Huqun et al. BMC Bioinformatics 2010, 11(Suppl 7):S5
http://www.biomedcentral.com/1471-2105/11/S7/S5

BMC Bioinformatics

PROCEEDINGS Open Access

A quantitatively-modeled homozygosity mapping algorithm, qHomozygosityMapping, utilizing whole genome single nucleotide polymorphism genotyping data

Huqun[1,2], Shun-ichiro Fukuyama[1†], Hiroyuki Morino[3], Hiroshi Miyazawa[1], Tomoaki Tanaka[1], Tomoko Suzuki[1], Masakazu Kohda[4], Hideshi Kawakami[3], Yasushi Okazaki[4], Kuniaki Seyama[5], Koichi Hagiwara[1*†]

From Asia Pacific Bioinformatics Network (APBioNet) Ninth International Conference on Bioinformatics (InCoB2010)
Tokyo, Japan. 26-28 September 2010

Abstract

Homozygosity mapping is a powerful procedure that is capable of detecting recessive disease-causing genes in a few patients from families with a history of inbreeding. We report here a homozygosity mapping algorithm for high-density single nucleotide polymorphism arrays that is able to (i) correct genotyping errors, (ii) search for autozygous segments genome-wide through regions with runs of homozygous SNPs, (iii) check the validity of the inbreeding history, and (iv) calculate the probability of the disease-causing gene being located in the regions identified. The genotyping error correction restored an average of 94.2% of the total length of all regions with run of homozygous SNPs, and 99.9% of the total length of them that were longer than 2 cM. 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. We confirmed the power of this algorithm using 6 patients with Siiyama-type α1-antitrypsin deficiency, a rare autosomal recessive disease in Japan. Our procedure will accelerate the identification of disease-causing genes using high-density SNP array data.

Background

Identification of the genetic factors underlying disease causation provides crucial information for disease prevention and treatment. Nevertheless, genetic factors have not yet been elucidated for many diseases [1,2].

Homozygosity mapping [3] enables the detection of recessive disease-causing genes in a few patients from families with a history of inbreeding; this mapping technique is especially useful for the detection of rare genes. With this technique, chromosomal segments in which all polymorphic markers are homozyogous are considered autozygous segment (AS) [4]. If a patient's coefficient of consanguinity is F, and the frequency of the disease-causing gene in the population is p, then the chance that the recessive disease-causing gene is located in an AS (P_{AS}) is.

$$P_{AS} = \frac{F}{(1-F)p+F} \quad (1)$$

[3]

If a patient is from an inbred family (i.e., F is large) and the disease is rare (i.e., p is small), then $P_{AS} \approx 1$, indicating that the gene is located in an AS. There are implementations that utilize single-nucleotide polymorphism (SNP) genotyping data obtained by high-density arrays [5,6]. The usable implementation should (i) correct genotyping errors because thousands of SNPs are mistyped per high-density SNP array, adversely

* Correspondence: hagiwark@saitama-med.ac.jp
† Contributed equally
[1]Department of Respiratory Medicine, Saitama Medical University, 38 Morohongo, Moroyama, Saitama 350-0495, Japan
Full list of author information is available at the end of the article

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 α1-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

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

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 nth SNP, respectively, then the average frequencies of the major alleles ($\overline{F}_{major\ allele}$) and the minor alleles ($\overline{F}_{minor\ allele}$) are

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

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

$$N_{homozygousSNP} \approx \left(\overline{F}_{major\ allele}\right)^2 N_{Pt} + \left(\overline{F}_{minor\ allele}\right)^2 N_{Pt}, \text{ and}$$

$$N_{heterozygousSNP} \approx 2\left(\overline{F}_{major\ allele}\right)\left(\overline{F}_{minor\ allele}\right) N_{Pt},$$

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

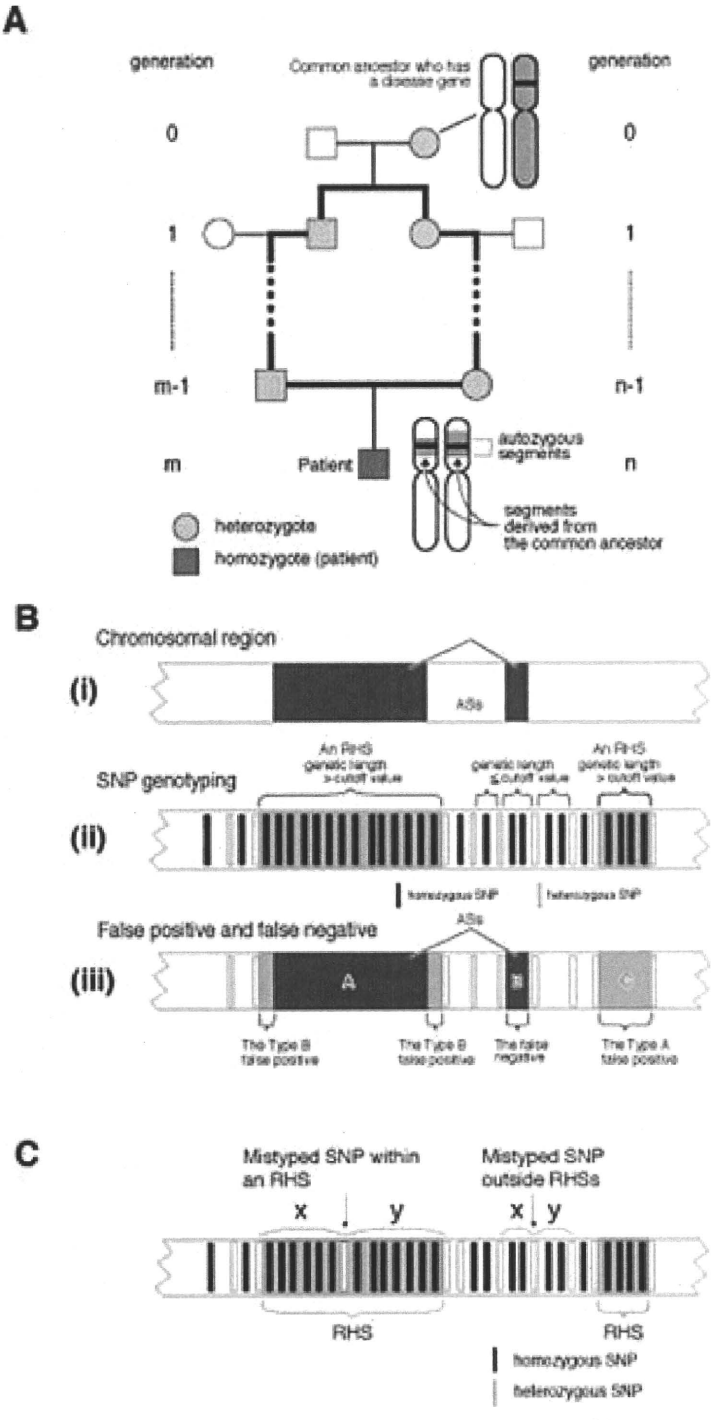


Figure 1 Connections between AS, RHS, false negative, type A false positive, and type B false positive values. (A) In a family with a consanguineous marriage, a loop is formed in the pedigree (bold lines). A chromosomal segment that is separately inherited through both sides of the arc becomes homozygous in an offspring and forms an autozygous segment. (B) (i) a chromosomal region with 2 ASs (dark gray boxes). (ii) An RHS is a region whose genetic length greater than the cutoff value. (iii) Relationship of an RHS and an AS. ASs are shown by dark gray boxes, and RHSs are shown by light gray boxes. Three types of errors are defined: false negative, type A false positive, and type B false positive. (C) Principle used for the genotyping error correction. If a homozygous SNP in an RHS is mistyped and becomes heterozygous, it is likely to have a greater distance (i.e. $x + y$) from the adjacent heterozygous SNPs than a heterozygous SNP that exists in another part of the autosomes. Therefore, heterozygous SNPs with a large $x + y$ are likely to be mistyped.

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

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

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

$$R_{Type\ A\ false\ positive} = \frac{\int_{0}^{c} xf(x)dx}{\int_{0}^{\infty} xf(x)dx} = (1+\lambda c)e^{-\lambda c}. \tag{5}$$

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

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

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

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

$$\begin{aligned} R_{false\ positive} &= R_{Type\ A\ false\ positive} + R_{Type\ B\ false\ positive} \\ &= (1+\frac{N_{heterozygousSNP}}{L_{autosome}}c)e^{-\frac{N_{heterozygousSNP}}{L_{autosome}}c} \\ &+ \frac{number\ of\ RHS}{2}\times\frac{L_{autosome}}{N_{heterozygousSNP}}. \end{aligned} \tag{7}$$

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**.

$$\begin{aligned} P_{GeneIsInRHS} &= (1-R_{false\ negative})\times P_{AS} \\ &= (1-R_{false\ negative})\times\frac{F}{(1-F)p+F}. \end{aligned} \tag{8}$$

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

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

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

$$P_{GeneIsInOverlap} = \prod_{All\ patients} P_{GeneIsInRHS}. \tag{10}$$

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 α1-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 ($P_{genotypingError}$) 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 ($P_{distanceOccurredByChance}$). A SNP with

$P_{distanceOccreceByChance} < 0.01$ was considered a mistyped SNP and these data were removed. This algorithm may erroneously remove 20 correctly genotyped heterozygous SNPs ($N_{homozygousSNP}$ x $P_{genotypingError}$ x 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

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

has a standard normal distribution, where $\hat{P}_2^* = \frac{x_2+0.5}{n_2+1}$, $\hat{P}_2^* = \frac{x_2+0.5}{n_2+1}$, $\hat{P}^* = \frac{x_1+x_2+0.5}{n_1+n_2+1}$. 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 α1-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 ($P_{GeneIsInRHS}$) 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

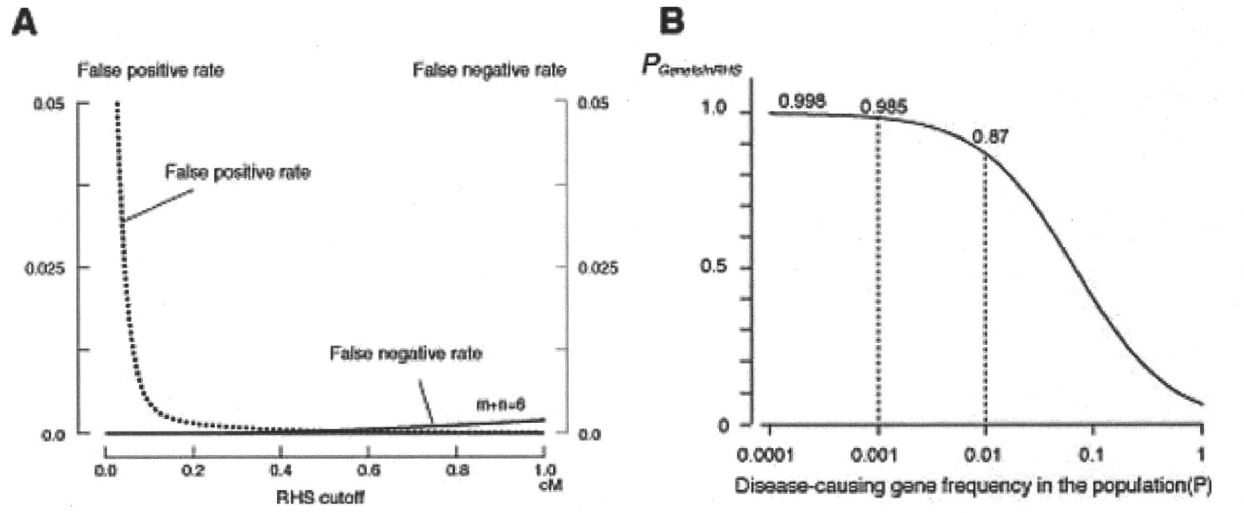


Figure 2 Determination of the RHS cutoff and the probability that the disease-causing gene is contained in RHSs. (A) The false negative rate ($R_{false\ negative}$) and the false positive rate ($R_{false\ positive}$) were calculated using **equations 3** and **7** using the genotyping data for 5 α1-antitrypsin deficiency patients. The false negative rate shown is for a child from a first-cousin marriage (m + n = 6). (B) The probability that RHSs contain the disease gene ($P_{GeneIsInRHS}$) calculated for a child from a first-cousin marriage. The coefficient of consanguinity (*F*) used was 1/16, which was calculated according to Wright {Wright, S. Systems of Mating. V. General Considerations *Genetics* 1921: **6**:167-178}. *F* can be more precisely calculated as the total length of RHSs divided by the total length of the autosomes for the actual calculation (**equation 9**). $P_{GeneIsInRHS}$ varies depending on the frequency of the gene in the population.

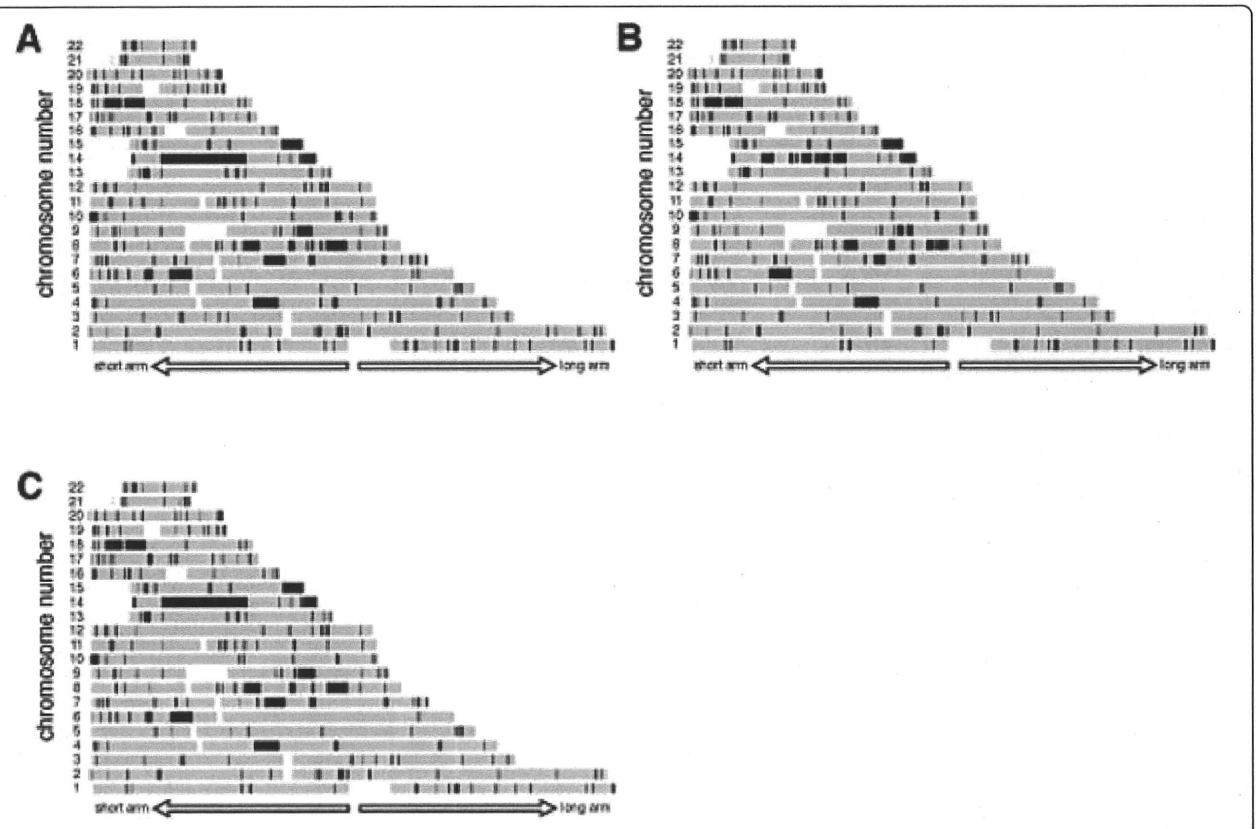


Figure 3 Genotyping error corrections. (A) RHSs for NA18987. (B) RHSs detected after introducing genotyping errors to 2,105 SNPs. (C) RHSs after the genotyping error correction algorithm was applied.

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 × 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 α1-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.

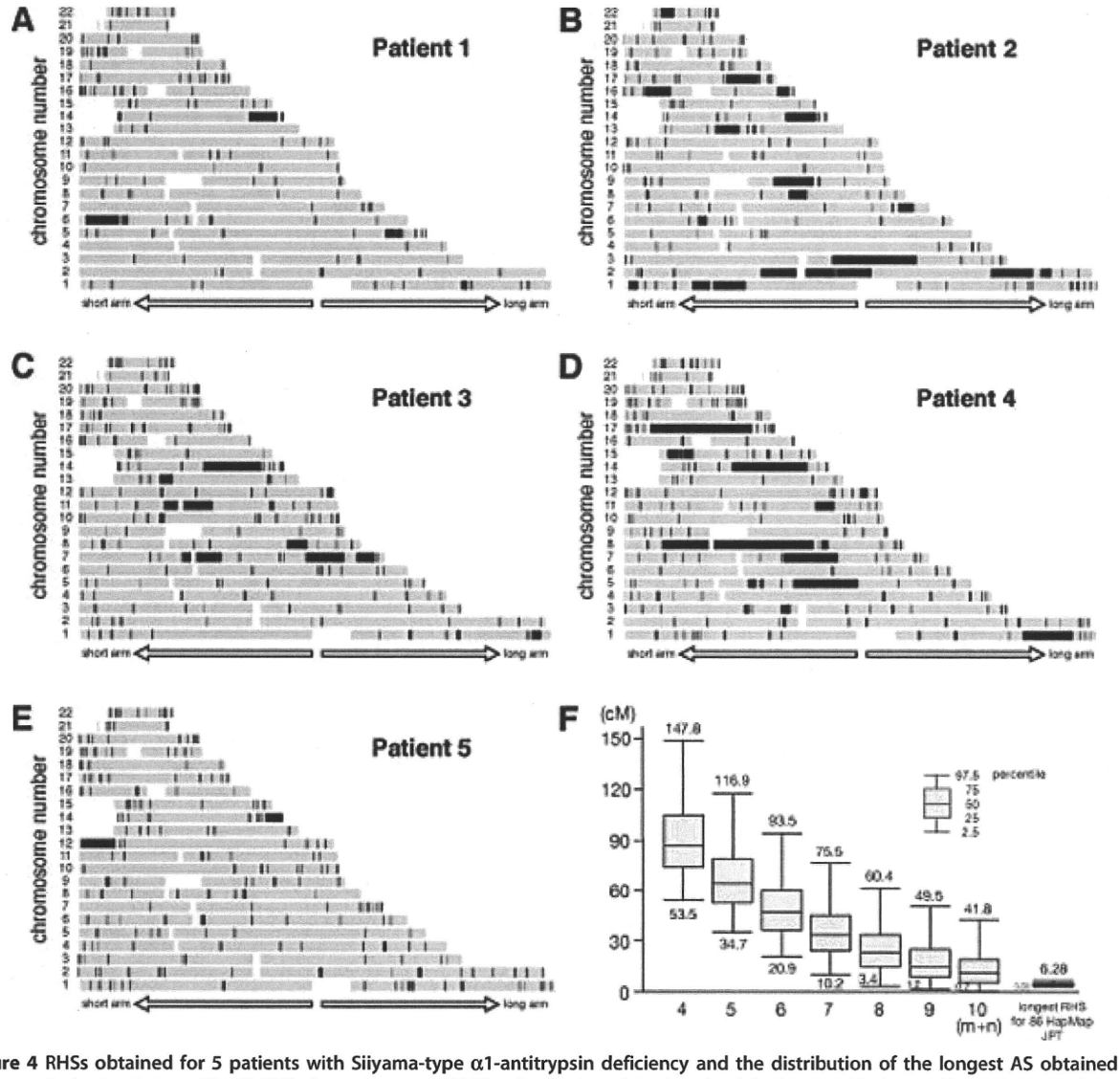


Figure 4 RHSs obtained for 5 patients with Siiyama-type α1-antitrypsin deficiency and the distribution of the longest AS obtained by a Monte Carlo simulation. (A) - (E) RHSs for each patient. (F) The distribution of the length of the longest AS obtained by a Monte Carlo Simulation. The distribution for 86 HapMap JPT patients is also shown in the right side.

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 ($P_{GeneIsInOverlap}$) was calculated by **equation 10** and is shown in Figure 5B. The prevalence of Siiyama-type α1-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

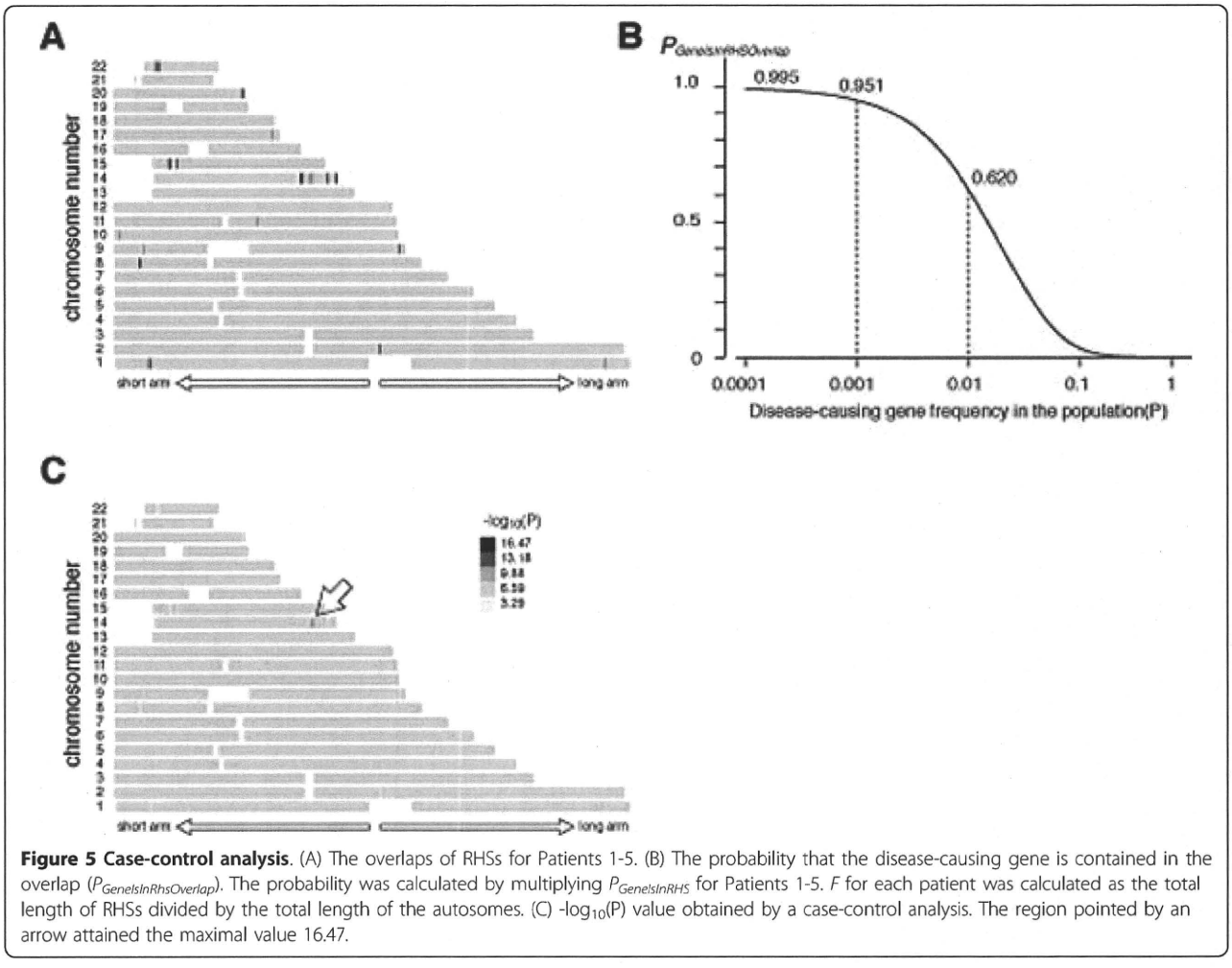


Figure 5 Case-control analysis. (A) The overlaps of RHSs for Patients 1-5. (B) The probability that the disease-causing gene is contained in the overlap ($P_{GeneIsInRhsOverlap}$). The probability was calculated by multiplying $P_{GeneIsInRHS}$ for Patients 1-5. *F* for each patient was calculated as the total length of RHSs divided by the total length of the autosomes. (C) -$\log_{10}$(P) value obtained by a case-control analysis. The region pointed by an arrow attained the maximal value 16.47.

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 antiproteinase, antitrypsin), member 10
SERPINA6	serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 6
LOC10028	Description: hypothetical protein LOC100287997
SERPINA2	serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 2
SERPINA1	serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 1
SERPINA11	serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 11
SERPINA9	serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 9
SERPINA12	serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, 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 α1-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 α1-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**

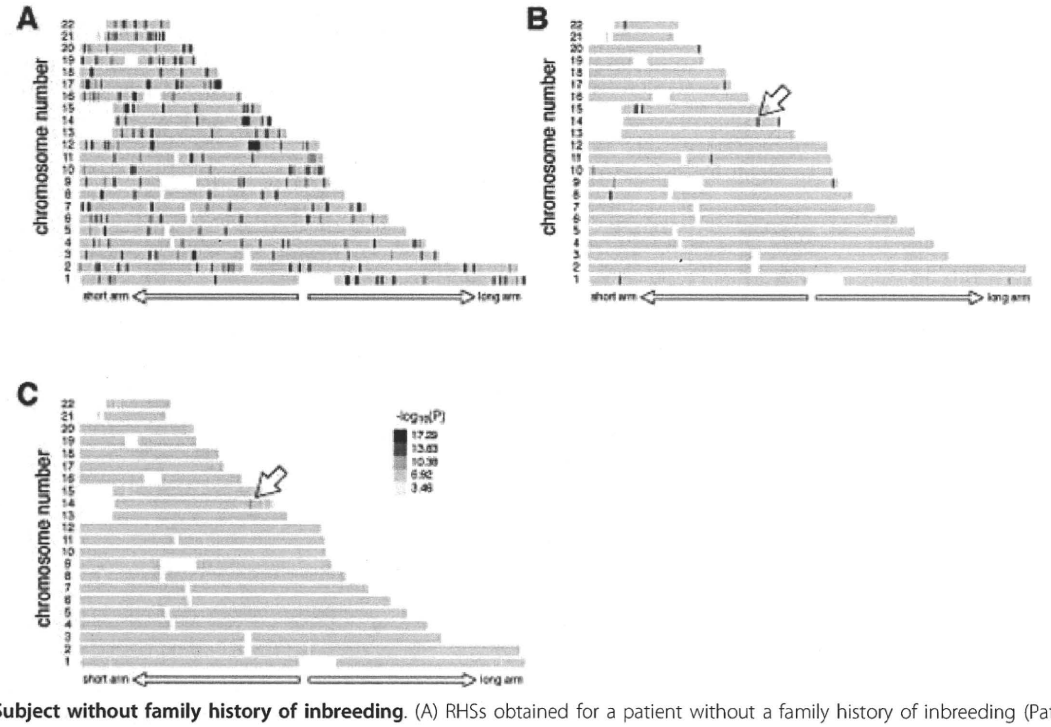


Figure 6 Subject without family history of inbreeding. (A) RHSs obtained for a patient without a family history of inbreeding (Patient 6). (B) RHS overlaps for Patients 1-6. Addition of data for Patient 6 further narrowed the overlapped regions (compare with **Figure 5**). The disease-causing gene was 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=S7.

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

Id4, a New Candidate Gene for Senile Osteoporosis, Acts as a Molecular Switch Promoting Osteoblast Differentiation

Yoshimi Tokuzawa[1☯], Ken Yagi[1☯], Yzumi Yamashita[1], Yutaka Nakachi[1], Itoshi Nikaido[1], Hidemasa Bono[1], Yuichi Ninomiya[1], Yukiko Kanesaki-Yatsuka[1], Masumi Akita[2], Hiromi Motegi[3], Shigeharu Wakana[3], Tetsuo Noda[3,4], Fred Sablitzky[5], Shigeki Arai[6], Riki Kurokawa[6], Toru Fukuda[7], Takenobu Katagiri[7], Christian Schönbach[8,9], Tatsuo Suda[1], Yosuke Mizuno[1], Yasushi Okazaki[1]*

1 Division of Functional Genomics and Systems Medicine, Research Center for Genomic Medicine, Saitama Medical University, Hidaka, Saitama, Japan, **2** Division of Morphological Science, Biomedical Research Center, Saitama Medical University, Iruma-gun, Saitama, Japan, **3** RIKEN BioResource Center, Tsukuba, Ibaraki, Japan, **4** The Cancer Institute of the Japanese Foundation for Cancer Research, Koto-ward, Tokyo, Japan, **5** Developmental Genetics and Gene Control, Institute of Genetics, University of Nottingham, Queen's Medical Center, Nottingham, United Kingdom, **6** Division of Gene Structure and Function, Research Center for Genomic Medicine, Saitama Medical University, Hidaka, Saitama, Japan, **7** Division of Pathophysiology, Research Center for Genomic Medicine, Saitama Medical University, Hidaka, Saitama, Japan, **8** Division of Genomics and Genetics, Nanyang Technological University School of Biological Sciences, Singapore, Singapore, **9** Department of Bioscience and Bioinformatics, Kyushu Institute of Technology, Iizuka, Fukuoka, Japan

Abstract

Excessive accumulation of bone marrow adipocytes observed in senile osteoporosis or age-related osteopenia is caused by the unbalanced differentiation of MSCs into bone marrow adipocytes or osteoblasts. Several transcription factors are known to regulate the balance between adipocyte and osteoblast differentiation. However, the molecular mechanisms that regulate the balance between adipocyte and osteoblast differentiation in the bone marrow have yet to be elucidated. To identify candidate genes associated with senile osteoporosis, we performed genome-wide expression analyses of differentiating osteoblasts and adipocytes. Among transcription factors that were enriched in the early phase of differentiation, Id4 was identified as a key molecule affecting the differentiation of both cell types. Experiments using bone marrow-derived stromal cell line ST2 and *Id4*-deficient mice showed that lack of *Id4* drastically reduces osteoblast differentiation and drives differentiation toward adipocytes. On the other hand knockdown of *Id4* in adipogenic-induced ST2 cells increased the expression of *Pparγ2*, a master regulator of adipocyte differentiation. Similar results were observed in bone marrow cells of femur and tibia of *Id4*-deficient mice. However the effect of *Id4* on *Pparγ2* and adipocyte differentiation is unlikely to be of direct nature. The mechanism of Id4 promoting osteoblast differentiation is associated with the Id4-mediated release of Hes1 from Hes1-Hey2 complexes. Hes1 increases the stability and transcriptional activity of Runx2, a key molecule of osteoblast differentiation, which results in an enhanced osteoblast-specific gene expression. The new role of Id4 in promoting osteoblast differentiation renders it a target for preventing the onset of senile osteoporosis.

Citation: Tokuzawa Y, Yagi K, Yamashita Y, Nakachi Y, Nikaido I, et al. (2010) Id4, a New Candidate Gene for Senile Osteoporosis, Acts as a Molecular Switch Promoting Osteoblast Differentiation. PLoS Genet 6(7): e1001019. doi:10.1371/journal.pgen.1001019

Editor: Gregory S. Barsh, Stanford University School of Medicine, United States of America

Received January 27, 2010; **Accepted** June 4, 2010; **Published** July 8, 2010

Funding: This work was supported by grants-in-aid of the Genome Network Project and Support Project of Strategic Research Center in Private Universities from the Ministry of Education, Culture, Sports, Science, and Technology (MEXT) to Saitama Medical University Research Center for Genomic Medicine. The funders had no role in study design, data collection and analysis, 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: okazaki@saitama-med.ac.jp

☯ These authors contributed equally to this work.

Introduction

Senile osteoporosis or age-related osteopenia is accompanied by increased bone marrow tissue adiposity [1]. Bone marrow adipocytes and osteoblasts are thought to originate from common mesenchymal stem cells (MSCs). Therefore, it has been suggested that the excessive accumulation of marrow adipocytes following bone loss is caused by unbalanced differentiation of MSCs into marrow adipocytes and osteoblasts [2]. Support for this hypothesis comes from studies of *peroxisome proliferators-activated receptor-γ* (*Pparγ*), a master regulator of adipocyte differentiation, deficient embryonic stem cells that showed an increase in osteoblast differentiation [3]. In contrast, calvarial adipocyte differentiation is augmented when *runt-related transcription factor 2 (Runx2)*, a master regulator of osteoblast differentiation has been knocked out [4]. Transcription factors Runx2 and *Sp7 transcription factor 7* (*Sp7*) regulate MSC commitment to osteoblast differentiation along with bone morphogenetic protein (BMP) signaling pathway [5]. Conversely, Pparγ and CCAAT/enhancer binding protein (C/EBP) transcription factor family members drive MSCs differentiation toward adipocytes [6]. Other proteins that regulate the balance between adipocyte and osteoblast differentiation are tafazzin, Wnt5a, Wnt10b, Msx2, C/EBPβ and basic helix-loop-helix (bHLH) family member e40 (Bhlhe40) [6–8]. Aforementioned transcription factors suppress adipocyte differentiation and promote osteoblast differentiation. Regardless of

Author Summary

Increased bone marrow adiposity is observed in the bone marrow of senile osteoporosis patients. This is caused by unbalanced differentiation of mesenchymal stem cells (MSCs) into osteoblast or adipocyte. Previous reports have indicated that several transcription factors play important roles in determining the direction of MSCs differentiation into osteoblast or adipocyte. So far, little is known about the overall dynamics and regulation of transcription factor expression changes leading to the imbalance of osteoblast and adipocyte differentiation inside the bone marrow. We have performed genome-wide gene expression analyses during the differentiation of MSCs into osteoblast or adipocyte. We identified basic helix-loop-helix transcription factor family member Id4 as a leading candidate controlling the differentiation toward adipocyte or osteoblast. Suppression of Id4 expression in MSCs repressed osteoblast differentiation and increased adipocyte differentiation. In contrast, overexpression of Id4 in MSCs promoted osteoblast differentiation and attenuated adipocyte differentiation. Moreover, *Id4*-mutant mice showed abnormal accumulation of lipid droplets in bone marrow and impaired bone formation activity. In summary, we have demonstrated a molecular function of Id4 in osteoblast differentiation. The findings revealed that Id4 is a molecular switch enhancing osteoblast differentiation at the expense of adipocyte differentiation.

these studies, the precise molecular mechanisms that regulate the balance between osteoblast and adipocyte differentiation in the bone marrow has yet to be elucidated. Hence, we aimed to identify transcription factors that regulate the direction of differentiation toward osteoblast or adipocyte by analyzing their genome-wide expression profiles in differentiation time series experiments.

We noticed in early phases a subgroup of transcription factors that appeared to function in both osteoblasts and adipocytes differentiation. Particularly, bHLH superfamily transcription factors were significantly enriched and up-regulated in the early phase of osteoblast differentiation.

The bHLH superfamily comprises transcription factors that form homo- or heterodimers and typically bind to a consensus sequence (CANNTG) called an E-box [9]. It is well known that bHLH transcription factors play important roles in development and cell differentiation. For example, Myod1 is a key differentiation factor of myoblasts and Srebf1 is involved in adipocyte differentiation [10,11]. Hairy and enhancer of split (Hes) family members of bHLH superfamily are crucial regulators of cortical development [12].

Here, we have identified Inhibitor of DNA binding 4 (Id4), which also belongs to the bHLH superfamily as a key molecule that regulates the direction of differentiation toward osteoblast or adipocyte *in vitro* and *in vivo* using genome wide expression study. Furthermore, we established that Id4 promotes osteoblast differentiation by enhancing Runx2 transcriptional activity through stabilization of Runx2 protein. The new role of Id4 in directing osteogenic and adipogenic cell fate makes it a likely target for preventing the onset of senile osteoporosis.

Results

Genome-wide expression profile predicts *Id4* as a candidate molecular switch in osteoblast and adipocyte differentiation

To delineate the sequential changes of transcription factors activating and repressing downstream osteogenic and/or adipogenic target genes, we evaluated the differentiation capability toward both osteoblasts and adipocytes using six cell lines (ST2, C2C12, DFAT-D1, PA6, 10T1/2, NRG). Of these, bone marrow-derived stromal cell line ST2 differentiated most efficiently into both osteoblasts and adipocytes (data not shown). Using Affymetrix mouse GeneChip, we aimed to identify clusters of transcription factors that are temporally co-regulated in one but not in another cluster (CIBEX Accession number: CBX90). Of 1,270 transcription factors, 407 genes were significantly up- or down- regulated in either osteoblast or adipocyte differentiation compared to the non-induced control (Table 1 and Table S1). Hierarchical clustering analysis of transcription factor gene expression data at 15 osteoblast and seven adipocyte differentiation time points (Figure 1A) revealed distinct clusters that represent phases of sequentially expressed transcription factors (Figure 1B). Differentiation into osteoblasts is characterized by five phases (Figure S1) whereas adipocyte differentiation resulted in four phases (Table S1). The early phases of osteoblast (1 hr) and adipocyte (48 hr) differentiation showed the greatest variability in transcription factor expression levels (Figure 2A and Table S1).

Chi-square testing for over-representation of transcription factors in each differentiation phase supported only six up-regulated bHLH superfamily members (*Id1*, *Id2*, *Npas4*, *Id4*, *Hes1* and *Bhlhe40*) of the immediate early phase osteoblast differentiation (1 hr) as significantly ($p<0.01$) enriched (Figure 2B). Since *Id4*, *Hes1* and *Bhlhe40* expression increased (decreased) twofold or greater during osteoblast (adipocyte) differentiation compared to the control (Figure 2C and Table 2), these transcription factors are likely to play a pivotal role in the regulation of osteoblast and adipocyte differentiation.

Indeed, Hes1 and Bhlhe40 are known to be involved in both differentiation pathways [8,13,14], whereas Id4 has not yet been implicated in either differentiation pathways. Additionally, *Id4* expression patterns in osteoblast and adipocyte differentiation were also compared by quantitative real-time PCR (qRT-PCR). Expression of *Id4* significantly increased during osteoblast differentiation, attained a peak on day 4 and decreased thereafter (Figure S2A). In contrast, *Id4* expression decreased during adipocyte differentiation (Figure S2B). Expression levels of *Id1* and *Id2* were also up-regulated in the early stage (1 hr) of osteoblast differentiation, but thereafter their expression dropped

Table 1. Expression behavior of 1,270 transcription factors selected from all mouse genes (Ensembl release 52) based on GO IDs (Table 3) during osteoblast and adipocyte differentiation.

Transcription Factors		Osteoblast Differentiation				
		↑	↓	↑↓	≈	Total
Adipocyte Differentiation	↑	22	6	4	70	102
	↓	25	40	21	132	218
	↑↓	3	5	2	2	12
	≈	42	26	7	863	938
	Total	92	77	34	1,067	1,270

Arrows indicate up- (↑), down-regulated (↓), or up- and down-regulated (↑↓) transcription factors. No change in expression is symbolized by the almost equal sign (≈).

doi:10.1371/journal.pgen.1001019.t001

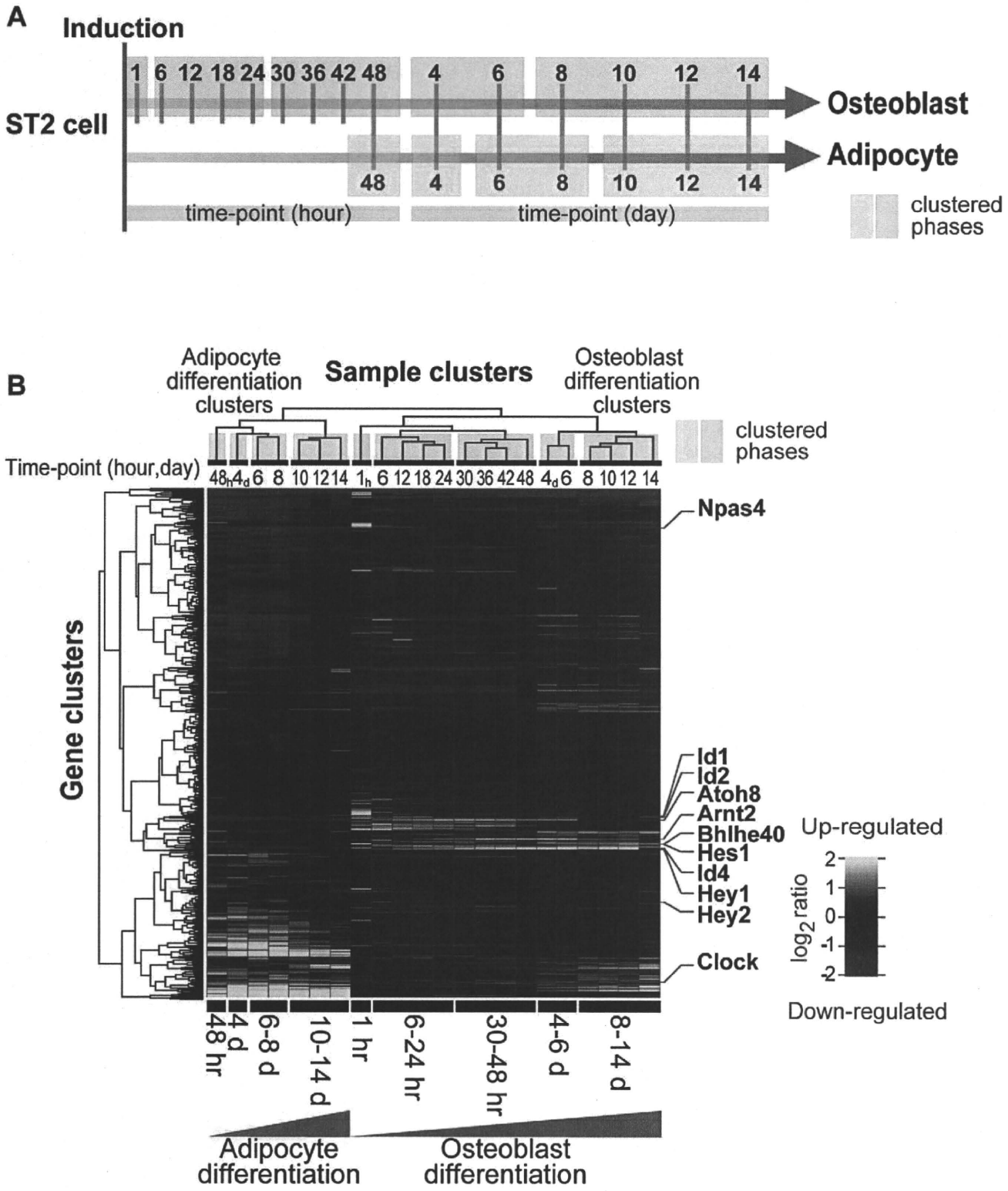
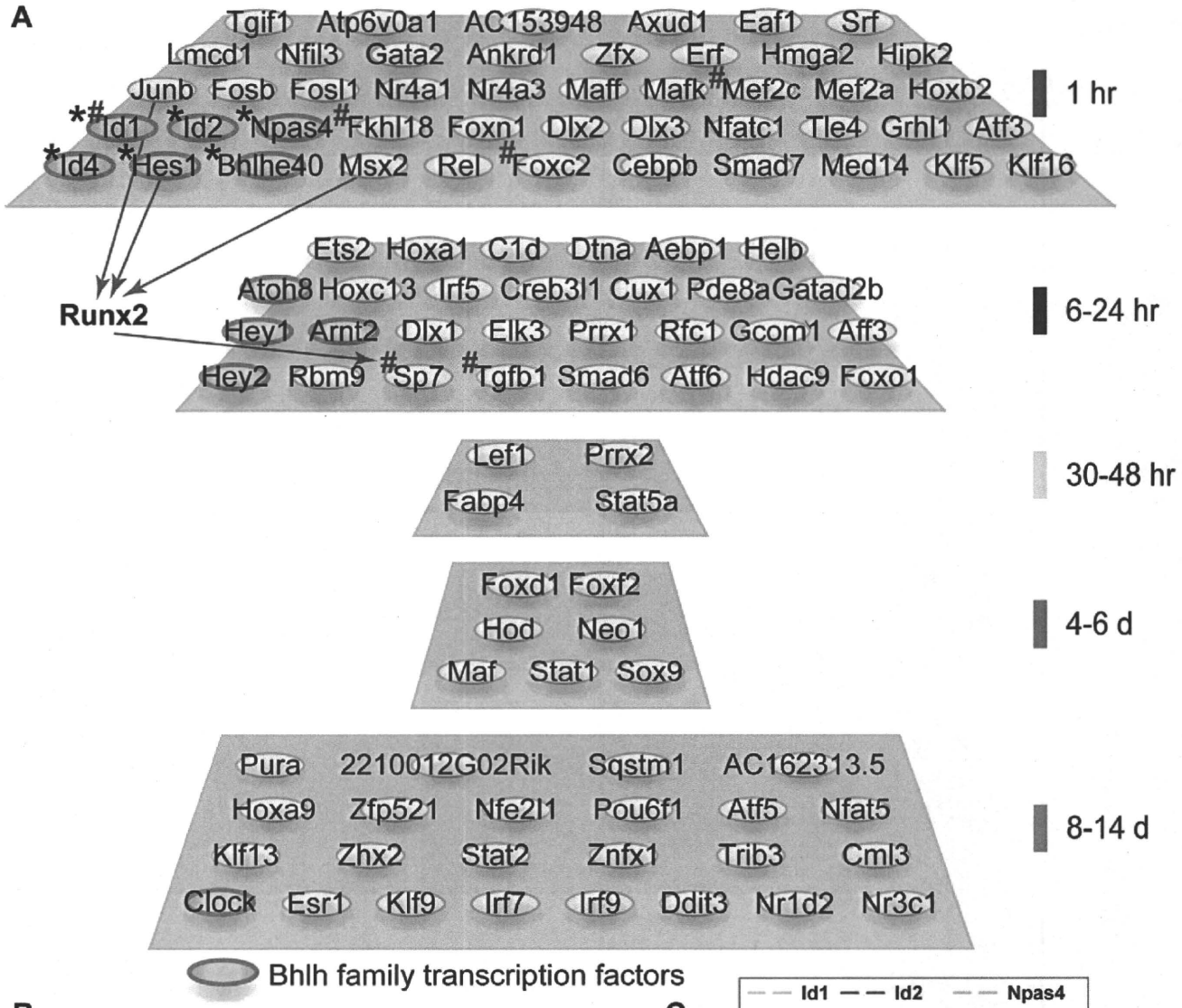


Figure 1. Transcription factor gene expression framework of differentiating ST2 cells. (A) Time-course sampling for gene expression analyses in ST2 cells. Vertical bars crossing the gradated blue and orange arrows indicate the sampling times during osteoblast and adipocyte differentiation, respectively. Five blue (osteoblasts) and four orange (adipocytes) boxes indicate the hierarchical clustering-derived phases of differentiation. (B) Gene expression heat map and clustering results of 1,270 transcription factors in 22 time-course samples.
doi:10.1371/journal.pgen.1001019.g001

to base levels (Figure 2C, upper panel). Therefore, we hypothesized that Id4 may act as a novel molecular switch in osteoblast and adipocyte differentiation.

Aside from *Id4*, *Hes1* and *Bhlhe40*, we identified additional bHLH members and various hypothetical and non-bHLH transcription factors as phase-specific candidate regulators of



B

Observerd	Up-regulated at time point 1h	All genes other than up-regulated ones at 1h	Total no.
Bhlh family	6	106	112
All other genes	249	19,039	19,288
Total no.	255	19,145	19,400

Expected	Up-regulated at time point 1h	All genes other than up-regulated ones at 1h	Total no.
Bhlh family	1.47	110.53	112
All other genes	253.53	19,034.47	19,288
Total no.	255	19,145	19,400

C

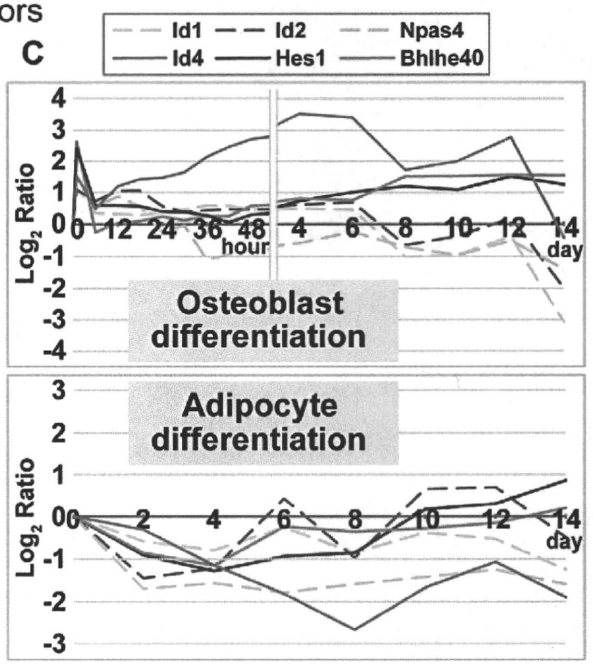


Figure 2. Enrichment of bHLH family in early phase of osteoblast differentiation and their expression pattern. (A) Summary of up-regulated transcription factors (Figure S1) associated with five phases of osteoblast differentiation. Asterisk (*): bHLH transcription factors that are overrepresented only during the early phase of osteoblast differentiation ($p<0.01$, Figure 2B). Sharp (#): transcription factors known to be associated with osteoblast differentiation and/or bone development. (B) Chi-square (χ^2) test for independence of bHLH superfamily genes in the immediate early phase (1 hr) of osteoblast differentiation. Null hypothesis: the occurrence of up-regulated bHLH transcription factors observed during the immediate early phase of osteoblast differentiation among all genes associated with probe set of Affymetrix GeneChip Mouse Genome 430 2.0 array is independent. Results of the contingency table (2×2 cells, degree of freedom = 1) allowed to reject the Null hypothesis. The expression of bHLH transcription factors was significantly up-regulated during the immediate early phase of osteoblast differentiation ($p<0.01$). (C) Expression pattern of six bHLH family genes in osteoblast and adipocyte differentiation. The chart represents expression changes of six bHLH genes during osteoblast and adipocyte differentiation. The horizontal axis indicates each time point after induction. The vertical axis indicates the logarithmic expression ratio (base 2) of each time point in comparison to the control (0 day).
doi:10.1371/journal.pgen.1001019.g002

osteo- and adipogenesis (Figure 2A and Table S1). The genes listed in Table S1 await further functional characterization regarding their involvement in osteoblast and/or adipocyte differentiation.

Id4 knockdown suppresses osteoblast differentiation of MSCs, and overexpression promotes osteoblast differentiation

To evaluate the potential role of Id4 in ST2 osteoblast differentiation, Id4 was suppressed by siRNA knockdown. As shown in Figure 3A, *Id4* siRNA (si*Id4*)-treated ST2 cells differentiating into osteoblasts showed a significant decrease in Id4 expression. The decline of *Id4* expression was accompanied by weak alkaline phosphatase (ALP) activity and reduced *bone γ carboxyglutamate protein 1* (*Bglap1* also called *osteocalcin*) expression (Figure 3B and 3C). Since both are markers of osteoblast differentiation, it appeared that Id4 is important in osteoblast differentiation of MSCs. We next evaluated whether forced expression of Id4 can promote osteoblast differentiation in MSCs by retroviral systems (Figure 3D). ST2 cells infected with Id4 recombinant retrovirus showed increased expression levels of *ALP* and *Bglap1* compared to cells infected with control virus independent of the presence or absence of BMP4 (Figure 3E and 3F). Taken together, Id4 promotes osteoblast differentiation of MSCs.

Id4 knockdown stimulates adipocyte differentiation, and overexpression attenuates lipid accumulation

Id4 knockdown in adipogenesis-induced ST2 cells (Figure 4A) significantly increased expression levels of other adipogenic marker genes such as *Pparγ2* [15] and *Adipoq* [16] (Figure 4B and 4C). Concomitantly, the number of Oil Red O stained lipid droplets and triglyceride levels also increased (Figure 4D and 4E). Forced expression of Id4 was confirmed by transfection into Cos7 cells with Id4 expression vector (Figure 4F). Adipocytes differentiated from ST2 cells transfected with Id4 expression vector showed slightly but significantly decreased lipid accumulation compared to empty vector transfectants (Figure 4G). The combined results of *Id4* siRNA knockdown and overexpression in ST2 suggest that Id4 attenuates differentiation of MSCs into adipocytes.

Id4-knockout mice ($Id4^{-/-}$) showed impaired osteoblast differentiation

Previously $Id4^{-/-}$ mice were studied only in context of neural development [17]. We confirmed that *Id4* expression was highest in the brain followed by cortical bone, kidney, thymus and bone marrow of C57BL/6J mice (Figure 5A). The body length and weight of 4 weeks old $Id4^{-/-}$ mice was 13–15% shorter and 35–40% lower compared to wild-type ($Id4^{+/+}$) littermates (Figure 5B and 5C). $Id4^{-/-}$ mice showed severe growth retardation and died by 5 weeks. In addition, we observed visible skeletal phenotypes of $Id4^{-/-}$ mice, but no skeletal deformities (data not shown). Altogether, our data hint at an important role of Id4 in bone formation.

Bone histological analysis of $Id4^{-/-}$ mice revealed significantly decreased bone volume (BV) in the 6th lumbar (Figure 5D and 5E).

Table 2. Expression of eleven bHLH transcription factors shown in Figure 2A.

	Differentiation		Differentiation Clusters	
Gene Symbol	**Osteoblast**	**Adipocyte**	**Osteoblast**	**Adipocyte**
Bhlhe40	↑	↓	1h ↑	4d ↓
Hes1	↑	↓	1h ↑	4d ↓
Id1	↑ ↓	↓	1h ↑	10d–14d ↓
Id2	↑ ↓	↓	1h ↑	2d ↓
Id4	↑	↓	1h ↑	4d ↓
Npas4	↑ ↓	↓	1h ↑	2d ↓
Arnt2	↑	≈	6h–24h ↑	≈
Atoh8	↑ ↓	↓	6h–24h ↑	2d ↓
Hey1	↑	≈	6h–24h ↑	≈
Hey2	↑	≈	6h–24h ↑	≈
Clock	↑	↑	8d–14d ↑	10d–14d ↑

Of eleven transcription factors, *Bhlhe40*, *Hes1* and *Id4* were two-fold or greater up-regulated during osteoblast differentiation. On the other hand, these genes were one-half fold or greater down-regulated during adipocyte differentiation. Arrows indicate up- (↑), down-regulated (↓), or up- and downregulated (↑ ↓) bHLH genes. No change in expression is symbolized by the almost equal sign (≈). Crosstalk functions between osteoblast and adipocyte differentiation have been reported for Bhlhe40 [8], Hes1 [13,14] and Id4 (this study), only.
doi:10.1371/journal.pgen.1001019.t002

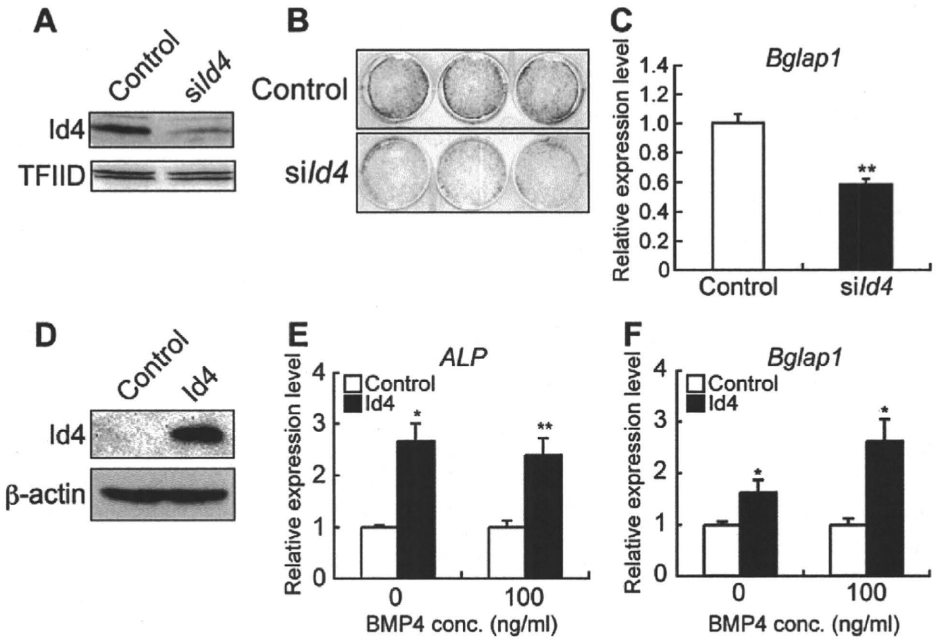


Figure 3. *Id4* promotes osteoblast differentiation in ST2 cells. (A) Western blotting of anti-Id4 antibody using ST2 nuclear extract. TFIID was detected as a loading control. Specific reduction of Id4 expression at day 2 after induction of ST2 cells treated with *Id4* siRNA (si*Id4*) for osteoblast differentiation. (B,C) Decrease in ALP activity (B) and *Bglap1* expression (C) at day 6 after induction of osteoblast differentiation in ST2 cells treated with si*Id4* as compared to the negative control. (D) Western blotting of anti-Id4 antibody using ST2 cell lysate, infected with Id4 retrovirus or control retrovirus (control). β-actin was detected as a loading control. (E,F) Increase in *ALP* (E) and *Bglap1* (F) expression at day 4 after infection with *Id4* virus as compared to the control virus in non-induced ST2 and ST2 osteoblast. Relative expression levels of *ALP* and *Bglap1* mRNA were measured by qRT-PCR. qRT-PCR data were subjected to Student's t-tests. Each error bar represents the mean ± SE of triplicates. *$p<0.05$ versus control, **$p<0.01$ versus control.

doi:10.1371/journal.pgen.1001019.g003

The bone formation rate (BFR) was decreased in $Id4^{-/-}$ mice compared to $Id4^{+/+}$ mice (Figure 5F and 5G). The BV to total volume ratio, BFR to bone surface (BS) ratio and mineral apposition rate (MAR) of $Id4^{-/-}$ mice were 57.3%, 28.1% and 30.7% lower compared to $Id4^{+/+}$ mice in the 6th lumbar, respectively (Figure 5N–5P).

In $Id4^{+/+}$ mice, active cuboidal-shaped osteoblasts (type II osteoblasts) were distributed in a row along the lumbar BS (Figure 5H), whereas in the corresponding region of $Id4^{-/-}$ mice osteoblasts were predominantly flat and resting (type IV osteoblasts; lining cells) (Figure 5I). The number of osteoblasts as a whole did not change significantly between $Id4^{+/+}$ and $Id4^{-/-}$ mice (data not shown). However, in $Id4^{-/-}$ mice the population of active osteoblasts was reduced (type II, Figure 5Q), whereas inactive osteoblasts accumulated (type IV, Figure 5R). These findings imply that Id4 modulates both differentiation of osteoblasts from pre-osteoblasts and regulation of osteoblast maturation.

Impaired bone formation was also observed in the lateral calvaria of $Id4^{-/-}$ mice (Figure 5K and 5M). In wild type mice, osteoblasts were closely lined up along the calvarial BS (Figure 5J). In contrast, no osteoblasts were observed along the calvarial bone of $Id4^{-/-}$ mice (Figure 5K). The osteoid thickness, BFR to BS ratios and MAR of $Id4^{-/-}$ mice calvarial bones were 61.5%, 49.1% and 65.2% of $Id4^{+/+}$ mice, respectively (Figure 5S, 5O, and 5P). These results suggest that Id4 is important for both endochondorial and membranous ossification. Growth Plate Width and Longitudinal Growth Rate (Lo. G. R) of $Id4^{-/-}$ mice tibia were 68.4% and 57.1% of $Id4^{+/+}$ mice, respectively (Figure S3A, S3B, S3C, S3D), which may have caused the growth retardation of $Id4^{-/-}$ mice.

Id4 interacts with Hey2 and inhibits the transcriptional repression of Hey2

Id family members are known to heterodimerize with other bHLH transcriptional factors, thus inhibiting the binding to the E-box motif [18]. To explore whether heterodimerized Id4 switches the direction of osteoblast and adipocyte differentiation, we assayed Id4 protein-protein interactions and analyzed their effects. Using immunoprecipitation we attempted to capture for candidate bHLH transcription factors that bind to Id4. Out of four tested bHLH transcription factors (Hes1, hairy/enhancer-of-split related with YRPW motif 1; Hey1, hairy/enhancer-of-split related with YRPW motif 2; Hey2 and Bhlhe40) (Figure S4A), only Hey2 bound to Id4 (Figure 6A, Figure S4A and S4B). An earlier study demonstrated that Hey2 is forming heterodimers with Hes1, which then bind to the E-box motif and repress transcription [19]. Therefore, we tested the effect of Id4 on transcriptional repression of Hey2/Hes1 heterodimers against the E-box element. Transcriptional repression onto E-box element in the presence of either Hey2 or Hes1 showed no effect or 39% inhibition of E-box transcriptional activity relative to control (empty expression vector), respectively (Figure 6B). Although luciferase activity was lowest in the presence of both Hey2 and Hes1, inhibition of transcriptional repression by Id4 increased in dose-dependent manner (Figure 6B). We also demonstrated that Hey2-Hes1 binding was abrogated with the dose-dependent increase of Id4 (Figure 6C, lane 5 and lane 6). Taken together, we confirmed that Id4 reverses the transcriptional repression by Hey2-Hes1 heterodimer in a dose-dependent manner.

Id4/Hey2 complex indirectly enhances Runx2 transcriptional activity

The presence of inactive osteoblasts (Figure 5I and 5R) in the bone tissues of the $Id4^{-/-}$ mice let us assume that Id4 may affect

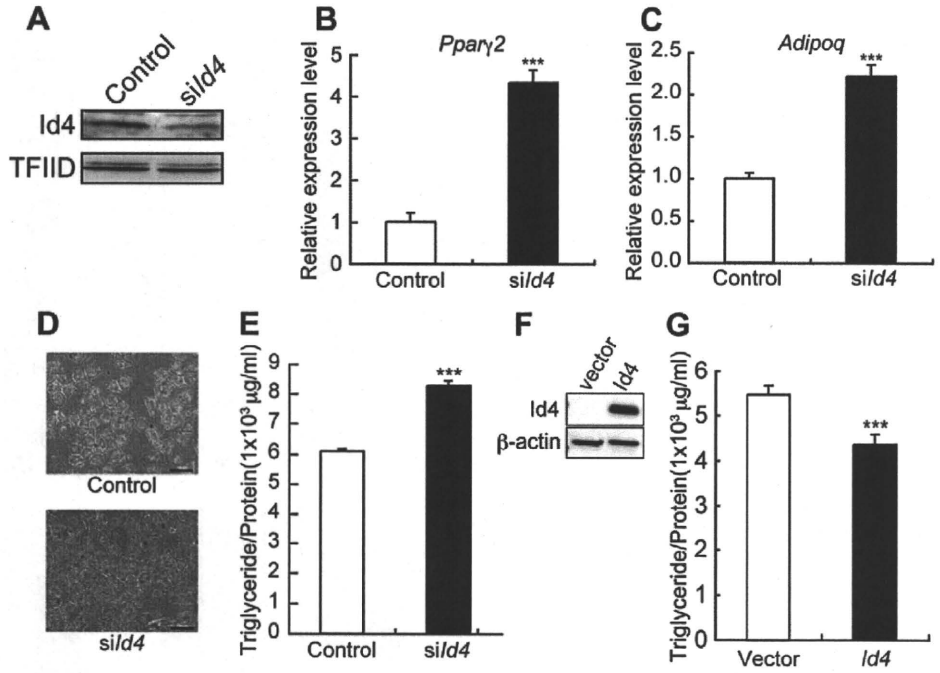


Figure 4. *Id4* weakly suppresses adipocyte differentiation in ST2 cells. (A) Western blotting of anti-Id4 antibody using ST2 nuclear extract. TFIID was detected as a loading control. Specific reduction of Id4 expression at day 2 after induction of ST2 cells treated with si*Id4* for adipocyte differentiation. (B,C) Increase in *Pparγ2* (B) and *Adipoq* (C) expression at day 1 after induction of adipocyte differentiation and treatment with si*Id4*. Relative expression levels of *Pparγ2* and *Adipoq* mRNA were measured by qRT-PCR. (D) Oil Red O staining of ST2 cells at day 4 after induction of adipocyte differentiation in presence of control or si*Id4*. Original magnification: ×200; bar = 100 μm. (E) Lipid content of ST2 cells measured as triglyceride level at day 5 after adipogenic induction in presence of control or si*Id4*. (F) Western blotting of anti-Id4 antibody using Cos7 cell lysate, transfected with Id4 expression vector or empty vector (Vector) as a control. β-actin was detected as a loading control. (G) Lipid content of ST2 cells transfected with Id4 expression vector or empty vector measured as triglyceride level at day 5 after adipogenic induction. qRT-PCR and lipid content data were subjected to Student's t-tests. Each error bar represents the mean ± SE of triplicates. *$p<0.05$ versus control and ***$p<0.005$ versus control.

doi:10.1371/journal.pgen.1001019.g004

the actions of Runx2 and Sp7, a key osteogenic differentiation molecule [20–22]. Osteoblast marker gene *Bglap1* expression level decreased not only in primary osteoblasts but also in embryonic day18.5 (E18.5) limb of $Id4^{-/-}$ mice (Figure S5A and S5B). *Bglap1*, a target gene of Runx2 has an E-box element other than osteoblast-specific element 2 (OSE2), which binds Runx2 to the promoter [23]. Therefore, we measured the promoter activity to examine the influence of Id4 on *Bglap1* E-box promoter-dependent transcriptional activity of Runx2. Although the suppression of *Bglap1* promoter activity by addition of Hes1 and Hey2 was not detected, a dose-dependent increase of transcriptional activity by Id4 was observed when testing the OSE2 element-containing promoter. When using the promoter without OSE2 the increase in transcriptional activity was not seen (Figure 7A). Since direct interaction between Id4 and Runx2 was ruled out experimentally (data not shown), Id4 may indirectly influence Runx2 transcriptional activity through Hes1. Hes1 is known to stimulate the transcriptional activity of Runx2 protein by increasing its stability during osteoblast differentiation [24]. We also confirmed that the addition of Hes1 stabilizes Runx2 protein. Interestingly, the addition of Id4 further increased the stabilization and accumulation of Runx2 (Figure 7B). Taken together, it appears that Id4 enhances Runx2 transcriptional activity through stabilization of Runx2 protein.

Taking into account the timing of the elevated *Id4* expression (Figure 2C and Figure S2A), our results strongly suggest that Id4 is indirectly driving the Hes1-mediated Runx2 stabilization during osteoblast differentiation. The facts that both Hes1 and Hey2 bind to the OSE2 element-containing *Bglap1* promoter region assessed by ChIP-qPCR, and that the amount of bound Hes1 and Hey2 decreased in ST2 osteoblasts (Figure 7C) further support this idea. However, the exact binding site of Hes1-Hey2 heterodimer remains to be identified. The proposed mechanism of Id4 action during osteoblast differentiation is illustrated in Figure 7D.

$Id4^{-/-}$ mice showed increased adipocytes

Id4 knockdown promoted adipocyte differentiation in ST2 cells (Figure 4B–4E). Histological analysis of $Id4^{-/-}$ tibia bones revealed elevated numbers of adipocytes in epiphyseal bone marrow of tibia compared to $Id4^{+/+}$ mice (Figure 8A–8D). The entire analyzed area of epiphyseal tibia bone marrow was occupied by adipocytes in $Id4^{-/-}$ mice (Figure 8E). Moreover, *Pparγ2* expression levels were also increased in bone marrow cells of femur and tibia of $Id4^{-/-}$ mice (Figure 8F). In comparison to $Id4^{+/+}$ mice, the number of adipocytes in the lateral calvaria was markedly increased in $Id4^{-/-}$ mice (Figure 5K). These aberrant traits observed in $Id4^{-/-}$ mice implicate Id4 as a crucial molecule in the lineage choice of MSCs differentiating into either osteoblasts or adipocytes (Figure 8G).

Discussion

In this study, we have delineated clusters of transcription factors that act as key regulators in the osteoblast and adipocyte differentiation network. The observation of sometimes disparately regulated transcription factors, led us to hypothesize a molecular

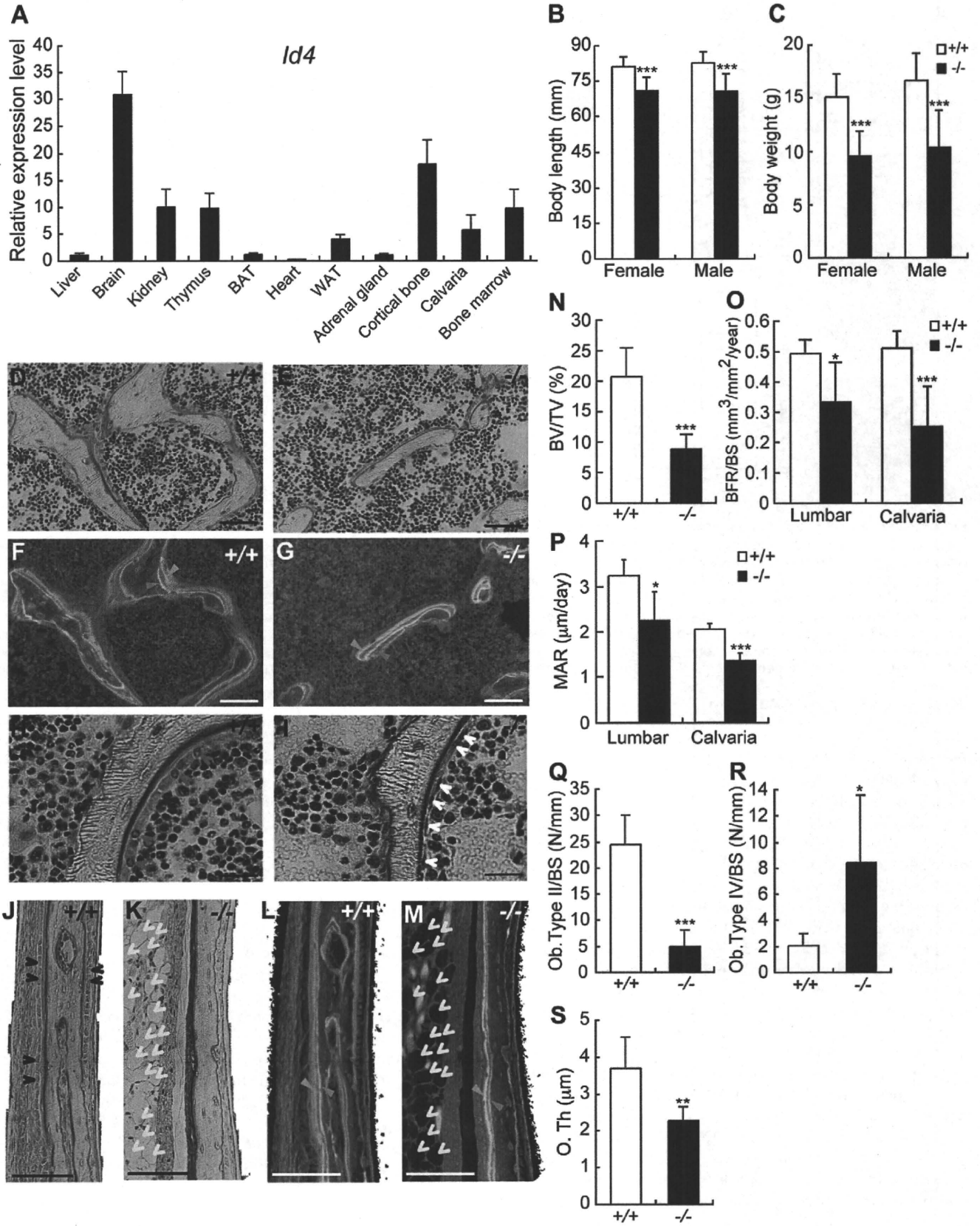


Figure 5. 4-week-old *Id4* knockout (*Id4*$^{-/-}$) mice show disturbance in growth and impaired osteoblast differentiation. (A) Tissue distribution of *Id4* mRNA in C57BL/6J male mice was measured by qRT-PCR. Brown adipose tissues; BAT, White adipose tissues; WAT. (B,C) Body length (B) and body weight (C) of wild-type (*Id4*$^{+/+}$) and *Id4*$^{-/-}$ mice. (D,E,H,I) Villanueva staining of the 6th lumbar bone of *Id4*$^{+/+}$ mice (D,H) and *Id4*$^{-/-}$ mice (E,I). Red and white arrowheads indicate active cuboidal-shaped osteoblasts (type II osteoblasts) and flattened resting osteoblasts (type IV

osteoblasts), respectively. (F,G,L,M) Bone formation rate (BFR) in the lumbar regions (F,G) and the lateral calvarial bone (L,M) of $Id4^{+/+}$ (F,L) and $Id4^{-/-}$ (G,M) mice. The BFR was measured using fluorescence microscopy following double-staining with tetracycline hydrochloride (blue arrowhead) and calcein (green arrowhead). Yellow arrowheads indicate adipocytes. (J,K) Villanueva staining of the lateral calvarial bone of $Id4^{+/+}$ (J) and $Id4^{-/-}$ mice (K). Black arrowheads indicate osteoblasts. The original magnifications and scale bars of the images are ×400 and 50 μm (D–G), ×400 and 30 μm (H,I) and ×200 and 50 μm (J–M). (N) Bone volume (BV) to total volume (TV) ratio in the 6th lumbar bone. (O,P) BFR to bone surface (BS) ratio (O) and mineral apposition rate (MAR) (P) in the 6th lumbar bone and lateral calvarial bone. (Q,R) Number of type II osteoblasts (Q) and type IV osteoblasts (R) on the corresponding area of the lumbar BS. (S) Osteoid thickness (O. Th) in the lateral calvarial bone. N-S: $Id4^{+/+}$ (n = 6), $Id4^{-/-}$ (n = 6). All data were subjected to Student's t-tests. Each error bar represents the mean ± SE. *$p<0.05$ versus control, **$p<0.01$ versus control and ***$p<0.005$ versus control.
doi:10.1371/journal.pgen.1001019.g005

switch function that promotes lineage-specific differentiation (e.g. osteoblast differentiation) and inhibits alternative differentiation pathways (e.g. adipocyte differentiation). To substantiate the hypothesis, we clustered gene expression profiles during osteoblast/adipocyte differentiation and identified Id4 as a candidate regulator of cell lineage choice. Id family members Id1–4 also belong to the bHLH superfamily, but lack the DNA binding domain. Heterodimerization of Id proteins with other bHLH proteins facilitates dominant negative regulation. A study by Bedford *et al.* using $Id4^{-/-}$ mice established Id4 as a regulator of proliferation and differentiation of neural precursor cells [17]. Until now, however, bone and skeletal abnormalities of this mouse model have not been reported. Besides our report, expression profile studies of BMP-independent osteoblast differentiation of human MSCs and BMP2-induced osteoblast differentiation using calvarial cells derived from Runx2-deficient mice [25,26] demonstrated that the expression level of *Id4* gene increases during osteoblast differentiation. These results suggest that Id4 plays an important role during osteoblast differentiation. Yet, the precise mechanism of Id4 action and regulation remained enigmatic. We have clearly demonstrated by *in vitro* and *in vivo* loss-of-function analysis that Id4 enhances osteogenic differentiation. Furthermore, we established a model of Id4 playing the role of molecular switch in osteoblast differentiation of MSCs (Figure 7D). Id4 expression increases upon BMP-induced osteoblast differentiation. Accumulating Id4 proteins transiently interact with Hes1-bound Hey2, thus triggering the release of Hes1 and the formation of Id4-Hey2 heterodimers and Hes1-Runx2 complexes. The binding of Hes1 to Runx2 potentiates the transcriptional activity of Runx2 and therefore osteoblast differentiation.

Hes1 and Hey2 are the target molecules of Notch signaling [19]. Until now, the relationship between Notch and BMP signaling pathways has been characterized by conflicting reports. On one hand experiments using ST2 cells showed that Notch1 suppresses the BMP-induced differentiation into osteoblasts [27]. On the other hand, Nobta et al. [28] reported that Notch signaling enhances BMP-induced osteoblast differentiation of C2C12 or MC3T3-E1 cells. At this point, there are insufficient data to resolve the controversy about the role of Notch signaling in BMP-induced osteoblast differentiation. However, our model (Figure 7D) may help to clarify the function of Notch signaling. In the absence of Id4, Hes1-Hey2 heterodimer just occupies the promoter region of the Runx2 target gene, whereas in the presence of Id4, Id4-Hey2 and Hes1-Runx2 complexes increase simultaneously. The concomitant increase of both complexes enhances the transcriptional activity of Runx2. Thus, we propose that availability and concentration of Id4 might account for the disparate roles of Notch signaling.

In $Id4^{-/-}$ mice, the drop of calvarial BFR is consistent with the phenotype of *Id1/Id3* heterozygous knockout mice. After BMP stimulation, *Id1/Id3* heterozygous knockout mice-derived calvarial cells showed reduced proliferation activity compared to calvarial cells derived from wild-type mice [29]. Thus, the decreased rate of calvarial bone formation in $Id4^{-/-}$ mice might be the consequence of reduced osteoblast proliferation. The expression levels of *Id1*and *Id2* were also up-regulated in the early stage of BMP4-induced osteoblast differentiation (Figure 2A). Indeed, it has been reported that *Id1* is an important early gene in osteoblasts after BMP stimulation [30,31]. Although the biological significance of Id1 in the regulation of MSCs has to be elucidated in future studies, in ST2 cells, expression levels of *Id1* and *Id2* immediately returned to base levels (Figure 2C). In contrast, *Id4* expression levels in ST2 and 3T3-E1 cells continued to rise until 4 days (Figure 2C and Figure S2A) and 7days, respectively [32].

Systemic hormones and local cytokines are known to be central regulators of bone formation. Serum levels of growth hormone (IGF1) and thyroid hormones (T3 and T4) did not change significantly between $Id4^{+/+}$ and $Id4^{-/-}$ mice (data not shown). Hence, impaired bone formation is most likely independent of hormonal factors and caused by repression of osteoblast differentiation.

Since the expression level of *Id4* decreases during adipocyte differentiation (Figure 2C and Figure S2B), Id4 was believed to inhibit MSCs differentiation into adipocytes. We demonstrated that Id4 suppression promoted adipocyte differentiation (Figure 4B–4E and Figure 8A–8E), but Id4 overexpression slightly decreased lipid accumulation level in ST2 adipocytes (Figure 4G). In an effort to shed light on the molecular mechanism, we assayed the expression level of *Pparγ2*, a master regulator of adipocyte differentiation. *Pparγ2* expression increased in adipogenic-induced ST2 cells when *Id4* was knocked down (Figure 4B) and in bone marrow cells of femur and tibia of $Id4^{-/-}$ mice (Figure 8F). However, the results of luciferase reporter assays ruled out effects of Id4 on the promoter activity of *Pparγ2* (data not shown). *Pparγ2* down-regulation by Id4 might involve a yet unknown, indirect regulatory mechanism. We noticed that the number of osteoclasts increased in the tibial epiphyseal regions and the 6th lumbar vertebra (data not shown). In view of an earlier report on Pparγ promoting osteoclast differentiation by activating *c-fos* [33], we interpret the increased number of adipocytes and osteoclasts in 6th lumbar and tibial bone of $Id4^{-/-}$ mice as a result of indirect activation of *Pparγ1* and/or *Pparγ2* transcriptional activity by lack of Id4. To corroborate these assumptions, additional analyses of the relationships and interactions of Id4, Pparγ and c-fos in context of adipocyte and osteoclast differentiation are necessary.

In summary, delineating clusters of transcription factors is a powerful strategy to identify cell fate-determining members of regulatory networks. Concerted application of our genome-wide expression profiling analyses and validation of transcription factor candidate regulators synthesize knowledge of specific molecular mechanism underlying osteoporosis and/or metabolic disease. In case of Id4, our findings reflect the potential pen-ultimate position of Id4 in the Runx2 activation/repression, which permits the differential integration of various upstream signals. Since BMP and Notch signaling affect osteoblast differentiation at different phases of differentiation, modulation of Id4 expression may create new venues for treating the onset of osteoporosis.

Materials and Methods

Cell culture

ST2 cells were obtained from RIKEN BioResource Center (BRC, Tsukuba, Japan) and cultured as described [34]. Primary