Chr	SNP ID (pos Build 36.3)	Coded/Other allele	Nearby Gene(s)	Stage	*N*	Coded allele freq.	SBP		DBP	
							Beta (SE), mm Hg	*P*	Beta (SE), mm Hg	*P*
				1+2	26,721	0.77	**1.18 (0.17)**	**1.0E-11**	**0.77 (0.11)**	**5.9E-13**
				3	20,236	0.74	**2.13 (0.21)**	**2.6E-23**	**1.34 (0.13)**	**6.5E-27**
				Joint Analysis	46,957	0.75	**1.56 (0.13)**	**7.9E-31**	**1.01 (0.08)**	**1.3E-35**
12	rs35444 (114,036,820)	A/G	*TBX3*	1	19,286	0.75	0.69 (0.19)	3.7E-04	0.54 (0.12)	8.1E-06
				2	10,460	0.75	0.49 (0.28)	0.077	0.48 (0.16)	0.003
				1+2	29,746	0.75	0.63 (0.16)	8.6E-05	0.52 (0.10)	9.6E-08
				3	20,238	0.75	0.64 (0.21)	0.003	0.46 (0.13)	3.0E-04
				Joint Analysis	49,984	0.75	0.63 (0.13)	7.5E-07	**0.50 (0.08)**	**1.3E-10**
Loci previously identified in Europeans										
1	rs880315 (10,719,453)	C/T	*CASZ1*	1	10,765	0.61	0.29 (0.24)	0.226	0.26 (0.14)	0.073
				Follow-up [a]	21,846	0.67	1.03 (0.19)	8.1E-08	**0.72 (0.11)**	**5.9E-11**
				Joint Analysis	32,611	0.65	0.74 (0.15)	7.3E-07	**0.56 (0.09)**	**3.1E-10**
4	rs16998073 (81,541,520)	T/A	*FGF5*	1		N/A	N/A		N/A	
				Follow-up [a]	21,864	0.30	**1.43 (0.20)**	**3.9E-13**	**0.76 (0.11)**	**2.0E-11**
10	rs11191548 (104,836,168)	T/C	*CNNM2 NT5C2 CYP17A1*	1	19,457	0.74	0.91 (0.19)	2.1E-06	0.49 (0.12)	5.6E-05
				Follow-up [a]	21,858	0.73	**1.47 (0.20)**	**4.9E-13**	**0.66 (0.12)**	**1.6E-08**
				Joint Analysis	41,315	0.74	**1.18 (0.14)**	**3.9E-17**	**0.58 (0.08)**	**6.6E-12**
12	rs17249754 (88,584,717)	G/A	*ATP2B1*	1	18,856	0.65	**1.38 (0.18)**	**7.6E-15**	**0.83 (0.11)**	**3.2E-13**
				Follow-up [a]	21,863	0.63	0.94 (0.19)	4.2E-07	0.35 (0.11)	1.2E-03
				Joint Analysis	40,719	0.64	**1.17 (0.13)**	**7.7E-20**	**0.58 (0.08)**	**1.9E-13**

Beta is the effect size on blood pressure in mm Hg per coded allele based on an additive genetic model.

Shown is the top SNP for each independent locus significantly (P<5E-8) associated with systolic and/or diastolic BP on joint analysis in up to 50,373 individuals of East Asian ancestry. Detailed results for the individual loci separately by study are presented in Supplementary Table 2.

For the Fukuoka study and KING study in stage 3, we used genotype data of rs12413409 for rs11191548 (r^2=0.98) at *CNNM2-NT5C2* and a proxy (rs2681472, r^2=0.98) for rs17249754 at *ATP2B1*; linkage disequilibrium coefficient (r^2) was estimated based on 5331 Japanese samples (CAGE-Amagasaki study).

Nat Genet. Author manuscript; available in PMC 2011 August 19.

Four loci—*CASZ1*, *FGF5*, *CNNM2-NT5C2*, and *ATP2B1*—in the table were previously reported to associate with BP in the Japanese replication study,[13] the samples of which constitute the participants in the present GWAS meta-analysis.

[a]Lead SNPs at the loci previously identified in Europeans were directly genotyped for follow-up in part of stage 2 and stage 3 samples. Imputed data were unavailable for an SNP rs16998073 at *FGF5* and only the results for the follow-up analysis are demonstrated in the table.

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,

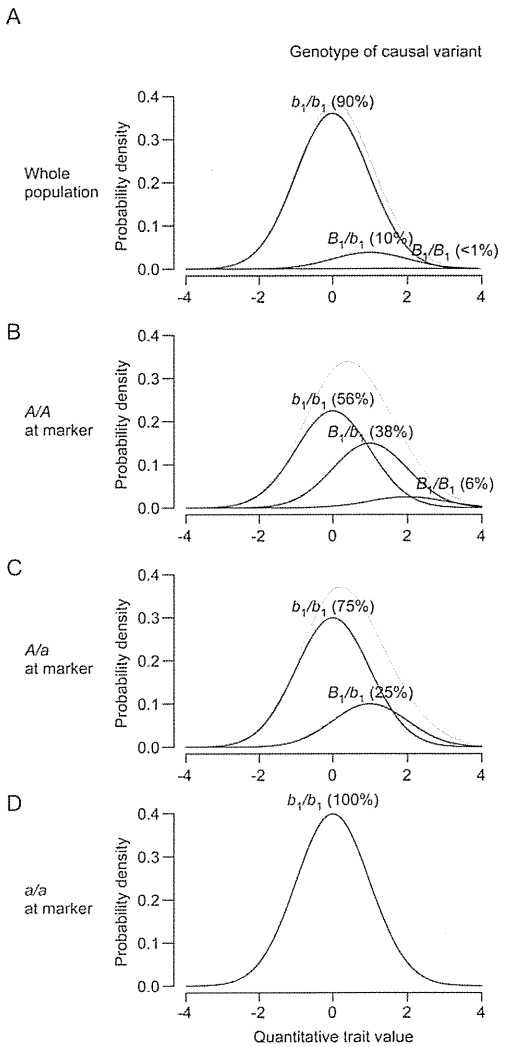


Figure 1. Probability distribution of the QT value within subgroups classified by marker SNP genotypes. (*A*) In the whole population, the total QT distribution (gray curve) comprises a mixture of normal distributions (black curves) with unit variance and the mean 0, 1, or 2, which correspond to genotypes 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%.

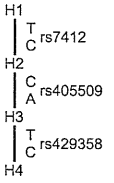
Test of skewness

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

B



C

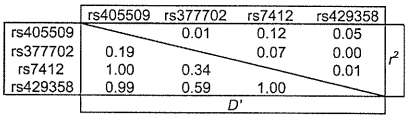


Figure 2. Haplotype classes (*A*), their phylogeny (*B*) for the marker SNP rs405509 showing synthetic association of functional variants rs7412 and rs429358 in the *APOE* locus. LD coefficients between the SNPs associated with LDL-C (*C*). Haplotype frequencies were calculated using the PLINK software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

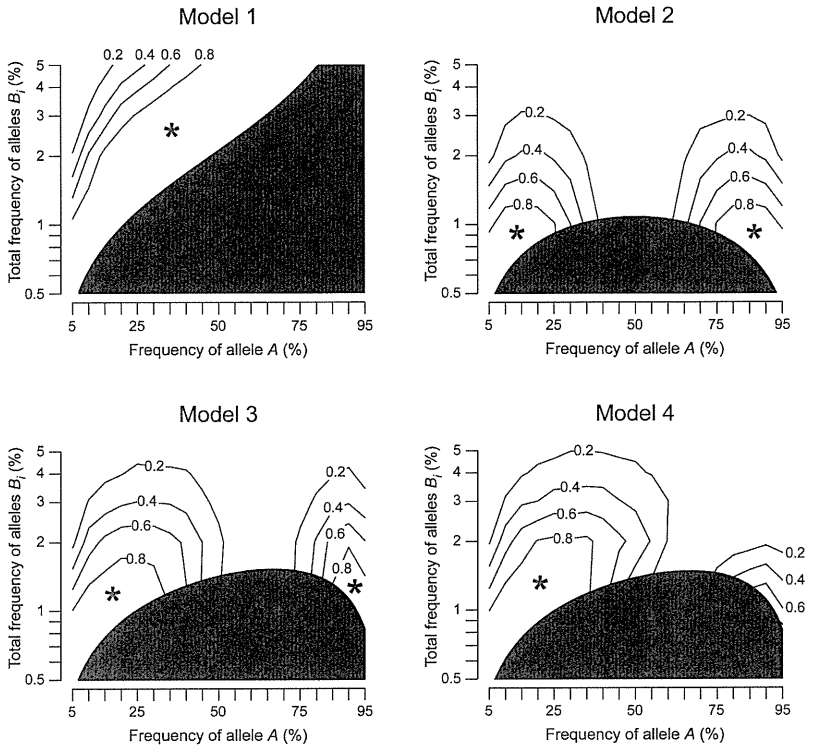


Figure 3. Power for detecting synthetic association by testing heteroscedasticity. The power was computed from simulation under four representative genetic models of synthetic association (see Methods), assuming the strength of marker association (R^2_{mrk}) of 0.00592. Horizontal and vertical axes represent 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

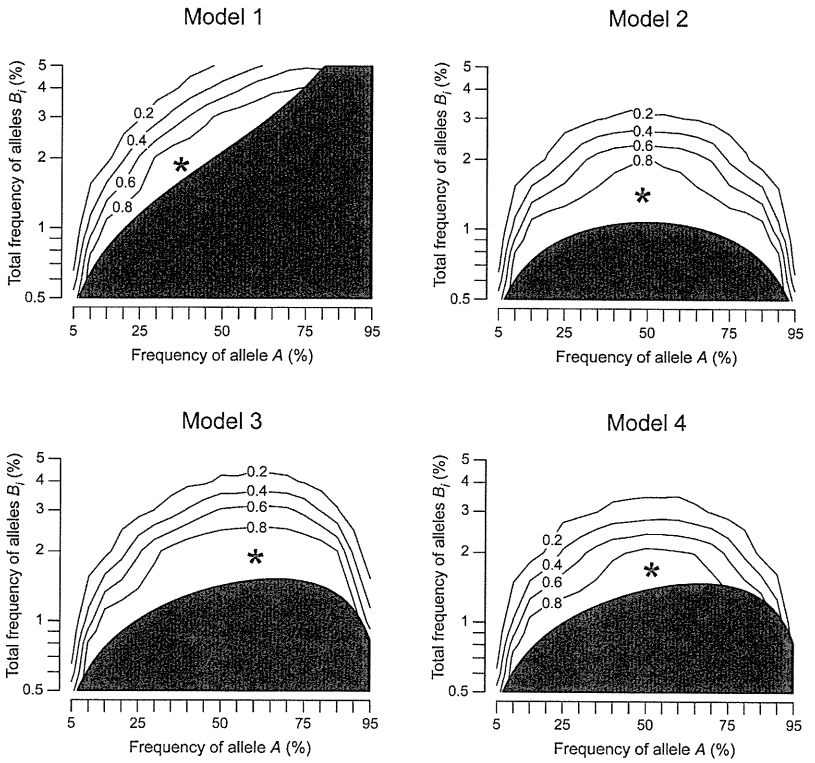


Figure 4. Power for detecting synthetic association by testing skewness. The power was computed from simulation under four representative genetic models, assuming the strength of marker association (R^2_{mrk}) of 0.00592. The format of the figure is the same as Figure 3.

(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.

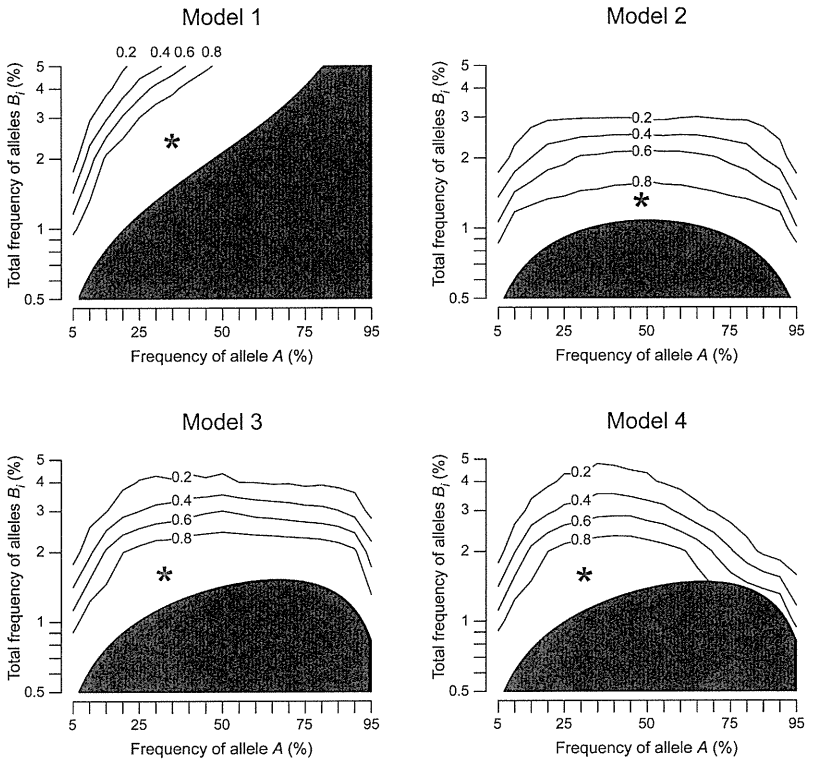


Figure 5. Power for detecting synthetic association by the combined test of heteroscedasticity and skewness. The power was computed from simulation under four representative genetic models, assuming the strength of marker association (R^2_{mrk}) of 0.00592. The format of the figure is the same as Figure 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 $p_A, 1-p_A >> p_{B_i}, p_{c_j}$. 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 (2012) 55:981–995
DOI 10.1007/s00125-011-2370-7

ARTICLE

Association of genetic variation in *FTO* with risk of obesity and type 2 diabetes with data from 96,551 East and South Asians

H. Li · T. O. Kilpeläinen · C. Liu · J. Zhu · Y. Liu · C. Hu · Z. Yang · W. Zhang · W. Bao · S. Cha · Y. Wu · T. Yang · A. Sekine · B. Y. Choi · C. S. Yajnik · D. Zhou · F. Takeuchi · K. Yamamoto · J. C. Chan · K. R. Mani · L. F. Been · M. Imamura · E. Nakashima · N. Lee · T. Fujisawa · S. Karasawa · W. Wen · C. V. Joglekar · W. Lu · Y. Chang · Y. Xiang · Y. Gao · S. Liu · Y. Song · S. H. Kwak · H. D. Shin · K. S. Park · C. H. D. Fall · J. Y. Kim · P. C. Sham · K. S. L. Lam · W. Zheng · X. Shu · H. Deng · H. Ikegami · G. V. Krishnaveni · D. K. Sanghera · L. Chuang · L. Liu · R. Hu · Y. Kim · M. Daimon · K. Hotta · W. Jia · J. S. Kooner · J. C. Chambers · G. R. Chandak · R. C. Ma · S. Maeda · R. Dorajoo · M. Yokota · R. Takayanagi · N. Kato · X. Lin · R. J. F. Loos

Received: 24 August 2011 / Accepted: 10 October 2011 / Published online: 23 November 2011

Abstract

Aims/hypothesis *FTO* harbours the strongest known obesity-susceptibility locus in Europeans. While there is growing evidence for a role for *FTO* in obesity risk in Asians, its association with type 2 diabetes, independently of BMI, remains inconsistent. To test whether there is an association of the *FTO* locus with obesity and type 2 diabetes, we conducted a meta-analysis of 32 populations including 96,551 East and South Asians.

Methods All studies published on the association between *FTO*-rs9939609 (or proxy [$r^2>0.98$]) and BMI, obesity or type 2 diabetes in East or South Asians were invited. Each study group analysed their data according to a standardised analysis plan. Association with type 2 diabetes was also adjusted for BMI. Random-effects meta-analyses were performed to pool all effect sizes.

Results The *FTO*-rs9939609 minor allele increased risk of obesity by 1.25-fold/allele ($p=9.0\times10^{-19}$), overweight by

Electronic supplementary material The online version of this article (doi:10.1007/s00125-011-2370-7) contains peer-reviewed but unedited supplementary material, which is available to authorised users.

H. Li · C. Liu · J. Zhu · D. Zhou · X. Lin (✉)
Institute for Nutritional Sciences,
Shanghai Institutes for Biological Sciences,
Chinese Academy of Sciences,
294 Tai-Yuan Road,
Shanghai 200031, People's Republic of China
e-mail: xlin@sibs.ac.cn

T. O. Kilpeläinen · R. J. F. Loos (✉)
MRC Epidemiology Unit, Institute of Metabolic Science Box 285,
Addenbrooke's Hospital,
Hills Road,
Cambridge CB2 0QQ, UK
e-mail: ruth.loos@mrc-epid.cam.ac.uk

Y. Liu
Institutes of Biomedical Sciences, Fudan University,
Shanghai, People's Republic of China

C. Hu · W. Jia
Shanghai Diabetes Institute,
Department of Endocrinology and Metabolism,
Shanghai Clinical Center of Diabetes,
Shanghai Jiao Tong University Affiliated Sixth People's Hospital,
Shanghai, People's Republic of China

Z. Yang · R. Hu
Department of Endocrinology and Metabolism, Huashan Hospital,
Institute of Endocrinology and Diabetology at Fudan University,
Shanghai Medical School, Fudan University,
Shanghai, People's Republic of China

W. Zhang · J. C. Chambers
Department Epidemiology and Biostatistics,
School of Public Health, Imperial College London,
London, UK

1.13-fold/allele ($p=1.0\times10^{-11}$) and type 2 diabetes by 1.15-fold/allele ($p=5.5\times10^{-8}$). The association with type 2 diabetes was attenuated after adjustment for BMI (OR 1.10-fold/allele, $p=6.6\times10^{-5}$). The *FTO*-rs9939609 minor allele increased BMI by 0.26 kg/m^2 per allele ($p=2.8\times10^{-17}$), WHR by 0.003/allele ($p=1.2\times10^{-6}$), and body fat percentage by 0.31%/allele ($p=0.0005$). Associations were similar using dominant models. While the minor allele is less common in East Asians (12–20%) than South Asians (30–33%), the effect of *FTO* variation on obesity-related traits and type 2 diabetes was similar in the two populations.

Conclusions/interpretation *FTO* is associated with increased risk of obesity and type 2 diabetes, with effect sizes similar in East and South Asians and similar to those observed in Europeans. Furthermore, *FTO* is also associated with type 2 diabetes independently of BMI.

W. Bao · L. Liu
Department of Nutrition and Food Hygiene and MOE Key Lab of Environment and Health, School of Public Health, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, People's Republic of China

S. Cha · J. Y. Kim
Division of Constitutional Medicine Research, Korea Institute of Oriental Medicine, Daejeon, South Korea

Y. Wu
Department of Genetics, University of North Carolina, Chapel Hill, NC, USA

T. Yang
Key Laboratory of Biomedical Information Engineering of Ministry of Education, and Institute of Molecular Genetics, School of Life Science and Technology, Xi'an Jiaotong University, Xi'an, People's Republic of China

A. Sekine · K. Hotta
EBM Research Center, Kyoto University Graduate School of Medicine, Kyoto, Japan

B. Y. Choi
Department of Preventive Medicine, HanYang University College of Medicine, Seoul, South Korea

C. S. Yajnik · C. V. Joglekar
Diabetology Research Centre, KEM Hospital and Research Centre, Pune, India

F. Takeuchi · N. Kato
National Center for Global Health and Medicine, Tokyo, Japan

K. Yamamoto
Division of Genome Analysis, Medical Institute of Bioregulation, Kyushu University, Fukuoka, Japan

Keywords Asians · *FTO* · Meta-analysis · Obesity · Type 2 diabetes

Abbreviations

GWAS	Genome-wide association study
MAF	Minor allele frequency
PAR	Population-attributable risk
SNP	Single-nucleotide polymorphism

Introduction

Large-scale genome-wide association studies (GWAS) in mainly white Europeans have identified at least 50 genetic loci to be robustly associated with obesity-related traits [1–

J. C. Chan · R. C. Ma
Department of Medicine and Therapeutics, Chinese University of Hong Kong, Prince of Wales Hospital, Shatin, Hong Kong Special Administrative Region, People's Republic of China

K. R. Mani · G. R. Chandak
Centre for Cellular and Molecular Biology (CCMB-CSIR), Hyderabad, India

L. F. Been · D. K. Sanghera
University of Oklahoma Health Sciences Center, Oklahoma City, OK, USA

M. Imamura · S. Maeda
Laboratory for Endocrinology and Metabolism, RIKEN Center for Genomic Medicine, Yokohama, Japan

E. Nakashima
Department of Diabetes and Endocrinology, Chubu Rosai Hospital, Nagoya, Japan

N. Lee
USC Office of Population Studies Foundation, University of San Carlos, Cebu, Philippines

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

S. Karasawa · M. Daimon
Third Department of Internal Medicine, and Global Center of Excellence Program Study Group, Yamagata University School of Medicine, Yamagata, Japan

W. Wen · W. Zheng · X. Shu
Division of Epidemiology, Department of Medicine, Vanderbilt School of Medicine, Nashville, TN, USA

12]. A cluster of common variants in the first intron of the fat mass and obesity-associated gene (*FTO*) was the first obesity-susceptibility locus to be identified by two independent GWAS in 2007 [1, 2] and has since been consistently replicated by many others and for a variety of obesity-related traits [7, 9, 13–15]. Of all currently identified obesity-susceptibility loci, the *FTO* locus has the most pronounced effect on BMI and obesity risk, at least in individuals of European descent. Each minor allele of any commonly investigated variant in *FTO* increases BMI by 0.30–0.40 kg/m^2 (equivalent to 870–1,150 g for a person 1.7 m tall) and risk of obesity by ~20% [7, 15]. The minor allele of the *FTO* variant is common (minor allele frequency (MAF)=~42%) in white Europeans, such that 66% of Europeans carry at least one risk allele and 18% carry two risk alleles. Because of the high prevalence of the risk allele and its relatively strong effect on BMI, the *FTO* locus explains most (0.34%), yet little, of the variation in BMI in Europeans [7].

FTO has also been examined as an obesity-susceptibility locus in populations of non-white European origin. While the initial replication efforts in East Asian populations were inconsistent [16, 17], a growing number of studies have provided evidence that genetic variation in *FTO* influences

W. Lu
Shanghai Institute of Preventive Medicine,
Shanghai, People's Republic of China

Y. Chang
National Taiwan University Hospital Bei-Hu branch,
Taipei, Taiwan

Y. Xiang · Y. Gao
Department of Epidemiology, Shanghai Cancer Institute,
Shanghai, People's Republic of China

S. Liu
Center for Metabolic Disease Prevention, School of Public Health and David Geffen School of Medicine, UCLA,
Los Angeles, CA, USA

Y. Song
Division of Preventive Medicine, Brigham & Women's Hospital,
Harvard Medical School,
Boston, MA, USA

S. H. Kwak · K. S. Park
Department of Internal Medicine, Seoul National University College of Medicine,
Seoul, South Korea

H. D. Shin
Department of Life Science, Sogang University,
Seoul, South Korea

C. H. D. Fall
MRC Lifecourse Epidemiology Unit, University of Southampton,
Southampton General Hospital,
Southampton, Hampshire, UK

P. C. Sham · K. S. L. Lam
Li Ka Shing Faculty of Medicine, University of Hong Kong,
Hong Kong, Hong Kong Special Administrative Region,
People's Republic of China

H. Deng
School of Medicine, University of Missouri,
Kansas City, MO, USA

H. Deng
Center of Systematic Biomedical Research, University of Shanghai for Science and Technology,
Shanghai, People's Republic of China

H. Deng
Institute of Bioscience and Biotechnology, School of Science,
Beijing Jiaotong University,
Beijing, People's Republic of China

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

G. V. Krishnaveni
Epidemiology Research Unit, Holdsworth Memorial Hospital,
Mysore, India

L. Chuang
Department of Internal Medicine,
National Taiwan University Hospital,
Taipei, Taiwan

Y. Kim
Department of Preventive Medicine,
Dong-A University College of Medicine,
Busan, South Korea

J. S. Kooner
National Heart & Lung Institute, Hammersmith Hospital,
Hammersmith Campus, Faculty of Medicine,
Imperial College London,
London, UK

R. Dorajoo
Genome Institute of Singapore, Agency for Science,
Technology and Research,
Singapore, Republic of Singapore

R. Dorajoo
Department of Genomics of Common Disease, School of Public Health, Hammersmith Hospital, Imperial College London,
London, UK

M. Yokota
Department of Genome Science, School of Dentistry,
Aichi-Gakuin University,
Nagoya, Japan

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

BMI and obesity risk also in Chinese, Japanese, Korean and Filipino populations [18–27]. A GWAS for BMI in 7,861 Koreans identified variation in *FTO* (rs9939609) as the most significantly associated locus, nearly reaching genome-wide significance ($p=1.5\times10^{-7}$) [28]. Furthermore, literature-based meta-analyses in Asians reported that the minor allele for the rs9939609 *FTO* single-nucleotide polymorphism (SNP) significantly ($p=9\times10^{-9}$) increased the risk of obesity, but no other obesity-related traits were examined [18, 29, 30]. Fewer studies in South Asians have been reported, two of which confirmed the association between the *FTO* locus and obesity susceptibility [31, 32], whereas one did not [33]. The prevalence of the risk allele in East Asians (~20%) and South Asians (~30%) is substantially lower than in Europeans, and the reported effect sizes in both East and South Asians vary widely for BMI (OR 0.13–0.83 kg/m^2 per minor allele) and obesity risk (OR 1.02–1.48 per minor allele) [16, 18, 20–25, 27, 34–39].

FTO was first identified as a type 2 diabetes-susceptibility gene, but, as further adjustment for BMI abolished the association with type 2 diabetes [1], it was suggested that *FTO* is primarily an obesity-susceptibility locus. However, the BMI-independent role of *FTO* in type 2 diabetes remains a matter of debate, particularly in Asians but also in white Europeans. While several studies have reported that the association between the *FTO* locus and risk of type 2 diabetes remained significant after adjustment for BMI [15, 18, 33, 35, 40, 41], others could not confirm this [21, 30, 32, 37, 42].

To firmly establish the association between the *FTO* locus and obesity susceptibility in East and South Asians and to assess its effect size and potential heterogeneity across Asian populations, we performed a systematic meta-analysis of data from 32 populations, including 96,551 men and women, using standardised study-specific association analyses. Furthermore, we examined whether the *FTO* locus is associated with type 2 diabetes independently of its association with BMI.

Methods

Literature search and study identification We designed a meta-analysis based on de novo analyses of data according to a standardised plan to achieve the greatest consistency possible across studies. We identified all published studies (before September 2010) that had examined the association of genetic variation in *FTO* with risk of obesity and type 2 diabetes and with obesity-related continuous traits in East and South Asian adults (age ≥18 years) by a PubMed literature search using the key words 'FTO', 'fat mass and obesity associated gene' and 'genome-wide association study'. References from the identified papers were subsequently screened to identify additional studies and to ensure that the list of eligible studies was complete. The literature search was carried out by two investigators independently, who cross-checked their search results for completeness.

Our literature search identified 38 publications, one of which was excluded because it was a subsample of another identified study. We invited the corresponding authors of the remaining 37 publications to join our meta-analysis, of which 26 agreed to participate and eventually 22 submitted raw data or summary statistics. We also included a Korean population with previously unpublished data (Y. M. Kim, J. Shin, C.B. Lee, M.K. Kim, Y. Tabara, T. Miki and B.Y. Choi), which was presented by a contributing author.

Taken together, our meta-analysis included data for 31 populations from 22 publications and one unpublished study, with 96,551 individuals altogether. The study identification and selection process is illustrated in Fig. 1.

All studies were conducted according to the Declaration of Helsinki. Informed consent was obtained from all participants, and the studies were approved by the ethics committees of the participating institutions.

Genotyping The rs9939609 *FTO* SNP was examined in 18 studies, whereas proxy SNPs were used in 14 studies. More specifically, the rs8050135 SNP was genotyped in 11 studies of East Asians and one of South Asians, and the rs3751812 and rs17817449 SNP were each genotyped once in studies of East Asians (electronic supplementary material [ESM] Table 1). The linkage disequilibrium between rs9939609 and the three proxies (rs8050135, 3751812, rs17817449) is perfect ($r^2=1$) in populations of East Asian origin, based on CHB+JPT data from the HapMap (Rel 24/Phase II). The linkage disequilibrium between rs9939609 and rs8050135 in Indian Asians is very high ($r^2>0.98$), based on a subsample ($n=305$) of the participating Lolipop study.

The genotyping success rate and concordance rate were >95%, and genotype distributions were in Hardy–Weinberg equilibrium ($p>0.01$) in all participating studies (ESM Table 1).

Statistical analysis As case–control definitions and statistical analyses used in the published papers were inconsistent, we asked analysts of each of the participating cohorts to re-analyse their data according to a standardised analysis plan. Summary statistics of each study were subsequently meta-analysed.

Obesity-susceptibility traits and type 2 diabetes Overweight was defined as a BMI ≥24 kg/m^2, and obesity as a BMI ≥28 kg/m^2 according to the definition proposed by the Working Group on Obesity in China [43]. Anthropometric data, including weight, height, waist circumference, hip

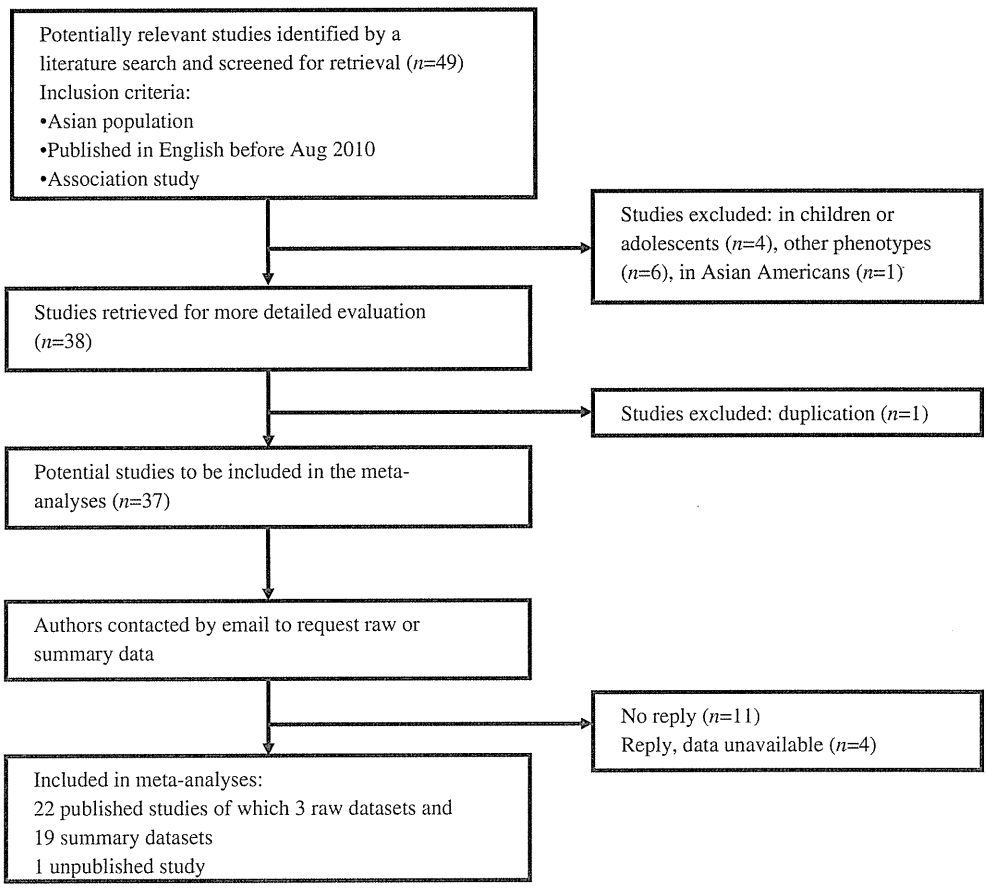


Fig. 1 Study identification and inclusion in the meta-analyses

circumference and body fat percentage, were collected in each study as described previously (ESM Table 1), BMI was calculated as weight (kg) divided by height squared (m^2), and WHR as waist circumference (cm) divided by hip circumference (cm). Raw data were used for analyses.

Type 2 diabetes was defined as meeting one or more of the following criteria: (1) fasting glucose ≥7.0 mmol/l; (2) 2-h glucose ≥11.1 mmol/l; (3) previous diagnosis of type 2 diabetes; (4) HbA_{1c} ≥6.5% (48 mmol/mol); (5) self-reported type 2 diabetes (ESM Table 1).

Study-specific de novo data analyses Association analyses within each study were performed for the total population and for men and women separately using additive and dominant genetic models. The associations of *FTO*-rs9939609 (or proxy) with risk of obesity and type 2 diabetes were assessed with multiple logistic regression models. Generalised linear models were used to assess the associations of *FTO*-rs9939609 (or proxy) with obesity-related continuous traits. In studies with a case–control design, analyses for continuous traits were conducted in control samples only. All analyses were adjusted for age and sex (sex-stratified analyses were only adjusted for age). The association with type 2 diabetes was also analysed with adjustment for BMI. Adjustments were performed by including the covariates (age, sex and/or BMI) as a linear term in the association model.

Summary statistics from the study-specific association analyses were reported in a standardised Excel form by the analysts of each study and collected centrally for meta-analyses.

Meta-analyses Data extraction from the forms and meta-analyses was performed independently by two investigators and cross-checked for consistency. All ambiguities were clarified with the respective analysts before the final meta-analyses.

ORs and beta coefficients from the individual studies were pooled using DerSimonian and Laird random-effects meta-analyses [44]. Meta-analyses were performed of all studies combined. Because of differences in genetic background as well as in susceptibility to obesity and type 2 diabetes, meta-analyses were also stratified by East Asian and South Asian origin of the populations. Furthermore, East Asians were further stratified according to their country of origin.

Between-study heterogeneity was tested by Cochrane's Q test and quantified by the I^2 index. I^2 values of <25%, 25–75% and >75% were defined as low, moderate and high heterogeneity, respectively [45]. To examine the sources of

heterogeneity in our meta-analyses, we performed random-effects meta-regressions, where the between-study variance was estimated with the restricted maximum likelihood approach. Meta-regressions included the following study-specific variables as covariates: year of publication, country of origin, sample size, study design, mean age and mean BMI.

A funnel plot, along with Begg's and Egger's tests, was used to test for the presence of publication bias.

Statistical analyses were performed with the Stata 9.0 software (StataCorp LP, College Station, TX, USA). Meta-analyses and meta-regressions were implemented by the *metan* and *metareg* commands of Stata, respectively. $p<0.05$ was considered to be significant, except for Cochrane's Q test for heterogeneity and Begg's and Egger's tests for publication bias, where a level of $p<0.10$ was used.

The variation in obesity-related continuous traits explained by the *FTO* variant was evaluated using the equation $2f(1-f)a^2$, where f is the frequency of the variant and a is its additive standardised effect [5]. Population-attributable risk (PAR) was calculated as $\mathrm{PAR} = (X-1)/X$. Assuming a multiplicative model, $X = (1-f)^2 + 2f(1-f)\gamma + f^2\gamma^2$, where γ is the estimated OR, and f is the frequency of risk allele [46].

Results

Characteristics of populations included in the meta-analyses Analyses were conducted in Chinese Hans (China Mainland: $n=10$; Singapore: $n=2$), Japanese ($n=7$), Indians ($n=7$), Koreans ($n=4$), Singapore Malays ($n=1$) and Filipinos ($n=1$; Table 1). Fifteen of the populations were case–control designed for obesity ($n=3$) or type 2 diabetes ($n=8$) or both ($n=4$), whereas 17 populations were population-based. The mean age and BMI of the populations ranged from 27.9 to 66.8 years and from 20.5 to 27.1 kg/m^2, respectively. The prevalence in population-based studies ranged from 3.1% to 37.9% for obesity and from 2.9% to 41.9% for type 2 diabetes.

The MAF of *FTO*-rs9939609 (or proxy) is 12–14% in Chinese Hans and Koreans, 18–20% in Japanese and Filipinos, and 30–33% in Singapore Malays and Indians (Table 1).

Associations with obesity and overweight A total of 24 populations ($n_{\mathrm{obese}}=13{,}032$; $n_{\mathrm{overweight}}=22{,}474$; $n_{\mathrm{normalweight}}=35{,}767$) were available for meta-analyses of the association between the *FTO* variant and risk of obesity and overweight.

Each additional *FTO*-rs9939609 minor (A) allele increased the odds of obesity by 1.25 ($p=9.0\times10^{-19}$) compared with normal weight individuals (Fig. 2), and by 1.17 ($p=7.4\times10^{-11}$) compared with non-obese individuals (ESM Fig. 1). Each additional minor allele increased the odds of overweight by 1.13 ($p=1.0\times10^{-11}$; ESM Fig. 2). The odds of obesity and overweight were the same in both East Asian and South Asian populations ($p=0.18$ and 0.84, respectively; ESM Table 2). Associations were similar in men and women (ESM Table 3). The heterogeneity across all studies was low ($13\%\leq I^2\leq19\%$).

When a dominant genetic model was used, the odds were only slightly higher than for the additive genetic model (ESM Table 4).

Association with type 2 diabetes In our meta-analysis of 22 populations ($n_{\mathrm{cases}}=33{,}744$, $n_{\mathrm{controls}}=43{,}549$), each additional *FTO*-rs9939609 minor allele increased the odds of type 2 diabetes by 1.15 ($p=5.5\times10^{-8}$) when adjusted for age and sex (Fig. 3). Further adjustment for BMI attenuated, but did not abolish, the association with type 2 diabetes (OR 1.10, $p=6.6\times10^{-5}$) (Fig. 4). Results were similar in East Asians and South Asians (ESM Table 2), in men and women (ESM Table 3), and when a dominant model was used (ESM Table 4).

The association results across all studies showed moderate heterogeneity ($44\%\leq I^2\leq48\%$; Figs 3 and 4). Meta-regression analyses revealed that the difference in study design contributed to some of the heterogeneity. Subsequent subgroup analyses showed that the association with type 2 diabetes was more pronounced in studies with a case–control design (OR [95% CI]=1.19 [1.14, 1.23], $p=3.7\times10^{-19}$, $I^2=0.0\%$) than in cohort studies (OR [95% CI]=1.09 [0.99, 1.20], $p=0.07$, $I^2=54.4\%$), which showed moderate heterogeneity (ESM Table 5).

Associations with obesity-related continuous traits The meta-analyses of the association of *FTO*-rs9939609 with BMI, waist circumference, hip circumference, WHR and body fat percentage included 30 ($n=71{,}022$), 22 ($n=51{,}543$), 20 ($n=48{,}508$), 20 ($n=48{,}508$) and nine ($n=19{,}580$) populations, respectively.

Each additional *FTO*-rs9939609 minor allele was associated with a 0.26 kg/m^2 higher BMI ($p=2.8\times10^{-17}$; equivalent to ~750 g/allele for a person 1.7 m tall) (Fig. 5), 0.51 cm larger waist circumference ($p=3.0\times10^{-9}$) (ESM Fig. 3), 0.36 cm larger hip circumference ($p=0.0003$) (ESM Fig. 4), 0.003 greater WHR ($p=1.2\times10^{-6}$; ESM Fig. 5), and 0.31% higher body fat percentage ($p=0.0005$) (ESM Fig. 6). All associations were very similar between East and South Asians (ESM Table 2), between men and women (ESM Table 3), or when a dominant genetic model was used (ESM Table 4).

We observed moderate heterogeneity across studies in the associations with BMI and hip circumference (BMI: $I^2=33\%$; hip circumference: $I^2=51\%$; Fig. 5; ESM Fig. 4). Meta-regression suggested that, for BMI, the heterogeneity was

Table 1 Descriptive information of studies included in the meta-analyses, sorted by ethnicity, study design and publication year

Paper	Study	Publication year	Ethnicity	Country	Study design	Sample size						Mean age (years)	Mean BMI (kg/m^2)	*FTO* SNP	MAF
						Obese	OW	NW	T2DM	NFG	QT analyses				
Li et al. [16]	NHAPC	2008	East Asian	China	Population based	472	1,215	1,503	423	1,893	3,190	58.62	24.43	rs9939609	0.11
Sha et al. [55]	GSBC	2009	East Asian	China	Population based	78	326	1,223	n.a.	n.a.	1,627	34.49	22.21	rs9939609	0.12
Hu et al. [56]	SHDS	2009	East Asian	China	Case–control[b]	n.a.	n.a.	n.a.	1,759	1,791	1,791	57.33	23.57	rs8050136	0.12
Li et al. [35]	WDS	2010	East Asian	China	Case–control[a, b]	243	976	1,368	877	1,405	1,405	44.23	21.45	rs9939609	0.12
Cheung et al. [24]	CRISPS	2010	East Asian	China	Case–control[a]	419	n.a.	691	n.a.	n.a.	691	44.98	21.19	rs8050136	0.12
Liu et al. [18]	n.a.	2010	East Asian	China	Case–control[a, b]	277	794	893	1,767	1,961	1,961	58.09	24.52	rs9939609	0.12
Ng et al. [21]	CUHK	2010	East Asian	China	Case–control[b, c]	1,147	2,293	2,432	5,872	583	583	41.31	22.87	rs3751812	0.12
Shu et al. [42]	SGWAS	2010	East Asian	China	Case–control[b]	n.a.	n.a.	n.a.	1,043	2,170	2,170	49.24	23.30	rs9939609	0.12
Wen et al. [57]	FLSGS	2010	East Asian	China	Case–control[b]	n.a.	n.a.	n.a.	1,160	1,127	1,127	59.09	24.13	rs8050136	0.12
Chang et al. [23]	NTUH	2008	East Asian	Taiwan	Case–control[a, b]	737	677	719	881	1,254	1,254	61.19	21.60	rs9939609	0.14
Cha et al. [25]	Kirin	2008	East Asian	Korea	Population based	252	304	361	n.a.	n.a.	917	27.91	26.39	rs17817449	0.14
Cha et al. [58]	KCMS	2009	East Asian	Korea	Population based	61	261	688	n.a.	n.a.	1,010	43.14	22.77	rs8050136	0.12
Kim et al. (unpublished data)	YangPyeung Cardiovascular Cohort Study		East Asian	Korea	Population based	339	995	1,092	194	2,061	2,426	57.60	24.49	rs9939609	0.12
Ng et al. [34]	Korea SNUH	2008	East Asian	Korea	Case–control[b]	n.a.	n.a.	n.a.	758	629	629	64.70	23.52	rs8050136	0.12
Takeuchi et al. [59]	CAGE-Amagasaki	2009	East Asian	Japan	Population based	388	1,562	3,719	n.a.	n.a.	5,660	48.86	22.99	rs9939609	0.19
Takeuchi et al. [59]	CAGE-Fukuoka	2009	East Asian	Japan	Population based	721	3,763	8,076	n.a.	n.a.	12,560	62.59	23.05	rs9939609	0.19
Takeuchi et al. [59]	CAGE-BMI	2009	East Asian	Japan	Population based	168	607	1,006	n.a.	n.a.	1,781	66.82	23.69	rs9939609	0.20
Karasawa et al. [19]	Takahata	2010	East Asian	Japan	Population based	220	886	1,533	215	2,306	2,639	63.04	23.48	rs9939609	0.20
Hotta et al. [20]	GWASJPN obesity	2008	East Asian	Japan	Case–control[a]	1,559	n.a.	1,541	n.a.	n.a.	1,541	47.52	21.21	rs9939609	0.18
Omori et al. [37]	RIKEN T2D	2008	East Asian	Japan	Case–control[b]	n.a.	n.a.	n.a.	4,584	2,262	2,262	44.84	22.86	rs8050136	0.20
Takeuchi et al. [59]	CAGE-T2DM	2009	East Asian	Japan	Case–control[b]	n.a.	n.a.	n.a.	6,781	7,307	n.a.	64.35	23.47	rs9939609	0.19
Marvelle et al. [27]	CLHNS	2008	East Asian	Philippines	Population based	321	560	836	155	1,463	1,717	48.51	24.31	rs9939609	0.18
Tan et al. [22]	SP2	2008	East Asian	Singapore (Chinese)	Population based	195	624	1,609	145	2,248	2,430	48.11	22.88	rs8050136	0.12
Tan et al. [22]	SiMES	2008	East Asian	Singapore (Malays)	Population based	848	826	846	787	1,248	2,520	59.04	26.38	rs8050136	0.30
Tan et al. [22]	SDCS	2008	East Asian	Singapore (Chinese)	Case–control[c]	426	809	757	n.a.	n.a.	n.a.	64.27	25.34	rs8050136	0.14
Chambers et al. [6]	LOLIPOP (IA317)	2008	South Asian	India	Population based	727	858	536	434	1,651	2,247	48.22	26.83	rs8050136	0.33
Chambers et al. [6]	LOLIPOP (IA610)	2008	South Asian	India	Population based	2,479	2,647	1,423	1,780	4,715	7,060	55.38	27.14	rs8050136	0.32
Tan et al. [22]	SINDI	2008	South Asian	India	Population	760	910	858	974	1,348	2,528	58.01	26.20	rs8050136	0.33

Table 1 (continued)

Paper	Study	Publication year	Ethnicity	Country	Study design	Sample size						Mean age (years)	Mean BMI (kg/m^2)	*FTO* SNP	MAF
						Obese	OW	NW	T2DM	NFG	QT analyses				
Yajnik et al. [33]	Parthenon	2009	South Asian	India	Population based	136	320	511	n.a.	n.a.	967	32.44	23.76	rs9939609	0.33
Yajnik et al. [33]	PMNS	2009	South Asian	India	Population based	59	271	1,546	50	1,681	1,876	32.71	20.83	rs9939609	0.31
Sanghera et al. [40]	Sikh Diabetes Study	2008	South Asian	India	Case–control[b]	n.a.	n.a.	n.a.	1,138	765	765	50.85	26.25	rs9939609	0.31
Yajnik et al. [33]	WELLGEN	2009	South Asian	India	Case–control[b]	n.a.	n.a.	n.a.	1,967	1,681	1,681	32.39	20.50	rs9939609	0.31

Individuals from CAGE-T2DM study were selected from other three CAGE population-based studies

[a] Obese case–control study

[b] T2DM case–control study

[c] Obese case–control study conducted in T2DM cases

n.a., data not available or not used in meta-analysis; NFG, normal fasting glucose; NW, normal weight; OW, overweight; QT, quantitative trait; T2DM, type 2 diabetes

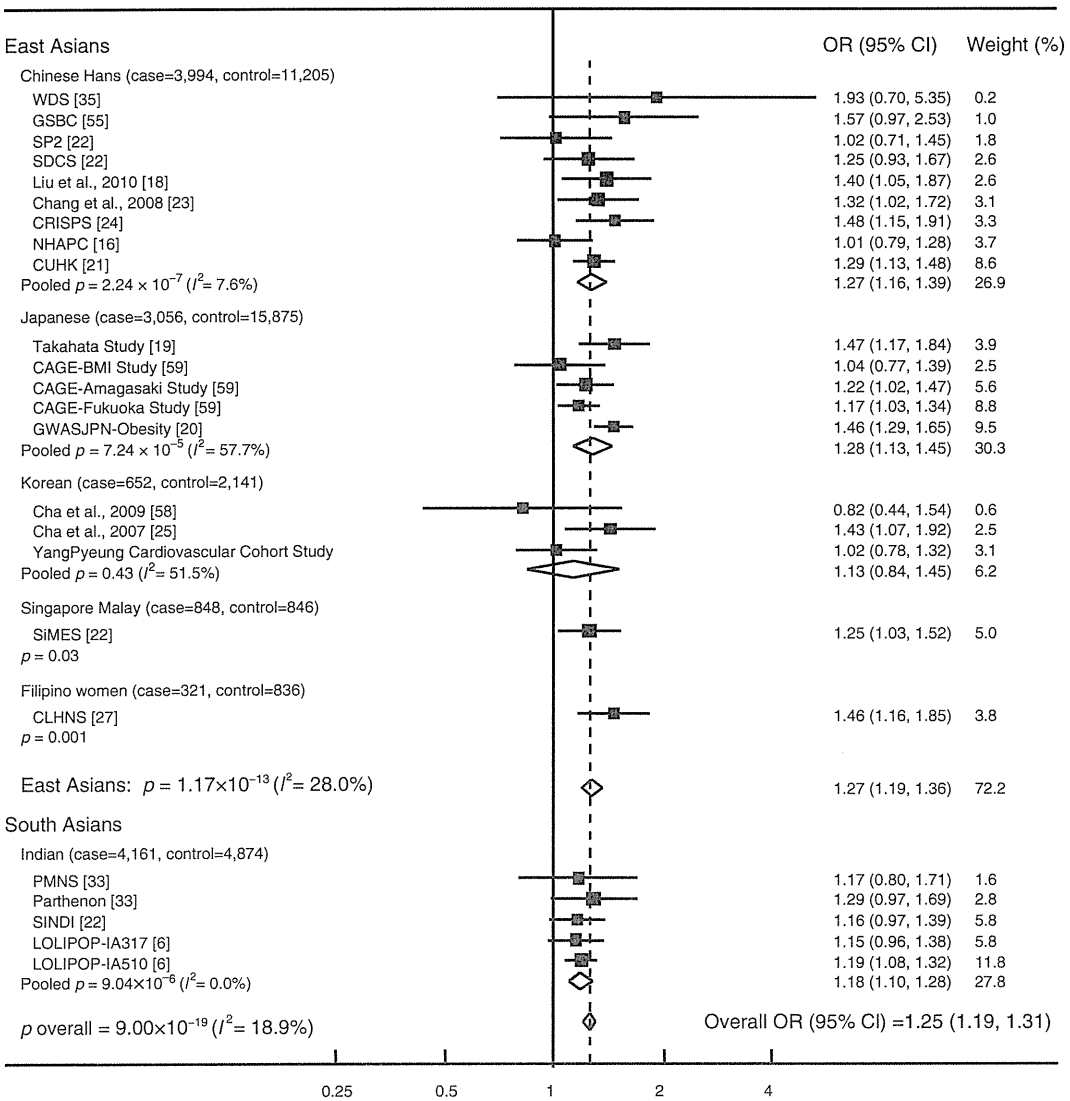


Fig. 2 Association of *FTO*-rs9939609 (or proxy) with obesity. Study-specific association analyses assumed an additive genetic model, comparing obese with normal-weight individuals, adjusted for age and sex. Effect sizes were combined using random-effects meta-analyses (DerSimonian–Laird method)

mainly due to difference in mean age and mean BMI among different populations. For hip circumference, the heterogeneity was mainly attributed to difference in mean BMI, i.e. the effect of the *FTO* minor allele tended to be larger in populations with a mean BMI ≥24 kg/m^2, compared with those with a mean BMI <24 kg/m^2.

FTO-rs9939609 explained 0.16% and 0.20% of the inter-individual variation in BMI in East and South Asian populations, respectively. The proportion of variation in other obesity-related continuous traits explained by *FTO*-rs9939609 was <0.10% (ESM Table 2).

Publication bias The funnel plots for the associations with obesity, type 2 diabetes, waist circumference, WHR and body fat percentage were symmetrical and the results for Begg's and Egger's tests were non-significant ($p \geq 0.10$), indicating that our results were not affected by publication bias (ESM Fig. 7). However, there was some evidence of publication bias and/or genetic heterogeneity for BMI (Begg's test, $p=0.08$; Egger's test, $p=0.07$) and hip circumference (Begg's test, $p=0.03$; Egger's test, $p=0.08$; ESM Fig. 7).

Discussion

This meta-analysis, combining data of 96,551 Asians from 32 populations, further confirms that genetic variation in *FTO* is associated with increased risk of obesity in East and South Asians. Despite differences in genetic background and obesity susceptibility between East and South Asians, the effect of *FTO* on obesity and related traits was generally