Human Molecular Genetics, 2011, Vol. 20, No. 20 ***4082–4092***
doi:10.1093/hmg/ddr328
Advance Access published on July 28, 2011

Allelic heterogeneity and more detailed analyses of known loci explain additional phenotypic variation and reveal complex patterns of association

Andrew R. Wood[1], Dena G. Hernandez[2,3], Michael A. Nalls[2], Hanieh Yaghootkar[1], J. Raphael Gibbs[2,3], Lorna W. Harries[4], Sean Chong[2], Matthew Moore[2], Michael N. Weedon[1], Jack M. Guralnik[5], Stefania Bandinelli[6], Anna Murray[1], Luigi Ferrucci[7], Andrew B Singleton[2], David Melzer[4] and Timothy M. Frayling[1],*

[1]Genetics of Complex Traits, Peninsula College of Medicine and Dentistry, University of Exeter, Exeter EX1 2LU, UK, [2]Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, 35 Lincoln Drive, Bethesda, MD, USA, [3]Department of Molecular Neuroscience and Reta Lila Laboratories, Institute of Neurology, UCL, Queen Square House, Queen Square, London WC1N 3BG, UK, [4]Institute of Biomedical and Clinical Sciences, Peninsula College of Medicine and Dentistry, University of Exeter, Barrack Road, Exeter EX2 5DW, UK, [5]Department of Epidemiology and Public Health, University of Maryland School of Medicine, Baltimore, MD, USA, [6]Geriatric Unit, Azienda Sanitaria di Firenze, Florence, Italy and [7]Clinical Research Branch, National Institute on Aging NIA-ASTRA Unit, Harbor Hospital, MD, USA

Received March 3, 2011; Revised July 14, 2011; Accepted July 22, 2011

The identification of multiple signals at individual loci could explain additional phenotypic variance ('missing heritability') of common traits, and help identify causal genes. We examined gene expression levels as a model trait because of the large number of strong genetic effects acting in *cis*. Using expression profiles from 613 individuals, we performed genome-wide single nucleotide polymorphism (SNP) analyses to identify *cis*-expression quantitative trait loci (eQTLs), and conditional analysis to identify second signals. We examined patterns of association when accounting for multiple SNPs at a locus and when including additional SNPs from the 1000 Genomes Project. We identified 1298 *cis*-eQTLs at an approximate false discovery rate 0.01, of which 118 (9%) showed evidence of a second independent signal. For this subset of 118 traits, accounting for two signals resulted in an average 31% increase in phenotypic variance explained (Wilcoxon $P < 0.0001$). The association of SNPs with *cis* gene expression could increase, stay similar or decrease in significance when accounting for linkage disequilibrium with second signals at the same locus. Pairs of SNPs increasing in significance tended to have gene expression increasing alleles on opposite haplotypes, whereas pairs of SNPs decreasing in significance tended to have gene expression increasing alleles on the same haplotypes. Adding data from the 1000 Genomes Project showed that apparently independent signals could be potentially explained by a single association signal. Our results show that accounting for multiple variants at a locus will increase the variance explained in a substantial fraction of loci, but that allelic heterogeneity will be difficult to define without resequencing loci and functional work.

INTRODUCTION

Genome-wide association studies (GWAS) have identified many novel associations between common variants and traits. However, the amount of phenotypic variance explained by these variants remains smaller than that predicted from heritability estimates. The presence of multiple association signals at individual loci, possibly as a result of

[*]To whom correspondence should be addressed. Tel: +44 1392722935; Fax: +44 1392722926; Email: tim.frayling@pms.ac.uk

© *The Author 2011. Published by Oxford University Press.*
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.5), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

allelic heterogeneity, may explain additional phenotypic variation of common traits, and therefore account for some of the 'missing heritability'. Allelic heterogeneity is defined as the presence of multiple alleles that act through one gene to influence a trait. Until recently, few GWAS had performed conditional and multivariable analyses to test the possibility that multiple independent common variants at a locus were associated with a trait. Exceptions include recent studies of height (1), Parkinson's disease (2) and fetal haemoglobin levels (3) and two studies of *cis* expression loci (4,5).

Allelic heterogeneity can be difficult to define for two main reasons. First, variants at the same locus tend to be correlated to varying degrees due to linkage disequilibrium (LD). At any one locus, this correlation often results in the association of many single nucleotide polymorphisms (SNPs) with the trait of interest. Usually, only the SNP with the strongest evidence of association is included to represent a new finding, and other SNPs are not considered independently associated if they are within a certain distance or correlated with the lead SNP above a certain r^2 threshold. Secondly, even if two SNPs in the same region are identified as independently associated with a trait, it is difficult to prove that they act on the same gene.

As well as potentially explaining additional phenotypic variation, more detailed analysis of loci identified by GWAS is important for a second reason—the identification of additional alleles at a locus may help identify the causal gene in the region. An example is the association between common SNPs at the *IFIH1* region and type 1 diabetes. Resequencing of genes in the region identified several low-frequency *IFIH1* coding variants independently associated with type 1 diabetes (6), strongly suggesting that the common SNP acts through *IFIH1* rather than another gene in that region.

To help understand the extent to which multiple signals at the same loci, possibly as a result of allelic heterogeneity, could contribute to common traits we used *cis* gene expression levels. Several GWAS have identified many expression quantitative trait loci (eQTLs) (4,7–11). There are several advantages to using gene expression levels to test allelic heterogeneity. First, eQTLs tend to have relatively strong phenotypic effects—especially *cis*-eQTLs (a variant that influences expression levels of a closely linked gene). Secondly, a large proportion of genes have *cis*-eQTLs. Thirdly, if analyses are limited to *cis*-eQTL associations, then the problem of identifying the causal gene at a locus is largely eliminated—SNPs associated with expression levels of a nearby gene are most likely to be acting directly on that gene. In contrast, some features of the genetics of *cis* gene expression levels may not be representative of common traits—most notably *cis*-effects could predominate the polygenic component.

We performed a *cis*-eQTL analysis using 613 individuals from the InCHIANTI study from whom fresh whole blood mRNA was available as well as genome-wide SNP data. We tested the hypothesis that multiple signals at known loci would explain more phenotypic variation and attempted to identify patterns of association consistent with allelic heterogeneity using *cis* gene expression phenotypes.

RESULTS

Identification of *cis*-SNPs in 1298 loci

We defined a *cis*-SNP as a SNP 1 Mb ± of a probe's start site. Using this definition, and based on HapMap-imputed genotyped data, we identified 1298 probes with at least one *cis*-SNP at $P < 1 \times 10^{-6}$ [approximate false discovery rate (FDR) < 1.0%]. We termed the most strongly associated SNP at each *cis*-signal the 'Index HapMap SNP'.

Conditional analysis identifies 118 (9%) *cis*-eQTLs with evidence of a second signal

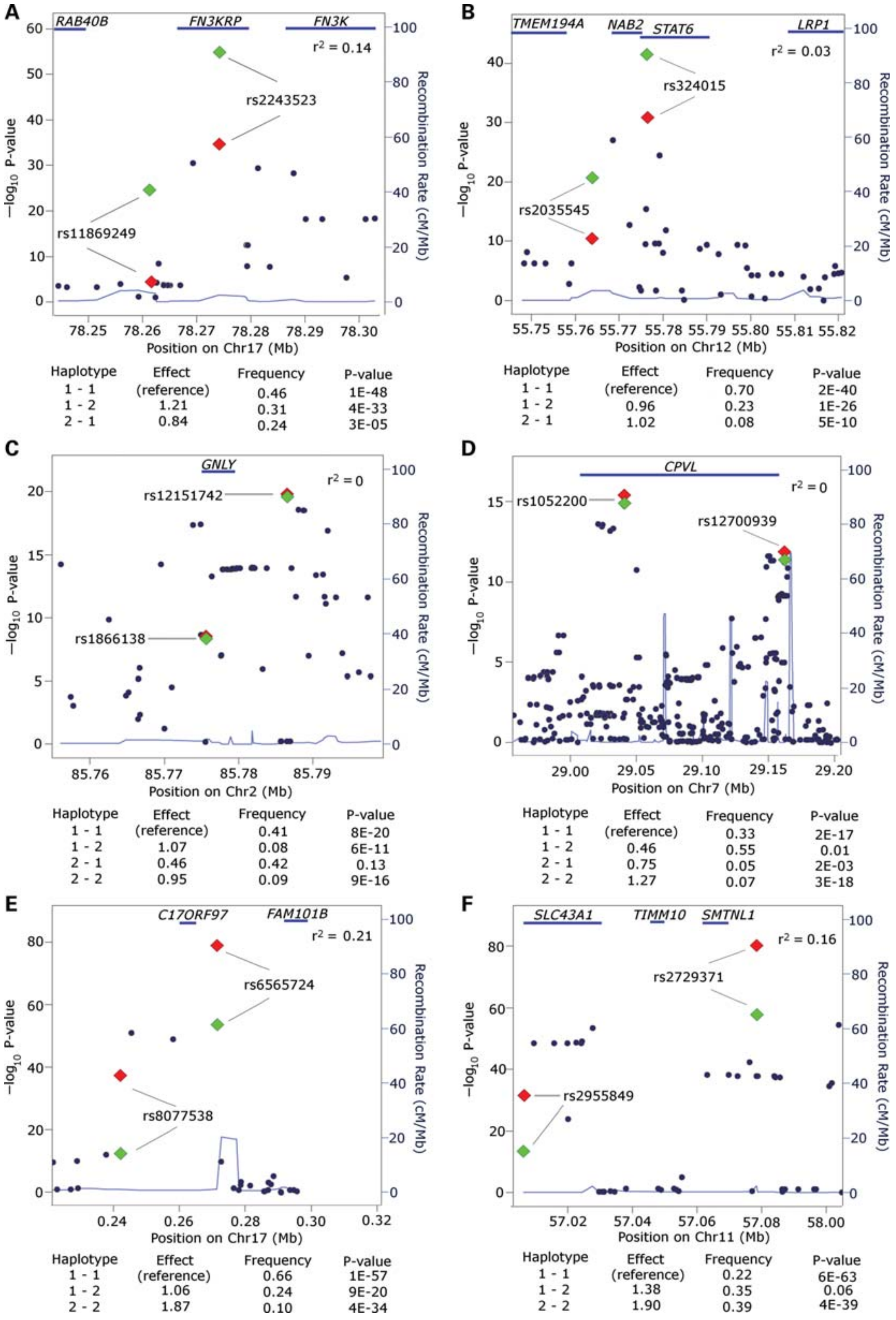
We repeated the association analysis for each of the relevant 1298 probes, but conditioned each on the Index HapMap SNP. We found evidence of a second signal for 118 (~9%) probes (113 separate loci), based on $P < 1 \times 10^{-6}$ (approximate FDR ~1%). We termed this SNP the 'Second HapMap SNP'.

Multivariable analyses explain additional phenotypic variance in loci with evidence of second signals

For each of the 118 probes with evidence of a second signal, we next tested the hypothesis that including SNPs representing both signals would explain more of the variance in the relevant gene expression trait. For each probe, we calculated the variance in *cis* gene expression explained by both SNPs and compared this figure to the variance explained by the Index HapMap SNP alone. We used a model that included both the Index HapMap SNP and the Second HapMap SNP as independent variables and the relevant *cis* gene expression levels as the dependent variable. This multivariable model provides an estimate of the effect of each SNP when taking into account any correlation (due to LD or interaction) with the other. For all 118 loci, including both SNPs increased the phenotypic variance explained compared with the single Index SNP (Supplementary Material, Table S1). The average phenotypic variance explained by the Index SNP alone was 17.5% (range: 3.8–63.9%) and this figure rose to 22.9% (range: 8.3–66.4%) when accounting for both SNPs and the correlation between them, an average increase of 31% (Wilcoxon $P < 0.0001$).

Testing two SNPs in the same statistical model increases, decreases or changes very little their strength of association compared with analyses of single SNPs

For the 118 probes with evidence of a second signal, we observed that the strength of association of the Second HapMap SNP at each *cis*-eQTL locus could increase, decrease or stay very similar when accounting for the LD with the Index HapMap SNP. Examples of the Second HapMap SNPs that increased the most ('jumpers'), reduced the most ('fallers') and changed the least ('stickers') after accounting for the correlation with the Index HapMap SNP are shown in Figure 1 and Table 1. The differences in association statistics between the one-SNP (univariable) and two-SNP (multivariable) analyses were highly correlated to the extent of LD between the two SNPs—association statistics changed least



A

Haplotype	Effect	Frequency	P-value
1 - 1	(reference)	0.46	1E-48
1 - 2	1.21	0.31	4E-33
2 - 1	0.84	0.24	3E-05

B

Haplotype	Effect	Frequency	P-value
1 - 1	(reference)	0.70	2E-40
1 - 2	0.96	0.23	1E-26
2 - 1	1.02	0.08	5E-10

C

Haplotype	Effect	Frequency	P-value
1 - 1	(reference)	0.41	8E-20
1 - 2	1.07	0.08	6E-11
2 - 1	0.46	0.42	0.13
2 - 2	0.95	0.09	9E-16

D

Haplotype	Effect	Frequency	P-value
1 - 1	(reference)	0.33	2E-17
1 - 2	0.46	0.55	0.01
2 - 1	0.75	0.05	2E-03
2 - 2	1.27	0.07	3E-18

E

Haplotype	Effect	Frequency	P-value
1 - 1	(reference)	0.66	1E-57
1 - 2	1.06	0.24	9E-20
2 - 2	1.87	0.10	4E-34

F

Haplotype	Effect	Frequency	P-value
1 - 1	(reference)	0.22	6E-63
1 - 2	1.38	0.35	0.06
2 - 2	1.90	0.39	4E-39

Figure 1. Effects of including two associated *cis*-eQTL SNPs in multivariable analyses. Plots show position of SNPs on the X-axis and $-\log_{10}$ *P*-values for association with *cis* gene expression on the Y-axis. The red diamonds represent the individual (univariable) statistics for the Index HapMap SNP and the second HapMap SNP. The two green diamonds represent the associations of the same two SNPs when accounting for the correlation between the two SNPs using a multivariable model. Estimated haplotype effects are shown underneath each plot, where '2' represents an allele associated with increased gene expression, and '1' represents an allele associated with reduced gene expression. Alleles on haplotypes are ordered by chromosomal position. (**A** and **B**) Examples of 'jumpers', pairs of SNPs that both increase in significance in the multivariable compared with univariable models at the *FN3KRP* and *STAT6* loci, respectively. (**C** and **D**) Examples of 'stickers', pairs of SNPs that remain very similar in significance in the multivariable compared with univariable models at the *GNLY* and *CPVL* loci, respectively. (**E** and **F**) Examples of 'fallers', pairs of SNPs that both reduce in significance in multivariable compared with univariable models at the *C17ORF97* and *TIMM10* loci, respectively.

Table 1. Examples of single (univariable) and two-SNP (multivariable) association statistics in loci with evidence of two signals

Gene	Index HapMap SNP	Second HapMap SNP	Index HapMap *P* Univariable	Index HapMap *P* Multivariable	Second HapMap *P* Univariable	Second HapMap *P* Multivariable	Index SNP—Second SNP r^2
Jumpers							
FN3KRP	rs2243523	rs11869249	4.56E−35	2.32E−55	4.96E−05	1.16E−25	0.14
BTNL3	rs4700774	rs4444930	1.85E−86	5.92E−103	4.63E−05	5.29E−22	0.03
LPCAT2	rs883180	rs2287072	1.91E−102	2.83E−116	0.89	1.05E−15	0.11
DHRS1	rs10134537	rs4568	8.75E−24	1.62E−36	0.54	1.81E−14	0.26
PRMT2	rs2070435	rs11910707	3.26E−28	9.07E−39	2.89E−08	5.59E−19	0.06
STAT6	rs324015	rs2035545	1.05E−31	3.26E−42	8.76E−11	2.03E−21	0.03
KRT72	rs626758	rs681812	1.27E−32	2.75E−42	0.08	5.66E−12	0.11
HOXB2	rs1042815	rs1553748	5.61E−72	1.38E−81	0.32	1.43E−11	0.05
IRF5	rs10229001	rs17424921	1.38E−137	7.30E−147	0.98	3.89E−11	0.05
RNASE2	rs11156734	rs4982386	4.77E−23	2.71E−32	9.04E−04	2.75E−13	0.08
Stickers							
GNLY	rs12151742	rs1866138	1.56E−20	1.66E−20	3.97E−09	4.05E−09	0.00
SUPT3H	rs9395049	rs7773444	3.09E−08	2.51E−08	1.01E−06	8.15E−07	0.00
CA2	rs2930553	rs2548281	4.32E−09	2.79E−09	2.43E−07	1.56E−07	0.00
KLHL5	rs10021255	rs2060005	3.80E−11	2.20E−11	2.11E−07	1.20E−07	0.00
EHD4	rs17686769	rs8034944	2.41E−09	5.74E−09	7.88E−08	1.88E−07	0.00
C17ORF60	rs6504230	rs16947956	3.78E−17	9.04E−17	2.97E−08	6.97E−08	0.00
NSFL1C	rs6105165	rs6079325	9.75E−25	2.77E−24	5.35E−09	1.46E−08	0.01
PLA2G7	rs9472830	rs10081169	3.74E−10	1.10E−09	3.91E−09	1.15E−08	0.00
RSPH3	rs12207795	rs9457532	7.95E−15	3.67E−14	1.83E−08	8.46E−08	0.00
CPVL	rs1052200	rs12700939	1.71E−15	3.66E−16	5.99E−12	1.26E−12	0.00
Fallers							
C17ORF97	rs6565724	rs8077538	2.59E−83	1.82E−57	1.86E−38	1.65E−12	0.21
TIMM10	rs2729371	rs2955849	9.26E−80	5.35E−57	4.90E−37	3.36E−14	0.16
ZP3	rs17718122	rs11505688	4.31E−127	4.72E−106	1.22E−29	1.65E−08	0.15
NAAA	rs11732759	rs11934638	1.19E−48	5.66E−34	3.91E−40	1.90E−25	0.10
DDT	rs5751777	rs11703881	5.48E−49	2.15E−38	4.31E−38	1.69E−27	0.04
SLCO3A1	rs2270059	rs6496898	2.49E−21	6.87E−11	1.96E−19	5.67E−09	0.15
NAAA	rs2242471	rs11934638	1.51E−42	3.13E−32	5.12E−33	1.08E−22	0.08
PLA2G4C	rs274883	rs2307279	5.55E−19	3.79E−09	6.35E−19	4.33E−09	0.20
VSTM1	rs10500316	rs612529	1.85E−25	6.68E−16	2.12E−23	7.82E−14	0.10
NQO2	rs2518581	rs9405188	1.20E−32	1.08E−23	2.43E−22	2.36E−13	0.07

'Jumpers' refers to loci where pairs of SNPs exhibit a relatively large increase in significance in multivariable compared with univariable analysis, 'Fallers' where pairs of SNPs decrease in significance by a comparatively large amount, and 'Stickers' where the significance of pairs of SNPs remain similar. Details from all 118 probe-*cis* gene associations, including effect sizes, are given in Supplementary Material, Table S2.

when LD between SNPs was weakest (Fig. 2, and Supplementary Material, Table S2 and Fig. S1).

To test why some second HapMap SNPs would increase in significance, some decrease and some stay very similar, we performed haplotype analyses. For each of the 118 probes, we calculated associations between haplotypes formed by the two SNPs and *cis* gene expression levels. Examples of these two-SNP haplotypes are shown in Figure 1, where '2' represents an allele associated with increased gene expression, and '1' represents an allele associated with reduced gene expression. We observed two types of haplotype. First, we observed haplotypes where the two alleles associated with increased gene expression tended to occur on opposite haplotypes. These '1–2' or '2–1' haplotypes were most common in the 'jumping' SNPs because a multivariable analysis will adjust for the cancelling out effect of the other SNP. Secondly, we observed haplotypes where the two alleles associated with increased gene expression tended to occur on the same haplotype. These '2–2' haplotypes were most common among the 'falling' SNPs because multivariable analysis will adjust for the correlated effect of the other SNP. 'Sticking' SNPs tended to have more of a mixture of both forms of haplotype (as expected due to the lower LD between them).

Allelic heterogeneity or one variant explains all?

The identification of a second signal at $P < 1 \times 10^{-6}$ after conditioning on the Index HapMap SNP does not necessarily mean there are two independent signals. It is possible that two apparently independent signals are tagging a third, unknown, variant. To investigate this possibility, we re-analysed the 118 probes with evidence of two signals using a denser set of SNPs—those imputed from the 1000 Genomes Project. We compared *cis*-eQTL association results based on HapMap-imputed data with those from 1000 Genomes Project imputed data. We termed the most strongly associated SNP from the 1000 Genome project the '1000G SNP'. We used this approach to test the proof of principle that previously untyped 'novel' variants could potentially explain two apparently independent association signals.

Adding 1000 Genomes imputed data to the 118 probes with evidence of two signals resulted in a range of patterns of association. For most probes, the evidence for two signals remained strong, but we also observed probes where including the most strongly associated 1000 Genomes SNP appeared to appreciably reduce the evidence for two independent signals. Examples of how association statistics change for the probes

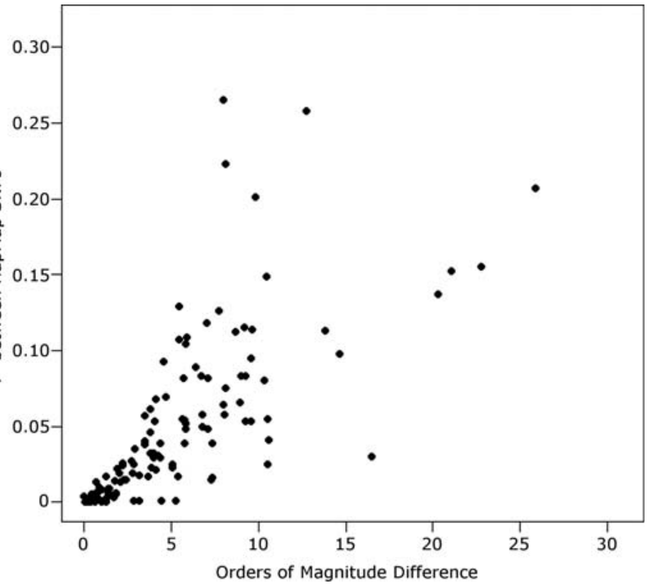


Figure 2. The correlation between how pairs of SNPs change in significance between univariable (single SNP) and multivariable (two SNP) models and the LD between them.

with strong evidence of two independent signals are shown in Figures 3A–D and 4A and Supplementary Material, Table S3. At each of these probes, the statistical strength of the association between the Second HapMap SNP and *cis* gene expression remains very similar when including the Index HapMap SNP and/or the 1000G SNP in the same multivariable test. This robust evidence of a second signal is consistent with the weak LD between the second signal and the other two SNPs. In contrast, the most strongly associated 1000G SNP appears to represent the same signal as the Index HapMap SNP—when these two SNPs are in the same multivariable statistical model, the evidence of association for each falls appreciably, due to the strong LD between them.

Examples of how association statistics change for the probes where two apparently independent signals are more likely to represent a single association signal are shown in Figures 3E–H and 4B and Supplementary Material, Table S3. For all three SNPs at these probes (two HapMap SNPs and the 1000G SNP), the evidence of association with *cis* gene expression falls appreciably when the correlation between all three is accounted for in a multivariable statistical model. These probes are notable in that the association statistics of the HapMap SNPs also fall when each is placed into a two-SNP multivariable model with the 1000G SNP. An example is at the *STAT6* locus. The two *cis* HapMap SNPs at the *STAT6* locus increase in significance when accounting for the LD between them, but including the 1000G SNP results in a pattern more consistent with a single association signal. Each of the two HapMap SNPs at *STAT6* is in moderate LD with the 1000G SNP.

Features of 1000 Genomes SNPs that better explain an association signal compared with HapMap SNPs

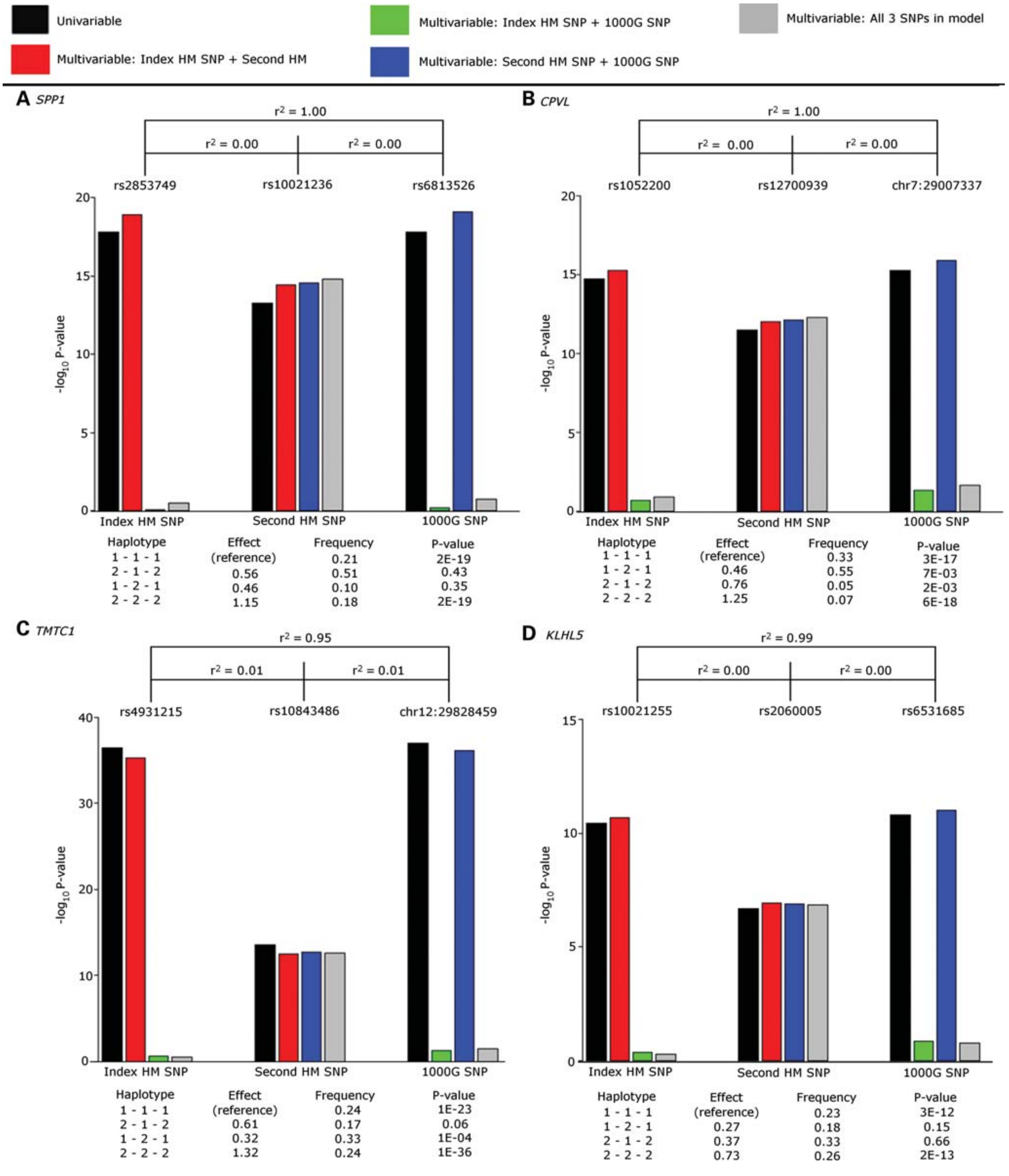
We collapsed the total of 1298 probes with a *cis*-eQTL and the 118 of those probes with evidence of two signals into 1188 and 113 genes, respectively. We examined in more detail the features of the loci where a 1000G SNP was more strongly associated with *cis* gene expression by greater than three orders of magnitude compared with the Index HapMap SNP. We identified 43/1188 genes that matched this criteria and observed 20/43 genes (47%) that overlapped with the 113 genes with at least two signals where we would expect approximately four under the null (Fisher's exact test $P <$ 0.0001). This suggests stronger 1000 Genomes signals are more likely to occur in regions with evidence of two independent signals. Compared with the Index HapMap SNPs, the 1000 Genomes SNPs at these 43 gene loci were not significantly closer to the *cis*-gene (longest transcript identified), no different in allele frequency [1000 Genomes SNP minor allele frequency (MAF) median = 0.29; HapMap Index SNP MAF median = 0.33], and no more likely to be conserved across species [based on the Genomic Evolutionary Rate Profiling (GERP) score] (all $P > 0.05$).

DISCUSSION

We have used *cis* gene expression levels to test the hypothesis that conditional and multivariable analysis of genotypes at known loci will explain additional phenotypic variance, and by inference, more of the 'missing heritability' for common traits. We found evidence of a second signal in 9% of *cis*-eQTLs. Accounting for these second signals resulted in an average increase of 31% in phenotypic variance explained at this subset of loci. Our results are consistent with other *cis*-eQTL analyses (4,5) and a recent analysis of height that identified evidence of a second signal in 5–10% of loci (1). Our results add to these previously published studies because we have performed full conditional and multivariable analyses at each locus.

Whether or not a second signal represents a genuine independent allele or is in partial LD with a single causative allele does not affect our conclusion that these types of analysis can explain more of the variance (and therefore more heritability) of a trait. The distinction between two independent signals and one partially tagged signal is more important when trying to use association results to identify causal genes, or when choosing SNPs for functional studies. Therefore, a second aim of our study was to investigate patterns of association consistent with allelic heterogeneity. Our results have implications for these types of analyses in disease and other clinically relevant traits. The identification of multiple alleles associated with a clinically relevant trait could be extremely important in narrowing the search for likely causal genes. Our study suggests full deep sequencing to identify all variants at a locus will be critical to distinguish between genuine allelic heterogeneity and artefacts that appear as evidence of separate association signals. We identified loci whereby seemingly independent signals represented by SNPs in low LD, and that remain statistically significant after conditional analyses, may not be independent. Instead, these loci may be explained by partially tagging untyped variants.

We observed several patterns of association when analysing single SNPs compared with multiple SNPs at the same loci. The frequency with which trait raising alleles segregated on



A SPP1

Haplotype	Effect	Frequency	P-value
1 - 1 - 1	(reference)	0.21	2E-19
2 - 1 - 2	0.56	0.51	0.43
1 - 2 - 1	0.46	0.10	0.35
2 - 2 - 2	1.15	0.18	2E-19

B CPVL

Haplotype	Effect	Frequency	P-value
1 - 1 - 1	(reference)	0.33	3E-17
1 - 2 - 1	0.46	0.55	7E-03
2 - 1 - 2	0.76	0.05	2E-03
2 - 2 - 2	1.25	0.07	6E-18

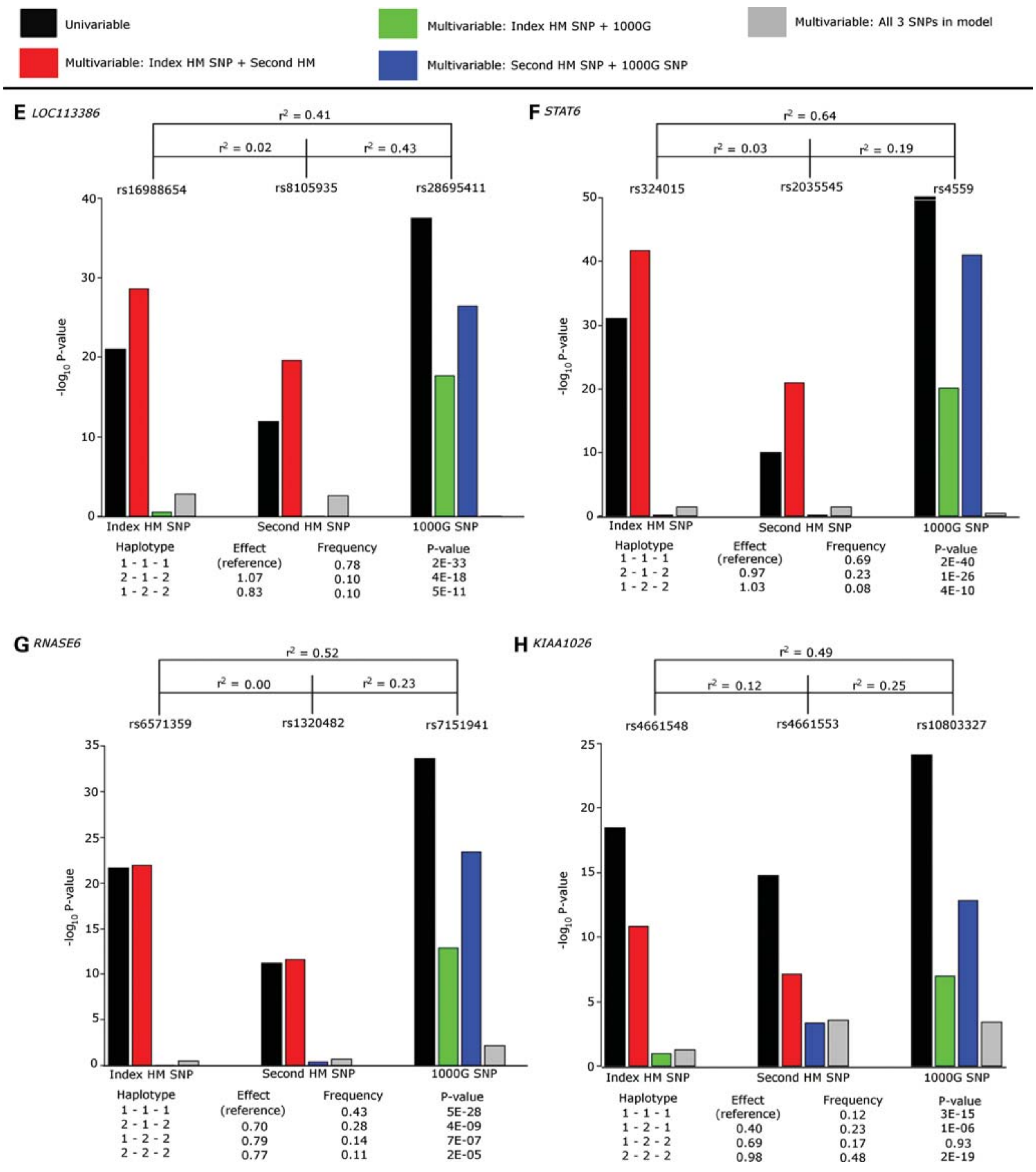
C TMTC1

Haplotype	Effect	Frequency	P-value
1 - 1 - 1	(reference)	0.24	1E-23
2 - 1 - 2	0.61	0.17	0.06
1 - 2 - 1	0.32	0.33	1E-04
2 - 2 - 2	1.32	0.24	1E-36

D KLHL5

Haplotype	Effect	Frequency	P-value
1 - 1 - 1	(reference)	0.23	3E-12
1 - 2 - 1	0.27	0.18	0.15
2 - 1 - 2	0.37	0.33	0.66
2 - 2 - 2	0.73	0.26	2E-13

Figure 3. Effects of including three associated *cis*-eQTL SNPs in multivariable analyses. Bars represent association with *cis* gene expression for three SNPs per locus—from left to right: the Index HapMap SNP; the Second HapMap SNP; and the 1000G SNP. Black bars represent the association of the SNP with *cis* gene expression without taking into account correlation (due to LD) with any other SNPs (univariable analysis). The remaining bars represent the association of the SNP with *cis* gene expression while taking into account any correlation with the other two SNPs, both separately in two SNP models and as a single model with all three SNPs (multivariable analyses). (**A**–**D**) *cis*-eQTL with strongest evidence of allelic heterogeneity. (**E**–**H**) *cis*-eQTL loci where two apparently independent signals could represent a single association signal. Estimated haplotype effects are shown underneath each plot, where '2' represents an allele associated with increased gene expression, and '1' represents an allele associated with reduced gene expression. Alleles on haplotypes are ordered by the Index HapMap SNP, the second HapMap SNP and the 1000G SNP.



E

Haplotype	Effect	Frequency	P-value
1 - 1 - 1	(reference)	0.78	2E-33
2 - 1 - 2	1.07	0.10	4E-18
1 - 2 - 2	0.83	0.10	5E-11

F

Haplotype	Effect	Frequency	P-value
1 - 1 - 1	(reference)	0.69	2E-40
2 - 1 - 2	0.97	0.23	1E-26
1 - 2 - 2	1.03	0.08	4E-10

G

Haplotype	Effect	Frequency	P-value
1 - 1 - 1	(reference)	0.43	5E-28
2 - 1 - 2	0.70	0.28	4E-09
1 - 2 - 2	0.79	0.14	7E-07
2 - 2 - 2	0.77	0.11	2E-05

H

Haplotype	Effect	Frequency	P-value
1 - 1 - 1	(reference)	0.12	3E-15
1 - 2 - 1	0.40	0.23	1E-06
1 - 2 - 2	0.69	0.17	0.93
2 - 2 - 2	0.98	0.48	2E-19

Fig. 3 *Continued*

the same or opposite haplotypes affected the degree to which association statistics ‘fell’ or ‘jumped’ in multivariable compared with univariable analyses. These patterns suggest that in the presence of moderate LD between two functional SNPs, the ability to distinguish between both signals will depend on the haplotype structure. In an extreme example, performing conditional or multivariable analysis on two independent functional SNPs in perfect ($r^2 = 1.0$) LD will not produce any evidence of a second signal.

We examined in more detail the features of loci where a 1000 Genome imputed SNP was more than three orders of magnitude stronger than the best HapMap-imputed SNP.

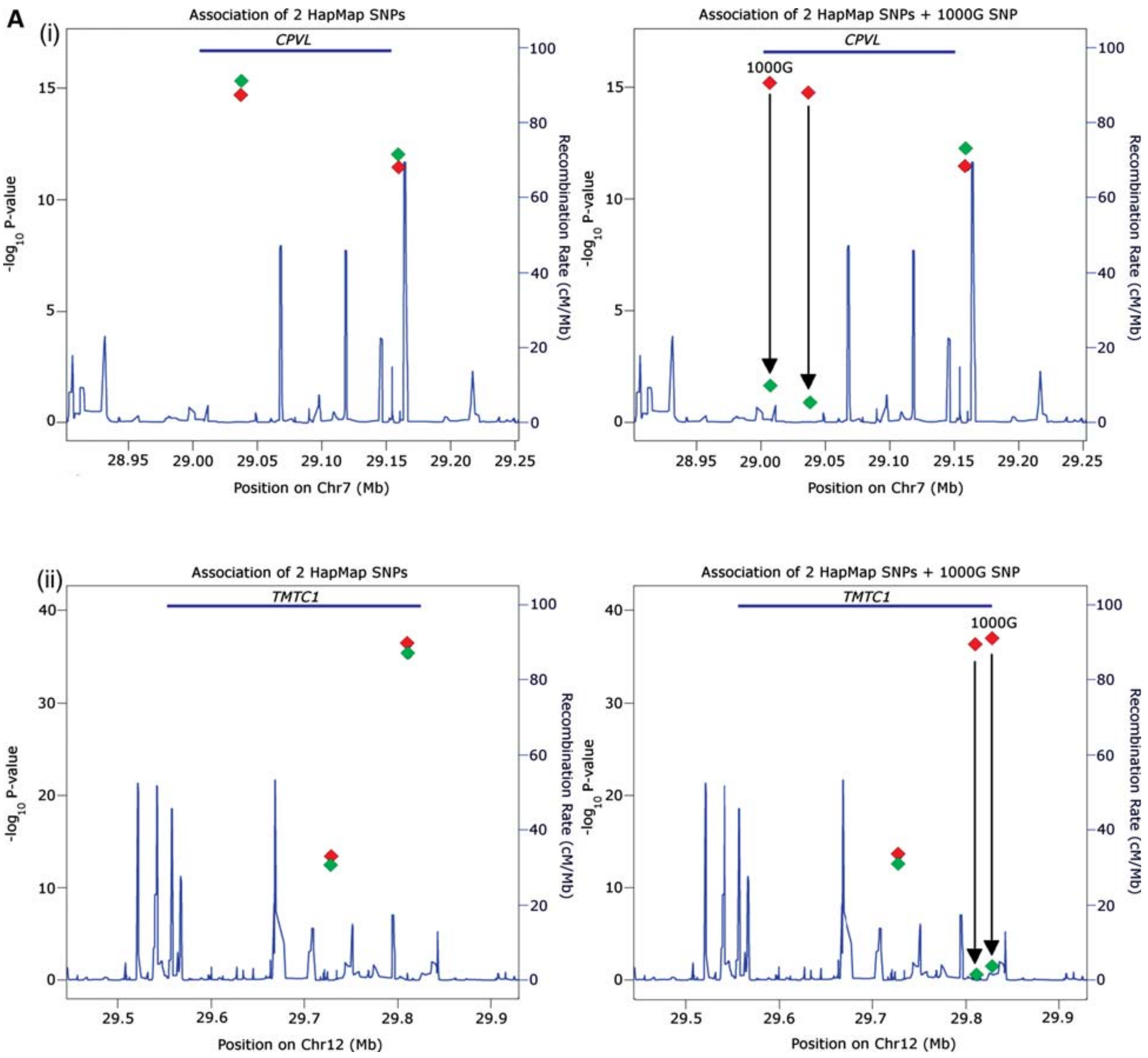


Figure 4. Effects of including three associated *cis*-eQTL SNPs in multivariable analyses shown as regional plots. Plots show position of SNPs on the X-axis and $-\log_{10}$ *P*-value for association with *cis* gene expression on the Y-axis. The red diamonds represent the individual (univariable) statistics. The green diamonds represent the associations of the same SNPs when accounting for the correlation between each other using a multivariable model. Arrows emphasize the directional change of significance; (**A**) examples of two loci where inclusion of the 1000G SNP does not reduce the evidence for two independent signals; (**B**) examples of two loci where a single 1000G SNP appears to account for two apparently independent HapMap *cis*-eQTL SNP associations.

These loci were much more likely to be among those with evidence of two signals. We suggest this enrichment means that loci with evidence of two apparently independent loci should be examined in much more detail before deciding whether or not one, two or more signals are responsible for the association between locus and trait. The 1000 Genomes imputed SNPs were similar in frequency and conservation score to the best HapMap-imputed SNPs and so did not reveal any evidence for the ‘synthetic association caused by rare variants’ hypothesis put forward by Goldstein and colleagues (12).

There are some limitations to our study. First, we have not definitively proven or disproven cases of genuine allelic heterogeneity. We have only identified example loci where two SNPs are more likely to represent two independent signals compared with other example loci, where the results are more consistent with a single association signal. Secondly, we have largely used imputed data, which results in more error in the estimation of LD compared with directly typed variants. However, results were similar at a subset of loci where direct genotypes were available (data not shown). Thirdly, this analysis has been limited to common variants identified by the HapMap and 1000 Genomes Project. Further studies are needed to assess the variance explained, and patterns of association that occur, when low frequency and rare SNPs are included. Finally, we have not been able to measure heritability as limited family data was available. However, through explaining more phenotypic variation

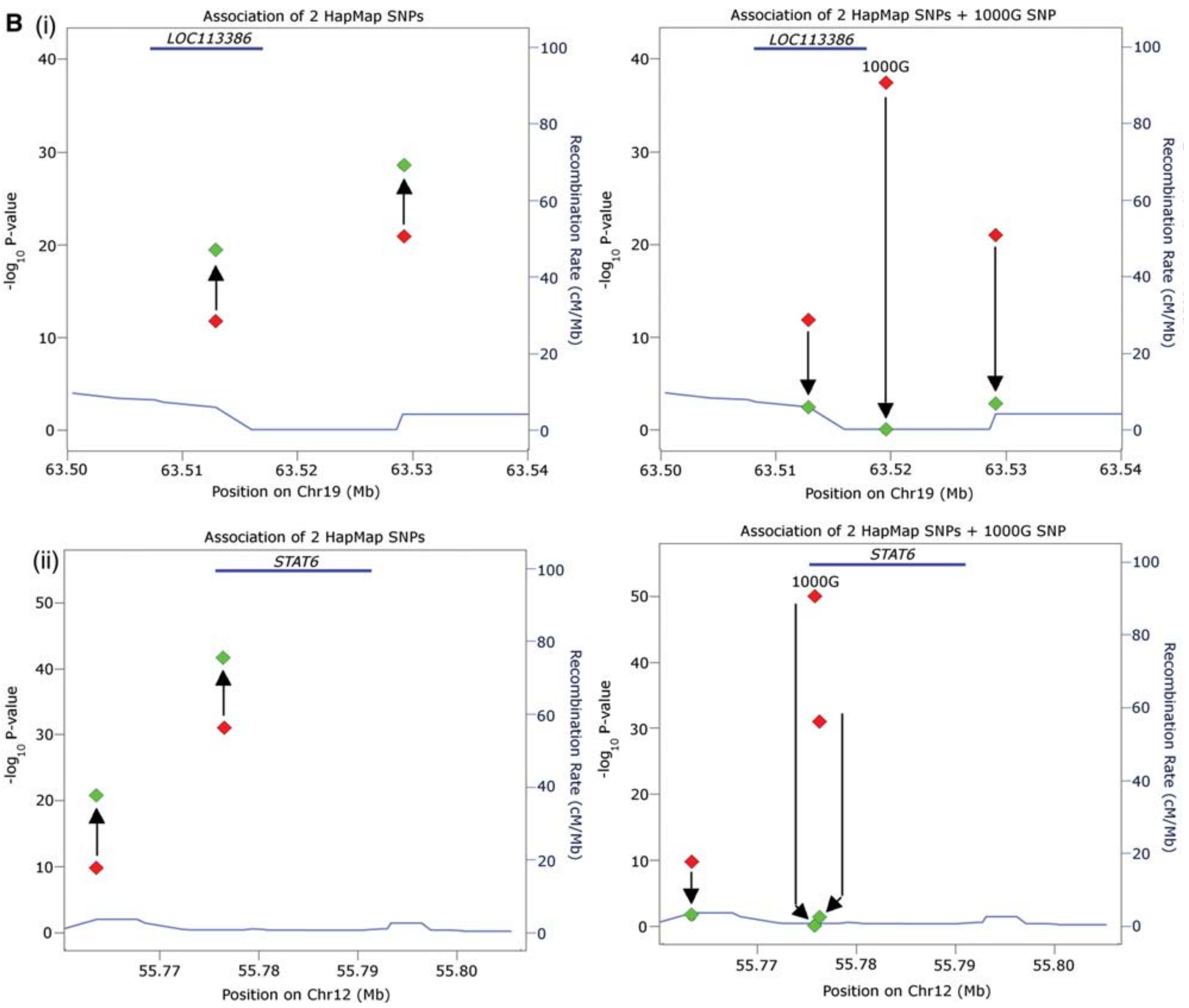


Fig. 4 *Continued*

through multiple additive effects (assuming no population stratification), we must have explained more of the heritability.

MATERIALS AND METHODS

Samples

We used individuals from the InCHIANTI study; a study of aging from the Chianti region in Tuscany, Italy (13,14).

Expression profiling

Peripheral blood specimens were collected from 712 individuals using the PAXgene tube technology to preserve levels of mRNA transcripts. RNA was extracted from peripheral blood samples using the PAXgene Blood mRNA kit (Qiagen, Crawley, UK) according to the manufacturer's instructions.

RNA was biotinylated and amplified using the Illumina® TotalPrep(tm) -96 RNA Amplification Kit and directly hybed with HumanHT-12_v3 Expression BeadChips that include 48 803 probes. Image data were collected on an Illumina iScan and analysed using Illumina GenomeStudio software. These experiments were performed as per the manufacturer's instructions and as previously described (11).

Genotyping and imputation

Genome-wide genotyping was performed using the Illumina Infinium HumanHap550 genotyping chip. Standard quality-control procedures were used to filter out SNPs with MAF $<1\%$, Hardy–Weinberg $P < 1 \times 10^{-4}$ and a call rate $<99\%$. This resulted in 495 343 directly genotyped SNPs.

We used MACH 1.0.16 to impute missing genotypes not captured by the Illumina chip. We formed two imputed data sets using the HapMap r22 build-36 reference panel to impute 2 543 887 SNPs and the June 2010 release of the 1000 Genomes Project (15) build-36 reference panel to impute 6 858 242 SNPs. Imputed SNPs were excluded from the analysis if their MAF was <0.01 and if their r^2

[a measure of imputation quality in MaCH (16)] was <0.3 or <0.5 depending on whether they were based on the HapMap or 1000 Genomes Project reference panels, respectively, as recommended and consistent with other literature (16).

Quality-control analysis of gene expression levels

The BeadChip microarrays were normalized using a cubic-spline normalization algorithm using Illumina's BeadStudio software in an attempt to minimize environmental factors during microarray processing that may affect levels of expression.

Next, each probe's expression value for a given individual that had a BeadStudio detection $P < 0.01$ was flagged as not differentially expressed from noise.

Using probe intensity values that were classified as above background noise, subjects were removed if their mean intensities were outside 3 standard deviations, or the proportion of probes expressed was less than 3 standard deviations. This left 698 individuals, of which 613 had genome-wide SNP data available.

Using the 698 individuals with good quality gene expression data, we removed probes that were not differentially expressed above background noise in over 5% of samples. Of the 16 571 probes that passed this threshold, we removed a further 2 682 probes that were reported to have had at least one SNP within the 50 bp probe region, based on dbSNP 129. We subsequently removed a further 68 probes as their start sites could not be identified. This left 13 821 probes with which to perform association analyses.

Cis-eQTL association testing

We transformed the post-quality-control expression levels for each probe by inverse normalization of expression residuals that adjusted for age, sex, amplification batch and hybridization batch.

To identify *cis*-SNP–probe associations, we initially performed association analysis of the 13 821 probes using directly genotyped data and the program PLINK. We defined a *cis*-SNP as a SNP 1 Mb ± of a probe's start site. Probe coordinates were mapped to HG18 using ReMOAT; a re-annotation pipeline set up for enhancing the annotation of Illumina BeadArrays (17). For association testing, we used a *cis*-SNP significance threshold of $P < 1 \times 10^{-6}$. We estimated the false discovery rate using two methods. First, we used the widely accepted criteria for GWAS that the genome contains ~1 million independent common variants. If the genome is 3000 Mb then a *cis* locus of 2 Mb of sequence is ~1/1500 of the genome and each *cis* locus will contain 1000000/1500 = 667 independent variants. We have tested 13 821 probes so performed 13 821 × 667 = 9.2 million tests. A *P*-value of $\sim 5 \times 10^{-9}$ therefore provides a Bonferroni-based corrected *P*-value of 0.05 and a *P*-value = 1×10^{-6} the threshold at which we would expect ~10 associations by chance. For second signals, we used the same principle but based the number of tests on the number of loci with primary signals ×667. Secondly, we selected all SNPs (best-guess genotypes for imputed SNPs) in the 1298 2 Mb regions with evidence of a *cis*-eQTL. We estimated the number of independent *cis*-SNPs in these regions using the indep-pairwise command in PLINK and an r^2 cut-off of 0.5. We repeated this calculation for the subset of 118 probes with evidence of two signals. The 2 Mb regions flanking the 1298 and 118 probes contained an average of 280 and 307 'independent' SNPs, respectively. These figures are still estimates because defining independence at $r^2 = 0.5$ is arbitrary and so we used the more conservative figure of 667 independent signals per *cis* locus.

We then followed up probes that showed potential associations with at least 1 *cis*-SNP $P < 1 \times 10^{-6}$ by repeating the analyses using the imputed HapMap data set using MACH2QTL. We termed the SNPs showing the strongest *cis* associations in this analysis the 'Index HapMap SNP'. The Index HapMap SNP was placed into a univariable model for analysis ($y = c + \beta_1 \cdot \mathrm{SNP}_1$) in STATA.

As a quality-control step, we also compared our *cis*-eQTL data to that from European HapMap samples. Of 293 *cis*-eQTLs reported by Stranger *et al*. (8), we identified 54 at $P < 5 \times 10^{-9}$ (Supplementary Material, Table S4) despite differences in cell type between our data (gene expression from whole blood) and the HapMap sample data (lymphoblastoid cell lines).

Conditional analyses using HapMap-imputed data

We next performed conditional analyses on each probe using the Index HapMap SNP. For each of the 1298 associations with evidence of a *cis*-eQTL, we performed a conditional analysis, where we included the Index HapMap SNP as a covariate in a reanalysis of the relevant chromosome, using the program MACH2QTL.

Probes with evidence of a second *cis*-signal associated at $P < 1 \times 10^{-6}$ (we termed the most strongly associated SNP from this analysis the 'Second HapMap SNP') were then re-analysed in a multivariable model. We placed both the Index HapMap SNP and the Second HapMap SNP into a multivariable model ($y = c + \beta_1 \cdot \mathrm{SNP}_1 + \beta_2 \cdot \mathrm{SNP}_2$) using STATA. This allowed us to assess the association between each SNP and *cis* gene expression, as well as the amount of variance explained by each SNP, when accounting for any correlation between the two. We compared the amount of variance explained to that when using a univariable analysis consisting of only the Index HapMap SNP. We used estimated counts of the reference allele derived from posterior probabilities, or 'dosages', at each SNP to represent genotypes in each individual.

Adding 1000 genomes project data

We next selected the 118 loci with two signals associated with *cis* gene expression at $P < 1 \times 10^{-6}$. We performed association testing in MACH2QTL using an imputed InCHIANTI data set based on the June 2010 release of the 1000 Genomes Project.

Using dosages derived from 1000 Genomes Project-based imputation, we performed further multivariable analyses. We included three SNPs in the statistical model: the Index HapMap SNP, the Second HapMap SNP and the most strongly associated 1000 Genomes SNP (where different from the top HapMap SNP) as independent variables and *cis* gene expression levels as the dependent variable ($y = c + \beta_1 \cdot \mathrm{SNP}_1 + \beta_2 \cdot \mathrm{SNP}_2 + \beta_3 \cdot \mathrm{SNP}_3$). We termed the most strongly associated 1000 Genomes SNP the '1000G SNP'. We also performed two further multivariable analyses using just two SNPs—the most strongly associated 1000 Genomes SNP with each of the two HapMap SNPs.

Haplotype analysis

We performed analyses of estimated haplotypes using UNPHASED (18) to examine their association against levels of expression. Prior to haplotype construction, we generated best-guess SNP genotypes based on dosage data and coded expression increasing alleles as '2' and expression decreasing alleles as '1'. We examined their frequency, significance and main effects by testing each one while adjusting for all others in the same model.

Directly genotyped SNPs

We attempted to verify data from the analyses described above based on SNP dosages using directly typed SNP genotypes if available. These were put into the same two-HapMap SNP multivariable models described above where possible in order to evaluate the sensitivity of the data.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at *HMG* online.

Conflict of Interest statement. None declared.

FUNDING

This work was supported by the Wellcome Trust 083270/Z/07/Z. The InCHIANTI study was supported by contract funding from the U.S. National Institute on Aging (NIA), and the research was supported in part by the Intramural Research Program, NIA, and National Institute of Health (NIH). A.R.W. supported by the Peninsula NIHR Clinical Research Facility. Funding to pay the Open Access publication charges for this article was provided by the Wellcome Trust.

REFERENCES

1. Lango Allen, H., Estrada, K., Lettre, G., Berndt, S.I., Weedon, M.N., Rivadeneira, F., Willer, C.J., Jackson, A.U., Vedantam, S., Raychaudhuri, S. *et al.* (2010) Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, **467**, 832–838.
2. Spencer, C.C., Plagnol, V., Strange, A., Gardner, M., Paisan-Ruiz, C., Band, G., Barker, R.A., Bellenguez, C., Bhatia, K., Blackburn, H. *et al.* (2011) Dissection of the genetics of Parkinson's disease identifies an additional association 5′ of SNCA and multiple associated haplotypes at 17q21. *Hum. Mol. Genet.*, **20**, 345–353.
3. Galarneau, G., Palmer, C.D., Sankaran, V.G., Orkin, S.H., Hirschhorn, J.N. and Lettre, G. (2010) Fine-mapping at three loci known to affect fetal hemoglobin levels explains additional genetic variation. *Nat. Genet.*, **42**, 1049–1051.
4. Dimas, A.S., Deutsch, S., Stranger, B.E., Montgomery, S.B., Borel, C., Attar-Cohen, H., Ingle, C., Beazley, C., Gutierrez Arcelus, M., Sekowska, M. *et al.* (2009) Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science*, **325**, 1246–1250.
5. Nica, A.C., Parts, L., Glass, D., Nisbet, J., Barrett, A., Sekowska, M., Travers, M., Potter, S., Grundberg, E., Small, K. *et al.* (2011) The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. *PLoS Genet.*, **7**, e1002003.
6. Nejentsev, S., Walker, N., Riches, D., Egholm, M. and Todd, J.A. (2009) Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science*, **324**, 387–389.
7. Dixon, A.L., Liang, L., Moffatt, M.F., Chen, W., Heath, S., Wong, K.C., Taylor, J., Burnett, E., Gut, I., Farrall, M. *et al.* (2007) A genome-wide association study of global gene expression. *Nat. Genet.*, **39**, 1202–1207.
8. Stranger, B.E., Nica, A.C., Forrest, M.S., Dimas, A., Bird, C.P., Beazley, C., Ingle, C.E., Dunning, M., Flicek, P., Koller, D. *et al.* (2007) Population genomics of human gene expression. *Nat. Genet.*, **39**, 1217–1224.
9. Myers, A.J., Gibbs, J.R., Webster, J.A., Rohrer, K., Zhao, A., Marlowe, L., Kaleem, M., Leung, D., Bryden, L., Nath, P. *et al.* (2007) A survey of genetic human cortical gene expression. *Nat. Genet.*, **39**, 1494–1499.
10. Ge, B., Pokholok, D.K., Kwan, T., Grundberg, E., Morcos, L., Verlaan, D.J., Le, J., Koka, V., Lam, K.C., Gagne, V. *et al.* (2009) Global patterns of cis variation in human cells revealed by high-density allelic expression analysis. *Nat. Genet.*, **41**, 1216–1222.
11. Gibbs, J.R., van der Brug, M.P., Hernandez, D.G., Traynor, B.J., Nalls, M.A., Lai, S.L., Arepalli, S., Dillman, A., Rafferty, I.P., Troncoso, J. *et al.* (2010) Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. *PLoS Genet.*, **6**, e1000952.
12. Dickson, S.P., Wang, K., Krantz, I., Hakonarson, H. and Goldstein, D.B. (2010) Rare variants create synthetic genome-wide associations. *PLoS Biol.*, **8**, e1000294.
13. Ferrucci, L., Bandinelli, S., Benvenuti, E., Di Iorio, A., Macchi, C., Harris, T.B. and Guralnik, J.M. (2000) Subsystems contributing to the decline in ability to walk: bridging the gap between epidemiology and geriatric practice in the InCHIANTI study. *J. Am. Geriatr. Soc.*, **48**, 1618–1625.
14. Melzer, D., Perry, J.R., Hernandez, D., Corsi, A.M., Stevens, K., Rafferty, I., Lauretani, F., Murray, A., Gibbs, J.R., Paolisso, G. *et al.* (2008) A genome-wide association study identifies protein quantitative trait loci (pQTLs). *PLoS Genet.*, **4**, e1000072.
15. Durbin, R.M., Abecasis, G.R., Altshuler, D.L., Auton, A., Brooks, L.D., Durbin, R.M., Gibbs, R.A., Hurles, M.E. and McVean, G.A. (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.
16. Li, Y., Willer, C.J., Ding, J., Scheet, P. and Abecasis, G.R. (2010) MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.*, **34**, 816–834.
17. Barbosa-Morais, N.L., Dunning, M.J., Samarajiwa, S.A., Darot, J.F., Ritchie, M.E., Lynch, A.G. and Tavare, S. (2010) A re-annotation pipeline for Illumina BeadArrays: improving the interpretation of gene expression data. *Nucleic Acids Res.*, **38**, e17.
18. Dudbridge, F. (2008) Likelihood-based association analysis for nuclear families and unrelated subjects with missing genotype data. *Hum. Hered.*, **66**, 87–98.