Published online 21 February 2013 *Nucleic Acids Research, 2013, Vol. 41, No. 7* ***e88***
doi:10.1093/nar/gkt069

Resolving the polymorphism-in-probe problem is critical for correct interpretation of expression QTL studies

Adaikalavan Ramasamy[1,2], Daniah Trabzuni[2,3], J. Raphael Gibbs[2,4], Allissa Dillman[4], Dena G. Hernandez[2,4], Sampath Arepalli[4], Robert Walker[5], Colin Smith[5], Gigaloluwa Peter Ilori[1], Andrey A. Shabalin[6], Yun Li[6,7], Andrew B. Singleton[4], Mark R. Cookson[4] for NABEC[#], John Hardy[2] for UKBEC, Mina Ryten[2,*] and Michael E. Weale[1,*]

[1]Department of Medical & Molecular Genetics, King's College London, 8th Floor, Tower Wing, Guy's Hospital, London SE1 9RT, UK, [2]Reta Lila Weston Institute and Department of Molecular Neuroscience, UCL Institute of Neurology, Queen Square, London WC1N 3BG, UK, [3]Department of Genetics, King Faisal Specialist Hospital and Research Centre, PO Box 3354, Riyadh 11211, Saudi Arabia, [4]Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, MD, USA, [5]MRC Sudden Death Brain Bank Project, University of Edinburgh, Department of Neuropathology, Wilkie Building, Teviot Place, Edinburgh EH8 9AG, USA, [6]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA and [7]Department of Genetics, Department of Biostatistics, Department of Computer Science, University of North Carolina, Chapel Hill, NC 27599, USA

Received October 24, 2012; Revised January 16, 2013; Accepted January 17, 2013

ABSTRACT

Polymorphisms in the target mRNA sequence can greatly affect the binding affinity of microarray probe sequences, leading to false-positive and false-negative expression quantitative trait locus (QTL) signals with any other polymorphisms in linkage disequilibrium. We provide the most complete solution to this problem, by using the latest genome and exome sequence reference data to identify almost all common polymorphisms (frequency >1% in Europeans) in probe sequences for two commonly used microarray panels (the gene-based Illumina Human HT12 array, which uses 50-mer probes, and exon-based Affymetrix Human Exon 1.0 ST array, which uses 25-mer probes). We demonstrate the impact of this problem using cerebellum and frontal cortex tissues from 438 neuropathologically normal individuals. We find that although only a small proportion of the probes contain polymorphisms, they account for a large proportion of apparent expression QTL signals, and therefore result in many false signals being declared as real. We find that the polymorphism-in-probe problem is insufficiently controlled by previous protocols, and illustrate this using some notable false-positive and false-negative examples in MAPT and PRICKLE1 that can be found in many eQTL databases. We recommend that both new and existing eQTL data sets should be carefully checked in order to adequately address this issue.

INTRODUCTION

Expression quantitative trait locus (eQTL) studies look for association signals between genetic variation (typically single nucleotide polymorphisms, or SNPs) and gene expression. Here, we use 'eQTL' to refer to all kinds of expression quantitative trait loci, whether arising from association with gene-level or exon-level expression patterns. These eQTL studies have provided insights into the mode and regulatory action of gene-level expression and the differential expression of alternatively spliced transcripts, and have provided important insights into the causal mechanisms behind some genome-wide association study signals (1–3).

It is anticipated that RNA-seq will become the future platform of choice for these studies. However, the protocols for this technology are still immature, and the costs for assaying large numbers of individuals are still high.

*To whom correspondence should be addressed. Tel: +44 20 7188 2601; Fax: +44 20 7188 2585; Email: michael.weale@kcl.ac.uk
Correspondence may also be addressed to Mina Ryten. Tel: +44 20 7679 4297; Fax: +44 20 3448 4017; Email: mina.ryten@ucl.ac.uk
[#]NABEC investigators listed in Acknowledegements.

For now, older platforms relying on microarrays remain important, not least because large and valuable repositories of data exist based on this technology. Expression microarrays work through the binding of oligonucleotide probe sequences to expressed mRNA. Two widely used examples are the Illumina Human HT12 array, which uses 50-mer probes biased towards the 3′ end of mRNA transcripts to estimate whole-gene levels of expression, and the Affymetrix Human Exon 1.0 ST array, which uses 25-mer probes, typically grouped in sets of four per exon, to estimate exon-specific levels of expression.

All microarrays are susceptible to a polymorphism-in-probe problem, which arises because probes are typically designed to match one reference sequence only. Thus sequences which depart from this reference, either due to the presence of different nucleotides (i.e. SNPs) or (particularly) due to the presence or absence of nucleotides (i.e. indels), are likely to exhibit a weaker binding affinity for the probe in question (4). This results in an apparent association between genotype and expression, confounding eQTL studies that are looking for just such a signal. Furthermore, this problem will not only generate a false eQTL signal with the polymorphism in the probe sequence, but also localized linkage disequilibrium (LD) will ensure that all polymorphisms in high LD with the offending polymorphism will likewise display a false association signal.

We note that an analogous problem can arise in RNA sequence-based eQTL studies. Allele-specific biases may occur when aligning RNA sequence reads to a single reference genome. Addressing these biases is an active line of enquiry in the field of allele-specific expression studies (5–7). We anticipate that aligning reads to personal genomes (e.g. via exome sequencing) will provide the best solution to this problem in the context of RNA sequencing.

Previous microarray-based eQTL studies have dealt with the polymorphism-in-probe problem with varying degrees of thoroughness and with considerable differences both in how investigators sought to identify suspect probes and also in how they then chose to remove suspect eQTL signals based on this information (see Supplementary Table S1 for selected examples). Several studies have attempted to quantify this problem empirically, either explicitly or as part of a biological result paper (Table 1), but again they vary in protocol (8–14). Several factors are likely to have led investigators to underestimate the scale of this issue. These include incomplete reference information on the location of all common SNPs and indels, associated difficulties in applying high quality imputation techniques to enable the prediction of non-genotyped SNPs and a lack of appreciation for the possibility of false associations due not only to genotyped polymorphisms within the probe sequences, but also polymorphisms in LD located outside the probe sequence and other polymorphisms like indels.

This article aims to deal with this problem comprehensively. First, using the most recent releases of the 1000 Genomes (March 2012) and NHLBI Exome Sequencing Project (NHLBI-ESP) together with data generated on two popular platforms, we provide the most comprehensive method to date for the identification and removal of suspect eQTL signals due to the polymorphism-in-probe problem. Second, we conduct a systematic investigation of the effect of the signal removal protocol on the quality of downstream eQTL signals. Third, we consider and evaluate the available solution for this problem. And finally, we provide some guidance and a website for users on how to identify probes that may contain polymorphisms.

MATERIALS AND METHODS

Data source

To demonstrate the extent of this problem, we used data from two consortia that have genotyped and expression profiled human cerebellum (CRBL) and frontal cortex (FCTX) from neuropathologically normal individuals using two popular platforms. Details on the data set generation and characteristics are given in the Supplementary Methods and summarized in Supplementary Table S2. Briefly, the Illumina HT12 data set consists of 304 individuals profiled by the North American Brain Expression Consortium (NABEC) (15–17), whereas the Affymetrix Human Exon data set consists of 134 individuals profiled by the UK Brain Expression Consortium (UKBEC) (18). In terms of probe design, the Illumina HT12-v3 BeadChip Array uses 50-mer probes whereas the Affymetrix Human Exon 1.0 ST array uses sets of 25-mer probes designed to target individual exons (usually 4 probes per set), the basic unit of expression in this case. These two arrays are used in the large majority of expression QTL studies to date (Supplementary Table S1).

The two consortia used different genome-wide genotyping arrays but both imputed additional markers (~5.8 million SNPs) using the 1000 Genomes (March 2012) data, thereby improving the coverage in SNPs between both data sets for eQTL analysis. The eQTL analysis was restricted to the autosomal regions of the genome for expression and genotype data.

Expression QTL analysis and LD-resolved signal identification

We tested the association between each SNP and each expression profile assuming an additive genetic model for SNPs. The computation was done using MatrixEQTL software (19) and R (http://www.r-project.org/) on a high performance linux-based computer cluster.

The process of imputation and natural LD across the genome, while useful to identify causal variants, does create a problem in that eQTLs from SNP-rich high LD regions would be represented several times by LD proxy. Therefore, we treated multiple associations for a given probe/probeset as a single signal if the associated SNPs were in pairwise LD of $r^2 > 0.5$ with each other, and the SNP with the smallest *P*-value as the ‘LD-resolved’ eQTL.

We consider an eQTL signal as *cis*-acting if the hit SNP is located within 1 Mb of the transcription start site of the associated transcript.

Table 1. Studies that have provided an empirical assessment of the polymorphism-in-probe problem

Article (PMID)	Tissues and sample size	Expression chip (probe length)	SNP set used to check SNP-in-probe (# SNPs in set)	Method of assessment and reported findings
Walter *et al.* (2007)[METHOD] Nature Methods PMID: 17762873	Whole brain from six C57BL/6J strain mice and six DBA/2J strain mice	Affymetrix MOE430 2.0 chips (25-mer probes but only transcript-level was analysed)	NIEHS/Perlegen Mouse Resequencing Project & Mouse Phenome Database SNP Tool & Sanger resequencing (~3.9 m SNPs)	Compared results before and after masking SNP-containing probes. 22% false-positive rate and 12% false-negative rates (RMA) or 36% false-positive rate and 13% false-negative rates (MAS 5.0)
Meyers *et al.* (2007) Nature Genetics PMID: 17982457	193 neuropathologically normal human brains (pooled regions)	Illumina Human Refseq-8 Expression (50-mer probes)	Genotyped SNPs (366 140 SNPs)	Discarded associations if probe contained a SNP 13% of significant *cis*-eQTLs discarded 5% of significant *trans*-eQTLs discarded
Benovoy *et al.* (2008)[METHOD] Nucleic Acids Research PMID: 18596082	57 CEU HapMap individuals, LCLs	Affymetrix Human Exon 1.0 ST (25-mer probes)	HapMap II release 21 (~ 4 million SNPs)	Compared results before and after masking SNP-containing probes. 86.6% false-positive rate and 0.3% false-negative rate (exon-level) 8.1% false-positive rate and 0.05% false-negative rate (gene-level)
Heinzen *et al.* (2008) PLoS Biology PMID: 19222302	93 frontal cortex 80 blood cell	Affymetrix Human Exon 1.0 ST (25-mer probes)	Genotyped SNPs (<550 thousand SNPs)	Discarded associations if the hit SNP was inside the probe sequence or in high LD ($r^2 > 0.50$) with a SNP inside the probe sequence 36.6% of significant *cis*-eQTLs (exon-level) discarded
Gamazon *et al.* (2010)[METHOD] PLoS One PMID: 20186275	57 CEU HapMap individuals 56 YRI HapMap individuals, LCLs	Affymetrix Human Exon 1.0 ST (25-mer probes)	1000 Genomes Pilot 1 (April 2009) + dbSNP v129 (unclear on number of SNPs)	Focused on 782 differentially spliced probesets from their previous published study and reports that ~15% of these could be affected by novel SNPs in 1000Genomes Pilot 1 (compared with dbSNP v129).
Stranger *et al.* (2012) PLoS Genetics PMID: 22532805	726 individuals from 8 HapMap populations, LCLs	Illumina Sentrix Human-6 Expression BeadChip version 2 (50-mer probes)	1000 Genomes Pilot 1 (Aug 2010) with MAF > 5% (unclear on number of SNPs)	6.5% of probes contained SNP(s) within the probe sequence while 7.4% of the significant probes (i.e. has at least one significant *cis*-eQTL) also contained SNP(s). Therefore, concluded no significant enrichment.
Ramasamy *et al.* (2013)[METHOD] (current article)	130 cerebellum 127 frontal cortex	Affymetrix Human Exon 1.0 ST (25-mer probes)	1000 Genomes Integrated Phase 1 version 3 (March 2012) and NHLBI-ESP (~9.3 million SNPs ~1 million indels)	Proportion of *cis* eQTLs discarded (depending on *P*-value): 60.2–90% in FCTX and 49.7–72.7% in CRBL
	301 cerebellum 304 frontal cortex	Illumina HT12 (50-mer probes)		31–52.6% in FCTX and 20–46.7% in CRBL

[METHOD] Indicates methodological articles that explicitly studied this problem in greater detail.

Polymorphism reference data sources and identification of polymorphism-containing probes

We define a 'suspect *cis*-eQTL' to be any *cis*-eQTL signal where the relevant probe contains a polymorphism with a minor allele frequency >1% in Europeans, regardless of the LD between the hit SNP and the polymorphism-in-probe. To identify probes containing polymorphisms, we considered several different genetic variation reference data sets, which differ in their completeness. The smallest is the set of SNPs available on the genotyping chip (Illumina HumanHap550 for the NABEC data set and Illumina Omni-1 M Quad for the UKBEC data set). We then considered the CEU panel of the final release of HapMap (release #28, merged Phase I + II + III data), although one should note that most of the studies listed in Supplementary Table S1 used earlier versions of HapMap. Next, we considered the SNP and indel data of the European panel ($n = 381$) of the latest version of the 1000 Genomes Project (March 2012: Integrated Phase I haplotype release version 3, based on the 2010–11 data freeze and 2012-03-14 haplotypes). Finally, we considered the SNPs (average read depth ≥ 10) from the Exome Variant Server, NHLBI-ESP, Seattle, WA (URL: http://evs.gs.washington.edu/EVS/) (accessed 11 May 2012), taken from 3510 European Americans drawn from multiple ESP cohorts. In all reference data sources, we restricted to polymorphisms that were identified with at least 1% allele frequency in European descent samples.

The list of probes and probesets used in Affymetrix Exon 1.0 ST and Illumina HT12 in this article along with the positions of the polymorphism-in-probe (if any) is given in Supplementary Table S5.

Probe masking for Affymetrix Exon 1.0 ST Array data

Affymetrix probes are grouped into probesets of typically four probes, which measure the expression of a given exon. If one of the four probes contains a polymorphism and three good one remain, we re-estimated the exon signal from the remaining three. We refer to these as 'altered' probesets. If less than three probes remain (either because more than one probe has a polymorphism or because of other QC-related drop out of probes) then the remaining information was considered insufficient and the probeset was discarded. Masking was done using Affymetrix Power Tools (see Supplementary Table S2 for codes). Probe masking has the advantage that both false-positive and false-negative eQTL signals can be recovered.

Conditional analysis for rescuing suspect *cis*-eQTLs from discarded probes/probesets

We applied conditional analysis by including the genotype dosage (number of minor alleles in the genotype) of the polymorphism in probe as a covariate in the linear model regressing the expression of a discarded probe/probeset against the SNP of interest. Multiple covariates were used if more than one polymorphism in the probe or probeset was found. We note that this method can in principle correct both false-positive signals (where the only signal is from the polymorphism-in-probe) and false-negative signals (where the polymorphism-in-probe counteracts the true signal). However, unlike probe masking, true signals can only be recovered if the truly associated SNP is in low LD with the polymorphism in the probe. High-LD SNPs are irretrievably confounded and unrecoverable by this method. Indeed, any SNPs that are in perfect LD with the corresponding polymorphism-in-probe will fail to fit in the conditional model, and must be assigned a conditional association *P*-value of 1 regardless of whether they are a true hit or not. The method also requires that the polymorphism-in-probe genotype be known for all individuals in the eQTL study (either via imputation or more directly via sequencing).

LD filtering for rescuing suspect *cis*-eQTLs from discarded probes/probesets

We applied LD filtering by choosing an arbitrary threshold for pairwise LD between the SNP of interest and the polymorphism-in-probe, to rescue *cis*-eQTLs with low LD. In contrast to conditional analysis, LD values can be obtained directly from the reference data source and therefore knowledge of the polymorphism-in-probe genotype for individuals in the eQTL data set is not required. We note that this method can only rescue false-positive signals, not false-negative ones, and furthermore, the rescued signals still carry some probability of being false positives via LD (and indeed we shall show this probability remains high even for very stringent LD thresholds).

An efficient approach to identifying probes containing polymorphism

The start and stop positions of probes from commercial arrays are generally available from the microarray chip manufacturer's websites. The positions of the variants are available from the latest releases of public projects such as the 1000 Genomes or NHLBI-ESP or other in-house sequencing projects. After obtaining this, one could then scan for overlapping variants in between the start and stop positions of every probe. Although this can be coded in many ways, we found the intersectBED tool, which uses the concept of an interval tree from the BEDtools suite (20), to be efficient. For example, it took approximately 3 s to search through 6 million SNPs and indels for 5000 probes. The codes for implementing this are given in Supplementary Methods. Special care is required when dealing with insertion polymorphisms. A user-friendly implementation (PiP Finder) is available at http://bit.ly/pipfinder using the final variation set defined here.

RESULTS

Proportion of probes (and probesets) containing polymorphism(s) in probe sequence

Using different reference data sources for defining polymorphisms, we identified the number of probes/probesets affected by the polymorphism-in-probe problem in both

Table 2. Classification of probes /probesets in both data sets with progressively more comprehensive polymorphism reference data source

		Affymetrix Human Exon 1.0 ST (~1.2 million 25-mer probes grouped into 298 k probesets) based on the UK Human Brain Expression Consortium (UKBEC, N = 134)				Illumina Human HT12 (43 009 50-mer probes) based on the North American Brain Expression Consortium (NABEC, N = 304)		
Polymorphism reference data source (restricted to autosomes and frequency > 1%)	No. of variants	No. of unique variants in probe sequence	No. of core probesets unaltered	No. of core probesets altered (%)	No. of core probesets discarded (%)	No. of unique variants in probe sequence	No. of probes unaltered	No. of probes discarded (%)
Illumina Infinium HumanHap550 (after QC)	512 771 SNPs					362 SNPs	42 638	371 (0.86%)
Illumina Omni 1M (after QC)	795 391 SNPs	20 926 SNPs	278 585	13 515 (4.5%)	6260 (2.1%)			
CEU panel of HapMap release 28 (August 2010) [unrelated N = 60 (Phase I/II), 112 (Phase III)]	2 602 611 SNPs	24 911 SNPs	275 010	16 162 (5.4%)	7188 (2.4%)	1557 SNPs	41 448	1561 (3.6%)
SNPs from European panel of 1000 Genomes Integrated Phase 1 version 3 (March 2012) (N = 381)	9 013 135 SNPs	50 813 SNPs	254 277	28 932 (9.7%)	15 151 (5.1%)	5186 SNPs	38 356	4653 (10.8%)
+ SNPs info from European Americans from the NHLBI-ESP (N = 3,510)	9 025 738 SNPs	52 843 SNPs	252 692	29 808 (10.0%)	15 860 (5.3%)	5361SNPs	38 243	4766 (11.1%)
+ indels from European panel of 1000 Genomes Integrated Phase 1 version 3 (March 2012) (N = 381)	9 025 738 SNPs + 927 779 indels	52 843 SNPs + 2097 indels	251 313	30 621 (10.3%)	16 426 (5.5%)	5361 SNPs + 332 indels	37 993	5016 (11.7%)

Table 3. Number of LD-resolved *cis*-eQTLs ($P < 10^{-12}$) for the two data sets, using polymorphisms (present with minor allele frequency >1% in Europeans) from the combined 1000 Genomes (March 2012) plus NHLBI Exome Sequence Project data sources

Affymetrix Human Exon 1.0 (25-mer probe design) based on the UK Human Brain Expression Consortium (UKBEC, $N = 134$)			Illumina Human HT-12v3 (50-mer probe design) based on the North American Brain Expression Consortium (NABEC, $N = 304$)		
	CRBL	FCTX		CRBL	FCTX
Total number of *cis*-eQTLs	1275	705	Total number of *cis*-eQTLs	1192	1018
Type of probeset giving rise to the *cis*-eQTL			Type of probe giving rise to the *cis*-eQTL		
None of the corresponding probes contain a polymorphism ('unaltered')	517	227	Probe does not contain a polymorphism ('unaltered')	793	681
Only one corresponding probe contains polymorphism(s) ('altered')	119	54	Probe contains polymorphisms(s) ('discarded')	396	337
Two or more corresponding probes contain polymorphism(s) ('discarded')	639	424			
Proportion of eQTLs discarded (excluding altered) = discarded / (discarded + unaltered)	55.2%	65.1%	Proportion of eQTLs discarded	33.2%	33.1%
Expected proportion of eQTLs to be discarded	6.1%		Expected proportion of eQTLs to be discarded	11.7%	

The expected proportion to be discarded is the proportion of all probe/probesets discarded (including ones without a *cis*-eQTL signal).

datasets (Table 2). The difference between the two datasets in the proportion of probes / probesets affected is roughly proportional to the amount of mRNA sequence covered. As a result of probe drop out and overlapping probes, each Affymetrix Human Exon 1.0 ST Array probeset covers on average 72.3 unique nucleotides, compared with the 50 nucleotides of each Illumina HT12 Expression Array, which explains why the proportion of altered plus discarded Affymetrix probesets is higher than the proportion of discarded Illumina probes (i.e. 15.8% vs. 11.7% for the latest polymorphism reference data source) (see 'Materials and Methods' section for definitions of 'altered probeset' versus 'discarded probeset').

As one might expect, as the number of polymorphisms available in the reference data source increases, so the true extent of the polymorphism-in-probe problem becomes more evident. Since the majority of expression QTL studies listed in Supplementary Table S1 have attempted to identify SNP-containing probes using earlier HapMap information, it is important to note the considerable increase in the number of SNPs and the availability of data on indels between the final release of HapMap and the current release of 1000 Genomes (March 2012). Therefore, even findings from these studies have to be rigorously checked for any residual polymorphism-in-probe problem.

Proportion of LD-resolved *cis*-acting eQTLs arising from probes containing polymorphism(s) in probe sequence

We investigated the number of LD-resolved *cis*-acting eQTLs (<1 Mb from transcription start site of associated transcript, SNPs in a single LD block counted as one signal) that can be considered suspect because they are associated with polymorphism-containing probes/probesets. We considered a wide range of significance thresholds, polymorphism reference sources and different brain regions (Figure 1).

We found that the proportion of the LD-resolved *cis*-eQTLs affected by polymorphism-in-probe is much larger than the overall proportion of probes affected (Tables 2 and 3). This finding is consistent across the two brain regions and across data sets. In the frontal cortex of the Affymetrix Human Exon data set, we found that up to 90% of the declared eQTL signals involved polymorphism-containing probes when we should have expected only 6.1% based on the overall proportion of such probes. Table 3 illustrates this point by tabulating the number of suspect LD-resolved *cis*-eQTLs at *P*-value $< 10^{-12}$ when using the European ancestry panels of the 1000 Genomes Project (March 2012) and NHLBI-ESP.

The exon-level *cis*-eQTL results generated using the Affymetrix array (25-mer probe design) are much more affected by the polymorphism-in-probe problem than those results generated using the Illumina array (50-mer probe design). This is in agreement with previous studies showing that the presence of a polymorphism in a longer sequence has a less pronounced effect on the binding affinity than in a shorter sequence (21). However, the enrichment of false positives at gene-level by averaging exon-level data is comparable with the performance of the Illumina array (Supplementary Table S3).

Finally, we note that the proportion of suspect *cis*-eQTLs generally increases with more stringent *P*-value cut-offs. Therefore, and somewhat counter-intuitively, the more significant a result is the more likely it is to be a false positive.

When we repeated the analysis with *trans*-eQTL signals, we also saw a small, but noticeable, enrichment of false positives (Supplementary Figure S1). This affects some of the commonly presented statistics from eQTLs studies such as *cis*- to *trans*-eQTL ratios (Supplementary Figure S2).

Approaches to dealing with suspect *cis*-eQTLs from discarded probes/probesets

For Affymetrix Exon 1.0 ST arrays, where a probeset expression value is typically estimated from four constituent

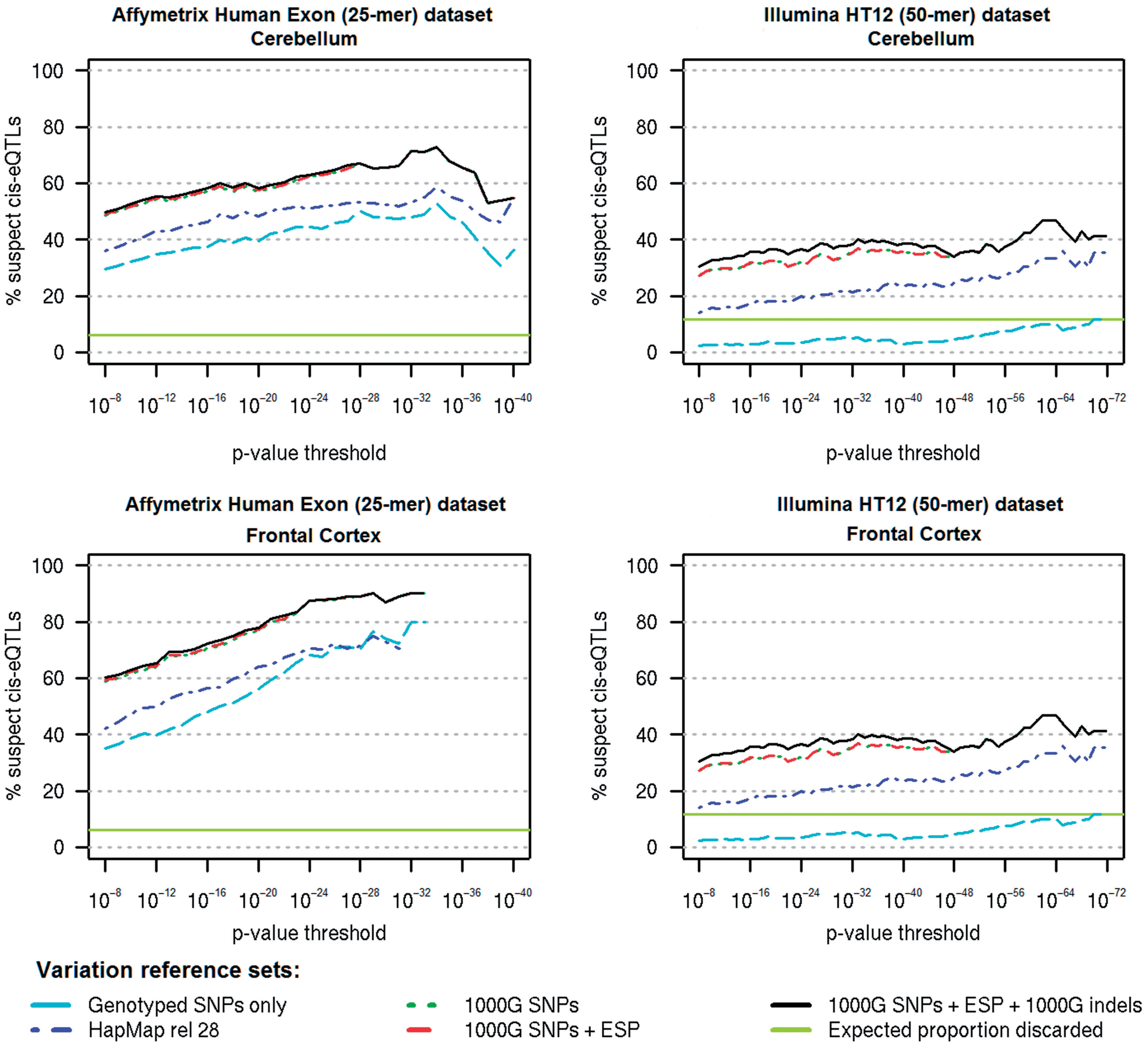

Figure 1. The proportion of LD-resolved *cis*-eQTL signals discarded because of the polymorphism-in-probe sequence problem using different polymorphism reference data sources and *P*-value thresholds. Multiple significant associations with a probe/probeset caused by SNPs in high LD ($r^2 \geq 0.5$) were treated as a single 'LD-resolved' signal. Also shown is the expected proportion that would be discarded if the rate was the same as the proportion of all probe/probesets (including ones without a *cis*-eQTL signal) discarded using the 1000 genomes (March 2012) plus Exome Sequencing Project reference data source.

probes, we apply probe masking (10) to exclude the polymorphism-containing probe and re-estimate the probeset expression value. If less than three probes remain free of polymorphism(s) for estimation, the probeset is still discarded. This solution recovers about two thirds of the polymorphism-containing probesets (to become 'altered' probesets).

This still leaves a large number of suspect *cis*-eQTLs from discarded probesets, especially for Illumina array data where probe masking is not applicable. Three alternatives to dealing with these suspect *cis*-eQTLs are (i) to remove all suspect *cis*-eQTLs; (ii) apply conditional association analysis; and (iii) apply LD filtering (see Materials and Methods section for details). Conditional association is a better motivated approach than LD filtering, but requires either imputation or sequencing to obtain the polymorphism-in-probe genotypes, a laborious step for existing eQTL data sets.

We find that we can rescue <6% of the suspect *cis*-eQTLs from both discarded Affymetrix probesets and discarded Illumina probes via the conditional analysis method (Supplementary Figure S3), so there is little additional benefit to using this method over probe masking (which does not require genotype imputation). LD filtering does a poor job of identifying these true eQTLs, regardless of the r^2 threshold applied. In all conditions considered, even if a stringent LD threshold of $r^2 < 0.1$ is used, at least 80% of the *cis*-eQTLs signals 'rescued' by LD filtering are in fact false according to the more proper conditional association method (Supplementary Table S4), and this percentage increases with the *P*-value stringency used to declare eQTL signals. This contraindicates the use of LD filtering to recover suspect *cis*-eQTLs.

Examples of false positives and false negatives due to polymorphism-in-probe

We selected three examples in two genes of relevance in brain disorders to demonstrate this problem in cerebellum (Figure 2). Common genetic variation at the MAPT gene has been associated with multiple neurodegenerative disorders (22) including Parkinson's Disease, progressive supranuclear palsy and corticobasal degeneration while PRICKLE1 has been implicated in progressive myoclonic epilepsy (23).

The first example (Figure 2A) shows a false-positive association between rs650927 and exon 6 of MAPT (probeset 3723733) in the Affymetrix Human Exon data set. The target sequence contains two SNPs that affect all four constituent probes, but the hit SNP is in high LD only with rs10445337 ($r^2 = 0.98$), which affects two of the probes. After excluding these two probes and re-calculating the probeset expression value via probe masking, the eQTL signal is no longer significant (*P*-value changes from 4.2×10^{-20} to 4.6×10^{-5}). Conditioning on both SNPs in the probe sequence also results in a non-significant result ($P = 0.644$).

The second example (Figure 2B) shows a false-negative association involving exon 8 of PRICKLE1 (probeset 3412103). One of the probes contains an SNP that is in high LD ($r^2 = 0.96$) with the hit SNP, which results in an opposite association compared with the other three probes. Excluding this probe results in the discovery of a significant eQTL signal (*P*-value changes from 1.4×10^{-5} to 4.3×10^{-18}), which would have been missed otherwise. Conditional analysis, using the original probeset expression, is not useful here, as the truly associated SNP is in too high LD with the polymorphism-in-probe, resulting in too high a level of confounding.

The final example (Figure 2C) is the false-positive association between rs1751739 (which tags the H1/H2 haplotype) and the probe ILMN_1710903 in the 3′UTR region of the MAPT gene. This influential finding was first reported in 2007(9) and has since been replicated in a number of other high profile studies (16,24). The hit SNP is in high LD with this common 2-base pair deletion (labelled as chr17:44102741:D in 1000 Genomes or as rs67759530 in dbSNP) within the probe ($r^2 = 0.91$, minor allele frequency = 23%), giving rise to a highly significant association in the Illumina HT12 data set (*P*-value = 8×10^{-31}). Since there are no constituent probes like there are in the Human Exon array, we investigated the association of the hit SNP with ILMN_2310814, which is located 2738 base pairs away and also in the 3′UTR of the MAPT gene, and observed no significant associations (*P*-value = 0.76). We discuss more about the eQTLs in this gene elsewhere (25).

DISCUSSION

In this article, we show that the presence of a small proportion of probes binding to sequences containing common polymorphisms massively inflates the number of *cis*-eQTL signals. These false eQTL signals tend to generate large effect sizes in relation to true signals, and so the problem only becomes worse as one increases the stringency of the *P*-value threshold used to define significance. Furthermore, these false signals will appear to replicate across studies if one uses the same array platform. We show here that previous eQTL studies are likely to have failed to adequately correct for this problem. This is primarily because of incomplete reference data on the location of all common polymorphisms at the time the studies were performed, but other factors have also played a part. For example, we show here that LD filtering introduces a large number of false positives, even if very low r^2 thresholds are used.

Our study suggests we are close to reaching a 'saturation point' in cataloguing all common exonic polymorphisms in the human genome. Although the number of polymorphisms with minor allele frequencies >1% goes up with every new reference data source we considered, often doubling or tripling compared with previous definitions, the proportion of *cis*-eQTLs discarded showed smaller and smaller changes. For example, the proportion of *cis*-eQTLs discarded in the frontal cortex samples of UKBEC at $P < 10^{-12}$ is 39.8% using genotype only data, 49.8% using HapMap release 28, 64.0% using 1000 Genomes SNPs, 64.4% using 1000 Genomes SNPs plus NHLBI-ESP exomes and 65.1% using 1000 Genomes (SNPs and indels) plus NHLBI-ESP exomes. This suggests we are close to having the full list of common polymorphisms, and that the ones we are missing are likely to be close to 1% in frequency and so with less of a tendency to generate false *cis*-eQTL signals. It is also worth noting that the number of probes/probesets discarded owing to the addition of nearly 1 million indels in the latest release of the 1000 Genomes Project is relatively low. One possible explanation for this is that, unlike SNPs, the existence of indels within exons is more likely to lead to deleterious frameshift changes in peptide sequences and thus are under negative selection.

Although the present study is the most complete analysis of the polymorphism-in-probe problem to date, it has limitations. We have not considered all types of polymorphisms, namely inversions and copy number variations, which are currently not as well characterized as SNPs or indels. We have also not attempted to model the binding affinity of probes as a function of the number of polymorphisms within a sequence; the position, nucleotide type and length of polymorphisms relative to the probe sequence; and the surrounding nucleotide types (4,21,26). We are, however, confident that the current size of reference data from the 1000 Genomes and Exome Sequence projects means that we are close to a complete catalogue of all common point-mutation polymorphisms in the major human populations.

Although the technology for assaying gene expression is now moving away from microarray-based methods and towards RNA sequencing, properly addressing the problem of polymorphism-in-probe remains important. Sample size remains the biggest driver for eQTL discovery, and while microarrays remain cheaper than RNA

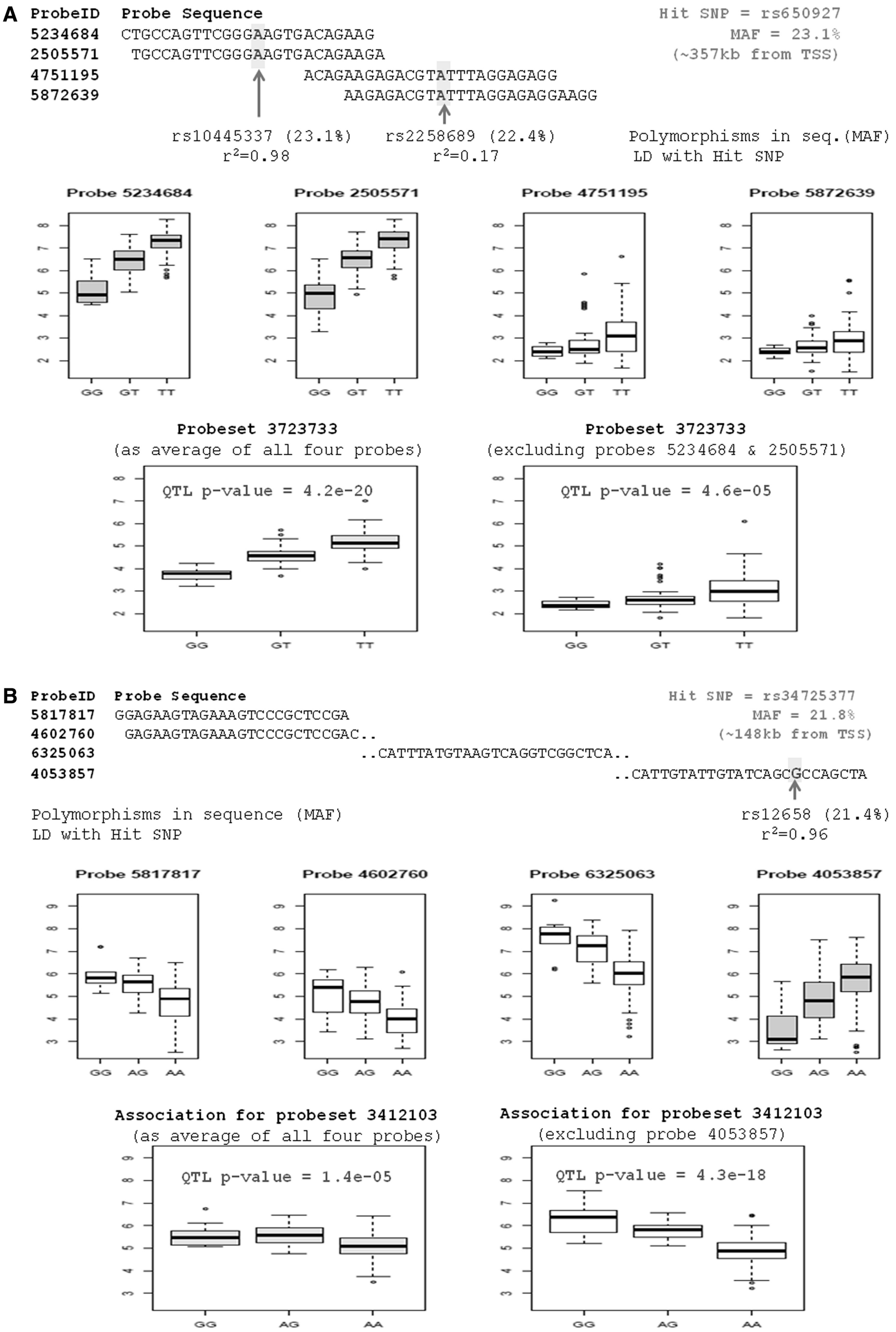

Figure 2. Illustrative examples of eQTLs with relevance to brain disorders. (**A**) Boxplots show the false-positive association between rs650927 genotypes and the measured expression of each of the four probes contained within the probeset 3723733 (exon 6 of MAPT) due to an SNP in the probe sequences. Two SNPs are present in the target sequence but only one is in high LD with the hit SNP. (**B**) Boxplots show the false-negative association between rs34725377 and probeset 3412103 (exon 8 of PRICKLE1) due to an SNP in one of the probe sequences. (**C**) Boxplots show the false-positive association between rs1751739 and the probe ILMN_1710903 (3′UTR region of the MAPT) due to a common 2-base pair deletion. The association between this SNP and ILMN_2310814, which also targets the 3′UTR of MAPT but is free of polymorphisms, is shown.

(continued)

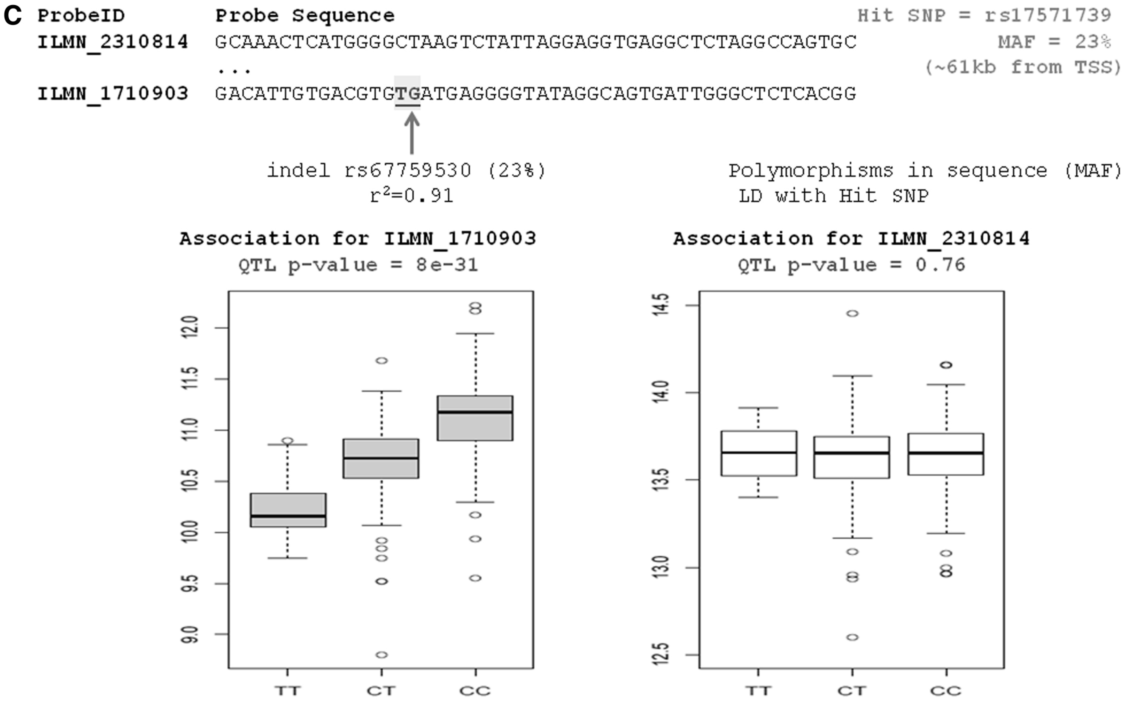

Figure 2. Continued.

sequencing, they will continue to be used for large-scale studies. In particular, low-expressed genes may require prohibitively large amounts of RNA sequencing to capture, but remain cheaply detectable via microarrays. Indeed, we are aware of only two eQTL studies based on RNA-sequence conducted in humans (both published in 2010), which use 69 HapMap West Africans (27) and 60 HapMap Europeans (28). In contrast, there are numerous recent studies, some involving thousands of samples, using microarray technology [e.g. 2355 samples from Grundberg *et al.* (29); 1490 samples from Zeller *et al.* (30)].

There is also a large and important body of existing microarray-based eQTL data spanning multiple tissue types and using large sample sizes. We believe that these data should be reassessed for potential false signals caused by polymorphism-in-probe issues, especially given the widespread distribution of these data via catalogues such as the Phenotype–Genotype Integrator, eQTLbrowser, seeQTL (31), SNPexpress (11) and GeneVar (32). Ideally, these catalogues should automatically flag up any suspect signals arising from polymorphism-containing probes. We have also written a web tool, PiP Finder (http://bit.ly/pipfinder), to provide researchers with an easy-to-use interface to identify this issue in any given eQTL signal. For an example of how we used this tool to check a recent publication for suspect eQTLs, please see our online comments to Zou *et al.* (33).

The polymorphism-in-probe problem is widely recognised in the eQTL literature, and various solutions have been proposed and implemented. Despite this, we show that false eQTL signals are likely to be widespread both in the literature and in extant eQTL databases. We note that the pervasive presence of false eQTL signals may have implications for, *inter alia*, the overlap of eQTL signals with genome-wide association study signals; the empirical distribution of eQTL signals relative to the transcription start site of genes; and apparent ratios of tissue-specific to cross-tissue eQTL signals. More generally, the findings of our study act as a cautionary tale for the interpretation of all types of genomic data, illustrating that even a relatively well-understood problem can be inadequately corrected.

Although we show that a large proportion of published *cis*-eQTL signals could be false, we also show that this problem can now be identified and resolved. From our own experience, meaningful, exciting and valid insights into the regulation of gene expression emerge once the polymorphism-in-probe problem is properly addressed.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1–5, Supplementary Figures 1–3, Supplementary Methods and Supplementary references [34–39].

ACKNOWLEDGEMENTS

Computing facilities used at King’s College London were supported by the National Institute for Health Research (NIHR) Biomedical Research Centre based at Guy’s and St Thomas’ NHS Foundation Trust and King’s College London. We would like to thank AROS Applied Biotechnology AS company laboratories and Affymetrix for their valuable input. The authors would like to thank the NHLBI GO Exome Sequencing Project and its ongoing studies, which produced and provided exome variant calls for comparison: the Lung GO Sequencing Project (HL-102923), the WHI Sequencing Project (HL-102924), the Broad GO Sequencing Project (HL-102925), the Seattle GO Sequencing Project (HL-102926) and the Heart GO Sequencing Project (HL-103010). The members of the North American Brain Expression Consortium (NABEC) are: (1) Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, MD, USA: Sampath Arepalli, Mark R Cookson, Allissa Dillman, J Raphael Gibbs, Dena G Hernandez, Michael A Nalls, Andrew Singleton, Bryan Traynor & Marcel van der Brug; (2) Clinical Research Branch, National Institute on Aging, Baltimore, MD, USA: Luigi Ferrucci; (3) Department of Molecular Neuroscience, UCL Institute of Neurology, London, UK: J Raphael Gibbs & Dena G Hernandez; (4) NICHD Brain and Tissue Bank for Developmental Disorders, University of Maryland Medical School, Baltimore, Maryland 21201, USA: Robert Johnson & Ronald Zielke; (5) Lymphocyte Cell Biology Unit, Laboratory of Immunology, National Institute on Aging, National Institutes of Health, Baltimore, MD, USA: Dan L Longo; (6) Brain Resource Center, Johns Hopkins University, Baltimore, MD, USA: Juan Troncoso; (7) ITGR Biomarker Discovery Group, Genentech, South San Francisco, CA, USA: Marcel van der Brug; (9) Research Resources Branch, National Institute on Aging, National Institutes of Health, Bethesda, MD, USA: Alan Zonderman.

FUNDING

Medical Research Council, UK, through the MRC Sudden Death Brain Bank (to C.S.); Project Grant [G0901254 to J.H. and M.W.]; Training Fellowship [G0802462 to M.R.]; King Faisal Specialist Hospital and Research Centre, Saudi Arabia (to D.T.); Intramural Research Program of the National Institute on Aging, National Institutes of Health, part of the US Department of Health and Human Services [ZIA AG000932-04] (in part) (to work performed by the North American Brain Expression Consortium). Funding for open access charge: Medical Research Council, UK.

Conflict of interest statement. None declared.

REFERENCES

1. Cookson,W., Liang,L., Abecasis,G., Moffatt,M. and Lathrop,M. (2009) Mapping complex disease traits with global gene expression. *Nat. Rev. Genet.*, **10**, 184–194.
2. Cheung,V.G. and Spielman,R.S. (2009) Genetics of human gene expression: mapping DNA variants that influence gene expression. *Nat. Rev. Genet.*, **10**, 595–604.
3. Gilad,Y., Rifkin,S.A. and Pritchard,J.K. (2008) Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends Genet.*, **24**, 408–415.
4. Naiser,T., Ehler,O., Kayser,J., Mai,T., Michel,W. and Ott,A. (2008) Impact of point-mutations on the hybridization affinity of surface-bound DNA/DNA and RNA/DNA oligonucleotide-duplexes: comparison of single base mismatches and base bulges. *BMC Biotechnol.*, **8**, 48.
5. Rozowsky,J., Abyzov,A., Wang,J., Alves,P., Raha,D., Harmanci,A., Leng,J., Bjornson,R., Kong,Y., Kitabayashi,N. *et al.* (2011) AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Mol. Syst. Biol.*, **7**, 522.
6. Vijaya Satya,R., Zavaljevski,N. and Reifman,J. (2012) A new strategy to reduce allelic bias in RNA-Seq readmapping. *Nucleic Acids Res*, **40**, e127.
7. Degner,J.F., Marioni,J.C., Pai,A.A., Pickrell,J.K., Nkadori,E., Gilad,Y. and Pritchard,J.K. (2009) Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics*, **25**, 3207–3212.
8. Walter,N.A., McWeeney,S.K., Peters,S.T., Belknap,J.K., Hitzemann,R. and Buck,K.J. (2007) SNPs matter: impact on detection of differential expression. *Nat. Methods*, **4**, 679–680.
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. Benovoy,D., Kwan,T. and Majewski,J. (2008) Effect of polymorphisms within probe-target sequences on olignonucleotide microarray experiments. *Nucleic Acids Res.*, **36**, 4417–4423.
11. Heinzen,E.L., Ge,D., Cronin,K.D., Maia,J.M., Shianna,K.V., Gabriel,W.N., Welsh-Bohmer,K.A., Hulette,C.M., Denny,T.N. and Goldstein,D.B. (2008) Tissue-specific genetic control of splicing: implications for the study of complex traits. *PLoS Biol.*, **6**, e1.
12. Alberts,R., Terpstra,P., Li,Y., Breitling,R., Nap,J.P. and Jansen,R.C. (2007) Sequence polymorphisms cause many false cis eQTLs. *PLoS One*, **2**, e622.
13. Gamazon,E.R., Zhang,W., Dolan,M.E. and Cox,N.J. (2010) Comprehensive survey of SNPs in the Affymetrix exon array using the 1000 Genomes dataset. *PLoS One*, **5**, e9366.
14. Stranger,B.E., Montgomery,S.B., Dimas,A.S., Parts,L., Stegle,O., Ingle,C.E., Sekowska,M., Smith,G.D., Evans,D., Gutierrez-Arcelus,M. *et al.* (2012) Patterns of cis regulatory variation in diverse human populations. *PLoS Genet.*, **8**, e1002639.
15. (IPDGC), I.P.s.D.G.C. and (WTCCC2), W.T.C.C.C. (2011) A two-stage meta-analysis identifies several new loci for Parkinson's disease. *PLoS Genet*, **7**, e1002142.
16. 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.
17. Hernandez,D.G., Nalls,M.A., Moore,M., Chong,S., Dillman,A., Trabzuni,D., Gibbs,J.R., Ryten,M., Arepalli,S., Weale,M.E. *et al.* (2012) Integration of GWAS SNPs and tissue specific expression profiling reveal discrete eQTLs for human traits in blood and brain. *Neurobiol. Dis.*, **47**, 20–28.
18. Trabzuni,D., Ryten,M., Walker,R., Smith,C., Imran,S., Ramasamy,A., Weale,M.E. and Hardy,J. (2011) Quality control parameters on a large dataset of regionally dissected human control brains for whole genome expression studies. *J. Neurochem.*, **119**, 275–282.
19. Shabalin,A.A. (2012) Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics*, **28**, 1353–1358.
20. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
21. Rennie,C., Noyes,H.A., Kemp,S.J., Hulme,H., Brass,A. and Hoyle,D.C. (2008) Strong position-dependent effects of sequence mismatches on signal ratios measured using long oligonucleotide microarrays. *BMC Genomics*, **9**, 317.
22. Vandrovcova,J., Anaya,F., Kay,V., Lees,A., Hardy,J. and de Silva,R. (2010) Disentangling the role of the tau gene locus in sporadic tauopathies. *Curr. Alzheimer Res.*, **7**, 726–734.
23. Bassuk,A.G., Wallace,R.H., Buhr,A., Buller,A.R., Afawi,Z., Shimojo,M., Miyata,S., Chen,S., Gonzalez-Alegre,P., Griesbach,H.L. *et al.* (2008) A homozygous mutation in human PRICKLE1 causes an autosomal-recessive progressive myoclonus epilepsy-ataxia syndrome. *Am. J. Hum. Genet.*, **83**, 572–581.

24. Nalls,M.A., Plagnol,V., Hernandez,D.G., Sharma,M., Sheerin,U.M., Saad,M., Simon-Sanchez,J., Schulte,C., Lesage,S., Sveinbjornsdottir,S. *et al*. (2011) Imputation of sequence variants for identification of genetic risks for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet*, **377**, 641–649.
25. Trabzuni,D., Wray,S., Vandrovcova,J., Ramasamy,A., Walker,R., Smith,C., Luk,C., Gibbs,J.R., Dillman,A., Hernandez,D.G. *et al.* (2012) MAPT expression and splicing is differentially regulated by brain region: relation to genotype and implication for tauopathies. *Hum. Mol. Genet*., **21**, 4094–4103.
26. Naiser,T., Kayser,J., Mai,T., Michel,W. and Ott,A. (2008) Position dependent mismatch discrimination on DNA microarrays - experiments and model. *BMC Bioinformatics*, **9**, 509.
27. Pickrell,J.K., Marioni,J.C., Pai,A.A., Degner,J.F., Engelhardt,B.E., Nkadori,E., Veyrieras,J.B., Stephens,M., Gilad,Y. and Pritchard,J.K. (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, **464**, 768–772.
28. Montgomery,S.B., Sammeth,M., Gutierrez-Arcelus,M., Lach,R.P., Ingle,C., Nisbett,J., Guigo,R. and Dermitzakis,E.T. (2010) Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature*, **464**, 773–777.
29. Grundberg,E., Small,K.S., Hedman,A.K., Nica,A.C., Buil,A., Keildson,S., Bell,J.T., Yang,T.P., Meduri,E., Barrett,A. *et al.* (2012) Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nat. Genet*., **44**, 1084–1089.
30. Zeller,T., Wild,P., Szymczak,S., Rotival,M., Schillert,A., Castagne,R., Maouche,S., Germain,M., Lackner,K., Rossmann,H. *et al.* (2010) Genetics and beyond–the transcriptome of human monocytes and disease susceptibility. *PLoS One*, **5**, e10693.
31. Xia,K., Shabalin,A.A., Huang,S., Madar,V., Zhou,Y.H., Wang,W., Zou,F., Sun,W., Sullivan,P.F. and Wright,F.A. (2012) seeQTL: a searchable database for human eQTLs. *Bioinformatics*, **28**, 451–452.
32. Yang,T.P., Beazley,C., Montgomery,S.B., Dimas,A.S., Gutierrez-Arcelus,M., Stranger,B.E., Deloukas,P. and Dermitzakis,E.T. (2010) Genevar: a database and Java application for the analysis and visualization of SNP-gene associations in eQTL studies. *Bioinformatics*, **26**, 2474–2476.
33. Zou,F., Chai,H.S., Younkin,C.S., Allen,M., Crook,J., Pankratz,V.S., Carrasquillo,M.M., Rowley,C.N., Nair,A.A., Middha,S. *et al*. (2012) Brain expression genome-wide association study (eGWAS) identifies human disease-associated variants. *PLoS Genet*., **8**, e1002707.
34. Millar,T., Walker,R., Arango,J.C., Ironside,J.W., Harrison,D.J., MacIntyre,D.J., Blackwood,D., Smith,C. and Bell,J.E. (2007) Tissue and organ donation for research in forensic pathology: the MRC Sudden Death Brain and Tissue Bank. *J. Pathol*., **213**, 369–375.
35. Beach,T.G., Sue,L.I., Walker,D.G., Roher,A.E., Lue,L., Vedders,L., Connor,D.J., Sabbagh,M.N. and Rogers,J. (2008) The Sun Health Research Institute Brain Donation Program: description and experience, 1987-2007. *Cell Tissue Bank*, **9**, 229–245.
36. Wu,Z.J., Irizarry,R.A., Gentleman,R., Martinez-Murillo,F. and Spencer,F. (2004) A model-based background adjustment for oligonucleotide expression arrays. *J. Am. Stat. Assoc*., **99**, 909–917.
37. Li,Y., Willer,C., Sanna,S. and Abecasis,G. (2009) Genotype imputation. *Annu. Rev. Genomics Hum. Genet*., **10**, 387–406.
38. 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.
39. 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.