# **Supplementary Material for** Schlattl *et al.*

## Supplementary Text

**Genotyping additional CNVs with CopySeq.** We used CopySeq (Waszak et al. 2010) to genotype 2,843 CNVs greater than 500bp in length (the CNV size cutoff of CopySeq), which the 1000 Genomes Project released without genotype information (Mills et al. 2011). CopySeq is a depth-of-coverage (i.e., read-depth) based algorithm that genotypes given genomic regions by relating their sample-specific read depth to an expected read-depth value inferred by assessing the reference genome for average sequencing coverage, repeat content, and GC content (Waszak et al. 2010). We inferred copy-number genotypes using CopySeq, applied with default parameters, by analyzing low-coverage whole genome Illumina sequence reads released by the 1000 Genomes Project in July 2010 (1000 Genomes Project Consortium, 2010). Specifically, we analyzed SAM/BAM format (Li et al. 2009) files that were generated by mapping DNA reads onto the hg18 assembly of the reference genome using the MAQ algorithm (Li et al. 2008). We note that in rare cases CNVs released as deletions by the 1000 Genomes Project were genotyped as duplications by CopySeq (Table 1). One explanation for this is that a portion of the CNVs released by the 1000 Genomes Project were ascertained by comparison to a population reference (Mills et al. 2011), rather than by comparison to the human reference genome as in CopySeq (distinct reference systems can lead to distinct interpretations as to whether a locus represents a deletion or duplication). CNV genotypes generated with CopySeq are in Supplementary Table 5.

**Inference of associations between SNPs and expressed gene loci**. In order to measure associations between SNP and expressed genes in the 200kb search-range defined through our CNVs, we calculated Spearman correlation coefficients based on our normalized gene expression data, using the same approach as we used when correlating CNVs with expression. When comparing the SNP-based correlations with the CNV-based correlations, we considered a CNV association to be ‘better’ only if no SNP within the search-range displayed an absolute Spearman correlation coefficient higher than that reported for our CNV with the highest absolute Spearman correlation for the same gene (see results in Supplementary Tables 2, 8). We decided to use the best absolute Spearman correlation for this analysis since it provided an unbiased comparison metric between CNV-gene and SNP-gene correlation given the unequal number of samples for which genotype information was available for SNPs compared to CNVs. As a source of SNP genotypes we used the comprehensive genotype lists recently released by the 1000 Genomes Project, July 2010 release (1000 Genomes Project Consortium, 2010); this release encompassed SNP genotype information for 56 of the CEU and 54 of the YRI individuals that were examined in this study.

**Comparison to earlier studies relating CNVs to expression.** We first compared our CNV-gene associations to a study that associated CNVs with the expression of three gene deletions (McCarroll et al. 2006) and confirmed all three associations in our data. Of the ten reported genes we only compared those three that were expressed also in our study. We next compared our results to the analysis of large-scale, CNV-associated eQTLs from Stranger et al. (2007), by collecting the set of unique genes for associations with CNV-clones (*i.e.*, eQTLs from microarray-based CNV calls (Stranger et al. 2007)). Specifically, we assessed all 42 genes reported by Stranger and coworkers that were unambiguously mapped onto our Ensembl gene IDs.

**Refinement of CNV boundaries by read-depth analysis and assessment of dosage compensation effects.** For assessing dosage compensation we considered expressed genes for which a deletion fully encompassing the gene was present in at least 5% of the samples in the YRI or CEU. Since both CNV breakpoint annotations and CNV genotypes may contain errors in rare cases (Waszak et al. 2010; Mills et al. 2011), we sought within all genes of interest for definite evidence concerning the predicted full deletion of these genes. To this end, we used CopySeq (Waszak et al. 2010) to determine the copy-number for each gene locus of interest employing population-scale sequence data (1000 Genomes Project Consortium 2010) in all CEU and YRI samples. Namely, we predicted the copy-number between the start- and end-coordinates for each gene in question, which we refer to as *gene copy-number*. We discarded genes in our dosage compensation analysis if more than half of the samples with copy-number (CN)=1 (according to the respective CNV genotype source) displayed a *gene copy-number* different to 1 for the gene in question, since this would suggest that the gene is actually not heterozygously deleted by the CNV. Additionally we required ≥80% overall concordance between *gene copy-number* and CN between all samples of a population. Furthermore, we discarded genes for which less than 2 samples were found for either CN=1 or CN=2.

We obtained point estimates of relative expression levels (indicated by circles in Figure 6 and by values indicated in Supplementary Table 9) by dividing the median of the normalized expression values from all individuals with CN=1 by the median of the expression number from individuals with CN=2. We further performed bootstrap sampling from the value pairs of genotype and normalized expression value of all individuals with CN=1 or CN=2 and calculated non-parametric BCa bootstrap confidence intervals (Efron 1987) using the R package "bootstrap". When generating Figure 6, we first merged samples from both populations. We did so by first computing the 25% trimmed median of the normalized expression values of samples with CN=1 and of samples with CN=2 separately for both populations. Next, we determined the ratios between the calculated median of CN=1 and CN=2 between the two populations. These two ratios should be very similar for most genes but may deviate because of background noise. We therefore calculated the average of the two ratios for both copy-numbers and used this as a factor to scale the expression values of one of the two populations. Genes with a confidence interval spanning both relative expression values 0.5 and 1 were not included in Figure 6 (we included these in

Supplementary Table 9).

We also performed a careful gene copy-number analysis to assess a bi-allelic deletion CNV predicted to partially disrupt *ULK1*, for which our approach detected a negative correlation between copy-number genotype and *ULK1* expression (Table 2). To examine whether the CNV truly disrupts *ULK1* we applied CopySeq to specifically assessed the 958bp region where the deletion and *ULK1* had been reported to overlap, by inferring copy-number genotypes between the start- and end-coordinates of this overlap region. We compared these newly inferred copy-numbers to the CNV copy-numbers from the respective CNV genotype source, and calculated the concordance between all samples of the population (YRI) in which the CNV-associated eQTL was reported. We found that the genotyping concordance was unexpectedly low (12%), suggesting that the deletion breakpoints were mis-annotated. We further did not identify a negative correlation between *ULK1* expression and the copy-number genotype inferred for the CNV-gene overlap region, but instead identified a weak (not significant) positive correlation, which supported our suspicion that the overlapping CNV-gene segment was actually not responsible for the inferred, negatively correlated, CNV-associated eQTL. By comparison, using CopySeq we confirmed that the remaining parts of the CNV, a 621bp segment lying in non-coding regions, showed a significant negative correlation with *ULK1*. Thus, the previously inferred region of overlap between the examined CNV and ULK1 displays variation in strong dis-agreement with the variation observed for the remaining CNV region, indicating that the deletion we associated with *ULK1* (for which so far only approximate breakpoint coordinates had been reported) has most likely mis-annotated breakpoints.

# Supplementary Tables

**Supplementary Table 1. Gene functional category enrichment analysis for CNV-associated eQTLs.**

| | Genes | P-value | adjusted P-value | FDR |
|---|---|---|---|---|
| **(A) GO-term** | | | | |
| MHC class II protein complex | *HLA-DRB1, HLA-DQB1, HLA-DRB5, HLA-DQA1, HLA-DQA2* | 1.2E-6 | 1.5E-4 | 0.001 |
| MHC protein complex | *HLA-DRB1, HLA-J, HLA-DQB1, HLA-DRB5, HLA-DQA1, HLA-DQA2* | 1.3E-6 | 7.5E-5 | 0.001 |
| antigen processing and presentation of peptide or polysaccharide antigen via MHC class II | *HLA-DRB1, HLA-DQB1, HLA-DRB5, HLA-DQA1, HLA-DQA2* | 7.1E-6 | 0.005 | 0.011 |
| antigen processing and presentation | *HLA-DRB1, HLA-J, HLA-DQB1, HLA-DRB5, HLA-DQA1, HLA-DQA2* | 4.3E-5 | 0.014 | 0.064 |
| **(B) KEGG-pathway** | | | | |
| has05310:Asthma | *HLA-DRB1, HLA-DQB1, HLA-DRB5, HLA-DQA1, AL713890.1* | 2.8E-05 | 0.002 | 0.03 |
| hsa05330:Allograft rejection | *HLA-DRB1, HLA-DQB1, HLA-DRB5, HLA-DQA1, AL713890.1* | 6.6E-05 | 0.002 | 0.07 |
| has05332:Graft-versus-host disease | *HLA-DRB1, HLA-DQB1, HLA-DRB5, HLA-DQA1, AL713890.1* | 9.1E-05 | 0.002 | 0.09 |

All 110 significantly associated genes were analyzed with the DAVID server (http://david.abcc.ncifcrf.gov/) v 6.7 (Huang da et al. 2009) by applying functional annotation clustering for default GO-terms (A) and KEGG-pathways (B). As background, we used all 12,622 genes that were expressed in at least one of the two analyzed populations.

**Supplementary Table 2. Comparison between CNV-associated eQTLs and SNP-associated eQTLs.**

| | | | | |
|---|---|---|---|---|
| **(A) eQTLs reported both by our approach as well as by Montgomery or Pickrell (including SNP-associated eQTLs outside our search-range)** | | | | |
| | #genes | CNV better | SNP better | % CNV better |
| **Montgomery (CEU)** | 23 | 16 | 7 | 70% |
| **Pickrell (YRI)** | 27 | 17 | 10 | 63% |
| **(B) All 1000 Genomes Project (1000GP) SNPs within 200kb of CNV-associated eQTLs that were also reported as eQTLs (genes) by Montgomery or Pickrell** | | | | |
| | #genes | CNV better | SNP better | % CNV better |
| **Montgomery (CEU)** | 24 | 13 | 11 | 54% |
| **Pickrell (YRI)** | 35 | 14 | 21 | 40% |
| **(C) All 1000GP SNPs within 200kb of CNV-associated eQTLs (genes) that were neither reported as eQTLs by Montgomery nor by Pickrell** | | | | |
| | #genes | CNV better | SNP better | % CNV better |
| **Montgomery (CEU)** | 26 | 21 | 5 | 81% |
| **Pickrell (YRI)** | 38 | 22 | 16 | 58% |
| **(D) All 1000GP SNPs within 200kb of CNV-associated eQTLs** | | | | |
| | #genes | CNV better | SNP better | % CNV better |
| **Montgomery (CEU)** | 50 | 34 | 16 | 68% |
| **Pickrell (YRI)** | 73 | 36 | 37 | 49% |
| **Non-redundant sum** | 110 | 63 | 47 | 57% |

The table summarizes results from systematic comparisons of the magnitude of Spearman correlation coefficients with normalized expression values, *i.e.*, comparing best absolute Spearman correlations involving CNVs *vs.* those involving SNPs. As genome-wide surveys of SNP-associated eQTLs have already been performed elsewhere (Pickrell et al. 2010; Montgomery et al. 2010), we focused, in our analysis, on such gene loci that were identified as CNV-associated eQTLs by our approach. In (A), we compared SNP-based correlations with CNV-based correlations by considering SNP-gene pairs reported in Montgomery et al. and Pickrell et al. (including such >200kb apart from genes). We used SNP genotypes from the 1000GP (1000 Genomes Project Consortium, 2010) for this analysis. The analysis in (B) considered all 1000GP SNPs within the 200kb search-range of eQTLs that were both identified by our approach as well as by Montgomery et al. (2010) or by Pickrell et al. (2010). In (C), we examined all 1000GP SNPs in the search-range of genes that were neither identified by Montgomery et al. (2010) nor by Pickrell et al. (2010). (D) summarizes information for all CNV-associated eQTLs (B) and (C) (regardless of their identification in previous SNP-based eQTL surveys).

**Supplementary Table 3: List of CNV-associated eQTLs**. This table displays all significantly associated CNV-gene pairs we identified with our FDR threshold (FDR≤10%). Some genes were associated with more than one CNV in their search range, and *vice versa*. Unless mentioned otherwise, we refer to the most strongly correlated CNV-expression (*i.e.*, CNV-gene) pairs throughout the paper when referring to CNV-associated eQTLs – thereby ranking CNV-gene pairs based on their multiple testing corrected *P*-values.

*(external data table)*

**Supplementary Table 4. SNP sets used in different eQTL surveys**

| Study | Size of SNP set* | Origin of SNP set |
|---|---|---|
| Stranger et al. 2007 | 1.2 million** | HapMap phase I (International HapMap Consortium 2005) |
| Montgomery et al. 2010 | 1.2 million*** | HapMap phase III (Altshuler et al. 2010) |
| Pickrell et al. 2010 | 3.8 million** | HapMap phase II and III (Frazer et al. 2007; Altshuler et al. 2010) |
| 1000 Genomes Project Consortium, 2010 (our analysis made use of a subset of these SNPs, *i.e.* the ones in the search-range of our CNV-associated eQTLs) | 15.3 million** | 1000 Genomes Project (1000 Genomes Project Consortium 2010) |

* denotes size of SNP set, according to publication.
** includes SNPs with minor allele frequency (MAF)>5%, as well as those with MAF≤5% (*i.e.*, rare as well as common SNPs).
*** includes common SNPs with MAF>5% only.

**Supplementary Table 5: CopySeq genotype calls**. The table displays CopySeq genotype calls generated for CNVs recently released without genotype information (Mills et al. 2011). CopySeq requires at least 500 bp of mappable DNA sequence for inferring copy-number genotypes (Waszak et al. 2010), and we thus focused on CNVs meeting this criterion. We inferred copy-number genotypes based on the depth-of-coverage of 1000 Genomes Project Illumina GAII reads (1000 Genomes Project Consortium, 2010). In 38 cases, we removed CNVs when most CopySeq-based copy-number genotype assignments in a population resulted in an uncertain value (denoted "NA").

*(external data table)*

## Supplementary Table 6. Enrichment of large CNVs amongst CNV-associated eQTLs

| | **All CNVs (based on all possible pairwise comparisons)** | **CNVs associated with eQTLs** | **Intergenic CNVs (pairwise comparisons)** | **Intergenic CNVs associated with eQTLs** |
|---|---|---|---|---|
| YRI | Av=7,135; M=1,136 (N=21,328) | Av=23,851; M=3,218 (N=73); *P*<7e-06 | Av=3,186; M=1,220 (N=10,066) | Av=14,750; M=3,121 (N=29); *P*=0.009 |
| CEU | Av=9,432; M=1,174 (N=16,433) | Av=30,467; M=1,590 (N=50); *P*=0.037 | Av=3,665; M=1,205 (N=7,937) | Av=3,188; M=1,400 (N=18); *P*=0.689 |
| YRI+ CEU | Av=8,135; M=1,156 (N=37,761) | Av=24,806; M=1,978 (N=110); *P*<0.0002 | Av=3,397; M=1,212 (N=18,003) | Av=10,553; M=1,554 (N=42); *P*=0.027 |
| | **All deletions (pairwise comparisons)** | **Deletions associated with eQTLs** | **Intergenic deletions (pairwise comparisons)** | **Intergenic deletions associated with eQTLs** |
| YRI | Av=4,876 M=1,060 (N=19,823) | Av=14,460; M=1,978 (N=54); *P*=0.004 | Av=2,841; M=1,175 (N=9,506) | Av=6,790; M=1,920 (N=25); *P*=0.013 |
| CEU | Av=6,839; M=1,062 (N=14,996) | Av=21,344; M=1,400 (N=43); *P*=0.15 | Av=3,139; M=1,120 (N=7,361) | Av=3,122; M=1,268 (N=17); *P*=0.586 |
| YRI+ CEU | Av=5,722; M=1,060 (N=34,819) | Av=15,671; M=1,560 (N=87); *P*=0.017 | Av=2,971; M=1,155 (N=16,867) | Av=4,889; M=1,400 (N=37); *P*=0.04 |
| | **All duplications (pairwise comparisons)** | **Duplications associated with eQTLs** | **Intergenic duplications (pairwise comparisons)** | **Intergenic duplications associated with eQTLs** |
| YRI | Av=36,887; M=5,808 (N=1,505) | Av=50,542; M=29,920 (N=19); *P*=0.059 | Av=9,056; M=2,169 (N=560) | Av=64501; M=15,357 (N=4); P=0.145 |
| CEU | Av=36,492; M=5,134 (N=1,437) | Av=86,434; M=29,920 (N=7); *P*=0.124 | Av=10,394; M=1,923 (N=576) | Av=4,320 M=4320 (N=1); *P*=0.812 |
| YRI+ CEU | Av=36,694; M=5,632 (N=2,942) | Av=59,358; M=29,212 (N=23); *P*=0.049 | Av=9,734; M=2,020 (N=1,136) | Av=52,464; M=13,170 (N=5); *P*=0.077 |

Large duplications and deletions were significantly enriched in CNV-associated eQTLs. When relating CNV sizes to the ‘whole set’ (depicted in column “CNVs (based on all possible pairwise comparisons)”) to generate a “baseline CNV size distribution”, we considered CNVs that were present in the search-range of several (*i.e.*, *X*) genes several times (X) in order to control for the fact that large CNVs were frequently in the search range of multiple genes. As an additional test, we identified significant enrichments of large CNVs also when confining the analysis to a specific data source, namely, the Mills et al. (2011) CNV dataset or the Conrad et al. (2010) CNV dataset (data not shown); therefore, the observed enrichments are independent of the data source. Av: average (mean) in basepairs (bp); M: median (bp); N: number of entries; *P*: *P*-values computed with Kolmogorov-Smirnov tests (significant *P*-values that combine results from CEU and YRI samples were underlined).

## Supplementary Table 7. Enrichment of particular CNV types amongst our list of CNV-associated eQTLs

| Population | CNV type | CNV-associated eQTLs per type | Enrichment relative to entire list of CNVs (corrected *P*-value, if significant) |
|---|---|---|---|
| YRI | biallelic deletion | 47 | 0.9 |
| YRI | biallelic duplication | 14 | 1.9 (*P*=0.002) |
| YRI | multiallelic dups & dels | 7 | 1.9 (*P*=0.006) |
| YRI | multiallelic duplication | 5 | 1.4 |
| CEU | biallelic deletion | 39 | 1 |
| CEU | biallelic duplication | 3 | 0.8 |
| CEU | multiallelic dups & dels | 4 | 1.8 (P=0.02) |
| CEU | multiallelic duplication | 4 | 1.2 |

The data presented in the table shows an enrichment of bi-allelic duplications in the YRI and specifically of multi-allelic CNVs showing signatures of both deletions and duplications (dups & dels) in the YRI and CEU populations amongst CNV-associated eQTLs relative to our entire CNV set. *P*-values indicating significant enrichment, displayed in the rightmost column, were computed based on 10,000 permutations. We also controlled for the occurrence of CNVs in the search range of multiple genes, using the same approach as described in the Supplementary Table 6 caption. Duplications were generally enriched for variants fully overlapping a gene, an enrichment that may account for the enrichment of duplications (including multi-allelic CNVs) among CNV-associated eQTLs.

**Supplementary Table 8. CNVs frequently display a stronger correlation with the expression at eQTL loci than nearby SNPs**

| Type of CNV-gene overlap | Number (percentage) of cases in which CNV correlates better than SNPs with gene expression variation |
|---|---|
| Gene deletion | 7 (100%) |
| Other deletion (intronic, partial, or intergenic) | 43 (51%) |
| Gene duplication | 3 (100%) |
| Other duplication (intronic, partial, or intergenic) | 10 (62.5%) |
| **Total** | **63 (57%)** |

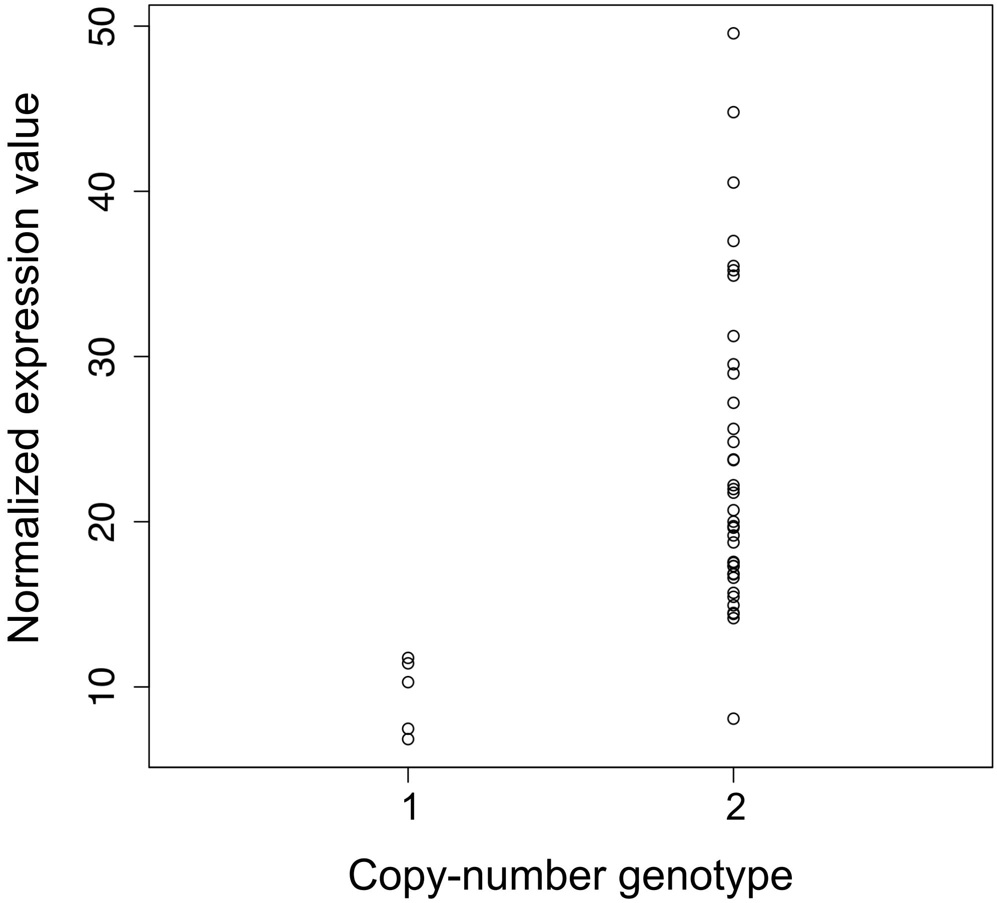
Similar to the analysis displayed in Supplementary Table 2, we correlated normalized expression values with CNVs *vs.* 1000GP SNPs in the 200kb search-range of CNV-associated eQTLs and ranked the results according to the resulting absolute Spearman correlation coefficients. For each gene the SNP and the CNV with the highest absolute Spearman correlation of both populations were compared. The category, 'gene deletion,' encompasses bi-allelic deletions, as well as multi-allelic CNVs showing signatures of both deletions and duplications (dups & dels).

## Supplementary Table 9. Evaluation of dosage compensation effects separately for CEU and YRI populations

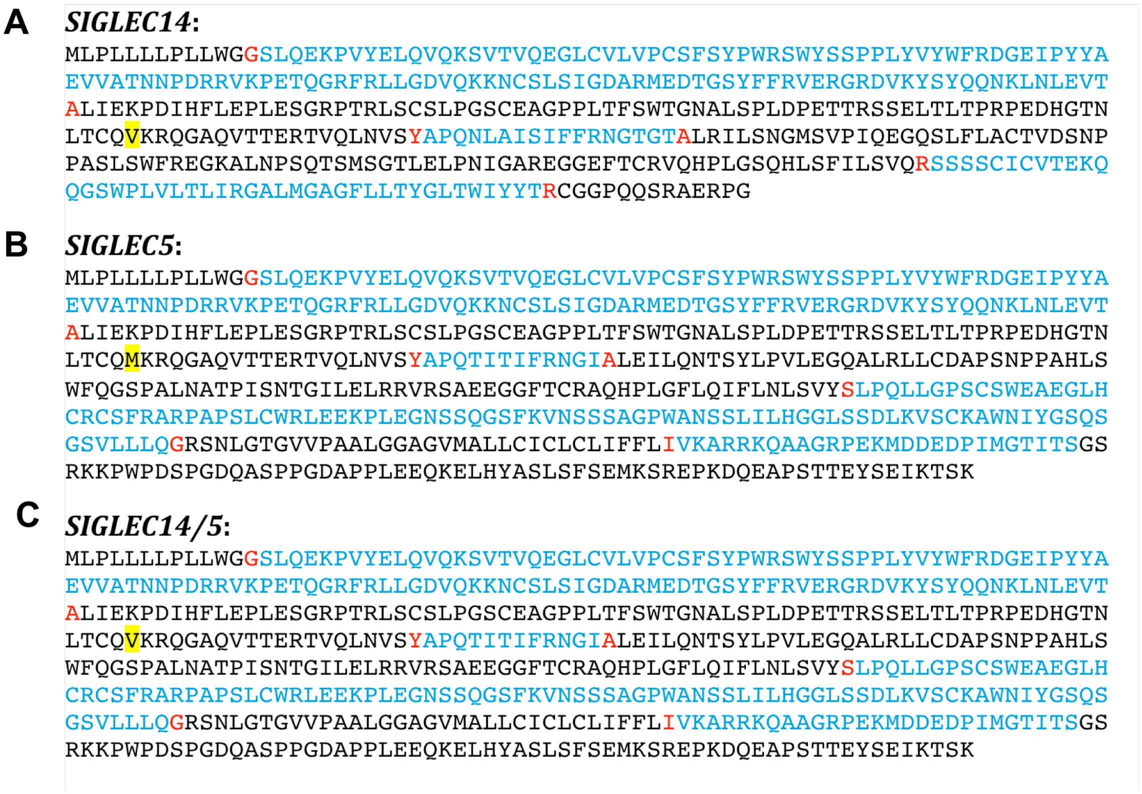
| | RATIO | | 68% down | | 68% up | | 95% down | | 95% up | |
|---|---|---|---|---|---|---|---|---|---|---|
| GENE | CEU | YRI | CEU | YRI | CEU | YRI | CEU | YRI | CEU | YRI |
| AC120036.5-1 | NA | 1.168 | NA | 0.938 | NA | 1.419 | NA | 0.553 | NA | 1.581 |
| AC083906.23 | 0.932 | NA | 0.63 | NA | 1.275 | NA | 0.455 | NA | 1.635 | NA |
| ACOT1 | 0.737 | 1.129 | 0.524 | 0.932 | 1.096 | 1.311 | 0.265 | 0.746 | 1.614 | 1.516 |
| APOBEC3BP | 0.466 | 0.494 | 0.373 | 0.362 | 0.528 | 0.534 | 0.291 | 0.326 | 0.578 | 0.649 |
| ARHGAP11B | 0.691 | NA | 0.451 | NA | 1.056 | NA | 0.397 | NA | 1.409 | NA |
| CDK11A | 0.506 | 0.514 | 0.427 | 0.378 | 0.59 | 0.554 | 0.342 | 0.31 | 0.66 | 0.659 |
| CRYBB2P1 | 0.73 | 0.466 | 0.397 | 0.298 | 0.902 | 0.634 | 0.136 | 0.284 | 1.064 | 0.658 |
| D87018.1-3 | 0.083 | NA | 0.046 | NA | 0.501 | NA | 0 | NA | 1.051 | NA |
| FAM86DP | 0.451 | NA | 0.381 | NA | 0.509 | NA | 0.289 | NA | 0.608 | NA |
| GSTM1 | 0.348 | 0.478 | 0.2 | 0.421 | 0.622 | 0.518 | 0.158 | 0.365 | 1.05 | 0.584 |
| HLA-DRB5 | 0.554 | 0.37 | 0.464 | 0.272 | 0.681 | 0.472 | 0.413 | 0.217 | 0.912 | 0.554 |
| IGLL3 | 0.726 | 1.066 | 0.366 | 0.198 | 1.006 | 1.951 | 0 | 0.134 | 1.417 | 2.162 |
| LRP5L | 0.693 | 0.73 | 0.574 | 0.439 | 0.772 | 1.027 | 0.491 | 0.404 | 1.13 | 1.129 |
| PI4KAP1 | 0.592 | 0.293 | 0.431 | 0.229 | 0.817 | 0.372 | 0.269 | 0.182 | 0.958 | 0.461 |
| RHD | NA | 0.61 | NA | 0.227 | NA | 0.771 | NA | 0.161 | NA | 0.966 |
| SC-9C5.12 | 2.293 | 0.708 | 1.033 | 0.521 | 10.494 | 0.911 | 0.503 | 0.342 | 22.1 | 1.193 |
| UGT2B17 | 0.528 | 0.426 | 0.436 | 0.341 | 0.66 | 0.483 | 0.346 | 0.278 | 0.82 | 0.625 |
| ZNF280B | 0.599 | 0.605 | 0.528 | 0.516 | 0.772 | 0.649 | 0.444 | 0.389 | 0.855 | 0.773 |

The table shows detailed information for genes tested for dosage-compensation effects (see Supplementary text). Results are displayed separately for the CEU and YRI samples. The ‘ratio’ is given by the median normalized expression values of samples with copy-number 1 divided by those of samples with copy-number 2. The subsequent 4 columns provide the lower and upper limits of the 68% and 95% confidence intervals. “NA”: gene was not expressed in a given population or CNV did not display copy-number variation in at least 5% of the individuals (see Supplementary text).

# Supplementary Figures



**Supplementary Figure 1. Relationship between gene copy number and expression in a gene locus previously associated with type 2 diabetes.** Relationship between gene copy number and expression at the *CDK11A* locus, a locus at which particular SNPs have previously been associated with type 2 diabetes (see main text). Normalized expression values were recorded in the CEU.



A *SIGLEC14*:

```
MLPLLLLPLLWGGSLQEKPVYELQVQKSVTVQEGLCVLVPCSFSYPWRSWYSSPPLYVYWFRDGEIPYYA
EVVATNNPDRRVKPETQGRFRLLGDVQKKNCSLSIGDARMEDTGSYFFRVERGRDVKYSYQQNKLNLEVT
ALIEKPDIHFLEPLESGRPTRLSCSLPGSCEAGPPLTFSWTGNALSPLDPETTRSSELTLTPRPEDHGTN
LTCQVKRQGAQVTTERTVQLNVSYAPQNLAISIFFRNGTGTALRILSNGMSVPIQEGQSLFLACTVDSNP
PASLSWFREGKALNPSQTSMSGTLELPNIGAREGGEFTCRVQHPLGSQHLSFILSVQRSSSSCICVTEKQ
QGSWPLVLTLIRGALMGAGFLLTYGLTWIYYTRCGGPQQSRAERPG
```

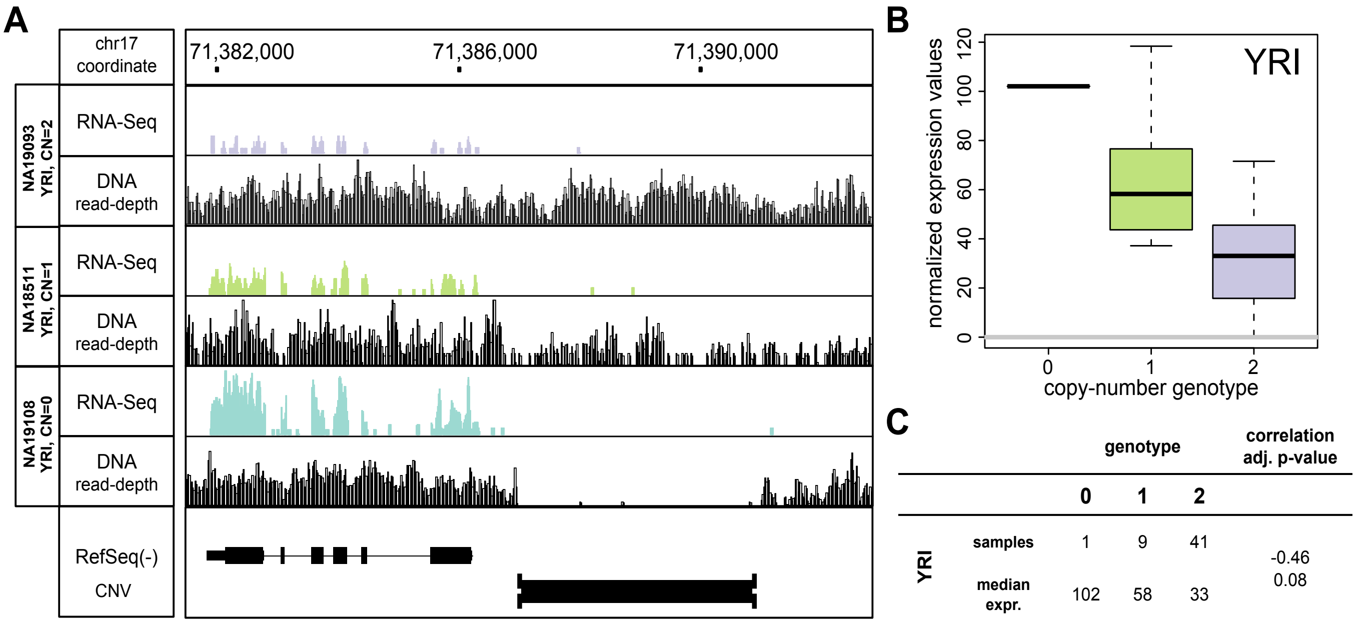
B *SIGLEC5*:

```
MLPLLLLPLLWGGSLQEKPVYELQVQKSVTVQEGLCVLVPCSFSYPWRSWYSSPPLYVYWFRDGEIPYYA
EVVATNNPDRRVKPETQGRFRLLGDVQKKNCSLSIGDARMEDTGSYFFRVERGRDVKYSYQQNKLNLEVT
ALIEKPDIHFLEPLESGRPTRLSCSLPGSCEAGPPLTFSWTGNALSPLDPETTRSSELTLTPRPEDHGTN
LTCQMKRQGAQVTTERTVQLNVSYAPQTITIFRNGIALEILQNTSYLPVLEGQALRLLCDAPSNPPAHLS
WFQGSPALNATPISNTGILELRRVRSAEEGGFTCRAQHPLGFLQIFLNLSVYSLPQLLGPSCSWEAEGLH
CRCSFRARPAPSLCWRLEEKPLEGNSSQGSFKVNSSSAGPWANSSLILHGGLSSDLKVSCKAWNIYGSQS
GSVLLLQGRSNLGTGVVPAALGGAGVMALLCICLCLIFFLIVKARRKQAAGRPEKMDDEDPIMGTITSGS
RKKPWPDSPGDQASPPGDAPPLEEQKELHYASLSFSEMKSREPKDQEAPSTTEYSEIKTSK
```

C *SIGLEC14/5*:

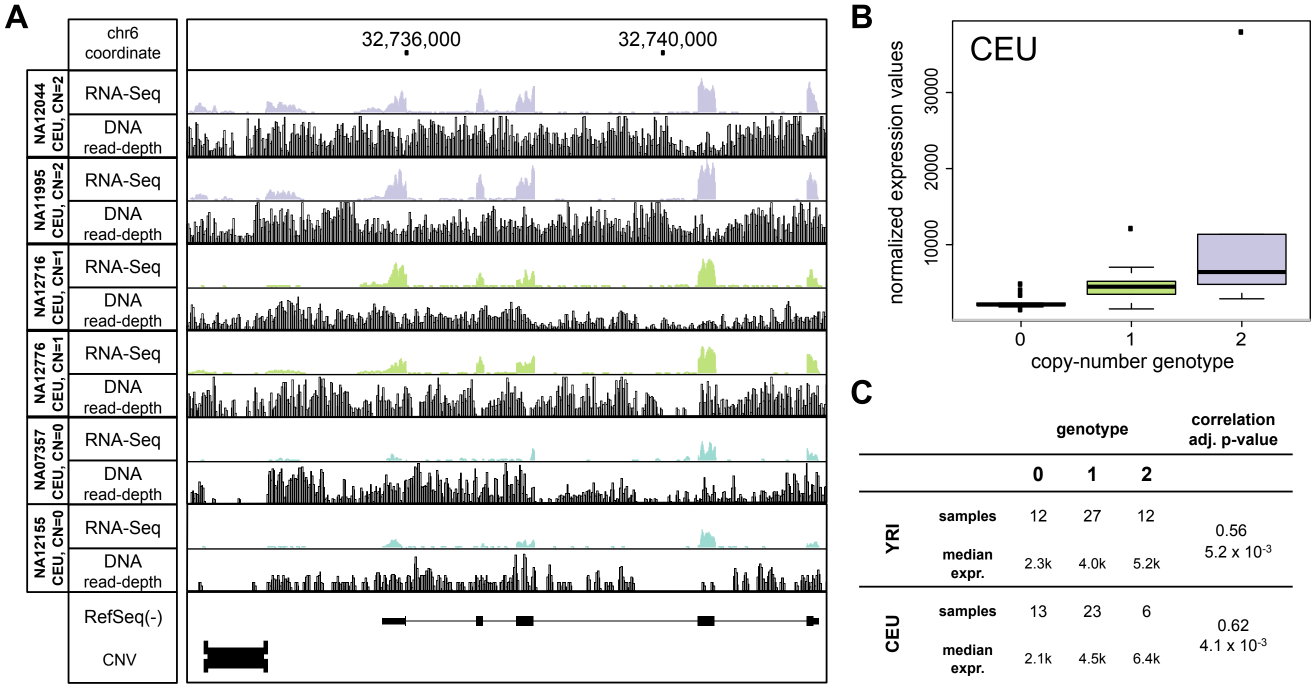
```
MLPLLLLPLLWGGSLQEKPVYELQVQKSVTVQEGLCVLVPCSFSYPWRSWYSSPPLYVYWFRDGEIPYYA
EVVATNNPDRRVKPETQGRFRLLGDVQKKNCSLSIGDARMEDTGSYFFRVERGRDVKYSYQQNKLNLEVT
ALIEKPDIHFLEPLESGRPTRLSCSLPGSCEAGPPLTFSWTGNALSPLDPETTRSSELTLTPRPEDHGTN
LTCQVKRQGAQVTTERTVQLNVSYAPQTITIFRNGIALEILQNTSYLPVLEGQALRLLCDAPSNPPAHLS
WFQGSPALNATPISNTGILELRRVRSAEEGGFTCRAQHPLGFLQIFLNLSVYSLPQLLGPSCSWEAEGLH
CRCSFRARPAPSLCWRLEEKPLEGNSSQGSFKVNSSSAGPWANSSLILHGGLSSDLKVSCKAWNIYGSQS
GSVLLLQGRSNLGTGVVPAALGGAGVMALLCICLCLIFFLIVKARRKQAAGRPEKMDDEDPIMGTITSGS
RKKPWPDSPGDQASPPGDAPPLEEQKELHYASLSFSEMKSREPKDQEAPSTTEYSEIKTSK
```

**Supplementary Figure 2**. ***SIGLEC14/5* fusion gene coding regions.** The formation of the fusion gene *SIGLEC14/5* involved a deletion (chr19:56,824,392-56,840,815; hg18) fusing exons 1, 2, and 3 of *SIGLEC14* with exons 4 to 9 of *SIGLEC5*. Blue and black colors indicate alternate exons in *SIGLEC14* (**A**), *SIGLEC5* (**B**), and the *SIGLEC14/5* fusion gene (**C**). Red indicates amino acids encoded across a splice junction. Yellow highlighting indicates an amino acid in exon 3, which is the single amino acid distinguishing the *SIGLEC5* and *SIGLEC14/5* coding regions. We examined the ancestral state of the region spanning *SIGLEC5* and *SIGLEC14* by evaluating the presence or absence of the respective region in the orangutan, chimpanzee, and macaque in the UCSC browser (http://genome.ucsc.edu/). In particular, the corresponding orthologous region is present in all three primate genomes, which suggests that the locus containing both *SIGLEC5* and *SIGLEC14* corresponds to the ancestral locus, whereas *SIGLEC14/5* was recently formed in humans involving a deletion.



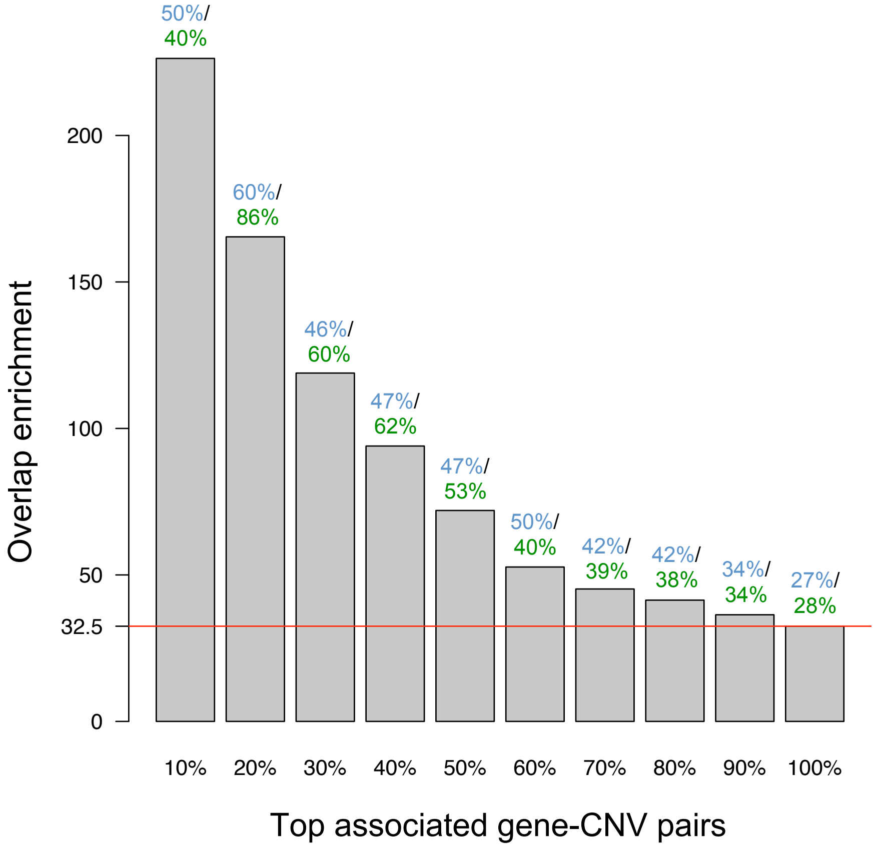
| | | genotype | | | correlation<br>adj. p-value |
|---|---|---|---|---|---|
| | | 0 | 1 | 2 | |
| YRI | samples | 1 | 9 | 41 | -0.46<br>0.08 |
| | median expr. | 102 | 58 | 33 | |

**Supplementary Figure 3. Deletion upstream of *TRIM47* associated with the gene's expression**. (A) A bi-allelic deletion (chr17:71,386,845-71,391,026) upstream of the *TRIM47* gene is associated with *TRIM47* expression, with the transcript abundance negatively correlating with the copy-number genotype. The black bar in the lower panel denotes the deletion, with a vertical dashed line indicating the mapped breakpoints. (B) Correlations between CNV genotype and normalized gene expression values for *TRIM47.* (C) Summary of observed sample abundance, median normalized expression value, and correlation *P*-values in the YRI samples. The analyzed CEU samples did not display a CNV at this locus.

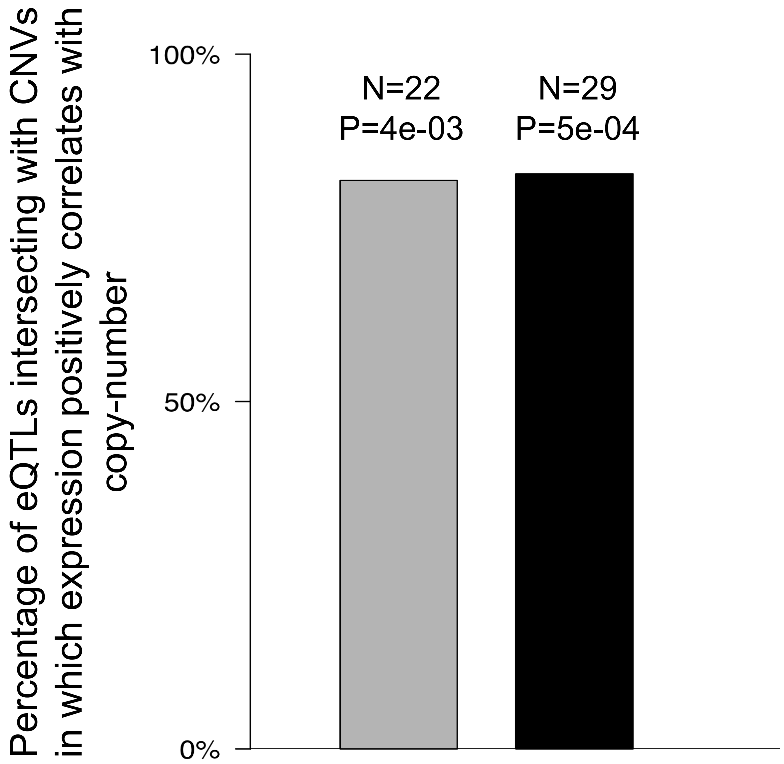


| | | genotype | | | correlation adj. p-value |
|---|---|---|---|---|---|
| | | **0** | **1** | **2** | |
| **YRI** | **samples** | 12 | 27 | 12 | 0.56<br>$5.2 \times 10^{-3}$ |
| | **median expr.** | 2.3k | 4.0k | 5.2k | |
| **CEU** | **samples** | 13 | 23 | 6 | 0.62<br>$4.1 \times 10^{-3}$ |
| | **median expr.** | 2.1k | 4.5k | 6.4k | |

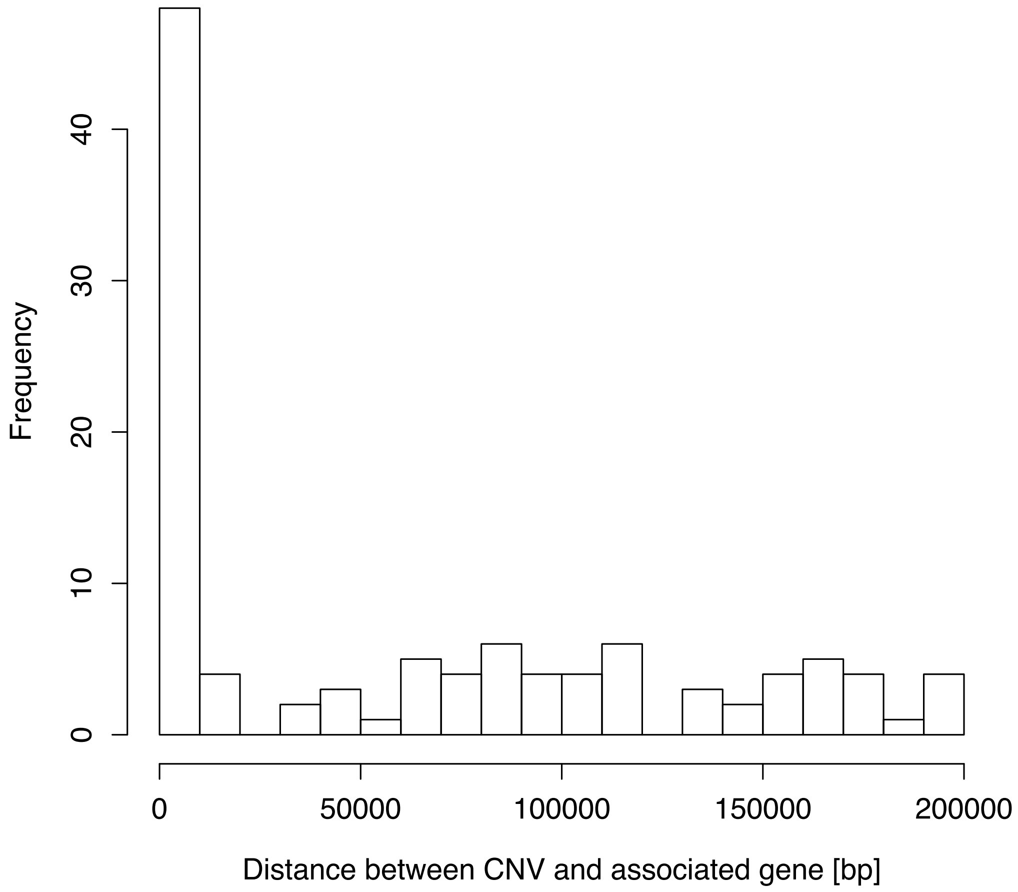
**Supplementary Figure 4. CNV downstream of *HLA-DQB1* is associated with *HLA-DQB1* gene expression.** (A) A bi-allelic deletion (chr6:32,726,466-32,728,025) downstream of *HLA-DQB1* is associated with the gene's expression. (B) Correlations between copy-number genotype and normalized gene expression values for *HLA-DQB1*, measured in the CEU population. The same CNV was also significantly positively correlated with *HLA-DQA1* expression in the CEU individuals and was borderline significant (adjusted *P*-value: 0.18, raw Spearman *P*-value<0.003) in the YRI. Both *HLA-DQB1* and *HLA-DQA1* have been associated with narcolepsy (reviewed by Maret el al. 2005). (C) Summary of observed sample abundance, median normalized expression value, and correlation *P*-values.



**Supplementary Figure 5. Enrichment of overlapping CNV-gene pair associations between the YRI and the CEU populations.** Comparisons between associated CNV-gene pairs identified in the YRI and the CEU populations revealed that significantly associated CNV-genes pairs identified in both populations are strongly enriched (by 32.5x) when compared to the expected number of shared pairs. The enrichment steadily increased when limiting the comparison to upper percentiles of the set of significant CNV-gene associations. Bars display the enrichment factors for different percentiles. Figures in blue indicate the percentage of associated unique YRI genes also observed in the CEU individuals, whereas green indicates the fraction of associated unique CEU genes observed with the YRI individuals for each percentile. To generate this figure we required the genes to be associated with the same CNV in both populations for a valid concordance; further, we considered genes only if they were expressed in both populations; also, we considered CNVs only if they were neither monomorphic in the CEU nor the YRI populations. We considered all CNV-gene pairs for this comparison and did not select for the strongest associated CNV for each gene.



**Supplementary Figure 6. Percentage of positive correlations between the copy-number genotype of exonic-sequence-affecting CNVs associated with an eQTL and the measured expression at these loci**. The black bar shows the percentage of CNV-associated eQTLs with genes disrupted or duplicated by a CNV, which displayed a positive correlation between the CNVs' copy-number genotype and the genes' normalized expression values. The gray bar corresponds to the non-redundant percentage of positive correlations*, i.e.*, only accounting for associations that were observed both in the CEU samples and the YRI samples once (the underlying data are summarized in detail in Table 2). The observed enrichment of positive correlations between copy-number and expression was significant (using a binomial test (two-sided) with success probability set to *p*=0.5; computed *P*-values are displayed on top of each bar). No significant enrichment of positive or negative correlations was observed for other classes of CNV-gene overlap *(e.g.*, intergenic CNVs; Table 2 and data not shown).



**Supplementary Figure 7. Distribution of distances between CNVs involved in CNV-associated eQTLs and their respective associated expressed genes.** The median of the distance distribution was 41.8kb.

## References of the Supplementary Information

1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. *Nature* **467**(7319): 1061-1073.

Altshuler, D.M., Gibbs, R.A., Peltonen, L., Dermitzakis, E., Schaffner, S.F., Yu, F., Bonnen, P.E., de Bakker, P.I., Deloukas, P., Gabriel, S.B. et al. 2010. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**(7311): 52-58.

Efron, B. 1987. Better bootstrap confidence intervals. *Journal of the American Statistical Association* **82**(397): 171.

Frazer, K.A. Ballinger, D.G. Cox, D.R. Hinds, D.A. Stuve, L.L. Gibbs, R.A. Belmont, J.W. Boudreau, A. Hardenbol, P. Leal, S.M. et al. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**(7164): 851-861.

Huang da, W., Sherman, B.T., and Lempicki, R.A. 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**(1): 44-57.

International HapMap Consortium. 2005. A haplotype map of the human genome. *Nature* **437**(7063): 1299-1320.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**(16): 2078-2079.

Li, H., Ruan, J., and Durbin, R. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* **18**(11): 1851-1858.

Maret, S., and Tafti, M. 2005. Genetics of narcolepsy and other major sleep disorders. *Swiss Med Wkly* **135**(45-46): 662-665.

McCarroll, S.A., Hadnott, T.N., Perry, G.H., Sabeti, P.C., Zody, M.C., Barrett, J.C., Dallaire, S., Gabriel, S.B., Lee, C., Daly, M.J. et al. 2006. Common deletion polymorphisms in the human genome. *Nat Genet* **38**(1): 86-92.

Mills, R.E., Walter, K., Stewart, C., Handsaker, R.E., Chen, K., Alkan, C., Abyzov, A., Yoon, S.C., Ye, K., Cheetham, R.K. et al. 2011. Mapping copy number variation by population-scale genome sequencing. *Nature* **470**(7332): 59-65.

Stranger, B.E., Forrest, M.S., Dunning, M., Ingle, C.E., Beazley, C., Thorne, N., Redon, R., Bird, C.P., de Grassi, A., Lee, C. et al. 2007. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315**(5813): 848-853.

Waszak, S.M., Hasin, Y., Zichner, T., Olender, T., Keydar, I., Khen, M., Stutz, A.M., Schlattl, A., Lancet, D., and Korbel, J.O. 2010. Systematic inference of copy-number genotypes from personal genome sequencing data reveals extensive olfactory receptor gene content diversity. *PLoS Comput Biol* **6**(11): e1000988.