

affecting the homozygosity mapping analysis; (ii) search for ASs genome-wide; (iii) check the validity of the inbreeding history, which is vital for homozygosity mapping but is often erroneous, and (iv) calculate the probability of the disease-causing gene being located in the regions identified. At the end of the analysis, we would know the probability that regions identified contain a disease-causing gene, and we would be able to determine how much effort should be devoted to scrutinizing the regions.

In the current study, we present an algorithm that implements the capabilities described in the above paragraph. We confirmed the power of this algorithm using 6 patients with Siiyama-type α 1-antitrypsin deficiency, a rare autosomal recessive disease in Japan [7,8]. The preliminary version of the algorithm described here has been used to prove that the *SLC34A2* gene is responsible for pulmonary alveolar microlithiasis [9]; the current version has been used to show that the *OPTN* gene is responsible for amyotrophic lateral sclerosis [10].

Implementation

Crossover model

We used the Haldane's Poisson process model for the occurrence of crossovers and performed all calculations based on this model [11]. Information on SNPs used by Affymetrix's Genome-Wide Human SNP Array 6.0 (hereafter referred to as SNP Array 6.0) was summarized in the annotation file, [12], in which the genetic distance from the telomere of the short arm of a chromosome to each SNP was obtained by interpolation using the sex-averaged data published by deCODE Genetics [13]. We restricted our analysis to a total of 890,625 autosomal SNPs with assigned dbSNP refIDs [14].

Monte Carlo simulation

The average number, the average length, and the maximal length of the ASs derived from a common ancestor were calculated for a range of $m + n$ values (Figure 1A) using a Monte Carlo simulation. The trial was repeated until we observed 100,000 events in which at least 1 AS appeared in the autosomal region.

The length of AS

The subject is removed from the common ancestor m generations on the paternal side and n generations on the maternal side (Figure 1A). Assuming that the length of each autosome is infinite, the length of AS conforms to an exponential distribution with a probability density function of

$$f(x) = \lambda e^{-\lambda x} \lambda = \frac{m+n}{100} (cM^{-1}). \quad (2)$$

In actuality, the autosomes have finite length; however, **equation 2** provides a good approximation when the length of an AS is much shorter than the length of an autosome.

RHS (run of homozygous SNPs), false negative, type A false positive and type B false positive

An RHS is defined as a run of homozygous SNPs with a genetic length greater than the RHS cutoff value (Figure 1B). All SNPs in an AS are homozygous, and therefore an RHS suggests the presence of an AS. We defined 3 types of errors. False negatives are ASs that are not contained in RHSs. Type A false positives are RHSs that do not contain ASs. Type B false positives are the spaces within an RHS that do not contain an AS. The false negative rate ($R_{false\ negative}$) is the ratio of false negatives to the total length of the AS. The false positive rate ($R_{false\ positive}$) is the ratio of false positives (the type A false positives plus the type B false positives) to the total length of the autosomes.

(1) $R_{false\ negative}$, the ratio of the total length of false negatives to the total length of the AS

According to the **equation 2**,

$$R_{false\ negative} = \frac{\int_0^c xf(x)dx}{\int_0^{\infty} xf(x)dx} = 1 - e^{-\lambda c} (1 + \lambda c). \quad (3)$$

(2) $R_{Type\ A\ false\ positive}$, the ratio of the total length of type A false positives to the total length of the autosomes

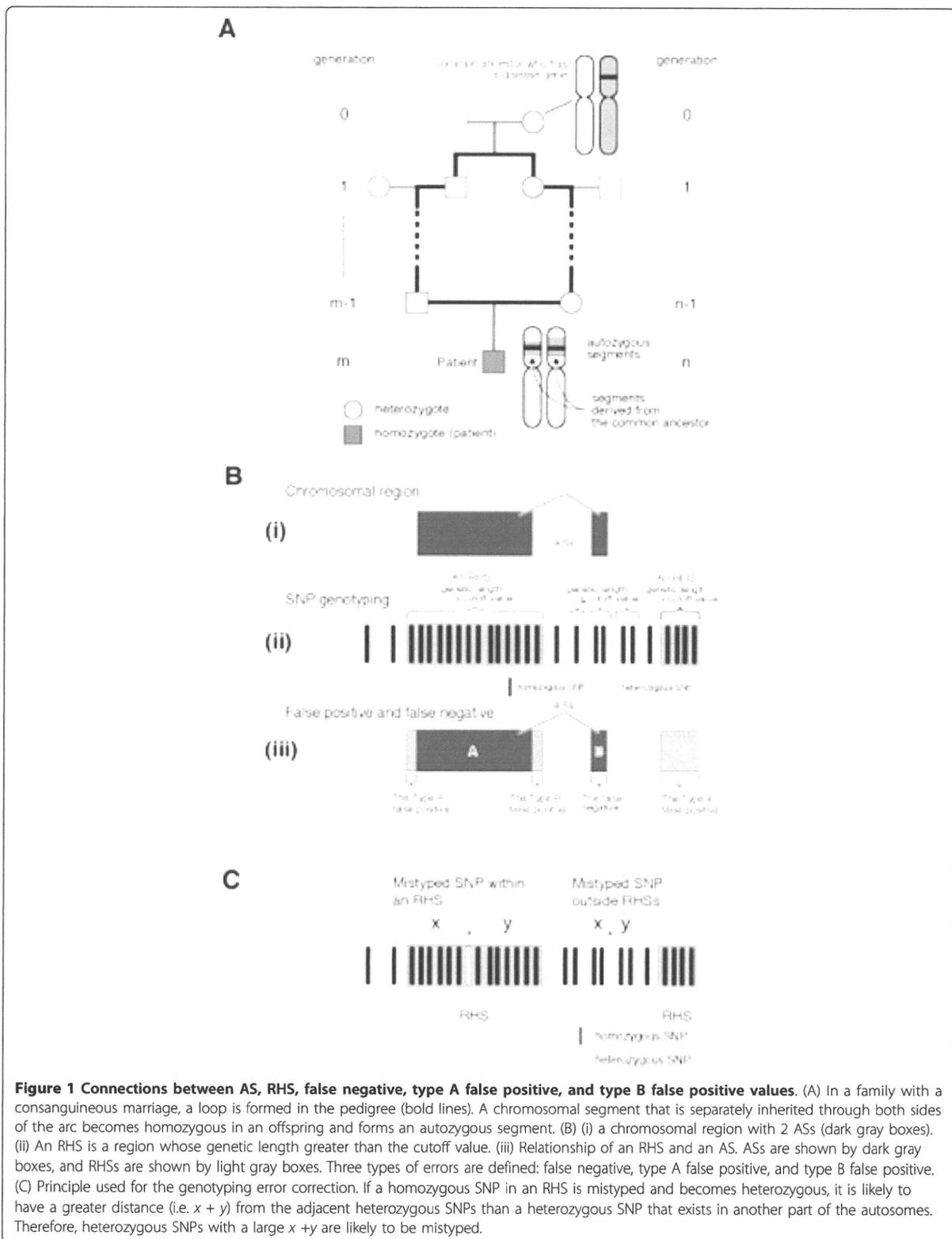
Given that N_{SNP} is the total number of SNPs on a genotyping array, and P_n and Q_n are the frequencies of the major and minor alleles for the n th SNP, respectively, then the average frequencies of the major alleles ($\bar{F}_{major\ allele}$) and the minor alleles ($\bar{F}_{minor\ allele}$) are

$$\bar{F}_{major\ allele} = \frac{\sum_{n=1}^{N_{SNP}} P_n}{N_{SNP}} \quad \text{and} \quad \bar{F}_{minor\ allele} = \frac{\sum_{n=1}^{N_{SNP}} Q_n}{N_{SNP}},$$

respectively. The numbers of homozygous SNPs ($N_{homozygous\ SNP}$) and heterozygous SNPs ($N_{heterozygous\ SNP}$) are approximated by

$$N_{homozygous\ SNP} \approx (\bar{F}_{major\ allele})^2 N_{pt} + (\bar{F}_{minor\ allele})^2 N_{pt}, \quad \text{and} \\ N_{heterozygous\ SNP} \approx 2(\bar{F}_{major\ allele})(\bar{F}_{minor\ allele}) N_{pt},$$

where N_{pt} is the number of SNPs successfully genotyped. Assuming that heterozygous SNPs are randomly



located, then the length between 2 heterozygous SNPs conforms to an exponential distribution with a probability density function of

$$f(x) = \lambda x \quad \lambda = \frac{N_{\text{heterozygousSNP}}}{L_{\text{autosome}}} (cM^{-1}), \quad (4)$$

where L_{autosome} is the entire length of the autosomes. Therefore, at a cutoff value of c cM,

$$R_{\text{Type A false positive}} = \frac{\int_0^{\infty} xf(x)dx}{\int_0^{\infty} xf(x)dx} = (1 + \lambda c)e^{-\lambda c}. \quad (5)$$

(3) $R_{\text{Type B false positive}}$, the ratio of the total length of type B false positives to the total length of the autosomes

$R_{\text{Type B false positive}}$ is not calculated mathematically but is calculated according to the actual data. An RHS containing an AS is expected to have type B false positives with an average length of $\frac{1}{2} \times \frac{L_{\text{autosome}}}{N_{\text{heterozygousSNP}}}$ on each end. It is impossible to distinguish RHSs that contain ASs from those that do not. We calculated $R_{\text{Type B false positive}}$ under the assumption that every RHS contains an AS. Therefore, the $R_{\text{Type B false positive}}$ calculation results in an overestimation, which we consider better than an underestimation for determination of the appropriate RHS cutoff. Therefore,

$$R_{\text{Type B false positive}} = \frac{\text{number of RHS}}{2} \times \frac{L_{\text{autosome}}}{N_{\text{heterozygousSNP}}}. \quad (6)$$

(4) $R_{\text{false positive}}$, the ratio of the total length of false positives to the total length of the autosomes

$$R_{\text{false positive}} = R_{\text{Type A false positive}} + R_{\text{Type B false positive}} \\ = \left(1 + \frac{N_{\text{heterozygousSNP}}}{L_{\text{autosome}}}\right)e^{-\frac{N_{\text{heterozygousSNP}}}{L_{\text{autosome}}}} + \frac{\text{number of RHS}}{2} \times \frac{L_{\text{autosome}}}{N_{\text{heterozygousSNP}}}. \quad (7)$$

Probability that a disease-causing gene is contained in RHSs, or the overlap of RHSs

The probability that RHSs obtained contains a disease-causing gene is calculated using **equation 1**.

$$P_{\text{GenesInRHS}} = (1 - R_{\text{false negative}}) \times P_{\text{AS}} \\ = (1 - R_{\text{false negative}}) \times \frac{F}{(1-F)p+F}. \quad (8)$$

Here, F is the coefficient of consanguinity and is calculated by

$$F \approx \frac{\text{total length of RHSs}}{\text{total length of the autosomes}}. \quad (9)$$

The probability that the overlap of RHSs among multiple patients contain the gene is calculated by

$$P_{\text{GenesInOverlap}} = \prod_{\text{All patients}} P_{\text{GenesInRHS}}. \quad (10)$$

Human Subjects and genotyping

This study was approved by the Institutional Review Boards of Saitama Medical University and Juntendo University. After obtaining written informed consent, DNA samples from 6 patients with $\alpha 1$ -antitrypsin deficiency were purified from peripheral blood. These patients were not related and lived in different areas of Japan. Patients 1-5 were from families with a history of inbreeding because their parents were first cousins. Patient 6 did not have any family history of inbreeding. These 6 patients were genotyped using the SNP Array 6.0. The genotyping data for 86 HapMap JPT were available in the HapMap3 draft release 2 <http://www.hapmap.org>, and were downloaded from the Wellcome Trust Sanger Institute web site <http://www.sanger.ac.uk/humgen/hapmap3/>. The genotyping data for NA18987, a subject in HapMap JPT, was also distributed from Affymetrix and was used in the current study.

Genotyping error correction

Genotyping errors may convert homozygous SNPs to heterozygous SNPs and erroneously terminate an RHS, resulting in the failure to detect a portion of an RHS. According to Affymetrix, SNP Array 6.0 has an accuracy of > 0.997 , implying that the genotyping error rate ($P_{\text{genotypingError}}$) may be 0.003 at maximum. A mistyped heterozygous SNP occurring in an RHS is separated by a large distance from neighboring heterozygous SNPs (Figure 1C). Therefore, if a heterozygous SNP is separated from neighboring SNPs by a distance that is rarely observed by chance, we speculated that the SNP was mistyped. Using **equation 4**, we calculated the probability of a heterozygous SNP being separated from neighboring SNPs at the observed distance ($P_{\text{distanceOccurredByChance}}$). A SNP with

$P_{distanceOccreByChance} < 0.01$ was considered a mistyped SNP and these data were removed. This algorithm may erroneously remove 20 correctly genotyped heterozygous SNPs ($N_{homozygousSNP} \times P_{genotypingError} \times 0.01$) from a single SNP array analysis data, which we considered acceptable.

Statistical analysis

The number of patients and controls who shared an RHS at each SNP position was compared. The assumption was made that

$$u = \frac{\hat{P}_1^* - \hat{P}_2^*}{\sqrt{\hat{P}^*(1-\hat{P}^*)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

has a standard normal distribution, where

$$\hat{P}_1^* = \frac{x_2 + 0.5}{n_2 + 1}, \hat{P}_2^* = \frac{x_2 + 0.5}{n_2 + 1}, \hat{P}^* = \frac{x_1 + x_2 + 0.5}{n_1 + n_2 + 1}. \text{ Here, } x_1$$

and x_2 represent the numbers of patients and controls sharing RHSs, respectively, and n_1 and n_2 represent the total numbers of patients and controls, respectively. The P value was calculated by

$$P = \int_{u_0}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx.$$

Computer program

The computer program was written in the ANSI standard C programming language. The program was compiled by the GNU C compiler 4.2 and run on a MacBook Pro (CPU: 2.53 GHz Intel Core 2 Duo, 4 GB RAM) computer. The command line programs and the programs equipped by graphic user interface are both available from our web site at <http://www.hhanalysis.com>.

Result

Strategy

Our aim was to establish an algorithm for homozygosity mapping that uses SNP genotyping data obtained by high-density arrays, is equipped by a powerful genotyping error correction algorithm, detects ASs genome-wide, allows investigation into the family inbreeding history, and is able to calculate the probability that the identified regions contain the target gene.

The algorithm searches for the ASs (Figure 1A, B(i)) through runs of homozygous SNPs, or RHSs, that are formed by consecutively homozygous SNPs and are

longer than the RHS cutoff value (Figure 1B(ii)). RHSs are presumably the autozygous segments (ASs). Three types of errors were defined; false negative, type A false positive, and type B false positive (Figure 1B(iii)). The main determinants of the false negative rate ($R_{false\ negative}$), which is the ratio of the total length of false negatives to the total length of ASs, are the number of SNPs investigated and the genotyping error rate. The main determinants of the false positive rate ($R_{false\ positive}$), which is the ratio of the total length of type A false positives plus type B false positives to the entire length of the autosomes, are the positioning of SNPs, local haplotype block structure [15], and population substructure [16].

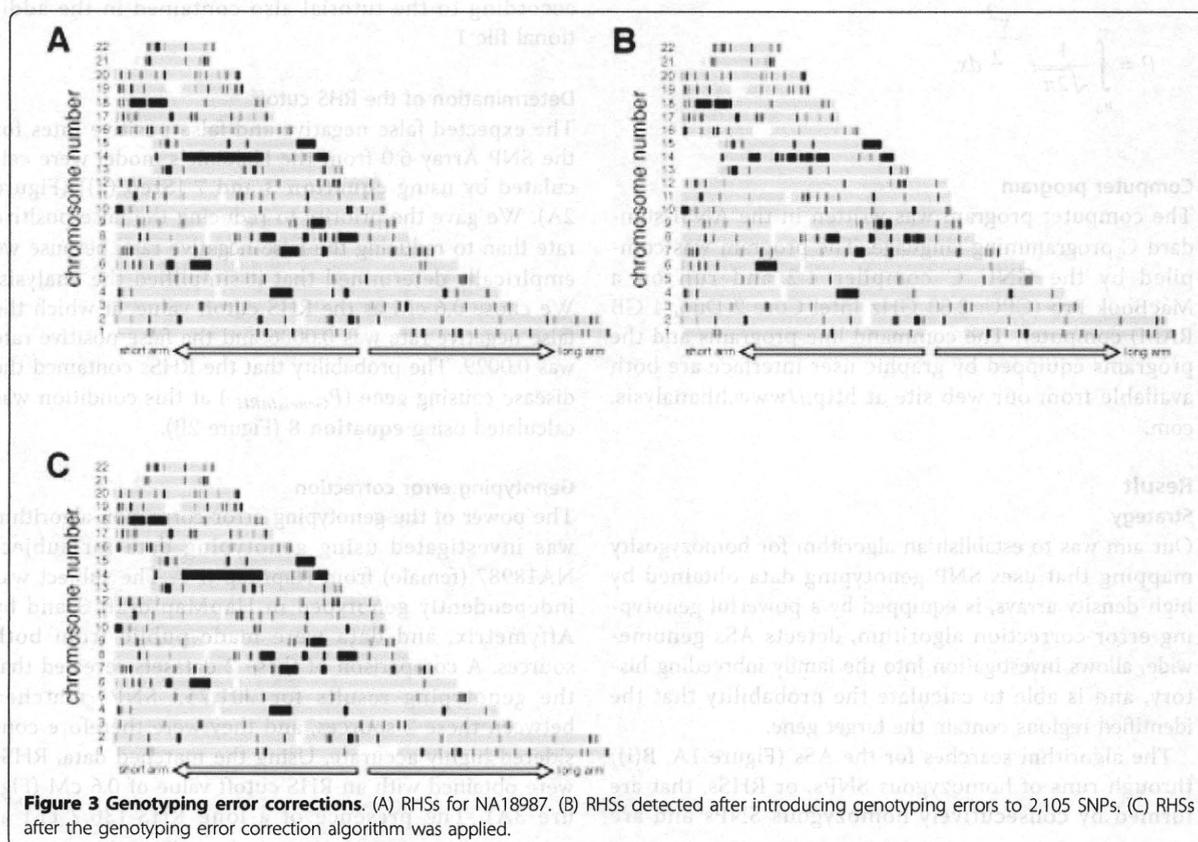
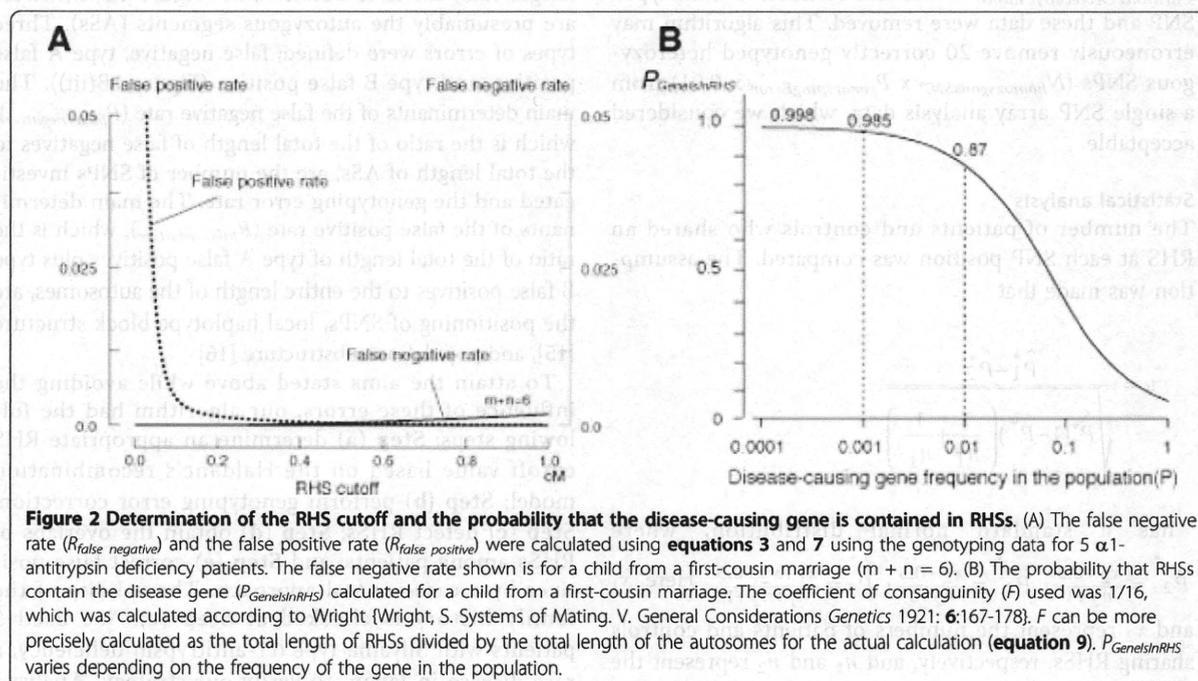
To attain the aims stated above while avoiding the influence of these errors, our algorithm had the following steps: **Step (a)** determine an appropriate RHS cutoff value based on the Haldane's recombination model; **Step (b)** perform genotyping error correction; **Step (c)** detect RHSs; **Step (d)** obtain the overlaps of RHSs among patients; and **Step (e)** correct false positives by a case-control approach. The validity of the family history is checked at **Step (c)**. We used 5 patients with Siiyama-type $\alpha 1$ -antitrypsin deficiency, a rare disease in Japan, to verify our strategy. Analyses performed in the Result section can be reproduced using the program contained in additional file 1 according to the tutorial also contained in the additional file 1.

Determination of the RHS cutoff

The expected false negative and false positive rates for the SNP Array 6.0 from the Haldane's model were calculated by using **equation 3** and 7 [**Step (a)**] (Figure 2A). We gave the priority to reducing the false positive rate than to reducing the false negative rate, because we empirically determined that it simplified the analysis. We chose 0.6 cM as the RHS cutoff value, at which the false negative rate was 0.0006 and the false positive rate was 0.0029. The probability that the RHSs contained the disease-causing gene ($P_{GenelsInRHS}$) at this condition was calculated using **equation 8** (Figure 2B).

Genotyping error correction

The power of the genotyping error correction algorithm was investigated using genotyping data for subject NA18987 (female) from HapMap JPT. The subject was independently genotyped in HapMap draft 3 and by Affymetrix, and data were made public from both sources. A comparison of these 2 datasets revealed that the genotyping results for 701,753 SNPs matched between these 2 sources, and they were therefore considered highly accurate. Using the matched data, RHSs were obtained with an RHS cutoff value of 0.6 cM (Figure 3A). The presence of a long RHS (36.2 cM at



maximum) suggested that she had a family history of inbreeding, as described later. Considering the fact that the manufacturer (Affymetrix) claimed that the genotyping error rate for the SNP Array 6.0 is less than 0.003, we randomly introduced errors into selected 2,105 SNPs (701,753 SNPs \times 0.003) and obtained RHSs. These error hampered the detection of RHSs, especially the long ones (Figure 3B). Following application of the genotyping error correction algorithm (Figure 1C), RHSs were restored (Figure 3C). The same trial repeated 100 times revealed that the genotyping error correction restored an average of 94.2% of the total length of all RHSs, and 99.9% of the total length of RHSs that were longer than 2 cM. This indicated that 99.9% of the total length of

ASs resulting from first- or second cousin marriages would be correctly detected as RHSs after the correction. The total length of the regions that were erroneously detected as RHSs amounted to only 0.2% of the total length of the autosomes. These results indicated that the performance of the genotyping error correction algorithm was excellent.

RHSs in the patients

We applied the genotyping error correction algorithm to the data for 5 patients with Siiyama-type α 1-antitrypsin deficiency [Step (b)], and then obtained RHSs [Step (c)] (Figure 4A-E). All patients had long RHSs, which were likely to be the result of first-cousin marriages.

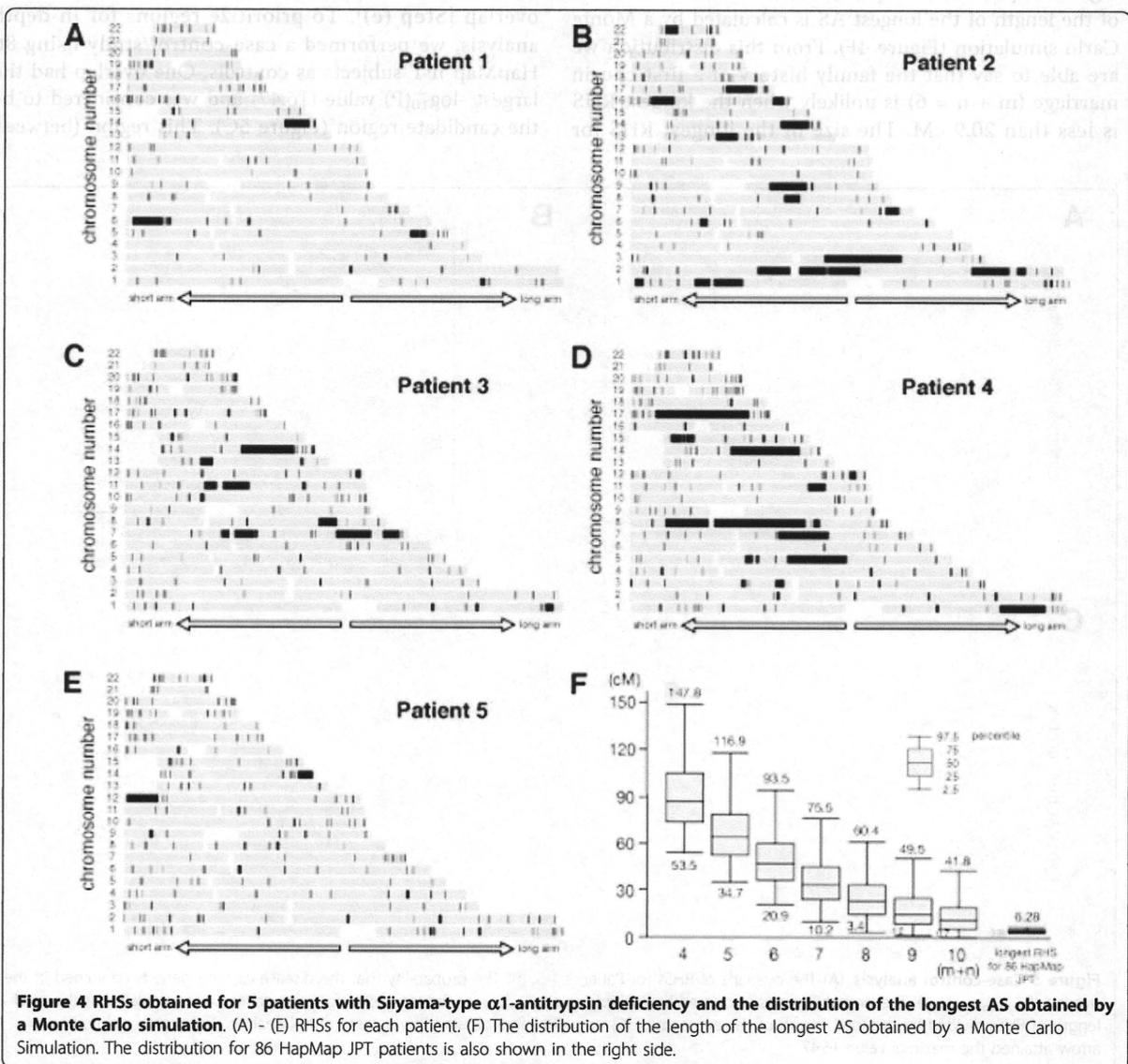


Table 1 Size of the longest RHS for each patient

	Length of the longest RHS (cM)
Patient 1	36.2
Patient 2	39.6
Patient 3	22.1
Patient 4	40.3
Patient 5	30.2

Statistics of AS

We investigated whether the RHSs obtained for each patient were consistent with family history [Step (d)]. We focused on the size of the longest AS because they are an index of the most recent occurrence of inbreeding in the patient's family (equation 2). The distribution of the length of the longest AS is calculated by a Monte Carlo simulation (Figure 4F). From this distribution we are able to say that the family history of a first cousin marriage ($m + n = 6$) is unlikely when the longest RHS is less than 20.9 cM. The size of the longest RHS for

Patients 1-5 were consistent with what expected from their family histories (Table 1).

Overlap of RHSs

We then obtained the overlaps of the RHSs for Patients 1-5 whose parents were first cousins [Step (d)] (Figure 5A). The probability that these regions contained the disease-causing gene ($P_{GenesInOverlap}$) was calculated by equation 10 and is shown in Figure 5B. The prevalence of Siiyama-type α 1-antitrypsin deficiency is less than 1 in a million in Japan, and the frequency of the gene is suspected to be less than 0.001 in the general population, indicating that the overlaps likely contained the disease-causing gene.

Some of the autosomal regions are prone to type A or type B false positives, and thus are likely to appear as an overlap [Step (e)]. To prioritize regions for in-depth analysis, we performed a case-control study using 86 HapMap JPT subjects as controls. One overlap had the largest $-\log_{10}(P)$ value (16.47) and was considered to be the candidate region (Figure 5C). This region (between

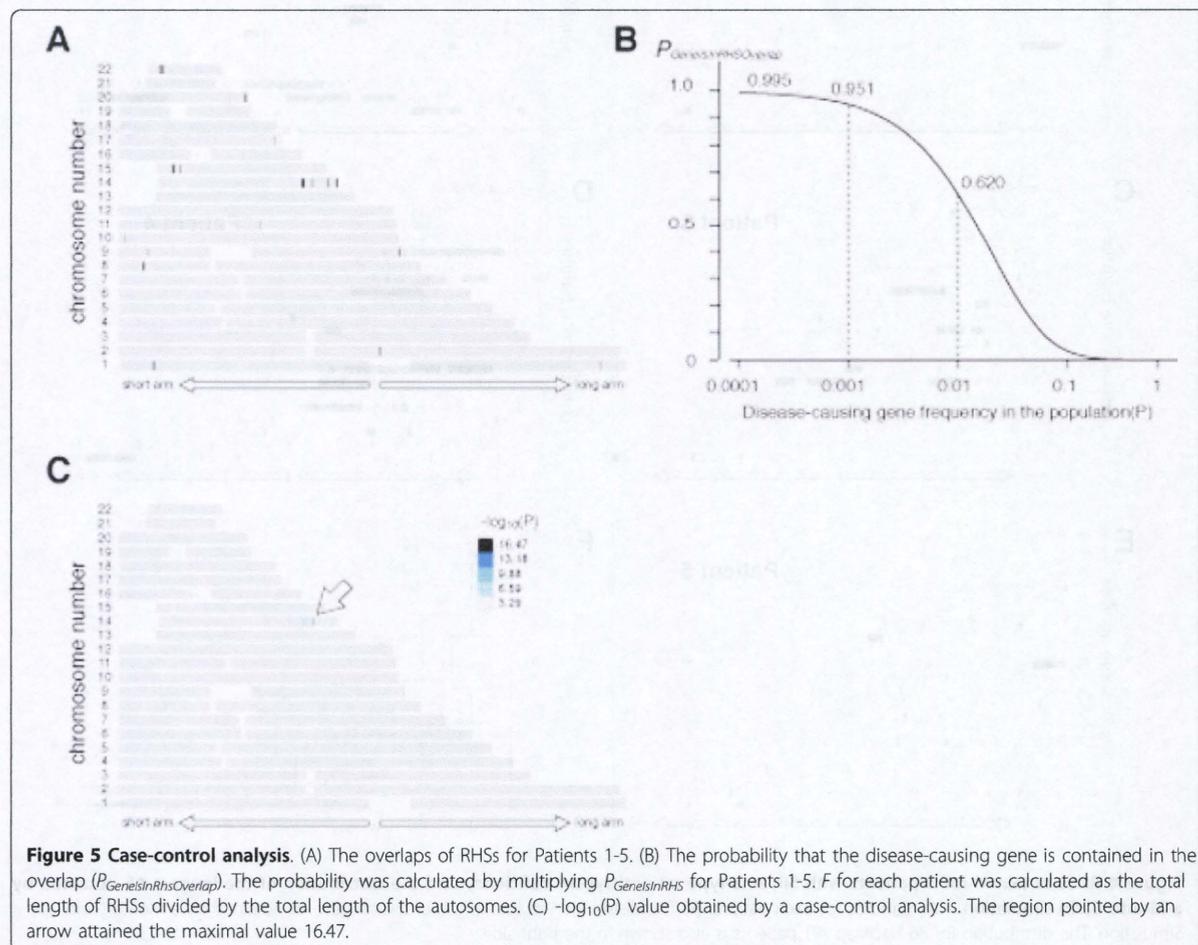


Table 2 Genes present in the candidate RHS overlap

C14orf48	chromosome 14 open reading frame 48
OTUB2	OTU domain, ubiquitin aldehyde binding 2
DDX24	DEAD (Asp-Glu-Ala-Asp) box polypeptide 24
IFI27L1	interferon, alpha-inducible protein 27-like 1
IFI27	interferon, alpha-inducible protein 27
IFI27L2	interferon, alpha-inducible protein 27-like 2
PPP4R4	protein phosphatase 4, regulatory subunit 4
SERPINA10	serpin peptidase inhibitor, clade A (alpha-1 antiprotease, antitrypsin), member 10
SERPINA6	serpin peptidase inhibitor, clade A (alpha-1 antiprotease, antitrypsin), member 6
LOC10028	Description: hypothetical protein LOC100287997
SERPINA2	serpin peptidase inhibitor, clade A (alpha-1 antiprotease, antitrypsin), member 2
SERPINA1	serpin peptidase inhibitor, clade A (alpha-1 antiprotease, antitrypsin), member 1
SERPINA11	serpin peptidase inhibitor, clade A (alpha-1 antiprotease, antitrypsin), member 11
SERPINA9	serpin peptidase inhibitor, clade A (alpha-1 antiprotease, antitrypsin), member 9
SERPINA12	serpin peptidase inhibitor, clade A (alpha-1 antiprotease, antitrypsin), member 12

rs10134551 and rs910349) had a genetic length of 1.44 cM, and contained 15 genes (Table 2), one of which was the disease-causing gene for Siiyama-type α 1-antitrypsin deficiency, SERPIN1.

A patient without family history of inbreeding

We occasionally encounter patients who do not have a family history of inbreeding while searching for a

recessive disease-causing gene. Data from such patients are not used in the main analysis, but these data may be used for prioritizing the overlaps of RHSs as obtained in Figure 5 for an in-depth search. Patient 6 had Siiyama-type α 1-antitrypsin deficiency but did not have a family history of inbreeding. The length of the longest RHS (6.8 cM, Figure 6A) was outside of the 95% range for the Japanese population (Figure 4F, **rightmost bar and**

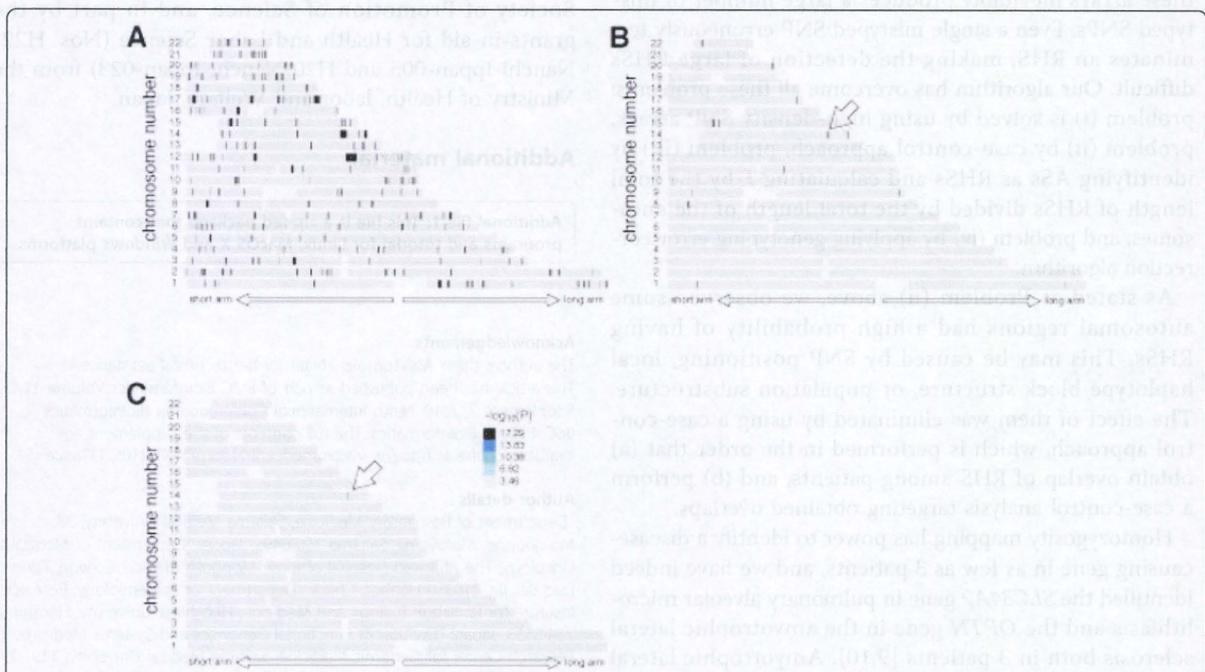


Figure 6 Subject without family history of inbreeding. (A) RHSs obtained for a patient without a family history of inbreeding (Patient 6). (B) RHS overlaps for Patients 1-6. Addition of data for Patient 6 further narrowed the overlapped regions (compare with Figure 5). The disease-causing gene was contained in the region indicated by a white arrow. (C) $-\log_{10}(P)$ value obtained by a case-control analysis. The region pointed by an arrow attained the maximal value 17.29.

whisker). We reasoned that the patient's family might have had forgotten inbreeding history, and that the RHSs for the patient may have a high probability of containing the disease-causing gene. This was indeed the case; addition of the data from Patient 6 excluded several overlapped regions (Figure 6B, **compare with** Figure 5A) and increased $-\log_{10}(P)$ (Figure 6C, **compare with** Figure 5C), although the list of the genes was the same as Table 2. If the length of the longest RHS suggested a hidden inbreeding history, the data for subjects without an inbreeding history could be used to prioritize some RHS overlaps for an in-depth search.

Discussion

In the current report, we described the quantitatively-modeled homozygosity mapping algorithm that uses high density array SNP genotyping data.

Homozygosity mapping is simple in principle, but many pitfalls were discovered when it was actually applied. Problems that included (i) unexpected allelic heterogeneity, (ii) identification of a homozygous identical-by-descent (IBD) region to the disease locus, (iii) underestimation of the extent of inbreeding, were pointed out in the analyses using microsatellite markers [17] and are still observed in the analyses using SNPs. Moreover, use of high-density SNP arrays introduced a novel problem, (iv) a large number of mistyped SNPs. Although the genotyping error rate is low for high-density arrays, the huge number of SNPs in these arrays inevitably produces a large number of mistyped SNPs. Even a single mistyped SNP erroneously terminates an RHS, making the detection of large RHSs difficult. Our algorithm has overcome all these problems: problem (i) is solved by using high-density SNP arrays, problem (ii) by case-control approach, problem (iii) by identifying ASs as RHSs and calculating F by the total length of RHSs divided by the total length of the autosomes, and problem (iv) by applying genotyping error correction algorithm.

As stated as Problem (ii) above, we observed some autosomal regions had a high probability of having RHSs. This may be caused by SNP positioning, local haplotype block structure, or population substructure. The effect of them was eliminated by using a case-control approach, which is performed in the order that (a) obtain overlap of RHS among patients, and (b) perform a case-control analysis targeting obtained overlaps.

Homozygosity mapping has power to identify a disease-causing gene in as few as 3 patients, and we have indeed identified the *SLC34A2* gene in pulmonary alveolar microlithiasis and the *OPTN* gene in the amyotrophic lateral sclerosis both in 3 patients [9,10]. Amyotrophic lateral sclerosis has multiple causative genes. In the latter report, we were able to identify one of the genes by investigating each combination of 3 patients from 7 patients with a

history of inbreeding, seeking for 3 patients harboring the same disease-causing gene. Our algorithm worked fine in this approach. During the process, it was quite helpful that the algorithm provided the probability that the identified regions contain the disease-causing gene, which determined how much effort should be further devoted. To our knowledge, the algorithm presented in the current study is the first to provide this information.

Conclusions

We described an algorithm that enables homozygosity mapping to be performed based on a quantitative model using SNP genotyping data. Our procedure will accelerate the identification of disease-causing genes using high-density SNP array data.

Availability and requirements

Project name: qHomozygosityMapping
Project home page: <http://www.hhanalysis.com>
Operating system(s): Mac, Linux and Windows.
Programming language: C
License: GNU GPL.

Any restrictions to use by non-academics: The software is for academic purpose only.

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 and H20-Nanchi-Ippan-023) from the Ministry of Health, labor and Welfare, Japan.

Additional material

Additional file 1: This file is a zipped package that contains programs and tutorial for Linux, MacOS X and Windows platforms.

Acknowledgements

The authors thank Ms. Tomoko Hirata for her technical assistance. This article has been published as part of BMC Bioinformatics Volume 11 Supplement 7, 2010: Ninth International Conference on Bioinformatics (InCoB2010): Bioinformatics. The full contents of the supplement are available online at <http://www.biomedcentral.com/1471-2105/11?issue=57>.

Author details

¹Department of Respiratory Medicine, Saitama Medical University, 38 Morohongo, Moroyama, Saitama 350-0495, Japan. ²Department of Medical Oncology, The Affiliated Hospital of Inner Mongolia Medical College, Tong Dao Bei Jie, 010050 Hohhot, China. ³Department of Epidemiology, Research Institute for Radiation Biology and Medicine, Hiroshima University, Hiroshima 734-8553, Japan. ⁴Division of Functional Genomics and Systems Medicine, Research Center for Genomic Medicine, Saitama Medical University, 1397-1 Yamane, Hidaka City, Saitama 350-1241, Japan. ⁵Department of Respiratory Medicine, Juntendo University, School of Medicine, 2-1-1 Hongo, Bunkyo-ku, Tokyo 113-8421, Japan.

Authors' contribution

Huqun, S.F., H.M., T.T., T.S., M.K., H.K., Y.O., and K.S. tested the programs, did genetic analyses and provided ideas to improve the program. K.S. collected the patients' samples. K.H. provided basic ideas, wrote the program, and prepared manuscript.

Competing interests

The authors declare that they have no competing interests.

Published: 15 October 2010

References

1. McKusick VA: **Mendelian Inheritance in Man and its online version, OMIM.** *Am J Hum Genet* 2007, **80**: 588-604.
2. **OMIM-Online Mendelian Inheritance in Man.** [<http://www.ncbi.nlm.nih.gov/Omim/mimstats.html>].
3. Lander ES, Botstein D: **Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children.** *Science* 1987, **236**:1567-1570.
4. Clark AG: **The size distribution of homozygous segments in the human genome.** *Am J Hum Genet* 1999, **65**:1489-1492.
5. Woods CG, Valente EM, Bond J, Roberts E: **A new method for autozygosity mapping using single nucleotide polymorphisms (SNPs) and EXCLUDEAR.** *J Med Genet* 2004, **41**:e101.
6. Seelow D, Schuelke M, Hildebrandt F, Nurnberg P: **HomozygosityMapper—an interactive approach to homozygosity mapping.** *Nucleic Acids Res* 2009, **37**:W593-599.
7. Seyama K: **State of alpha1-antitrypsin deficiency in Japan.** *Respirology* 2001, **6**(Suppl):S35-38.
8. Seyama K, Nukiwa T, Souma S, Shimizu K, Kira S: **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* 1995, **152**:2119-2126.
9. Izumi S, Miyazawa H, Ishii K, Uchiyama B, Ishida T, Tanaka S, Tazawa R, Fukuyama S, Tanaka T, Nagai Y, Yokote A, Takahashi H, Fukushima T, Kobayashi K, Chiba H, Nagata M, Sakamoto S, Nakata K, Takebayashi Y, Shimizu Y, Kaneko K, Shimizu M, Kanazawa M, Abe S, Inoue Y, Takenoshita S, Yoshimura K, Kudo K, Tachibana T, Nukiwa T, Hagiwara K: **Mutations in the SLC34A2 gene are associated with pulmonary alveolar microlithiasis.** *Am J Respir Crit Care Med* 2007, **175**:263-268.
10. Maruyama H, Morino H, Ito H, Izumi Y, Kato H, Watanabe Y, Kinoshita Y, Kamada M, Nodera H, Suzuki H, Komure O, Matsuura S, Kobatake K, Morimoto N, Abe K, Suzuki N, Aoki M, Kawata A, Hirai T, Kato T, Ogasawara K, Hirano A, Takemi T, Kusaka H, Hagiwara K, Kaji R, Kawakami H: **Mutations of optineurin in amyotrophic lateral sclerosis.** *Nature* 2010, **465**:223-226.
11. Haldane J: **The combination of linkage values, and the calculation of distances between the loci of linked factors.** *J Genet* 1919, **8**:299-309.
12. **Affymetrix - Home.** [<http://www.affymetrix.com/index.affx>].
13. Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G, Shlien A, Palsson ST, Frigge ML, Thorgeirsson TE, Gulcher JR, Stefansson K: **A high-resolution recombination map of the human genome.** *Nat Genet* 2002, **31**:241-247.
14. **National Center for Biotechnology Information.** [<http://www.ncbi.nlm.nih.gov>].
15. International HapMap Consortium: **The International HapMap Project.** *Nature* 2003, **426**:789-796.
16. Overall AD, Nichols RA: **A method for distinguishing consanguinity and population substructure using multilocus genotype data.** *Mol Biol Evol* 2001, **18**:2048-2056.
17. Miano MG, Jacobson SG, Carothers A, Hanson I, Teague P, Lovell J, Cideciyan AV, Haider N, Stone EM, Sheffield VC, Wright AF: **Pitfalls in homozygosity mapping.** *Am J Hum Genet* 2000, **67**:1348-1351.

doi:10.1186/1471-2105-11-S7-S5

Cite this article as: Huqun et al.: A quantitatively-modeled homozygosity mapping algorithm, qHomozygosityMapping, utilizing whole genome single nucleotide polymorphism genotyping data. *BMC Bioinformatics* 2010 **11**(Suppl 7):S5.

Submit your next manuscript to BioMed Central and take full advantage of:

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at
www.biomedcentral.com/submit



Homozygosity haplotype analysis improves the identify novel genes



John M. Brehm, MD, MPH^{1,2,4,5}, Koichi Hagiwara⁸, Nadia Boutaoui, PhD¹, Manuel E. Soto-Quiros, MD⁹, Jacobson, MD, MPH⁵, Soma Datta¹, Edwin K. Silverman,

¹Channing Laboratory, ²Division of Pulmonary/Critical Care Medicine, ³Center for Genomic Medicine, ⁴Department of Medicine, and ⁵Medical School, Boston, Massachusetts, 02115; ⁷Division of Pediatric Pulmonology, Hospital N

Background

-Genome-wide association studies (GWAS) of large cohorts of subjects with and without chronic obstructive pulmonary disease have identified *FAM13A1* and *CHRNA3/5* as likely COPD susceptibility genes.^{1,2} While other genes may affect COPD risk, they may not be detected by GWAS with usual sample size.

- Various methods have been proposed to weight the results of GWAS using prior information, but many of these strategies use prior knowledge that is biased as to the results (such as up-weighting exonic SNPs or promoter regions).

- We employed an unbiased novel methodology called Homozygosity Haplotype Analysis³ in a genetically isolated population to construct weights for the combined GWAS of three COPD studies (Eclipse, NETT-NAS, and Norway). These weights were then used to adjust the significance level of each SNP to identify novel loci that would not meet genome-wide significance in a standard GWAS.⁴

Methods

Study:

We recruited 58 cases with moderate to severe emphysema based on spirometry and 57 smoking controls without pre-existing lung disease from the Central Valley region of Costa Rica. To include a genetically homogenous population, all subjects were required to have 6 of 8 great-grandparents born in the Central Valley.

Genotyping and Homozygosity Haplotype Construction

Cases and controls were genotyped on the Illumina-Quad 610k platform. Regions of conserved homozygosity haplotype (RCHH) were constructed by comparing subjects at loci that are homozygous, while ignoring loci that are heterozygous. The number of cases with a given run is then compared to the number of controls, and a P value is constructed.

Application to GWAS

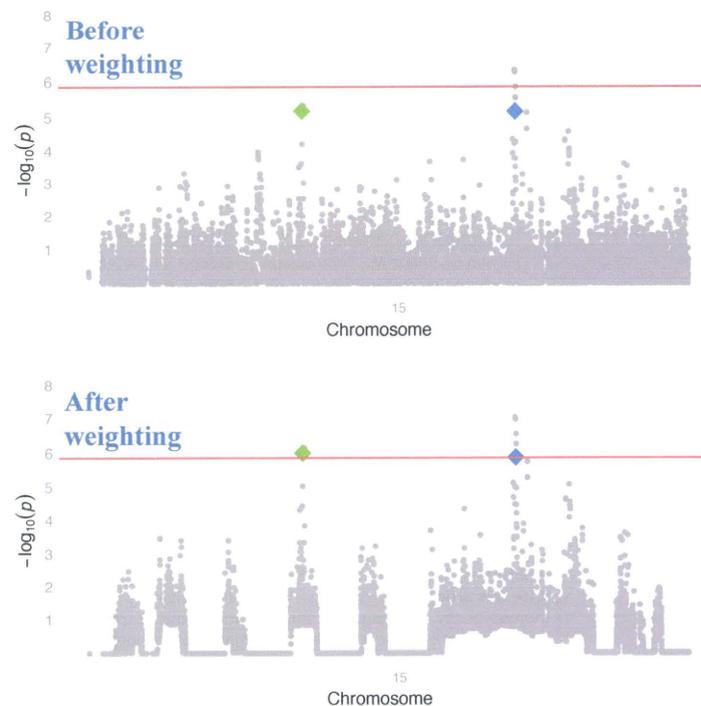
The P values for every RCHH were used to create weights which were then applied to the combined cohorts of Eclipse, NETT-NAS, and Norway (2,940 cases and 1,380 controls) using software from Roeder et al.⁴ A false discovery rate corrected alpha was used to determine genome-wide significance. The original and weighted P values of all FDR significant SNPs is presented in the table. Also included are P values for associations in the individual cohorts, and the significance of the RCHH identified in the Costa Rica population.

Results

FDR significant SNPs: FGF7 and PSMA4 are novel genes

SNP	Location (Chr: BP)	AI	OR	Original P	Original rank	Weighted P	Gene (distance from gene)
rs1903003	4:90105320	C	0.75	7.18E-08	1	6.87E-08	FAM13A1(0)
rs7671167	4:90103002	C	0.76	8.59E-08	2	8.22E-08	FAM13A1(0)
rs1062980	15:76579582	C	0.76	4.81E-07	3	9.53E-08	IREB2(0)
rs13180	15:76576543	C	0.76	5.01E-07	4	9.93E-08	IREB2(0)
rs8034191	15:76593078	C	1.32	5.37E-07	5	1.06E-07	IREB2(+12.2)
rs12914385	15:76685778	T	1.29	1.42E-06	9	2.81E-07	CHRNA3(0)
rs1051730	15:76681394	A	1.29	2.80E-06	14	5.54E-07	CHRNA3(0)
rs17404727	15:47791375	C	1.28	4.71E-06	15	7.17E-07	.
rs996414	9:26570067	G	0.76	1.80E-06	11	8.76E-07	.
rs4480740	15:47543134	A	1.27	6.75E-06	17	1.03E-06	FGF7(0)
rs12591300	15:47492033	A	1.27	8.78E-06	21	1.34E-06	FGF7(-10.72)
rs2656069	15:76532762	C	0.75	6.82E-06	18	1.35E-06	IREB2(0)
rs2036534	15:76614003	C	0.75	6.98E-06	19	1.38E-06	PSMA4(-5.75)
rs2869967	4:90088355	C	1.29	1.48E-06	10	1.41E-06	FAM13A1(0)

FGF7 (green) and PSMA4 (blue) are genome-wide significant after applying homozygosity weights



The power of genome-wide association studies to detect genetic loci in COPD



MD, PhD⁷, Lydiana Avila, MD⁷, Michael Cho, MD, MPH^{1,2,4}, Blanca Himes, PhD¹, Francine
n, MD, PhD^{1,4}, Juan C. Celedón, MD, Dr.PH¹⁻⁴

¹Department of Radiology, Brigham and Women's Hospital, Boston, Massachusetts, 02115; ⁶Department of Medicine, Harvard
Nacional de Niños, San José, Costa Rica; ⁸Saitama Medical University, Saitama, Japan

Conclusions

Results after weighting by homozygosity haplotypes

Region (kb)	Norway P value	NETT-NAS P value	Eclipse P value	Costa Rica RCHH		
				Cases (%)	Controls (%)	RCHH P value
0-22kb	4.3E-04	1.4E-03	9.1E-03	1 (2%)	0 (0%)	0.212
	7.9E-04	2.7E-04	7.8E-03	1 (2%)	0 (0%)	0.212
	9.9E-03	1.0E-02	3.6E-02	5 (9%)	0 (0%)	0.016
	7.9E-03	1.6E-02	4.3E-02	5 (9%)	0 (0%)	0.016
	1.5E-04	8.7E-03	8.2E-01	5 (9%)	0 (0%)	0.016
	1.4E-03	9.8E-03	9.5E-01	5 (9%)	0 (0%)	0.016
	4.3E-04	2.1E-02	8.4E-01	5 (9%)	0 (0%)	0.016
	1.9E-02	2.2E-02	3.4E-02	7 (12%)	0 (0%)	0.004
	6.2E-04	8.8E-01	2.6E-01	2 (3%)	0 (0%)	0.106
	4.0E-02	2.5E-02	2.1E-02	7 (12%)	0 (0%)	0.004
	3.9E-02	8.3E-02	2.5E-02	7 (12%)	0 (0%)	0.004
	1.6E-01	2.6E-03	1.0E-02	5 (9%)	0 (0%)	0.016
	5.8E-02	7.4E-03	9.4E-02	5 (9%)	0 (0%)	0.016
	4.7E-04	7.6E-03	4.4E-03	1 (2%)	0 (0%)	0.212

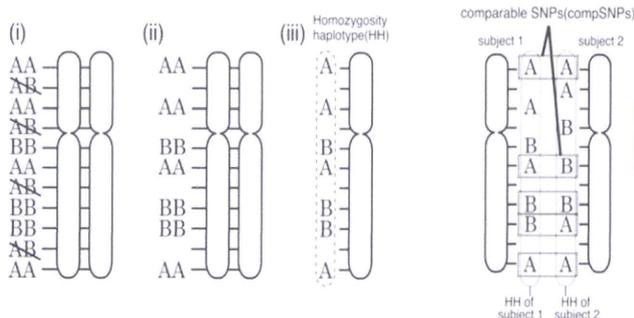
-By identifying regions of conserved homozygosity haplotype in a genetically homogenous population in Costa Rica, we were able to construct weights that improved the power to detect two novel loci that may play a role in COPD pathogenesis.

- The previously identified genes *CHRNA3*² and *IREB2*⁵ were contained within significant regions of homozygosity, which validates this method.

- *FGF7* may play a role in protecting against the oxidant injury induced by cigarette smoke, and/or influence airway development during embryogenesis. Differences in fetal development may play a role in susceptibility to COPD later in life.

- Further work needs to be done to examine the role of *FGF7* and *PSMA4* in expression studies in mice and appropriate human tissues. We also plan to perform fine-mapping or sequencing to elucidate the haplotype or SNPs that confer COPD risk.

Construction of RCHH



Novel loci identified

FGF7 or fibroblast growth factor 7, was identified in cultured human embryonic lung fibroblasts, and plays a role in promoting wound healing and protecting airway epithelium from oxidant injury in mice

PSMA4 is in the same chromosomal region as *CHRNA3* and *IREB2*. *PSMA4* mRNA levels are increased in lung tumors, and *PSMA4* expression induces apoptosis in cancer cells.

References

1. Cho MH, Boutaoui N, Klanderman BJ, Sylvia JS, Ziniti JP, Hersh CP, Demeo DL, et al. (2010) Variants in *FAM13A* are associated with chronic obstructive pulmonary disease. *Nat Genet* 42:200-202
2. Pillai SG, Ge D, Zhu G, Kong X, Shianna KV, Need AC, Feng S, Hersh CP, Bakke P, Gulsvik A, Ruppert A, Lodrup Carlsen KC, Roses A, Anderson W, Rennard SI, Lomas DA, Silverman EK, Goldstein DB (2009) A genome-wide association study in chronic obstructive pulmonary disease (COPD): identification of two major susceptibility loci. *PLoS Genet* 5:e1000421
3. Miyazawa H, Kato M, Awata T, Kohda M, Iwasa H, Koyama N, Tanaka T, Kyo S, Okazaki Y, Hagiwara K (2007) Homozygosity Haplotype Allows a Genomewide Search for the Autosomal Segments Shared among Patients. *The American Journal of Human Genetics* 80:1090-1102
4. Roeder K, Bacanu SA, Wasserman L, Devlin B (2006) Using linkage genome scans to improve power of association in genome scans. *Am J Hum Genet* 78:243-252
5. DeMeo DL, Mariani T, Bhattacharya S, Srisuma S, Lange C, Litonjua A, Bueno R, Pillai SG, Lomas DA, Sparrow D, Shapiro SD, Criner GJ, Kim HP, Chen Z, Choi AM, Reilly J, Silverman EK (2009) Integration of genomic and genetic approaches implicates *IREB2* as a COPD susceptibility gene. *Am J Hum Genet* 85:493-502

研究成果の刊行に関する一覧表

平成 22 年度 研究報告

英文論文 (9 報)

1. Satoh H, Inoue A, Kobayashi K, Maemondo M, Oizumi S, Isobe H, Gemma A, Saijo Y, Yoshizawa H, Hagiwara K, Nukiwa T. Low-dose gefitinib treatment for patients with advanced non-small cell lung cancer harboring sensitive epidermal growth factor receptor mutations. *J Thorac Oncol* 2011;in press.
2. Yamaguchi T, Soma T, Takaku Y, Nakagome K, Hagiwara K, Kanazawa M, Nagata M. Salbutamol modulates the balance of Th1 and Th2 cytokines by mononuclear cells from allergic asthmatics. *Int Arch Allergy Immunol* 2010;152 Suppl 1:32-40.
3. Tanaka T, Matsuoka M, Sutani A, Gemma A, Maemondo M, Inoue A, Okinaga S, Nagashima M, Oizumi S, Uematsu K, Nagai Y, Moriyama G, Miyazawa H, Ikebuchi K, Morita S, Kobayashi K, Hagiwara K. Frequency of and variables associated with the EGFR mutation and its subtypes. *Int J Cancer* 2010;126:651-655.
4. Takaku Y, Nakagome K, Kobayashi T, Yamaguchi T, Nishihara F, Soma T, Hagiwara K, Kanazawa M, Nagata M. Changes in airway inflammation and hyperresponsiveness after inhaled corticosteroid cessation in allergic asthma. *Int Arch Allergy Immunol* 2010;152 Suppl 1:41-46.
5. Sadakata R, Hatamochi A, Kodama K, Kaga A, Yamaguchi T, Soma T, Usui Y, Nagata M, Ohtake A, Hagiwara K, Kanazawa M. Ehlers-danlos syndrome type IV, vascular type, which demonstrated a novel point mutation in the col3A1 gene. *Intern Med* 2010;49:1797-1800.
6. Nakada H, Nakagome K, Takaku Y, Nishihara F, Yamaguchi T, Soma T, Hagiwara K, Kanazawa M, Kase Y, Nagata M. [questionnaire for determining relationship between nasal and asthma symptoms]. *Alerugi* 2010;59:688-698.
7. Maruyama H, Morino H, Ito H, Izumi Y, Kato H, Watanabe Y, Kinoshita Y, Kamada M, Nodera H, Suzuki H, Komure O, Matsuura S, Kobatake K, Morimoto N, Abe K, Suzuki N, Aoki M, Kawata A, Hirai T, Kato T, Ogasawara K, Hirano A, Takumi T, Kusaka H, Hagiwara K, Kaji R, Kawakami H. Mutations of optineurin in amyotrophic lateral sclerosis. *Nature* 2010;465:223-226.
8. Maemondo M, Inoue A, Kobayashi K, Sugawara S, Oizumi S, Isobe H, Gemma A, Harada M, Yoshizawa H, Kinoshita I, Fujita Y, Okinaga S, Hirano H, Yoshimori K, Harada T, Ogura T, Ando M, Miyazawa H, Tanaka T, Saijo Y, Hagiwara K, Morita S, Nukiwa T. Gefitinib or chemotherapy for non-small-cell lung cancer with mutated EGFR. *N Engl J Med* 2010;362:2380-2388.
9. Huqun, Fukuyama S, Morino H, Miyazawa H, Tanaka T, Suzuki T, Kohda M, Kawakami H, Okazaki Y, Seyama K, Hagiwara K. A quantitatively-modeled homozygosity mapping algorithm, qhomozygositymapping, utilizing whole genome single nucleotide polymorphism genotyping data. *BMC Bioinformatics* 2010;11 Suppl 7:S5.

和文論文（3報）

1. 萩原弘一 薬剤性肺障害 日本人における特性 治療学 44 巻 5 号 Page591-593, 2010
2. 萩原弘一 肺胞微石症 成人病と生活習慣病 40 巻 6 号 Page699-703, 2010
3. 萩原弘一 呼吸器疾患のゲノム解析 日本内科学会雑誌 100 巻 Suppl. Page99-100, 211

厚生労働科学研究補助金
難治性疾患克服研究事業
特発性肺線維症急性増悪及び薬剤性肺障害に関する日本人特異的遺伝素因に関する研究
平成 22 年度研究報告書

編集, 印刷, 発行

〒 350-0495 埼玉県入間郡毛呂山町毛呂本郷 38
埼玉医科大学 呼吸器内科 萩原弘一
TEL 049-276-1319, FAX 049-276-1635, E-mail hagiwark@saitama-med.ac.jp
