nature genetics

Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits

Jian Yang[1,2], Teresa Ferreira[3], Andrew P Morris[3], Sarah E Medland[1], Genetic Investigation of ANthropometric Traits (GIANT) Consortium[4], DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium[4], Pamela A F Madden[5], Andrew C Heath[5], Nicholas G Martin[1], Grant W Montgomery[1], Michael N Weedon[6], Ruth J Loos[7], Timothy M Frayling[6], Mark I McCarthy[3,8], Joel N Hirschhorn[9–13], Michael E Goddard[14,15] & Peter M Visscher[1,2,16]

© 2012 Nature America, Inc. All rights reserved.

We present an approximate conditional and joint association analysis that can use summary-level statistics from a meta-analysis of genome-wide association studies (GWAS) and estimated linkage disequilibrium (LD) from a reference sample with individual-level genotype data. Using this method, we analyzed meta-analysis summary data from the GIANT Consortium for height and body mass index (BMI), with the LD structure estimated from genotype data in two independent cohorts. We identified 36 loci with multiple associated variants for height (38 leading and 49 additional SNPs, 87 in total) via a genome-wide SNP selection procedure. The 49 new SNPs explain approximately 1.3% of variance, nearly doubling the heritability explained at the 36 loci. We did not find any locus showing multiple associated SNPs for BMI. The method we present is computationally fast and is also applicable to case-control data, which we demonstrate in an example from meta-analysis of type 2 diabetes by the DIAGRAM Consortium.

Genome-wide association studies have been successful in identifying genes and pathways involved in the development of human complex traits and diseases[1,2]. For many traits, such as height and BMI, and diseases, such as type 2 diabetes (T2D) and breast cancer, an increasing number of genetic variants have been identified that are associated with trait variation by performing GWAS with continually increasing sample sizes or meta-analyses of multiple studies[3–6], in line with a pattern of polygenic inheritance. Usually, SNPs are tested for associations with a trait on the basis of a single-SNP model, and the SNP showing the strongest statistical evidence for association in a genomic region (for example, a 2-Mb window centered on the locus) is reported to represent the association in this region. Implicit assumptions, often untested, are that the detected association at the top SNP captures the maximum amount of variation in the region by its LD with an unknown causal variant and that other SNPs in the vicinity show association because they are correlated with the top SNP. There are a number of reasons why these assumptions may not be met. First, even if there is a single underlying, causal variant, a single genotyped or imputed SNP may not capture the overall amount of variation at this locus[7,8]. Second, there may be multiple causal variants at the locus, in which case, a single SNP is unlikely to account for all the LD between the unknown causal variants and the genotyped or imputed SNPs at the locus. Therefore, the total variation that could be explained at a locus may be underestimated if only the most significant SNP in the region is selected.

Conditional analysis has been used as a tool to identify secondary association signals at a locus[3,9,10], involving association analysis conditioning on the primary associated SNP at the locus to test whether there are any other SNPs significantly associated. A more general and comprehensive strategy would be to perform a conditional analysis, starting with the top associated SNP, across the whole genome followed by a stepwise procedure of selecting additional SNPs, one by one, according to their conditional *P* values. Such a strategy would allow the discovery of more than two associated SNPs at a locus[7,11]. For meta-analysis of a large number of participating studies, however, pooled individual-level genotype data are usually unavailable, such that conditional analysis can only be performed at the level of individual studies. Summary results from individual studies are then collected and combined through a second round of meta-analysis. This procedure is administratively

[1]Queensland Institute of Medical Research, Brisbane, Queensland, Australia. [2]University of Queensland Diamantina Institute, University of Queensland, Princess Alexandra Hospital, Brisbane, Queensland, Australia. [3]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK. [4]A full list of members is provided in the **Supplementary Note**. [5]Department of Psychiatry, Washington University, Saint Louis, Missouri, USA. [6]Genetics of Complex Traits, Peninsula College of Medicine and Dentistry, University of Exeter, Exeter, UK. [7]Medical Research Council (MRC) Epidemiology Unit, Institute of Metabolic Science, Addenbrooke's Hospital, Cambridge, UK. [8]Oxford Centre for Diabetes, Endocrinology and Metabolism, Headington, Oxford, UK. [9]Program in Genomics, Children's Hospital, Boston, Massachusetts, USA. [10]Division of Genetics, Children's Hospital, Boston, Massachusetts, USA. [11]Division of Endocrinology, Children's Hospital, Boston, Massachusetts, USA. [12]Broad Institute, Cambridge, Massachusetts, USA. [13]Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA. [14]Department of Food and Agricultural Systems, The University of Melbourne, Parkville, Victoria, Australia. [15]Biosciences Research Division, Department of Primary Industries, Bundoora, Victoria, Australia. [16]The Queensland Brain Institute, The University of Queensland, Brisbane, Queensland, Australia. Correspondence should be addressed to P.M.V. (peter.visscher@uq.edu.au).

Received 19 August 2011; accepted 6 February 2012; published online 18 March 2012; doi:10.1038/ng.2213

Table 1 Summary of 87 multiple associated SNPs at 36 loci for height with $P < 5 \times 10^{-8}$ in the joint analysis using the ARIC cohort as a reference sample for LD

					GIANT single-SNP meta-analysis				Joint analysis, LD from ARIC cohort				
SNP	Chr.	Location (bp)	Nearest gene	Allele 1	Freq.	β	P	q_L^2 (%)	Freq.	b	P	r	q^2 (%)
rs17346452	1	170319910	*DNM3*	T	0.727	–0.038	2.5×10^{-19}	0.059	0.706	–0.040	2.8×10^{-21}	0.109	0.062
rs2421992	1	170507874	*DNM3*	T	0.701	0.021	2.6×10^{-6}		0.713	0.025	2.6×10^{-8}		0.023
rs6684205	1	216676325	*TGFB2*	A	0.714	–0.033	1.6×10^{-15}	0.045	0.711	–0.035	1.9×10^{-17}	0.051	0.048
rs11118171	1	217114492	*LYPLAL1*	A	0.631	0.025	2.2×10^{-10}		0.644	0.025	2.4×10^{-10}	–0.084	0.030
rs11118346	1	217810342	*SLC30A10*	T	0.464	–0.026	1.7×10^{-12}	0.036	0.469	–0.025	6.9×10^{-12}		0.035
rs4665736	2	25041103	*RBJ*	T	0.535	0.034	1.3×10^{-18}	0.058	0.533	0.029	3.7×10^{-14}	0.123	0.050
rs11694842	2	25336474	*DNMT3A*	A	0.669	0.028	2.6×10^{-12}		0.660	0.026	1.1×10^{-10}		0.033
rs1367226	2	55943044	*EFEMP1*	A	0.434	–0.005	2.0×10^{-1}		0.428	–0.027	5.0×10^{-11}	–0.421	0.007
rs3791675	2	55964813	*EFEMP1*	T	0.234	–0.050	1.1×10^{-28}	0.091	0.249	–0.063	3.0×10^{-37}		0.116
rs1541777	2	219295535	*TTLL4*	A	0.526	0.025	1.0×10^{-11}		0.513	0.022	6.4×10^{-9}	0.062	0.028
rs6741325	2	219615943	*CCDC108*	C	0.902	0.048	2.0×10^{-14}	0.043	0.902	0.044	5.2×10^{-12}	0.060	0.039
rs16859517	2	219657428	*NHEJ1*	T	0.036	0.073	5.8×10^{-12}		0.038	0.064	2.6×10^{-9}		0.034
rs7598759	2	232030200	*NCL*	T	0.454	–0.022	6.2×10^{-8}		0.419	–0.023	1.2×10^{-8}	–0.013	0.025
rs2580816	2	232506210	*NPPC*	T	0.197	–0.041	7.3×10^{-17}	0.055	0.185	–0.041	2.7×10^{-17}	–0.016	0.056
rs7571716	2	233149664	*EIF4E2*	T	0.292	0.028	7.1×10^{-12}		0.287	0.027	3.6×10^{-11}		0.033
rs4676386	2	241423659	*KIF1A*	A	0.483	0.023	1.5×10^{-9}		0.477	0.022	5.7×10^{-9}	–0.028	0.027
rs12694997	2	241911659	*37500*	A	0.244	–0.027	5.4×10^{-10}	0.029	0.234	–0.027	6.2×10^{-10}		0.028
rs7652177	3	173451771	*FNDC3B*	C	0.496	–0.031	3.9×10^{-16}		0.491	–0.029	1.2×10^{-14}	–0.060	0.045
rs572169	3	173648421	*GHSR*	T	0.313	0.036	1.0×10^{-18}	0.056	0.315	0.034	4.8×10^{-17}		0.053
rs6784185	3	186955759	*IGF2BP2*	A	0.203	–0.008	8.8×10^{-2}		0.205	–0.034	2.2×10^{-10}	0.523	0.009
rs720390	3	187031377	*IGF2BP2*	A	0.386	0.031	1.8×10^{-14}	0.046	0.378	0.045	1.5×10^{-22}		0.067
rs16896276	4	17624254	*LCORL*	A	0.269	0.041	8.2×10^{-23}		0.256	0.030	1.8×10^{-12}	0.248	0.051
rs2061455	4	17644348	*LCORL*	A	0.847	0.072	7.8×10^{-39}	0.139	0.841	0.063	7.2×10^{-29}		0.122
rs17720281	4	145763226	*HHIP*	T	0.406	0.047	1.7×10^{-30}		0.437	0.031	9.2×10^{-13}	–0.389	0.074
rs7689420	4	145787802	*HHIP*	T	0.163	–0.069	1.1×10^{-41}	0.133	0.164	–0.054	2.9×10^{-23}		0.106
rs7731703	5	32730699	*NPR3*	T	0.312	–0.030	1.4×10^{-10}		0.324	–0.032	1.4×10^{-11}	–0.143	0.043
rs1173735	5	32807136	*NPR3*	A	0.739	–0.030	1.3×10^{-12}		0.738	–0.042	5.1×10^{-23}	0.181	0.051
rs1173727	5	32866278	*C5orf23*	T	0.394	0.036	1.5×10^{-20}	0.063	0.409	0.040	1.4×10^{-24}	–0.071	0.070
rs11745439	5	33265791	*TARS*	A	0.288	–0.028	1.3×10^{-11}		0.271	–0.026	9.9×10^{-10}		0.031
rs4620037	5	170807702	*FGF18*	A	0.801	0.032	3.6×10^{-12}		0.788	0.035	4.2×10^{-14}	0.049	0.037
rs1529701	5	170933582	*FGF18*	T	0.292	–0.023	4.9×10^{-8}		0.309	–0.024	1.6×10^{-8}	0.001	0.023
rs12153391	5	171136043	*FBXW11*	A	0.255	–0.033	2.4×10^{-13}	0.042	0.244	–0.033	4.3×10^{-13}	0.027	0.042
rs4868126	5	171216074	*FBXW11*	T	0.396	–0.031	3.3×10^{-13}		0.397	–0.030	6.8×10^{-13}		0.046
rs4246079	6	6834817	*RREB1*	A	0.118	–0.039	1.9×10^{-9}		0.124	–0.039	2.5×10^{-9}	0.001	0.033
rs3812163	6	7670759	*BMP6*	A	0.540	–0.037	5.3×10^{-22}	0.069	0.541	–0.037	1.7×10^{-22}	0.023	0.070
rs9942510	6	7745305	*BMP6*	A	0.162	0.029	1.3×10^{-8}		0.156	0.030	4.3×10^{-9}		0.024
rs12204421	6	33736841	*ITPR3*	A	0.738	0.030	5.6×10^{-12}		0.747	0.030	1.0×10^{-11}	0.001	0.036
rs12214804	6	34296844	*HMGA1*	T	0.925	–0.079	3.5×10^{-26}	0.090	0.922	–0.082	9.9×10^{-28}	0.064	0.093
rs3800461	6	34724300	*C6orf106*	C	0.124	0.045	6.4×10^{-15}		0.126	0.046	7.3×10^{-16}	–0.054	0.046
rs6899744	6	35394273	*DEF6*	T	0.016	–0.138	2.2×10^{-15}	0.063	0.018	–0.131	5.5×10^{-14}		0.060
rs648831	6	81012927	*BCKDHB*	T	0.493	0.028	2.7×10^{-13}		0.509	0.025	3.5×10^{-11}	–0.064	0.036
rs310402	6	81857211	*FAM46A*	T	0.462	–0.030	2.6×10^{-15}	0.045	0.452	–0.028	5.6×10^{-14}		0.042
rs6569648	6	130390812	*L3MBTL3*	T	0.761	–0.036	3.7×10^{-16}	0.048	0.768	–0.036	2.4×10^{-16}	–0.011	0.049
rs6921207	6	131369649	*EPB41L2*	A	0.366	0.021	3.0×10^{-8}		0.362	0.021	4.9×10^{-8}		0.022
rs543650	6	152152636	*ESR1*	T	0.396	–0.032	6.4×10^{-13}	0.050	0.399	–0.029	3.5×10^{-11}	–0.115	0.046
rs3020418	6	152386855	*ESR1*	A	0.300	0.028	5.9×10^{-12}		0.286	0.026	1.7×10^{-10}		0.032
rs4470914	7	19583047	*TWISTNB*	T	0.181	0.033	6.7×10^{-11}	0.033	0.180	0.033	2.3×10^{-11}	0.010	0.034
rs12538581	7	20365642	*ITGB8*	A	0.502	–0.024	4.7×10^{-10}		0.497	–0.024	1.5×10^{-10}		0.030
rs10958476	8	57258362	*PLAG1*	T	0.787	–0.042	1.4×10^{-18}		0.833	–0.036	1.6×10^{-13}	–0.167	0.053
rs7460090	8	57356717	*RDHE2*	T	0.873	0.055	8.6×10^{-22}	0.068	0.882	0.047	2.2×10^{-16}		0.059
rs473902	9	97296056	*PTCH1*	T	0.922	0.074	7.0×10^{-20}	0.082	0.913	0.064	7.3×10^{-15}	–0.208	0.071
rs10512248	9	97299524	*PTCH1*	T	0.660	–0.033	1.5×10^{-16}		0.658	–0.026	1.2×10^{-10}	–0.020	0.040
rs10990303	9	97450226	*PTCH1*	T	0.227	0.030	1.4×10^{-10}		0.208	0.029	2.4×10^{-10}	0.033	0.031
rs2025151	9	98201333	*ZNF367*	C	0.810	–0.041	3.1×10^{-16}		0.806	–0.041	5.0×10^{-16}		0.053
rs13302480	9	117505134	*37226*	C	0.115	–0.036	7.2×10^{-9}		0.109	–0.037	1.9×10^{-9}	0.041	0.028
rs7869550	9	118174617	*PAPPA*	A	0.796	0.030	1.5×10^{-10}	0.029	0.799	0.031	1.6×10^{-11}		0.031
rs7849585	9	138251691	*QSOX2*	T	0.334	0.032	4.6×10^{-15}	0.048	0.318	0.031	3.7×10^{-14}	–0.031	0.046
rs8413	9	138443132	*INPP5E*	T	0.599	–0.026	3.0×10^{-11}		0.584	–0.024	4.2×10^{-10}		0.031
rs77993	10	8058852	*ZMIZ1*	A	0.439	–0.023	2.1×10^{-9}		0.426	–0.024	3.6×10^{-10}	–0.037	0.028

(continued)

© 2012 Nature America, Inc. All rights reserved.

Table 1 Continued

					GIANT single-SNP meta-analysis				Joint analysis, LD from ARIC cohort				
SNP	Chr.	Location (bp)	Nearest gene	Allele 1	Freq.	β	P	q_L^2 (%)	Freq.	b	P	r	q^2 (%)
rs2145998	10	80791702	*PPIF*	A	0.480	–0.025	2.0×10^{-11}	0.033	0.490	–0.026	7.2×10^{-12}		0.034
rs11107116	12	92502635	*SOCS2*	T	0.223	0.052	4.0×10^{-32}	0.098	0.216	0.056	4.2×10^{-37}	0.095	0.106
rs2885691	12	92646350	*CRADD*	T	0.436	–0.029	2.3×10^{-14}		0.439	–0.034	1.1×10^{-18}		0.049
rs7980687	12	122388664	*SBNO1*	A	0.206	0.036	6.7×10^{-14}	0.043	0.192	0.034	4.0×10^{-13}	0.045	0.041
rs1809889	12	123367179	*FAM101A*	T	0.290	0.032	1.1×10^{-12}		0.293	0.030	7.7×10^{-12}		0.040
rs12440667	15	72018492	*LOXL1*	T	0.467	–0.028	1.4×10^{-12}		0.481	–0.030	2.2×10^{-14}	–0.051	0.043
rs5742915	15	72123686	*PML*	T	0.527	–0.031	8.2×10^{-14}	0.049	0.516	–0.032	3.4×10^{-15}		0.051
rs11259936	15	82371586	*ADAMTSL3*	A	0.484	–0.042	1.4×10^{-29}	0.091	0.484	–0.038	7.5×10^{-24}	–0.164	0.082
rs12148239	15	82432022	*ADAMTSL3*	T	0.728	0.031	1.1×10^{-13}		0.724	0.024	8.8×10^{-9}		0.031
rs4932429	15	87164536	*ACAN*	C	0.510	–0.025	2.7×10^{-10}		0.520	–0.022	2.3×10^{-8}	0.047	0.028
rs16942341	15	87189909	*ACAN*	T	0.031	–0.134	2.2×10^{-24}	0.110	0.027	–0.116	1.8×10^{-18}	–0.122	0.095
rs2280470*	15	87196630	*ACAN*	A	0.334	0.039	1.5×10^{-22}		0.330	0.036	6.3×10^{-19}		0.065
rs12916269	15	98347739	*ADAMTS17*	A	0.426	–0.028	9.8×10^{-13}		0.409	–0.030	9.6×10^{-15}	–0.069	0.043
rs2035344	15	98507671	*ADAMTS17*	A	0.687	–0.024	1.6×10^{-8}		0.699	–0.024	1.9×10^{-8}	0.049	0.025
rs4965598	15	98577137	*ADAMTS17*	T	0.681	–0.035	1.4×10^{-18}	0.056	0.680	–0.035	1.9×10^{-18}		0.056
rs8182364	17	44373024	*SNF8*	A	0.461	0.023	7.8×10^{-10}		0.462	0.021	6.8×10^{-9}	0.044	0.025
rs2072153	17	44745013	*ZNF652*	C	0.306	0.026	7.4×10^{-11}	0.031	0.308	0.026	4.6×10^{-10}		0.030
rs227724	17	52133816	*C17orf67*	A	0.657	–0.027	7.0×10^{-12}	0.035	0.656	–0.026	6.9×10^{-11}	–0.029	0.033
rs4794665	17	52205328	*C17orf67*	A	0.485	0.024	2.0×10^{-10}		0.490	0.023	7.5×10^{-10}		0.028
rs2079795	17	56851431	*C17orf82*	T	0.329	0.040	4.0×10^{-23}	0.071	0.331	0.039	6.4×10^{-23}	–0.005	0.071
rs12451513	17	56997110	*NACA2*	T	0.602	–0.022	2.0×10^{-7}		0.605	–0.022	4.2×10^{-8}		0.024
rs2137143	17	59159133	*LYK5*	T	0.055	0.049	3.1×10^{-8}		0.059	0.062	2.8×10^{-12}	0.147	0.033
rs2727300	17	59319130	*GH2*	A	0.275	0.037	4.1×10^{-18}	0.056	0.281	0.041	4.7×10^{-22}		0.062
rs12458127	18	44911356	*DYM*	T	0.076	–0.056	1.4×10^{-13}		0.055	–0.042	3.9×10^{-8}	0.218	0.034
rs9967417	18	45213498	*DYM*	C	0.565	–0.038	1.4×10^{-22}	0.074	0.563	–0.033	5.2×10^{-17}		0.065
rs6060154	20	33063262	*TRPC4AP*	A	0.264	0.023	5.9×10^{-8}		0.278	0.033	9.5×10^{-15}	0.040	0.030
rs143384	20	33489170	*GDF5*	A	0.581	–0.064	2.5×10^{-55}	0.206	0.595	–0.053	4.5×10^{-33}	–0.390	0.171
rs6060739	20	34031006	*C20orf152*	T	0.188	0.052	2.8×10^{-25}		0.184	0.032	2.4×10^{-9}		0.053
Total								2.5					4.1

SNPs on different chromosomes or more than 10 Mb distant from each other are assumed to be uncorrelated. The empirical variance of phenotypes of height estimated from the summary statistic is 0.967. Chr., chromosome; allele 1, reference allele; freq.: frequency of the reference allele; β, marginal effect; $q_L{}^2$, variance explained based on marginal SNP effect for the leading SNPs; b, joint effect; q^2, variance explained based on the joint SNP effect; r, LD correlation between a SNP and the next adjacent SNP at a locus.

onerous. It often takes months to organize and perform a single round of this kind of conditional meta-analysis, and it would be extremely time-consuming and therefore impractical to implement a stepwise selection procedure in this manner.

We propose an approximate conditional and joint analysis approach using summary-level statistics from a meta-analysis and LD corrections between SNPs estimated from a reference sample, such as a subset of the meta-analysis sample, using an approach similar to one previously described[12]. We adopt a genome-wide stepwise selection procedure to select SNPs on the basis of conditional P values and estimate the joint effects of all selected SNPs after the model has been optimized. We applied this method to meta-analysis for height and BMI from the GIANT Consortium and validated results by prediction analysis in independent samples. We extended the procedure to the analysis of case-control data and demonstrate its power with an example of meta-analysis data for T2D.

RESULTS

Loci with multiple associated variants

Using summary statistics (effect size, standard error and allele frequency) of ~2.5 million SNPs from the GIANT meta-analysis of 133,653 individuals for height[3] and 123,865 individuals for BMI[4] along with SNP LD estimated in 6,654 unrelated European-Americans selected from the Atherosclerosis Risk in Communities (ARIC) study (Online Methods), we identified 247 jointly associated SNPs for height and 33 for BMI with $P < 5 \times 10^{-8}$ (**Supplementary Tables 1–3**).

For the convenience of presentation and the summary of results, we define a locus as a chromosomal region at which adjacent pairs of associated SNPs are less than 1 Mb distant, and we define alleles of two SNPs to be positively (negatively) correlated if their disequilibrium parameter D is positive (negative)[13]. Of the 247 SNPs associated with height, 87 at 36 loci represent multiple associated SNPs within a single locus (**Table 1** and **Supplementary Tables 1** and **2**), and all 36 loci are located in the genomic regions known to be associated with height[3]. We did not find any locus with multiple associated SNPs for BMI (**Supplementary Tables 1** and **3**). For most of the height-associated loci, multiple associated variants were detected, mainly because of their very low LD ($r^2 < 0.01$), despite their relatively close physical proximity (**Table 1**). In this case, the effect sizes from a joint analysis were little different from those from single-SNP analyses. For some loci, SNPs were in modest LD and their increasing alleles were positively correlated. One example of this is the rs17720281 and rs7689420 SNP pair at the *HHIP* locus on chromosome 4 (**Table 1**), where effect sizes for these SNPs were therefore overestimated in single-SNP analyses. In the joint analysis, although the joint effects were smaller compared to the marginal effects, these SNPs still reached genome-wide significance, and the variance explained by them collectively was larger than that if we only considered the leading SNP at that locus. For the loci at which the increasing alleles of at least two SNPs were negatively correlated, the SNP effects were underestimated in single-SNP analyses, meaning that some associated variants may be undetected. In other words, the joint

© 2012 Nature America, Inc. All rights reserved.

Table 2 Prediction analysis based on the SNPs at the 36 loci with multiple associated SNPs

	Prediction in ARIC		Prediction in QIMR	
	g_{87}	g_{49}	g_{87}	g_{49}
Slope	0.979	0.953	1.076	0.880
S.e.	0.060	0.095	0.076	0.123
P	1.6×10^{-58}	2.3×10^{-23}	4.3×10^{-44}	8.9×10^{-13}
R^2	0.038	0.015	0.048	0.013

Shown are the results of a linear regression analysis of the observed height phenotype on a single predictor based upon all 87 multiple associated SNPs (g_{87}) and that based on the 49 additional SNPs (g_{49}) in both the ARIC and QIMR cohorts. The predictors in one cohort are created based on SNP effects estimated from the approximate joint analysis using the other cohort as a reference sample.

analysis is more powerful than the single-SNP analysis in detecting such SNPs[7,8]. For example, rs1367226 at the *EFEMP1* locus on chromosome 2 (P value from the single-SNP meta-analysis (P_M) = 0.198) and rs6784185 at the *IGF2BP2* locus on chromosome 3 P_M = 0.088) did not even show nominally significant association in single-SNP analyses, but they both reached genome-wide significance when fitted jointly with the leading SNPs at these two loci (**Table 1**). At the same time, the significance and effect sizes of the leading SNPs at the two loci also increased where $P_M = 1.1 \times 10^{-28}$ versus the P value from the joint analysis (P_J) = 3.0×10^{-37} for rs3791675 and $P_M = 1.8 \times 10^{-14}$ versus $P_J = 1.5 \times 10^{-22}$ for rs720390 (**Table 1**). There were 11 loci harboring more than 2 associated SNPs, with a maximum number of 4, and the length of each locus varied substantially. For examples, three associated SNPs at a locus on chromosome 1 covered a genomic region of 1,134 kb, and the *ACAN* locus on chromosome 15, with three associated SNPs, only has a length of 32 kb (**Table 1**). Only considering the leading SNP(s) at each of the 36 loci (**Table 1**), there were 38 leading SNPs that, in total, explained 2.5% of phenotypic variance. However, taking all 87 jointly associated SNPs into account (38 leading and 49 additional), 4.1% of variance was explained, with the additional 49 SNPs explaining an additional 1.6% of the variance.

We extracted the GIANT summary statistics of the 247 associated SNPs and performed a joint analysis of these SNPs with their LD estimated in 3,924 unrelated Australians of British Isles ancestry[14] selected from a GWAS cohort at the Queensland Institute of Medical Research (QIMR) (Online Methods). The allele frequencies of the 247 SNPs estimated from either the ARIC or the QIMR cohort were consistent with those reported by the GIANT meta-analysis (**Supplementary Fig. 1**). The joint effects and their corresponding P values, obtained from the joint analysis using the ARIC cohort as a reference sample, showed good agreement with those obtained using the QIMR cohort as the reference sample (**Supplementary Fig. 2** and **Supplementary Table 2**), suggesting that the results are robust with respect to the choice of reference sample.

We created two predictors in the QIMR cohort by PLINK[15], one based on all 87 multiple associated SNPs and the other based on the 49 additionally associated SNPs only, with SNP effects estimated from the joint analysis using the ARIC cohort as a reference sample, and then regressed the observed height phenotypes on the predictors. We performed the same prediction analysis in the ARIC cohort but with SNP effects estimated from the joint analysis using the QIMR cohort as the reference sample, acknowledging that the ARIC cohort is part of the discovery sample of the GIANT meta-analysis. The regression slopes were not significantly different from 1 (**Table 2**), suggesting that the estimates of joint SNP effects are unbiased, the prediction R^2 of all 87 SNPs was ~3.8–4.8%, consistent with the estimate of 4.1% of variance explained in the discovery sample, and the prediction R^2 of the 49 additional associated SNPs was ~1.3–1.5%, in line with the estimate of 1.6% of variance explained by these SNPs in the discovery sample. Hence, by performing a prediction analysis in an independent sample, we confirmed that the 49 additional associated variants explain approximately 1.3% of phenotypic variation.

The GIANT Consortium performed a conditional meta-analysis for height[3] in a subset of stage 1 studies including 106,336 individuals and identified 19 secondary signals at 19 loci at $P < 3.3 \times 10^{-7}$. The GIANT conditional meta-analysis only reported one secondary signal at each of the 19 loci, because it was too time-consuming to conduct a single run of the conditional meta-analysis to take the process further. There are 16 loci associated with height that were reported by GIANT[3] with secondary SNPs at $P < 5 \times 10^{-8}$, all of which overlapped with our set of 36 loci with multiple associated SNPs. The concordance of these results provides a technical replication of the analysis methods.

Associated SNPs more than 1 Mb away can be in substantial LD

The GIANT meta-analysis identified 180 loci that were associated with height[3]. When we fitted all 180 hit SNPs simultaneously in a joint analysis, the majority of them seemed to be independently associated, because they had been deliberately selected to be at least 1 Mb away from each other to render them unlikely to be in strong LD. However, there was an exception. Two SNPs, rs1814175 and rs5017948, were reported as independently associated SNPs by the GIANT Consortium[3] with P values in the discovery set of 1.9×10^{-8} and 4.6×10^{-8}, respectively (**Table 3**). These SNPs are ~1.76 Mb distant but in substantial LD (r = 0.61 and 0.59 in the ARIC and QIMR cohorts, respectively), suggesting that, in some specific cases, the commonly used 1-Mb window is not big enough to guarantee that two SNPs are independently associated with a trait and that the stepwise conditional analysis is a more general approach to refine association signals and to identify additional associated SNPs. Given either of these SNPs in the model, the other SNP is not found to reach genome-wide significance with $P_J > 0.001$ (**Table 3**). In our conditional and joint analysis, only the rs1814175 SNP was selected, and no additional signals were detected in this region.

Case-control studies

Our method is applicable to case-control studies (Online Methods). We demonstrate this by using the summary-level statistics of the DIAGRAM meta-analysis for T2D from a discovery set of 8,130 affected individuals (cases) and 38,987 controls. In our example analysis, we focused only on the *CDKN2B* region, where there was some previous evidence of multiple signals[16]. We analyzed the DIAGRAM meta-analysis data, with allele frequencies and LD structure estimated from the ARIC cohort, and replicated the findings by a joint analysis

Table 3 Joint analysis of two GIANT hit SNPs on chromosome 11

			GIANT meta-analysis			Joint analysis with LD from ARIC				Joint analysis with LD from QIMR			
SNP	Location (bp)	Allele 1	Freq.	β	P	Freq.	b	P	r	Freq.	b	P	r
rs1814175	49515748	T	0.339	0.023	1.9×10^{-8}	0.330	0.015	4.0×10^{-3}	0.611	0.357	0.015	3.1×10^{-3}	0.591
rs5017948	51270794	A	0.186	0.027	4.6×10^{-8}	0.196	0.016	1.0×10^{-2}		0.194	0.016	7.7×10^{-3}	

Allele 1, reference allele; freq.: frequency of the reference allele; β, marginal effect; b, joint effect; r, LD correlation between a SNP and the next adjacent SNP at a locus.

© 2012 Nature America, Inc. All rights reserved.

Table 4 Joint analysis of two T2D SNPs at the *CDKN2B* locus on chromosome 9

			DIAGRAM meta-analysis		Joint analysis with LD from ARIC				Joint analysis with LD from QIMR			
SNP	Location (bp)	Allele 1	log(OR)	P	Freq.	log(OR)	P	r	Freq.	log(OR)	P	r
rs10965250	22,123,284	G	0.181	1.0×10^{-10}	0.827	0.273	1.3×10^{-18}	−0.530	0.828	0.299	9.6×10^{-21}	−0.590
rs10757282	22,123,984	C	0.097	3.1×10^{-4}	0.432	0.208	2.5×10^{-12}		0.373	0.235	1.8×10^{-14}	

Allele 1, reference allele; freq.: frequency of the reference allele; *r*, LD correlation between a SNP and the next adjacent SNP at a locus. log(OR), log odds ratio.

using the QIMR cohort as reference sample. Two SNPs, rs10965250 and rs10757282, which are only 700 bp apart, were retained in stepwise model selection as jointly associated SNPs with $P < 5 \times 10^{-12}$, using either the ARIC or QIMR cohort as a reference sample for LD (**Table 4**), consistent with the result from a previous analysis that the two SNPs define a haplotype association[16]. The risk alleles of these two SNPs are negatively correlated (r = −0.53 and −0.59 in the ARIC and QIMR cohorts, respectively); therefore, the secondary SNP (rs10757282; $P_M = 3.1 \times 10^{-4}$) was masked by the primary SNP (rs10965250; $P_M = 1.0 \times 10^{-10}$) in single-SNP analyses. When they were fitted jointly, their effects, as well as statistical significance, were substantially increased compared to what was obtained in single-SNP analyses.

DISCUSSION

We have presented a method of approximate conditional and joint genome-wide association analysis that is powerful, versatile and computationally fast. The method does not require any additional genotyping or phenotyping and does not rely on individual-level genotype and phenotype data, except for a reference population with individual genotypes—either from one of the participating studies of the meta-analysis or from genotype data in the public domain—that is required for LD estimation. The effect sizes of most SNPs that are associated with complex traits are very small, such that there is a great benefit in using estimates from a large-scale meta-analysis, and a reasonably large reference sample is sufficient to estimate LD between SNPs located near to one another. We believe that this method is useful to refine independent GWAS associations and to identify additional associated variants in large-scale meta-analysis where the pooled individual-level genotype data are unavailable for analysis.

The method is built upon the assumption that the reference sample is from the same population as the samples from which the genotype-phenotype associations are estimated, meaning that the LD correlations that are estimated from the reference sample are unbiased. We show by simulation (**Supplementary Note**) that the P values from our approximate approach are highly consistent with those from the conditional meta-analysis (correlation of >0.99 in 1,000 simulation replicates), given a reference sample of 6,654 individuals (**Supplementary Fig. 3** and **Supplementary Table 4**). The simulation results did not change if the reference sample was independent of the discovery sample, as long as both were sampled from the same general population (**Supplementary Fig. 4**). We recommend that a reference sample be chosen with a large sample size, so that the LD correlations are estimated with little error[17]. The simulation results indicate that a reference sample with a size of at least 2,000 is required and that little additional accuracy is gained beyond a sample size of 5,000 (**Supplementary Fig. 4**). The reference sample needs to be checked for cryptic relatedness and population stratification, which could cause correlations between SNPs that do not exist in the discovery set. In the present study, we included only the individuals of European descent in the ARIC cohort[18] and of British Isles descent in the QIMR cohort[14] and removed one of each pair of individuals with a SNP-derived relatedness estimate of >0.025 in both cohorts (Online Methods). If the expected value of the LD correlation between two SNPs is zero in the general population, the sampling variance of an observed LD correlation in a sample is proportional to the sample size (m), with var[r | E(r) = 0] = 1 / m. Thus, given a random sample of 6,654 unrelated individuals from the population, the probability of observing a LD correlation greater than 0.1 or smaller than −0.1 ($r^2 > 0.01$) is 3.4×10^{-16}. In order to investigate possible false positives resulting from errors in LD estimation, we first performed the analysis using the ARIC cohort as a reference sample, and we then performed a joint analysis of the selected SNPs using the QIMR cohort as a reference sample (**Supplementary Tables 2** and **3**). We show that the LD correlations between adjacent pairs of the 247 height-associated SNPs are in very good agreement across the ARIC and QIMR cohorts (**Supplementary Fig. 5**). Therefore, our main results were unlikely to be driven by errors in estimating the LD structure in the reference sample. This is consistent with our technical replication of all secondary signals reported by the GIANT Consortium from conditional analysis using individual-level genotype and phenotype data[3]. We reported most results based on the ARIC cohort (for example, **Table 1**) because it has a larger sample size and is more genetically similar to the whole GIANT meta-analysis sample relative to the QIMR cohort, where ancestry is restricted to the British Isles[14]. We could also only consider results that were consistent using two independent cohorts as reference samples. However, this might be too conservative, as some real associations identified using one cohort with P values that just passed the arbitrary cutoff value of 5×10^{-8}, might be eliminated from the analysis in another cohort due to random errors in the estimates of LD correlations. Our method is not limited to meta-analysis summary data but can also be applied to a single GWAS cohort with individual-level genotypes, in which case, the whole discovery sample is used as the reference sample, and our method then becomes equivalent to a multiple regression analysis (Online Methods and **Supplementary Fig. 6**). In this case, the automated stepwise selection procedure implemented in our software tool[19], which has, to our knowledge, not been implemented in any other GWAS analysis tools in the public domain, would still be useful for data applications.

As with any fixed-effect model selection strategy, such as stepwise linear multiple regression analysis, there is a risk of over-fitting effects. This can be a particular problem for the analysis of GWAS SNP data because the number of SNPs is typically much larger than the experimental sample size. The effects of selected SNPs tend to be overestimated (sometimes called the winner's curse) and, if the threshold for inclusion is less stringent, false positives could be included in the model. In both cases, the estimated residual variance will be too low. This can, in theory, be a runaway process, because the more SNPs that are selected in the model, the lower the apparent residual variance and the greater the number of remaining SNPs that will become significant and will be added to the model. In the general population, the expected value of the LD correlation between SNPs on different chromosomes or more than d Mb distant is zero, even though, in a particular sample, the observed value is nonzero due to finite sample size. In our method, we set the LD correlation between distant SNPs

© 2012 Nature America, Inc. All rights reserved.

to zero, because it is inappropriate to represent a randomly sampled correlation in the discovery sample by another randomly sampled correlation in the reference sample. In the conditional analysis, if a SNP to be tested is more than d Mb distant from all the top SNPs fitted in the model, we are therefore unable to model and adjust for the variability in the estimate of the SNP effect due to the sampling variation of correlations in the discovery sample. Thus, the conditional effect will be the same as the marginal effect, whereas the standard error of this SNP effect decreases as the residual variance is reduced because of the selected SNPs in the model. This signifies that test statistics will be inflated and the false positive rate will increase. This problem will be dramatically exacerbated if the discovery set is not very large, for example, coming from a single GWAS cohort, with there being a higher chance of observing a substantial correlation due to random sampling and, further, if the selected SNPs fitted in the model explain a large proportion of variance. In our method implementation, we keep the residual variance constant at the same level of the phenotypic variance, even after fitting SNPs that cumulatively explain a substantial proportion of phenotypic variation in the model (Online Methods). Although this approach is conservative, because we know that fitting the 180 known height-associated SNPs in the model reduces the residual variance by ~10% and therefore increases power to detect additional variants, it has the benefit of keeping the type-I error rate at the same level as that in the meta-analysis and thus avoids over-fitting.

To demonstrate the conservative nature of our model selection strategy, we performed selection of SNPs with a less stringent P value threshold. We constrained the analysis to region 1 Mb up- or downstream of the 180 known height-associated SNPs and chose a P value threshold of 5×10^{-7}, as only ~13% of the genome is covered by these 2-Mb regions. We identified 85 additional associated SNPs at 60 loci, which explained 2.4% of variance in the discovery set. We validated the joint effects of these 85 SNPs by our prediction analyses. The prediction R^2 were 2.4% ($P = 8.6 \times 10^{-37}$) and 1.9% ($P = 5.3 \times 10^{-18}$) in the ARIC and QIMR cohorts, respectively, suggesting that we could detect more associated variants with a less stringent threshold but, of course, might increase the risk of including false positives. Nevertheless, this analysis confirms more heritability can be explained at a substantial proportion of loci that affect the trait. It also suggests a model of genetic architecture of a large number of loci and multiple causal variants at many of these regions.

In the GIANT meta-analysis for height and BMI[3,4], the summary statistics were adjusted by the genomic control method[20] in each of the participating studies, and the test statistics were adjusted by the genomic control method for a second time in the combined analysis of all studies, which is sometimes called 'double-GC' correction. In the present study, we did not perform the second genomic control correction (although we have provided an option to implement this in our software tool) for two reasons. First, the purpose of genomic control correction is to adjust for the effect of population stratification, but, in the absence of population stratification and presence of polygenic inheritance, genomic inflation is expected[21]; therefore, double-GC correction might be too conservative and overkill. Second, if the genomic inflation is due to stratification, there is no reason to find additional associated SNPs at known loci, whereas under the hypothesis that the genomic inflation is consistent with polygenic inheritance and that there are multiple variants at the same loci segregating in the population, we would expect to see what we found empirically. The GIANT conditional meta-analysis detected 16 loci with additional associated SNPs, and we identified 20 more such loci, which is partly because the GIANT conditional meta-analysis used only a subset of the discovery sample (106,336 out of 133,653 individuals) due to the difficulty of managing the large number of participating studies in a fixed time period and partly because the GIANT conditional meta-analysis implemented a double-GC correction that substantially reduced the power of detection.

The results for height and BMI seem to be very different. For height, we identified 36 loci with multiple associated SNPs, whereas, for BMI, we did not find any such loci. It seems unlikely that this large difference can be entirely explained by the greater power to detect associations with height compared to BMI because of the greater heritability of height. The narrow-sense heritability for height is estimated to be ~80% by pedigree analyses[22], and the heritability for BMI is ~40–60% (refs. 23,24). If we assume the heritability of BMI is 50%, then 4% (2%/50%) of narrow-sense heritability for BMI has been explained by GWAS[4], a much lower proportion than that for height, which is approximately 12.5% (10%/80%)[3]. When considering all the SNPs simultaneously, 32% (16%/50%) of narrow-sense heritability for BMI can be captured by all common SNPs using the whole-genome estimation approach we recently developed, which is also lower than the corresponding explained heritability for height (~56%)[14,18]. In a previous analysis partitioning genetic variance onto individual chromosomes, the variance explained by each chromosome showed a strong linear relationship with chromosome length for height, but such a relationship was rather weak for BMI[18]. To investigate whether additional variants for BMI could be detected, we performed a conditional and joint analysis with a less stringent P value threshold of 5×10^{-6}, with the LD structure estimated from the ARIC cohort. We identified 19 multiple associated SNPs (9 leading and 10 additional SNPs) at 9 loci (**Supplementary Tables 1** and **5**), which is still much lower than the number of additional variants detected for height. The ten additional SNPs explained 0.21% of the variance in the discovery set. When using these SNPs to predict the BMI phenotypes in the QIMR cohort, the prediction R^2 was 0.13%, which is nominally significant ($P = 0.037$). Taken together, the previous and current results are consistent and suggest that the genetic architectures for height and BMI might be different in terms of the allelic spectrum of causal variants within and between loci, the distribution of effect sizes and the robustness of effect sizes to environments and gene modifiers.

We identified 36 loci with multiple associations for height. We have shown by examples of pairs of multiple associated SNPs that marginal SNP effects will be underestimated (overestimated) if their trait increasing alleles are negatively (positively) correlated, consistent with the findings from a GWAS of gene expressions[25]. However, this is not necessarily always the case when there are more than two associated SNPs in LD with each other, and the generality of these results depends on the actual LD correlations of all segregating causal variants at a locus. If one of the associated SNPs at each locus is causative, then there must be multiple causal variants in that region, because the joint effects have already taken the LD into account, such that, conditional on the causal variant in the model, the effects of any of its proxies would not be statistically significant. However, it is unlikely that the SNPs themselves are causative, because the ~2.5 million SNPs in the HapMap 2 panel of Utah residents of Northern and Western European ancestry (CEU)[26] represent only a fraction of all the polymorphisms segregating in the human population[27]. Assuming that multiple associated SNPs at a particular locus are not causative, it is unlikely that they are in LD with a single rare causal variant, especially for SNPs with minor allele frequency (MAF > 0.1), and it is also implausible that they are in LD with a single common causal variant (**Supplementary Note**). Therefore, it seems likely that there are multiple causal variants segregating at the same locus; however,

© 2012 Nature America, Inc. All rights reserved.

this inference is indirect and inconclusive. With whole-genome sequence data, the conditional and joint analysis approach we present here will be helpful in identifying causal variants.

URLs. GCTA, http://gump.qimr.edu.au/gcta/massoc.html.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturegenetics/.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

The authors thank the staff and participants of the ARIC study for their important contributions. We thank all referees for many constructive comments and suggestions. We acknowledge funding from the Australian National Health and Medical Research Council (NHMRC; 241944, 389875, 389891, 389892, 389938, 442915, 442981, 496739, 496688, 552485, 613672, 613601 and 1011506), the US National Institutes of Health (AA07535, AA10248, AA014041, AA13320, AA13321, AA13326 and DA12854) and the Australian Research Council (ARC; DP0770096 and DP1093502). A.P.M. acknowledges financial support from the Wellcome Trust (WT081682/Z/06/Z). M.I.M. acknowledges financial support from the Wellcome Trust (WT083270). The Atherosclerosis Risk in Communities Study is carried out as a collaborative study supported by contracts with the National Heart, Lung, and Blood Institute (HHSN268201100005C, HHSN268201100006C, HHSN268201100007C, HHSN268201100008C, HHSN268201100009C, HHSN268201100010C, HHSN268201100011C and HHSN268201100012C), the XXX (R01HL087641, R01HL59367 and R01HL086694), the US National Human Genome Research Institute (U01HG004402) and the US National Institutes of Health (HHSN268200625226C). Infrastructure was partly supported by a component of the US National Institutes of Health and NIH Roadmap for Medical Research (UL1RR025005).

AUTHOR CONTRIBUTIONS

P.M.V. and M.E.G. designed the study. M.E.G., J.Y. and P.M.V. derived the analytical results. J.Y. performed all statistical analyses. J.Y. and P.M.V. wrote the first draft of the paper. M.N.W., R.J.L., T.M.F., M.I.M. and J.N.H. contributed the summary data of the height and BMI meta-analyses on behalf of the GIANT Consortium and provided comments that improved earlier versions of the manuscript. T.F., A.P.M. and M.I.M. contributed the summary data of the T2D meta-analysis on behalf of the DIAGRAM Consortium. S.E.M., P.A.F.M., A.C.H., N.G.M. and G.W.M. contributed the individual-level and imputed genotypes and phenotype data of the QIMR cohort.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturegenetics/.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

© 2012 Nature America, Inc. All rights reserved.

1. Hindorff, L.A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. USA* **106**, 9362–9367 (2009).
2. Manolio, T.A. Genomewide association studies and assessment of the risk of disease. *N. Engl. J. Med.* **363**, 166–176 (2010).
3. Lango Allen, H. *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832–838 (2010).
4. Speliotes, E.K. *et al.* Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat. Genet.* **42**, 937–948 (2010).
5. Voight, B.F. *et al.* Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat. Genet.* **42**, 579–589 (2010).
6. Turnbull, C. *et al.* Genome-wide association study identifies five new breast cancer susceptibility loci. *Nat. Genet.* **42**, 504–507 (2010).
7. Galarneau, G. *et al.* Fine-mapping at three loci known to affect fetal hemoglobin levels explains additional genetic variation. *Nat. Genet.* **42**, 1049–1051 (2010).
8. Sanna, S. *et al.* Fine mapping of five loci associated with kow-density lipoprotein cholesterol detects variants that double the explained heritability. *PLoS Genet.* **7**, e1002198 (2011).
9. Ripke, S. *et al.* Genome-wide association study identifies five new schizophrenia loci. *Nat. Genet.* **43**, 969–976 (2011).
10. Sklar, P. *et al.* Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near *ODZ4*. *Nat. Genet.* **43**, 977–983 (2011).
11. Trynka, G. *et al.* Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nat. Genet.* **43**, 1193–1201 (2011).
12. Sonesson, A.K., Meuwissen, T.H. & Goddard, M.E. The use of communal rearing of families and DNA pooling in aquaculture genomic selection schemes. *Genet. Sel. Evol.* **42**, 41 (2010).
13. Weir, B.S. *Genetic Data Analysis.* (Sinauer Associates, Sunderland, Massachusetts, 1990).
14. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).
15. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
16. Shea, J. *et al.* Comparing strategies to fine-map the association of common SNPs at chromosome 9p21 with type 2 diabetes and myocardial infarction. *Nat. Genet.* **43**, 801–805 (2011).
17. Pardo, L. *et al.* Global similarity with local differences in linkage disequilibrium between the Dutch and HapMap-CEU populations. *Eur. J. Hum. Genet.* **17**, 802–810 (2009).
18. Yang, J. *et al.* Genome partitioning of genetic variation for complex traits using common SNPs. *Nat. Genet.* **43**, 519–525 (2011).
19. Yang, J., Lee, S.H., Goddard, M.E. & Visscher, P.M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
20. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997–1004 (1999).
21. Yang, J. *et al.* Genomic inflation factors under polygenic inheritance. *Eur. J. Hum. Genet.* **19**, 807–812 (2011).
22. Visscher, P.M., Hill, W.G. & Wray, N.R. Heritability in the genomics era—concepts and misconceptions. *Nat. Rev. Genet.* **9**, 255–266 (2008).
23. Magnusson, P.K. & Rasmussen, F. Familial resemblance of body mass index and familial risk of high and low body mass index. A study of young men in Sweden. *Int. J. Obes. Relat. Metab. Disord.* **26**, 1225–1231 (2002).
24. Schousboe, K. *et al.* Sex differences in heritability of BMI: a comparative study of results from twin studies in eight countries. *Twin Res.* **6**, 409–421 (2003).
25. Wood, A.R. *et al.* Allelic heterogeneity and more detailed analyses of known loci explain additional phenotypic variation and reveal complex patterns of association. *Hum. Mol. Genet.* **20**, 4082–4092 (2011).
26. The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).
27. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).

© 2012 Nature America, Inc. All rights reserved.

ONLINE METHODS

Summary statistics of meta-analysis and individual-level genotype data. The GIANT Consortium performed a meta-analysis of GWAS data in a discovery set with 133,653 and 123,865 individuals of recent European ancestry from 46 studies for height[3] and BMI[4], respectively. In each of the participating studies, genotype data were imputed to ~2.8 million SNPs present in the HapMap Phase 2 European-American reference panel[26], and the standard errors of all SNPs were adjusted by the genomic control method[20]. We calculated the effective sample size for each SNP and excluded SNPs with effective sample sizes of >2 s.d. from the mean. We also excluded SNPs with MAF of <0.01, retaining ~2.5 million SNPs for both height and BMI.

We also obtained access to the individual-level genotype and phenotype data of the ARIC cohort, a population-based study of Americans[28], and the QIMR cohort, a twin study of Australians[29]. The ARIC samples were genotyped by Affymetrix 6.0 SNP array, and the QIMR samples were genotyped by Illumina 610K or 370K array. After quality control filtering of SNPs, 593,521 and 274,604 genotyped SNPs were retained in the ARIC (excluding SNPs with missingness of >2%, MAF of <0.01 or Hardy-Weinberg equilibrium (HWE) P value of $<1 \times 10^{-3}$) and QIMR cohorts (excluding SNPs with missingness of >5%, MAF of <0.01 and HWE P value of $<1 \times 10^{-6}$), respectively. After sample quality control analysis, 8,682 and 11,742 individuals of European ancestry in the ARIC and QIMR cohorts, respectively, were included for further analysis. The quality control protocol has been detailed previously for the ARIC cohort[18,30] and for the QIMR cohort[14,29]. We then estimated pairwise genetic relationships between individuals[14] and removed one of each pair of individuals with an estimated relatedness of >0.025. After these quality control steps, 6,654 and 3,924 unrelated individuals were retained in the ARIC and QIMR cohorts, respectively. All the ARIC samples were from adults and the QIMR samples were from 3,247 adults and 677 16-year-old adolescents. The SNP data for both ARIC and QIMR cohorts were imputed to the HapMap Phase 2 CEU panel by MACH[31]. We used the best guess genotypes of the imputed SNPs and excluded imputed SNPs with HWE P value of $<1 \times 10^{-6}$, imputation R^2 of <0.3 or MAF of <0.01 and retained 2,406,652 and 2,410,957 SNPs in the ARIC and QIMR cohorts, respectively. The ARIC cohort is part of the discovery sample of the GIANT meta-analysis, whereas the QIMR cohort is not. In the prediction analyses, the height and BMI phenotypes in the ARIC and QIMR cohorts were adjusted for age and sex effects and standardized to z scores[14,18]. In the QIMR cohort, only samples from 3,247 adults were used in the prediction analysis for BMI.

Estimating the joint effects of multiple SNPs for a quantitative trait. Under the assumption that a quantitative trait is affected by multiple genetic variants, we can express phenotypes in a sample of unrelated individuals by a multi-SNP model as

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e} \tag{1}$$

where $\mathbf{y} = \{y_i\}$ is an $n \times 1$ vector of phenotypes, with n being the sample size, $\mathbf{X} = \{x_{ij}\}$ is an $n \times N$ genotype matrix, with $x_{ij} = -2p_j$, $1 - 2p_j$ or $2 - 2p_j$ for the jth SNP of the ith individual, with p_j being the allele frequency of a SNP j and N being the number of SNPs fitted in the model, and $\mathbf{b} = \{b_j\}$, an $N \times 1$ vector of joint SNP effects. For simplicity, we subtract the mean of the phenotype from y_i, such that we do not need to fit the intercept in the model. We therefore can estimate the joint effects of multiple SNPs by the least-squares approach as

$$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \text{ and } \operatorname{var}(\hat{\mathbf{b}}) = \sigma_J^2(\mathbf{X}'\mathbf{X})^{-1} \tag{2}$$

where σ_J^2 is the residual variance in the joint analysis.

In a GWAS or meta-analysis, however, each SNP is usually tested for association separately based on a single-SNP model

$$\mathbf{y} = \mathbf{x}_j\beta_j + \mathbf{e} \tag{3}$$

where $\mathbf{x}_j$ is the jth column of $\mathbf{X}$ and β_j is the marginal effect of SNP j. The marginal effects of multiple SNPs estimated from a single SNP–based genome scan can be written in matrix form as

$$\hat{\boldsymbol{\beta}} = \mathbf{D}^{-1}\mathbf{X}'\mathbf{y} \text{ and } \operatorname{var}(\hat{\boldsymbol{\beta}}) = \sigma_M^2\mathbf{D}^{-1} \tag{4}$$

where $\beta = \{\beta_j\}$ is an $N \times 1$ vector of marginal SNP effects, $\mathbf{D} = \{D_j\}$ is the diagonal matrix of $\mathbf{X}'\mathbf{X}$ with $D_j = \sum_i^n x_{ij}^2$ and σ_M^2 is the residual variance in the single-SNP analyses. The marginal SNP effects do not take the LD correlations between SNPs into account compared with the joint SNP effects. There are two issues involved in such single-SNP analyses for SNPs a short distance from each other: (i) if the increasing (or risk) alleles of two SNPs is negatively correlated, the effects of both SNPs will be attenuated; therefore, the single-SNP analysis is underpowered, and one SNP or both SNPs may be undetected and (ii) if both SNPs reach genome-wide significance, it is difficult to determine their degree of dependency by interrogating the LD afterwards.

With the summary statistics from single-SNP analyses and individual-level genotype data of the discovery sample, we can convert the marginal effects to joint effects without using the phenotype data. We know from equation (4) that $\mathbf{X}'\mathbf{y} = \mathbf{D}\hat{\boldsymbol{\beta}}$, and we therefore can rewrite equation (2) with respect to $\hat{\boldsymbol{\beta}}$

$$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{D}\hat{\boldsymbol{\beta}} \text{ and } \operatorname{var}(\hat{\mathbf{b}}) = \sigma_J^2(\mathbf{X}'\mathbf{X})^{-1} \tag{5}$$

The proportion of phenotypic variance explained by all the SNPs (coefficient of determination of a multiple regression model) is

$$R_J^2 = \frac{\hat{\mathbf{b}}'\mathbf{X}'\mathbf{y}}{\mathbf{y}'\mathbf{y}} = \frac{\hat{\mathbf{b}}'\mathbf{D}\hat{\boldsymbol{\beta}}}{\mathbf{y}'\mathbf{y}} \tag{6}$$

giving the following equation:

$$\hat{\sigma}_J^2 = \frac{(1 - R_J^2)\mathbf{y}'\mathbf{y}}{n - N} = \frac{\mathbf{y}'\mathbf{y} - \hat{\mathbf{b}}'\mathbf{D}\hat{\boldsymbol{\beta}}}{n - N} \tag{7}$$

In an association analysis of a single SNP j,

$$\hat{\sigma}_{M(j)}^2 = \frac{\mathbf{y}'\mathbf{y} - D_j\hat{\beta}_j^2}{n - 1} \tag{8}$$

and the squared standard error of the estimate of the effect size is $S_j^2 = \hat{\sigma}_{M(j)}^2 / D_j$ so that $\mathbf{y}'\mathbf{y} = D_jS_j^2(n - 1) + D_j\hat{\beta}_j^2$. Although the phenotypes of a quantitative trait are often standardized to z scores, we take the median of $D_jS_j^2(n - 1) + D_j\hat{\beta}_j^2$ across all the SNPs to calculate $\mathbf{y}'\mathbf{y}$ instead of relying on the variance being known.

For a meta-analysis of a large number of cohorts, such as the GIANT Consortium meta-analysis[3,4], we are usually unable to obtain pooled individual-level genotype data of the whole discovery set; hence, we do not have the $\mathbf{X}'\mathbf{X}$ matrix. $\mathbf{X}'\mathbf{X}$ is essentially a variance-covariance matrix of SNP genotypes, which can be estimated from the allele frequencies in the meta-analysis sample along with LD correlations between SNPs from a reference sample, such as one of the meta-analysis cohorts for which individual-level genotype data are available. We let $\mathbf{W} = \{w_{ij}\}$ denote the genotype matrix of the reference sample with sample size of m, where $w_{ij} = -2f_j$, $1 - 2f_j$ or $2 - 2f_j$ for the three genotypes, with f_j being the allele frequency of a SNP j in the reference sample, and we let $\mathbf{D}_W$ denote the diagonal matrix of $\mathbf{W}'\mathbf{W}$ with $D_{W(j)} = \sum_i^m w_{ij}^2$. If the reference sample is from the same population as the meta-analysis sample, the LD correlation between a pair of SNPs j and k should be similar in the two samples[32,33], with

$$\frac{\sum_i^n x_{ij}x_{ik}}{\sqrt{\sum_i^n x_{ij}^2 \sum_i^n x_{ik}^2}} \approx \frac{\sum_i^m w_{ij}w_{ik}}{\sqrt{\sum_i^m w_{ij}^2 \sum_i^m w_{ik}^2}} \tag{9}$$

so that $\mathbf{X}'\mathbf{X}$ is approximately equal to $\mathbf{B}$, with the jkth element of B being

$$B_{jk} \approx \sqrt{\frac{D_jD_k}{D_{W(j)}D_{W(k)}}} \sum_i^m w_{ij}w_{ik} \tag{10}$$

We have defined above that $D_j = \sum_i^n x_{ij}^2$ and, as x_{ij} is not available in this case, we thus take $D_j = 2p_j\,(1 - p_j)n$, assuming HWE and can show this in matrix form:

$$\mathbf{B} = \mathbf{D}^{1/2}\mathbf{D}_W^{-1/2}\mathbf{W}'\mathbf{W}\mathbf{D}_W^{-1/2}\mathbf{D}^{1/2} \tag{11}$$

Therefore, we can approximate a joint analysis of multiple SNPs as

$$\tilde{\mathbf{b}} = \mathbf{B}^{-1}\mathbf{D}\hat{\boldsymbol{\beta}} \text{ and } \mathrm{var}(\tilde{\mathbf{b}}) = \sigma_J^2\mathbf{B}^{-1} \qquad (12)$$

where $\tilde{\mathbf{b}} = \{\tilde{b}_j\}$ is an $N \times 1$ vector of approximate estimates of joint SNP effects. If a SNP is uncorrelated with all other SNPs in the model, then the estimate of the effect size from the joint analysis will be identical to that from the meta-analysis. In a genetically homogenous population of large effective size, the expected value of LD correlation between two SNPs on different chromosomes or a large distance apart is approximately zero, and the observed LD correlations between such pairs of SNPs in a sample are just a result of random sampling. We show with empirical data that the observed LD correlation between SNPs more than 10 Mb apart is consistent with what we would expect by chance (**Supplementary Fig. 7**). We use the expected values (zeros) in the matrix **B** for such pairs of distant SNPs, because it is more appropriate to represent a sampled correlation observed in the meta-analysis sample by its expected value rather than another sampled correlation observed in the reference sample unless the whole meta-analysis sample is used as the reference sample.

In addition, the sample size varies for different SNPs due to imputation failures for different SNPs in different participating studies. Therefore, n is no longer constant across different SNPs, and we need to rescale the elements of **B** and **D** according to the different effective sample sizes of different SNPs. For any SNP j,

$$\hat{n}_j = \mathbf{y}'\mathbf{y} / D_j S_j^2 - \hat{\beta}_j^2 / S_j^2 + 1 \qquad (13)$$

where we take the variance explained by a single SNP into account, considering that the effect sizes of some particular SNPs are large for some traits. We use the estimated effective sample size rather than the reported sample size, because the effective sample size will be smaller than the reported sample size if there is some degree of relatedness in the data. We then adjust the jkth element of **B** for the sample size variability of the SNPs as

$$B_{jk} = \min(\hat{n}_j, \hat{n}_k) 2 \sqrt{\frac{p_j(1-p_j)p_k(1-p_k)}{\sum_i^m w_{ij}^2 \sum_i^m w_{ik}^2}} \sum_i^m w_{ij}w_{ik} \qquad (14)$$

and adjust the jth diagonal element of **D** as $D_j = 2p_j(1-p_j)\hat{n}_j$.

Conditional analysis. In a linear regression analysis of multiple SNPs, the least-squares estimates of the joint effects of one set of SNPs conditional on another set of SNPs ($\mathbf{b}_2 \mid \mathbf{b}_1$) are

$$\hat{\mathbf{b}}_2 \mid \hat{\mathbf{b}}_1 = (\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{y} - (\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y} \qquad (15)$$

$$\mathrm{var}(\hat{\mathbf{b}}_2 \mid \hat{\mathbf{b}}_1) = \sigma_C^2(\mathbf{X}_2'\mathbf{X}_2)^{-1} - \sigma_C^2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1} \qquad (16)$$

where σ_C^2 is the residual variance in the conditional analysis and all the other variables and parameters are defined as above, with the subscripts 1 and 2 indicating the two SNP sets. We can perform a multi-SNP conditional analysis using summary data from single-SNP analyses and individual-level genotype data of the sample without accessing the phenotype data by

$$\hat{\mathbf{b}}_2 \mid \hat{\mathbf{b}}_1 = (\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{D}_2\hat{\boldsymbol{\beta}}_2 - (\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{D}_1\hat{\boldsymbol{\beta}}_1 \qquad (17)$$

$$\hat{\sigma}_C^2 = \frac{\mathbf{y}'\mathbf{y} - \hat{\mathbf{b}}_1'\mathbf{D}_1\hat{\beta}_1 - (\hat{\mathbf{b}}_2 \mid \hat{\mathbf{b}}_1)'\mathbf{D}_2\hat{\beta}_2}{(n - N_1 - N_2)} \qquad (18)$$

where N_1 and N_2 are the number of SNPs in the two sets. If there is only one SNP to be tested in the conditional analysis ($N_2 = 1$), then equations (17) and (18) simplify to

$$\hat{b}_2 \mid \hat{\mathbf{b}}_1 = \hat{\beta}_2 - (\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{D}_1\hat{\boldsymbol{\beta}}_1 \qquad (19)$$

$$\mathrm{var}(\hat{b}_2 \mid \hat{\mathbf{b}}_1) = \sigma_C^2[D_2 - \mathbf{X}_2'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2] / D_2^2 \qquad (20)$$

where $\hat{b}_2$, $\hat{\beta}_2$ and D_2 are scalars.

As in a joint analysis, if the individual-level genotype data of the discovery sample are unavailable, we can estimate the LD correlations from the reference sample and approximate a conditional analysis as

$$\tilde{\mathbf{b}}_2 \mid \tilde{\mathbf{b}}_1 = \mathbf{B}_2^{-1}\mathbf{D}_2\hat{\boldsymbol{\beta}}_2 - \mathbf{B}_2^{-1}\mathbf{C}\mathbf{B}_1^{-1}\mathbf{D}_1\hat{\boldsymbol{\beta}}_1 \qquad (21)$$

$$\mathrm{var}(\tilde{\mathbf{b}}_2 \mid \tilde{\mathbf{b}}_1) = \sigma_C^2\mathbf{B}_2^{-1} - \sigma_C^2\mathbf{B}_2^{-1}\mathbf{C}\mathbf{B}_1^{-1}\mathbf{C}'\mathbf{B}_2^{-1} \qquad (22)$$

where $\mathbf{B}_1$ and $\mathbf{B}_2$ are similar as in equation (11) and $\mathbf{C} \approx \mathbf{X}_2'\mathbf{X}_1$ with the jkth element.

$$C_{jk} = \min(\hat{n}_{2j}, \hat{n}_{1k}) 2 \sqrt{\frac{p_{2j}(1-p_{2j})p_{1k}(1-p_{1k})}{\sum_i^m w_{2ij}^2 \sum_i^m w_{1ik}^2}} \sum_i^m w_{2ij}w_{1ik} \qquad (23)$$

Model selection. There are many ways of performing model selection in a multiple regression framework. We use the following stepwise selection strategy to select the associated SNPs iteratively over all the SNPs across the whole genome, regardless of their P values from the meta-analysis, except for the most significant SNP, which was used for model initiation.

(1) Start with a model with the most significant SNP in the single-SNP meta-analysis across the whole genome with P value below a cutoff P value, such as 5×10^{-8}.
(2) For the tth step, calculate the P values of all the remaining SNPs conditional on the SNP(s) that have already been selected in the model. To avoid problems due to colinearity, if the squared multiple correlation between a SNP to be tested and the selected SNP(s) is larger than a cutoff value, such as 0.9, the conditional P value for that SNP will be set to 1.
(3) Select the SNP with minimum conditional P value that is lower than the cutoff P value. However, if adding the new SNP causes new colinearity problems between any of the selected SNPs and the others, we drop the new SNP and repeat this process.
(4) Fit all the selected SNPs jointly in a model and drop the SNP with the largest P value that is greater than the cutoff P value.
(5) Repeat processes (2), (3) and (4) until no SNPs can be added or removed from the model.

A multiple regression analysis with model selection, such as that presented above, might suffer from over-fitting of effects, because the residual variance decreases as more and more SNPs are included in the model, such that the false positive rate for the inclusion of new SNPs in the model would be inflated. In practice, we keep the residual variance constant to the phenotypic variance, even if we added significant SNPs into the model, which may be too conservative and may therefore result in a loss of power to detect additional associated variants but has the benefit of keeping the false positive rate in the conditional and joint analysis at the same level as in the meta-analysis. If a SNP has no correlation with any of the SNPs selected in the model, its P value in the conditional or joint analysis will remain the same as it is in the meta-analysis.

Case-control studies. We know from the methods above that the scale of measurement of a quantitative trait is not important, as it can be dropped from the equations. We therefore extend these methods to be applied to the case-control study design, assuming that the disease liabilities (**L**) of all the individuals are known, and model the effects of multiple SNPs.

$$\mathbf{L} = \mathbf{X}\mathbf{b} + \mathbf{e} \qquad (24)$$

There are two distributions that are often assigned to the residuals to transform the underlying liability to the probability of being affected or unaffected, the standard normal distribution (probit model) and the logistic distribution (logistic regression). Given the logistic probability function of $f(l_i) = \exp(l_i)/[1 + \exp(l_i)]$ with l_i being the liability of the ith individual, the odds ratio (OR)

© 2012 Nature America, Inc. All rights reserved.

for a SNP j in a multiple-SNP analysis is $\exp(b_j)$, with b_j being the log(OR) in a joint analysis, and is $\exp(\beta_j)$ for a single-SNP model $\mathbf{L} = \mathbf{x}_j\beta_j + \mathbf{e}$, with β_j being the log(OR) in a single-SNP analysis. Even though the residuals follow a logistic distribution, the least-squares estimates of effect sizes are unbiased, because the least-squares approach does not rely on the assumption of normality. Hence, we can apply the same methods as described above for a quantitative trait to a case-control study, as long as the effect sizes and standard errors are expressed on the log(OR) scale.

Software tool. The method described above has been implemented as an option in the GCTA software package (see URLs)[19].

28. Rimm, E.B. *et al.* Prospective study of alcohol consumption and risk of coronary disease in men. *Lancet* **338**, 464–468 (1991).
29. Medland, S.E. *et al.* Common variants in the Trichohyalin gene are associated with straight hair in Europeans. *Am. J. Hum. Genet.* **85**, 750–755 (2009).
30. Laurie, C.C. *et al.* Quality control and quality assurance in genotypic data for genome-wide association studies. *Genet. Epidemiol.* **34**, 591–602 (2010).
31. Li, Y., Willer, C.J., Ding, J., Scheet, P. & Abecasis, G.R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.* **34**, 816–834 (2010).
32. Ke, X. *et al.* Efficiency and consistency of haplotype tagging of dense SNP maps in multiple samples. *Hum. Mol. Genet.* **13**, 2557–2565 (2004).
33. Teo, Y.Y. *et al.* Genome-wide comparisons of variation in linkage disequilibrium. *Genome Res.* **19**, 1849–1860 (2009).

© 2012 Nature America, Inc. All rights reserved.