Molecular Genetics and Metabolism 105 (2012) 382–389

Contents lists available at SciVerse ScienceDirect

Molecular Genetics and Metabolism

journal homepage: www.elsevier.com/locate/ymgme

Recombination mapping using Boolean logic and high-density SNP genotyping for exome sequence filtering

Thomas C. Markello [a,b,*], Ted Han [a,1], Hannah Carlson-Donohoe [c], Chidi Ahaghotu [a], Ursula Harper [d], MaryPat Jones [d], Settara Chandrasekharappa [d], Yair Anikster [a,e], David R. Adams [a,b], NISC Comparative Sequencing Program [f], William A. Gahl [a,b], Cornelius F. Boerkoel [b]

[a] National Human Genome Research Institute, NIH, Bethesda, MD 20892-1611, USA
[b] NIH Undiagnosed Diseases Program, Office of the Director, NIH, Bethesda, MD 20892-1851, USA
[c] University of Minnesota School of Medicine, Minneapolis, MN, 55455, USA
[d] Genomics Core, Genome Technology Branch, NHGRI, NIH, Bethesda, MD, USA
[e] The Edmond and Lily Safra Children's Hospital, Sheba Medical Center, and Sackler School of Medicine, Tel Aviv University, Tel Aviv, Israel
[f] National Intramural Sequencing Center, Rockville, MD 20852, USA

ARTICLE INFO

Article history:
Received 18 December 2011
Accepted 19 December 2011
Available online 23 December 2011

Keywords:
Linkage
Recombination
Mapping
Exome sequencing
Single nucleotide variants

ABSTRACT

Whole genome sequence data for small pedigrees has been shown to provide sufficient information to resolve detailed haplotypes in small pedigrees. Using such information, recombinations can be mapped onto chromosomes, compared with the segregation of a disease of interest and used to filter genome sequence variants. We now show that relatively inexpensive SNP array data from small pedigrees can be used in a similar manner to provide a means of identifying regions of interest in exome sequencing projects. We demonstrate that in those situations where one can assume complete penetrance and parental DNA is available, SNP recombination mapping using Boolean logic identifies chromosomal regions identical to those detected by multipoint linkage using microsatellites but with much less computation. We further show that this approach is successful because the probability of a double crossover between informative SNP loci is negligible. Our observations provide a rationale for using SNP arrays and recombination mapping as a rapid and cost-effective means of incorporating chromosome segregation information into exome sequencing projects intended for disease-gene identification.

Published by Elsevier Inc.

1. Introduction

Identification of genetic loci segregating with disease is often crucial for identification of sequence variants associated with molecularly uncharacterized diseases. Such disease-associated loci have traditionally focused subsequent molecular analyses; however, with the advent of genome and exome sequencing, they are extremely helpful for prioritizing or filtering the thousands of sequence variants detected.

Traditionally, potential disease loci have been identified by linkage analysis using a relatively small number of polymorphic, widely spaced genetic markers. These markers were initially simple biochemical or cytogenetic polymorphisms, then restriction fragment length polymorphisms (RFLPs) or variable number of tandem repeats (VNTRs) and finally microsatellite markers (small tandem repeats or STRs). Recently, widely spaced groups of tightly linked single nucleotide polymorphisms (SNPs) have been employed as genetic markers for linkage studies [1]. These surrogate and actual genotyping techniques produced the first autosomal gene linkage for the Duffy blood group on chromosome 1 [2], later for the CFTR locus on chromosome 7 [3,4] and continue to provide valuable results [5]. These classical linkage strategies generally use widely spaced genotyping markers initially and then follow that with fine mapping. Although computer algorithms such as LIPED [6], MENDEL [7], MLINK [8] and GENEHUNTER [9] have been written to use pedigree and genotyping information to identify genomic regions segregating with the disease phenotype [6], computational limitations and difficulty coping with markers in strong linkage disequilibrium limits use of high density markers in a one-step mapping strategy.

Recent advances in the use of high-density genomic data, such as that obtained by whole genome sequencing, have provided a new means of identifying recombination sites. Whole genome data has been used to map recombination intervals and establish detailed haplotypes using small pedigrees. Recombination intervals can be defined precisely and reduce the “search space for candidate genes” [10,11]. Although methods for using SNP data extracted from exome sequence have also been developed [12], exome sequencing provides

* Corresponding author at: Building 10/10C107, 10 Center Drive, NHGRI, NIH, Bethesda MD 20892-1851, USA. Fax: +1 301 480 3015.
E-mail address: markellot@mail.nih.gov (T.C. Markello).
[1] Current address: GeneDx, Gaithersburg, MD, 20877-2142, USA.

1096-7192/$ – see front matter. Published by Elsevier Inc.
doi:10.1016/j.ymgme.2011.12.014

less contiguous data; large unsequenced introns and intergenic spaces will not provide variants to use for recombination mapping. However, since high-density SNP arrays profile these regions, we hypothesized that using such SNP data would solve this problem at a lower per-individual cost than genome sequencing. Additionally, if SNP arrays could be used for recombination mapping, they would also delineate chromosomal regions of homozygosity [13,14], duplications/deletions, and mosaicism.

The goal of such recombination mapping using SNPs would be to determine all the cases where a series of loci that segregates in a manner consistent with a fully penetrant Mendelian model abruptly changes to a series of loci that is inconsistent with that model. For sufficiently closely spaced loci, these transitions will directly and completely reflect the meiotic recombinations in the parental gametes. In a human genome of length 3200 centiMorgans (cM), a single crossover should occur on average every 30 to 32 Mb if one assumes 1 cM is approximately equivalent to 1 Mb. The random walk probability of a mean free path length between two recombinations, however, makes two recombination events in smaller intervals probable with some real frequency. Such double crossover events have the potential to complicate SNP-based recombination mapping. For any specific pedigree and model of inheritance, the presence of a double crossover occurring between two informative SNP markers will cause a segregation analysis (even homozygosity mapping) to fail to exclude or include genomic regions that could contain the disease-causing mutation. With increasing marker density, however, we hypothesized that there will be a point at which the markers are so close together that the possibility of a double crossover may be ignored with confidence. The question, therefore, is whether the informative SNPs in a high-density SNP array form a dense-enough array of markers to replicate the situation for whole-genome data, i.e., where double crossovers can be ignored during recombination mapping.

Assuming full penetrance, DNA from both parents, and absence of double recombinants between informative SNPs, we further hypothesized that a Boolean logic strategy applied to high density SNP data would produce the same results as a classical linkage strategy, require minimal computational infrastructure, and integrate readily into a SNP array analysis pipeline (e.g., as a plug-in software to the Illumina Genome Studio). This Boolean logic approach to recombination mapping makes use of logical rule sets that are straightforward to design and implement for small pedigrees. Boolean logic operations can produce the truth tables for all logical genotype states that are, or are not,

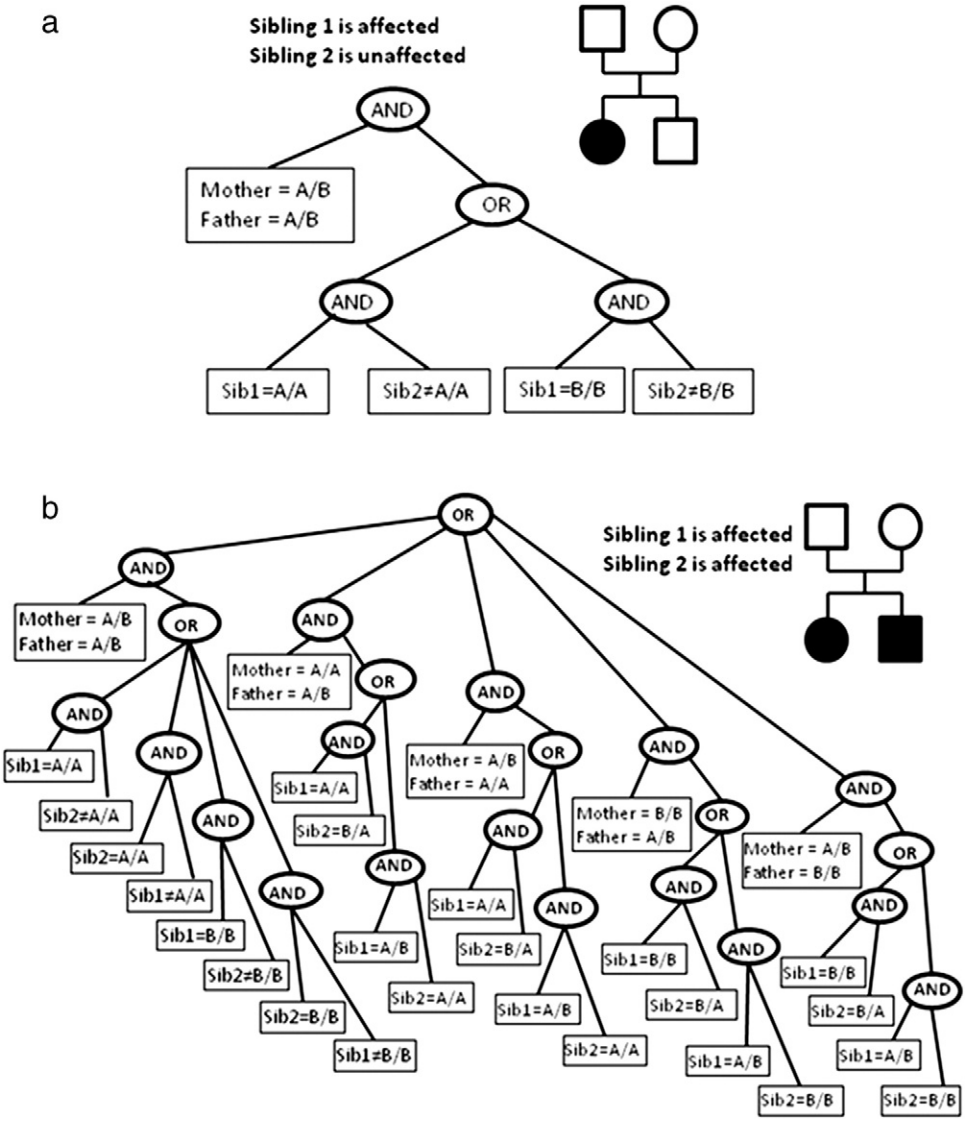

Fig. 1. Boolean filters for families with two children. a: recombination mapping inclusion filter logic tree for one affected and one unaffected sibling with both parents available for genotyping. b: recombination mapping exclusion filter logic tree for two affected siblings and both parents available for genotyping.