Takeuchi F, Kato N et al.	Common variants at the *GCK, GCKR, G6PC2-ABCB11*, and *MTNR1B* loci are associated with fasting glucose in two Asian populations.	***Diabetologia***	53(2)	299-308	2010
Takeuchi F, Kato N et al.	Evaluation of genetic loci influencing adult height in the Japanese population.	***J Hum Genet.***	54(12)	749-52	2009
Nabika T *et al.*	The stroke-prone spontaneously hypertensive rat: still a useful model for post-GWAS genetic studies?	***Hypertens Res***	35(5)	477-484	2012
Hashimoto M, Nabika T *et al.*	Effects of hydrogen-rich water on abnormalities in a SHR.Cg-Leprcp/ NDmcr rat - a metabolic syndrome rat model.	***Medical Gas Research***	3(1)	26	2011
Sakurai-Yamashita Y, Nabika T *et al.*	Blood pressure-independent factors determine the susceptibility to delayed neuronal death in the stroke-prone spontaneously hypertensive rats.	***Cell Mol Neurobiol***	30(2)	283-7	2010
Nabika T	Quality over quantity? No, quality and quantity.	***Hypertens Res***	33(2)	110-1	2010
Han Y, Fukuda N *et al.*	Role of complement 3a in the synthetic phenotype and angiotensin II-production in vascular smooth muscle cells from spontaneously hypertensive rats.	***Am J Hypertens***	25(3)	284-9	2012
Yoshida Y, Fukuda N *et al.*	Treatment with valsartan stimulates endothelial progenitor cells and renal label-retaining cells in hypertensive rats.	***J Hypertens***	29(1)	91-101	2011

Yamamoto C, Fukuda N *et al.*	Protective effects of statin on cardiac fibrosis and apoptosis in adrenomedullin knockout mice with the angiotensin II and high salt loading.	***Hypertens Res***	34(3)	348-353	2011
Fukuda N	Transforming growth factor-β as a treatment target in renal diseases.	***J Nephrol***	22(6)	708-15	2009

Method

Detection of common single nucleotide polymorphisms synthesizing quantitative trait association of rarer causal variants

Fumihiko Takeuchi,[1,2,6] Shotai Kobayashi,[3] Toshio Ogihara,[4] Akihiro Fujioka,[5] and Norihiro Kato[1]

[1]*Department of Gene Diagnostics and Therapeutics, Research Institute, National Center for Global Health and Medicine, Tokyo 162-8655, Japan;* [2]*Pathogen Genomics Center, National Institute of Infectious Diseases, Tokyo 162-8640, Japan;* [3]*Shimane University Hospital, Izumo 693-8501, Japan;* [4]*Department of Geriatric Medicine and Nephrology, Osaka University Graduate School of Medicine, Suita 565-0871, Japan;* [5]*Amagasaki Health Medical Foundation, Amagasaki 661-0012, Japan*

Genome-wide association (GWA) studies have identified hundreds of common (minor allele frequency ≥5%) single nucleotide polymorphisms (SNPs) associated with phenotype traits or diseases, yet causal variants accounting for the association signals have rarely been determined. A question then raised is whether a GWA signal represents an "indirect association" as a proxy of a strongly correlated causal variant with similar frequency, or a "synthetic association" of one or more rarer causal variants in linkage disequilibrium ($D' \approx 1$, but r^2 not large); answering the question generally requires extensive resequencing and association analysis. Instead, we propose to test statistically whether a quantitative trait (QT) association of an SNP represents a synthetic association or not by inspecting the QT distribution at each genotype, not requiring the causal variant(s) to be known. We devised two test statistics and assessed the power by mathematical analysis and simulation. Testing the heterogeneity of variance was powerful when low-frequency causal alleles are linked mostly to one SNP allele, while testing the skewness outperformed when the causal alleles are linked evenly to either of the SNP alleles. By testing a statistic combining these two in 5000 individuals, we could detect synthetic association of a GWA signal when causal alleles sum up to 3% in frequency. Such signal only partially explains the heritability contributed by the whole locus. The proposed test is useful for designing fine mapping after studying association of common SNPs exhaustively; we can prioritize which GWA signal and which individuals to be resequenced, and identify the causal variants efficiently.

[Supplemental material is available for this article. The synthetic association test software is freely available at http://www.fumihiko.takeuchi.name/PUBLICATIONS/synthetic.R.]

Genome-wide association (GWA) studies have identified hundreds of common (minor allele frequency [MAF] ≥5%) single nucleotide polymorphisms (SNPs) associated with a few hundred traits or diseases, yet the associated SNPs and their proxies mostly do not show evident function related to the target trait, and eventual identification of causal variants accounting for GWA signals has been challenging (Wellcome Trust Case Control Consortium 2007; McCarthy et al. 2008). A question that is then raised is whether a common SNP identified in a GWA study represents an "indirect association" as a proxy of a strongly correlated causal variant with similar frequency, or a "synthetic association" of one or more rarer causal variants that are in linkage disequilibrium (LD) ($D' \approx 1$, but r^2 not large) with the common SNP (Cirulli and Goldstein 2010; Dickson et al. 2010).

Synthetic association accounted for GWA signals in several studies. In a GWA study for dose of anticoagulant drug warfarin, the strongest association signal in the *CYP2C9* gene was observed at an SNP rs4917639, whose minor allele (frequency 18%) is a composite of two functional alleles *CYP2C9*2* (rs1799853, frequency 11%) and *CYP2C9*3* (rs1057910, frequency 7%) (Wadelius et al. 2007; Takeuchi et al. 2009). In a GWA study for anemia in patients treated for chronic hepatitis, the strongest signal was observed at an SNP rs6051702 in *C20orf194*, whose minor allele (frequency 19%) is almost exactly a composite of two causal variants in the neighboring *ITPA* gene (frequency 8% and 12%) (Fellay et al. 2010). When there are many rare causal variants, but no common one, as in the *HBB* gene for sickle cell anemia or the *GJB2/GJB6* locus for hearing loss, the association of common SNPs detected in GWA studies were attributable to the rare variants (Dickson et al. 2010). Using simulations, Dickson and colleagues showed that synthetic association is likely to occur when there are multiple rare variants in a locus (Dickson et al. 2010).

In general, identification of the causal variants accounting for a synthetic association requires extensive resequencing and association analysis. Instead, here we propose to test statistically whether a quantitative trait (QT) association of an SNP represents a synthetic association or not by inspecting only the QT distribution at each genotype of the SNP, without a priori knowledge about rarer causal variants. We focus on two statistics of the QT distribution: the heterogeneity of variance (i.e., heteroscedasticity) among SNP genotypes and the skewness. The statistical tests were examined in real data of the apolipoprotein E (*APOE*) gene, and in simulated data for representative models of synthetic association. Moreover, we formulated a general mathematical model of synthetic association, and assessed the test statistics theoretically. The two statistics were suitable for complementary scenarios: Heteroscedasticity was more sensitive than skewness when low-frequency

[6]Corresponding author.
E-mail fumihiko@takeuchi.name.
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.115832.110.

21:1122–1130 © 2011 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/11; www.genome.org

(<5%) causal alleles were linked mostly to one SNP allele, while skewness outperformed when the causal alleles were linked in balance to either of the two SNP alleles. We thus devised a test combining the two statistics, which was powerful for any of the assumed models.

Results

Test of heteroscedasticity

We first show a schematic example of synthetic association and illustrate how QT variance can differ among individuals classified by marker SNP genotypes. We assume a common marker SNP with alleles A and a, and a single causal variant with alleles B_1 and b_1. The allele B_1 (5% in frequency) is always linked to allele A (20% in frequency); thus, existing haplotype classes are AB_1, Ab_1, and ab_1. We assume the QT is normally distributed with the unit variance and the mean equal to 2, 1, and 0 within a subgroup of individuals having genotype B_1/B_1, B_1/b_1, and b_1/b_1, respectively. The QT distribution in the whole population becomes a mixture of the normal distributions combined according to the frequency of genotypes B_1/B_1, B_1/b_1, and b_1/b_1 (Fig. 1A). Individuals with A/A genotype at the marker SNP are enriched with the genotypes of B_1/B_1 and B_1/b_1 at the causal variant, thus their QT distribution widens (Fig. 1B). On the contrary, individuals with a/a genotype at the marker all have b_1/b_1 genotype at the causal variant, and their QT variance equals one (Fig. 1D). The QT variance is the largest in the subgroup with A/A genotype, which is linked more frequently to the low-frequency causal allele B_1, and the smallest in the subgroup with a/a genotype. Indeed, QT variance among individuals of a specific marker genotype enlarges proportionally to two factors: the variance of the causal genotype within the subgroup, and the squared effect-size of the causal allele (equation M2). The low-frequency causal variant causes the synthetic association of the marker SNP, and the heteroscedasticity of QT distribution among the marker genotypes.

We could exemplify the detection of synthetic association using heteroscedasticity in the *APOE* gene, which is known to associate with LDL cholesterol (LDL-C) level through three classical isoforms coded by two functional (or causal) variants—rs7412 (Arg158Cys) and rs429358 (Cys112Arg). As compared with E3 (the most common isoform), E2 (coded by rs7412) and E4 (coded by rs429358) decreased and increased the LDL-C level, respectively (Weisgraber et al. 1981; Weisgraber 1994; Bennet et al. 2007). The two variants had MAF <10% (in Europeans and East Asians) and were not included in SNP chips of GWA scan (except for the recent ones containing >1 million SNPs). In a GWA study for lipids in 1210 Japanese (F Takeuchi, et al., in prep.), we initially found four SNPs near *APOE* to attain locus-wise significant P-values for LDL-C association, although any of these were not significant after adjustment for the two functional variants. When only the chip SNPs were analyzed, rs405509 and rs377702 showed statistically independent signals of association (Supplemental Fig. 1). In a larger panel of 4840 individuals, the association signals remained at the two chip SNPs, and heteroscedasticity was significant for rs405509 (P = 0.019) (Table 1). Indeed, the causal minor alleles of rs7412 (T) and rs429358 (C) were linked to alternate alleles of rs405509 (C and A, respectively), demonstrating synthetic association (Fig. 2). The two causal variants could simultaneously enlarge the QT variance at all three genotypes of rs405509, and consequently, diminish heteroscedasticity (equation M5). However, in this case, as the effect-size of rs7412 was much larger than that of rs429358,

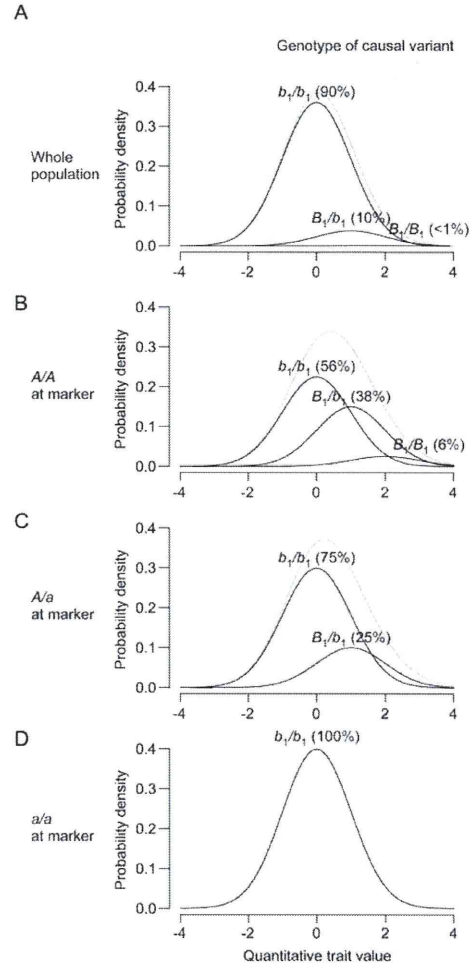

Figure 1. Probability distribution of the QT value within subgroups classified by marker SNP genotypes. (*A*) In the whole population, the total QT distribution (gray curve) comprises a mixture of normal distributions (black curves) with unit variance and the mean 0, 1, or 2, which correspond to genotypes b_1/b_1, B_1/b_1, and B_1/B_1 at the causal variant. As genotype B_1/B_1 is rare (0.25%), the corresponding curve appears flat. (*B*) QT distribution among individuals with A/A genotype at the marker. As B_1/B_1 and B_1/b_1 genotypes are enriched in this subgroup due to LD, the variance is enlarged, as noticeable from the lower peak and wider distribution of the gray curve. (*C*) Individuals with the A/a genotype have either genotypes b_1/b_1 or B_1/b_1, and the QT variance is moderately enlarged. (*D*) All individuals with a/a genotype at the marker have b_1/b_1 genotype at the causal variant. The QT variance is 1.10 in *A*, 1.38 in *B*, 1.19 in *C*, and 1 in *D*.

heteroscedasticity remained detectable; rs7412 enlarged the QT variance at C/C genotype of rs405509 to 1.182, whereas rs429358 kept the QT variance at A/A genotype at 0.978. On the other hand, the heteroscedasticity of rs377702 did not reach statistical significance due to its recombination with rs7412 (D' = 0.34). Thus, even if we identified the association signals at rs405509 and rs377702 via the GWA scan, by detecting heteroscedasticity we could notice the presence of synthetic association and the necessity to search for variants not on the chip.

We next estimated the power to detect synthetic association at an SNP that could be identified in a GWA study. We assumed

Table 1. Testing heteroscedasticity of SNPs in the *APOE* locus associated with LDL-C

SNP	Genotype	Number of individuals	Distribution of LDL-C level		Association with LDL-C level			Heteroscedasticity
			Mean	Variance	Beta	*P*-value	R^2	*P*-value
rs405509	C/C	462	−0.153	1.182	−0.117	1.0×10^{-7}	0.006	0.019
(GWAS SNP)	C/A	2035	−0.050	0.976				
	A/A	2343	0.073	0.978				
rs377702	T/T	32	−0.487	1.231	−0.191	5.1×10^{-7}	0.005	0.583
(GWAS SNP)	T/C	677	−0.149	1.025				
	C/C	4131	0.028	0.991				
rs7412	T/T	12	−1.302	1.079	−0.651	2.0×10^{-44}	0.040	0.92
(causal variant)	T/C	452	−0.584	0.981				
	C/C	4376	0.064	0.960				
rs429358	T/T	3954	−0.042	0.987	−0.212	1.4×10^{-9}	0.008	0.73
(causal variant)	T/C	850	0.185	1.023				
	C/C	36	0.214	1.104				

We first adjusted LDL-C level for body mass index and categories by sex and age (≤40, 41–50, 51–60, ≥61 yr) and then applied rank-based inverse normal transformation. Individuals under lipid treatment were excluded. Data are shown for 4840 individuals with complete observation from the Amagasaki study in Takeuchi et al. (2010).

that the marker SNP has MAF ≥5%, and that the proportion of QT variance explained by the marker is $R^2_{mrk} = 0.00592$, a borderline level to attain genome-wide significance (see Supplemental Notes). Figure 3 illustrates the statistical power for detecting heteroscedasticity in 5000 individuals. We examined four representative models of synthetic association by simulation. Under Model 1, there are l causal variants with alleles B_1 and b_1, B_2 and b_2, up to B_l and b_l, and the low-frequency causal alleles B_i have a uniform effect (e.g., increase QT) and are all linked to marker allele A. The QT variance enlarges for individuals with *A/A* genotype at the marker since they carry various numbers of the causal alleles, whereas individuals with *a/a* genotype at the marker carry none. Heteroscedasticity of the marker was detectable (power >0.8) in the region marked with an asterisk: For example, when the A allele frequency, $p_A \geq 45\%$, or alternatively when $p_A = 25\%$ and the cumulative frequency of causal alleles is <3%. For a fixed value of p_A, the power for detecting heteroscedasticity increases as the cumulative frequency of causal alleles decreases. When p_A becomes small, the detectable range narrows; the highest cumulative frequency in the detectable range changes proportionally to $\sqrt{p_A/(1-p_A)}$, as estimated in equation M12.

We next examine Models 2–4, where both of the marker alleles are loaded with low-frequency causal alleles. In addition to l causal variants with alleles B_1 and b_1, B_2 and b_2, up to B_l and b_l, there are m other causal variants with alleles C_1 and c_1, C_2 and c_2, up to C_m and c_m, and we designate the low-frequency alleles B_i and c_j as causal. The two groups of causal alleles, B_i and c_j, affect the QT in opposing directions and are linked to alternate alleles A and a of the marker, respectively, and thus synthetically generate the marker association. The QT variance at marker genotype *A/A* enlarges due to the causal alleles B_i, and the variance at marker genotype *a/a* enlarges due to the causal alleles c_j (equation M3). Indeed, the variances for all marker genotypes increase and become less heterogeneous than under Model 1. Under Model 2, there is exact balance in effect-size and cumulative frequency between the two groups of causal alleles. The heteroscedasticity disappears if $p_A = 50\%$ (equation M5), and became undetectably weak around the frequency (Fig. 3). The heteroscedasticity was detectable when p_A is close to 5% or 95%: For example, when $p_A = 15\%$ or 85% and the cumulative frequency of the causal alleles B_i, which equals the cumulative frequency of c_j, is <1%. Under Models 3 and 4, where the causal alleles B_i and c_j are not balanced, heteroscedasticity still disappeared, but around a different marker allele frequency. Under Model 3, the effect-size of the causal variants is uniform, yet the cumulative frequency of alleles c_j is half that of alleles B_i, and under Model 4, the cumulative frequencies are identical, yet the effect-size of alleles C_j is half that of alleles B_i. Heteroscedasticity was undetectable around $p_A = 65\%$ and 80% under Models 3 and 4, respectively. At $p_A = 25\%$ heteroscedasticity was detectable when the cumulative frequency of B_i alleles was <2%.

Test of skewness

As the test of heteroscedasticity could not detect synthetic association at a certain marker allele frequency around $p_A = 50\%$, when both alleles of the marker were loaded with low-frequency causal alleles (Fig. 3, Models 2–4) we introduced the test of skewness to cope with such a case. We observed that synthetic association skews the QT distribution at the marker genotypes *A/A* and *a/a* oppositely (equation M7): QT distribution among individuals with marker genotype *A/A* is skewed toward the effect direction of causal alleles B_i, and the QT distribution at genotype *a/a* is skewed toward the opposite direction, which is the effect direction of causal alleles

A

Haplotype class	rs405509 (GWAS SNP)	rs7412 (causal variant)	rs429358 (causal variant)	Frequency	Coded isoform
H1	C	T	T	0.049	E2
H2	C	C	T	0.256	E3
H3	A	C	T	0.599	E3
H4	A	C	C	0.096	E4

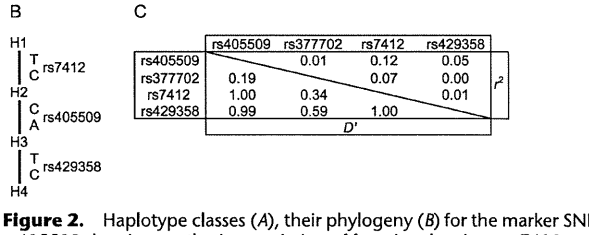

Figure 2. Haplotype classes (*A*), their phylogeny (*B*) for the marker SNP rs405509 showing synthetic association of functional variants rs7412 and rs429358 in the *APOE* locus. LD coefficients between the SNPs associated with LDL-C (C). Haplotype frequencies were calculated using the PLINK software (Purcell et al. 2007).

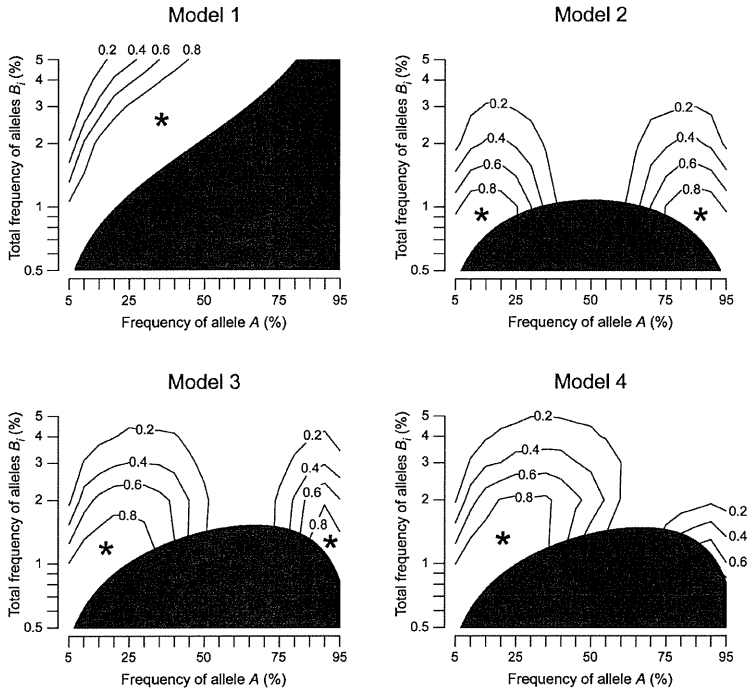

Figure 3. Power for detecting synthetic association by testing heteroscedasticity. The power was computed from simulation under four representative genetic models of synthetic association (see Methods), assuming the strength of marker association (R^2_{mrk}) of 0.00592. Horizontal and vertical axes represent the frequency of the marker allele *A*, and the cumulative frequency of causal alleles B_i (linked to allele *A*), respectively. The asterisk indicates the region where synthetic association is detectable with power >0.8. The black region of the parameter space should be neglected, as it does not include causal variants accounting for the marker association.

c_j. Thus, we added the skewness test statistics for the two genotypes, taking the direction into account (equation M8). Accordingly, under Model 2, the test of skewness could detect synthetic association around p_A = 50% (Fig. 4). The detectable range with regard to the cumulative frequency of causal alleles B_i is wide (extends up to 2%) when p_A is around 50% and is narrow when near 5% or 95%; as estimated in equation M14, the maximum detectable cumulative frequency is proportional to $(p_A(1-p_A))^{\frac{3}{4}}$. On the other hand, the skewness test was less powerful than the heteroscedasticity test when all causal alleles are linked to one marker allele (Model 1 in Fig. 4 vs. Fig. 3).

Combined test

Between the two tests to detect synthetic association, the test of heteroscedasticity was more powerful when one marker allele was loaded with the causal alleles (Model 1), and the test of skewness was more powerful when both of the marker alleles were loaded with a balanced amount of causal alleles (Model 2). By combining the two tests (equation M10), we devised the third test that was powerful under all of the models (Fig. 5). The detectable range for the cumulative frequency of B_i alleles exceeds 1% when the causal variants are exactly balanced (Model 2), and is up to 2% otherwise. Overall, we could detect synthetic association if the cumulative frequency of all causal alleles, B_i and c_j altogether, is <3%.

The power to detect synthetic association is influenced by the strength of marker SNP association and sample size. So far, we studied association at a borderline level of genome-wide significance, which is much weaker than some reported SNPs, for example, of lipid traits (Chasman et al. 2009). When the strength of association is doubled to R^2_{mrk} = 0.0118, synthetic association could be detected if the cumulative frequency of all causal alleles is <6%, in a wider range (Supplemental Fig. 2). When the sample size is halved to 2500 individuals, the detectable region narrowed (Supplemental Fig. 3), because the χ^2 statistics of the tests are proportional to the sample size (equations M5 and M9).

Discussion

As seen through mathematical analysis and simulation, we could detect an SNP representing synthetic association of rarer causal variant(s) by testing heteroscedasticity and skewness. The test only requires the genotype–phenotype data obtained in association studies, and the causal variants can be unknown. Whereas previous studies of synthetic association were based on empirical results and simulation (Dickson et al. 2010), we introduced a general mathematical formulation (see Methods) and estimated the variance and skewness of the marker SNP. We also performed computer simulations under representative models of synthetic association and obtained concordant results. The test of heteroscedasticity outperformed the test of skewness when low-frequency causal alleles were linked mostly to one SNP allele, while the test of skewness was better when the causal alleles were linked in balance to either of the two SNP alleles. The test combining the two could detect synthetic association if the cumulative frequency of causal alleles is <3% when tested in 5000 individuals for a marker SNP associated with QT at a borderline level of genome-wide significance (Fig. 5).

Genetic or environmental factors not correlated or interacting with the tested marker SNP do not skew the proposed test statistics. Thus, even when there is allelic heterogeneity, the variants not in LD with the marker SNP have no effects on the test. Although we modeled the causal variants to have an additive effect on QT, the mode of inheritance does not change the results, because homozygotes for a low-frequency allele are rare and negligible. In the power assessment by simulation, we modeled the causal variants to be in complete LD (D' = 1) with the marker. When LD decays, the heteroscedasticity or skewness at the marker becomes weaker and less detectable. However, since the marker is associated with QT, causal variant(s) of the same directional effect should be loaded mostly to one allele of the marker, thus the decay of LD would be limited.

There are a few limitations in using heteroscedasticity and skewness to detect synthetic association. False positives arise if a causal variant itself shows heteroscedasticity. This can result from a strong gene–environment interaction. Indeed, the test of heteroscedasticity has been used for detecting such interaction

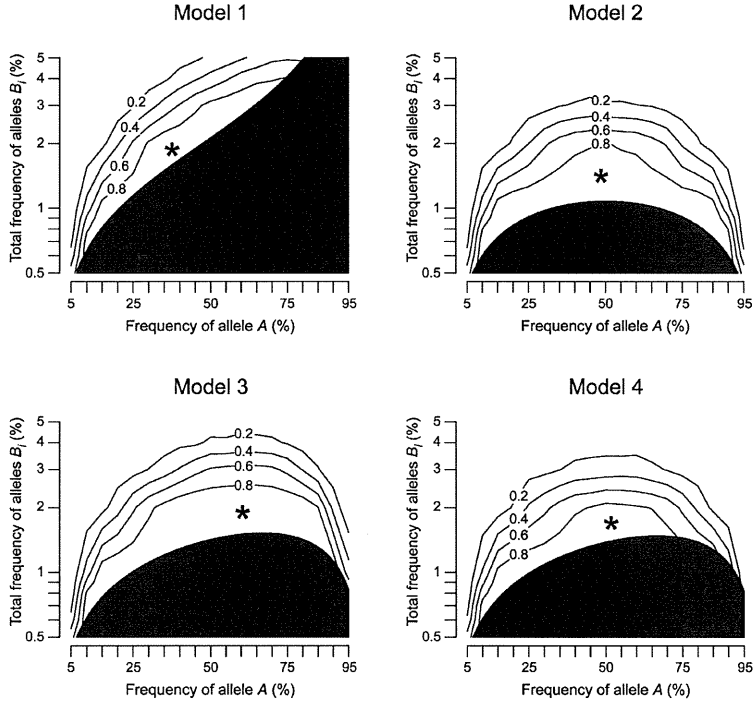

Figure 4. Power for detecting synthetic association by testing skewness. The power was computed from simulation under four representative genetic models, assuming the strength of marker association (R^2_{mrk}) of 0.00592. The format of the figure is the same as Figure 3.

(Pare et al. 2010). Another possible source of false positives is population stratification; if two subpopulations have a different mean QT at a specific causal genotype, the QT variance enlarges when the subpopulations are combined. Although a realistic level of population stratification is unlikely problematic (Supplemental Table 1), we recommend applying the test to each cohort separately.

Although the tests we proposed are limited to QTs, the idea of stratifying individuals by marker genotype leads to another test of synthetic association, which is applicable to quantitative as well as dichotomous traits. Here, we compare the association of neighboring common SNPs among the strata. If rare causal alleles are loaded onto the marker allele *A*, but not on the allele *a*, neighboring SNPs in LD with the causal variants will show association in the individuals with *A/A* or *A/a* genotypes, but not in the individuals with *a/a* genotype. As a result, the association *P*-value of neighboring SNPs would be distributed differently among the strata. Contrarily, if the marker SNP represents indirect association, no neighboring SNPs will be associated in any of the strata.

The proposed test can be helpful in understanding the "missing heritability" that GWA studies failed to account for (Maher 2008). If synthetic association is detected at a GWA signal SNP, the heritability of the SNP is likely an underestimate of the heritability of the whole locus (Dickson et al. 2010). Actually, in the *APOE* example, the explained variance was much smaller for the leading SNPs on a GWA chip ($R^2 = 0.006$ and 0.005) than for the two causal variants ($R^2 = 0.040$ and 0.008; see Table 1). In other words, our method can identify loci that contribute to a trait more than what we would expect from GWA study results.

It is unknown what proportion of GWA signals are due to synthetic association rather than indirect association. The situation is likely to differ by the function and molecular evolution of the genes. For example, several common causal variants are known for pharmacological traits that have not been under evolutionary selection (Cirulli and Goldstein 2010). On the contrary, only rare causal variants are found in genes for renal salt reabsorption, which have been under purifying selection; homozygotes of mutant alleles are susceptible to severe renal salt wasting and hypotension, although heterozygotes confer health benefits from lower blood pressure in postreproductive ages (Ji et al. 2008). Although the proposed test would help detecting synthetic association of a marker SNP, the discovery of the causal variants can require resequencing of a large number of individuals if the causal variants are rare. The aim of the proposed test is to assess potential synthetic association at a particular locus and then use the information to help design future resequencing studies.

Testing synthetic association is useful for designing fine mapping after exhaustively interrogating association of common SNPs at a locus. Exhaustive analysis of common SNPs (MAF ≥5%) is becoming accomplishable by genotyping with SNP chips of the GWA test and by imputing the unassayed SNPs using the HapMap or the 1000 genomes project data. The next focus is to explore rarer variants by resequencing and to identify the causal variants. Since resequencing is still expensive, we need to prioritize which GWA loci and which individuals are to be resequenced. Such information is obtainable by testing synthetic association of the common leading SNP(s), showing the strongest association in a locus. If synthetic association is detected for the leading SNP(s), rarer variants need to be examined in order to pinpoint the variants causing synthetic association. Moreover, if heteroscedasticity is detected, we can discover the causal variants efficiently by resequencing individuals having the homozygote genotype with larger QT variance, and especially those having extreme QT values, who are enriched with the rare causal alleles. Alternatively, if the test for synthetic association is not significant (in >5000 samples), the leading SNP(s) or their proxies are likely causal. Whereas conventional fine-mapping techniques aim to find the causal SNP(s) or haplotype(s) from a set of SNPs tested for association (McCarthy et al. 2008), our method is unique in suggesting that causal variants can be discovered if the study is extended to rarer variants.

Numerous SNP associations have been identified in recent GWA studies, yet our understanding of causal variants is very limited; it is not easy to prove functional changes, let alone the causality with the associated phenotype (Cirulli and Goldstein 2010). We proposed a simple statistical test, which helps to detect whether a common SNP associated with a QT is a noncausal marker in LD with rarer causal variant(s). The proposed test statistic can serve as a milepost in fine mapping and help understand the genetic structure of complex traits.

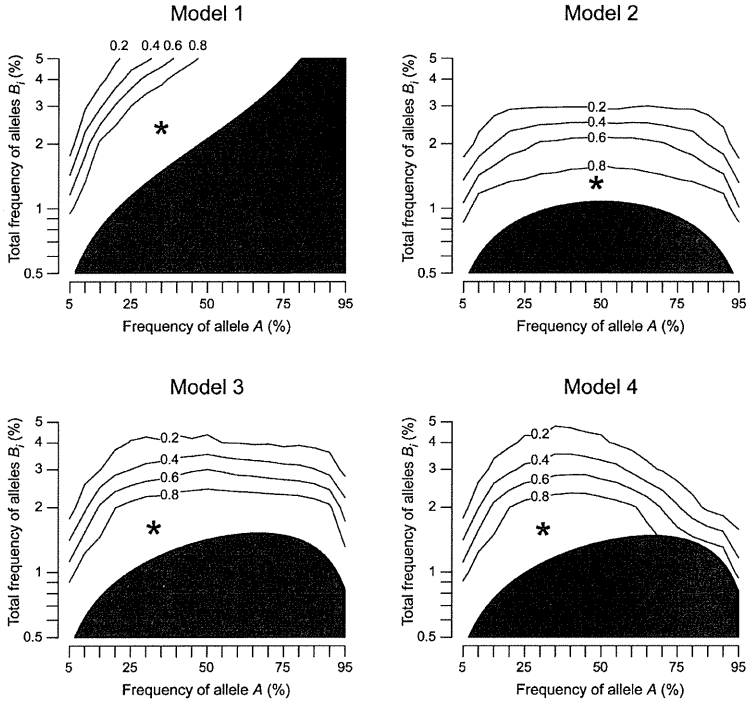

Figure 5. Power for detecting synthetic association by the combined test of heteroscedasticity and skewness. The power was computed from simulation under four representative genetic models, assuming the strength of marker association (R^2_{mrk}) of 0.00592. The format of the figure is the same as Figure 3.

the QT value is represented by a random variable q. The allele B_i (or C_j) is modeled to affect the QT by d_i (or e_j, respectively); specifically, the QT value q of an individual with multivariant genotype $(y_1,\ldots,y_l,z_1,\ldots,z_m)$ has the probability density of a normal distribution with the unit variance and the mean of $\sum_{i=1}^{l} d_i y_i + \sum_{j=1}^{m} e_j z_j$. Thus, the probability density function of the genotype and QT level $(x,y_1,\ldots,y_l,z_1,\ldots,z_m,q)$ becomes

$$p(x,y_1,\ldots,y_l,z_1,\ldots,z_m,q)=p(x,y_1,\ldots,y_l,z_1,\ldots,z_m)\times\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{\left(q-\sum_{i=1}^{l} d_i y_i-\sum_{j=1}^{m} e_j z_j\right)^2}{2}\right),$$

where $p(x,y_1,\ldots,y_l,z_1,\ldots,z_m)$ represents the frequency of the multivariant genotype $(x,y_1,\ldots,y_l,z_1,\ldots,z_m)$. The expectation of the QT value is

$$E[q]=\sum_{i=1}^{l} d_i E[y_i]+\sum_{j=1}^{m} e_j E[z_j],$$

where $E[\cdot]$ represents the expectation (see equation S5 in the Supplemental Notes for derivation). Similarly, when conditioned on a specific genotype x_0 at the marker,

$$E[q\mid x=x_0]=\sum_{i=1}^{l} d_i E[y_i\mid x=x_0]+\sum_{j=1}^{m} e_j E[z_j\mid x=x_0], \quad \text{(M1)}$$

where $E[\cdot|x=x_0]$ represents the conditional expectation.

Methods

Modeling the probability distribution of genotype and QT

We model a QT-associated marker SNP with alleles (referred to as the marker alleles), A and a. We assume the low-frequency allele of each causal variant—which we call the causal allele—is linked exclusively to one of the marker alleles; l causal variants each have alleles B_1 and b_1, B_2 and b_2, up to B_l and b_l, where the causal allele B_i is linked to marker allele A; m other causal variants each have alleles C_1 and c_1, C_2 and c_2, up to C_m and c_m, where the causal allele c_j is linked to marker allele a. We impose one assumption for mathematical convenience. Among the haplotype classes sharing a specific marker allele (A or a), we assume the probability distribution of causal variant alleles are independent among the variants; for example, among the haplotype classes carrying the A allele, the frequency (conditional on the marker allele being A) of the haplotype class carrying both B_1 and B_2 should equal the product of the frequencies of classes carrying B_1 and B_2, which is very small (e.g., 0.01%, if the frequency is 1% both for B_1 and B_2). The assumed frequency would differ only marginally from the actual frequency: The haplotype class carrying both B_1 and B_2 does not exist initially if the two causal variants arose separately in the phylogeny, and increases to the assumed frequency by recombination. This assumption enables us to rewrite the test statistics into simple forms (see Supplemental Notes). As we assume Hardy-Weinberg equilibrium, the frequencies of multivariant genotypes can be calculated from those of the haplotype classes.

Each individual's dose of the capital letter alleles, A, B_i, or C_j, is represented by random variables, x, y_i, or z_j, respectively, and

Variance

The variance of QT among individuals having a specific marker genotype x_0 becomes

$$\mathrm{Var}[q\mid x=x_0]=1+\sum_{i=1}^{l} d_i^2\mathrm{Var}[y_i\mid x=x_0]+\sum_{j=1}^{m} e_j^2\mathrm{Var}[z_j\mid x=x_0], \quad \text{(M2)}$$

where $\mathrm{Var}[\cdot|x=x_0]$ represents the conditional variance (see Supplemental equation S6 for derivation). Equation M2 indicates that the inflation (above one) of QT variance decomposes into a sum of terms, each corresponding to one causal variant and determined by the square of the effect-size, d_i or e_j, and the conditional variance of genotype. Designating the frequencies of alleles A, B_i and c_j as p_A, p_{B_i}, and p_{c_j}, respectively, the genotype variance becomes

$$\mathrm{Var}[y_i\mid x=2]=2\cdot\frac{p_{B_i}}{p_A}\left(1-\frac{p_{B_i}}{p_A}\right)\approx 2\cdot\frac{p_{B_i}}{p_A},$$
$$\mathrm{Var}[y_i\mid x=1]=\frac{p_{B_i}}{p_A}\left(1-\frac{p_{B_i}}{p_A}\right)\approx\frac{p_{B_i}}{p_A},$$
$$\mathrm{Var}[y_i\mid x=0]=0,$$
$$\mathrm{Var}[z_j\mid x=2]=0,$$
$$\mathrm{Var}[z_j\mid x=1]=\frac{p_{c_j}}{1-p_A}\left(1-\frac{p_{c_j}}{1-p_A}\right)\approx\frac{p_{c_j}}{1-p_A},$$
$$\mathrm{Var}[z_j\mid x=0]=2\cdot\frac{p_{c_j}}{1-p_A}\left(1-\frac{p_{c_j}}{1-p_A}\right)\approx 2\cdot\frac{p_{c_j}}{1-p_A},$$

where the approximation is under $p_A, 1-p_A >> p_{B_i}, p_{c_j}$. By substituting the genotype variance into equation M2, we obtain

$$\begin{aligned}\mathrm{Var}[q \mid x=2] &= 1+2\sum_{i=1}^{l} d_i^2 \frac{p_{B_i}}{p_A},\\ \mathrm{Var}[q \mid x=1] &= 1+\sum_{i=1}^{l} d_i^2 \frac{p_{B_i}}{p_A}+\sum_{j=1}^{m} e_j^2 \frac{p_{c_j}}{1-p_A},\\ \mathrm{Var}[q \mid x=0] &= 1+2\sum_{j=1}^{m} e_j^2 \frac{p_{c_j}}{1-p_A}.\end{aligned} \tag{M3}$$

Thus, the QT variance at marker genotypes $x = 2$ (*A/A*) and $x = 0$ (*a/a*) is determined by the contribution of alleles B_i and c_j, respectively, and the average of the two variances equals the variance at genotype $x = 1$ (*A/a*). The average of three variances weighted by marker genotype frequency becomes

$$\begin{aligned}E[\mathrm{Var}[q \mid x]] &= p_A^2\mathrm{Var}[q \mid x=2] + 2p_A(1-p_A)\mathrm{Var}[q \mid x=1]\\ &\quad + (1-p_A)^2\mathrm{Var}[q \mid x=0] = 1+2\left(\sum_{i=1}^{l} d_i^2 p_{B_i}+\sum_{j=1}^{m} e_j^2 p_{c_j}\right).\end{aligned} \tag{M4}$$

We test the heterogeneity of QT variance (i.e., heteroscedasticity) among marker genotypes using Bartlett's test (Bartlett 1937). The χ^2 statistic (two degrees of freedom) is expected to become

$$\begin{aligned}&\chi^2_{\text{heteroscedasticity}}\\ &= N[\ln(E[\mathrm{Var}[q \mid x]])-p_A^2\ln(\mathrm{Var}[q \mid x=2])\\ &\quad -2p_A(1-p_A)\ln(\mathrm{Var}[q \mid x=1]) - (1-p_A)^2\ln(\mathrm{Var}[q \mid x=0])]\\ &\approx N\left[-2\left(\sum_{i=1}^{l} d_i^2 p_{B_i}+\sum_{j=1}^{m} e_j^2 p_{c_j}\right)^2+2\left(\sum_{i=1}^{l} d_i^2 p_{B_i}\right)^2\right.\\ &\quad \left.+\left(\sqrt{\frac{1-p_A}{p_A}}\sum_{i=1}^{l} d_i^2 p_{B_i}+\sqrt{\frac{p_A}{1-p_A}}\sum_{j=1}^{m} e_j^2 p_{c_j}\right)^2+2\left(\sum_{j=1}^{m} e_j^2 p_{c_j}\right)^2\right]\\ &= N\left\{\left(\sqrt{\frac{1-p_A}{p_A}}\sum_{i=1}^{l} d_i^2 p_{B_i}\right)-\left(\sqrt{\frac{p_A}{1-p_A}}\sum_{j=1}^{m} e_j^2 p_{c_j}\right)\right\}^2,\end{aligned} \tag{M5}$$

when the sample size N is large (thus, the constant for the Bartlett's test statistic equals one); for the second equality, equations M3 and M4 were substituted, and $\ln(1+x)$ was approximated as $x-x^2/2$. In the curly brackets of the final formula in equation M5, the contribution by causal alleles B_i (linked to marker allele *A*) is subtracted by the contribution of alleles c_j (linked to allele *a*). Thus, the statistic for heteroscedasticity is maximized when all low-frequency causal alleles are linked to the same marker allele, and diminishes when they are linked evenly to both of the marker alleles.

Skewness

The third central moment of the QT distribution among the individuals having a specific marker genotype x_0 is

$$\mu_3[q \mid x=x_0] = \sum_{i=1}^{l} d_i^3\mu_3[y_i \mid x=x_0]+\sum_{j=1}^{m} e_j^3\mu_3[z_j \mid x=x_0], \tag{M6}$$

which decomposes into a sum of terms, each contributed by one causal variant (see Supplemental equation S7 for derivation); $\mu_3[\cdot|x=x_0]$ represents the conditional third central moment. The third central moment of the genotypes y_i and z_j conditional on a marker genotype becomes

$$\begin{aligned}&\mu_3[y_i \mid x=2] = 2\cdot\frac{p_{B_i}}{p_A}\left(1-\frac{p_{B_i}}{p_A}\right)\left(1-\frac{2p_{B_i}}{p_A}\right)\approx 2\cdot\frac{p_{B_i}}{p_A},\\ &\mu_3[y_i \mid x=1] = \frac{p_{B_i}}{p_A}\left(1-\frac{p_{B_i}}{p_A}\right)\left(1-\frac{2p_{B_i}}{p_A}\right)\approx \frac{p_{B_i}}{p_A},\\ &\mu_3[y_i \mid x=0] = 0,\\ &\mu_3[z_j \mid x=2] = 0,\\ &\mu_3[z_j \mid x=1] = -\frac{p_{c_j}}{1-p_A}\left(1-\frac{p_{c_j}}{1-p_A}\right)\left(1-\frac{2p_{c_j}}{1-p_A}\right)\approx -\frac{p_{c_j}}{1-p_A},\\ &\mu_3[z_j \mid x=0] = -2\cdot\frac{p_{c_j}}{1-p_A}\left(1-\frac{p_{c_j}}{1-p_A}\right)\left(1-\frac{2p_{c_j}}{1-p_A}\right)\approx -2\cdot\frac{p_{c_j}}{1-p_A},\end{aligned}$$

where the approximation is under $p_A, 1-p_A \gg p_{B_i}, p_{c_j}$. By substituting the genotype moment into equation M6, we obtain

$$\begin{aligned}\mu_3[q \mid x=2] &= \frac{2}{p_A}\sum_{i=1}^{l} d_i^3 p_{B_i},\\ \mu_3[q \mid x=1] &= \left(\frac{1}{p_A}\sum_{i=1}^{l} d_i^3 p_{B_i}\right)-\left(\frac{1}{1-p_A}\sum_{j=1}^{m} e_j^3 p_{c_j}\right),\\ \mu_3[q \mid x=0] &= -\frac{2}{1-p_A}\sum_{j=1}^{m} e_j^3 p_{c_j}.\end{aligned} \tag{M7}$$

The z statistic (standard normal distribution) for testing skewness (Stuart et al. 1999) of QT among the individuals with genotype x_0 becomes

$$\begin{aligned}z_{x=x_0} &= \sqrt{\frac{N_{x=x_0}}{6}}\frac{\mu_3[q \mid x=x_0]}{\mathrm{Var}[q \mid x=x_0]^{3/2}}\\ &\approx \sqrt{\frac{N_{x=x_0}}{6}}\mu_3[q \mid x=x_0],\end{aligned}$$

where $N_{x=x_0}$ is the number of the individuals, and the variance in the denominator is approximated as one for this statistic. By substituting equation M7, and converting $N_{x=x_0}$ into N multiplied by the marker genotype frequency,

$$\begin{aligned}z_{x=2} &= \sqrt{\frac{2N}{3}}\sum_{i=1}^{l} d_i^3 p_{B_i},\\ z_{x=1} &= \left(\sqrt{\frac{N}{3}\cdot\frac{1-p_A}{p_A}}\sum_{i=1}^{l} d_i^3 p_{B_i}\right)-\left(\sqrt{\frac{N}{3}\cdot\frac{p_A}{1-p_A}}\sum_{j=1}^{m} e_j^3 p_{c_j}\right),\\ z_{x=0} &= -\sqrt{\frac{2N}{3}}\sum_{j=1}^{m} e_j^3 p_{c_j}.\end{aligned}$$

As the QT distributions at the two homozygote marker genotypes should be skewed to opposite directions under synthetic association, $z_{x=2}$ and $z_{x=0}$ would have opposite signs. We take their difference and obtain a χ^2 statistic with one degree of freedom,

$$\chi^2_{\text{skewness}} = \frac{(z_{x=2}-z_{x=0})^2}{2}, \tag{M8}$$

which we adopt as the test statistic for skewness. The test statistic is expected to become

$$\chi^2_{\text{skewness}} = \frac{N}{3}\left(\sum_{i=1}^{l} d_i^3 p_{B_i}+\sum_{j=1}^{m} e_j^3 p_{c_j}\right)^2, \tag{M9}$$

reflecting the contribution by causal alleles B_i and c_j.

Statistical tests of synthetic association and type I error rate

We combine two types of statistics, heteroscedasticity and skewness, to devise the combined test. Using Fisher's method, the *P*-values for testing heteroscedasticity and skewness, $p_{\text{heteroscedasticity}}$ and p_{skewness}, respectively, are combined as a χ^2 statistic (four degrees of freedom),

$$\chi^2_{\text{combined}} = -2\ln(p_{\text{heteroscedasticity}}\cdot p_{\text{skewness}}). \tag{M10}$$

The significance level was set to 0.05 for all tests.

Before testing, we applied rank-based inverse normal transformation (Blom 1958) to the whole QT distribution. The transformation avoids detecting spurious signals when the QT distribution is skewed as a whole. The transformed QT value q_i of the i-th individual is

$$q_i = \Phi^{-1}\left(\frac{r_i-c}{N-2c+1}\right),$$

where r_i is the rank of the individual, N is the total number of individuals, $c = 3/8$, and Φ^{-1} is the standard normal quantile. We strongly recommend applying the transformation, although it can

cause false positives when the marker association is extremely strong, as explained below.

Type I error rate of the tests were assessed from simulated and empirical data. Under the "null hypothesis" of indirect association, we inspected the distribution of nominal *P*-value, and assessed the test as accurate, conservative, or anticonservative, if the actual type I error rate was equal, smaller, or larger than the nominal *P*-value, respectively; an anticonservative test cannot be used. For a marker showing association at a borderline level of genome-wide significance ($R^2_{mrk} = 0.00592$), the heteroscedasticity test was accurate, but the skewness test tended to be conservative, due to the inverse normal transformation (Supplemental Fig. 4); this was not calibrated. As the tests for heteroscedasticity and skewness were not correlated, they could be combined using Fisher's method. When the marker association was as large as $R^2_{mrk} = 0.1$, which is exceptional for GWA signals, the inverse normal transformation caused spurious heteroscedasticity and skewness, thus the proposed tests were not valid. For gene expression data (Stranger et al. 2007), the heteroscedasticity test was accurate, and the skewness test was slightly conservative (Supplemental Fig. 5; Supplemental Table 2).

Models for simulation

If the strength of the marker SNP association R^2_{mrk}, and the frequency of variants (p_A, p_{B_i}, p_{c_j}) are specified, we can calculate the effect-size of the causal variants (d_i, e_j) by solving Supplemental equation S3, and determine the genetic model. We systematically explore four representative models of synthetic association by simulation. (Plots of d_i according to variant frequency are shown in Supplemental Fig. 6.)

Model 1

All causal alleles linked to marker allele *A* have identical effect-size, and there are no causal alleles linked to allele *a*. By solving Supplemental equation S3 under $d_1 = \cdots = d_l$ and $m = 0$,

$$d_i = \frac{1}{\sqrt{\frac{1-R^2_{mrk}}{R^2_{mrk}} \cdot \frac{2(1-p_A)}{p_A} \left(\sum_{i=1}^{l} p_{B_i}\right)^2 - 2\sum_{i=1}^{l} p_{B_i}}}. \quad \text{(M11)}$$

By substituting this to the test statistic of heteroscedasticity (equation M5), and solving for $\sum_{i=1}^{l} p_{B_i}$,

$$\sum_{i=1}^{l} p_{B_i} = \frac{R^2_{mrk}}{1-R^2_{mrk}} \frac{p_A}{2(1-p_A)} \left(\sqrt{\frac{N}{\chi^2_{\text{heteroscedasticity}}} \frac{1-p_A}{p_A}} + 2 \right).$$

Thus, when $R^2_{mrk} = 0.00592$ and $N = 5000$, heteroscedasticity is detectable at a power >0.8 under a significance level of 0.05 (which requires a noncentrality parameter of $\chi^2 > 9.64$ for the χ^2 distribution with two degrees of freedom; "+2" in the parenthesis is negligible) if

$$\sum_{i=1}^{l} p_{B_i} < 0.068 \sqrt{\frac{p_A}{1-p_A}}. \quad \text{(M12)}$$

Heteroscedasticity is detectable if the cumulative frequency of causal alleles (left term) is smaller than a certain function of the frequency of the marker allele *A* (right term); the detectable range with regard to $\sum_{i=1}^{l} p_{B_i}$ is wider when p_A is larger.

Model 2

Causal alleles are linked to the two marker alleles in a balanced way, such that the effect-size is uniform, as $d_1 = \cdots = d_l = e_1 = \cdots = e_m$, and the cumulative frequencies equal between causal alleles B_i (linked to marker allele *A*) and causal alleles c_j (linked to marker allele *a*), as $\sum_{i=1}^{l} p_{B_i} = \sum_{j=1}^{m} p_{c_j}$. By solving Supplemental equation S3 under the constraints,

$$d_i = e_j = \frac{1}{\sqrt{\frac{1-R^2_{mrk}}{R^2_{mrk}} \cdot \frac{2}{p_A(1-p_A)} \left(\sum_{i=1}^{l} p_{B_i}\right)^2 - 4\sum_{i=1}^{l} p_{B_i}}}. \quad \text{(M13)}$$

By substituting this (approximating the last term in the square-root as zero) to the test statistic of skewness (equation M9), and solving for $\sum_{i=1}^{l} p_{B_i}$,

$$\sum_{i=1}^{l} p_{B_i} = \left(\frac{R^2_{mrk}}{1-R^2_{mrk}}\right)^{\frac{3}{4}} \left(\frac{N}{6\chi^2_{\text{skewness}}}\right)^{\frac{1}{4}} (p_A(1-p_A))^{\frac{3}{4}}.$$

Thus, when $R^2_{mrk} = 0.00592$ and $N = 5000$, skewness is detectable at a power >0.8 under a significance level of 0.05 (which requires a noncentrality parameter of $\chi^2 > 7.85$ for the χ^2 distribution with one degree of freedom) if

$$\sum_{i=1}^{l} p_{B_i} < 0.069 \, (p_A(1-p_A))^{\frac{3}{4}}. \quad \text{(M14)}$$

Skewness is detectable if the cumulative frequency of causal alleles B_i (left term) is smaller than a certain function of the frequency of the marker allele *A* (right term); the detectable range with regard to $\sum_{i=1}^{l} p_{B_i}$ is widest when p_A is around 0.5.

Model 3

The effect-size of causal alleles is uniform, as $d_1 = \cdots = d_l = e_1 = \cdots = e_m$, yet the cumulative frequency of the causal alleles B_i is twice the cumulative frequency of causal alleles c_j, as $\sum_{i=1}^{l} p_{B_i} = 2\sum_{j=1}^{m} p_{c_j}$. Then,

$$d_i = e_j = \frac{1}{\sqrt{\frac{1-R^2_{mrk}}{R^2_{mrk}} \cdot \frac{(2-p_A)^2}{2p_A(1-p_A)} \left(\sum_{i=1}^{l} p_{B_i}\right)^2 - 3\sum_{i=1}^{l} p_{B_i}}}. \quad \text{(M15)}$$

Model 4

The cumulative frequencies are equal between causal alleles linked to the two marker alleles, as $\sum_{i=1}^{l} p_{B_i} = \sum_{j=1}^{m} p_{c_j}$, yet the effect-size of causal alleles B_i is twice the effect-size of causal alleles c_j, as $d_1 = \cdots = d_l = 2e_1 = \cdots = 2e_m$. Then,

$$d_i = 2e_j = \frac{1}{\sqrt{\frac{1-R^2_{mrk}}{R^2_{mrk}} \cdot \frac{(2-p_A)^2}{2p_A(1-p_A)} \left(\sum_{i=1}^{l} p_{B_i}\right)^2 - \frac{5}{2}\sum_{i=1}^{l} p_{B_i}}}. \quad \text{(M16)}$$

Power assessment by simulation

We assessed the power of the three tests under each of the four models by simulation. In any of the models, we assumed that effect-size equals among the causal alleles linked to the same marker allele (i.e., $d_1 = \cdots = d_l$ and $e_1 = \cdots = e_m$). In such a case, the tests remain the same if instead there was one composite allele of B_i's and another composite of c_j's. Using this property, we actually simulated the special case with one causal allele linked to each one of the marker alleles; the simulation results apply to the general case with multiple causal alleles.

Simulations were performed under the following parameter values; $R^2_{mrk} = 0.00592, 0.0118$; $p_A = 0.05, 0.10, \ldots, 0.95$; $\sum_{i=1}^{l} p_{B_i} = 0.005, 0.006, \ldots, 0.01, 0.02, \ldots, 0.05$. Other parameters—$\sum_{j=1}^{m} p_{c_j}$, d_i, and e_j—were determined according to constraints. We randomly generated 5000 (or 2500) individuals using simulation and applied the tests. The power was assessed from 1000 simulation trials. We used the R software for computation.

Acknowledgments

We thank the participants in the lipid study and Drs. Toru Nabika (Shimane University), Tomohiro Katsuya (Osaka University), and Yukio Yamori (Mukogawa Women's University). We also thank anonymous reviewers for their constructive comments. This work was supported by the Program for Promotion of Fundamental Studies in Health Sciences of the National Institute of Biomedical Innovation Organization; a Grant of the National Center for Global Health and Medicine; and the Ministry of Health, Labor and Welfare.

References

Bartlett MS. 1937. Properties of sufficiency and statistical tests. *Proc R Soc Lond A Math Phys Sci* **160:** 268–282.

Bennet AM, Di Angelantonio E, Ye Z, Wensley F, Dahlin A, Ahlbom A, Keavney B, Collins R, Wiman B, de Faire U, et al. 2007. Association of apolipoprotein E genotypes with lipid levels and coronary risk. *JAMA* **298:** 1300–1311.

Blom G. 1958. Statistical estimates and transformed β-variables. Wiley, New York.

Chasman DI, Pare G, Mora S, Hopewell JC, Peloso G, Clarke R, Cupples LA, Hamsten A, Kathiresan S, Malarstig A, et al. 2009. Forty-three loci associated with plasma lipoprotein size, concentration, and cholesterol content in genome-wide analysis. *PLoS Genet* **5:** e1000730. doi: 10.1371/journal.pgen.1000730.

Cirulli ET, Goldstein DB. 2010. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat Rev Genet* **11:** 415–425.

Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB. 2010. Rare variants create synthetic genome-wide associations. *PLoS Biol* **8:** e1000294. doi: 10.1371/journal.pbio.1000294.

Fellay J, Thompson AJ, Ge D, Gumbs CE, Urban TJ, Shianna KV, Little LD, Qiu P, Bertelsen AH, Watson M, et al. 2010. ITPA gene variants protect against anaemia in patients treated for chronic hepatitis C. *Nature* **464:** 405–408.

Ji W, Foo JN, O'Roak BJ, Zhao H, Larson MG, Simon DB, Newton-Cheh C, State MW, Levy D, Lifton RP. 2008. Rare independent mutations in renal salt handling genes contribute to blood pressure variation. *Nat Genet* **40:** 592–599.

Maher B. 2008. Personal genomes: The case of the missing heritability. *Nature* **456:** 18–21.

McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN. 2008. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet* **9:** 356–369.

Pare G, Cook NR, Ridker PM, Chasman DI. 2010. On the use of variance per genotype as a tool to identify quantitative trait interaction effects: a report from the Women's Genome Health Study. *PLoS Genet* **6:** e1000981. doi: 10.1371/journal.pgen.1000981.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81:** 559–575.

Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, de Grassi A, Lee C, et al. 2007. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315:** 848–853.

Stuart A, Ord J, Arnold S. 1999. *Kendall's advanced theory of statistics*. Arnold, London, United Kingdom.

Takeuchi F, McGinnis R, Bourgeois S, Barnes C, Eriksson N, Soranzo N, Whittaker P, Ranganath V, Kumanduri V, McLaren W, et al. 2009. A genome-wide association study confirms VKORC1, CYP2C9, and CYP4F2 as principal genetic determinants of warfarin dose. *PLoS Genet* **5:** e1000433. doi: 10.1371/journal.pgen.1000433.

Takeuchi F, Isono M, Katsuya T, Yamamoto K, Yokota M, Sugiyama T, Nabika T, Fujioka A, Ohnaka K, Asano H, et al. 2010. Blood pressure and hypertension are associated with 7 loci in the Japanese population. *Circulation* **121:** 2302–2309.

Wadelius M, Chen LY, Eriksson N, Bumpstead S, Ghori J, Wadelius C, Bentley D, McGinnis R, Deloukas P. 2007. Association of warfarin dose with genes involved in its action and metabolism. *Hum Genet* **121:** 23–34.

Weisgraber KH. 1994. Apolipoprotein E: structure-function relationships. *Adv Protein Chem* **45:** 249–302.

Weisgraber KH, Rall SC Jr, Mahley RW. 1981. Human E apoprotein heterogeneity. Cysteine-arginine interchanges in the amino acid sequence of the apo-E isoforms. *J Biol Chem* **256:** 9077–9083.

Wellcome Trust Case Control Consortium. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447:** 661–678.

Received September 24, 2010; accepted in revised form March 9, 2011.

Diabetologia (2011) 54:1350–1359
DOI 10.1007/s00125-011-2086-8

ARTICLE

Association of genetic variants for susceptibility to obesity with type 2 diabetes in Japanese individuals

F. Takeuchi · K. Yamamoto · T. Katsuya · T. Nabika · T. Sugiyama · A. Fujioka · M. Isono · K. Ohnaka · T. Fujisawa · E. Nakashima · H. Ikegami · J. Nakamura · Y. Yamori · S. Yamaguchi · S. Kobayashi · T. Ogihara · R. Takayanagi · N. Kato

Received: 5 September 2010 / Accepted: 25 January 2011 / Published online: 3 March 2011

Abstract
Aims/hypothesis In populations of East Asian descent, we performed a replication study of loci previously identified in populations of European descent as being associated with obesity measures such as BMI and type 2 diabetes.

Methods We genotyped 14 single nucleotide polymorphisms (SNPs) from 13 candidate loci that had previously been identified by genome-wide association meta-analyses for obesity measures in Europeans. Genotyping was done in 18,264 participants from two general Japanese populations. For SNPs showing an obesity association in Japanese individuals, we further examined diabetes associations in up to 6,781 cases and 7,307 controls from a subset of the original, as well as from additional populations.

Electronic supplementary material The online version of this article (doi:10.1007/s00125-011-2086-8) contains supplementary material, which is available to authorised users.

F. Takeuchi · M. Isono · N. Kato (✉)
Department of Gene Diagnostics and Therapeutics, Research Institute, National Center for Global Health and Medicine,
1-21-1 Toyama,
Shinjuku-ku, Tokyo 162-8655, Japan
e-mail: nokato@ri.ncgm.go.jp

K. Yamamoto
Department of Molecular Genetics,
Medical Institute of Bioregulation, Kyushu University,
Fukuoka, Japan

T. Katsuya
Department of Clinical Gene Therapy,
Osaka University Graduate School of Medicine,
Suita, Japan

T. Katsuya · T. Fujisawa · T. Ogihara
Department of Geriatric Medicine and Nephrology,
Osaka University Graduate School of Medicine,
Suita, Japan

T. Nabika
Department of Functional Pathology,
Shimane University School of Medicine,
Izumo, Japan

T. Sugiyama
Institute for Adult Diseases, Asahi Life Foundation,
Tokyo, Japan

A. Fujioka
Amagasaki Health Medical Foundation,
Amagasaki, Japan

K. Ohnaka
Department of Geriatric Medicine, Graduate School of Medical Sciences, Kyushu University,
Fukuoka, Japan

E. Nakashima · J. Nakamura
Division of Endocrinology and Diabetes, Department of Internal Medicine, Nagoya University Graduate School of Medicine,
Nagoya, Japan

E. Nakashima
Department of Metabolism and Endocrine Internal Medicine,
Chubu Rosai Hospital,
Nagoya, Japan

H. Ikegami
Department of Endocrinology, Metabolism and Diabetes,
Kinki University School of Medicine,
Osaka-Sayama, Japan

Y. Yamori
Mukogawa Women's University Institute for World Health Development,
Nishinomiya, Japan

Results Significant obesity associations ($p<0.1$ two-tailed, concordant direction with previous reports) were replicated for 11 SNPs from the following ten loci in Japanese participants: *SEC16B*, *TMEM18*, *GNPDA2*, *BDNF*, *MTCH2*, *BCDIN3D–FAIM2*, *SH2B1–ATP2A1*, *FTO*, *MC4R* and *KCTD15*. The strongest effect was observed at *TMEM18* rs4854344 ($p=7.1\times10^{-7}$ for BMI). Among the 11 SNPs showing significant obesity association, six were also associated with diabetes (OR 1.05–1.17; $p=0.04–2.4\times10^{-7}$) after adjustment for BMI in the Japanese. When meta-analysed with data from the previous reports, the BMI-adjusted diabetes association was found to be highly significant for the *FTO* locus in East Asians (OR 1.13; 95% CI 1.09–1.18; $p=7.8\times10^{-10}$) with substantial inter-ethnic heterogeneity ($p=0.003$).

Conclusions/interpretation We confirmed that ten candidate loci are associated with obesity measures in the general Japanese populations. Six (of ten) loci exert diabetogenic effects in the Japanese, although relatively modest in size, and independently of increased adiposity.

Keywords Asians · Association study · Ethnicity · Obesity · Type 2 diabetes

Abbreviations

CAGE	Cardiovascular Genome Epidemiology
GWA	Genome-wide association
SNP	Single nucleotide polymorphism

Introduction

Obesity is a major risk factor for type 2 diabetes, dyslipidaemia, hypertension and cardiovascular disease, and has a strong genetic component [1]. Twin studies have generally found heritability estimates of 0.75 to 0.85 for BMI and approximately 0.70 for weight [2]. Genome-wide association (GWA) studies have provided evidence that several loci are associated with common obesity mostly in populations of European descent [3–15]. The first such loci reported was the fat-mass and obesity-associated gene (*FTO*) [3, 16, 17]. A common variant in the *FTO* locus, rs9939609, was originally identified as part of a GWA study for type 2 diabetes [3]; people with homozygous risk alleles weighed approximately 3 kg more than those without a risk allele. It has been suggested that this variant can predispose individuals to type 2 diabetes and metabolic disorders through its primary effect on BMI in Europeans [18].

Ethnic differences have been assumed to exist between Europeans and Asians in terms of the components and impacts of genetic factors for obesity, and traits or disorders related to obesity (e.g. type 2 diabetes) [2, 19, 20]. For instance, the relationship between BMI and body fat per cent differs between these populations, with Asians in general having a higher body fat per cent at a lower BMI than Europeans [20]. It has also been estimated that the absolute genetic variances for BMI and weight are greater in Europeans than in East Asians, according to an adolescent twin study [2]. From an epidemiological viewpoint, moreover, it has been hypothesised that the overall impact of obesity on type 2 diabetes is greater in Asians than in Europeans [21]. Accordingly, it is of interest to compare the genetic associations between populations of European descent and Japanese populations.

To date, several studies on non-European populations have replicated the associations for a number of novel obesity loci previously identified by GWA studies in Europeans [13, 22–25]; nevertheless, statistical power has not been sufficient to make strong conclusions. Therefore to test associations between obesity measures (BMI and weight) and type 2 diabetes for 14 single nucleotide polymorphisms (SNPs) from 13 candidate loci recently reported by two GWA meta-analyses [10, 11], we performed a replication study in the general Japanese populations.

Methods

Study populations

We performed a replication study of previously identified variants in the general Japanese populations (Table 1, electronic supplementary material [ESM] Study samples for BMI association analysis). Specifically, 5,695 Japanese participants (referred to hereafter as the Amagasaki panel) were consecutively enrolled in the population-based setting as described previously [26] and 12,569 other Japanese

S. Yamaguchi
Department of Internal Medicine III,
Shimane University School of Medicine,
Izumo, Japan

S. Kobayashi
Shimane University Hospital,
Izumo, Japan

T. Ogihara
Osaka General Medical Center,
Osaka, Japan

R. Takayanagi
Department of Medicine and Bioregulatory Science, Graduate School of Medical Sciences, Kyushu University,
Fukuoka, Japan

Table 1 Clinical characteristics of study participants for BMI

Variables	Amagasaki panel	Fukuoka panel
Both sexes (*n*)	5,695	12,569
Women (*n*)	2,290	6,898
Men (*n*)	3,405	5,671
Age (years)	48.8±12.6	62.6±6.8
BMI (kg/m^2)	23.0±3.2	23.1±3.0
Body weight (kg)	61.8±11.5	58.4±10.2
Alcohol drinking (%)		
None	24.1	48.6
Previous drinker	1.2	5.0
Chance drinker [a]	35.6	–
Current drinker	39.1	46.4
Smoking (%)		
None	55.2	59.9
Previous smoker	9.9	23.1
Current smoker	34.8	17.0
Blood chemistry		
Fasting plasma glucose (mmol/l)	5.22±0.54	N/A
HbA_{1c} (%)[b]	5.41±0.84	5.23±0.77
LDL-cholesterol (mmol/l)[c]	3.21±0.81	N/A
Triacylglycerol (mmol/l)	1.24±0.97	1.66±1.12
HDL-cholesterol (mmol/l)	1.63±0.46	1.62±0.44
Blood pressure (mmHg)		
Systolic blood pressure	124.3±18.1	138.8±21.2
Diastolic blood pressure	75.9±11.6	83.9±11.7
Prevalence of metabolic diseases (%)[d]		
Hypertension	23.4	56.9
Diabetes	6.1	7.6
Dyslipidaemia	42.2	N/A

Values are means±SD unless otherwise indicated

All clinical assessments were performed using uniform standards in each population; blood samples were taken after ≥6 h fast in the Amagasaki panel and without setting strict fasting condition in the Fukuoka panel

N/A, not applicable

[a] Since the questionnaire did not differentiate between chance drinker and current drinker, the corresponding participants were combined in the current drinker category in the Fukuoka panel

[b] HbA_{1c} was measured in 1,288 participants in the Amagasaki panel and in all participants in the Fukuoka panel

[c] LDL-cholesterol was calculated in the Amagasaki panel using the Friedewald formula, with missing values assigned to individuals with triacylglycerol >4.52 mmol/l. Since blood samples were taken without setting strict fasting condition, the values for LDL-cholesterol and prevalence of dyslipidaemia are not shown for the Fukuoka panel, in accordance with the Japan Atherosclerosis Society Guidelines [49]

[d] Hypertension was defined as systolic blood pressure ≥140 mmHg and/or diastolic blood pressure ≥90 mmHg, or taking antihypertensive medication. Diabetes was defined as fasting plasma glucose ≥7.0 mmol/l and/or HbA_{1c}≥6.5%, or taking blood glucose-lowering medication. Dyslipidaemia was defined according to the Japan Atherosclerosis Society Guidelines [49]. The criteria for diabetes are not identical to those adopted for selecting diabetic cases in the Fukuoka panel (see details in ESM Study samples for type 2 diabetes case–control studies)

participants (referred to hereafter as the Fukuoka panel) were randomly selected from residents aged 50 to 74 years in the general population [27]. Among the candidate loci tested in the two panels, those showing a replication of association ($p<0.05$ one-tailed, i.e. $p<0.1$ two-tailed, in concordant direction with previous reports) were subjected to tests to examine type 2 diabetes associations in a Japanese case–control study panel (ESM Table 1, ESM Study samples for type 2 diabetes case–control studies). For the purpose of uniformity, two tailed p values are shown throughout the text, unless otherwise indicated. Of 2,041 cases and 2,418 controls enrolled from the Cardiovascular Genome Epidemiology (CAGE) Network [28], 931 cases and 1,404 controls were included in stage 1 and 1,110 cases and 1,014 controls in stage 2. In addition, participants in the Fukuoka panel were included in stage 1 and Biobank Japan (http://biobankjp.org/ [in Japanese], accessed 1 February 2011) cases were included in stage 2 (ESM Table 1). From the CAGE Network, type 2 diabetes cases were enrolled according to the 1999 WHO criteria, while unaffected controls were enrolled according to the following criteria: (1) no past history of urinary glucose or glucose intolerance; (2) $HbA_{1c}<5.6\%$ or a normal result from 75 g glucose tolerance test; and (3) age at examination ≥55 years. From the population-based participants in the Fukuoka panel, diabetic participants ($n=740$) and unaffected controls ($n=4,889$) were selected for case–control analysis; among these samples, there was some overlap between the BMI/weight and type 2 diabetes studies. Here, diabetes was defined as $HbA_{1c}\geq7.0$ or under treatment for type 2 diabetes; a relatively stringent criterion for HbA_{1c} (≥7.0%) was adopted because of the lack of fasting plasma glucose data for participants in the Fukuoka panel. The controls were chosen as non-diabetic participants who met the following conditions: age ≥55 years; $HbA_{1c}\leq5.0\%$; no previous and/or current treatment for diabetes; and absence of renal failure (serum creatinine<265.2 μmol/l), as previously described [27]. In total, 6,781 cases and 7,307 controls were used for the type 2 diabetes case–control study in Japanese. The CAGE Network samples were categorised, according to the stages of a previous GWA study [28], into two panels: CAGE–GWAS (stage 1); and CAGE–replication (stage 2) (ESM Table 1); genotyping in the CAGE-GWAS panel was performed with Infinium assay (Illumina, San Diego, CA). All participants enrolled in these different studies provided written informed consent. Local Ethics Committees approved the protocols used.

Height and body weight were measured by trained personnel using standard anthropometric techniques for all participants other than Biobank Japan type 2 diabetes participants, for whom the relevant data were self-reported from questionnaire.

SNP genotyping and quality control

Samples (except for the CAGE–GWAS panel) were genotyped using the TaqMan assay (Applied Biosystems by Life Technologies, Carlsbad, CA, USA) for 14 SNPs from 13 unique obesity loci previously identified in populations of European descent [10, 11]. These SNPs included: rs2815752 (*NEGR1*), rs10913469 (*SEC16B*), rs4854344 (*TMEM18*), rs7647305 (*ETV5*), rs10938397 (*GNPDA2*), rs2844479 (*NCR3–AIF1*), rs6265 (*BDNF*), rs10838738 (*MTCH2*), rs7138803 (*BCDIN3D–FAIM2*), rs4788102 (*SH2B1–ATP2A1*), rs6499640 (*FTO*), rs9939609 (*FTO*), rs12970134 (*MC4R*) and rs29941 (*KCTD15*). The genotype distribution of all tested SNPs was in Hardy–Weinberg equilibrium ($p > 10^{-4}$). We obtained successful genotyping call rates of >99.6% for all SNPs and >99.8% for all included samples (across 14 SNPs).

Statistical analysis

SNP association analysis BMI and weight were inverse-normal transformed separately by sex in each panel before association analysis. In addition, in the Fukuoka study panel, we examined the WHR as a variable of fat distribution, which was also inverse-normal transformed. We tested SNPs for the trait association using linear regression analysis in an additive genotype model after adjustment for age classes separately by sex. Age classes were defined according to age distribution in the individual panels, and included ≤40, 41–50, 51–60 and >60 years for the Amagasaki panel, and ≤55, 56–60, 61–65, 66–70 and >70 years for the Fukuoka panel. A one tailed value of $p<0.05$ ($p<0.1$ two-tailed) was considered statistically significant. We combined association results for the two Japanese panels by using the inverse variance method. We used PLINK (version 1.06; http://pngu.mgh.harvard.edu/~purcell/plink/) [29], *R* software (version 2.8.1; www.r-project.org) and rmeta (version 2.16; http://cran.r-project.org) for association test and meta-analysis (websites accessed 6 February 2011).

Assessment of genetic effect of obesity variants To assess the proportion of variance for BMI that was explained by each SNP, we calculated a coefficient of determination R^2 as: $2f(1-f)\beta^2$, where f is the minor allele frequency and β is the per-allele effect on the standardised values of BMI. We measured the cumulative effect of multiple SNPs by summing the R^2 values for individual SNPs.

Test of ethnic diversity and sex specificity We compared the per-allele effect size of each SNP on inverse-normal transformed BMI between the ethnic groups (Japanese vs Europeans) and between sexes. Taking into account the well-known male–female differences in body composition (e.g. fat distribution, deposition and accumulation) [30], we tested the potential sex specificity in the genetic associations with obesity. We examined the heterogeneity of the effect size with Cochran's *Q*-test [31].

Association analysis of type 2 diabetes While we tested the primary association with obesity measures, we tested type 2 diabetes association as a secondary analysis in the present study. Thus, we performed two-staged analysis for diabetes case–control study. That is, all SNPs were tested for association with type 2 diabetes in stage 1 samples and only SNPs with $p<0.1$ for BMI or weight were then tested in stage 2 samples (ESM Table 1). Using logistic regression analysis, we tested association of candidate SNPs with type 2 diabetes, with and without adjustment for BMI. We adjusted the diabetes trait for sex and BMI, but not for age because age distribution differed between case and control groups; cases were younger than controls, who were defined as being ≥55 years of age.

Results

BMI and weight association at reported SNPs

In this study, we tested the association of BMI and weight with 14 SNP loci that had attained genome-wide significance levels ($p<5\times10^{-8}$) in previously performed GWA meta-analyses for obesity in Europeans [10, 11]. In the combined sample (n=18,264), we found that 11 of the 14 SNPs had nominally significant ($p<0.1$ two-tailed) associations with BMI and/or weight; these were rs10913469 (*SEC16B*), rs4854344 (*TMEM18*), rs10938397 (*GNPDA2*), rs6265 (*BDNF*), rs10838738 (*MTCH2*), rs7138803 (*BCDIN3D–FAIM2*), rs4788102 (*SH2B1–ATP2A1*), rs6499640 (*FTO*), rs9939609 (*FTO*), rs12970134 (*MC4R*) and rs29941 (*KCTD15*) (Table 2). Furthermore, significant associations between BMI and two SNPs at *FTO* were determined to be independent (i.e. $p=0.016$ for rs6499640 and $p=3.9\times10^{-6}$ for rs9939609, both two-tailed, when these were simultaneously included in the linear regression model). Given the much stronger association for rs9939609 than rs6499640 (Table 2), rs9939609 was considered to be the key SNP for *FTO*, although rs6499640 still appears to have some role. The strength of association did not appear to significantly differ between the study panels (Amagasaki vs Fukuoka), except at *BDNF* ($p<0.01$ for heterogeneity; ESM Table 2). The presence of sexual dimorphism was indicated for two loci, *NEGR1* and *GNPDA2* ($p<0.05$ for heterogeneity; ESM Table 3).

Table 2 Association of previously reported SNPs with BMI and type 2 diabetes in Japanese individuals

SNP details				Allele testing		Obesity measures				Type 2 diabetes (by BMI-adjusted status)				Cases/controls (*n*/*n*)
						BMI (*n*=18,264)		Weight (*n*=18,264)		Not adjusted		Adjusted		
Chr[a]	Neighbouring gene(s)	SNP	Position (B36)	Allele	Frequency	Beta (%)	*p* value	Beta (%)	*p* value	OR	*p* value	OR	*p* value	
1	*NEGR1*	rs2815752	72,585,028	T	0.92	0.53	0.785	0.25	0.895	0.86	0.028	0.88	0.097	1,671/6,293
1	*SEC16B*	rs10913469	176,180,142	C	0.23	3.00	0.014[b]	2.20	0.069 [b]	1.04	0.151	1.02	0.576	6,781/7,307
2	*TMEM18*	rs4854344	628,144	A	0.90	8.38	7.1×10^{-7}[b]	7.58	6.3×10^{-6}[b]	1.18	3.2×10^{-5}[b]	1.16	2.9×10^{-4}[b]	6,781/7,307
3	*ETV5*	rs7647305	187,316,984	C	0.96	0.87	0.755	0.88	0.753	1.01	0.931	1.04	0.739	1,671/6,293
4	*GNPDA2*	rs10938397	44,877,284	G	0.30	4.36	1.3×10^{-4}[b]	3.64	1.3×10^{-3}[b]	1.07	0.007[b]	1.07	0.017[b]	6,781/7,307
6	*NCR3-AIF1*	rs2844479	31,680,935	A	0.51	0.59	0.569	0.82	0.428	0.96	0.280	0.98	0.544	1,671/6,293
11	*BDNF*	rs6265	27,636,492	G	0.59	5.41	3.1×10^{-7}[b]	4.21	6.3×10^{-5}[b]	1.09	0.001[b]	1.07	0.010[b]	6,781/7,307
11	*MTCH2*	rs10838738	47,619,625	G	0.32	3.77	0.001[b]	1.95	0.078 [b]	0.96	0.076	0.94	0.016	6,781/7,307
12	*BCDIN3D-FAIM2*	rs7138803	48,533,735	A	0.34	1.63	0.137	1.92	0.078 [b]	1.06	0.020[b]	1.05	0.041[b]	6,781/7,307
16	*SH2B1-ATP2A1*	rs4788102	28,780,899	A	0.14	3.52	0.018[b]	3.12	0.035 [b]	1.04	0.276	1.03	0.378	6,781/7,307
16	*FTO*	rs6499640	52,327,178	A	0.15	3.47	0.018[b]	1.96	0.180	1.02	0.659	1.00	0.936	6,781/7,307
16	*FTO*	rs9939609	52,378,028	A	0.19	6.05	4.6×10^{-6}[b]	4.81	2.5×10^{-4}[b]	1.20	4.3×10^{-10}[b]	1.17	2.4×10^{-7}[b]	6,781/7,307
18	*MC4R*	rs12970134	56,035,730	A	0.16	2.12	0.135	4.30	0.002[b]	1.11	0.002[b]	1.11	0.002[b]	6,781/7,307
19	*KCTD15*	rs29941	39,001,372	C	0.21	0.54	0.671	2.42	0.055[b]	1.01	0.710	1.01	0.754	6,781/7,307

Results for Japanese participants from the Amagasaki (*n*=5,695) and Fukuoka (*n*=12,569) panels were combined by meta-analysis; we tested the allele reported to increase BMI in Europeans

Effect sizes are indicated as beta per SD unit of trait (for BMI and weight) and OR (for type 2 diabetes)

The association with type 2 diabetes was tested after adjustment for BMI and sex (see 'Methods' section)

Cohort-wise results are shown in ESM

[a] Chromosome

[b] Suggestive association ($p<0.1$, two-tailed)

Ethnic heterogeneity in effect sizes

To test for the potential presence of ethnic heterogeneity, we compared the effect sizes (β for BMI) between the Japanese and European populations (Fig. 1a). The sample size was smaller in the Japanese group, and consequently the 95% CI was wider in this group, although the direction of association was concordant between the ethnic groups for all SNPs. However, we did find significant ($p<0.05$) inter-ethnic heterogeneity in per-allele effect at rs2844479 (*NCR3–AIF1*) and rs29941 (*KCTD15*) for standardised BMI, and at rs9939609 (*FTO*) for non-standardised BMI (in kg/m^2; ESM Table 4). We calculated R^2 as the proportion of phenotypic variance explained by a SNP (see Methods). Based on these R^2 measurements, the association of three replicated loci, rs4854344 (*TMEM18*), rs6265 (*BDNF*) and rs10838738 (*MTCH2*), was stronger in the Japanese population than in the population of European descent. We detected no significant gene×gene interactions by regression analysis for BMI associations with 14 SNPs (data not shown); the genotypes of each SNP were concomitantly included assuming an additive effect when combined. For the 14 SNPs, the cumulative R^2 totalled 0.65% and 1.2% in the Japanese and Europeans, respectively. Thus, the explained variance tended to be smaller in Japanese than in Europeans.

Association of obesity susceptibility SNPs with type 2 diabetes

For the case–control study of type 2 diabetes, we genotyped 11 obesity-associated SNPs in the stage 1 and stage 2 panels; the remaining three SNPs were genotyped only in the stage 1 panel (ESM Table 1). Some evidence of association with type 2 diabetes in a direction consistent with the BMI–SNP associations ($p=0.02–4.3\times10^{-10}$ and OR 1.06–1.20 before adjustment for BMI; $p=0.04–2.4\times10^{-7}$ and OR 1.05–1.17 after adjustment for BMI) was provided for six (of 11) SNPs in the Japanese population (Fig. 1b, Table 2). After adjusting for BMI, most of the observed associations between obesity-associated SNPs and diabetes were slightly attenuated, although nominal significance remained (Table 2, ESM Table 5). Considering the possibility of some selection bias in the case–control study design, we also tested the diabetes association by nested case–control comparison within the same population (i.e. Fukuoka panel) and verified a fair consistency in the OR for type 2 diabetes (ESM Table 6). Further, to examine the possibility that the BMI-independent diabetes associations derive from genetic susceptibility to fat distribution, we tested the association with WHR among the control participants in the Fukuoka panel (n=4,889), where none of the six SNPs showed significant association with WHR

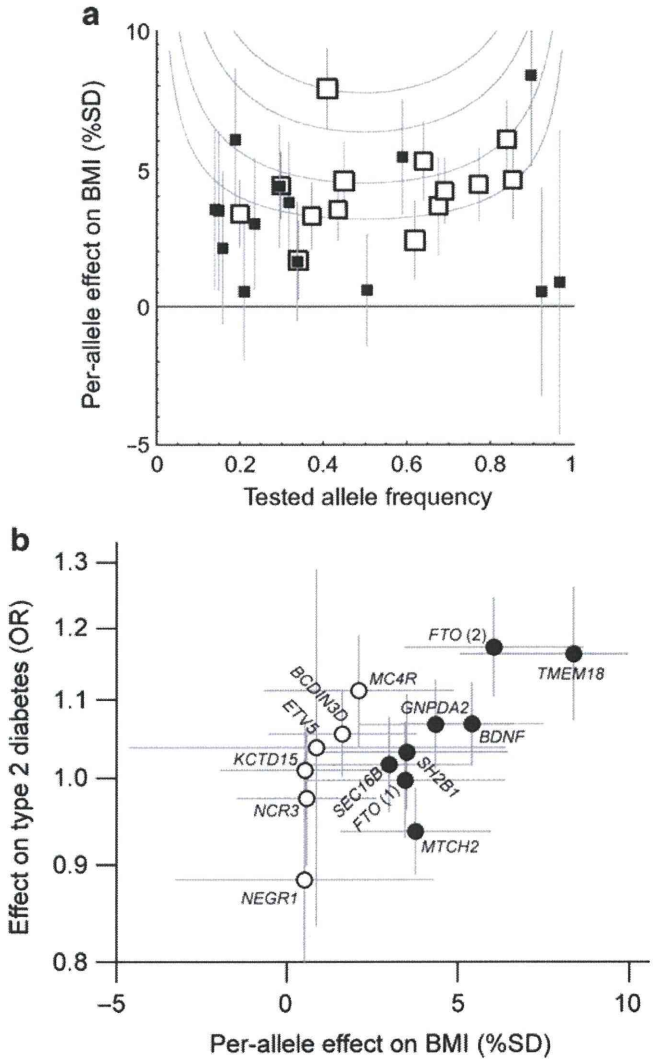

Fig. 1 Effect size for BMI and type 2 diabetes at SNPs previously reported to be associated with BMI in Europeans. **a** Cross-population comparison of per-allele effect of 14 BMI-associated SNPs between the Japanese and European populations. The per-allele effects (β in % SD) of each variant on BMI are shown by squares (proportional in size to tested sample size) and vertical lines (representing 95% CI) for the Japanese (black squares) and Europeans (white squares). Curves (in grey) indicate R^2=0.003, 0.002, 0.001 and 0.0005 (from top to bottom). **b** Comparison of genetic impacts on BMI (β in *x*-axis) and type 2 diabetes adjusted for BMI (OR in *y*-axis) for the 14 SNPs. Those showing significant (black circles) and non-significant (white circles) association with BMI in Japanese are depicted. SNP rs numbers of individual loci are as follows: rs2815752 for *NEGR1*, rs10913469 for *SEC16B*, rs4854344 for *TMEM18*, rs7647305 for *ETV5*, rs10938397 for *GNPDA2*, rs2844479 for *NCR3*, rs6265 for *BDNF*, rs10838738 for *MTCH2*, rs7138803 for *BCDIN3D*, rs4788102 for *SH2B1*, rs6499640 for *FTO*(1), rs9939609 for *FTO*(2), rs12970134 for *MC4R* and rs29941 for *KCTD15*

in a direction consistent with the SNP–diabetes associations (ESM Table 7).

Assuming the potential presence of 'unadjusted' confounding influences of BMI, we plotted (Fig. 2) the data from the genetic studies examining type 2 diabetes

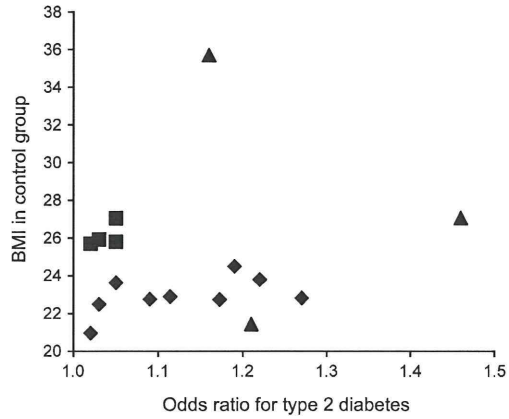

Fig. 2 Relationship between BMI in the control group (or derived population) and the OR for type 2 diabetes in genetic studies examining type 2 diabetes association at the *FTO* locus (by using rs9939609 or its proxies). The studies are categorised into three ethnic groups: East Asians (diamonds), Europeans (squares) and others (triangles). For further details, see ESM Table 8

association at the *FTO* locus [3, 11, 24, 32–42], our aim being to evaluate the relationship between BMI in the control group (or derived population) and the OR for type 2 diabetes. We detected no apparent relationship between the tested variables (p=0.15 for East Asians and p=0.41 for Europeans; Fig. 2, ESM Table 8). While there was a significant cross-population difference in the OR for type 2 diabetes (e.g. OR 1.13 and 1.04 for East Asians and Europeans, respectively; I^2=89%; p=0.003), the BMI-adjusted association with type 2 diabetes proved to be highly significant (OR 1.09, $p=3.0\times10^{-11}$) in the meta-analysis involving 36,064 cases and 61,234 controls of various ethnic origins.

Discussion

The present study investigated genetic susceptibility to common forms of obesity and its relevance to type 2 diabetes in Japanese populations. Replicating a study of candidate loci previously identified by GWA meta-analyses of Europeans [10, 11], we found some ethnic diversity in obesity variants (Fig. 1a). The top hit associations were consistently reported at *FTO* and *MC4R* in populations of European descent [5, 6, 10–12, 14, 15], whereas equivalent or more pronounced associations with obesity localised to *TMEM18* and *BDNF*, together with modest association at *SEC16B*, *GNPDA2*, *MTCH2*, *BCDIN3D–FAIM2*, *SH2B1–ATP2A1* and *KCTD15*, were found in the Japanese populations (Table 2, ESM Table 4). In addition, we found that the direction of BMI association was concordant between the ethnic groups at all tested loci, although the cumulative effect of the associated SNPs was smaller in Japanese than in Europeans. Of particular note is the fact that, in the present study, we highlighted the genetic impact of six loci on type 2 diabetes, this impact being independent of obesity susceptibility in all six loci.

Recently, two GWA meta-analyses were performed for examination of obesity in populations of European descent; the studies involved >32,000 and >25,000 individuals, with follow-up analysis using genotypes from large cohorts (>50,000 samples in total) and computer-generated association results [10, 11]. Collectively, these studies revealed 13 unique obesity-associated loci in Europeans. Four studies in total, including our present study, have attempted to replicate the obesity association in East Asians; of these, two studies involved Japanese individuals and two involved Chinese individuals [23–25]. While the design of these studies was not identical (two case–control studies, one quantitative trait analysis and one study with both-types of analytical approaches combined), all four studies consistently reported significant association with obesity at the *GNPDA2* (rs10938397) and *FTO* (rs9939609 or its proxies) loci in East Asians. The present study, which is the largest of the four studies and the only one that involved quantitative trait analysis using the general population sample, has enabled us to validate nine other SNP loci. Eight of these SNP loci had previously shown obesity association in one or two studies [23–25], whereas replication at *MTCH2* (rs10838738) has not been previously reported in East Asians.

Compared with the data for Europeans [10, 11], the variance for BMI explained by individual SNP loci appeared to be modest as a whole in the Japanese population, except for *TMEM18* rs4854344 (R^2=0.13%), *BDNF* rs6265 (R^2=0.1%) and *MTCH2* rs10838738 (R^2= 0.06%; Fig. 1b, ESM Table 4). It should be kept in mind that the larger cumulative effect on BMI in Europeans than in Japanese is partly due to the 'winner's curse effect' [43], where the original meta-analysis tends to overestimate the true population effect, in addition to ethnic difference in at-risk alleles. We found that the effect size (β) was larger at *TMEM18* rs4854344 than at *FTO* rs9939609 in the independent panels of Japanese samples (ESM Table 2). Nevertheless, there seems to be an overall consistency of genetic variants for susceptibility to obesity between the examined ethnic groups.

A remaining issue of interest is whether the correlations between obesity-associated SNPs and type 2 diabetes can be explained by the SNP–trait and trait–type 2 diabetes associations [18]. A simple way to assess the causal direction of associations (i.e. whether increased adiposity resulting from the variant under investigation is causally related to type 2 diabetes) is to test the inconsistency of the SNP–diabetes association both with and without adjustment for BMI. In this regard, our results, although not conclusive at some loci, indicated

that six obesity variants could lead to altered fat mass and altered susceptibility to type 2 diabetes, at least in part, through separate mechanisms, based on the statistical evidence for independence following analysis of either the whole sample or separately by sex (Table 2, ESM Table 5). Interestingly, a previous study in Europeans [18] demonstrated that the associations between *FTO* variants and obesity-related metabolic traits (including type 2 diabetes) are likely to be entirely mediated by BMI; this study also emphasised that the use of appropriately powered studies was important in making such assessments. Our meta-analysis for the *FTO* locus, involving 36,064 cases and 61,234 controls, has shown that the BMI-adjusted diabetes association is highly significant (OR 1.09, $p=3.0\times10^{-11}$). Based on analysis of stratification of Japanese samples by extent of obesity (underweight, normal weight, overweight and obese), sex (men, women) and age (<60 and ≥60 years), we determined that the strength of type 2 diabetes association tended to be more pronounced among underweight people aged <60 years in both sexes, although it did not reach statistical significance due to relatively small sample size ($n<100$ for cases and controls) in the corresponding (age- and obesity-stratified) group ($p=0.18$ for a group of underweight people aged <60 years vs the others; Fig. 3, ESM Fig. 1). Previous epidemiological studies have reported that being underweight is likely to be associated with the risk of type 2 diabetes in the Japanese population [44, 45] and that Japanese people have low insulin secretory capacity and a high risk of type 2 diabetes at BMIs lower than the existing WHO cut-off point for being overweight, 25 kg/m^2 [46, 47]. Thus, together with other, unidentified confounding factors, the extent of obesity (>2/3 of the Japanese participants were underweight or normal weight) may account for cross- and within-population differences in type 2 diabetes susceptibility at obesity loci such as *FTO*. Despite limited data availability at present on the topic, the direction of BMI-adjusted diabetes association seems to be concordant in some five studies [3, 4, 10, 24, 42] that have investigated the 14 obesity variants (ESM Table 9).

Sex specificity is another issue of interest in the genetics of obesity. Some epidemiological studies have suggested that different genes influence variation of BMI in men and women [30]. However, little evidence has been provided to support this notion. In the present study, we found nominally significant sex-related differences in the effect sizes for two obesity variant loci, *NEGR1* and *GNPDA2* (ESM Table 3). Although heterogeneity was not statistically significant ($p=0.09$), obesity association only in women appeared to be replicated for *MC4R*. Notably, a previous study has reported the equivalent type of sexual dimorphism for the association between a *MC4R* variant and obesity measures in the Swedish population [42]. Despite no prior evidence of sex specificity at this locus in Europeans [4, 48], these findings warrant further investigation with reference to ethnic homogeneity and/or the specificity of target populations.

We acknowledge several limitations inherent in the present study. Specifically, the sample size used to screen obesity association was smaller for the Japanese group ($n=18,264$) than that used for the European GWA meta-analyses ($n>25,000$ in the discovery stage). Therefore, we assessed statistical power for each of the tested loci, assuming effect sizes equivalent to those reported in the original GWA studies for the allele frequencies in the Japanese (ESM Table 10). Apart from the replicated

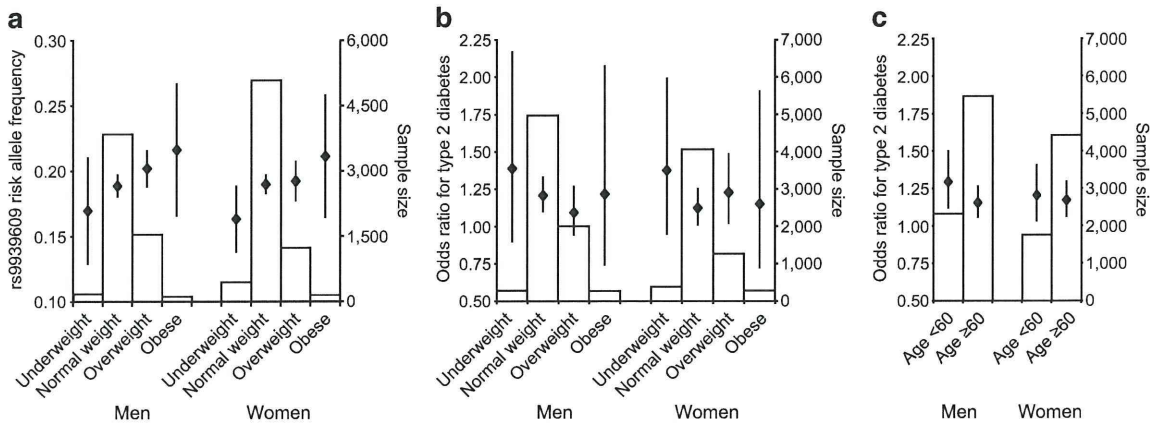

Fig. 3 Stratification of Japanese sample by extent of obesity, sex (men and women), and age (<60 and ≥60 years) in relation with genetic impacts at *FTO* rs9939609. Obesity was graded as: underweight (BMI <18.5 kg/m^2), normal weight (18.50–24.99 kg/m^2), overweight (25.00–29.99 kg/m^2) and obese (≥30 kg/m^2). The samples were analysed separately by sex. **a** Risk allele (A in the plus strand) frequencies of rs9939609, compared among four subgroups of Fukuoka panel ($n=12,569$) stratified by extent of obesity. **b** ORs for type 2 diabetes, compared among four subgroups of case–control sample (6,781 cases, 7,307 controls) stratified by extent of obesity. **c** ORs for type 2 diabetes, compared between two subgroups of case–control sample (6,781 cases, 7,307 controls) stratified by age. Whiskers indicate 95% CIs. For detailed stratification, see ESM Fig. 1

associations, sufficient power (i.e. >0.8) was not attainable for two (of 14) SNP loci in the Japanese samples. In addition to the issue of power, differences in linkage disequilibrium patterns between the ethnic groups could also have contributed to the lack of replicated association at a given locus. We found some cross-population differences in linkage disequilibrium relations at several SNP loci, which could also attenuate (or strengthen) the association signals accordingly. Regional examination of SNP–obesity association, which is ongoing in Asians, but not point-wise studies, will resolve this issue.

In summary, we confirmed that 11 SNPs from ten candidate loci were significantly associated with BMI in Japanese individuals, as was previously reported in Europeans. We also found some cross-population differences in the effect sizes of individual obesity variants. These studies, moreover, highlight the genetic influences on type 2 diabetes that appear to be independent of BMI, as well as the potential presence of sex specificity in the genetics of common obesity.

Acknowledgements This work was supported by the Program for Promotion of Fundamental Studies in Health Sciences of the National Institute of Biomedical Innovation Organization (NIBIO), the Manpei Suzuki Diabetes Foundation, a grant from National Center for Global Health and Medicine (NCGM) and by the Ministry of Education, Cultures, Sports, Science and Technology, all of Japan. We acknowledge the outstanding contributions of the employees of NCGM, who provided technical and infrastructural support for this work. Above all, we thank the patients and study participants who made this work possible and who gave it value. We thank all the people who continuously support the Hospital-based Cohort Study at NCGM, the Amagasaki Study and the Kyushu University Fukuoka Cohort Study in Japan. We also thank S. Kono, M. Ogasawara, C. Makibayashi and the many physicians of the Amagasaki Medical Association for their contribution in collecting DNA samples and accompanying clinical information. Part of the DNA samples from type 2 diabetes cases used for this research were provided from the Leading Project for Personalized Medicine at the Ministry of Education, Culture, Sports, Science and Technology, Japan.

Duality of interest The authors declare that there is no duality of interest associated with this manuscript.

References

1. Haslam DW, James WP (2005) Obesity. Lancet 366:1197–1209
2. Hur YM, Kaprio J, Iacono WG et al (2008) Genetic influences on the difference in variability of height, weight and body mass index between Caucasian and East Asian adolescent twins. Int J Obes (Lond) 32:1455–1467
3. Frayling TM, Timpson NJ, Weedon MN et al (2007) A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. Science 316:889–894
4. Loos RJ, Lindgren CM, Li S et al (2008) Common variants near MC4R are associated with fat mass, weight and risk of obesity. Nat Genet 40:768–775
5. Scuteri A, Sanna S, Chen WM et al (2007) Genome-wide association scan shows genetic variants in the FTO gene are associated with obesity-related traits. PLoS Genet 3:e115
6. Hinney A, Nguyen TT, Scherag A et al (2007) Genome wide association (GWA) study for early onset extreme obesity supports the role of fat mass and obesity associated gene (FTO) variants. PLoS One 2:e1361
7. Fox CS, Heard-Costa N, Cupples LA, Dupuis J, Vasan RS, Atwood LD (2007) Genome-wide association to body mass index and waist circumference: the Framingham Heart Study 100K project. BMC Med Genet 8(Suppl 1):S18
8. Yanagiya T, Tanabe A, Iida A et al (2007) Association of single-nucleotide polymorphisms in MTMR9 gene with obesity. Hum Mol Genet 16:3017–3026
9. Liu YJ, Liu XG, Wang L et al (2008) Genome-wide association scans identified CTNNBL1 as a novel gene for obesity. Hum Mol Genet 17:1803–1813
10. Willer CJ, Speliotes EK, Loos RJ et al (2009) Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. Nat Genet 41:25–34
11. Thorleifsson G, Walters GB, Gudbjartsson DF et al (2009) Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity. Nat Genet 41:18–24
12. Meyre D, Delplanque J, Chevre JC et al (2009) Genome-wide association study for early-onset and morbid adult obesity identifies three new risk loci in European populations. Nat Genet 41:157–159
13. Chambers JC, Elliott P, Zabaneh D et al (2008) Common genetic variation near MC4R is associated with waist circumference and insulin resistance. Nat Genet 40:716–718
14. Sabatti C, Service SK, Hartikainen AL et al (2009) Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. Nat Genet 41:35–46
15. Scherag A, Dina C, Hinney A et al (2010) Two new loci for body-weight regulation identified in a joint analysis of genome-wide association studies for early-onset extreme obesity in French and German study groups. PLoS Genet 6:e1000916
16. Loos RJ, Bouchard C (2008) FTO: the first gene contributing to common forms of human obesity. Obes Rev 9:246–250
17. Fischer J, Koch L, Emmerling C et al (2009) Inactivation of the Fto gene protects from obesity. Nature 458:894–898
18. Freathy RM, Timpson NJ, Lawlor DA et al (2008) Common variation in the FTO gene alters diabetes-related metabolic traits to the extent expected given its effect on BMI. Diabetes 57:1419–1426
19. Allison DB, Heshka S, Neale MC, Heymsfield SB (1994) Race effects in the genetics of adolescents' body mass index. Int J Obes Relat Metab Disord 18:363–368
20. Deurenberg P, Deurenberg-Yap M, Guricci S (2002) Asians are different from Caucasians and from each other in their body mass index/body fat per cent relationship. Obes Rev 3:141–146
21. Stevens J, Truesdale KP, Katz EG, Cai J (2008) Impact of body mass index on incident hypertension and diabetes in Chinese Asians, American Whites, and American Blacks: the People's Republic of China Study and the Atherosclerosis Risk in Communities Study. Am J Epidemiol 167:1365–1374
22. Cho YS, Go MJ, Kim YJ et al (2009) A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits. Nat Genet 41:527–534
23. Hotta K, Nakata Y, Matsuo T et al (2008) Variations in the FTO gene are associated with severe obesity in the Japanese. J Hum Genet 53:546–553
24. Ng MC, Tam CH, So WY et al (2010) Implication of genetic variants near NEGR1, SEC16B, TMEM18, ETV5/DGKG, GNPDA2, LIN7C/BDNF, MTCH2, BCDIN3D/FAIM2, SH2B1, FTO, MC4R, and KCTD15 with obesity and type 2 diabetes in 7705 Chinese. J Clin Endocrinol Metab 95:2418–2425