European Journal of Human Genetics (2015) 23, 1229–1235

www.nature.com/ejhg

ARTICLE

# Effects of copy number variable regions on local gene expression in white blood cells of Mexican Americans

August Blackburn*[,1,2], Marcio Almeida[1], Angela Dean[3], Joanne E Curran[1], Matthew P Johnson[1], Eric K Moses[4], Lawrence J Abraham[4], Melanie A Carless[1], Thomas D Dyer[1], Satish Kumar[1], Laura Almasy[1], Michael C Mahaney[1], Anthony Comuzzie[1], Sarah Williams-Blangero[1,5], John Blangero[1], Donna M Lehman[6] and Harald HH Göring[1]

**Only few systematic studies on the contribution of copy number variation to gene expression variation have been published to date. Here we identify effects of copy number variable regions (CNVRs) on nearby gene expression by investigating 909 CNVRs and expression levels of 12059 nearby genes in white blood cells from Mexican-American participants of the San Antonio Family Heart Study. We empirically evaluate our ability to detect the contribution of CNVs to proximal gene expression (presumably in *cis*) at various window sizes (up to a 10 Mb distance) between the gene and CNV. We found a ~ 1-Mb window size to be optimal for capturing *cis* effects of CNVs. Up to 10% of the CNVs in this study were found to be significantly associated with the expression of at least one gene within their vicinity. As expected, we find that CNVs that directly overlap gene sequences have the largest effects on gene expression (compared with non-overlapping CNVRs located nearby), with positive correlation (except for a few exceptions) between estimated genomic dosage and expression level. We find that genes whose expression level is significantly influenced by nearby CNVRs are enriched for immunity and autoimmunity related genes. These findings add to the currently limited catalog of CNVRs that are recognized as expression quantitative trait loci, and have implications for future study designs as well as for prioritizing candidate causal variants in genomic regions associated with disease.**

*European Journal of Human Genetics* (2015) **23,** 1229–1235; doi:10.1038/ejhg.2014.280; published online 14 January 2015

## INTRODUCTION

Inter-individual variation in transcript abundance is known to be significantly heritable for many genes. Transcript level can be considered as a quantitative endophenotype whose genetic regulatory machinery can be mapped to the genome.[1] The expression quantitative trait loci (eQTL) with the strongest effects on gene expression act primarily in *cis*.[1] A critical bottleneck in the search for disease genes is the identification of the underlying causal variants, which are often initially localized in genome-wide association studies (GWAS). A promising hypothesis now being explored for complex traits and diseases is that functional alleles may be regulatory in nature and exert their effect by altering gene expression[2] (and thus making them detectable by genetic investigations of expression levels). This general hypothesis is supported by various observations, including the fact that most of the identified common disease-associated SNPs are not in protein coding regions and often are located far away from exons of known genes. Recent studies have shown that GWAS SNPs are enriched among eQTL.[3,4] Further, GWAS SNPs, that are known eQTL, often affect gene expression in the disease tissue.[3] Given these observations, gene transcript levels have received a high level of interest as endophenotypes that can be correlated with disease status, and whose genetic regulatory mechanisms can be mapped with considerable power.[2]

Copy number variation reported in the Database of Genomic Variants covers roughly 70% of the genome,[5] although this estimate is likely upward biased by inaccurate breakpoint identification.[6] Nonetheless, in an individual, copy number variants (CNVs) make up more variation than SNPs on a per nucleotide basis.[7] Thus, the potential effects of CNVs as eQTL are likely large. Effects caused by gene dosage should be less tissue specific than variation in gene expression caused by genetic variation in distant regulatory regions, which is important as we are often limited to studying surrogate tissues to identify eQTL. Only a few recent studies have sought to systematically identify CNVs which act as eQTL.[8,9] Stranger *et al*[9] identified 238 genes with expression levels that were significantly associated with copy number variation. More recently, Schlattl *et al*[8] identified 110 genes with expression affected by CNVs. Both studies investigated the relative proportions of eQTL attributable to CNVs and SNPs, but the effects by both variants are difficult to disentangle because of the linkage disequilibrium between CNVs and SNPs. This correlation among genetic variants located in genomic proximity to one another also suggests that some eQTL previously identified in SNP-based studies may be attributable to CNVs. Gamazon *et al*[10] found that SNPs tagging CNVs were enriched for *cis*-eQTL, and that these SNPs are overrepresented in the National Human Genome Research Institute's (NHGRI) catalog of GWAS SNPs.

[1]Department of Genetics, Texas Biomedical Research Institute, San Antonio, TX, USA; [2]Department of Cellular and Structural Biology, UT Health Science Center San Antonio, San Antonio, TX, USA; [3]Department of Computer Science, University of Texas San Antonio, San Antonio, TX, USA; [4]Centre for Genetic Origins of Health and Disease, University of Western Australia, Perth, WA, Australia; [5]Southwest National Primate Research Center, San Antonio, TX, USA; [6]Department of Medicine/Division of Clinical Epidemiology, UT Health Science Center San Antonio, San Antonio, TX, USA
*Correspondence: Dr A Blackburn, Department of Genetics, Texas Biomedical Research Institute, 7620 NW Loop 410, San Antonio, TX 78227, USA. Tel: +1 210 258 9208; Fax: +1 210 258 9444; E-mail: augustb@txbiomedgenetics.org
Received 25 March 2014; revised 25 September 2014; accepted 26 November 2014; published online 14 January 2015

In this study we seek to add to the growing catalog of eQTL by identifying genes whose expression level in white blood cells (mainly lymphocytes) is affected by CNVs in the San Antonio Family Heart Study (SAFHS).

## SUBJECTS AND METHODS

### Study design

Participants in the SAFHS[11] are members of extended, multigenerational families of Mexican-American descent. SAFHS is a family study where the subjects were not ascertained on disease status. The Institutional Review Board at the University of Texas Health Science Center San Antonio approved the current study, and informed consent was obtained from all participants. All study related clinical exams were conducted in San Antonio, TX, USA. Gene expression and copy number variation data were available for 1104 participants.

### Copy number variable regions

We recently identified 2937 copy number variable regions (CNVRs) in participants of the SAFHS using various Illumina (San Diego, CA, USA) Infinium Beadchips.[12] Our ability to characterize these CNVRs is limited in the absence of sequencing data. Some CNVRs fall within known complex regions, or fall within regions of the genome that are predisposed to recurrent copy-number-altering mutational events.[13,14] However, we reason that the majority are diallelic CNVs as we previously observed reasonable concordance in size and location between these CNVRs and those identified to be polymorphisms by HapMap3.[12,15]

Using Log R ratios for probes within each CNVR, we generated quantitative values representative of copy number using the principal components function implemented within CNVtools.[16] As described by Barnes *et al*,[16] this approach has the advantages of creating a single representative value for each region, as well as generally improving cluster separation when compared with using the mean or median of all probe intensities. Using this approach, cluster separation was only sufficient to allow us to 'bin' a relatively poor percentage of the CNVRs (186 CNVRs, 6.3%) into defined copy number states. However, the underlying quantitative values (from principal components analysis), which are representatives of copy number, are overwhelmingly heritable (95% have statistically significant heritability estimates). Further, a subset of these CNVRs (920 CNVRs, 31.3%) show evidence of linkage to their own genomic location. Taken together, these observations strongly support the assertion by Barnes *et al*[16] that this approach captures features representative of copy number, and provides support for using these quantitative measurements in the place of discrete copy number in this study.[12]

In the absence of accurately binned copy number states, testing of underlying quantitative measures as representations of copy number has been shown to be effective and often a more accurate strategy than binning.[17] This strategy was applied to study the effects of CNVs on gene expression by Stranger *et al*.[9] Quantitative values representative of copy number have previously been used in a variance component framework accounting for relatedness of individuals within pedigrees and applied to study variation in gene transcript expression.[18] With this established precedent, we chose to leverage the available quantitative copy number measurements to identify CNVs that act as eQTL. We chose to work with the subset of CNVRs for which these quantitative values show evidence of linkage to their own genomic location (920 CNVRs), as we reasoned that this subset is likely more robustly measured. The annotation for these 920 CNVRs was updated to hg19 using the liftOver utility from the UCSC genome browser.[19] Eleven CNVRs do not map uniquely to hg19, resulting in a total of 909 CNVRs for this study.

The quantitative measurements which are used in this study are based on the R function prcomp,[16] for which numerical sign is arbitrary. Thus, a positive correlation between these values and copy number is not obligatory. Accordingly, to accurately model the direction of effect of copy number on gene expression, the sign of $\beta$ from our statistical genetic analysis (which represents the direction of effect) was adjusted according the correlation between the values produced by CNVtools (using prcomp) and the mean of Log R ratios (which are positively correlated with copy number) for the probes in each region.

### Gene expression

For this study, we used gene expression values from Goring *et al*.[1] The ascertainment of transcript abundance measurements has been previously described in detail.[1] Briefly, genome-wide transcription profiles were created using Illumina Sentrix Human Whole Genome (WG-6) Series I BeadChips, and are archived under ArrayExpress accession number E-TABM-305. However, these data were re-processed using the following approach. Based on the number of probes with detectable expression (at 'detection *P*-value' $\leq 0.05$), the average of the raw expression levels across probes, and the average correlation across all probes between each sample and all others, 1244 samples were determined to yield expression profile data of acceptable quality. Among these samples, we tested whether there was an enrichment of samples with a detection *P*-value $\leq 0.05$ for each probe (the 'detection *P*-value' is a quantity provided by Illumina software for each probe in each sample, generated by comparing the expression level of a given probe to null control probes on the array) to determine which probes detected significant expression. We did this using a binomial test of the number of samples with 'detection *P*-value' $\leq 0.05$ (5% of the samples would be expected to have a *P*-value at this level by chance). To correct for multiple testing (as there are many probes being tested) we kept probes at a false discovery rate of 0.05. Subsequently, we performed background noise correction, log2 transformation, and quantile normalization. We have used this procedure previously and it is also described here.[20] For the sake of simplicity, tests were performed at the probe level, although in some cases genes were represented by more than one probe. illuminaHumanv1.db available through the Bioconductor website was used for probe annotation.

### Statistical genetic analysis

The relationship between copy number variation and probe-level gene expression was examined using a variance components model, as implemented in the software package SOLAR.[21,22] An additive autosomal polygenic model was used to allow for the non-independence of relatives attributable to their expected genome-wide genetic similarity because of kinship. Gene expression was the trait of interest whose expected value depends on several measured variables ('covariates'). Additional covariates used in all models were sex, age, sex $\times$ age interaction, age,$^2$ and sex $\times$ age$^2$ interaction. Before analysis, probe-level expression values and quantitative values representative of copy number were rank normalized to assure that the assumption of normality during maximum likelihood estimation was met. False discovery rate was controlled using the procedures defined by Storey and Tibshirani.[23]

### Additional analyses

Functional annotation clustering was performed using David bioinformatics resources[24] using a background gene set of the genes used in this study to correct for potential tissue-specific effects, an approach that has been used previously.[8] Briefly, the annotation used for this analysis were all RefSeq annotations available for the 13 546 probes in this study. David Bioinformatics Gene ID Conversion Tool, available at the DAVID bioinformatics website, was used to convert this list into DAVID gene ids and also to remove redundancy from this list. The background gene set is provided in the Supplementary Material. Clustering was performed using medium classification stringency and clusters with an Enrichment score (as defined by DAVID bioinformatics resources[24]) $>2.5$ (which corresponds to a *P*-value of 0.003) were considered significant.

## RESULTS

### Background information and terminology

In a previous study,[12] we identified 2937 CNVRs among participants of the SAFHS genotyped using various Illumina Infinium Beadchips. For most CNVRs, poor cluster separation did not allow for precise copy number determination. However, quantitative values representative of copy number (generated using Log R ratios for probes within each CNVR and the principal components function implemented within CNVtools[12,16]) were significantly heritable for an overwhelming percentage (95%) of these regions. Furthermore, 920 of the more

common heritable CNVRs showed linkage to their own genomic location, providing additional evidence that these CNVRs are real. In this study, we use the quantitative values for these 920 variants as substitutes to integer copy numbers, to identify CNVs that are eQTL. We compared the results of association with all gene expression values using these quantitative values and using binned copy number genotypes for 149 CNVRs for which binned copy number genotypes are available. The proportions of variance accounted for by the CNVRs were very consistent, especially for the tail end of the distribution with higher proportions of variance. The estimated proportions of variance accounted for by the 250 most significant results were highly correlated between quantitative values and binned copy number ($R^2 > 0.99$, Supplementary Figure 1).

The annotation for these 920 CNVRs was updated to hg19 using the liftOver utility from the UCSC genome browser.[19] Eleven CNVRs do not map uniquely to hg19, resulting in a total of 909 CNVRs for use in this study. Figure 1 provides an example of the relationship between these quantitative values, their underlying discrete copy number state, and gene expression. A portion of these CNVRs represents known complex regions. However, most represent diallelic copy number polymorphisms (presence or absence of a deletion/duplication) as we previously observed reasonable concordance in size and location between these CNVRs and those identified to be polymorphisms by HapMap3.[12,15] We will refer to the total set as CNVs for the ease of communication.

### Identification of *cis*-eQTL

In order to identify which CNVs are putative *cis*-eQTL, we tested the aforementioned 909 CNVs for association with transcript levels of genes within a symmetrical 10 Mb window of each CNV (a total window size of 20 Mb+CNV length). In total, we analyzed 89 893 CNV-expression probe pairs. As expected, we detected a substantial number of significant eQTL, after adjusting for multiple testing. As shown in Figure 2, significant findings were enriched for proximity between CNVs and genes. The most highly-associated CNV-gene pair was between GSTM1 and an overlapping duplication, and was detected using two separate gene expression probes. This duplication was estimated to account for ~52% of the variance in GSTM1 expression by both probes (Figure 1).

At a symmetrical window size of 10 Mb, 118 tests were significant ($q < 0.1$) representing 97 genes and 75 CNVs (Supplementary Material). Some genes were affected by multiple non-overlapping CNVs, and some CNVs had an effect on multiple genes. Fifteen (~15%) of these 97 genes were previously identified by Schlattl *et al.* (10 genes) or Stranger *et al.* (10 genes), which is certainly a greater proportion than would be expected by chance. Five genes (HLA-DRB5, HLA-DQA1, NAIP, RRP7B, and PDPR) were found in all three

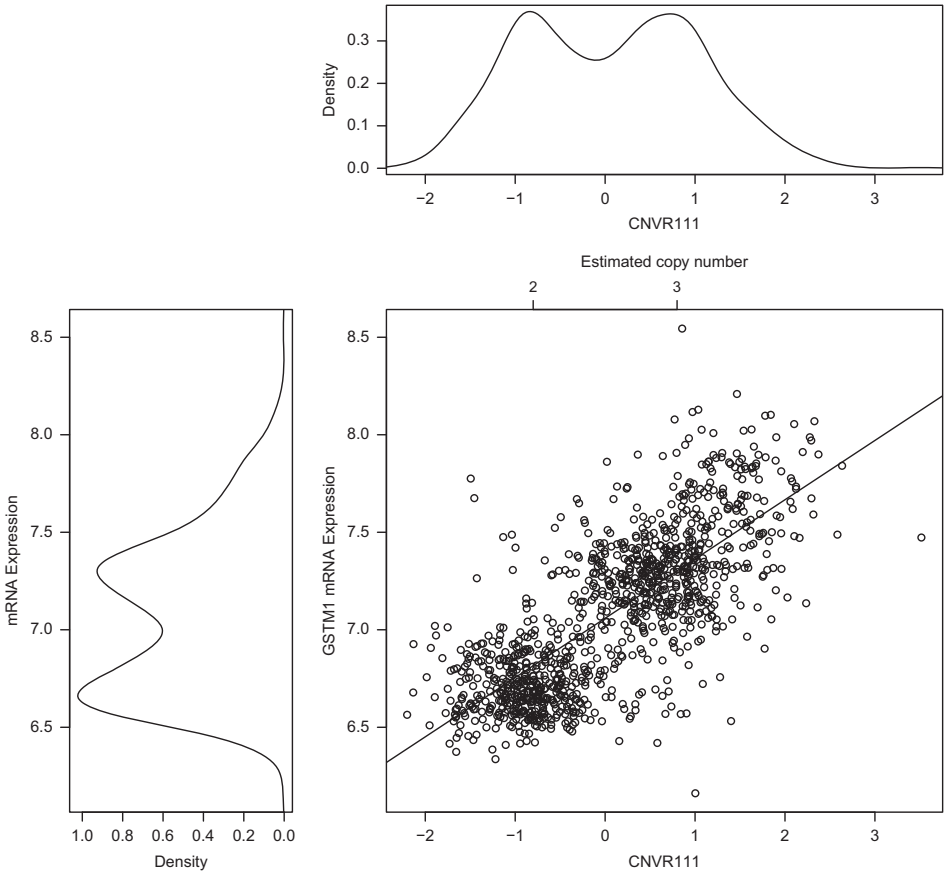


**Figure 1** CNVR111 and GSTM1 expression. Quantitative values representative of copy number (horizontal axis of the main panel) for CNVR111 (a duplication) are significantly associated with mRNA expression of GSTM1 (vertical axis of the main panel). A density plot shows that these quantitative values cluster in two overlapping distributions, which represent underlying discrete genotypes. A density plot of the gene expression values reveals that expression closely mirrors the underlying genotypes.

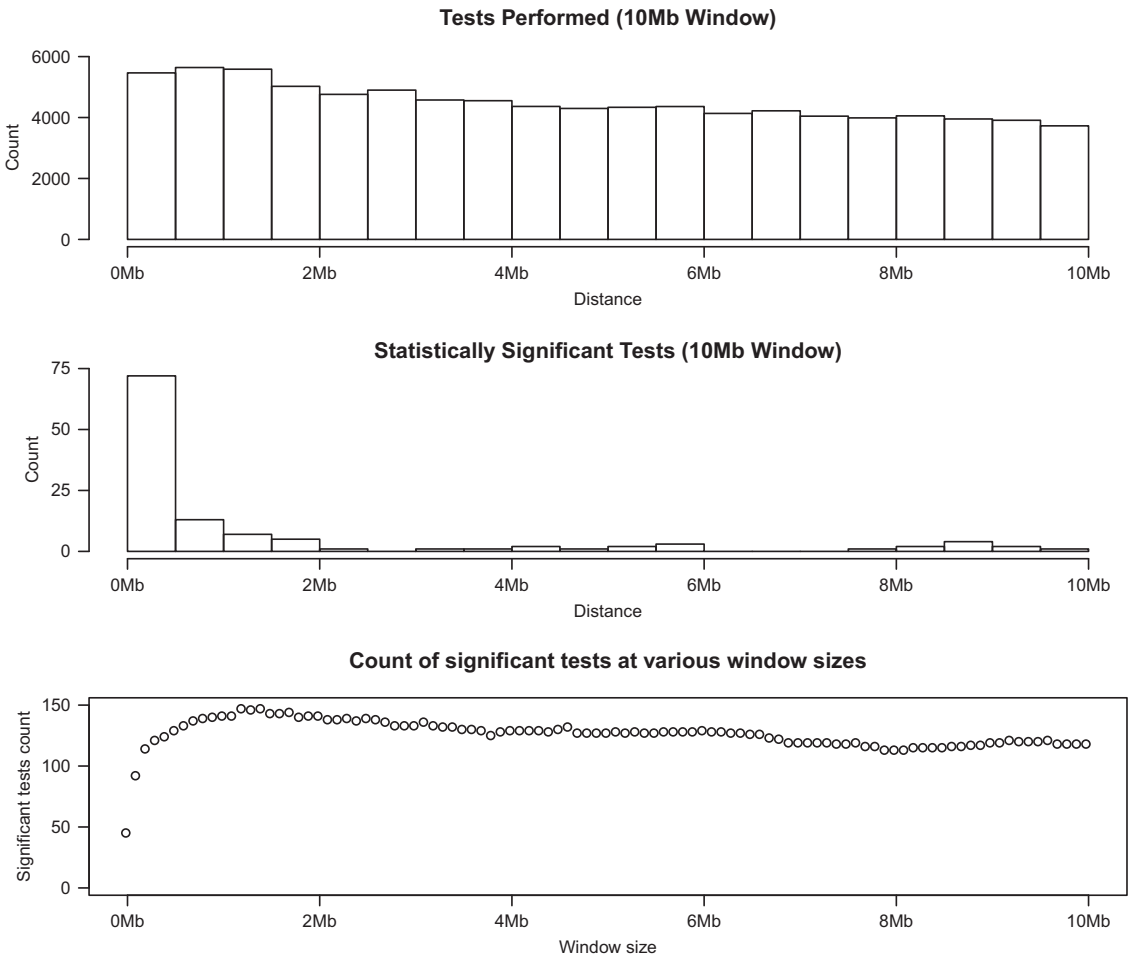


**Figure 2** Window size and statistically significant tests. The top panel shows the distribution of the distances between the gene and CNV for the tests performed using a 10 Mb window size. The middle panel shows the tests that were statistically significant ($q<0.1$) among the tests performed at a 10 Mb window size. The statistically significant results are clearly enriched for proximity between genes and CNVs. The bottom panel shows the number of statistically significant tests (vertical axis, $q<0.1$) for various window sizes in increments of 100 kb up to 10 Mb. The benefit of increasing window size to capture additional *cis* effects is outweighed by correction of multiple testing around a window size of 1.2 Mb.

studies. Most of the identified genes are novel, and given the limitations of this and previous studies (either in sample size or in methodology for identifying and genotyping CNVs) it is likely that we are only scratching the surface of the influence of CNVs on gene expression levels.

Interestingly, 33 and 15 significant CNV-gene pairs were separated by at least 1 and 5 Mb, respectively. Despite identifying these more distant effects, in general the closer the distance between CNV and gene, the higher the average proportion of variation in gene expression attributable to the CNV. There was, however, a notable exception, in which 28% of the variance in NUPR1L expression was accounted for by a ~50 kb duplication located ~9.1 Mb upstream from the transcription start site. It is important to note that with the available data we are not able to determine the insertion location of the duplicated sequence, which may be much closer to the NUPR1L gene, and could potentially explain its strong effect.

**Influence of CNVs on directly overlapping genes' expression levels**

Among CNVs, those directly overlapping genes are expected to have the most direct and largest (average) effects on variation in gene expression *a priori*. We sought to interpret the results at this more restricted window size. Only considering genes that overlap with CNVs, 45 of 350 (12.9%) tests were significant ($q<0.1$), representing 43 genes and 38 CNVs. When only considering genes entirely contained within CNVs, 32 of 157 (20.4%) tests were significant ($q<0.1$), representing 31 genes and 27 CNVs.

Among the 32 significant results, 29 were positively correlated indicating that the effects on gene expression are presumably due to a direct dosage effect due to increase or decrease in gene copy number. The three genes with apparent negative correlations, DGCR6L, LRRC14, and PCGF3, all appear to fall within complex regions of the genome, an observation that is corroborated by the complexity of the CNV calls from the 1000 genomes project in this region.[25] Schlattl *et al*[8] previously observed counterintuitive negative correlations between copy number and gene expression, and therefore these observations are unlikely due to chance, although significant negative correlations may be a result of the intrinsic difficulty of accurately genotyping in complex regions of the genome. Overall, we have replicated the observations by Schlattl *et al*[8] that gene expression is generally positively correlated with copy number.

Many genes appeared to be clinically relevant among the 45 significant findings (when considering genes overlapped by CNVRs). For example, point mutations in the *HBG2* gene, such as a G to A point mutation at position 202 (which causes a valine to methionine

substitution at codon 68), can cause neonatal cyanosis and anemia[26] by inhibiting or preventing binding of oxygen to hemoglobin. Deletions of the *HBG2* gene can cause complications during prenatal diagnosis of $\beta$-thalassemia[27] due to the absence of the potential compensatory effect of persistent fetal hemoglobin expression into adulthood, which often offsets effects of $\beta$-thalassemia. Conversely, duplications of the *HBG2* gene appear to be benign.[28] GSTM1 and GSTT1, both glutathione S-transferases, which are commonly over-expressed in multiple cancers, may aid in chemotherapeutic drug resistance through their role in drug metabolism,[29] and thus altered baseline expression may also have a similar role. TBXAS1 is involved in the conversion of prostaglandin endoperoxide into thromboxane A2, which is a potent vasoconstrictor and inducer of platelet aggregation,[30] and is thought to be responsible for the rare autosomal recessive bone density disorder, Ghosal hematodiaphyseal dysplasia.[31] Additionally, 15 genes appear under the Gene Ontology term 'immune response'.[24,32] Thus, indications are that copy number variable genes are major players in determining genetic risk for clinically relevant phenotypes.

### Optimization of window size

We sought to establish an optimal window size for this study that maximizes the number of statistically significant findings. To do this, we subset the data based on symmetrical window sizes incrementally increased by 100 kb. With each set of data, we calculated the number of findings that would be called statistically significant at $q<0.1$. As shown in Figure 2, the number of statistically significant tests increases until around ~1.2 Mb, after which the number of statistically significant tests slowly declines due to more stringent significance criteria necessary due to increased hypothesis testing. The results at this window size are not dissimilar to the results at a 1 Mb symmetrical window, which interestingly is commonly used in SNP-based eQTL studies.

At a symmetrical window size of 1.2 Mb, 147 tests were statistically significant, representing 88 CNVs and 117 genes. At this window size, 32 (~22%) significant tests represent cases in which genes overlap CNVs. A summary of the results at different window sizes described in this manuscript is provided in Table 1. Among the significant findings that were excluded at the smallest window size were clinically relevant genes such as GSTM2 and GSTM4, which are also glutathione S-transferases. Additionally, clinically relevant genes were excluded due to severe multiple testing correction necessitated when using the 10 Mb symmetrical window, including HBG1, which is a hemoglobin subunit very closely related to HBG2. It is important to note that the overall effect on HBG1 expression (and other genes not detected at the 10 Mb window size) appears to be small.

### Identification of *trans*-eQTL

We tested the aforementioned 909 CNVs for association with transcript levels of all genes. In total, we analyzed 12 364 146 CNV-expression probe pairs (including the aforementioned 89 893 *cis*-eQTL). We detected two significant ($P<4.04\times10^{-9}$, Bonferroni) *trans*-eQTL (not previously detected in our *cis*-eQTL analysis), a stark difference compared with the 44 *cis*-eQTL that are significant at this same threshold. Expression of MAPK8IP1 (on chromosome 11) is affected by copy number variation on chromosome 17, and EPB41L4A (on chromosome 5) is affected by copy number variation on chromosome 6. These observations support previous reports[1] that the effect sizes of putative *trans*-eQTL tend to be smaller than those typically observed for *cis*-eQTL.

### Ontology and pathway analysis of *eQTL* genes

We examined whether the genes whose expression levels were significantly impacted by nearby CNVs fall into specific categories using the results at a 10 Mb window. Using David Bioinformatics,[24] we found a cluster of genes enriched among KEGG pathways,[33] Gene Ontology[32] terms, and the Uniprot tissues[34] related to immunity and autoimmunity (Supplementary Information). A similar observation was made by Schlattl *et al.*,[8] and is in line with the known enrichment of immunity related genes in CNVRs.[25,35] Although it is possible that this observation is a result of working with blood cells, our results appear to be consistent with a growing body of evidence that supports a biological relationship between heritable copy number variation and the immune system. We also observed significant clusters enriched in KEGG pathways and Gene Ontology terms related to glutathione transferase activity and Gene Ontology terms related to the plasma membrane.

### Effect of expression level on experimental power to detect *cis* effects

We postulated that the power to detect true associations will be positively correlated with expression level, as the signal-to-noise ratio increases with increased expression level. This suggests that the power to detect true associations may be improved by limiting transcripts to those with higher expression levels. With this rationale, we subset the tests performed based on gene expression and calculated *q*-values[23] for each subset. Despite our expectation of an improvement, limiting the tests performed to more highly expressed genes did not improve the overall number of significant findings (results not shown). It is worth noting that the rate of positive findings did change, as there is a trade-off between power per test and the number of tests performed. With all transcripts included, 0.13% of the tests (at symmetrical window size 10 Mb) were statistically significant; this rate rose to 0.40% when tests were limited to the top 5% of transcript expression levels.

**Table 1 Summary of tests performed and statistically significant findings at various window sizes**

| | *Tests performed* | | | *Number of significant tests (number positive correlations in parentheses)* | | | | |
|---|---|---|---|---|---|---|---|---|
| *Distance between CNVR and gene TSS* | *Number of tests* | *Number of genes represented* | *Number of CNVRs represented* | *q < 0.01* | *q < 0.1* | *Number of genes significant (q < 0.1)* | *Number of CNVRs significant (q < 0.1)* | *Percent of tests significant (q < 0.1)* |
| Gene entirely contained by CNVR | 157 | 140 | 88 | 24 (23) | 32 (29) | 31 | 27 | 20.4% |
| CNVR overlaps gene | 350 | 301 | 234 | 33 (31) | 45 (36) | 43 | 38 | 12.9% |
| 1 Mb | 11106 | 5837 | 803 | 78 (65) | 141 (98) | 115 | 83 | 1.3% |
| 1.2 Mb | 13320 | 6621 | 824 | 79 (65) | 147 (103) | 117 | 88 | 1.1% |
| 10 Mb | 89893 | 12059 | 909 | 71 (61) | 118 (89) | 97 | 75 | 0.13% |

Abbreviations: CNVR, copy number variable region; Mb, megabase; TSS, transcription start site.

## DISCUSSION

One of the potential mechanisms through which CNVs may exert a causal effect on human health and disease is by altering gene transcription. There is now a large and growing list of CNVs (both recurrent and non-recurrent) associated with human diseases. By cataloging CNVs that are themselves *cis*-acting eQTL, the results of this study will aid in the design and interpretation of future studies.

The current study is subjected to two primary limitations. First, in order to increase the number of regions that could be examined in this study, and for technical reasons, we used a quantitative value representative of copy number instead of (estimated) integer copy number values. Second, we limited the study to 909 CNVs representing ~1.5% of the autosomal genome. A quick look at the Database of Genomic Variants[5] indicates that the portion of the autosomal genome which is likely to be copy number variable is much higher than 1.5%.

In addition to identifying CNVs that affect mRNA expression of nearby genes, we empirically evaluated the effect of expression levels and window size selection on power to detect CNVs that are eQTL. The effect of expression level on power to detect association was moderate compared with what we expected to observe, namely a more pronounced skew toward more highly expressed genes among the significant results. During our enrichment analysis using David Bioinformatics we corrected for potential tissue-specific effects as well as we could. This is an incomplete correction as some genes will be measured with a higher signal-to-noise ratio than others. It is possible that this could cause tissue-specific effects if more highly expressed genes are significant more often. This appears to be the case, but only slightly, and certainly not enough to skew the results of our enrichment analysis to the observed levels of enrichment in immunity and blood related traits.

Significant findings in which genes are themselves copy number variable lend themselves to the most straightforward interpretation. However, nearby *cis*-eQTL are also of interest and can be identified with considerable power. Indeed many researchers are aware of the effect of window size on power to detect eQTL, however the trade-offs involved are not clearly defined in the literature. There is a clear trade-off between newly discoverable *cis*-eQTL and increased multiple testing burden with increasing window sizes. This is not simply a statistical problem, in that there is a real enrichment of eQTL proximal to genes being investigated, and thus there are also truly fewer eQTL with large effects at further distances regardless of multiple testing burden. The sharp increase in the number of eQTL discovered up to 1.2 Mb in this dataset indicates that up to this window size, multiple testing correction is outweighed by newly discovered *cis*-eQTL. Conversely, at larger window sizes multiple testing burden begins to outweigh newly discovered *cis*-eQTL. The optimal window size may vary between studies due to their relative power, but the underlying biology is likely fairly consistent. The shape of the curve in Figure 2 indicates that for the purpose of identifying as many relationships as possible, choosing an overly broad window size is preferable to overly constraining a window size *a priori*. Although the power of individual studies may alter what can be detected, our empirical results indicate that there is a considerable amount of meaningful information to be found by looking at a symmetrical window size up to about 1 Mb from each gene. This also indicates that causal variants for clinically relevant traits may exert their effects on the trait through genes that are fairly distant from their location. Although straightforward, our empirical evaluation of window size selection will serve as a useful guide for future studies.

We discovered up to 117 genes for which transcript expression is significantly associated with copy number variation. The overwhelming majority of these findings is novel, and in many cases involves clinically relevant genes. Up to ~10% of the CNVRs examined in this study were found to be significantly associated with the expression of at least one nearby gene. This suggests a high overall functional role of CNVs in variation of gene expression and, by proxy, trait variation. Most of the significant findings account for moderate (<5%) proportions of variance in gene expression. This is at least partially an effect of allele frequency, which tends to be lower for larger CNVs (presumably because these tend to be deleterious and, hence, will be selected against). These variants tended to account for moderate proportion of gene expression variation is consistent with our expectation that most variants that affect trait variation are either rare or have low-to-moderate effect per allele.

Our results indicate that future studies investigating more comprehensive sets of CNVs with higher resolution data are likely to identify many more CNVs that are eQTL. These findings provide valuable information that will aid in the interpretation of future studies focused on investigating the genetic architecture of human disease.

## CONFLICT OF INTEREST

The authors declare no conflict of interest.

## ACKNOWLEDGEMENTS

This study was supported in part by grants from the National Institutes of Health (DK47482, DK70746, DK053889, HL045222, RR013556, MH059490, and HL80149) and the Department of Defense (DOD PC081025). We thank the participants of the SAFHS for their generous cooperation.

1 Goring HH, Curran JE, Johnson MP *et al*: Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. *Nat Genet* 2007; **39**: 1208–1216.
2 Cookson W, Liang L, Abecasis G, Moffatt M, Lathrop M: Mapping complex disease traits with global gene expression. *Nat Rev Genet* 2009; **10**: 184–194.
3 Grundberg E, Small KS, Hedman AK *et al*: Mapping *cis*- and *trans*-regulatory effects across multiple tissues in twins. *Nat Genet* 2012; **44**: 1084–1089.
4 Lappalainen T, Sammeth M, Friedlander MR *et al*: Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* 2013; **501**: 506–511.
5 Iafrate AJ, Feuk L, Rivera MN *et al*: Detection of large-scale variation in the human genome. *Nat Genet* 2004; **36**: 949–951.
6 Perry GH, Ben-Dor A, Tsalenko A *et al*: The fine-scale and complex architecture of human copy-number variation. *Am J Hum Genet* 2008; **82**: 685–695.
7 Levy S, Sutton G, Ng PC *et al*: The diploid genome sequence of an individual human. *PLoS Biol* 2007; **5**: e254.
8 Schlattl A, Anders S, Waszak SM, Huber W, Korbel JO: Relating CNVs to transcriptome data at fine resolution: assessment of the effect of variant size, type, and overlap with functional regions. *Genome Res* 2011; **21**: 2004–2013.
9 Stranger BE, Forrest MS, Dunning M *et al*: Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* 2007; **315**: 848–853.
10 Gamazon ER, Nicolae DL, Cox NJ: A study of CNVs as trait-associated polymorphisms and as expression quantitative trait loci. *PLoS Genet* 2011; **7**: e1001292.
11 Mitchell BD, Kammerer CM, Blangero J *et al*: Genetic and environmental contributions to cardiovascular risk factors in Mexican Americans. The San Antonio Family Heart Study. *Circulation* 1996; **94**: 2159–2170.
12 Blackburn A, Goring HH, Dean A *et al*: Utilizing extended pedigree information for discovery and confirmation of copy number variable regions among Mexican Americans. *Eur J Hum Genet* 2012; **21**: 404–409.
13 Mefford HC, Eichler EE: Duplication hotspots, rare genomic disorders, and common disease. *Curr Opin Genet Dev* 2009; **19**: 196–204.
14 Gokcumen O, Babb PL, Iskow RC *et al*: Refinement of primate copy number variation hotspots identifies candidate genomic regions evolving under positive selection. *Genome Biol* 2011; **12**: R52.
15 Altshuler DM, Gibbs RA, Peltonen L *et al*: Integrating common and rare genetic variation in diverse human populations. *Nature* 2010; **467**: 52–58.
16 Barnes C, Plagnol V, Fitzgerald T *et al*: A robust statistical method for case-control association testing with copy number variation. *Nat Genet* 2008; **40**: 1245–1252.
17 McCarroll SA, Altshuler DM: Copy-number variation and association studies of human disease. *Nat Genet* 2007; **39**: S37–S42.

18 Eleftherohorinou H, Andersson-Assarsson JC, Walters RG *et al*: famCNV: copy number variant association for quantitative traits in families. *Bioinformatics* 2011; **27**: 1873–1875.
19 Kent WJ, Sugnet CW, Furey TS *et al*: The human genome browser at UCSC. *Genome Res* 2002; **12**: 996–1006.
20 Sanders AR, Goring HH, Duan J *et al*: Transcriptome study of differential expression in schizophrenia. *Hum Mol Genet* 2013; **22**: 5001–5014.
21 Almasy L, Blangero J: Multipoint quantitative-trait linkage analysis in general pedigrees. *Am J Hum Genet* 1998; **62**: 1198–1211.
22 Boerwinkle E, Chakraborty R, Sing CF: The use of measured genotype information in the analysis of quantitative phenotypes in man. I. Models and analytical methods. *Ann Hum Genet* 1986; **50**: 181–194.
23 Storey JD, Tibshirani R: Statistical significance for genomewide studies. *Proc Natl Acad Sci USA* 2003; **100**: 9440–9445.
24 Huang da W, Sherman BT, Lempicki RA: Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 2009; **4**: 44–57.
25 Mills RE, Walter K, Stewart C *et al*: Mapping copy number variation by population-scale genome sequencing. *Nature* 2011; **470**: 59–65.
26 Crowley MA, Mollan TL, Abdulmalik OY *et al*: A hemoglobin variant associated with neonatal cyanosis and anemia. *N Engl J Med* 2011;**364**:1837–1843.
27 Phylipsen M, Amato A, Cappabianca MP *et al*: Two new beta-thalassemia deletions compromising prenatal diagnosis in an Italian and a Turkish couple seeking prevention. *Haematologica* 2009; **94**: 1289–1292.
28 Trent RJ, Bowden DK, Old JM, Wainscoat JS, Clegg JB, Weatherall DJ: A novel rearrangement of the human beta-like globin gene cluster. *Nucleic Acids Res* 1981; **9**: 6723–6733.
29 Townsend DM, Tew KD: The role of glutathione-S-transferase in anti-cancer drug resistance. *Oncogene* 2003; **22**: 7369–7375.
30 Chase MB, Baek SJ, Purtell DC, Schwartz S, Shen RF: Mapping of the human thromboxane synthase gene (TBXAS1) to chromosome 7q34-q35 by two-color fluorescence *in situ* hybridization. *Genomics* 1993; **16**: 771–773.
31 Genevieve D, Proulle V, Isidor B *et al*: Thromboxane synthase mutations in an increased bone density disorder (Ghosal syndrome). *Nat Genet* 2008; **40**: 284–286.
32 Ashburner M, Ball CA, Blake JA *et al*: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000; **25**: 25–29.
33 Kanehisa M, Goto S: KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000; **28**: 27–30.
34 UniProt C: Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Res* 2013; **41**: D43–D47.
35 Conrad DF, Pinto D, Redon R *et al*: Origins and functional impact of copy number variation in the human genome. *Nature* 2010; **464**: 704–712.

Supplementary Information accompanies this paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)