Population Differences in Transcript-Regulator Expression Quantitative Trait Loci

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Published Version</td>
<td>doi:10.1371/journal.pone.0034286</td>
</tr>
<tr>
<td>Citable link</td>
<td><a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:10058898">http://nrs.harvard.edu/urn-3:HUL.InstRepos:10058898</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at <a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA">http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA</a></td>
</tr>
</tbody>
</table>
Population Differences in Transcript-Regulator Expression Quantitative Trait Loci

Pierre R. Bushel1,2, Ray McGovern4, Liwen Liu2,3, Oliver Hofmann4,5, Ahsan Huda2,6, Jun Lu2,6, Winston Hide4,5, Xihong Lin5

1 Biostatistics Branch, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina, United States of America, 2 Microarray and Genome Informatics, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina, United States of America, 3 Microarray Core, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina, United States of America, 4 Bioinformatics Core, Harvard School of Public Health, Boston, Massachusetts, United States of America, 5 Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts, United States of America, 6 SRA International Inc., Research Triangle Park, North Carolina, United States of America

Abstract
Gene expression quantitative trait loci (eQTL) are useful for identifying single nucleotide polymorphisms (SNPs) associated with diseases. At times, a genetic variant may be associated with a master regulator involved in the manifestation of a disease. The downstream target genes of the master regulator are typically co-expressed and share biological function. Therefore, it is practical to screen for eQTLs by identifying SNPs associated with the targets of a transcript-regulator (TR). We used a multivariate regression with the gene expression of known targets of TRs and SNPs to identify TReQTLs in European (CEU) and African (YRI) HapMap populations. A nominal p-value of \( <1 \times 10^{-6} \) revealed 234 SNPs in CEU and 154 in YRI as TReQTLs. These represent 36 independent (tag) SNPs in CEU and 39 in YRI affecting the downstream targets of 25 and 36 TRs respectively. At a false discovery rate (FDR) = 45%, one cis-acting tag SNP (within 1 kb of a gene) in each population was identified as a TReQTL. In CEU, the SNP (rs16858621) in Pcx1 was found to be associated with the genes regulated by CREM whereas in YRI, the SNP (rs16909324) was linked to the targets of miRNA hsa-miR-125a. To infer the pathways that regulate expression, we ranked TReQTLs by connectivity within the structure of biological process subtrees. One TReQTL SNP (rs3790904) in CEU maps to Lphn2 and is associated (nominal \( p = 8.1 \times 10^{-7} \)) with the targets of the X-linked breast cancer suppressor Foxp3. The structure of the biological process subtree and a gene interaction network of the TReQTL revealed that tumor necrosis factor, NF-kappaB and variants in G-protein coupled receptors signaling may play a central role as communicators in Foxp3 functional regulation. The potential pleiotropic effect of the Foxp3 TReQTLs was gleaned from integrating mRNA-Seq data and SNP-set enrichment into the analysis.

Introduction
Phenotypic differences between populations have been shown to be associated with variation in genes, the epigenome, the environment and quantitative traits. Gene expression has been used as a quantitative phenotypic trait to locate regions in the genome that have polymorphisms governing differential transcription within populations [1,2,3,4]. This type of inference termed expression quantitative trait loci (eQTL) analysis has been used in genome-wide association studies (GWAS) to map single nucleotide polymorphisms (SNPs) to regions that affect gene expression [5]. Recently it has been shown that SNPs associated with a phenotypic trait are more likely to be eQTLs [6]. The advantage of understanding the contribution of genetic variations on the expression of genes has major implications on the manner in which pharmaceuticals are personalized for an individual and how complex diseases are investigated.

A typical eQTL approach entails modeling the expression of a single gene as a response variable with the genotypes of a single SNP as the predictor variable. Variants of eQTL modeling take the form of a pathway, network component, sparse factor, cluster or the average of a group of co-expressed genes as the response variable and/or predict the expression according to a set of SNPs selected by LASSO, canonical correlation or interval mapping [7,8,9,10,11,12,13,14]. The goal is to determine if there are “eQTL hotspots” [15] where a SNP leads to widespread changes in the expression of genes that are coordinately regulated. Hallmark examples of the power of eQTL analysis for determination of population differences are illustrated by several recent bodies of work. For example, several investigators have demonstrated the robustness of eQTLs to discern variation in gene expression between populations due to environmental exposures or geographic ancestry [16,17,18,19]. Others have shown that gene expression can vary according to particular genotypes, chemical agents and factors such as tissue type, gender, genotype and age [20,21,22,23,24,25]. Also, many have successfully linked genetic variants to transcriptional patterns within ethnic groups [26] although batch effects and biological noise confounding the
differences between the populations can distort the interpretation of the results [27,20]. For instance, the SCAN database [29] is a catalogue of the association of a given SNP to variations in gene expression between Yoruba in Ibadan, Nigeria (African: YRI) and CEPH-Utah residents with ancestry from northern and western Europe (European: CEU) HapMap populations [30]. These variants can be cis- or trans-acting whereby the effect is situated proximal to the expressed gene or it is located elsewhere in the genome, respectively. Although it has been suggested that SNPs residing in transcription factors (TFs) have no significant attributable effect on gene expression variation [11], it is unknown whether a variant that affects the genes regulated by a TF operates through a system of regulated pathways. Therefore, a more comprehensive way to better understand the genetic component of variation in gene expression within and between populations is to address the problem on a systems biology level. In other words, on a genome-wide scale, simultaneously model the expression of genes that are downstream targets (DSTs) of a transcript-regulator (TR) to perform genome-wide associations for SNPs that are genetically-wired regulatory mechanisms. We used a multivariate approach to leverage the covariance of gene expression differences between populations through genetically-wired regulatory mechanisms. We used a multivariate linear regression to model the DSTs of TRs with SNPs typed in CEU and YRI populations. The DSTs of the TRs were obtained from the TRANSFAC® knowledgebase - a biological resource that catalogs experimentally-proven binding and regulation of genes by various TRs [31,32]. For all TRs with two or more DSTs, we detected 234 SNPs in CEU and 154 in YRI representing 36 and 39 independent (tag) SNPs as TReQTLs and affecting the DSTs of 25 and 36 TRs respectively. The expression of the DSTs of 24 TRs was associated with SNPs in both populations. Mapping to within 1 kb of a gene and controlling for multiple testing revealed one cis-acting tag SNP in each population as a TReQTL. In CEU, a TReQTL SNP was found to be associated with the DSTs of the

X-linked breast cancer suppressor Foxp3 but is not significant in the YRI dataset. The Foxp3 TReQTL SNPs were overrepresented in evolutionary conserved regions (ECRs) of the genome in CEU and enriched in splice junctions (SJs) in YRI.

Results

Analysis Strategy

Typical expression quantitative trait loci (eQTL) analyses take the form of a strategy where a single gene is used as a response variable and individual single nucleotide polymorphisms (SNPs) as the predictor variable to determine if there is association of a particular phenotype with a variant. The correlation of co-regulated genes is not taken into consideration. We used a multivariate approach to leverage the covariance of the gene expression of downstream targets (DSTs) of a transcript-regulator (TR) to perform genome-wide associations for SNPs that are potentially linked to changes in gene expression across genotypes.

The genotype data (phase-II, release 24, forward strand, non-redundant) from the 60 Yoruba in Ibadan, Nigeria (African: YRI) and from the 60 CEPH-Utah residents with ancestry from northern and western Europe (European: CEU) populations were obtained from the International HapMap Project [30]. Gene expression data from the profiling of Epstein-Barr virus (EBV)-transformed lymphoblastoid cell lines from the individuals in each CEU and YRI HapMap population [33] were obtained from the National Center for Biotechnology Information Gene Expression Omnibus (GEO) database under accession number GSE10824. Figure 1 illustrates that using this model, TReQTLs can be identified which are associated with the downstream targets of TRs. The genetic variation attributed to the association are imbedded, and therefore discovered in the network of regulatory pathways that govern the co-regulation behind the phenotypic trait. The TReQTLs may be within a single gene (cis), spread across several genes (trans) or located in regions of unknown biological function. In addition, the case may be that several TReQTLs for the DSTs of TRs may share the same variants or portions of the same variants. To investigate the regulatory component of TReQTLs, we first sort out to determine if two populations (YRI and CEU) had shared or varied signaling transduction mechanisms robust enough for a more refined association analysis.

Figure 1. Strategy to identify transcript-regulator eQTLs (TReQTLs). The gene expression of downstream targets (DSTs) of a transcript-regulator (TR) is used as quantitative traits to associate with individual single nucleotide polymorphisms (SNPs). In some cases the SNPs map to the same gene, different genes, the TR or are intergenic. doi:10.1371/journal.pone.0034286.g001
Downstream Targets of Transcript-regulators

Using the TRANSFAC [31,32] and TRANPATH [34] databases of components of signal transduction and regulatory pathways respectively, 2,743 TRs were mapped from the approximately 9,000 probe sets on the Affymetrix Human HG-Focus Target GeneChip Array. 1,438 signaling pathways were identified as comprised of at least one of the TRs and 78 TRs were mapped to one or more pathways. As shown in Figure 2, 333 TRs were determined to have two or more DSTs. These regulate 1,931 DSTs. The TRs consist of transactors (TFs), cofactors, complexes and miRNAs. Three TFs, all stimulating proteins (Sp), regulate more than 60 DSTs. The median value for the TR DSTs is 3 and the mean is 5.8.

Significant Signaling Transcript-regulators of Individual Populations

To determine the extent of co-regulation of genes within CEU and YRI populations, the correlation of the gene expression of pairs of the genes which are DSTs of TRs was assessed. Significance (p-value<or=0.05) was determined from a non-parametric distribution of correlation scores obtained from 10,000 random cases. The correlation score is the sum of the squares of the Pearson correlations (r) among all pairs of genes determined to be DSTs of a TR (see the Materials and Methods section for more detail). For illustrative purposes, Table 1 reports the significance of the correlation of DSTs of only those TRs where disease-causing SNPs are located in the TR target site. The disease-causing SNPs were obtained from the NHGRI GWAS Catalog (www.genome.gov/gwasudies) and mapped to the TRANSFAC position weight matrix consensus sequence for the binding of at least one of the DSTs of the TR. Interferon-stimulated gene factor 3 (ISGF3), X-box binding protein 1 (XBP1) and hepatocyte nuclear factor 4-alpha (HNF4-α) are significant in the CEU and YRI populations. Signal transducers and activators of transcription −1 (STAT1), activating transcription factor 1 (ATF1) and peroxisome proliferator activating receptor gamma (PPAR-γ) are significant in YRI only whereas upstream transcription factor 1 (USF1), the Sp1:Sp3 complex and the retinoid X receptor alpha (RXR-α):PPAR-γ complex are significant in CEU only.

Figure 2. Distribution of the number of genes as downstream targets (DSTs) of transcript-regulators. The x-axis is the # of genes as DSTs and the y-axis is the count. The table inset is a summary of the frequency distribution for the count of the DSTs (two or more) per TR.

doi:10.1371/journal.pone.0034286.g002
Table 1. Co-regulation of DSTs of TRs where disease-causing SNPs are located in the TR binding site of at least one of the TR DSTs.

<table>
<thead>
<tr>
<th>TR ID</th>
<th>TR Symbol</th>
<th># of DSTs</th>
<th>CEU GCS</th>
<th>CEU p-value</th>
<th>YRI GCS</th>
<th>YRI p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>T00428</td>
<td>ISGF-3</td>
<td>3</td>
<td>0.796</td>
<td>0.0055</td>
<td>0.328</td>
<td>0.0532</td>
</tr>
<tr>
<td>T00221</td>
<td>E2F-DP</td>
<td>7</td>
<td>1.374</td>
<td>0.3006</td>
<td>1.463</td>
<td>0.0629</td>
</tr>
<tr>
<td>T09902</td>
<td>XBP-1</td>
<td>2</td>
<td>0.565</td>
<td>0.0009</td>
<td>0.533</td>
<td>0.0001</td>
</tr>
<tr>
<td>T09484</td>
<td>NF-E2p45</td>
<td>4</td>
<td>1.148</td>
<td>0.0073</td>
<td>0.427</td>
<td>0.1163</td>
</tr>
<tr>
<td>T09998</td>
<td>c-Myc</td>
<td>3</td>
<td>0.541</td>
<td>0.0276</td>
<td>0.284</td>
<td>0.0821</td>
</tr>
<tr>
<td>T01804</td>
<td>NF-YA</td>
<td>13</td>
<td>6.222</td>
<td>0.0766</td>
<td>4.749</td>
<td>0.0482</td>
</tr>
<tr>
<td>T04759</td>
<td>STAT1</td>
<td>11</td>
<td>2.076</td>
<td>0.9201</td>
<td>3.452</td>
<td>0.0482</td>
</tr>
<tr>
<td>T09328</td>
<td>usf1</td>
<td>6</td>
<td>1.769</td>
<td>0.0268</td>
<td>0.607</td>
<td>0.4227</td>
</tr>
<tr>
<td>T10359</td>
<td>sp1:sp3</td>
<td>2</td>
<td>0.252</td>
<td>0.0316</td>
<td>0.064</td>
<td>0.2137</td>
</tr>
<tr>
<td>T00167</td>
<td>ATF-2-xbb4</td>
<td>3</td>
<td>0.078</td>
<td>0.7219</td>
<td>0.073</td>
<td>0.6045</td>
</tr>
<tr>
<td>T03828</td>
<td>HNF-4alpha</td>
<td>12</td>
<td>6.445</td>
<td>0.0187</td>
<td>4.570</td>
<td>0.0222</td>
</tr>
<tr>
<td>T04670</td>
<td>MafG</td>
<td>2</td>
<td>0.284</td>
<td>0.0250</td>
<td>0.300</td>
<td>0.0062</td>
</tr>
<tr>
<td>T00968</td>
<td>ATF-1</td>
<td>4</td>
<td>0.544</td>
<td>0.1565</td>
<td>0.564</td>
<td>0.0459</td>
</tr>
<tr>
<td>T05351</td>
<td>PPARgamma</td>
<td>2</td>
<td>0.193</td>
<td>0.0660</td>
<td>0.184</td>
<td>0.0339</td>
</tr>
<tr>
<td>T08618</td>
<td>RXR-alpha:PPARgamma</td>
<td>2</td>
<td>0.237</td>
<td>0.0395</td>
<td>0.006</td>
<td>0.7065</td>
</tr>
</tbody>
</table>

GCS – Group correlation score. The disease-causing SNPs were obtained from the NHGRI GWAS Catalog (Available at: www.genome.gov/gwastudies. Accessed 3/3/2010) with selected SNP-trait associations limited to those with \( p \)-values \( \leq 1 \times 10^{-5} \).

doi:10.1371/journal.pone.0034286.t001

Figure 3. Manhattan plot displaying distribution of TReQTLs. The x-axis is the relative position of the SNPs across the genome in Mb. The chromosomes are illustrated by alternating shaded and unshaded sections of the plot. The order of the chromosomes is from #1 to #22 from left to right. The y-axis represents the \(-\log_{10} p\)-value of the SNP association with the gene expression of DSTs of the TRs. p-values of SNPs from CEU are denoted as red circles, p-values of SNPs from YRI are denoted as green triangles. For visualization purposes, only SNPs having a p-value < 0.002 in any of the 333 TRs are plotted.

doi:10.1371/journal.pone.0034286.g003
of the DSTs of 24 TRs was associated with SNPs in both populations (Table 2). These were SNPs mapped within or in proximity to genes involved in transcription regulation, cell communication, transport, kinase activity, growth and development. Interestingly, several of the TReQTL SNPs in CEU are mapped to pseudogenes.

Cohesive TReQTL Biological Process Subtrees Reveal Descriptive Molecular Events

Each TReQTL is comprised of a SNP, a TR and a set of DSTs. Each constituent is associated with a gene. We mapped the constituents, except for cases where the TR is a miRNAs, to genes and then determined the GO biological process term each was annotated to. The collection of terms was then used to construct a GO biological process subtree. All ancestors of a term were included in the subtree. Our adjusted cohesion score (ACS) is an ad hoc way to i) measure the amount of connectivity between terms, ii) account for the significance of the TReQTL and iii) consider the average number of paths per term. As listed in Table 3, the top ranked TReQTLs have the more cohesive subtrees and are more descriptive with respect to the term with most paths associated with it (The full list is in supplemental material Table S2 (CEU) Table S3 (YRI)). These associations within the subtree can lead to new insight into the possible role of the TReQTL SNP in the pathophysiology of diseases. For instance, in YRI, the rs12258754 allelic variant is associated with the DSTs of activating transcription factor 3 (ATF3) and produced a subtre with vascular smooth muscle cell (VSMC) contraction as the granular biological process node (Figure 5A). Interestingly, in CEU, four SNPs associated with the variation of expression for the DSTs of miRNAs hsa-mir-181b-1 (MI0000270) and hsa-mir-181b-2 (MI0000683) are mapped to the peptidyl-prolyl cis-trans isomerase A-like pseudo-gene and generated a subtree with synaptic transmission as the most descriptive biological process term (Figure 5B).

TReQTL interaction network

Many of the variants map to Online Mendelian Inheritance in Man (OMIM) associated disease genes (data not shown). For instance, one TReQTL in the CEU population is associated with the DSTs of the X-linked breast cancer suppressor gene Foxp3 (T04280) transfactor [39]. Foxp3 belongs to the Forkhead box family of genes, is located on chromosome X and is essential for the production and normal function of regulatory T-cells. As shown in Figure 6, interleukin 2 (IL2) and colony stimulating factor 2 (CSF2), the DSTs of Foxp3, are two cytokines whose gene expression co-regulation (correlation = 0.56) is significantly associated with the variants of tag SNP rs3790904 (p-value = 8.1 10^-7) which maps to the latrophilin homolog 1 (Lphh1/Lphn2) G-protein receptor (GPRC) gene. This association is not significant in YRI (p-value = 0.89). Other significant SNPs in CEU that are linked to the DST of Foxp3 map to an additional GPRC gene (Lphn3) as well as a membrane ion channel (KCNJ1), a phosphatidic acid phosphatase type 2 enzyme (Ppa2pdc1a) and an uncharacterized gene. An interaction network of Foxp3, its DSTs and genes that the TReQTLs map to revealed tumor necrosis factor (TNF) and NF-kB as central hubs along with carcinomaembryonic antigen-related cell adhesion molecule 3 (CEACAM3), Tgf-beta, and hepatocyte growth factor (HGF) as connectors (Figure 7). Interesting enough, negative regulation of NF-kappaB TF activity, cAMP response element-binding (CREB) activity and T-cell cytokine production/positive regulation of regulatory T-cell differentiation are biological processes within the Foxp3 TReQTL subtree that are highly connected (Data not shown).

Over-representation of the Foxp3 TReQTL SNPs in Genomic Regions

SNPs in multi-species/evolutionary conserved regions (ECRs) may imply similarity of function across species [40] and those within splice junctions (SJ) or splicing enhancers may play a role in gene expression regulation through exon splicing [21,41]. In other words, genetic variants that are related to gene expression differences between populations are more likely to be in genomic regions conserved across species and/or possibly involved in regulating transcription by altering splice forms of transcript messages. As an anecdotal example, we used 472 SNPs from the union of the Foxp3 TReQTL SNPs between CEU (n = 233) and YRI (n = 239) with a more liberal nominal p-value threshold <1*10^-6 to determine over-representation within ECRs and SJs. As displayed in Table 4, using SNP-set enrichment analysis [42], the SNPs are significantly over-represented in 5-way ECRs for CEU (p-value = 0.006) but not YRI (p-value = 0.9) and enriched in SJs for YRI (p-value = 1*10^-5) but not CEU (p-value = 0.9). Fisher exact tests confirm significance of the over-representation of these TReQTL SNPs in 5-way ECRs for CEU (p-value = 0.0079) and in SJs for YRI (p-value = 0.0001).

Discussion

Genetic and transcriptional variations are important key factors in the evolution of biology and the dispensation of diseases. Single nucleotide polymorphisms (SNPs) are one type of DNA sequence alteration that is commonly used as a marker for tracking genetic variation. The allelic frequency of a SNP at a given locus can vary between populations and the genotype may code for a SNP that results in a particular phenotype, trait or disease [43,44,45,46]. Within populations and under certain biological conditions genes are coordinately regulated by transcript-regulators (TRs) such as transcription factors (TFs), cofactors, complexes of TFs and miRNAs (Table 1). These co-expressed genes often times share biological functions and work in concert to mediate cellular events such as biological processes and molecular pathways. Although it has been shown that TFs do not harbor trans-acting variants [11], coupling coordinately regulated genes as a quantitative trait for a loci (eQTL) with the genotype of SNPs as a genome-wide association study (GWAS) can presumably help to elucidate variation in gene expression (TReQTLs) on a genomic and systems biology scale that code for particular phenotypes and complex diseases [9].

Tailoring the GWAS eQTL analysis by considering genes with coordinated expression is of added value to reveal master regulators of transcriptional genetic variation (Figure 1). We used a multivariate linear regression with the gene expression of known downstream targets (DSTs) of TRs (Figure 2) as the response variable and individual SNPs as predictor variables to identify TReQTLs in European (CEU) and African (YRI) HapMap populations. At a nominal p-value threshold of <1*10^-6 we discovered 234 SNPs in CEU and 154 in YRI as putative TReQTLs (Figure 4). These represent 36 and 39 independent (tag) SNPs in CEU and YRI affecting the DSTs of 25 and 36 TRs respectively. Two SNPs (one in each population) are co-acting TReQTLs (within 1 kb of a gene) at a false discovery rate (FDR) of 45%. One of them, a SNP in the peroxin-like 2 (Pmnl2) gene was found in CEU to be highly associated with the DSTs of the cAMP responsive element modulator (CREM) transcript whereas in the YRI dataset, a SNP was linked to the DSTs of miRNA hsa-miR-125a. Although the FDR may seem abnormally high and one would expect at least one if not both of the TReQTLs to be false positives, it can be misleading as others have demonstrated that
Figure 4. Dot plot of significant TReQTLs. A) CEU B) YRI. Each circle represents a TReQTL SNP with a p-value < 1 × 10^{-6}. The x-axis is the relative position of the TReQTL SNPs across the genome in Mb. The chromosomes are illustrated by alternating shaded and unshaded sections of the plot. The order of the chromosomes is from #1 to #22 from left to right. The y-axis is the relative position of the TR across the genome in Mb. The order of the chromosomes is from #1 to #22 from bottom to top. The points were jittered to enhance the display of TReQTLs in close proximity. TReQTLs near the diagonal line have the potential to be cis-regulated.

doi:10.1371/journal.pone.0034286.g004
Table 2. TRs in common between CEU and YRI TReQTLs.

<table>
<thead>
<tr>
<th>TR ID</th>
<th>TR Symbol</th>
<th>CEU SNP</th>
<th>Gene ID CEU SNP</th>
<th>Gene Symbol CEU SNP</th>
<th>YRI SNP</th>
<th>Gene ID YRI SNP</th>
<th>Gene Symbol YRI SNP</th>
</tr>
</thead>
<tbody>
<tr>
<td>T00250</td>
<td>Elk-1</td>
<td>rs9838549</td>
<td>131185</td>
<td>LOC131185</td>
<td>rs965676</td>
<td>6638</td>
<td>SNRPN</td>
</tr>
<tr>
<td>T00255</td>
<td>ENKTF-1</td>
<td>rs10510093</td>
<td>2263</td>
<td>FGFR2</td>
<td>rs6864839</td>
<td>4488</td>
<td>MSX2</td>
</tr>
<tr>
<td>T00498</td>
<td>alpha-enolase</td>
<td>rs1029741</td>
<td>54543</td>
<td>TOMM7</td>
<td>rs12358485</td>
<td>359779</td>
<td>MRP3S3P3</td>
</tr>
<tr>
<td>T00902</td>
<td>XBP-1</td>
<td>rs11668328</td>
<td>129563</td>
<td>DIS3L2</td>
<td>rs611734</td>
<td>27131</td>
<td>SN5X</td>
</tr>
<tr>
<td>T00910</td>
<td>YB-1</td>
<td>rs17586344</td>
<td>1956</td>
<td>EGRF</td>
<td>rs11120212</td>
<td>100050832</td>
<td>LOC100050832</td>
</tr>
<tr>
<td>T01814</td>
<td>panx1-isoform5a</td>
<td>rs3212143</td>
<td>11035</td>
<td>RIPK3</td>
<td>rs10755971</td>
<td>137902</td>
<td>PXDN1</td>
</tr>
<tr>
<td>T01931</td>
<td>ReIβ</td>
<td>rs9610774</td>
<td>29775</td>
<td>CARD10</td>
<td>rs289838</td>
<td>9111</td>
<td>NMI</td>
</tr>
<tr>
<td>T02689</td>
<td>GATA-6</td>
<td>rs13345832</td>
<td>55769</td>
<td>ZNF83</td>
<td>rs2937889</td>
<td>57509</td>
<td>MTUS1</td>
</tr>
<tr>
<td>T04870</td>
<td>MafG</td>
<td>rs1153303</td>
<td>150000</td>
<td>ABCC13</td>
<td>rs12691592</td>
<td>53353</td>
<td>LRP18</td>
</tr>
<tr>
<td>T04953</td>
<td>TFI1A</td>
<td>rs11692860</td>
<td>729009</td>
<td>FTH1P20</td>
<td>rs7755851</td>
<td>5071</td>
<td>PARK2</td>
</tr>
<tr>
<td>T04959</td>
<td>GKLf-isoform1</td>
<td>rs9484664</td>
<td>10042074</td>
<td>LOC10042074</td>
<td>rs16848653</td>
<td>55137</td>
<td>FIGN</td>
</tr>
<tr>
<td>T04996</td>
<td>ZBP89</td>
<td>rs6691852</td>
<td>467</td>
<td>ATF3</td>
<td>rs6549604</td>
<td>5067</td>
<td>CNTN3</td>
</tr>
<tr>
<td>T05324</td>
<td>LXR-alpha/RXR-alpha</td>
<td>rs11152748</td>
<td>6955</td>
<td>TRA-alpha</td>
<td>rs7072859</td>
<td>2894</td>
<td>GRID1</td>
</tr>
<tr>
<td>T06135</td>
<td>p63gamma</td>
<td>rs6670238</td>
<td>51018</td>
<td>RRP15</td>
<td>rs1558561</td>
<td>9717</td>
<td>SEC14LS</td>
</tr>
<tr>
<td>T08465</td>
<td>C/EBPalpha</td>
<td>rs9068</td>
<td>220988</td>
<td>HNRNP43</td>
<td>rs6570819</td>
<td>23328</td>
<td>SASH1</td>
</tr>
<tr>
<td>T08618</td>
<td>ROR-alpha/PPARgamma</td>
<td>rs1311584</td>
<td>150928</td>
<td>PTMAP5</td>
<td>rs4596085</td>
<td>11280</td>
<td>SCN11A</td>
</tr>
<tr>
<td>T09159</td>
<td>pitx2a</td>
<td>rs1983600</td>
<td>9742</td>
<td>IFT140</td>
<td>rs6966641</td>
<td>154664</td>
<td>ABCA13</td>
</tr>
<tr>
<td>T10331</td>
<td>NIF-1</td>
<td>rs7272098</td>
<td>6238</td>
<td>RBBP1</td>
<td>rs1347038</td>
<td>2043</td>
<td>EPH4</td>
</tr>
<tr>
<td>T10852</td>
<td>HIF2Aalpha</td>
<td>rs2741270</td>
<td>248</td>
<td>ALPI</td>
<td>rs28740902</td>
<td>4487</td>
<td>MSX1</td>
</tr>
<tr>
<td>T11264</td>
<td>CP2</td>
<td>rs1020344</td>
<td>100130101</td>
<td>LOC100130101</td>
<td>rs3819726</td>
<td>4121</td>
<td>MAN1A</td>
</tr>
<tr>
<td>T13796</td>
<td>TLS</td>
<td>rs10413078</td>
<td>55333</td>
<td>SYN2BP</td>
<td>rs870181</td>
<td>55275</td>
<td>VPS53</td>
</tr>
<tr>
<td>T14942</td>
<td>hsa-miR-181b</td>
<td>rs17543080</td>
<td>392285</td>
<td>LOC392285</td>
<td>rs10797531</td>
<td>148641</td>
<td>SLC35F3</td>
</tr>
<tr>
<td>T15206</td>
<td>N-Myc</td>
<td>rs2268943</td>
<td>4070</td>
<td>TACSTD2</td>
<td>rs1181164</td>
<td>148979</td>
<td>GLI5</td>
</tr>
<tr>
<td>T15913</td>
<td>RXR-alpha/NR1B1</td>
<td>rs1855625</td>
<td>643954</td>
<td>RPAPA43</td>
<td>rs17238405</td>
<td>4734</td>
<td>NEDD4</td>
</tr>
</tbody>
</table>

doi:10.1371/journal.pone.0034286.t002

adjusting for biases which arise from correlations in eQTL analysis is a major challenge and a substantial overestimation of the number of false positives [47,48,49].

Interestingly enough, the gene expression of the DSTs of 24 TRs was associated with SNPs (albeit different ones) in both populations (Table 2) but the majority differed (Figures 3 and 4). The overlap in the TReQTLs probably reflects the ubiquity of certain basic biological processes such as transcription regulation, cell communication, transport, kinase activity, growth and development. On the otherhand, one TReQTL tag SNP (rs3790904) in the CEU population is associated (p-value = 0.89) with the DSTs of the X-linked breast cancer suppressor gene Foxp3 (Figure 6) but is not significant in YRI (p-value = 0.89). The interaction network of the Foxp3 TReQTL in CEU revealed that tumor necrosis factor (TNF), NF-kappaB and variants in G-protein coupled receptors (GPCR) signaling may play a central role as communicators in Foxp3 functional regulation (Figure 7). Although the Foxp3 tumor suppressor is biologically relevant in the pathogenesis of breast cancer, some have shown that SNPs in the germline of the gene are not associated with the risk of the disease [50]. Our TReQTL analysis reveals other potentially interesting loci which might be causative in the etiology of complex diseases.

Another difference between the two populations based on the TReQTLs was the connectivity of the underlying Gene Ontology (GO) biological processes that the genes of the TReQTL represent (Figure 5). In CEU, several SNPs associated with the variation of expression for the DSTs of two miRNAs (hsa-miR-181b-1 (MI0000270) and hsa-miR-181b-2 (MI0000683)) are mapped to the peptidyl-prolyl cis-trans isomerase pseudo-gene and yields a subtre with synaptic transmission as the more cohesive descriptive GO term (Table 3). The activity of this enzyme has been suggested to be necessary for memory formation and may be involved in complex neurodegenerations such as Alzheimer’s disease [51]. In YRI, a SNP (rs12258754) controlling the variation of expression for the DSTs of activating transcription factor 3 (Atf3) yielded a subtre with vascular smooth muscle cell (VSMC) contraction as the more descriptive GO term (Table 3). Although much is not currently known about the function of Atf3 in VSMCs [52], mutations in the actin, alpha 2 (Acta2) smooth muscle gene have been shown to result in a variety of vascular diseases [53]. Transcriptional networks such as these have recently been shown to hubs with high connectivity and association with controlling higher-ordered biological function such as lipogenesis, lipid trafficking and surfactant homeostasis [54]. Our approach embraces this strategy by using the SNPs within the TReQTLs as an adjudicator for the identification of master regulators of these genetic networks. Although it is expected that a TR and its TReQTLs will share a common signaling pathway, what is not certain is that the SNP associated with the eQTL from the TR and DSTs will reside near or in a gene with biological functionality that forms a cohesive GO biological process subtre. Bear in mind that it is not known where the true regulating TR associated with a candidate TReQTL actually exerts its biological functionality and to date, there is no independent data set with gene expression and genotype calls from another sample of the YRI and CEU.
## Table 3. TReQTL GO subtree cohesiveness.

<table>
<thead>
<tr>
<th>ID of GO Term with Max. of Paths for Term</th>
<th>GO Term</th>
<th>Ave. of Paths per GO Term</th>
<th>Max. of Paths</th>
<th># of Paths</th>
<th>GO Term with # of Edges</th>
<th>SNP</th>
<th>p-value</th>
<th>ACS</th>
</tr>
</thead>
<tbody>
<tr>
<td>T00902 XBP-1 rs12664788</td>
<td>immune response</td>
<td>1.25</td>
<td>13.159</td>
<td>4</td>
<td>4</td>
<td>0.000000679</td>
<td>2</td>
<td>1.25</td>
</tr>
<tr>
<td>T14942 hsa-miR-181b rs10976413</td>
<td>synaptic transmission</td>
<td>2.05</td>
<td>3.592</td>
<td>21</td>
<td>31</td>
<td>7.63E-08</td>
<td>6</td>
<td>2.05</td>
</tr>
<tr>
<td>T15286 E2F-1:DP-1 rs1382606</td>
<td>DNA repair</td>
<td>2.64</td>
<td>2.210</td>
<td>29</td>
<td>41</td>
<td>0.000000159</td>
<td>7</td>
<td>2.64</td>
</tr>
<tr>
<td>T05444 RFX5:RFXAP:RFXANK rs6940715</td>
<td>DNA repair</td>
<td>2.31</td>
<td>2.111</td>
<td>32</td>
<td>41</td>
<td>11.54</td>
<td>10</td>
<td>2.31</td>
</tr>
<tr>
<td>T00902 XBP-1 rs6111734</td>
<td>DNA repair</td>
<td>2.97</td>
<td>0.788</td>
<td>54</td>
<td>113</td>
<td>0.000000199</td>
<td>15</td>
<td>2.97</td>
</tr>
</tbody>
</table>

McCauley et al. [40] reported that SNPs in multi-species conserved sequences (MCS) are useful as markers linking to complex diseases. Recent evidence suggests that SNPs that influence alternative splicing are enriched within splice junctions (SJ) or disrupt splicing enhancers [21,41]. Our analysis of Foxp3 TReQTLs revealed SNPs overrepresented within 5-way (human, mouse, rhesus monkey, chimp) evolutionary conserved regions (ECRs) in CEU and in SJs of YRI defined by RNA-Seq mapping (Table 4). These results support the notion that genomics, genetics and transcriptomics play an intricate role in sustaining population diversity and structure [16]. It would be interesting to determine how environmental factors, population structure and geographical differences affect transcript abundance as a quantitative trait when co-regulation of gene expression is considered.

Although the identification of TReQTLs is useful for determining genetic variants regulating gene expression, there are limitations to the approach and guidelines with interpretation of the results. First, there is a paucity of information about the genes which TRs control. We restricted our analysis to only 333 TRs with two or more DSTs known at a given time to be regulated by TRs. This does not capture the full array of genetic variants which might contribute to the gene expression differences between the two populations. However, as advances in functional genomics leads to improved knowledge about gene regulation and biological function on a genome-wide scale, the discovery of TReQTLs should advance and be more informative. In addition, the study of the transcript-regulation of genes by miRNA is in its infancy and there is a small number of miRNAs known to regulate genes. Furthermore, our analysis only tested the association of a single SNP with sets of coordinately expressed genes. It is very likely that the variation in expression is due to the synergistic effect of two or more SNPs. In fact, there may be other mediators of complex diseases other than SNPs acting alone or symbiotically. Finally, our work relied on samples from immortalized lymphoblastoid cell lines (LCLs) and not from a disease state. Therefore, it is debatable whether or not the genetic associations of SNPs with gene expression in LCLs will carry over to tissue samples from organs [55]. However, there is some indication, albeit a paucity of evidence, that the DNA repair capacity of LCLs from breast cancer samples is significantly lower than control subjects [56], that tumor-infiltrating Foxp3+ regulatory T cells can distinguish between high-risk breast cancer patients and those at risk of a late relapse [57] and that a fraction of eQTLs derived from the analysis of UK Adult Twin registry LCLs gene expression and genotype data overlap with those identified in a HapMap population [47]. Despite the caveats noted above, the advantages of associating genetic markers such as SNPs to quantitative traits such as co-regulated genes is promising and of value as an additional strategy when investigating the role of a genetic variant and master regulators in the etiology of a complex diseases.

## Materials and Methods

### Genotype Data

Genotype data (phase-II, release 24, forward strand, non-redundant) from the 60 Yoruba in Ibadan, Nigeria (African: YRI)
and from the 60 CEPH-Utah residents with ancestry from northern and western Europe (European: CEU) populations were obtained from the International HapMap Project [30]. SNPs with a call rate <95%, minor allele frequency (MAF) < 0.05, or Hardy-Weinberg equilibrium [58] p-value < 0.05 within each population separately were removed and we restricted our analysis to autosomal markers only. About 2 million SNPs in CEU and ~2.2 million SNPs in YRI were retained after filtering. The approximately 1.5 million SNPs in common between the two populations after filtering (common set) were used for TReQTL preliminary analysis using a nominal p-value threshold of \(1 \times 10^{-6}\). To account for correlated SNPs, we used the LRTag approach [59,60] with linkage disequilibrium (LD) correlation \(r^2 > 0.5\) and MAF \(> 0.05\) to tag 416,160 independent SNPs (tag set) of the 1.5 million in the common set. In addition, for multiple testing correction, we used the p-values of the tag SNPs that are within 1 kb of a gene (cis-acting set) to obtain an FDR. For the SNP association portion of the study, we focused the sample set on the 60 CEU and 55 YRI individuals that had corresponding gene expression data.

Figure 5. TReQTL Gene Ontology (GO) biological process subtrees. A) Based on the GO biological processes from the gene that the YRI TReQTL SNP rs12258754 map to and those of the DSTs of activating transcription factor 3 (Atf3) and of Atf3 itself. B) Based on the GO biological processes from the gene that the CEU TReQTL SNP rs10976413 map to and those of the DSTs of miRNAs hsa-mir-181b-1 (MI0000270) and hsa-mir-181b-2 (MI0000683).

doi:10.1371/journal.pone.0034286.g005
Microarray Gene Expression Data

Gene expression data from the profiling of Epstein-Barr virus (EBV)-transformed lymphoblastoid cell lines from the individuals in each CEU and YRI HapMap population were acquired by assaying total RNA on Affymetrix Human HG-Focus Target GeneChip Array [33] containing about 9,000 probe sets and representing about 8,600 UniGenes (NetAffx array annotation release 29, March 30, 2009 UniGene build 219). The data are available in the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) database [61,62] under accession number GSE10824. The data were log2 transformed and normalized using the robust multi-array average method [63]. The gene expression data from probe sets that mapped to the same UniGene cluster were averaged resulting in 8,399 unique UniGene transcripts represented by probe sets on the array.

RNA-Seq Data

We obtained raw sequence reads from the whole transcriptome sequencing of the CEU and YRI populations [64,65]. These reads were mapped to the reference human genome (hg18, NCBI 36) using the Bowtie program [66]. Splice junctions (SJs) were mapped using the Tophat program [67]. Transcript abundance was calculated by the Cufflinks program [68] and normalized using the FPKM (Fragment Per Kilobase of Exon per Million reads) method [69]. At a nominal p-value of 1x10^-4 there are a total of 7,149 and 7,040 TReQTL SNPs in the CEU and YRI populations respectively. Each SNP was evaluated for its presence in SJs in each individual from both populations. The total number of SNPs in each population was calculated based on their presence or absence in any individual of the population.

Signal Transduction Pathway Profiling

The Affymetrix GeneChip array probe sets were collapsed into approximately 8,600 UniGene transcript clusters using the March 30, 2009 release of UniGene (build 219). The gene expression data from probe sets that mapped to the same UniGene cluster were averaged resulting in 8,399 UniGene transcripts. The UniGene downstream targets (DSTs) of transcript-regulators (TRs: transcription factors (TFs), miRNAs, cofactors, and complexes) were obtained from the March 26, 2010 release (version 2010.1) of the TRANSFAC® database [31,32]. TRs were mapped to signaling pathways using the June 26, 2009 release (version 2009.2) of the TRANSPATH® database [34]. Significance of signal transduction pathway profiling was determined as previously described [70]. Briefly, for each population individually, significant TRs were based on a Group Correlation Score

\[
GCS = \sum_{i\neq j} r_{ij}^2
\]

defined as the sum of the squares of the Pearson correlations \(r\) among all pairs of genes \(i\) and \(j\) determined to be DSTs of the TR. Significant pathways made up of TRs are based on an Exclusive Group Correlation Score

\[
EGCS = \sum_{TR(i),TR(j)=0} r_{ij}^2
\]

defined as the sum of the squares of \(r\) over all pairs of genes \(i\) and \(j\) in a pathway that do not share any TR. This eliminates the contribution of the co-expression of DSTs that share TRs. The \(p-\)
value for a score was determined from a non-parametric distribution of correlation scores obtained from random cases ($B = 10,000$ reshuffles of the genes) and the number of times ($n$) one of these permuted scores is greater than the observed correlation score. Thus, $p-value = n/B$. For both correlation scores, GCS and EGCS, this null hypothesis keeps the structure and overlap of all pathways fixed, but changes the identity of the genes.

**TReQTL Analysis**

For each population and TR, we performed a genome-wide scan by regressing the log$_2$ expression levels of the DSTs on each SNP genotype ($Z$) (coded 0, 1, and 2 representing the number of minor alleles) separately across the genome. The following multivariate linear regression (MVR) model was used

$$Y_{ij} = \beta_0 + Z_i \beta_j + e_i$$

where $Y_{ij}$ denotes the log$_2$ expression levels of the DST $j$ ($j = 1, \ldots, m$) for a TR for subject $i$ ($i = 1, \ldots, n$), $m$ is the number of DSTs of the TR, $Z_i$ is a SNP genotype, $e_i$ is an error and $e_i = (e_{i1}, \ldots, e_{im})$ follows a multivariate normal distribution with mean 0 and covariance $\Sigma$. To test for the null hypothesis of association between a SNP and a TR, we performed the likelihood ratio test for testing the null hypothesis $H_0: \beta_{1j} = \ldots = \beta_{mj} = 0$, which follows a chi-square distribution with $m$ degrees of freedom for $m < n$. Let the chi-square test statistic $D = -2(\ln(\text{likelihood null model}) - \ln(\text{likelihood full model}))$ where the null model is the MVR model without the genotypes corresponding to the SNP and the full model is the MVR model with the genotypes corresponding to the

<table>
<thead>
<tr>
<th>Genomic Region</th>
<th>CEU</th>
<th>YRI</th>
</tr>
</thead>
<tbody>
<tr>
<td>5-way Conservation</td>
<td>0.006</td>
<td>0.900</td>
</tr>
<tr>
<td>17-way Conservation</td>
<td>0.071</td>
<td>0.870</td>
</tr>
<tr>
<td>Splice Junctions</td>
<td>0.893</td>
<td>1.0E-04</td>
</tr>
</tbody>
</table>

10 K permutations of 472 SNPs with a $p$-value $< 1 \times 10^{-6}$ in either CEU or YRI.

doi:10.1371/journal.pone.0034286.t004
SNP. The $p$-value for each association of a SNP and set of DSTs for a TR was obtained from the distribution of $D = \chi^2$ with degrees of freedom $= m$. For cases where $m > n$, an approximate F-statistic [9,35,36] was used in order to avoid situations where the covariance matrix from the MVR model is not full rank. In these cases, the $p$-value for statistical significance of each association of a SNP and a set of DSTs for a TR was assessed by permuting the $n$ rows and $n$ columns of the F-statistic G (Gower’s centered) matrix ($1 \times 10^6$ times) and determining the number of times one of these bootstrapped scores is greater than the observed score. We fit model (5) by regressing the DSTs of a TR on each SNP separately across the genome. For multiple testing correction, we used the $6.1 \times 10^6$ $p$-values from the regression of the DSTs of the 333 TRs on the 184,616 independent (tag) SNPs that are within 1 kb of a gene (cis-acting set) to control the FDR [71] at 45%.

SNP Set Enrichment Analysis

To determine whether a set of SNPs representing TReQTLs are enriched within the genome, a variation of gene set enrichment analysis [72] was used. Rather than enrich for SNPs within pathways as previously described [42], we test for enrichment of SNPs within particular genomic regions. The regions of interest are either evolutionary conserved regions (ECRs) or splice junctions (SJs). The rationale is that variants which are related to gene expression differences between populations are more likely to be in genomic regions conserved across species and/or possibly involved in regulating transcription by altering splice forms of transcript messages. The ECRs are from 5-way (human, mouse, chimp, rhesus monkey and dog) and 17-way (human, chimp, macaque, mouse, rat, rabbit, dog, cow, armadillo, elephant, tenrec, opossum, chicken, frog, zebrafish, Tetraodon, and Fugu) conservation scores for the +/− 20 kbp flanking regions of the genes. Overlapping chromosomal location intervals for phastCon scores [73] were segmented and the highest conservation score within the interval was obtained. SJs were mapped using TopHat [67]. Transcript abundance was calculated by Cufflinks [68] and normalized using FPKM [69] (see the RNA-Seq methods section). For each TR, given a combined set $L$ of SNPs associated with significant TReQTLs within the CEU or YRI population, their corresponding $p$-values and the +1 or −1 indication (flag) of the assignment of the SNP either within or not within genomic region set $S$, an enrichment score (ES) is obtained by the ranking of the SNPs in ascending order (most significant to least significant), and then summing up the assignment flags. The ES is calculated by screening this list from the top to the bottom and increasing (decreasing) a running sum Kholmogorov–Smirnov-like statistic [72] when encountering or not encountering a SNP in a genomic region. A normalized ES (NESS) is obtained by accounting for the size of the genomic region set $S$. The $p$-value for enrichment was determined from a non-parametric distribution of NESSs obtained from random cases (10,000 reshuffles of the assignment of the SNP (either within or not within a genomic region)) and the number of times one of these permuted scores is greater than the observed NESS. Significance of enrichment was also confirmed using a two-tailed Fisher’s exact (parametric) test.

Ranking of the TReQTL SNPs

The ranking of TReQTL SNPs was performed by first measuring the cohesion of GO biological process terms in TReQTL $i$.

\[
x_i = \frac{\mu_i}{N_i(N_i-1)}
\]

where $N_i$ is the number of nodes (biological process terms) represented in TReQTL $i$, and $\mu_i$ is the number of the edges between nodes. The edges were derived from the structure of GO subtree for each TReQTL $i$, created from the biological process terms of the gene that the SNP maps to or is in close proximity, those of the TRs (excluding miRNAs) and the DST genes. The cohesion measure $x_i$ is then weighted by the $p$-value of TReQTL $i$ to give a weighted rank. The weight is computed as $-2log_{10}(p$-value). Finally, the number of paths and the number of biological process terms within the subtree were used to derive of the adjusted cohesion score

\[
ACS_i = \frac{x_i(-2log_{10}(p$-value$_i))(N_i/M_i)}{(P_i/N_i)}
\]

where for the $i$th TReQTL $i$, $M_i$ is the maximum number of paths of a biological process term and $P_i$ is the number of paths.

Gene Interaction Network

Ingenuity Pathway Analysis (IPA) software version 8.8 and canonical pathway content version 3204 were used to build gene interactions from Foxp3, its DSTs (CSF2 and IL2) and the significant TReQTL genes from CEU and mapped on chromosome 1 (KCNJ1, LPHN2 and LPHN3).

Supporting Information

Table S1 The TReQTLs for the CEU and YRI populations. Tab-delimited text file. (TXT)

Table S2 Gene Ontology biological process subtrees from the CEU TReQTLs. Tab-delimited text file. (TXT)

Table S3 Gene Ontology biological process subtrees from the YRI TReQTLs. Tab-delimited text file. (TXT)

Acknowledgments

We would like to thank Yuehua Cui and Shaoyu Li for pseudo F-statistics details, Dmitri Zaykin for discussions about combining SNPs with gene expression for genetical genomics and Frank Dai at the NIEHS Computational Biology Facility for multiprocessor computing resources. We are extremely grateful to Lan Liu and Tao Jiang for providing us with the LRTag genome-wide and across populations SNP tagging application. We also thank Steve Kleeberger, Min Shi and Greg Gibson for their critical review of the manuscript.

Author Contributions

Contributed intellectually to the interpretation of results: PRB OH WH XL. Conceived and designed the experiments: PRB XL. Analyzed the data: PRB RM LL AH JL. Wrote the paper: PRB.

References


