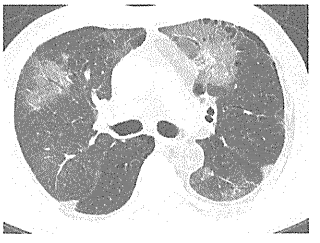
男性

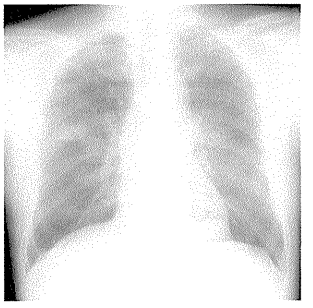
原因薬剤　　エンブレル、メトレート

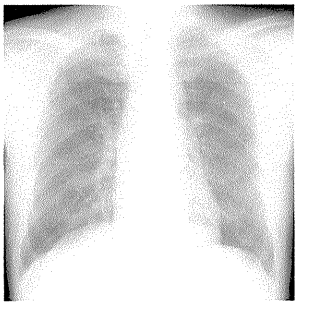
増悪前CT画像　　増悪時CT画像　　増悪後CT画像



2011/03/30

増悪前XP　　増悪時XP　　増悪後XP





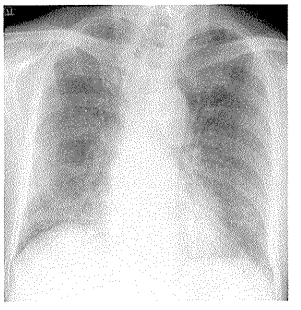
2011/03/30　　2011/05/09

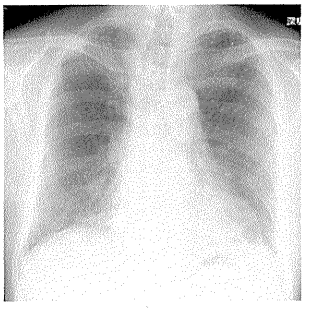
原因薬剤　　カソデックス

増悪前CT画像　　増悪時CT画像　　増悪後CT画像

2010/03/09　　2011/10/07

増悪前XP　　増悪時XP　　増悪後XP

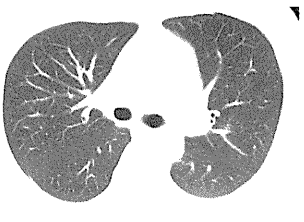


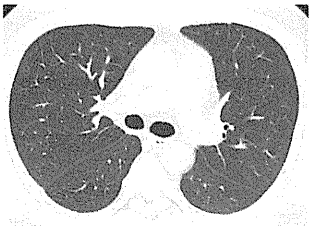


2010/03/09　　2011/10/07

男性　　原因薬剤　　シクロスポリン，オンコビン，ア

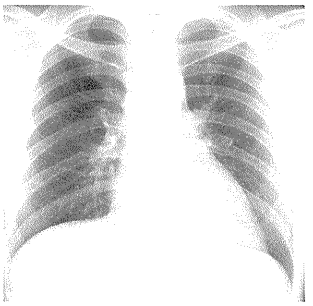
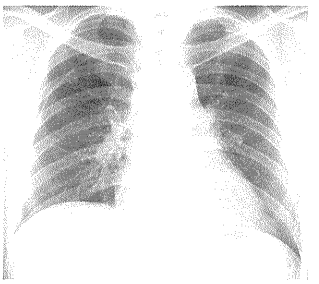
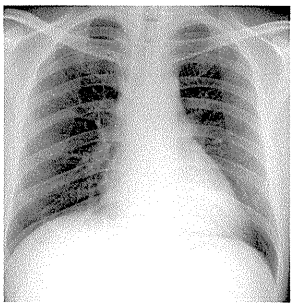
増悪前CT画像　　増悪時CT画像　　増悪後CT画像





2009/09/29　　2009/10/27　　2009/11/02

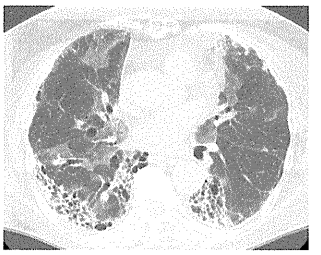
増悪前XP　　増悪時XP　　増悪後XP

2009/09/18　　2009/10/27　　2009/11/05

女性　　原因薬剤　　ヒュミラ

増悪前CT画像　　増悪時CT画像　　増悪後CT画像


2010/05/18　　2010/08/30　　2010/10/06

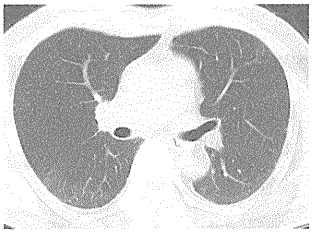
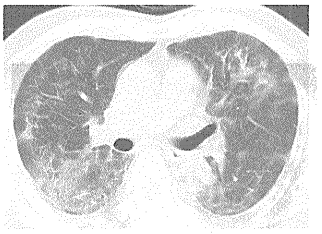
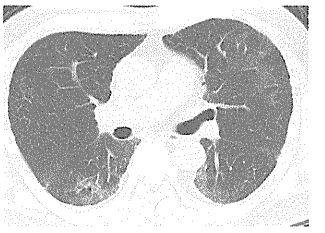
増悪前XP　　増悪時XP　　増悪後XP



2010/05/14　　2010/08/30　　2010/10/01

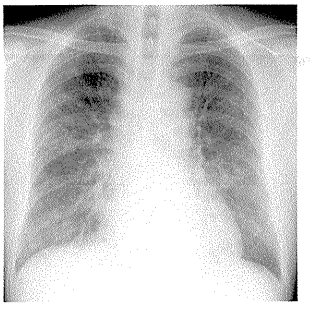
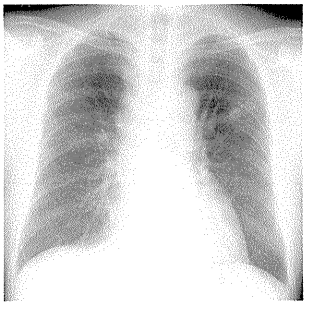
男性　　原因薬剤　　リリカ、ツムラ柴令湯

増悪前CT画像　　増悪時CT画像　　増悪後CT画像

2011/01/29　　2011/08/20　　2011/11/04

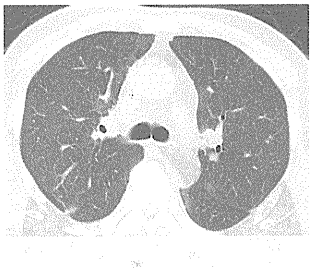
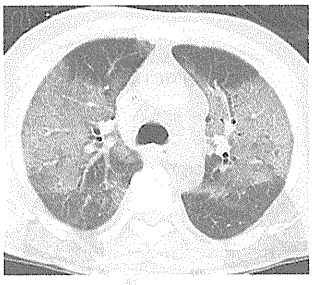
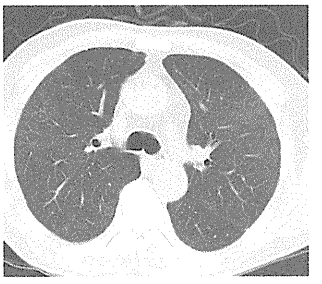
増悪前XP　　増悪時XP　　増悪後XP

2011/08/20　　2011/11/04

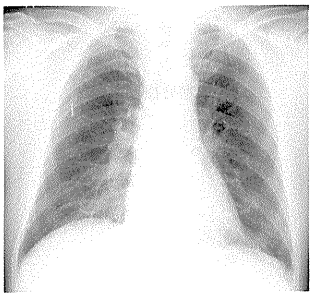
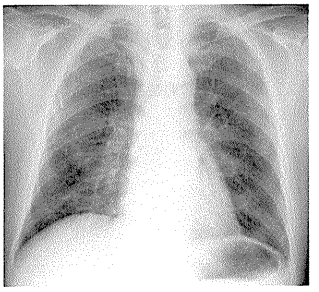
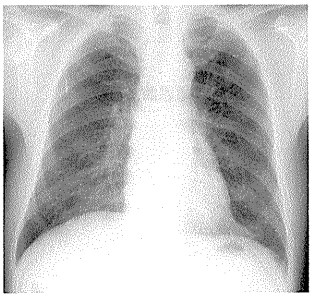
男性　　原因薬剤　　リリカ、半夏瀉心湯

増悪前CT画像　　増悪時CT画像　　増悪後CT画像

2011/09/25　　2011/10/01　　2011/10/14

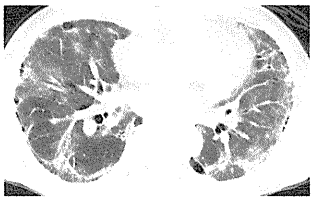
増悪前XP　　増悪時XP　　増悪後XP

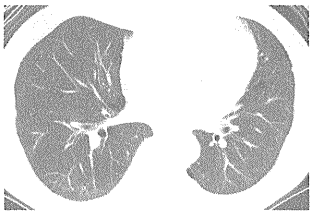




2011/09/25　　2011/09/30　　2011/10/14

男性　　原因薬剤　　ロキソニン

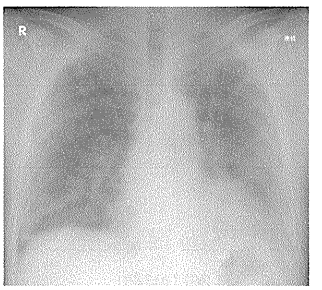
増悪前CT画像　　増悪時CT画像　　増悪後CT画像





2009/09/15　　2010/03/12

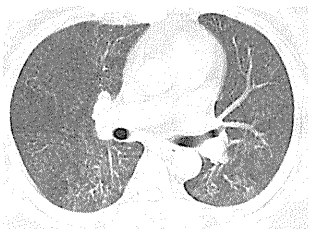
増悪前XP　　増悪時XP　　増悪後XP

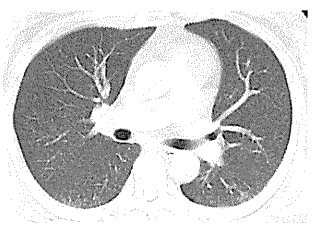


2009/09/15　　2010/03/12

女性　　原因薬剤　　感冒薬、ペンタサ

増悪前CT画像　　増悪時CT画像　　増悪後CT画像



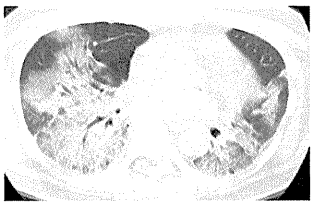


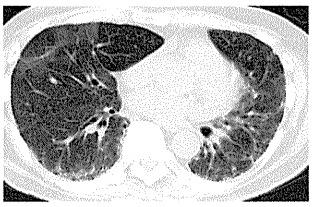
2009/06/23　　2009/06/29　　2009/07/24

増悪前XP　　増悪時XP　　増悪後XP

男性　　原因薬剤　　抗リウマチ薬．アザルフィジン，

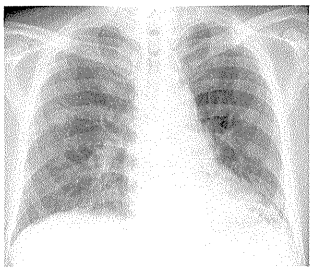
増悪前CT画像　　増悪時CT画像　　増悪後CT画像





2010/04/12　　2010/05/21

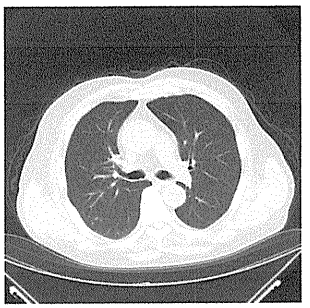
増悪前XP　　増悪時XP　　増悪後XP

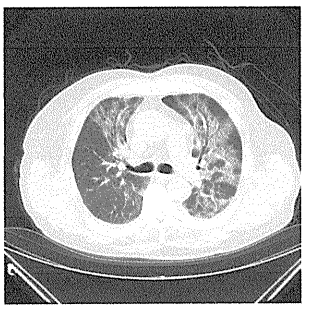


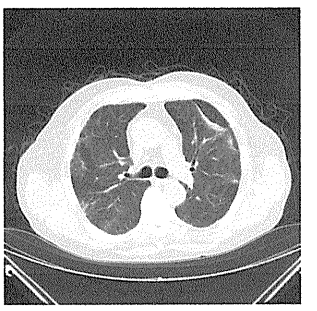
2010/04/12　　2010/05/27

男性　　原因薬剤　　柴朴湯

増悪前CT画像　　増悪時CT画像　　増悪後CT画像

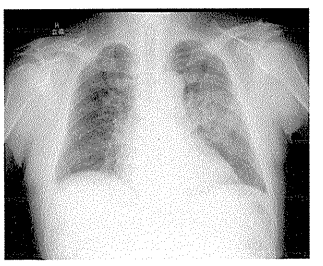


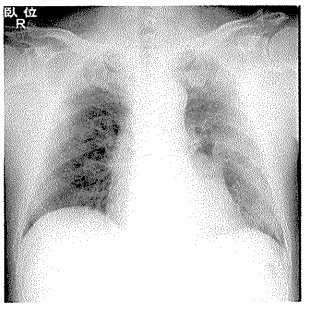




2011.05.10　　2011.05.30　　2011.06.14

増悪前XP　　増悪時XP　　増悪後XP





2011.05.10　　2011.05.30　　2011.06.13

女性　　原因薬剤　小柴胡湯

増悪前CT画像　　増悪時CT画像　　増悪後CT画像

2011/01/11　　2011/01/27

増悪前XP　　増悪時XP　　増悪後XP

遺伝子解析研究手法開発
―報告論文―

OPEN ACCESS Freely available online

PLoS one

Homozygosity Mapping on Homozygosity Haplotype Analysis to Detect Recessive Disease-Causing Genes from a Small Number of Unrelated, Outbred Patients

Koichi Hagiwara[1]*, Hiroyuki Morino[2], Jun Shiihara[1], Tomoaki Tanaka[1], Hitoshi Miyazawa[1], Tomoko Suzuki[1], Masakazu Kohda[3,4], Yasushi Okazaki[3,4], Kuniaki Seyama[5], Hideshi Kawakami[2]

1 Department of Respiratory Medicine, Saitama Medical University, Moroyama, Saitama, Japan, **2** Department of Epidemiology, Research Institute for Radiation Biology and Medicine, Hiroshima University, Hiroshima, Hiroshima, Japan, **3** Division of Functional Genomics and Systems Medicine, Research Center for Genomic Medicine, Research Center for Genomic Medicine, Saitama Medical University, Hidaka, Saitama, Japan, **4** Division of Translational Research, Research Center for Genomic Medicine, Research Center for Genomic Medicine, Saitama Medical University, Hidaka, Saitama, Japan, **5** Department of Respiratory Medicine, Juntendo University School of Medicine, Bunkyo-ku, Tokyo, Japan

Abstract

Genes involved in disease that are not common are often difficult to identify; a method that pinpoints them from a small number of unrelated patients will be of great help. In order to establish such a method that detects recessive genes identical-by-descent, we modified homozygosity mapping (HM) so that it is constructed on the basis of homozygosity haplotype (HM on HH) analysis. An analysis using 6 unrelated patients with Siiyama-type α1-antitrypsin deficiency, a disease caused by a founder gene, the correct gene locus was pinpointed from data of any 2 patients (length: 1.2–21.8 centimorgans, median: 1.6 centimorgans). For a test population in which these 6 patients and 54 healthy subjects were scrambled, the approach accurately identified these 6 patients and pinpointed the locus to a 1.4-centimorgan fragment. Analyses using synthetic data revealed that the analysis works well for IBD fragment derived from a most recent common ancestor (MRCA) who existed less than 60 generations ago. The analysis is unsuitable for the genes with a frequency in general population more than 0.1. Thus, HM on HH analysis is a powerful technique, applicable to a small number of patients not known to be related, and will accelerate the identification of disease-causing genes for recessive conditions.

Citation: Hagiwara K, Morino H, Shiihara J, Tanaka T, Miyazawa H, et al. (2011) Homozygosity Mapping on Homozygosity Haplotype Analysis to Detect Recessive Disease-Causing Genes from a Small Number of Unrelated, Outbred Patients. PLoS ONE 6(9): e25059. doi:10.1371/journal.pone.0025059

Editor: Kazutaka Ikeda, Tokyo Metropolitan Institute of Medical Science, Japan

Received July 29, 2011; **Accepted** August 26, 2011; **Published** September 20, 2011

Funding: This work is supported in part by the grant-in-aid for scientific research (No. 18390242) from the Japan Society of Promotion of Science, and in part by the grants-in-aid for Health and Labor Science [Nos. H22-Nanchi-Ippan-005 to K.H. and H20-Nanchi-Ippan-023 to K.H.] from the Ministry of Health, labor and Welfare, Japan. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: hagiwark@saitama-med.ac.jp

Introduction

Identification of susceptible genetic loci is of great importance for understanding the underlying mechanisms of a number of diseases, and thus aiding the development of their treatment. Whole-genome association studies using individuals not known to be related have been very successful for the analysis of common diseases [1], while linkage-based approaches have identified a number of genes with large effect sizes [2]. More lately, greater attention has been directed to diseases that cannot be investigated using these approaches, either because of the difficulty in collecting a large number of samples, or in finding a sizeable family with the disease [3]. Such diseases include those caused by multiple rare genetic variants or by genes with low penetrance or with effects that become apparent only in the elderly [4]. For unraveling the causes of such diseases, there is the need for an approach that is effective in the context of a small number of patients not known to be related.

The homozygosity mapping (HM) method was developed to identify a disease-causing gene through analyses of patients from inbred families [5]. This principle was later expanded and applied to patients from outbred families [6,7]. Moreover, the use of SNP data from genome-wide analyses has increased the sensitivity of the detection [8,9]. However, because the algorithm employed in HM is highly vulnerable to genotyping errors, an appropriate correction for such errors is required [10].

In contrast, the homozygosity haplotype (HH) method [9] is an imputation-free method for determining haplotypes, because it uses only a fraction of SNP genotyping data. When a region of conserved homozygosity haplotype (RCHH) is observed in different individuals, there is a reasonable possibility that these individuals share an identical-by-descent (IBD) fragment in 1 or both strands of the homologous chromosomes. The algorithm is robust to genotyping errors and thus requires very little or no correction for genotyping errors.

During a previous study that aimed to identify a disease-causing gene for amyotrophic lateral sclerosis (MIM 613435) [11], we encountered 2 unrelated patients who shared the same homozygous mutation in the *OPTN* gene (MIM 602432). In addition, the region of DNA encompassing the gene contained a number of SNPs that were homozygous in both patients (a runs of homozygous SNPs [RHS] [10]). Further, the RHS was contained

in a 0.9-Mb region of conserved HH (RCHH) [9]. In contrast, the length of RCHH shared between either of the 2 patients and each of the 85 control subjects was shorter than 0.9 Mb. We therefore concluded that these 2 patients are very likely shared the disease-causing IBD gene [11]. We considered that the reasoning had a general application and the presence of a long RCHH that contains an RHS strongly suggested the presence of an IBD fragment. We then encoded this reasoning into a computer program, thereby establishing HM on HH analysis. Here, we show here that this is a powerful method that can identify susceptible loci by identifying homozygous IBD fragments from a small number of outbred patients.

Methods

Ethics Statement

This study was approved by the Institutional Review Boards of Saitama Medical University, Tokyo University, and Juntendo University. All patients involved in the current study provided written informed consent.

HM on HH analysis

HM on HH analysis is a combination of HM analysis [5,10] employing controls and HH analysis [9] employing controls. The analysis does not presume that the patients are from inbred families, and can be performed on patients from the general population. It searches for an RHS overlap that is contained in an RCHH (see below). A candidate region thus obtained may contain a recessive disease-causing gene.

Most recent common ancestor (MRCA)

For patients sharing a disease-causing gene, the most recent common ancestor (MRCA) is the most recent ancestor from whom they inherited the recessive disease-causing gene (**Figure 1A**). Therefore, in the patients, the disease-causing gene is IBD. HM on HH analysis identifies 2 or more patients who are homozygous for this gene.

Structure formed by the IBD fragments

The IBD fragments generate characteristic regions in the genotyping data both in a single patient and between 2 patients.

In a single patient, the overlap of 2 IBD fragments forms an RHS if its length is greater than the RHS cutoff (**Figure 1B**) [10]. Between 2 patients, RHSs can form an overlap (RHS overlap, hereafter). In the RHS overlap, the genotypes of both subjects are identical, forming an RHS overlap in which 2 subjects share an identical genotype (RHS overlap IG, hereafter) (**Figure 1C**). In addition, the overlap of the "region in which at least 1 fragment is derived from the MRCA" generates an RCHH if its length is greater than the RCHH cutoff (**Figure 1C**) [9]. The RHS overlap IG is contained in the RCHH, and the structure is hereby called the RHS overlap IG-RCHH nest. An RHS overlap IG-RCHH nest may be formed by chance between a patient and a control due to a coincidence in the SNP genotype. However, the RHS overlap IG-RCHH nest between the patients is likely to be longer, both in the size of the RHS overlap IG and in the size of the RCHH, than that formed by chance between a patient and a control (**Figure 1D**). Consequently, if we detect an RHS overlap IG-RCHH nest between 2 patients and it is longer than any of that detected between each patient and each control both in the size of the RHS overlap IG and in the size of the RCHH, the RHS overlap IG-RCHH nest is likely to suggest the presence of the IBD fragments in these 2 patients.

HM on HH analysis

HM on HH analysis searches for the RHS overlap IG-RCHH nest. The analysis is composed of 4 steps. Step 1: HM. The RHSs are obtained, and the RHS overlaps are selected as candidate regions for a disease-causing gene (**Figure 2A**) [10]. Step 2: Intermediate analysis 2 (IM2). RHS overlap IGs are selected as candidate regions (**Figure 2B**). Step 3: Intermediate analysis 3 (IM3). For each SNP position contained in an RHS overlap IG detected in Step 2, the presence of an RHS overlap IG between a patient and a control is investigated. When the RHS overlap IG between the 2 patients is longer in size than any of those between a patient and a control, it is selected as a candidate region (**Figure 2C**). Step 4: HH analysis using controls. The RHS overlap IG-RCHH nest is determined between 2 patients. For each SNP position contained in the RHS overlap IG in the RHS overlap IG-RCHH nest, the presence of an RHS overlap IG-RCHH nest formed between a patient and a control is investigated. When the RHS overlap IG between the 2 patients is longer in length than any of those formed between a patient and a control, and the RCHH between the 2 patients is longer in length than any of those formed between a patient and a control, the RHS overlap IG is selected as a candidate region (**Figure 2D**).

Parameter values

The parameter values used in the current study were as follows. The RHS cutoff was 1.2 centimorgans. At this cutoff, the total length of the regions falsely identified as RHSs was less than 1.5 centimorgans in a genome-wide search [10]. Meanwhile, 8.4% of the total length of RHSs fail to be identified as RHSs when the MRCA occurred 20 generations ago; 25%, 40 generations ago; 42%, 60 generations ago; 57%, 80 generations ago, and 69%, 100 generations ago (**Figure S1A**). Before detecting the RHSs, a genotyping error correction algorithm was applied, with the suspected genotyping error rate set at 0.006 [10]. The RCHH cutoff was 0.0 centimorgans; thus, a match of HH of any length was considered to be an RCHH.

Human subjects

Patients with Siiyama-type α1-antitrypsin deficiency (MIM 107400.0039). Siiyama-type α1-antitrypsin deficiency is a rare recessive disease in Japan [12]. Whole-genome high-density SNP array genotyping data of 6 patients [10], who were not related and lived in different areas of Japan, were used in the current study. All patients provided written informed consent. The maximal likelihood estimates of the generational distance of the MRCA for each pair of patients ranged between 5 and 74 (median 61) generations.

Control subjects. The whole-genome high-density SNP genotyping data of 198 healthy Japanese subjects from the general population were provided by Prof. Tokunaga, Tokyo University. Additionally, the SNP genotyping data of 116 JPT (Japanese in Tokyo) subjects was obtained from the HapMap3 release 28 (http://hapmap.ncbi.nlm.nih.gov/), and data corresponding to the SNPs employed in the Genome-Wide Human SNP Array 6.0 were extracted. From these 314 subjects, we chose 261 subjects based on the number of SNPs genotyped (the number of successfully genotyped SNPs for the selected 261 subjects ranged between 707041 and 903804). These 261 subjects were randomly assigned as controls (200 subjects), as participants in a test population (20, 40, or 60 subjects), and a subject who served as the MRCA. The number of controls used was determined because 200 was the largest round number of controls that could be used. The number of the patients in the test population was determined so that the largest test population had 10 times the number of

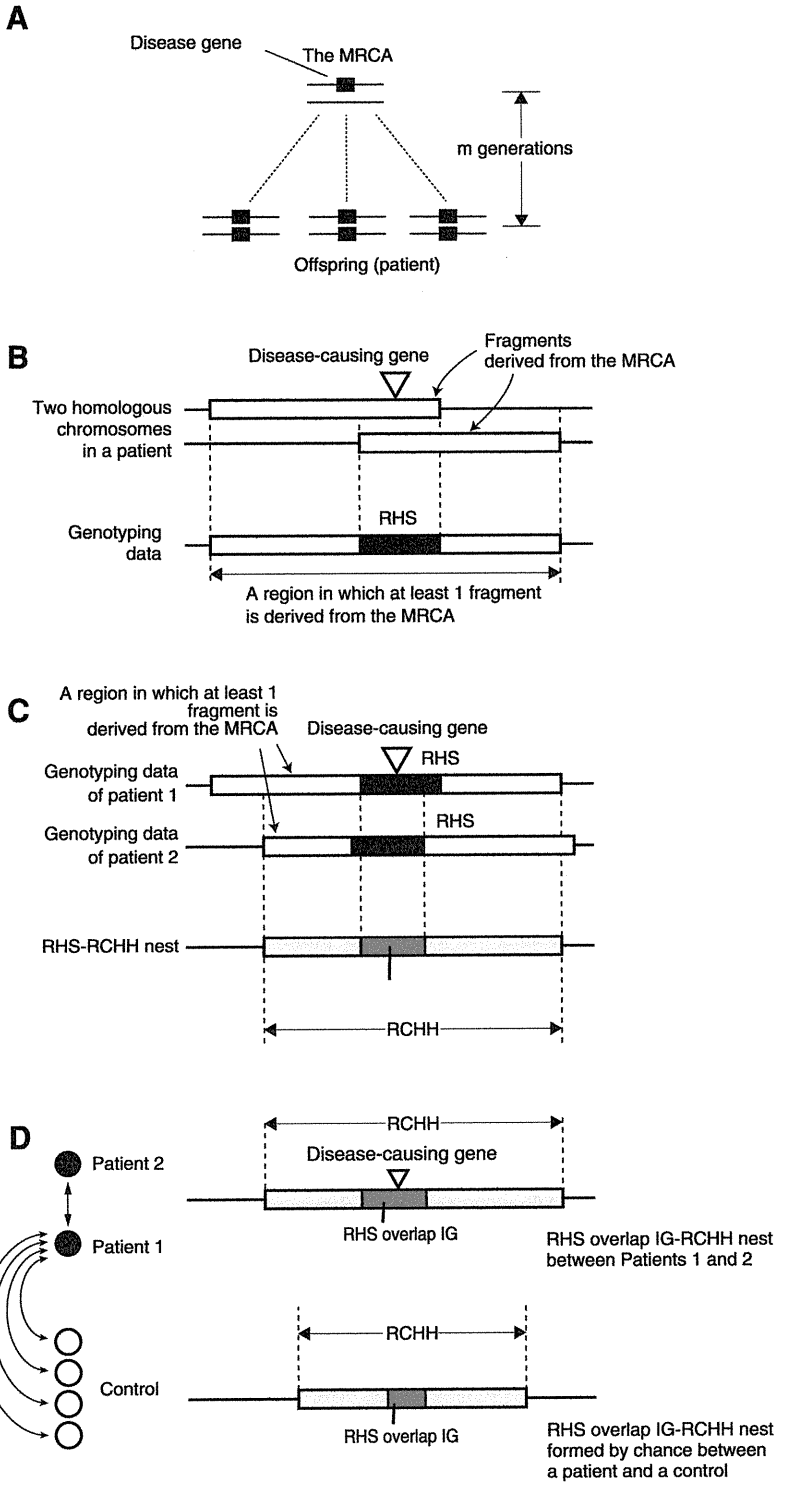


Figure 1. Structures formed by the fragments derived from the MRCA. (A) The MRCA has a single copy of the disease-causing gene. The gene is segregated to the patient through both the maternal and paternal lines, and thus the patients are homozygous for the disease-causing gene. (B) Each of the homologous chromosomes in the patient has a fragment derived from the MRCA. All SNPs in the overlap are homozygous, forming an RHS [10]. The union of the fragments generates "a region in which at least 1 fragment is derived from the MRCA." (C) Assume that there are 2 patients. The genotypes of these patients are identical in the RHS overlap, forming an RHS overlap IG. The overlap of "regions in which at least 1 fragment is derived from the MRCA" forms an RCHH [9]. This RCHH therefore contains the RHS overlap IG. This nested structure is hereby called an RHS overlap IG-RCHH nest. (D) The 2 patients are compared with subjects from a general population (controls). An RHS overlap IG-RCHH nest may be

formed between patients 1 or 2 and each of the controls due to a coincidence in the SNP genotype. However, the RHS overlap IG-RCHH nest between the patients is likely to be longer than any of the RHS overlap IG-RCHH nests accidentally formed between a patient and a control.
doi:10.1371/journal.pone.0025059.g001

patients with Siiyama-type α1-antitrypsin deficiency; it was believed that this number was suited for demonstrating the power of the analysis and for enabling an easy interpretation of the analysis results.

Genotyping

SNP genotyping was performed using the Genome-Wide Human SNP Array 6.0 (Affymetrix).

Synthetic data

The synthetic genotyping data of a patient who shared 2 IBD fragments that contain a disease-causing gene were made as follows: (i) A subject was randomly chosen from the 261 subjects (see above) to serve as the MRCA. (ii) An SNP was randomly chosen from an autosomal region and was considered to mark the position of the disease-causing gene. (iii) The range of the chromosomal region that contained the SNP and was inherited by the patient from the MRCA was calculated according to the Haldane's Poisson process model [13]. (iv) Step (iii) was repeated for the second fragment. (v) The genotyping data of the patient corresponding to the regions that were obtained at steps (iii) and (iv) were replaced with those of the MRCA.

Variables investigated in HM on HH analysis of a population

The variables investigated were the number of subjects in the test population (20, 40, and 60), proportion of patients in the test population (0, 5, 10, 15, 20, 25, and 30%), generational distance of the MRCA (20, 40, 60, 80, and 100 generations), and the gene frequency in the general population (0.0, 0.05, and 0.1). A gene frequency of 0.0 was considered to represent a rare variant, while gene frequencies of 0.05 and 0.1 were considered to represent common variants.

Computer program

The program was written in the Ruby programming language (http://www.ruby-lang.org/en/) with an extension library written in the C programming language (http://gcc.gnu.org/). The program was executed on a MacPro computer that ran on MacOS X 10.6.

Program

HM on HH program is available at Homozygosity Haplotype Analysis Web site, http://www.hhanalysis.com

Results

HM on HH analysis in patients with Siiyama-type α1-antitrypsin deficiency

We tested the performance of HM on HH analysis by using the SNP genotypes of 6 unrelated patients with Siiyama-type α1-antitrypsin deficiency, a rare autosomal recessive disease in Japan caused by a founder mutation of the *SERPINA1* gene (MIM 107400) [12]. As controls, we employed the genotypes of 200 Japanese individuals from the general population. The results obtained after each of the 4 steps that compose HM on HH analysis are shown for a pair of patients (**Figure 3A**). After the completion of the analysis, 2 closely located regions with a total length of 1.4 centimorgans were identified, 1 of which contained *SERPINA1* (**Figure 3A**). The results of the other 14 patient-pair combinations (note that ${}_6C_2 = 15$) were similar: each combination identified candidate regions (total length: 1.2 to 21.8 centimorgans, median: 1.6 centimorgans) that contained *SERPINA1*. Using the genotyping data of only 2 patients, HM on HH analysis was able to narrow the position of the disease-causing gene to a very short chromosomal interval.

HM on HH analysis of a pair of synthetic patients

We further examined the performance of HM on HH analysis of a pair of patients using synthetic data. We investigated the MRCA at 5 different generational distances (20, 40, 60, 80 and 100 generations). For each distance, we employed 60 randomly selected subjects, so that a total of 1770 pairs (${}_{60}C_2 = 1770$) were constructed. Each pair was investigated for 100 randomly selected SNP locations, which were assumed to be the location of a disease-causing gene. The number of trials was thus 177000 (1770 combinations × 100 SNPs) for each generational distance.

Detection of the region containing the RHS overlap depended on the generational distance of the MRCA (**Figure 3B**). This was a major limitation of HM on HH analysis: at the HM step, only RHSs that were longer in length than the RHS cutoff were detected (**Figure S1A**) [10]. The detection will be improved by genotyping more SNPs at a genomewide level, which will allow the use of a smaller RHS cutoff value (**Figure S1B**). Once an RHS overlap was detected at the HM step, HM on HH analysis rarely failed to track it (**Figure 3C**): for the MRCA that occurred 20 generations earlier, the RHS overlap was falsely excluded (false negative) in only 1.5% of the cases, while the falsely included areas (false positive) were reduced from 61.7 centimorgans after the HM step to 0.47 centimorgans after the completion of the HH step, indicating that a small false positive is a prominent feature of HM on HH analysis. Data for the other generations of the MRCA are presented in **Figure S2**.

HM on HH analysis of a population

HM on HH analysis of a population targets a population containing multiple patients sharing an IBD fragment (**Figure 4A**). This simulates a situation in which the population is a collection of patients with the same disease, and some of the patients share an IBD gene. We attempt to identify (1) a patient subgroup sharing an IBD fragment and (2) the chromosomal location of the shared IBD fragment. Here, we defined the analysis level: at analysis level n, the computer program searches for a subgroup consisting of n patients, any pair of which shares an IBD fragment at the same position on the chromosome (**Figure 4B**). To achieve the aims (1) and (2) as stated above, the program identifies (a) the topmost analysis level at which any subgroup is detected, (b) the members that are contained in the subgroup, and (c) the position of the IBD fragment on the chromosome.

First, we investigated the background signal that was detected in the general population (**Figure 4C**). For this purpose, we employed 260 normal subjects. Step (a): 260 normal subjects were randomly divided into a test population (60 subjects) and 200 controls. Step (b): HM on HH analysis of a population was performed. Steps (a) and (b) were repeated 500 times. The histogram of the topmost analysis level, at which any subgroup was detected (**Figure 4D**), demonstrated that a subgroup could be falsely detected (i.e., false positive) in the level 4 analysis and in an earlier analysis level. Conversely, when a positive result was

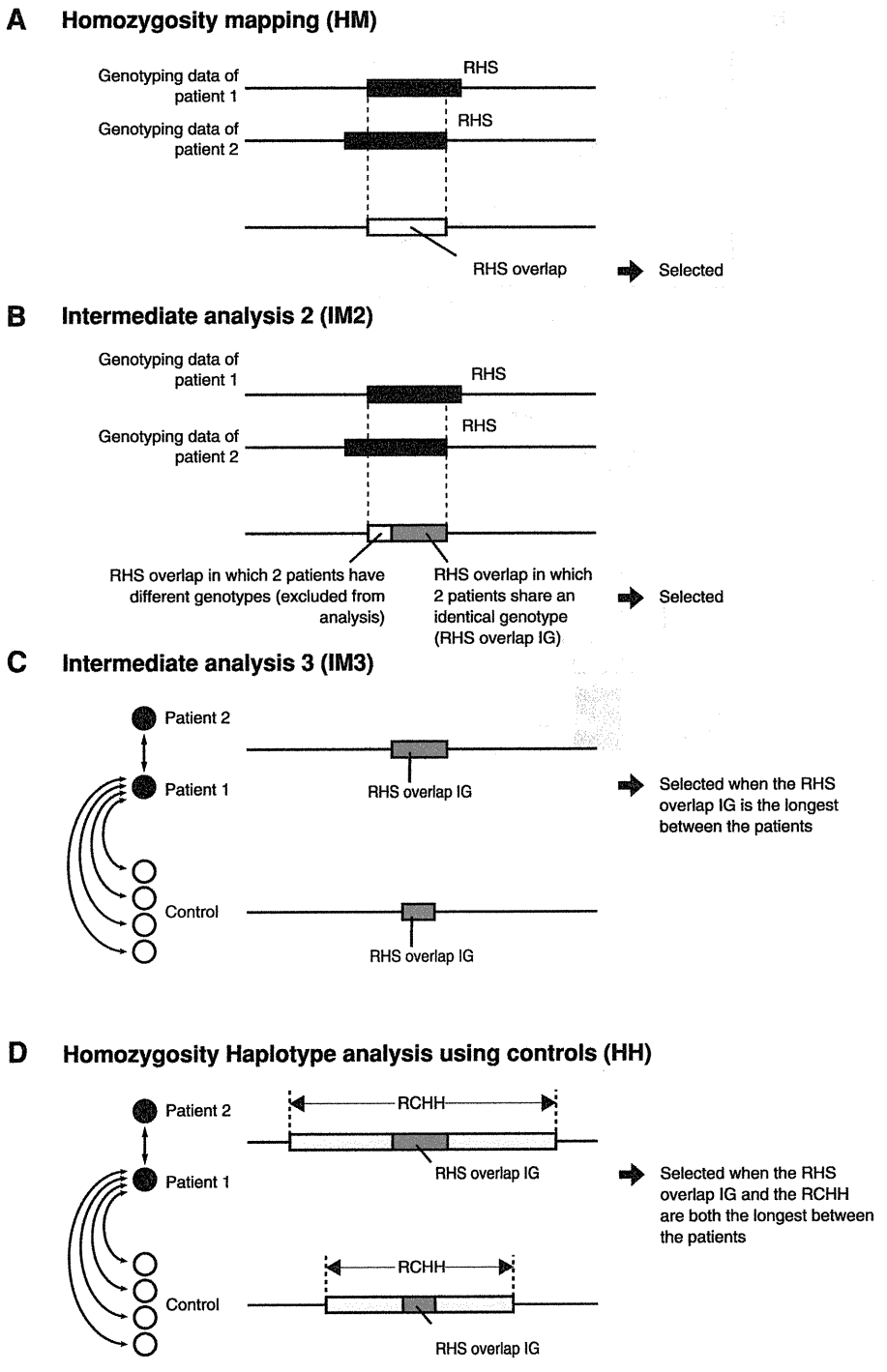


Figure 2. Steps of HM on HH analysis. HM on HH analysis is composed of 4 steps that are serially performed. (A) Homozygosity mapping (HM). The RHSs are determined for each patient, and the RHS overlaps are obtained. (B) Intermediate analysis 2 (IM2). The RHS overlap IGs are determined. (C) Intermediate analysis 3 (IM3). The RHS overlap IGs are compared. The RHS overlap IG is selected as a candidate region when the RHS overlap IG is the longest between the patients. (D) HH analysis using controls. RHS overlap IG-RCHH nests are compared. The RHS overlap IG is selected as a candidate region when the RHS overlap IG and the RCHH are both the longest between the patients.
doi:10.1371/journal.pone.0025059.g002

obtained in the level 5 analysis or in a later analysis, a subgroup sharing an IBD fragment was likely to be detected. Next, we investigated a test population comprising 6 unrelated patients with Siiyama-type α1-antitrypsin deficiency and 54 normal subjects (**Figure 4E**). A subgroup was detected at level 6 (**Figure 4F**); the members of the subgroup were the 6 patients with Siiyama-type

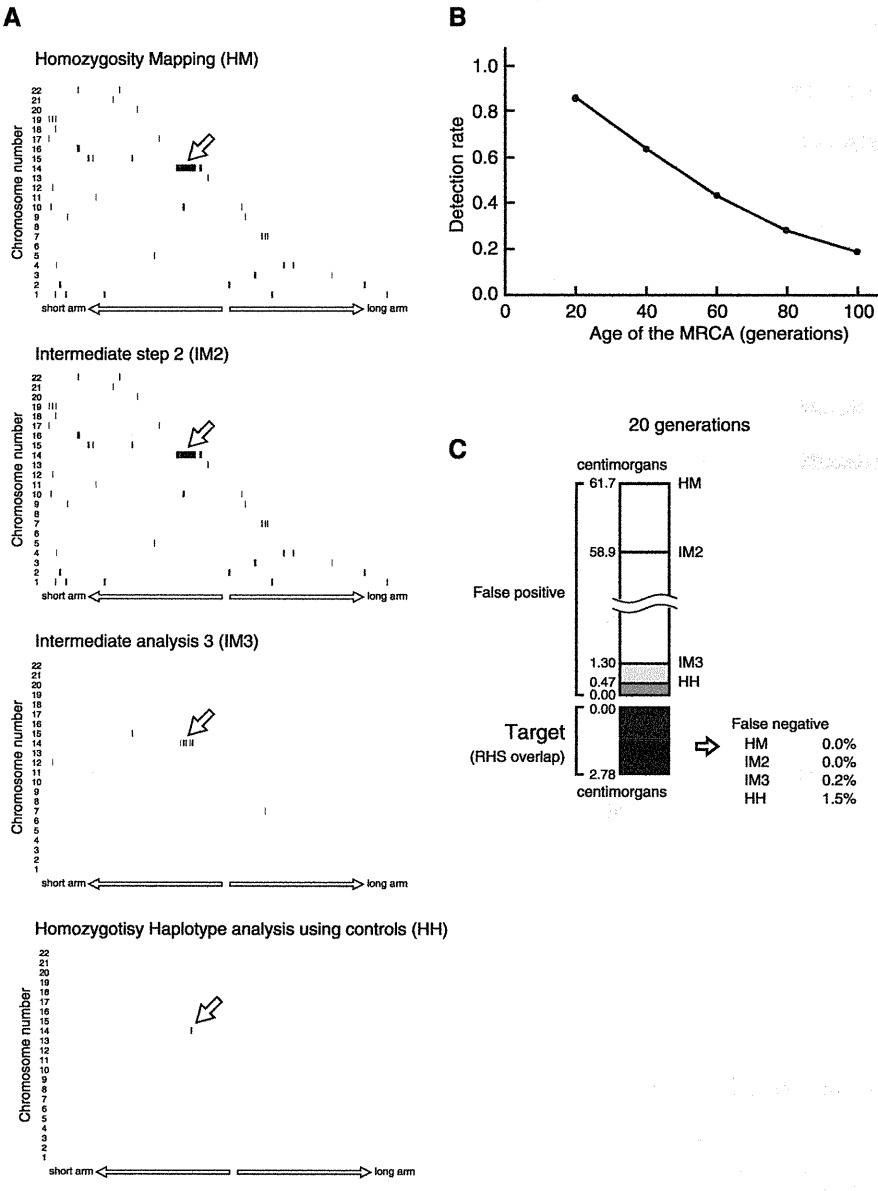


Figure 3. HM on HH analysis of a pair. (A) Analysis of patients 1 and 2 with Siiyama-type α1-antitrypsin deficiency. The position of the disease-causing gene (*SERPINA1*) is indicated by an arrow. HM on HH analysis is composed of 4 steps that are sequentially performed. The regions selected after each step are shown as black bands. The total length of the regions selected at the end was 1.36 centimorgans. (B) The rate at which the RHS overlap was detected by the HM step (i.e., the first step of the analysis) was the major determinant of HM on HH analysis. The detection rate will be improved by genotyping more SNPs genomewide. (C) False positives and false negatives for each analysis. False negatives are decreased with the progression of the analyses. False negatives are very few: 1.5% of the RHS overlap detected by the HM analysis is falsely excluded by HM on HH analysis.
doi:10.1371/journal.pone.0025059.g003

α1-antitrypsin deficiency. The candidate region, 1.2 centimorgans in width, was located on chromosome 14 and contained the *SERPINA1* gene. HM on HH accurately isolated a subpopulation that accounted for only 10% of the population and identified the position of an IBD fragment on the chromosome.

HM on HH analysis of a population containing synthetic patients

To study the performance of HM on HH analysis in more detail, we studied test populations containing synthetic patients. The synthetic patients (5, 10, 15, 20, 25, and 30% of the members of the population) were homozygous for the IBD fragment derived from MRCAs at generational distances of 20, 40, 60, 80, and 100 generations. For each combination of these parameters, the analysis was repeated 100 times by changing the disease-gene location, which was randomly selected from the SNP positions on the autosomes. The analysis was considered successful when (1) only a single candidate region was detected in the topmost level that detected any subgroup, and (2) the candidate region contained the locus of a disease-causing gene. The rates of successful trials (detection rate) were graphed for populations with 60, 40, and 20 subjects (**Figure 5A**). The results demonstrated

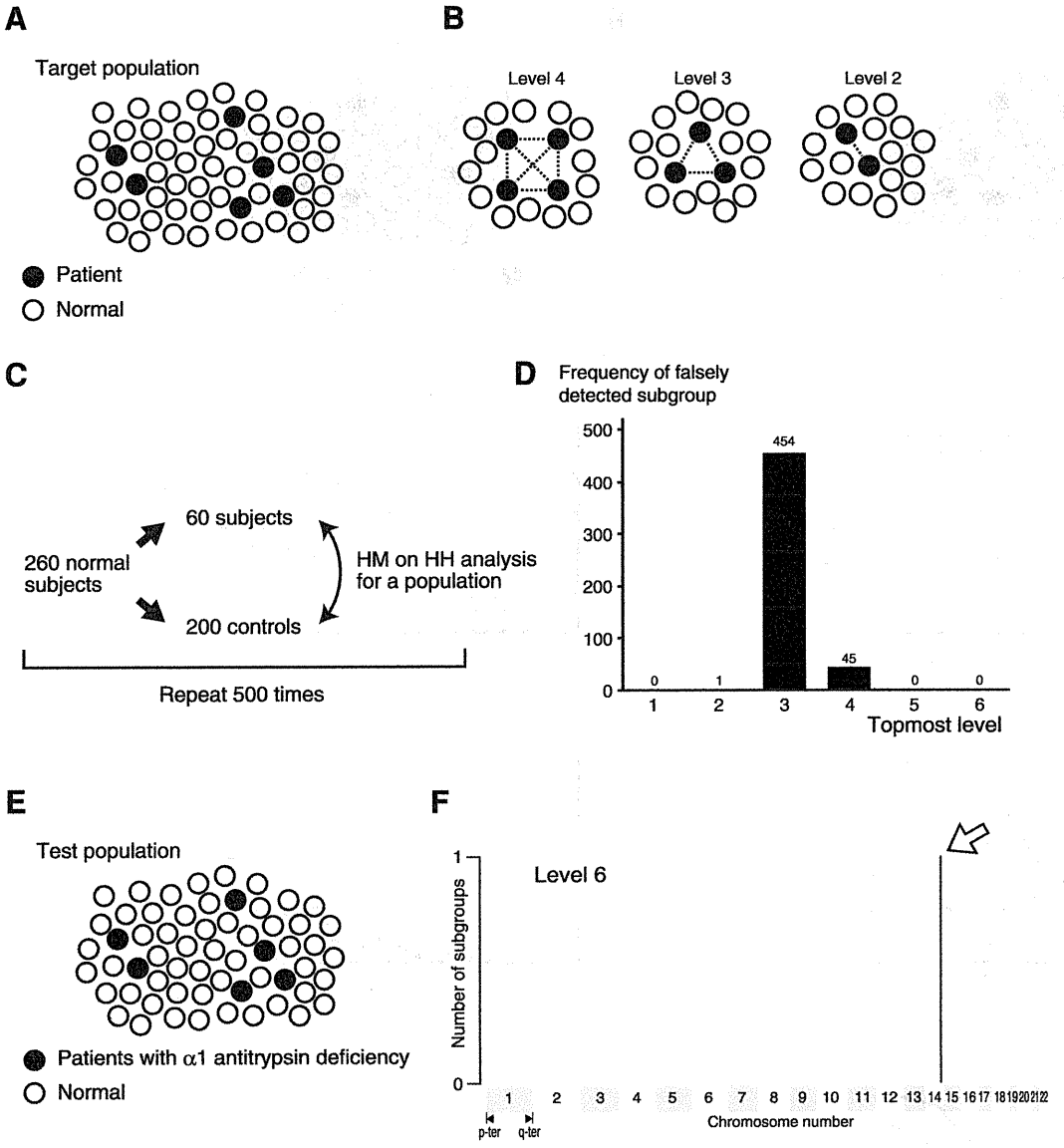


Figure 4. HM on HH analysis of a population. (A) Targets. Targets of HM on HH analysis of a population are populations containing multiple subjects sharing an IBD fragment. (B) Analysis level. At the level *n* analysis, a subgroup of *n* members, each pair of which shares an IBD fragment at the same position on the chromosome, are sought. (C) Scheme of the background determination. (D) Background. A subgroup(s) was falsely detected at level 3, 4, and 5 analyses. (E) A test population. The population is composed of 6 patients with Siiyama-type α1-antitrypsin deficiency (black circles) and 54 normal subjects (white circles). (F) Result. The horizontal position indicates the location on the autosomes, each of which is aligned from the p terminal (left side) to the q terminal (right side). A single subgroup was identified at the level 6 analysis, and the candidate region contained the *SERPINA1* gene. The members of the subgroup, which was the output on the computer console and thus is not shown here, were the 6 patients with Siiyama-type α1-antitrypsin deficiency.
doi:10.1371/journal.pone.0025059.g004

that HM on HH could identify a subpopulation sharing an IBD fragment that accounted for only a small fraction of the population.

The analysis described above assumed that the frequency of the allele containing the disease-causing gene was 0.0 in the general population. However, the disease-causing gene may be a common variant. We investigated the performance of HM on HH analysis when the frequency of the disease-causing gene in the general population was 0.05 or 0.1 (**Figure 5B**). The results indicated that the performance was severely degraded for a frequency of 0.1. HM on HH analysis was considered to work well for a frequency <0.1. Therefore, the HM on HH analysis targets a recessive gene that is the cause of a disease, in which less than 1% of the people in the general population are homozygous for the gene and thus may suffer from the disease somewhere in their lifetime.

Analysis without utilizing HH information

Analyses similar to HM on HH analysis may be performed by stopping the analysis after the HM, IM2, or IM3 steps (**Figure 3A**). When stopping after either the HM or the IM2,

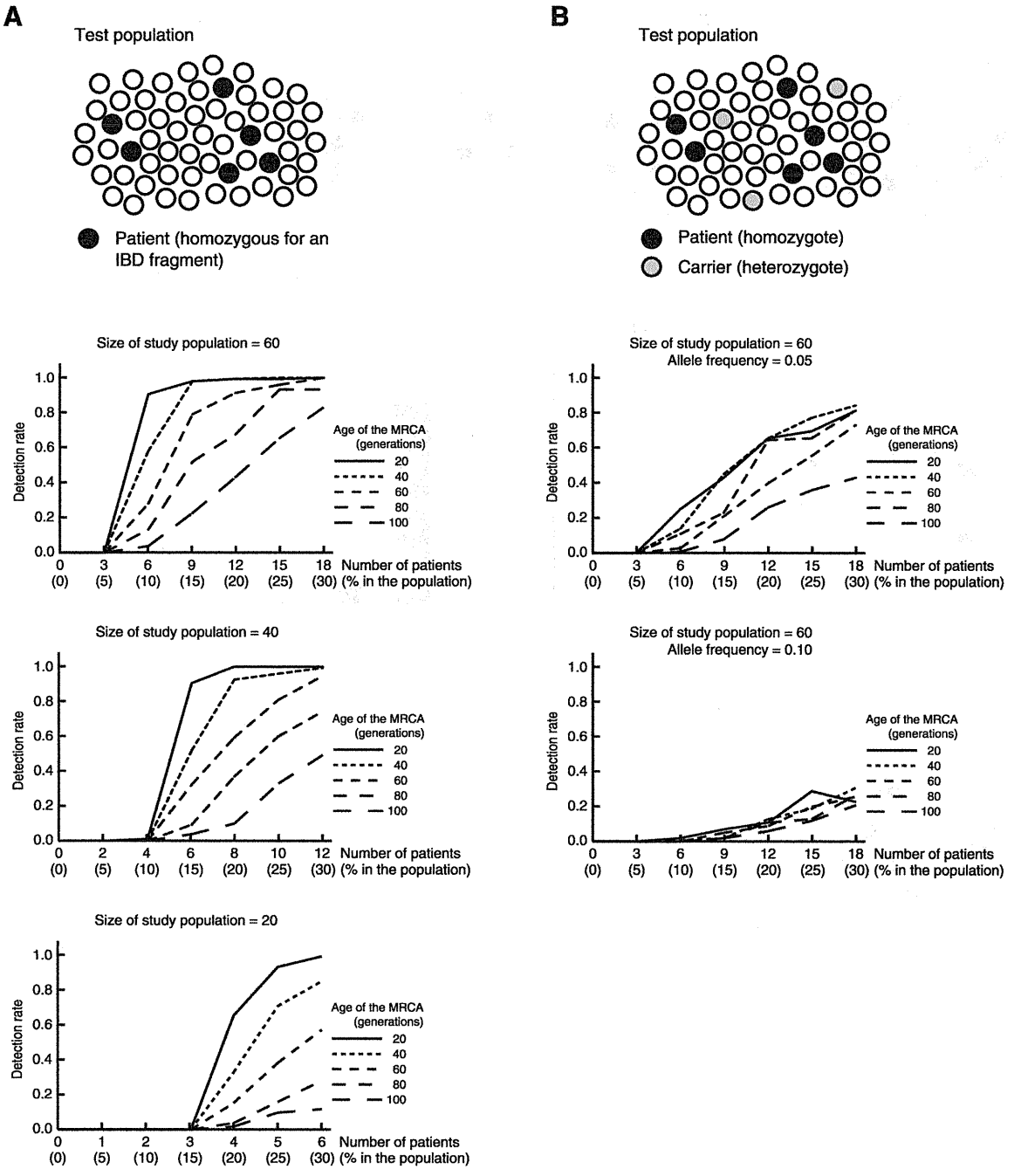


Figure 5. HM on HH analysis of a population performed on populations containing synthetic patients. (A) Scheme of the analysis and Result. The test population is composed of patients homozygous for a gene identical by descent (black circles) and subjects who do not have the gene (white circles). The horizontal line of each graph indicates the number (percentage) of patients homozygous for the gene in the test population. (B) Scheme of the analysis and Result for a gene that is widely shared in the general population. The test population is composed of patients homozygous for a gene identical by descent (black circles), subjects heterozygous for the gene (gray circles), and subjects who do not have the gene (white circles). In the case that the gene was widely shared in the general population (the frequency of the allele of the disease-causing gene was 0.05 and 0.1), the detection rate was decreased.
doi:10.1371/journal.pone.0025059.g005

the program ran out of memory from the explosion in the numbers of subgroups that resulted from a large amount of false positives. When stopping after the IM3 step, the detection rate was much lower than that after the HH analysis, because of a greater amount of false positives (**Figure s3**). These results demonstrated that the small amount of false positives attained by the HH step is important for the performance of HM on HH analysis of a population.

Discussion

In the current report, we demonstrated that HM on HH is able to narrow the candidate region for a disease-causing gene to a very small chromosomal interval either by employing 2 outbred patients sharing an IBD fragment, or by using a small population in which 10%≤ of the patients share an IBD fragment. Haplotype information obtained from the region that flanks the RHSs was the component of the HM on HH analysis that enabled them. By using the HM on HH analysis, genes with a recessive trait are exploited in the very early stage of a project attempting to identify a disease-causing gene.

It has been reported that HM is able to identify a candidate region from as few as 3 inbred patients [5,8]. Although this number is small, the clinical characteristics often do not provide information sufficient for selecting 3 patients who may have an IBD gene. Furthermore, the total length of the candidate regions detected by HM is usually large [8,11], which necessitate an enormous effort for an in-depth search of the regions. HM on HH analysis offers the advantage of being capable of using only 2 patients to obtain a relatively narrow candidate region, typically it is a few centimorgans in length. This may enable novel strategies for identifying disease-causing genes. One such strategy is to collect several patients who are likely to share a fragment IBD, identify the candidate regions by a pairwise comparison, and scrutinize all of these regions by high-throughput sequencing [14].

A small number of founder mutations often largely accounts for the occurrence of a recessive disease or its predisposition. Examples are α1-antitrypsin deficiency and cystic fibrosis in Europeans [15,16], and Gaucher disease and Tay-Sachs disease in Ashkenazi Jews [17]. The cause of the prevalence may be heterozygote advantage, a founder effect, or genetic drift [17,18]. Whatever is the cause, this suggests that the predominance of a limited number of founder mutations is worth taking into consideration in an attempt to search for disease-causing genes with a recessive trait. HM on HH analysis is suitable for pursuing the possibility.

The generational distance of the MRCA has a major effect on the performance of HM on HH analysis. In the analysis of pair of patients, the effect was large (**Figure 3B**). In the analysis of population, the effect was moderate (**Figure 5A**). Use of arrays with a greater number of SNPs will accomplish a better performance (**Figure S1B**). Data obtained using SNP Array 6.0 were investigated in the current study; they were considered suitable for the MRCA with a generational distance ≤60 generations (**Figures 3B and 5A**). Founder populations that settled in recent centuries are amenable to the analysis. These include the French-Canadian population that settled in Quebec in the 17th century [19,20], or the Icelandic population that was founded in the 10th century [21], because the generational distance of the MRCA may be less than 20–60 generations in many diseases.

Isolated populations may also be suitable for this analysis; in such population, a single IBD gene from an MRCA existed in a recent generation may predominate among patients with a specific disease. In many countries, there may be many geographical areas in which MRCAs for a disease-causing gene have a generational distance of 20–60 generations. A small number of patients that HM on HH analysis requires will make the analysis easily performable in small populations from such areas.

Inclusion of the subjects who share the IBD fragment degrades the performance of the analysis. The frequency of the gene in the control should be less than 0.1, i.e., less than 20% of the control subjects may be heterozygous for the gene and less than 1% of the control subjects may be homozygous for the gene. The analysis is not suitable for the common variants for the common diseases that are often the targets of the genome-wide association studies.

The calculation time of the HH on HH analysis is short. The analysis of a pair of patients is completed in a fraction of a second; the analysis for a test population of 60 subjects is completed in 15 seconds. Theoretically, an analysis of a study population of 60 subjects requires an investigation of 1.15×10^{18} subgroups. However, many of the comparison of 2 patients generate a result without any candidate region, and thus eliminate the need for investigating any subgroups containing a given pair. A small amount of false positive of HM on HH analysis enables an exhaustive search for the subgroups.

We used 200 controls in the current study, but it is possible to decrease this number with minimal loss in performance. When the analysis was performed with 100 controls, we found that the performance was only mildly decreased. Moreover, the International HapMap3 project (see International HapMap project Web page) has genotyped and released about 100 or more subjects for each of the 10 ethnic groups, and these data may be used for controls.

The RHS overlap IG-RCHH nest was selected when both the length of the RHS overlap IG and the length of the RCHH between the patients were both at the top. The criteria may be weakened to "top 1%," "top 10%," etc. However, we found that the condition of "at the top" worked best for almost all cases (data not shown). The current criterion is thus considered good for HM on HH analysis.

The RHS cutoff for the Genome-Wide Human SNP Array 6.0 was selected so that the total length of the false-positive RHS was acceptable (1.5 centimorgans per a patient). The equivalent RHS cutoff values for other high-density SNP arrays are 0.75 centimorgans for the Human Omni2.5 BeadChip (Illumina), 1.1 centimorgans for the Human1M-Duo BeadChip (Illumina), and 1.9 centimorgans for the GeneChip Human Mapping 500K Array Set (Affymetrix) [10].

In conclusion, HM on HH analysis used genetic information on both the RHS and the flanking regions, and thus detected the locus for a recessive, disease-causing genes with a very low background from a small number of patients. HM on HH analysis will accelerate the elucidation of the genetic causes of many diseases.

Supporting Information

Figure S1 Errors in the HM. (A) The false positive rate is the ratio of the total length of RHSs that are falsely detected along the entire length of the autosomes. The false negative rate is the ratio of the total length of the autozygous segments (i.e., chromosomal regions in which both chromosomal fragments are IBD) that fail to be detected as RHSs along the total length of the autozygous segments. The false positive rate is dependent on the kind of high-density array and thus is shown for each array. 2.5 M, Human Omni2.5 BeadChip (Illumina); 1 M, Human1M-duo BeadChip (Illuina); SNP6.0: Genome-Wide Human SNP Array 6.0 (Affymetrix); 500K, 500K GeneChips Mapping Array Set (Affymetrix). m: the age of the MRCA. (B) Detection rates for each array. The figure corresponds to **Figure 3B**; this figure summarizes the theoretical calculation, while the result in **Figure 3B** is the result using the actual genotyping data.
(EPS)

Figure S2 Errors in HM on HH for a pair of patients. The data corresponding to those of **Figure 3C** for the MRCAs with a generational distance of 40, 60, 80, and 100 generations.
(EPS)

Figure S3 Result of the analysis stopped after the IM3 step. (A) Background. The background was observed to a higher analysis level than that for the HH analysis (compare with **Figure 4D**). (B) Detection rate. Positive results obtained at a level 7 analysis or higher were considered successful. (C) Detection rate. Positive results obtained at a level 9 analysis or higher were considered successful. Figures (B) and (C) correspond with those shown in **Figure 5A**. In both conditions, the detection rate was lower than those shown in **Figure 5A**.
(EPS)

Acknowledgments

We thank Prof. Tokunaga and Dr. Nao Nishida, Department of Human Genetics, Graduate School of Medicine, University of Tokyo, for providing SNP data for 198 Japanese individuals.

Author Contributions

Conceived and designed the experiments: KH H. Morino HK. Performed the experiments: JS TT H. Miyazawa TS. Analyzed the data: MK YO. Contributed reagents/materials/analysis tools: KS. Wrote the paper: KH MK HK.

References

1. Manolio TA (2010) Genomewide association studies and assessment of the risk of disease. N Engl J Med 363: 166–176.
2. Ott J (1999) Analysis of Human Genetic Linkage. Baltimore, MD: Johns Hopkins University Press.
3. Hardy J, Singleton A (2009) Genomewide association studies and human disease. N Engl J Med 360: 1759–1768.
4. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, et al. (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 9: 356–369.
5. Lander ES, Botstein D (1987) Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. Science 236: 1567–1570.
6. Hildebrandt F, Heeringa SF, Ruschendorf F, Attanasio M, Nurnberg G, et al. (2009) A systematic approach to mapping recessive disease genes in individuals from outbred populations. PLoS Genet 5: e1000353.
7. Browning SR, Browning BL (2010) High-resolution detection of identity by descent in unrelated individuals. Am J Hum Genet 86: 526–539.
8. Huqun, Izumi S, Miyazawa H, Ishii K, Uchiyama B, et al. (2007) Mutations in the SLC34A2 gene are associated with pulmonary alveolar microlithiasis. Am J Respir Crit Care Med 175: 263–268.
9. Miyazawa H, Kato M, Awata T, Kohda M, Iwasa H, et al. (2007) Homozygosity haplotype allows a genomewide search for the autosomal segments shared among patients. Am J Hum Genet 80: 1090–1102.
10. Huqun, Fukuyama S, Morino H, Miyazawa H, Tanaka T, et al. (2010) A quantitatively-modeled homozygosity mapping algorithm, qHomozygosityMapping, utilizing whole genome single nucleotide polymorphism genotyping data. BMC Bioinformatics 11 Suppl 7: S5.
11. Maruyama H, Morino H, Ito H, Izumi Y, Kato H, et al. (2010) Mutations of optineurin in amyotrophic lateral sclerosis. Nature 465: 223–226.
12. Seyama K, Nukiwa T, Souma S, Shimizu K, Kira S (1995) Alpha 1-antitrypsin-deficient variant Siiyama (Ser53[TCC] to Phe53[TTC]) is prevalent in Japan. Status of alpha 1-antitrypsin deficiency in Japan. Am J Respir Crit Care Med 152: 2119–2126.
13. Haldane J (1919) The combination of linkage values, and the calculation of distances between the loci of linked factors. J Genet 8: 299–309.
14. Berg JS, Evans JP, Leigh MW, Omran H, Bizon C, et al. (2011) Next generation massively parallel sequencing of targeted exomes to identify genetic mutations in primary ciliary dyskinesia: implications for application to clinical testing. Genet Med 13: 218–229.
15. Janciauskiene SM, Bals R, Koczulla R, Vogelmeier C, Kohnlein T, et al. (2011) The discovery of alpha1-antitrypsin and its role in health and disease. Respir Med 105: 1129–1139.
16. Salvatore D, Buzzetti R, Baldo E, Forneris MP, Lucidi V, et al. (2011) An overview of international literature from cystic fibrosis registries. Part 3. Disease incidence, genotype/phenotype correlation, microbiology, pregnancy, clinical complications, lung transplantation, and miscellanea. J Cystic Fibrosis 10: 71–85.
17. Charrow J (2004) Ashkenazi Jewish genetic disorders. Familial Cancer 3: 201–206.
18. Dean M, Carrington M, O'Brien SJ (2002) Balanced polymorphism selected by genetic versus infectious human disease. Annu Rev Genomics Hum Genet 3: 263–292.
19. Heyer E, Tremblay M (1995) Variability of the genetic contribution of Quebec population founders associated to some deleterious genes. Am J Hum Genet 56: 970–978.
20. Laberge AM, Michaud J, Richter A, Lemyre E, Lambert M, et al. (2005) Population history and its impact on medical genetics in Quebec. Clin Genet 68: 287–301.
21. Williams JT (1993) Origin and population structure of the Icelanders. Hum Biol 65: 167–191.

ORIGINAL ARTICLE

Identification of FGF7 as a novel susceptibility locus for chronic obstructive pulmonary disease

John M Brehm,[1] Koichi Hagiwara,[2] Yohannes Tesfaigzi,[3] Shannon Bruse,[3] Thomas J Mariani,[4] Soumyaroop Bhattacharya,[4] Nadia Boutaoui,[1] John P Ziniti,[5] Manuel E Soto-Quiros,[6] Lydiana Avila,[6] Michael H Cho,[5,7,8] Blanca Himes,[5] Augusto A Litonjua,[5,7,8,9] Francine Jacobson,[10] Per Bakke,[11] Amund Gulsvik,[11] Wayne H Anderson,[12] David A Lomas,[13] Erick Forno,[14] Soma Datta,[5] Edwin K Silverman,[5,7,8,15] Juan C Celedón[1]

► Additional materials are published online only. To view these files please visit the journal online (http://thorax.bmj.com).

For numbered affiliations see end of article.

Correspondence to
Dr Juan C Celedón, Division of Pediatric Pulmonary Medicine, Allergy and Immunology, Children's Hospital of Pittsburgh of UPMC, 4401 Penn Avenue, Pittsburgh, PA 15224, USA; juan.celedon@chp.edu

Received 11 February 2011
Accepted 5 August 2011
Published Online First
15 September 2011

ABSTRACT

Rationale Traditional genome-wide association studies (GWASs) of large cohorts of subjects with chronic obstructive pulmonary disease (COPD) have successfully identified novel candidate genes, but several other plausible loci do not meet strict criteria for genome-wide significance after correction for multiple testing.

Objectives The authors hypothesise that by applying unbiased weights derived from unique populations we can identify additional COPD susceptibility loci.

Methods The authors performed a homozygosity haplotype analysis on a group of subjects with and without COPD to identify regions of conserved homozygosity haplotype (RCHHs). Weights were constructed based on the frequency of these RCHHs in case versus controls, and used to adjust the p values from a large collaborative GWAS of COPD.

Results The authors identified 2318 RCHHs, of which 576 were significantly ($p<0.05$) over-represented in cases. After applying the weights constructed from these regions to a collaborative GWAS of COPD, the authors identified two single nucleotide polymorphisms (SNPs) in a novel gene (fibroblast growth factor-7 (*FGF7*)) that gained genome-wide significance by the false discovery rate method. In a follow-up analysis, both SNPs (rs12591300 and rs4480740) were significantly associated with COPD in an independent population (combined p values of 7.9E−7 and 2.8E−6, respectively). In another independent population, increased lung tissue *FGF7* expression was associated with worse measures of lung function.

Conclusion Weights constructed from a homozygosity haplotype analysis of an isolated population successfully identify novel genetic associations from a GWAS on a separate population. This method can be used to identify promising candidate genes that fail to meet strict correction for multiple testing.

Key messages

What is the key question?

► Can information from isolated populations improve our ability to detect novel genetic variants in genome-wide association studies (GWASs)?

What is the bottom line?

► We identified statistically significant polymorphisms in a novel chronic obstructive pulmonary disease (COPD) gene (*FGF7*), which we replicated in an independent population.

Why read on?

► We demonstrate the use of a novel method (homozygosity haplotype analysis) for identifying genomic regions that are inherited from a common ancestor, and use this information to weight a GWAS of COPD to identify novel genetic variants that are associated with increased risk of disease.

INTRODUCTION

Traditional genome-wide association studies (GWASs) have identified novel susceptibility loci for complex diseases such as chronic obstructive pulmonary disease (COPD).[1–3] Because the effect size of most common disease-susceptibility variants is modest, GWASs of complex diseases require large sample sizes to achieve statistically significant results after correction for multiple testing. Weighting the results of GWASs according to prior information (eg, from linkage studies) may significantly improve the power to detect associations that do not meet genome-wide (GW) significance.[4]

Homozygosity mapping is a promising technique to identifying regions of the genome that are more likely to contain disease-susceptibility loci. Although initially developed to identify rare susceptibility mutations for monogenic traits in families,[5] homozygosity mapping has recently been successfully applied to the study of complex diseases.[6 7] While techniques vary, the concept underlying all homozygosity haplotype (HH) methods is that regions of homozygosity are more likely to contain disease-susceptibility loci in affected subjects than in unaffected individuals.[8]

Using high-density single nucleotide polymorphism (SNP) arrays, Miyazawa *et al* developed a novel variation of homozygosity mapping that tests whether multiple subjects share the same genotype among homozygous SNPs, and then constructed a region of conserved homozygosity

haplotype (RCHH) that reflects the transmission of the haplotype from a founder population. In theoretical simulations, this method was shown to be a viable method to detect disease-susceptibility loci in recently admixed populations.[9] We hypothesised that application of this method to a genetic isolate in Costa Rica would result in detection of an over-representation of regions of conserved homozygosity in subjects affected with COPD compared with unaffected subjects. In this report, we first identify regions of conserved homozygosity in Costa Ricans and then show that weights derived from these regions can be applied to GWASs in non-isolated populations to identify novel disease-susceptibility loci for COPD. Using this approach, we identify a novel COPD candidate gene (fibroblast growth factor-7 (*FGF7*)).

MATERIALS AND METHODS

Study population

The primary study population consisted of 58 subjects with COPD (cases) and 57 subjects without COPD (controls) in the Genetic Epidemiology of COPD in Costa Rica study. Cases were recruited from patients attending four adult hospitals in San José (Costa Rica) and their affiliated clinics, and through newspaper advertisements. Control subjects were recruited from individuals attending a smoking-cessation clinic at the Institute for Pharmaco-dependency in San José, and through newspaper advertisements. To ensure their descent from the founder population of the Central Valley of Costa Rica (which is predominantly of Spanish and Native American ancestry), all participants were required to have at least six great-grandparents born in the Central Valley. Additional inclusion criteria for cases were ages 21–71 years, physician-diagnosed COPD, ≥10 pack-years of cigarette smoking, a forced expiratory volume in one second (FEV_1) ≤65% predicted and an FEV_1/forced vital capacity (FVC) ratio of ≤70% after bronchodilator administration (180 μg of albuterol by metered dose inhaler). Controls were recruited on the basis of the same criteria for age and smoking history, but they had to have no physician-diagnosed COPD and normal spirometry. Exclusion criteria for cases and controls included history of chronic pre-existing chronic lung disease (eg, bronchiectasis) and severe α-1-antitrypsin deficiency (for cases), based on molecular phenotyping. The baseline characteristics of this cohort are in listed in the online supplementary table 2.

Written consent was obtained from participating subjects. The study was approved by the institutional review boards of the Hospital Nacional de Niños (San José, Costa Rica), Partners Healthcare System (Boston, Massachusetts, USA), and participating National Emphysema Treatment Trial (NETT), Evaluation of COPD Longitudinally to Identify Predictive Surrogate Endpoints (ECLIPSE) and Norway centres.

Genotyping of Costa Rican cohort

High-density SNP genotyping was performed using the Illumina Quad 610 platform at the Channing Laboratory, Boston, Massachusetts, USA. Cases and controls were randomly distributed among batches, and each batch contained a replicate sample. All subjects had an SNP call rate >95%. After quality control measures (see online supplementary table 1), a total of 558 929 SNPs were acceptable for analysis.

Collaborative COPD cohorts for the primary GWAS

Three populations with a total of 2940 cases and 1380 controls were used for the primary GWAS: (1) subjects in a case–control study of COPD in Norway (838 cases and 791 controls)[3]; (2) subjects in the NETT (366 cases) and the Normative Aging Study (414 controls)[10 11] and (3) 1736 cases and 175 controls from the multicentre ECLIPSE study.[12] All controls were current or former smokers with normal spirometry, and all cases with COPD had moderate to very severe disease according to the Global Initiative for Chronic Obstructive Lung Disease classification.[13]

Lovelace Smokers Cohort

The top SNPs in novel genes were replicated in a cohort of 1845 smoking adults in New Mexico, 424 (23%) of whom were classified with COPD based on an FEV_1/FVC ratio below the fifth percentile of the predicted value, also referred to as the lower limit of normal.[14] Of the 1845 participants, 1411 (77%) were Caucasian and 313 (17%) were Hispanic. The protocols for subject recruitment and data collection for the Lovelace Smokers Cohort have been previously described in detail.[15] The two SNPs (rs12591300 and rs4480740) were genotyped by allelic discrimination using Taqman assay (Applied Biosystems, Foster City, California, USA). The case–control association analysis was first performed in all subjects, and then separately in Caucasians and Hispanics. All analyses were adjusted for age, gender and pack-years of cigarette smoking; the analysis of all subjects was additionally adjusted for self-declared ethnicity.

Gene expression analysis

For the top novel candidate genes, we examined the correlation of gene expression in lung tissue with COPD intermediate phenotypes (FEV_1 and FEV_1/FVC ratio) in a previously published COPD biomarker discovery study.[16] This cohort consists of 56 subjects with varying degrees of obstruction who underwent lung resection for a solitary pulmonary nodule. RNA expression profiling was completed using the Affymetrix U133 Plus 2.0 array, as previously described.[16] Expression correlation with quantitative phenotypes was conducted as previously described.[16]

Statistical analysis

Construction of RCHHs

RCHHs were identified using the method described by Miyazawa *et al*.[9] In brief, for any given individual all heterozygous SNPs were ignored and the SNP location was scored with the value of the allele for that subject. Subjects are compared only across SNPs that are scored. RCHHs are defined by runs of SNPs that share the same allele at the homozygous locations across multiple subjects, ignoring heterozygous SNPs. The size of the shared segments between any two individuals was set at 3.0 cM (roughly and approximately three million base pairs), which in theoretical work conducted by Miyazawa *et al*[9] reduced the false positive and false negative rates of discovery. A theoretical ancestral segment was then constructed from the largest subgroup of subjects sharing a particular RCHH (see online supplementary figure 1). While any two subjects must have at least 3.0 cM of sharing, the size may be much smaller when comparing across multiple subjects (online supplementary figure 2). If more than one ancestral region is identified at a particular chromosomal location, the region shared by the most number of subjects is used (online supplementary figure 3). The total number of cases and controls sharing this ancestral allele is used to calculate a p value based on a standard normal distribution.

For the primary analysis of the collaborative COPD cohort, logistic regression analysis was performed under an additive genetic model for each SNP, adjusting for age, pack-years of smoking and the first 16 principal components (to adjust for population stratification). The p values from all RCHHs identified in Costa Rica were then used to construct a cumulative weight for each SNP from the recent GWAS of COPD in the combined cohort of Norway, ECLIPSE and NETT–Normative

Aging Study using the method developed by Roeder *et al*.[4] Briefly, the weighting method utilises prior information (in this case, the p value representing the degree of over-representation of a region of the genome in cases versus controls) to upweight or downweight p values from an association study (in this case, the GWAS of COPD in the collaborative cohort). In order to maintain an overall α level of 0.05, the assigned weights across the genome average to 1. For this study, SNPs that did not fall inside of an RCHH (and therefore did not have a p value) were assigned a p value of 1 (and therefore a weight approaching zero). This is a more conservative approach than excluding these SNPs from consideration. The method then calculates a false-discovery rate (FDR) using the method described by Benjamini and Hochberg[17] to correct for multiple testing.

The RCHHs were created and compared with HHAnalysis (available at http://www.hhanalysis.com). Association analysis was performed using PLINK V.1.07 (http://pngu.mgh.harvard.edu/purcell/plink). The weighting procedure was performed using software developed by Roeder *et al*[4] (http://wpicr.wpic.pitt.edu/wpiccompgen/). All other statistical analysis was performed using R V.2.9.0 (http://www.R-project.org).

RESULTS

Identification of RCHHs in Costa Rica and construction of weights

In total, 2318 RCHHs were identified in the Costa Rican cohort. Of these 2318 regions, 576 were significantly ($p<0.05$) over-represented in cases compared with controls; none of the regions were significantly more frequent in controls than cases. The median size of the significant regions was 105 kb, and the largest was 7.2 Mb. Online supplementary table 3 shows the top 20 p values representing 100 RCHHs in Costa Rica.

Each SNP in the combined collaborative COPD cohort was then mapped to an RCHH and assigned the p value of the whole region. SNPs that did not map to an RCHH were assigned a p value of 1. The mapped p values across all genotyped SNPs were then used to create weights using a cumulative distribution function. The algorithm is constructed so that the mean weight across all SNPs is 1: some SNPs are upweighted and a much larger fraction is downweighted. The nominal p value is divided by the weight to obtain the weighted p value.

Application of weights to the COPD GWAS

We applied the weights derived from the HH analysis above to reanalyse GW genotypic data in a cohort of subjects of European descent that was previously employed for a traditional GWAS of COPD. After weighting, 14 SNPs were significant at an FDR-corrected α of 0.05. The top five SNPs from the unweighted GWAS retained their original ranks, but several SNPs that did not achieve GW significance in the traditional GW association analysis became more statistically significant and moved higher in the list (table 1). Of these SNPs, those in the gene for *FAM13A* were identified in the original analysis of the GWAS,[1] and SNPs in *IREB2*[18] and *CHRNA3*[3] have been implicated in COPD affection status in prior candidate-gene and GWASs. Two of the other SNPs lie in two novel candidate genes for COPD, *FGF7* and proteasome subunit, α-type, 4 (*PSMA4*) (figure **1**). The RCHH in Costa Rica that contains *FGF7* was present in seven cases and no controls, and the RCHH containing *PSMA4* was present in five cases and no controls.

The regions containing the genes *CHRNA3* and *IREB2* were also over-represented in cases compared with controls ($p<0.05$), and after weighting they were GW significant by FDR. While there was an RCHH containing *FAM13A* identified in the Costa Rican cohort, it was only seen in one case and no controls.

Replication in Lovelace Smokers Cohort

The top two SNPs in or near *FGF7* were genotyped in the 1845 smoking adults in the Lovelace Smokers Cohort. The minor alleles of both SNPs conferred increased odds for COPD in the whole population in the same direction as the original collaborative COPD cohort (table 2). Among the Hispanic subgroup, the effect size was larger and in the same direction for both SNPs, but only rs12591300 showed a significant association with COPD affection status.

Gene expression analysis

Our previous studies indicate that gene expression patterns associated with quantitative, intermediate COPD phenotypes are most informative for the discovery of disease-associated genes.[16 18 19] We examined disease-associated expression patterns for our novel candidate genes in a previously published GW expression data set from 56 subjects with varying degrees of airflow obstruction (assessed by spirometric measures of lung function (FEV_1 and FEV_1/FVC ratio)).[16] Expression of *FGF7* (as defined by multiple and independent probe sets) was significantly negatively correlated with both FEV_1 (nominal p value <0.01) and FEV_1/FVC ratio (nominal p value <0.01), indicating increased expression associated with increased disease severity. Expression in COPD subjects was increased compared with control subjects, but the difference was not statistically significant. *PSMA4* expression was not correlated with lung function and was not differentially expressed in cases versus controls.

DISCUSSION

While successful in identifying novel candidate genes, GWASs of complex traits are unlikely to identify all potential common disease-susceptibility variants because of limited power if strict criteria for GW significance are applied. In the absence of a very large sample size, novel methods are needed to identify disease-susceptibility variants not meeting GW significance. We identified RCHHs for COPD in a GW case–control study in Costa Rica. After applying a weighting method based on the degree of significance of these regions to a GWAS of COPD cases and controls of European descent, we identified two SNPs in a novel candidate gene for COPD (*FGF7*) and demonstrated that several SNPs in the previously identified candidate genes *IREB2* and *CHRNA3* met GW criteria for statistical significance. An SNP in another novel gene (*PSMA4*) was GW significant after weighting. However, expression of *PSMA4* in the lung was not associated with COPD phenotypes, and thus the observed association is likely due to linkage disequilibrium with the nearby genes *CHRNA3* and *IREB2*. We then replicated the two *FGF7* SNPs in an independent cohort of smoking adults, and showed that they are both significantly associated in the same direction with COPD. Notably, the effect sizes in Hispanics are larger than in the overall cohort, suggesting that these alleles confer greater risk in this population. This Hispanic population in New Mexico has a similar proportion of European and Native American ancestry as the Costa Rican cohort,[20 21] so another likely possibility is that patterns of linkage disequilibrium may be different between Hispanics and Caucasians in this genomic region, and that these SNPs are tagging a haplotype or functional SNP in the Hispanic subjects. Additionally, there was a trend towards increased lung tissue expression of *FGF7* in an independent cohort of COPD subjects, in whom there was a significant negative correlation between *FGF7* expression and FEV_1 and FEV_1/FVC ratio.