Short Communication

LINKGEN: A new algorithm to process data in genetic linkage studies

Rodrigo Secolin, Cristiane S. Rocha, Fábio R. Torres, Marilza L. Santos, Cláudia V. Maurer-Morelli, Neide F. Santos, Iscia Lopes-Cendes*  

Department of Medical Genetics, Faculty of Medical Sciences, University of Campinas (UNICAMP), Tessália Vieira de Camargo, 126, Cidade Universitária Zeferino Vaz, Campinas SP, Brazil, CEP 13084-971

Received 24 July 2007; accepted 21 February 2008  
Available online 2 April 2008

Abstract

Genetic linkage studies using whole genome scans are useful approaches for identifying genes related to human diseases. In general, these studies require genotyping of a large number of markers, which are used in statistical analysis. Recent technology has allowed easy genotyping of a large number of markers in less time; therefore, interface programs are required for manipulation of these large data sets. We present a new algorithm, which processes input data in LINKAGE format from data analyzed by automated genotyping systems. The algorithm was implemented in PERL script and R environment. Validation was performed with genotyped data from 127 individuals and 720 microsatellite markers of two whole genome scans. Our results showed a significant decrease in data processing time. In addition, this algorithm provides unbiased allele frequency estimation used for linkage analysis. LINKGEN is a freely available online tool and allows easier, faster, and reliable manipulation of large genotyping data sets. © 2008 Elsevier Inc. All rights reserved.

Keywords: Linkage studies; Data handling; Bioinformatics

Introduction

Genetic linkage studies using whole genome scans are useful tools for finding human disease loci and genes [1,2]. Approximately 300–450 microsatellite markers or high-density SNP maps are required for these studies [3,4]. In parametric linkage studies, as well as meta-analysis genome scans, microsatellite markers are still the best choice for genotyping large single families or multiple small families [3,5,6]. Recently, high-throughput automated genotyping equipment has been used for analyzing these molecular markers, such as MegaBace (GE Healthcare, Chicago, IL) and ABI (Applied Biosystems, Foster City, CA) systems. However, data output format from most of this equipment is not compatible with input format required by several types of parametric and nonparametric genetic linkage software, such as LINKAGE, GENEHUNTER, MERLIN, and S.A.G.E. packages [7–10]. Although interface programs have been implemented for microarray-based SNP genotyping [11], there is no such tool available for processing of microsatellite genotype data. Furthermore, since not all individuals used in linkage studies are genotyped, the manual data processing of these large data sets leads to tedious and laborious work, which may increase the probability of clerical errors.

In this study we present a new algorithm, named LINKGEN, designed to generate an interface between most types of automated genotyping software and statistical packages for linkage analysis; in addition, it includes allele frequency estimates from pedigree data sets.

Results and discussion

LINKGEN achieves a faster and easier data processing for linkage analysis. Indeed, our results showed that all 720 microsatellite genotypes were correctly distributed across all 127 family members, using the LINKGEN algorithm in each validation study. Data manipulation can be carried out by the web; therefore, the software is able to operate in any operational system and for any type of analysis which uses genetic linkage statistical methods. In fact, several studies in our laboratory have used this algorithm [12–14], decreasing considerably the statistical processing time.
Since most output data formats from different automated genotyping equipment have at least a genotype call column, our software is able to accept output data from many genotyping systems currently in use, without a lot of preprocessing. In addition, most SNP genotyping equipment also includes genotype call columns, making LINKGEN also suitable for use in studies in which only a few SNP markers are genotyped.

Approximately 1500 disease genes have been identified and listed in the Online Mendelian Inheritance in Man (OMIM) using genetic linkage studies [2,15]. A large number of microsatellite or SNP markers are required for whole genome scans [2–4]. Since advanced genotyping techniques of microsatellite markers are now available, interface programs between genotyping and statistical software became an important tool for processing data in these studies, such as genome-wide scans and meta-analysis of genome scans [5,6]. In conclusion, this software allows construction of appropriate input files for most linkage analysis programs, providing an easier, faster, and reliable manipulation of large genotyping data sets in whole genome linkage analysis.

Materials and methods

Implementation

LINKGEN was implemented in PERL scripts and used two files, corresponding to two matrices defined as follows: a pedigree structure matrix \( P = (p_{i,j})_{m \times n} \) built in PRE-LINKAGE format [3], and a marker genotype matrix \( M = (m_{i,j})_{s \times g} \) from any genotyping system and any number of markers (Fig. 1). In \( P \) matrix, there are up to seven columns, as follows: \( p_{i,1} \) = pedigree identification (ID); \( p_{i,2} \) = individual ID; \( p_{i,3} \) = father ID; \( p_{i,4} \) = mother ID; \( p_{i,5} \) = sex; \( p_{i,6} \) = affection status; and \( p_{i,7} \) = optional liability class. Here we included a last column \( p_{i,g} \) with \( g = 7 \) or 8, depending on the presence of a liability class column. This last column indicates whether the individuals have been genotyped (\( p_{i,g} = 1 \)) or not (\( p_{i,g} = 0 \)). In the allele marker matrix, we included two initial columns, where \( m_{i,j} \) = pedigree ID, and \( m_{i,2} \) = individual ID. Therefore, our algorithm combines these two matrices in a matrix \( R = (r_{a,b})_{m \times ((n-1) + (z-2))} \), where each \( i \)-th row follows the formula:

\[
r_{i,1...((n-1)+(z-2))} = \begin{cases} 
(p_{i,1}...p_{i,g-1}) \ U (m_{i,3}...m_{i,2}) & \text{if } p_{i,1} = m_{i,1} \land p_{i,2} = m_{i,2} \land p_{i,g} = 1 \\
(p_{i,1}...p_{i,g-1}) \ U (m_{i,3} = 0...m_{i,2} = 0) & \text{if } p_{i,1} = m_{i,1} \land p_{i,2} = m_{i,2} \land p_{i,g} = 0 
\end{cases}
\]

Therefore, by this formula, each genotyped individual in pedigree data receives his/her corresponding group of allele markers, whereas each nongenotyped individual receives missing data code 0 (Fig. 1).

These files should have *.fam and *.gen extensions, regarding \( P \) and \( M \) matrices, respectively. Both files can be created in any text editor or calculus sheet processor software (i.e., Microsoft Word or Excel). In principle, the number of rows and columns is unlimited, depending on the file editor and computer configuration. As a result, a *.pre extension file in LINKAGE format is created.

Frequency estimates

Marker frequencies, denoted as \( p \), are estimated based on sibship data according to the formula [16]:

\[
p = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\sum_{j=1}^{k_i} (k_i + 1)},
\]

which, assuming that a marker can have a \( \beta \) allele, \( \beta ij \) is the number of \( \beta \) alleles carried by sibling \( j \) in family \( i \) (=0, 1, or 2); \( n \) is the number of families; and \( k_i \) denotes the number of siblings in \( i \)-th family. Considering Hardy-Weinberg equilibrium and absence of genotyping errors for markers studied, Broman [16] showed that this frequency estimation is unbiased and provides reliable results. These estimates were implemented in R environment and included to LINKGEN PERL scripts.

Validation

To validate the algorithm, we used genotyped data from two large whole genome scans performed in two Mendelian forms of partial epilepsy syndromes: familial mesial temporal lobe epilepsy (OMIM No. 608096) and autosomal dominant partial epilepsy with auditory features (OMIM No. 600512). These studies comprised 68 and 59 individuals and genotyping of 337 and 383 microsatellite markers, respectively. Genotyping was performed automatically using the MegaBACE 1000 (GE Healthcare) system and genotype call by the FRAGMENT PROFILE software. We used LINKAGE, GENEHUNTER, and S.A.G.E. packages for statistical analysis [7,8,10]. Pedigree structure matrix and genotyped allele matrix were built in Microsoft Excel software and saved as *.fam and *.gen files, respectively. Since genotypic data from MegaBace and ABI equipment do not have the same structure as *.gen file, this file is initially built with family and individual ID columns, followed by genotyped allele markers, which are copied and pasted from the genotype call column of FRAGMENT PROFILE output data of each marker file. Finally, these data are saved as *.gen file (Fig. 1).

Availability and requirements

Software home page: http://lgm.fcm.unicamp.br:9001/cgi-bin/linkgen/linkgen.cgi
Operating system(s): platform independent
Programming language: Perl (5.8.8) and R (2.5.1)
Other requirements: none
License: freely available
Fig. 1. Example of LINKGEN algorithm workflow. Squares and circles indicate males and females, respectively. White and black symbols represent unaffected and affected individuals, respectively. In this pedigree, individuals 1 and 6 are not genotyped.
Acknowledgments

This study was supported by FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo) and CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), Brazil. Rodrigo Secolin, Cristiane S. Rocha, Fábio R. Torres, and Neide F. Santos received scholarships from FAPESP and Cláudia V. Maurer-Morelli received a fellowship from CNPq.

References