Human Molecular Genetics, 2017, Vol. 26, No. 8 1444–1451

doi: 10.1093/hmg/ddx043
Advance Access Publication Date: 6 February 2017
Original Article

ORIGINAL ARTICLE

Conditional eQTL analysis reveals allelic heterogeneity of gene expression

Rick Jansen[1,*], Jouke-Jan Hottenga[2], Michel G. Nivard[2], Abdel Abdellaoui[2], Bram Laport[1], Eco J. de Geus[2], Fred A. Wright[3], Brenda W.J.H. Penninx[1,†] and Dorret I. Boomsma[2,†]

[1]Department of Psychiatry, Vrije Universiteit Medical Center, Amsterdam Neuroscience, Amsterdam, The Netherlands, [2]Department of Biological Psychology, Vrije Universiteit Amsterdam, Amsterdam Neuroscience, Amsterdam Public Health, Amsterdam, The Netherlands and [3]Departments of Statistics and Biological Sciences, Bioinformatics Research Center, North Carolina State University, NC, USA

*To whom correspondence should be addressed at: 1187, 1081 HL Amsterdam, The Netherlands. Tel: +31207884682; Fax: +31207885664; Email: ri.jansen@ggzingeest.nl

Abstract

In recent years, multiple eQTL (expression quantitative trait loci) catalogs have become available that can help understand the functionality of complex trait-related single nucleotide polymorphisms (SNPs). In eQTL catalogs, gene expression is often strongly associated with multiple SNPs, which may reflect either one or multiple independent associations. Conditional eQTL analysis allows a distinction between dependent and independent eQTLs. We performed conditional eQTL analysis in 4,896 peripheral blood microarray gene expression samples. Our analysis showed that 35% of genes with a cis eQTL have at least two independent cis eQTLs; for several genes up to 13 independent cis eQTLs were identified. Also, 12% (671) of the independent cis eQTLs identified in conditional analyses were not significant in unconditional analyses. The number of GWAS catalog SNPs identified as eQTL in the conditional analyses increases with 24% as compared to unconditional analyses. We provide an online conditional cis eQTL mapping catalog for whole blood (https://eqtl.onderzoek.io/), which can be used to lookup eQTLs more accurately than in standard unconditional whole blood eQTL databases.

Introduction

The genome and the transcriptome are highly interconnected. Currently for ~ 50% of all genes, cis eQTLs (SNPs < 1Mb distance from the associated gene expression gene) have been identified in whole blood micro array studies with sample sizes of around 5000 individuals (1,2). RNA-seq studies found cis eQTLs for 79% of all genes, using 922 peripheral blood samples (3), and for 49% in 462 lymphoblastoid cell line samples (4). Studies of other (non-blood) tissues often used smaller sample sizes but still discovered a large number of eQTLs. For example, in brain tissue 32% of all transcripts measured have an eQTL in one of 10 brain regions ($N < 131$) (5). Meta eQTL analyses across several brain regions ($N = 424$) found cis eQTLs for ~ 18% of all genes (6). The number of identified eQTLs depends on sample size and the threshold for significance, which varies from study to study, but the trend suggests that with the currently increasing sample sizes and variety of tissues (7), cis eQTLs are likely to be identified for all expressed genes.

[†]These authors contributed equally to this work.

Received: September 16, 2016. **Revised:** January 26, 2017. **Accepted:** January 27, 2017

© The Author 2017. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

Table 1. Number of probesets and genes with 1, 2, 3, . . ., or 13 independent cis eQTL effects. In total 44,241 probesets were measured, targeting 18,238 genes

# independent eQTLs	≥1	≥2	≥3	≥4	≥5	≥6	≥7	≥8	≥9	≥10	≥11	≥12	13
# probesets	13583	4039	1251	485	229	113	56	29	16	10	4	4	3
# genes	7120	2485	830	323	154	75	40	21	13	6	3	3	3

Many eQTL studies provide online eQTL databases that can be used, for example, to determine whether an SNP identified in a genome wide association study (GWAS SNP) is regulating gene expression. Most GWAS SNPs are non-coding and therefore are likely to function by regulating gene expression and protein levels. Looking up significant GWAS SNPs in eQTL databases has become part of standard post-hoc analysis performed for GWAS (8–11). GWAS SNPs are often in linkage disequilibrium (LD) with regions containing multiple genes; eQTL analysis can help identifying the causal gene. For example, a locus associated with myocardial infarction is associated with *SORT1* expression, while *SORT1* is located 40KB and two genes away from the causal SNP (12). *SORT1* upregulation leads to higher LDL-cholesterol levels, a risk factor for myocardial infarction. Identifying the SNP-expression-disease pathway may ultimately lead to treatments targeting gene or protein expression. For example, variants in *TSLP* are associated with asthma and *TSLP* expression: decreasing *TSLP* expression reduces asthma symptoms (13). Recent reports show that RNA levels contribute on average 73% to protein level variance (14,15), and eQTLs are often also associated with protein levels. This emphasizes the relevance of studying RNA, given the small sample size and/or small number of proteins in the current protein-QTL studies.

It is not always possible to determine whether a GWAS SNP is tagging a locus that contains the variant controlling gene expression in the current eQTL databases that only provide unconditional analysis results. Besides the difficulty to determine whether two association signals tag the same causal variant (16–18), gene expression is often regulated by multiple independent eQTL SNPs: >26% of all genes have at least two independent eQTLs (4,19). The GWAS SNP may tag any of the independent eQTL effects, but these independent eQTLs can only be identified by the conditional eQTL analysis. Several eQTL studies have computed independent eQTL effects using one or two conditioning steps and identified up to three independent eQTL effects per gene (20), but to the best of our knowledge none of them provided the full conditional analysis results. Here we perform and provide full conditional eQTL mapping based on 4,896 samples from the Dutch NESDA (21) and NTR (22) cohorts measured with Affymetrix U219 micro arrays. For some genes up to 12 independent eQTL SNPs were identified, and many of these eQTLs were not found in the unconditional analysis. We also demonstrate that conditional eQTL analyses increases the number of GWAS catalog SNPs that can be reliably classified as eQTLs.

Results

Peripheral blood gene expression was measured in 4,896 subjects with European ancestry (1,880 unrelated subjects from NESDA, 559 MZ twin pairs, 594 DZ twin pairs, 51 parent-sibling trios and 557 unrelated subjects from NTR). SNPs were imputed using the 1000 Genomes (phase 1) reference and coded additive-codominantly (0, 1 or 2). For each gene expression (44,241 probesets targeting 18,238 genes) - SNP ($N = 8{,}158{,}830$) pair at distance <1 Mb a linear model was fitted with gene expression as dependent, and genotype as independent variable after correcting for several technical and demographic covariates (Materials & Methods). FDR was computed based on permutations that accounted for relatedness. At a FDR of 5% ($P < 1e\text{-}5$), cis eQTLs were identified for 13,583 probesets targeting 7,120 genes (31% of all probesets, 39% of all measured genes). The number of cis eQTLs per probeset varies considerably (median = 115, mean = 237, SD = 424); there were 410 probesets with more than 1000 cis eQTLs. The cis eQTLs associated with the expression of the same gene may harbor multiple independent associations, or may contain only one signal on which all associations are dependent. In order to reveal dependent and independent associations, for each expression probeset the most significant association was identified (E1 SNP or primary eQTL SNP), and cis eQTL analysis was repeated for each probeset conditional on the corresponding E1 SNP. This second round of cis eQTL analysis revealed 4,039 probesets (targeting 2,485 genes) with significant cis eQTLs conditional on the E1 SNPs. Thus, 35% of the genes with a cis eQTL (14% of all genes) have at least one additional independent cis eQTL effect. For each probeset the most significant cis eQTL conditional on the E1 SNP was selected (E2 SNP or secondary eQTL SNP) and cis eQTL analysis was repeated conditional on the E1 and E2 SNPs. For 1251 probeset targeting 830 genes a third independent cis eQTL effect was identified (E3 SNP). The conditional cis eQTL analysis was repeated conditional on E1, E2, E3, . . ., E13 SNPs. After correcting for 13 independent cis eQTL effects no further associations were retrieved. Table 1 gives an overview of the number of independent cis eQTL effects per probeset and gene, Supplementary Material, Table S1 contains E1-E13 SNPs for each probeset. Six genes with 10 or more independent cis eQTL effects were identified: *HLA-C*, *HLA-B* (chr 6, MHC class 1), *HLA-DPA1* (chr 6, MHC class 2), *STAT6* (chr 10, a transcription factor involved in the innate immune system), *ZNF815P* (chr 7) and *KRT23* (keratin family, chr17).

LD between E1 and corresponding E2 SNPs was on average 0.37 (sd 0.17), see Supplementary Materials, Fig. S1 and Table S1. In order to estimate the replication rates of conditional eQTLs, we used the results of a recent eQTL study in whole blood, using RNA-seq (23), the BIOS study. This study provided top eQTLs after several conditioning steps (E1-EN SNPs). We computed LD between the corresponding E1-EN SNPs from the two studies. From the E1 SNPs we identified (top EQTL SNPs without conditioning) there were 65% in LD > 0.1 with the E1 SNPs from the BIOS study. From the E2 SNPs, there were 37% in LD > 0.1, and from the E3 SNPs 23%. We note that for the replication of a conditional eQTL, the conditioning should be the same in the two studies. From the eQTLs for which the E1 SNPs were in LD > 0.8 between the two studies (forcing a similar conditioning step), the corresponding E2 SNPs were in LD > 0.1 for 45% of the eQTLs.

Figure 1 shows that the number of independent eQTL effects is correlated with the number of eQTLs identified in the unconditional analysis. However, there are many probesets with only

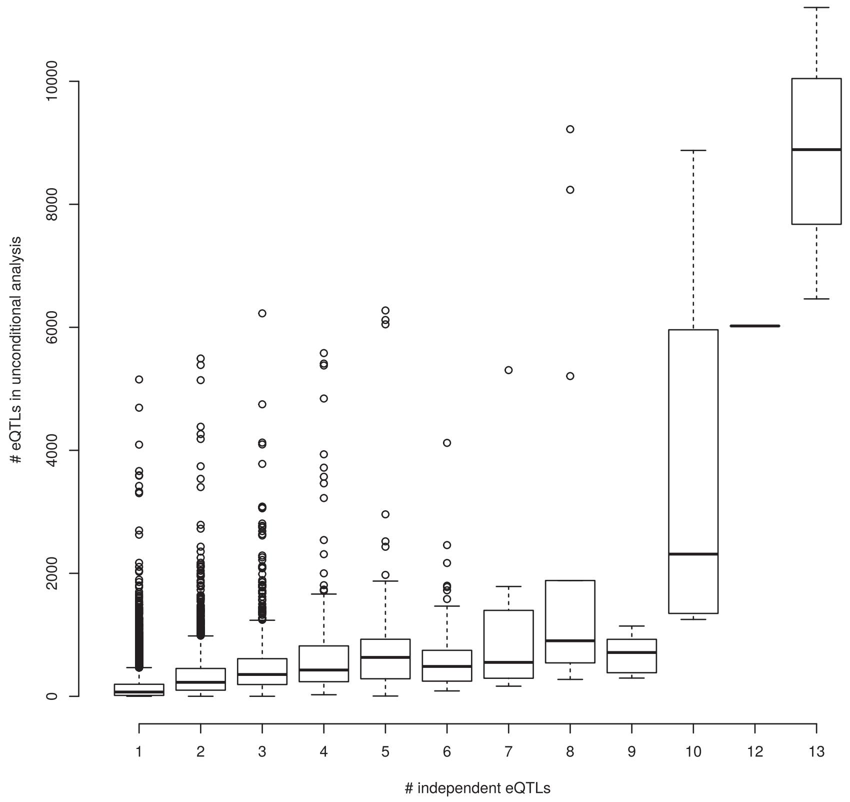

Figure 1. For each gene expression probeset, the number of independent cis eQTLs (identified using conditional eQTL analysis) are plotted versus the number of eQTLs identified in the unconditional analysis.

one independent eQTL which have many eQTLs in the unconditional analysis (9542 probesets have only one independent eQTL: 3960 of them have >100, 114 have >1000 eQTLs in the unconditional analysis). For these 9542 probesets with only one independent eQTL, all significant eQTLs identified in the unconditional analysis most likely reflect one underlying signal.

Conditional analysis reveals eQTLs not identified in unconditional analysis

From the 3720 unique E2 SNPs, 621 of the E2 SNP-gene pairs were not identified in the unconditional analysis (17%), and 366 (10%) of the E2 SNPs were not identified in the unconditional analysis. The E2 SNPs not identified in the unconditional analysis are strongly associated with corresponding gene expression, so these do not simply reflect false positives ($P < 1e\text{-}7$, for 124 out of 366 E2 SNPs). Likewise, after conditioning on E1 and E2 SNPs, 1158 E3 SNPs were identified, from which 164 (14%) were not significant in the unconditional analysis. From all E2–E13 SNPs, 671 (12%) were not significant in unconditional analysis. E2 SNPs only identified in conditional analysis are ‘masked’ by the effect of the corresponding E1 SNP: 34% of them are positively correlated with the E1 SNP, but have a direction of effect on expression opposite to the effect of the E1 SNP on expression. And 65% of them are negatively correlated with the E1 SNP, but have the same direction of effect on expression as the E1 SNP. For example, rs946262 is the E1 SNP for *CHI3L1* expression ($P = 2.2e\text{-}308$, beta=-0.9), and rs12023876 has no effect on unconditional expression of *CHI3L1* ($P > 0.05$, Supplementary Materials, Fig. S2). However, the association between *CHI3L1* expression and rs12023876 conditional on rs946262 is very strong ($P = 2.7e\text{-}39$, beta = 0.2). The correlation coefficient between rs12023876 and rs946262 is 0.23. Thus, the SNPs have a positive correlation but a different direction of effect on *CHI3L1* expression, thereby creating the masking effect.

A lookup of the 671 eQTL SNPs not identified in the unconditional analysis in the GWAS catalog (https://www.genome.gov/26525384) revealed 4 eQTL SNPs in strong LD ($r^2 > 0.8$) with a SNP reported in the GWAS catalog. For example, rs11676950 (E2 SNP for *FAM117B*, $P = 2.3e\text{-}34$), is in strong LD with a GWAS hit for total cholesterol levels (rs11694172, $r^2 = 0.8$ (24)). In the original analysis in the paper reporting this GWAS, the identified SNPs were associated with gene expression in an unconditional analysis, and like in the unconditional analysis performed here, rs11694172 was not identified as eQTL. *FAM117B* knockout mice have lower cholesterol levels ($P < 1.3e\text{-}8$ https://www.mousephenotype.org) and *FAM117B* expression is associated with cholesterol in our sample ($P < 1.7e\text{-}4$, Beta = 0.03, $N = 3306$).

The percentage, however, of the 671 eQTL SNPs in strong LD with GWAS hits (0.5%) is lower than expected by chance (based on LD of 100 random SNP sets with GWAS catalog SNPs: mean 3% overlap, sd 0.001, maximum 3%, minimum 2.5%). This may partially be due to the fact that most GWAS do not perform conditional analysis, and so the effect of these SNPs may be masked by primary GWAS hits, like we observed in the eQTL analysis.

Conditional eQTL analysis provides a more accurate GWAS SNP lookup

eQTL databases are often used to verify if an SNP identified in a GWAS (a GWAS SNP) is associated with gene expression. Due to the strong associations between SNPs and gene expression, a

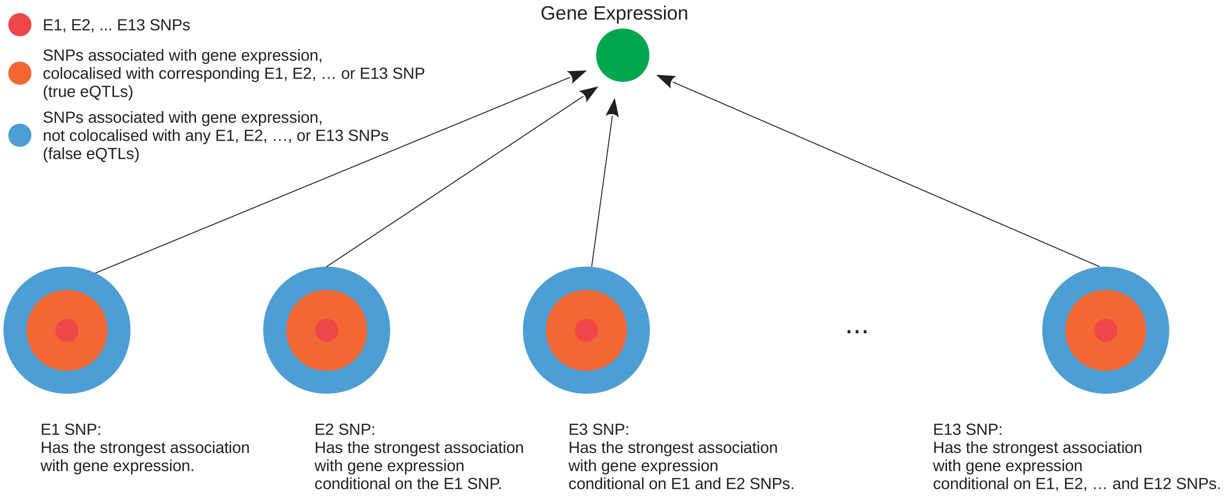
● E1, E2, ... E13 SNPs
● SNPs associated with gene expression, colocalised with corresponding E1, E2, … or E13 SNP (true eQTLs)
● SNPs associated with gene expression, not colocalised with any E1, E2, …, or E13 SNPs (false eQTLs)
Gene Expression
…
E1 SNP: Has the strongest association with gene expression.
E2 SNP: Has the strongest association with gene expression conditional on the E1 SNP.
E3 SNP: Has the strongest association with gene expression conditional on E1 and E2 SNPs.
E13 SNP: Has the strongest association with gene expression conditional on E1, E2, … and E12 SNPs.

Figure 2. Schematic overview of E1-E13 SNPs, SNPs colocalised with an E1, E2, . . ., or E13 SNP (true eQTLs), and SNPs not colocalised with any of the E1-E13 SNPs (false eQTLs).

significant association between the GWAS SNP and gene expression does not mean the SNP is a 'true eQTL', i.e. it may not tag the functional locus that regulates gene expression (the GWAS SNP and the eQTL are not 'colocalised' (16))· In many studies the LD is computed between the GWAS SNP and eQTLs to determine colocalisation of GWAS SNPs and eQTLs (2,9) but many other methods have been proposed (16–18). After conditional eQTL analysis, not only the colocalisation of the GWAS SNP and the E1 SNP can be determined (which is commonly done), but also of the GWAS SNP and the E2, E3, . . ., E13 SNPs. We demonstrate this by using LD computation as colocalisation method, but any other method can be used. If a GWAS SNP is in low LD with an E1 SNP, the association between the GWAS SNP and gene expression may be very significant, but solely caused by the low LD between the GWAS SNP and the E1 SNP. In this case, the association between the GWAS SNP and gene expression will disappear when conditioning on the E1 SNP. If the GWAS SNP is in strong LD (say $r^2 > 0.8$) with the E1 SNP, they are likely to tag the same functional locus that regulates gene expression, and we call it a 'true eQTL' (Fig. 2). The same applies to the E2, E3, . . ., or E13 SNPs. When performing a GWAS SNP lookup in our database, we propose the following procedure: 1) Verify if the GWAS SNP is associated with gene expression in the conditional or any of the unconditional eQTL analysis. 2) If so, select the corresponding E1, E2, . . ., or E13 SNP and use your favorite method (16–18) to determine colocalisation of GWAS SNPs and the E1, E2, ., or E13 SNP. 3) If the GWAS SNP is colocalised with the corresponding E1, E2, ., or E13 SNP, the GWAS SNP is likely to tag a functional locus regulating gene expression, and we call it a true eQTL. If not, we call it a false eQTL (Fig. 2).

To illustrate this procedure, we selected all SNPs from the GWAS catalog. From 11,966 GWAS SNPs, 3,132 (26%) are associated with gene expression in one of the eQTL analyses (unconditional, or conditional on E1, or E1 and E2, . . ., E12 SNPs). From these 3,132 GWAS SNPS, 902 (29%) are in LD ($r^2 > 0.8$) with an E1, E2, . . . or E12 SNP. So from the 3,132 GWAS SNPs associated with gene expression, we classify 902 (29%) as true eQTLs. From these 902 GWAS SNPs, 689 were in LD with an E1 SNP (identifiable with unconditional analysis) and 213 (24%) with an E2, E3, . . . , or E12 SNP (only identifiable with conditional analysis). Thus, without conditional eQTL analysis there would be 24% less GWAS SNPs identified as true eQTLs. Moreover, both E1 and E2 SNPs are enriched with GWAS SNPs (6.8% of the E1 SNPs, 4.1% of the E2 SNPs is in LD ($r^2 > 0.8$) with a GWAS SNP (enrichment $P < 0.01$ based on LD of 100 random SNP sets with GWAS catalog SNPs)). The conditional eQTL database we created has been used for follow up analysis of GWAS for metabolomics, fertility, heart rate variation, menarche, blood pressure and many others (11,25–27).

Conditional eQTLs contribute to SNP and family-based heritability of gene expression

The gene expression sample used for eQTL analysis consisted of 559 MZ twin pairs, 594 DZ twin pairs, 51 parent-sibling trios and 2437 unrelated subjects. This allowed estimation of the narrow sense heritability (h^2_{TOT}), and the portion of variance explained by common SNPs (h^2_{SNP}) and corresponding standard deviations in a single model (28), for each of the 44,241 gene expression probes (Supplementary Material, Table S2). For 1143 genes we found significant h^2_{SNP} (>0.14) and 5985 genes showed significant h^2_{TOT} (>0.07). From the 4039 probesets with two independent eQTLs, we estimated how much of h^2_{TOT} and h^2_{SNP} was explained by the primary cis eQTL, and how much by the primary and secondary cis eQTLs (Supplementary Material, Fig. S3). On average, the primary eQTL accounts for 23% of h^2_{TOT}, the primary and secondary eQTL together account for 31% of h^2_{TOT}. The primary eQTL explains on average 34% of h^2_{SNP} while the primary and secondary eQTL together explain 42% of h^2_{SNP}. Thus, although the primary cis eQTL explains more heritability than the secondary cis eQTL, the secondary cis eQTL increases the captured SNP and total heritability substantially, emphasizing the importance of not only considering primary, but also secondary cis eQTLs in variance or heritability decomposition.

Trans eQTL analysis

Gene expression corrected for all independent cis eQTL effects was subjected to trans eQTL analysis. eQTL effects were defined as trans when probeset–SNP pairs were at >5M base pairs (Mb). At an FDR of 5% ($P < 5e\text{-}10$, based on permutations similar to the cis eQTL analysis), for 434 probesets (targeting 267 genes) trans eQTLs were identified (138 unique trans eQTL SNPs, Supplementary Material, Table S3) after extensive QC (Sup. Methods). For all these trans eQTLs, associations were present for multiple probes per expression probeset (diminishing cross hybridization artifacts). Like in the cis eQTL analysis, expression was residualized with respect to the top trans eQTL, and trans eQTL analysis was repeated. For 129 probesets an additional independent trans eQTL was identified, and after another conditioning step, 39 probesets had three independent trans eQTLs (Supplementary Material, Table S3). After conditioning on up to nine independent trans eQTLs, no more trans eQTL effects were present. When two trans eQTL SNPs influence the same gene expression and are located on two different chromosomes or are located far away (>5 Mb) from each other, their independence is evident and conditional trans eQTL analysis is not necessary. For these cases, all trans eQTLs in the conditional analysis were also identified in unconditional analysis. Only a few trans eQTLs are located closer than 1Mb from each other and influence the same gene expression independently. This occurs at three loci on chromosome 6, 17 and 16, associated with in total 22 genes (Supplementary Material, Table S3). Without conditional analysis these cases could not have been differentiated from the cases where all SNPs at a locus represent only one signal. See Sup. Results for details on trans eQTL analysis, the role of (estimated) blood cell composition traits in eQTL analyses and the associations between CNV's and gene expression.

Discussion

In the conditional cis eQTL analysis, 35% of the cis-regulated gene expression appears to be allelic heterogeneous (14% of all genes), as indicated by multiple independent eQTLs per gene. The other 65% cis-regulated genes seem to be controlled by only one locus. Hundreds of cis eQTLs were only identified in conditional, and not in unconditional analysis. This finding has important implications: unconditional eQTL databases do not provide the complete picture. As an example, we highlighted an eQTL for *FAM117B* which was only identified in the conditional analysis and in strong LD with a GWAS hit for total cholesterol levels (24). The online conditional cis eQTL catalog we provide will help researchers to verify if a SNP controls gene expression more accurately than in standard eQTL databases.

Similarly as was reported previously (4,19), we identified multiple independent eQTL effects for 35% of the genes with an eQTL. Since secondary eQTLs are less significant than top eQTLs, this number is likely to increase with increases in sample size. If multiple eQTLs for a gene are associated with the same phenotype, and do so via the mediating gene expression, the accumulated signal in the gene expression may show a stronger association with the phenotype than the individual eQTLs. This makes gene expression, or the cis regulated component of gene expression, an interesting target for association studies (29,30). Besides revealing the dependency structure of eQTLs, the conditional eQTL analysis also identifies eQTLs which are not identified in the unconditional analysis (12%). This can be caused by a low positive correlation between the eQTLs, while having opposite effects on gene expression, or a negative correlation between them and the same direction of effect on gene expression. In trans eQTL analysis the independent eQTL effects are often located at different chromosomes, and uncorrelated, and therefore also identified in unconditional analysis.

In conditional eQTL analysis, gene expression is conditioned on the primary eQTL. If after this conditioning, gene expression is still associated with some other SNP(s), there are multiple possible scenarios: 1) there are multiple loci influencing gene expression independently 2) since we do not measure in one cell type but in whole blood, the two identified SNPs may each be active in a different cell type and 3) there is one locus influencing gene expressing, but this locus is not well tagged by the measured or imputed SNP, and the seemingly independent eQTL effects are both correlated with the functional locus and only reflect a single effect. With the current imputation resolution most functional loci are tagged and therefore the second scenario is unlikely to occur. Future eQTL studies using DNA sequence data, covering the complete genome, will be able to solve this issue.

We found that primary eQTLs (SNPs with the strongest effect on gene expression) are significantly overlapping with GWAS catalog SNPs, as was shown previously for all eQTLs identified in the unconditional analysis (2). Here, we showed that also secondary eQTLs (SNPs associated with gene expression after conditioning) are significantly overlapping with GWAS catalog SNPs, even though the power to detect secondary eQTLs in GWAS is probably lower: the secondary eQTL may also not be the strongest association in the GWAS. Or even only be identified after conditioning on the strongest association, which is not always done. In summary, conditional eQTL analysis increases the number of identified independent eQTLs, improves the look up of GWAS SNPs, and provides a better decomposition of gene expression heritability. This should be taken into account in future eQTL studies.

Materials and Methods

Subjects for eQTL analysis

The two parent projects that supplied data for the eQTL analysis are large-scale longitudinal studies: the Netherlands Study of Depression and Anxiety (NESDA) (21) and the Netherlands Twin Register (NTR) (22). NESDA and NTR studies were approved by the Central Ethics Committee on Research Involving Human Subjects of the VU University Medical Center, Amsterdam (institutional review board [IRB] number IRB-2991 under Federal wide Assurance 3703; IRB/institute codes: NESDA 03-183 and NTR 03-180). All participants provided written informed consent. The sample used for eQTL analysis consisted of 4,896 subjects with European ancestry (1,880 unrelated subjects from NESDA, 559 MZ twin pairs, 102 siblings of MZ twins (one per MZ twin pair), 594 DZ twin pairs, 111 siblings of DZ twins (one per DZ twin pair), 51 parent-sibling trios and 344 unrelated subjects from NTR). The age of the participants ranged from 17 to 88 years (mean = 38, *SD* = 13); 65% of the sample was female. The data used for this study largely overlaps with those used in our earlier study (1). We used all 4647 samples used by Wright *et al.*, plus 249 additional samples.

Blood sampling, RNA extraction, and RNA expression measurement

Study protocols and biological sample collection methods were harmonized between NTR and NESDA. RNA processing and measurements have been described in detail previously (1,31). Venous blood samples were drawn in the morning after an overnight fast. Heparinized whole blood samples were transferred within 20 min of sampling into PAXgene Blood RNA tubes (Qiagen, Valencia, California, USA) and stored at −20 °C. Gene expression assays were conducted at the Rutgers University Cell and DNA Repository. Samples were hybridized to Affymetrix U219 arrays (Affymetrix, Santa Clara, CA) containing 530,467 probes summarized in 49,293 probe sets. Array hybridization, washing, staining, and scanning were carried out in an

Affymetrix GeneTitan System per the manufacturer's protocol. Gene expression data were required to pass standard Affymetrix QC metrics (Affymetrix expression console) before further analysis. We excluded from further analysis proves that did not map uniquely to the hg19 (Genome Reference Consortium Human Build 37) reference genome sequence, as well as probes targeting a messenger RNA (mRNA) molecule resulting from a transcription of a DNA sequence containing a single nucleotide polymorphism (based on the dbSNP137 common database). After this filtering step, data for analysis remained for 423,201 probes, which could be summarized into 44,241 probe sets targeting 18,238 genes. Normalized probe set expression values were obtained using Robust Multi-array Average (RMA) normalization as implemented in the Affymetrix Power Tools software (APT, version 1.12.0, Affymetrix). Data for samples that displayed a low average Pearson correlation with the probe set expression values of other samples, and samples with incorrect sex-chromosome expression were removed, leaving 4,896 subjects for analysis.

Gene expression normalization

The inverse quantile normal transformation was applied for each expression probe set to obtain normal distributions. The transformed probeset data were then residualized by multiple linear regression with respect to the covariates sex, age, body mass index (kg/m^2), blood hemoglobin level, smoking status, several technical covariates (plate, well, hour of blood sampling, lab, days between blood sampling and RNA extraction and average correlation with other samples) and the scores on three principal components (PCs) as estimated from the imputed SNP genotype data (32) using the EIGENSOFT package. The residuals resulting from the linear regression analysis of the probe set intensity values onto the covariates listed above were subjected to a principal component analysis, with the aim to further filter out environmental variation from the data (33). For each principal component a genome-wide association study was performed, and the first 50 principal components without genome-wide significant SNP associations were removed from the residualized probeset data before eQTL analysis.

DNA extraction and SNP genotyping and imputation

DNA was extracted from peripheral blood as described previously (34). SNP genotype pre-imputation quality control, haplotype phasing and 1000 Genomes phase 1 imputation were performed as described previously (35). Imputed SNP genotypes were coded into the reference allele dosage format, and filtered at $\mathrm{MAF} > 0.01$ and HW $P > 1\mathrm{E}-04$ resulting in 8,158,830 remaining SNPs for eQTL analysis.

eQTL analysis and FDR based on permutations accounting for relatedness

eQTL effects were detected with a linear model approach using MatrixeQTL (36) with expression level as dependent variable and SNP genotype values as independent variable. To account for relatedness of the NTR subjects, permutations were performed where in each permutation the relatedness was preserved. In each permutation the genotypes of the MZ twin pairs were assigned the expression of a random MZ twin pair, the genotypes of the DZ twin pairs were assigned the expression of a random DZ twin pair, the genotypes of the MZ twin pairs with sibling were assigned the expression of a random MZ twin pair with sibling, the genotypes of the parent-sibling trios were assigned the expression of a random parent-sibling trios and the genotypes of the unrelated subjects were assigned the expression of a random subject from the group of unrelated subjects. For each permutation the complete cis eQTL analysis was repeated, and after each permutation the P-value threshold for rejecting at $\mathrm{FDR} < 0.05$ was computed. This can be done in two ways: 1) divide the total number of significant eQTLs in the permuted data by the total number of significant eQTLs in the unpermuted data (=false positives/true positives) or 2) divide the total number of probesets with a significant eQTL in the permuted data by the total number of probesets with a significant eQTLs in the unpermuted data. We used the second method which is more conservative and was proposed earlier (33) to account for large LD blocks with strong eQTL effects that inflate the FDR when using the first method. Similar as what was observed in Fehrman *et al.*, only 10 permutations were needed to have the P-value threshold corresponding to $\mathrm{FDR} < 5\%$ converging. Of note, the eQTL P-values reported in this manuscript are based on the complete sample with related subject and thus are too liberal: however the FDR takes into account the family structure and should be used to draw conclusions. The reported betas from the linear models can be correctly estimated from samples containing related subjects. For the conditional eQTL analysis, the same P-values threshold was used as for the unconditional analysis, in order to keep the threshold fixed across analysis. The conditional analysis consists of a new group of tests, and has to be considered separately from the unconditional analysis in terms of significance. During conditional analysis much less tests were performed compared to the unconditional analysis, so when using the same P-value threshold, we are sure to be conservative when stating $\mathrm{FDR} < 5\%$ for the conditional analysis.

eQTL effects were defined as *cis* when probe set–SNP pairs were at distance < 1M base pairs (Mb). For each probe set that displayed a statistically significant association with at least one SNP in the *cis* region, we identified the most significantly associated SNP (E1 SNP). Conditional eQTL analysis was carried out by first residualizing probeset expression using the corresponding E1 SNP and then repeating the eQTL analysis using the residualized data. Then, for each probe set the most significant SNP was selected (E2 SNP) and each probeset was residualized using the E1 and E2 SNPs, and eQTL analysis was repeated using the residualized expression. This was repeated until no more significant associations were found between residualized expression and SNP data (after up to 12 rounds of conditional analysis). We call the E1-E13 SNPs independent eQTL SNPs. So we define 'independent eQTL SNPs' as E1-E*N* SNPs ($N = 2$–13), for which the E*i* SNP is significantly associated with gene expression, while conditioning on the E1-E(*i*-1) SNP(s). This does not mean that LD between independent eQTL SNPs is close to zero: as we show in the results, independent eQTL SNPs can be significantly correlated.

Initially, the first 50 principal components without genome-wide significant SNP associations were removed from the residualized probeset data before eQTL analysis. Post hoc analysis was performed by repeating the eQTL analysis and correcting for all 50 principal components (PCs). From all top cis eQTLs identified in the initial analysis, 97% were still significant after removing all 50 PCs, using the same threshold for significance $P < 1\mathrm{e}{-5}$ and 99.7% when using $P < 1\mathrm{e}{-4}$ as a threshold.

SNP and family-based heritability of gene expression

Subjects were genotyped on the Affymetrix 6k chip. SNPs that passed basic QC, with Hardy Weinberg equilibrium *P*-value below 10e-5 and minor allele frequency above 0.01 were included in the following analysis. A genetic relatedness matrix (GRM) was computed based on genotyped SNPs using GCTA version 1.24.2 (37). As the sample used for this study contains a substantial proportion of closely related individuals, the total heritability and the variance attributable to SNPs can be estimated concurrently (28). To enable estimation of both the total additive genetic variance and the variance attributable to measured SNPs concurrently we constructed a second GRM ($GRM_{>0.05}$), all values above 0.05 are copied from the GRM to the $GRM_{>0.05}$, values below 0.05 are substituted with 0. GRM and $GRM_{>0.05}$ contain respectively all genetic relationships in the sample, and the close (i.e. familial) genetic relationships in the sample. We fitted the variance model as proposed by Zaitlen *et al.* (2013) to each of the probesets. This model partitions the total variance in a probeset into the variance attributable to SNPs (h^2_{SNP}) the narrow sense heritability (h^2_{TOT}) and the residual variation, i.e. variation not attributable to genetic effects. From these models standard deviations of h^2_{TOT} and h^2_{SNP} were computed to determine significance of heritability.

Data Submission

Full conditional eQTL results are accessible at https://eqtl.onderzoek.io/. Gene expression and genotype data used for this study are available at dbGaP, accession number phs000486.v1.p1 (http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000486.v1.p1).

Supplementary Material

Supplementary Material is available at *HMG* online.

Conflict of Interest statement. None declared.

Funding

The Netherlands Organization for Scientific Research (MagW/ZonMW grants 904-61-090, 985-10-002,904-61-193,480-04-004, 400-05-717, 912-100-20; Spinozapremie 56-464-14192; Geestkracht program grant 10-000-1002); the Center for Medical Systems Biology (NWO Genomics), Biobanking and Biomolecular Resources Research Infrastructure (BBMRI-NL), VU University's Institutes for Health and Care Research (EMGO$^+$) and Neuroscience Campus Amsterdam and the European Science Council (ERC, 230374), The US National Institute of Mental Health (RC2 MH089951) as part of the American Recovery and Reinvestment Act of 2009; Biobanking and Biomolecular Research Infrastructure, R.J. and M.V were funded by BBMRI-NL, NWO project 184.021.007.

References

1. Wright, F.A., Sullivan, P.F., Brooks, A.I., Zou, F., Sun, W., Xia, K., Madar, V., Jansen, R., Chung, W., Zhou, Y.H., *et al.* (2014) Heritability and genomics of gene expression in peripheral blood. *Nat. Genet.*, **46**, 430–437.
2. Westra, H.J., Peters, M.J., Esko, T., Yaghootkar, H., Schurmann, C., Kettunen, J., Christiansen, M.W., Fairfax, B.P., Schramm, K., Powell, J.E., *et al.* (2013) Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.*, **45**, 1238–1243.
3. Battle, A., Mostafavi, S., Zhu, X., Potash, J.B., Weissman, M.M., McCormick, C., Haudenschild, C.D., Beckman, K.B., Shi, J., Mei, R., *et al.* (2014) Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res.*, **24**, 14–24.
4. Lappalainen, T., Sammeth, M., Friedländer, M.R., Hoen, P.A.C., 't, Monlong, J., Rivas, M.A., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P.G., *et al.* (2013) Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, **501**, 506–511.
5. Ramasamy, A., Trabzuni, D., Guelfi, S., Varghese, V., Smith, C., Walker, R., De, T., UK Brain Expression Consortium, North American Brain Expression Consortium, Coin, L., *et al.* (2014) Genetic variability in the regulation of gene expression in ten regions of the human brain. *Nat. Neurosci.*, **17**, 1418–1428.
6. Kim, Y., Xia, K., Tao, R., Giusti-Rodriguez, P., Vladimirov, V., van den Oord, E. and Sullivan, P.F. (2014) A meta-analysis of gene expression quantitative trait loci in brain. *Transl. Psychiatry*, **4**, e459.
7. GTEx Consortium, (2015) Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*, **348**, 648–660.
8. den Hoed, M., Eijgelsheim, M., Esko, T., Brundel, B.J.J.M., Peal, D.S., Evans, D.M., Nolte, I.M., Segrè, A.V., Holm, H., Handsaker, R.E., *et al.* (2013) Identification of heart rate–associated loci and their effects on cardiac conduction and rhythm disorders. *Nat. Genet.*, **45**, 621–631.
9. Schizophrenia Working Group of the Psychiatric Genomics Consortium (2014) Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, **511**, 421–427.
10. Albert, F.W. and Kruglyak, L. (2015) The role of regulatory variation in complex traits and disease. *Nat. Rev. Genet.*, **16**, 197–212.
11. Vaez, A., Jansen, R., Prins, B.P., Hottenga, J.J., de Geus, E.J.C., Boomsma, D.I., Penninx, B.W.J.H., Nolte, I.M., Snieder, H. and Alizadeh, B.Z. (2015) In Silico Post Genome-Wide Association Studies Analysis of C-Reactive Protein Loci Suggests an Important Role for Interferons. *Circ. Cardiovasc. Genet.*, **8**, 487–497.
12. Musunuru, K., Strong, A., Frank-Kamenetsky, M., Lee, N.E., Ahfeldt, T., Sachs, K.V., Li, X., Li, H., Kuperwasser, N., Ruda, V.M., *et al.* (2010) From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature*, **466**, 714–719.
13. Gauvreau, G.M., O'Byrne, P.M., Boulet, L.P., Wang, Y., Cockcroft, D., Bigler, J., FitzGerald, J.M., Boedigheimer, M., Davis, B.E., Dias, C., *et al.* (2014) Effects of an anti-TSLP antibody on allergen-induced asthmatic responses. *N. Engl. J. Med.*, **370**, 2102–2110.
14. Battle, A., Khan, Z., Wang, S.H., Mitrano, A., Ford, M.J., Pritchard, J.K. and Gilad, Y. (2015) Genomic variation. Impact of regulatory variation from RNA to protein. *Science*, **347**, 664–667.
15. Li, J.J. and Biggin, M.D. (2015) Gene expression. Statistics requantitates the central dogma. *Science*, **347**, 1066–1067.
16. Giambartolomei, C., Vukcevic, D., Schadt, E.E., Franke, L., Hingorani, A.D., Wallace, C. and Plagnol, V. (2014) Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLOS Genet.*, **10**, e1004383.
17. Wallace, C. (2013) Statistical testing of shared genetic control for potentially related traits. *Genet. Epidemiol.*, **37**, 802–813.
18. Nica, A.C., Montgomery, S.B., Dimas, A.S., Stranger, B.E., Beazley, C., Barroso, I. and Dermitzakis, E.T. (2010) Candidate Causal Regulatory Effects by Integration of

Expression QTLs with Complex Trait Genetic Associations. *PLOS Genet.*, **6**, e1000895.

19. Fu, J., Wolfs, M.G.M., Deelen, P., Westra, H.J., Fehrmann, R.S.N., Te Meerman, G.J., Buurman, W.A., Rensen, S.S.M., Groen, H.J.M., Weersma, R.K., *et al.* (2012) Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression. *PLoS Genet.*, **8**, e1002431.
20. Ma, B., Huang, J. and Liang, L. (2014) RTeQTL: Real-time online engine for expression quantitative trait loci analyses. *Database J. Biol. Databases Curation*, **2014**, bau066. doi:10.1093/database/bau066.
21. Penninx, B.W.J.H., Beekman, A.T.F., Smit, J.H., Zitman, F.G., Nolen, W.A., Spinhoven, P., Cuijpers, P., De Jong, P.J., Van Marwijk, H.W.J., Assendelft, W.J.J., *et al.* (2008) The Netherlands Study of Depression and Anxiety (NESDA): rationale, objectives and methods. *Int. J. Methods Psychiatr. Res.*, **17**, 121–140.
22. Boomsma, D.I., de Geus, E.J.C., Vink, J.M., Stubbe, J.H., Distel, M.A., Hottenga, J.J., Posthuma, D., van Beijsterveldt, T.C.E.M., Hudziak, J.J., Bartels, M., *et al.* (2006) Netherlands Twin Register: from twins to twin families. *Twin. Res. Hum. Genet.*, **9**, 849–857.
23. Zhernakova, D.V., Deelen, P., Vermaat, M., van Iterson, M., van Galen, M., Arindrarto, W., van 't Hof, P., Mei, H., van Dijk, F., Westra, H.J., *et al.* (2017) Identification of context-dependent expression quantitative trait loci in whole blood. *Nat. Genet.*, **49**, 139–145.
24. Global Lipids Genetics Consortium (2013) Discovery and refinement of loci associated with lipid levels. *Nat. Genet.*, **45**, 1274–1283.
25. Draisma, H.H.M., Pool, R., Kobl, M., Jansen, R., Petersen, A.K., Vaarhorst, A.A.M., Yet, I., Haller, T., Demirkan, A., Esko, T., *et al.* (2015) Genome-wide association study identifies novel genetic variants contributing to variation in blood metabolite levels. *Nat. Commun.*, **6**, 7208.
26. Winkler, T.W., Justice, A.E., Graff, M., Barata, L., Feitosa, M.F., Chu, S., Czajkowski, J., Esko, T., Fall, T., Kilpeläinen, T.O., *et al.* (2015) The Influence of Age and Sex on Genetic Associations with Adult Body Size and Shape: A Large-Scale Genome-Wide Interaction Study. *PLOS Genet.*, **11**, e1005378.
27. Cousminer, D.L., Stergiakouli, E., Berry, D.J., Ang, W., Groen-Blokhuis, M.M., Körner, A., Siitonen, N., Ntalla, I., Marinelli, M., Perry, J.R.B., *et al.* (2014) Genome-wide association study of sexual maturation in males and females highlights a role for body mass and menarche loci in male puberty. *Hum. Mol. Genet.*, **23**, 4452–4464.
28. Zaitlen, N., Kraft, P., Patterson, N., Pasaniuc, B., Bhatia, G., Pollack, S. and Price, A.L. (2013) Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits. *PLoS Genet.*, **9**, e1003520.
29. Gamazon, E.R., Wheeler, H.E., Shah, K.P., Mozaffari, S.V., Aquino-Michaels, K., Carroll, R.J., Eyler, A.E., Denny, J.C., GTEx Consortium, Nicolae, D.L., *et al.* (2015) A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.*, **47**, 1091–1098.
30. Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B.W.J.H., Jansen, R., de Geus, E.J.C., Boomsma, D.I., Wright, F.A., *et al.* (2016) Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.*, **48**, 245–252.
31. Jansen, R., Batista, S., Brooks, A.I., Tischfield, J.A., Willemsen, G., van Grootheest, G., Hottenga, J.J., Milaneschi, Y., Mbarek, H., Madar, V., *et al.* (2014) Sex differences in the human peripheral blood transcriptome. *BMC Genomics*, **15**, 33.
32. Abdellaoui, A., Hottenga, J.J., Knijff, P., de, Nivard, M.G., Xiao, X., Scheet, P., Brooks, A., Ehli, E.A., Hu, Y., Davies, G.E., *et al.* (2013) Population structure, migration, and diversifying selection in the Netherlands. *Eur. J. Hum. Genet.*, **21**, 1277–1285.
33. Fehrmann, R.S.N., Jansen, R.C., Veldink, J.H., Westra, H.J., Arends, D., Bonder, M.J., Fu, J., Deelen, P., Groen, H.J.M., Smolonska, A., *et al.* (2011) Trans-eQTLs Reveal That Independent Genetic Variants Associated with a Complex Phenotype Converge on Intermediate Genes, with a Major Role for the HLA. *PLoS Genet.*, **7**, e1002197.
34. Boomsma, D.I., Willemsen, G., Sullivan, P.F., Heutink, P., Meijer, P., Sondervan, D., Kluft, C., Smit, G., Nolen, W.A., Zitman, F.G., *et al.* (2008) Genome-wide association of major depression: description of samples for the GAIN Major Depressive Disorder Study: NTR and NESDA biobank projects. *Eur. J. Hum. Genet. EJHG*, **16**, 335–342.
35. Nivard, M.G., Mbarek, H., Hottenga, J.J., Smit, J.H., Jansen, R., Penninx, B.W., Middeldorp, C.M. and Boomsma, D.I. (2014) Further confirmation of the association between anxiety and CTNND2: replication in humans. *Genes Brain Behav.*, **13**, 195–201.
36. Shabalin, A.A. (2012) Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinforma. Oxf. Engl.*, **28**, 1353–1358.
37. Yang, J., Lee, S.H., Goddard, M.E. and Visscher, P.M. (2013) Genome-wide complex trait analysis (GCTA): methods, data analyses, and interpretations. *Methods Mol. Biol. Clifton NJ.*, **1019**, 215–236.