramount in this particular case. Still, the LOD score of 5.49 for the

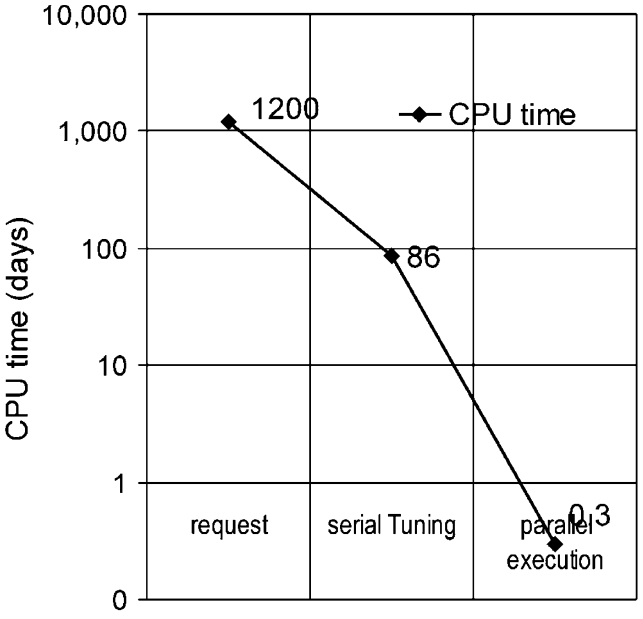

Figure 3 The combined effect of serial optimization and parallelization for the 17-bit pedigree.

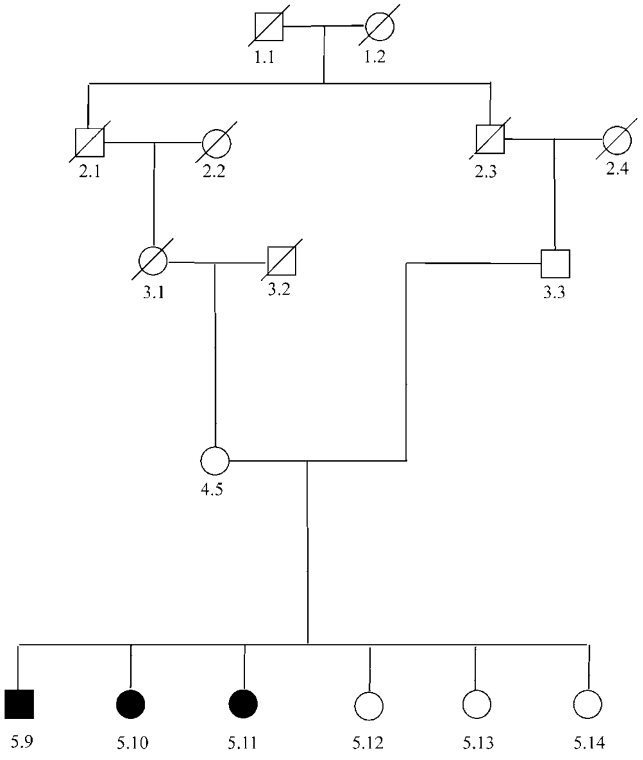

Figure 4 The complete pedigree of the hypercholesterolemia study analyzed in this paper.

multiplicative model is a remarkable result. In addition, the high LOD score difference between the multiplicative model and the heterogeneity or additive model (which

yield LODs of 3.08 and 2.87, respectively) clearly speaks in favor of multiplicative action of both loci, where only individuals with a homozygous-mutant genotype at both disease loci express the disease. Therefore, the finding obtained here corroborates and underlines the result presented by Al-Kateb *et al*[4] since now all available individuals in the pedigree could be included.

Three-dimensional plots of the LOD score

In addition to the optimization described in the previous section, we have implemented a functionality into GENEHUNTER-TWOLOCUS, which enables the user to calculate the LOD and NPL score on a two-dimensional grid extending over the plane spanned by the positions of the two disease loci on their respective chromosomes. This is contrary to the previous version where the position of just one-trait locus is varied, with the position of the other locus being fixed. Figure 5 shows the three-dimensional plot of the LOD score function for the data set of a study of high factor VIII levels in venous thromboembolism (M Berger *et al*, personal communication), which has been analyzed with the GENEHUNTER-TWOLOCUS version described in this paper. The position of the maximum LOD score is denoted in the headline of the plot. In order to get all of the information contained in this 3-D plot with the former version of GENEHUNTER-TWOLOCUS, the program would have to be started many times, once for each position of the first disease locus. With our improvement described here, the LOD or NPL score function is obtained in a single run over the complete range spanned by the two marker maps.

Table 1 Maximum LOD scores of 13- and 17-bit pedigrees

Maximum LOD score	*13-bit*	*17-bit*
Multiplicative model	5.41	5.49
Heterogeneity model	2.82	3.08
Additive model	2.75	2.86

Comparison of GENEHUNTER-TWOLOCUS, TLINKAGE and SUPERLINK

The Elston–Stewart-based TLINKAGE[14] and the newly developed SUPERLINK,[15] which uses Bayesian Networks,[16] are programs that also perform two-trait-locus linkage analysis. Performance tests with these programs concerning computation time and memory requirements and comparisons with GENEHUNTER-TWOLOCUS can be found in the supplementary electronic information.

To summarize the results of these tests, it can be said that SUPERLINK (version 1.2) is very well suited for large pedigrees, even if a considerable number of markers is involved, and thus can handle some of the cases that are unreachable for the Elston–Stewart-based TLINKAGE and

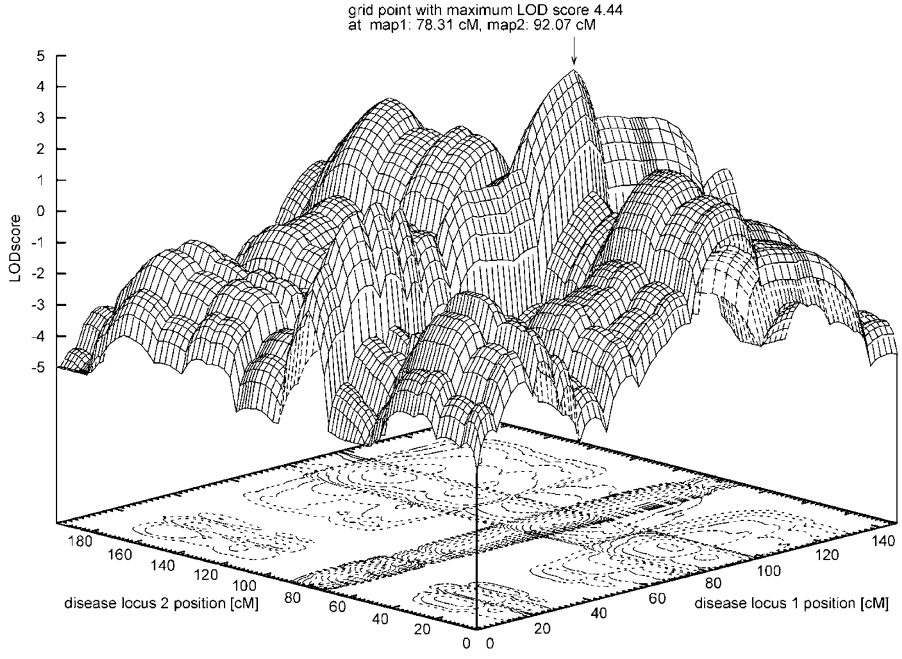

Figure 5 LOD score function for the data set of a study of high factor VIII levels in venous thromboembolism (M Berger *et al.*, personal communication). The dashed lines at the bottom of the plot indicate curves of constant height at the LOD-score surface.

the Lander–Green-based GENEHUNTER-TWOLOCUS. Still, the increase of computation time using SUPERLINK grows more than linearly with the number of markers. Also, the program does not perform as well if some of the individuals are untyped. Therefore, the data of the hypercholesterolemia study cannot be analyzed by SUPERLINK with the complete set of markers, as is possible with GENEHUNTER-TWOLOCUS. TLINKAGE can also handle large pedigrees, but with only very few markers. The new version of GENEHUNTER-TWOLOCUS, presented here, can analyze moderately large pedigrees, with at least up to 18 bits, while it is possible to handle large numbers of markers and disease-locus positions. The fact that GENEHUNTER-TWOLOCUS calculates complete LOD and NPL score surfaces at practically no additional computation time is particularly useful in the context of a two-disease-locus analysis.

Discussion

The genetic dissection of complex traits remains to be one of the great challenges of contemporary science. For the successful mapping of genes causing such diseases, two issues, among others, will be of major importance in a linkage analysis: (i) an adequate phenotype definition, and (ii) a correct model of the genotype–phenotype relation (see, eg Strauch *et al*[17]). The second issue includes the genetic architecture of the trait, that is, the number of loci that determine disease susceptibility, as well as the genotype frequencies and penetrances for all genotype combinations. Our work addresses this matter, more precisely, the modeling of the genotype–phenotype relation in the context of two trait loci. We have completely focused on dichotomous traits, or traits which have been dichotomized. Usually, a linkage analysis of genome-scan data is performed under the assumption of only one trait locus, even if it is known that the disease under study is governed by two or more loci. It is evident that the power to detect linkage is highest under a disease model that is sufficiently close to the true mode of inheritance. Therefore, it is definitely favorable to analyze a complex disease that is caused by at least two loci under a two-trait-locus model. Of course, this does not mean that such an analysis is best as the first step. Rather, an analysis with two trait loci should follow the standard single-disease-locus analysis, for a pair of genetic regions initially showing promising results. Under certain assumptions, it is possible to derive the parameters of the two-locus model from the best-fitting single-locus models.[18]

Like the original GENEHUNTER program, GENEHUNTER-TWOLOCUS is based on the Lander–Green algorithm that allows to incorporate many markers into the analysis. This benefit arises from the fact that the computational demands increase only linearly with the number of markers. The framework of multimarker analysis offers a great advantage: if a marker happens to be not completely informative, for example, due to homozygous or untyped individuals, the inheritance pattern can be reconstructed, often to a large degree, by spill-over of information from adjacent markers. This feature of multimarker analysis is already of advantage when highly polymorphic microsatellite markers are used. It will become a necessity once genome-scans for linkage are performed with a large number of less informative single nucleotide polymorphisms. On the other hand, computation time and memory requirements of the Lander–Green algorithm increase exponentially with the number of meioses in a pedigree. With GENEHUNTER-TWOLOCUS, where two trait loci must be considered in the calculations, the computation time becomes prohibitively long for larger pedigrees. Therefore, two-trait-locus analysis up to now was restricted to pedigrees not exceeding 12 or 13 bits. Here, we have presented major algorithmic improvements of GENEHUNTER-TWOLOCUS. To begin with, the optimization of time-critical parts of the source code decreases the computation time by one order of magnitude. On top of that, the parallelization results in a further speedup. Since no idle times or overhead due to interprocessor communication occur, all of the computation time spent by the processors fully contributes to the two-locus score calculation. Therefore, the new version of GENEHUNTER-TWOLOCUS scales perfectly, that is, the speedup due to parallelization is equal to the number of processors. This results in a further significant cutdown of the time needed for an analysis. Together, the algorithmic improvements have increased the size of pedigrees that can be analyzed by GENEHUNTER-TWOLOCUS from 12 or 13 bits to 18 bits and more, depending on the available computational resources. Thus, it is now possible to analyze pedigrees of reasonable size, almost the same as with the single-locus GENEHUNTER version, which would have taken years to be analyzed before. In many cases, clusters of personal computers running Linux are locally available at institutions. Alternatively, access to a large-scale supercomputing resource with many processors can be applied for. Even on a single-processor machine, users will benefit by the serial optimization.

In addition to the computational improvements, the two-trait-locus LOD and NPL scores are now calculated with the positions of both trait loci varied in their respective marker groups without the need of additional computation time. With the previous version, the position of the first disease locus was fixed as specified by the user. Now, one no longer has to guess the exact position of the first trait locus out of the blue, or perform several runs for different fixed positions. Instead, the two-locus score is obtained for the complete grid of disease-locus positions within both marker groups. With the new GENEHUNTER-TWOLOCUS version, these results are

readily prepared to be viewed with GNUPLOT (http://www.gnuplot.info/), or a different graphics package, as a three-dimensional picture, like the example shown in Figure 5. This allows to easily interpret the results, and to identify the maximum-LOD- or NPL-score, as well as the corresponding disease-locus positions, at a single glance. It gives researchers the 'complete picture' of an analysis with two trait loci.

GENEHUNTER-TWOLOCUS offers the possibility to compute both LOD and NPL scores simultaneously. The results obtained by the old and new version are identical, for both types of statistics. With two-locus LOD-score analysis, it is necessary to specify a considerable number of parameters. Under the assumption that both trait loci are di-allelic, there are two disease allele frequencies. In addition, for three possible genotypes at each locus, a 3×3 matrix of two-locus penetrances must be specified. This holds under the assumption that no imprinting takes place. Under imprinting, which is also called parent-of-origin effect, heterozygotes need to be distinguished by the parent who transmitted the mutation. Thus, an adequate single-disease-locus model with imprinting contains two heterozygote penetrances or four penetrances altogether, as implemented in the single-trait-locus GENEHUNTER-IMPRINTING program.[1] GENEHUNTER-TWOLOCUS also incorporates imprinting models, in the context of two-trait loci; here, the model extends to a 4×4 penetrance matrix. Since the disease model is often unknown for a complex trait, researchers may find it convenient to use two-locus NPL analysis. It comes to a result without a particular model. Still, although the number of parameters to specify is large, a two-trait-locus LOD score analysis offers the possibility to test several disease models. Such a procedure is clearly explorative. It gives researchers the opportunity to gain information not only about evidence for linkage and the positions of both disease loci, but also on how the two loci act, and interact, on the trait. As an example, we have reanalyzed the hypercholesterolemia data, as a follow-up of the study published by Al-Kateb *et al*,[4] with the complete 17-bit pedigree. With a two-trait-locus analysis jointly taking into account the two significantly linked loci at 1p and 13q in that family, we were able to rule out the heterogeneity and additive two-locus models, and clearly confirmed the multiplicative model as the correct two-locus mode of inheritance, with an associated LOD score of 5.49. The significant statistical evidence for interaction of these two loci in the pathogenesis of hypercholesterolemia is underscored by (1) identification of the causative mutations for hypercholesterolemia at 1p36.1–p35 that reside in the ARH gene encoding a putative LDL receptor adaptor,[19,4] and (2) by evidence that the second region at 13q22–q32 harbors a locus modifying LDL cholesterol levels in both normal individuals and in a family with autosomal dominant familial hypercholesterolemia.[20]

Our timing and memory comparisons between GENEHUNTER-TWOLOCUS, TLINKAGE, and SUPERLINK have shown that only GENEHUNTER-TWOLOCUS can handle the large number of markers available with the hypercholesterolemia data set. If there are not too many markers, it is also possible to use the program SUPERLINK, which can handle some of the cases unreachable for the other two programs. However, SUPERLINK does not perform as well if some of the individuals are untyped. While TLINKAGE is able to analyze large pedigrees, it can only cope with two or, at most, three markers. With SUPERLINK and TLINKAGE, the calculation time drastically increases if the two-locus LOD is to be computed for many combinations of disease-locus positions, while the time remains practically the same with GENEHUNTER-TWOLOCUS.

In conclusion, we offer a tool for genetic linkage analysis with two trait loci, that has been considerably improved and enhanced. This applies to convenience in usage and interpretation of the results, as well as to computational performance. Two-trait-locus linkage analysis with many markers is no longer restricted to small pedigrees. Therefore, we hope our new developments will help to maximize the power to detect linkage, which is otherwise low in the context of complex traits. The new version of GENEHUNTER-TWOLOCUS can be obtained without charge by sending an e-mail to the first author.

Acknowledgements

This work was supported by Grants Str643/1 (Project D2 of FOR423), GRK246/TP07, and SFB400/Z2 of the Deutsche Forschungsgemeinschaft, as well as by Grant GEM-Bonn-NGFN-01GS0201 of the Bundesministerium für Bildung und Forschung.

References

1 Strauch K, Fimmers R, Kurz T, Deichmann KA, Wienker TF, Baur MP: Parametric and nonparametric linkage analysis with imprinting and two-locus-trait models: application to mite sensitization. *Am J Hum Genet* 2000; **66**: 1945–1957.
2 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.
3 Kruglyak L, Lander ES: Faster multipoint linkage analysis using Fourier transforms. *J Comput Biol* 1998; **5**: 1–7.
4 Al-Kateb HA, Bähring S, Hoffmann K *et al*: Mutation in the ARH Gene and a chromosome 13q locus influence cholesterol levels in a new form of digenic-recessive familial hypercholesterolemia. *Circ Res* 2002; **90**: 951–958.
5 Nath SK, Chen CH, Schork NJ: Two-trait-locus linkage analyses of asthma susceptibility. *Genet Epidemiol* 2001; **21** (Suppl 1): S278–S283.
6 Lander ES, Green P: Construction of multilocus genetic linkage maps in humans. *Proc Natl Acad Sci USA* 1987; **84**: 2363–2367.
7 Kruglyak L, Daly MJ, Lander ES: Rapid multipoint linkage analysis of recessive traits in nuclear families, including homozygosity mapping. *Am J Hum Genet* 1995; **56**: 519–527.
8 Elston RC, Stewart J: A general model for the genetic analysis of pedigree data. *Hum Hered* 1971; **21**: 523–542.

9 Kothari K, Lopez-Benitez N, Poduslo SE: High-performance implementation and analysis of the Linkmap program. *J Biomed Inform* 2001; **34**: 405–414.
10 Lathrop GM, Lalouel JM: Easy calculation of lod scores and genetic risks on small computers. *Am J Hum Genet* 1984; **36**: 460–465.
11 Lathrop GM, Lalouel JM, Julier C, Ott J: Strategies for multilocus linkage analysis in humans. *Proc Natl Acad Sci USA* 1984; **81**: 3443–3446.
12 Matise TC, Schroeder MD, Chiarulli DM, Weeks DE: Parallel computation of genetic likelihoods using CRI-MAP, PVM and a network of distributed workstations. *Hum Hered* 1995; **45**: 103–116.
13 Conant G, Plimpton S, Old W *et al*: Parallel genehunter: implementation of a linkage analysis package for distributed-memory architectures. *J Parallel Distr Com* **63**: 674–682.
14 Lathrop GM, Ott J: Analysis of complex diseases under oligogenic models and intrafamilial heterogeneity by the LINKAGE programs. *Am J Hum Genet* 1990; **47**: A188, (abstract).
15 Fishelson M, Geiger D: Exact genetic linkage computations for general pedigrees. *Bioinformatics* 2002; **18** (Suppl.): 189–198.
16 Pearl J: *Probabilistic reasoning in intelligent systems*. San Francisco, CA: Morgan Kaufmann.
17 Strauch K, Fimmers R, Baur MP, Wienker TF: How to model a complex trait 1. General considerations and suggestions. *Hum Hered* 2003; **55**: 202–210.
18 Strauch K, Fimmers R, Baur MP, Wienker TF: How to model a complex trait 2. Analysis with two disease loci. *Hum Hered* 2003; **56**: 200–211.
19 Garcia CK, Wilund K, Arca M *et al*: Autosomal recessive hypercholesterolemia caused by mutations in a putative LDL receptor adaptor protein. *Science* 2001; **292** (5520): 1394–1398.
20 Knoblauch H, Muller-Myhsok B, Busjahn A *et al*: Cholesterol-lowering gene maps to chromosome 13q. *Am J Hum Genet* 2000; **66**: 157–166.

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