Published in final edited form as:
Hum Genet. 2009 February ; 125(1): 81–93. doi:10.1007/s00439-008-0601-x.

Identification of common genetic variants that account for transcript isoform variation between human populations

Wei Zhang, **Shiwei Duan**, **Wasim K. Bleibel**, **Steven A. Wisel**, **R. Stephanie Huang**, **Xiaolin Wu**, and **Lijun He**
Section of Hematology/Oncology, Department of Medicine, The University of Chicago, Box MC6091, 5841 S. Maryland Ave., Chicago, IL 60637, USA

Tyson A. Clark, **Tina X. Chen**, **Anthony C. Schweitzer**, and **John E. Blume**
Expression Research Laboratory, Affymetrix Inc., Santa Clara, CA 95051, USA

M. Eileen Dolan
Section of Hematology/Oncology, Department of Medicine, The University of Chicago, Box MC6091, 5841 S. Maryland Ave., Chicago, IL 60637, USA

Nancy J. Cox
Department of Human Genetics, The University of Chicago, Chicago, IL 60637, USA

Abstract

In addition to the differences between populations in transcriptional and translational regulation of genes, alternative pre-mRNA splicing (AS) is also likely to play an important role in regulating gene expression and generating variation in mRNA and protein isoforms. Recently, the genetic contribution to transcript isoform variation has been reported in individuals of recent European descent. We report here results of an investigation of the differences in AS patterns between human populations. AS patterns in 176 HapMap lymphoblastoid cell lines derived from individuals of European and African ancestry were evaluated using the Affymetrix GeneChip® Human Exon 1.0 ST Array. A variety of biological processes such as response to stimulus and transcription were found to be enriched among the differentially spliced genes. The differentially spliced genes also include some involved in human diseases that have different prevalence or susceptibility between populations. The genetic contribution to the population differences in transcript isoform variation was then evaluated by a genome-wide association using the HapMap genotypic data on single nucleotide polymorphisms (SNPs). The results suggest that local and distant genetic variants account for a substantial fraction of the observed transcript isoform variation between human populations. Our findings provide new insights into the complexity of the human genome as well as the health disparities between the two populations.

Introduction

The existence of health disparities between human populations, for example, the differential response to therapeutic treatments (Huang et al. 2007) and higher risks of certain common diseases has been reported by clinical scientists. However, the genetic basis for population differences in clinical outcomes and risk of common disease is not fully understood (Huang et

e-mail: ncox@medicine.bsd.uchicago.edu.
W. Zhang and S. Duan contributed equally to this work.
e-mail: wzhang1@uchicago.edu

Electronic supplementary material The online version of this article (doi:10.1007/s00439-008-0601-x) contains supplementary material, which is available to authorized users.

al. 2007; Ioannidis et al. 2004; Kurian and Cardarelli 2007). In the past few years, gene expression has been studied as a quantitative complex phenotype (Morley et al. 2004; Stranger et al. 2005), which sits between genetic/non-genetic variations and other more complicated cellular or whole-body phenotypes. Therefore, studying variation in gene expression between populations may help explain these health disparities. In addition to several differences between populations in transcriptional and translational regulation of genes, alternative premRNA splicing (AS) is also likely to play an important role in regulating gene expression and generating variation in mRNA and protein isoforms. The initial sequencing and analysis of the human genome suggested an unexpectedly low gene number of 30,000-35,000 (Lander et al. 2001; Venter et al. 2001), which raises the question of the source of the complexity of the human genome. Numerous studies such as those using expressed sequence tags (ESTs) and cDNAs aligned to the genomic sequences have shown that AS is prevalent in mammalian genomes (Sorek et al. 2004). It has been estimated that between one-third and two-thirds of all human genes undergo alternative splicing (Sorek et al. 2004) and the disruption of specific AS events has been implicated in several human genetic diseases including cancer (Brinkman 2004; Faustino and Cooper 2003; Novoyatleva et al. 2006).

Studies using the International HapMap Project (http://www.hapmap.org) resources (Frazer et al. 2007; International HapMap Consortium 2003, 2005; Zhang et al. 2008b) have shown that common genetic variants in the form of single nucleotide polymorphisms (SNPs) contribute to gene expression variation within the same HapMap population (Duan et al. 2008a) as well as between the different HapMap populations (Spielman et al. 2007; Storey et al. 2007; Stranger et al. 2007; Zhang et al. 2008a). The Phase I/II HapMap samples are comprised of a panel of lymphoblastoid cell lines (LCLs) derived from individuals of northern and western European ancestry collected by Centre d'Etude du Polymorphisme Humain (CEPH) (CEU: CEPH individuals from Utah, USA, 30 parents-child trios), individuals of African ancestry (YRI: Yoruba people from Ibadan, Nigeria, 30 parents-child trios) and individuals of eastern Asian ancestry (CHB: Han Chinese from Beijing, China, 45 unrelated samples; JPT: Japanese from Tokyo, Japan, 45 unrelated samples). Recently, studies have begun to demonstrate the genetic contribution to the transcript isoform variation in the unrelated CEU samples (Hull et al. 2007; Kwan et al. 2007, 2008). However, the systematic comparison of the transcript isoform variation including AS events between human populations and their regulation by common genetic variants have not been comprehensively investigated. We therefore utilized the Affymetrix GeneChip® Human Exon 1.0 ST Array (exon array), which contains probes for ~20,000 well-annotated human genes (~1.4 million annotated and predicted exons corresponding to 17,745 transcript clusters using the core set of exon-level probesets supported by RefSeq (Pruitt et al. 2007)), to study 176 HapMap samples (87 CEU and 89 YRI) from parents-offspring trios.

One potential problem with the use of oligonucleotide expression arrays is the possibility that SNPs located within probes could affect hybridization efficiency (Gilad et al. 2005) and lead to false expression quantitative loci (eQTLs) (Alberts et al. 2007). This effect was also observed in our exon array expression data. We described this effect in a previous publication using *HLA-DPB1* as an example (Zhang et al. 2008a). To reduce the potential variability associated with this effect, we filtered out probesets (exon-level) containing all known SNPs in the current dbSNP database (version 129) (Duan et al. 2008b) maintained by the National Center for Biotechnology Information (NCBI) before summarizing transcript cluster (gene-level) expression signals. In addition, a recent publication suggested that the effect of unannotated or undiscovered SNPs is quite small for the exon array using the unrelated CEU samples (Kwan et al. 2008). Our goals were then to identify probesets that showed transcript isoform variation between these two populations, to determine what biological processes or pathways were enriched in the genes containing differentially spliced probesets and to evaluate the contribution of local and distant genetic variants (SNPs) to the observed population differences

in transcript isoform variation (see Supplemental Fig. 1 for the workflow). Specifically, we focused on the differences between the CEU and YRI samples in simple cassette exon skipping events. Splicing index (SI), defined as the relative contribution of a probeset (exon-level) to transcript cluster (gene-level) expression (Affymetrix Inc. 2006; Gardina et al. 2006) was used to evaluate any transcript isoform variation between the two populations.

Materials and methods

Cell lines, RNA isolation and chip hybridization

Details for this part including our approach to avoid systematic bias were described in a previous publication (Zhang et al. 2008a). Briefly, HapMap cell lines (International HapMap Consortium 2003, 2005) (30 CEU trios and 30 YRI trios) were purchased from Coriell Institute for Medical Research (Camden, NJ). Two CEU samples (GM10855 and GM12236) were not available from Coriell at the time of the study. The viability of two lines (GM12716, GM18871) was below 85% at the sample collection time. Therefore, a total of 176 cell lines (87 CEU samples and 89 YRI samples) were included in this study. Total RNA was extracted using Qiagen Qiashredder and RNeasy plus kits (Qiagen, Germantown, MD) according to manufacturer's protocol. All 176 RNA samples had high quality and showed no signs of DNA contamination or RNA degradation. RNA samples were immediately frozen and stored at -80° C. For each cell line, ribosomal RNA was depleted and cDNA was generated, which was fragmented and end labeled. Approximately 5.5 μg of labeled DNA target was hybridized to the Affymetrix GeneChip® Human Exon 1.0 ST Array at 45°C for 16 h per manufacturer's recommendation (http://www.affymetrix.com/products/arrays/exon_application.affx). Hybridized arrays were then washed and scanned on a GCS3000 Scanner (Affymetrix, Santa Clara, CA).

Data Filtering for SNPs in probes, signal normalization and summarization

Expression arrays were analyzed using the Affymetrix PowerTools v1.8.6 (http://www.affymetrix.com/support/developer/powertools/index.affx). The start and end coordinates of all probes represented on the exon array were queried and determined against the human genome (hg18). The coordinates for all SNPs were then queried in the dbSNP database (version 129) (http://www.ncbi.nlm.nih.gov/projects/SNP) and used to identify probes harboring known SNPs. Of the ~1.4 million probesets on the exon array, 350,382 probesets contained at least one probe with a SNP (~600,000 probes). The probeset signal intensity files were filtered by removing those ~600,000 probes from the probesets harboring these known SNPs (Duan et al. 2008b). Probe intensities were then background corrected and quantile normalized over all 176 samples. The data were then $\log_2$ transformed with a median polish. Gene-level expression of 17,745 transcript clusters was summarized using the RMA (robust multi-array average) (Irizarry et al. 2003) method with signals generated on a core set [i.e., with RefSeq-supported (Pruitt et al. 2007) annotation] of exons (~110,000 probesets). A transcript cluster or probeset was defined to be reliably expressed in LCLs if the $\log_2$ transformed expression signal was greater than 6 in at least 80% of the 176 samples. A total of 8,565 of the 17,745 core transcript clusters met these criteria. To avoid annotation ambiguity, the final analysis dataset is comprised of 7,701 expressed transcript clusters (corresponding to 102,729 probesets, a minimum of 3 probesets for each transcript cluster) with unique gene annotations (based on NCBI Human Genome Build 34) as retrieved from the Affymetrix NetAffx Analysis Center (http://www.affymetrix.com/analysis/index.affx).

Detecting differentially spliced probesets between populations

Candidate probesets differentially spliced between the CEU and YRI samples were detected by calculating the splicing index (SI) (Affymetrix Inc. 2006; Gardina et al. 2006). The SI represents the log-transformed normalized exon-level probeset intensities by the gene-level

transcript cluster intensities in each sample. $\mathrm{SI}_{i,j}=\log\left(\frac{e_{i,j}}{g_i}\right)$, where $e_{i,j}$ is the intensity of the *j*th probeset of the *i*th transcript cluster, g_i is the intensity of the *i*th transcript cluster and $\mathrm{SI}_{i,j}$ is the splicing index of the *j*th probeset of the *i*th transcript cluster. The permutation-based free step-down approach of Westfall-Young (W-Y approach) (Westfall and Young 1993) was used to detect probesets with differential SI values between the CEU and YRI samples. The basic test was the standard pooled variance *t* statistic. Because of the relatedness among family members, trios were permuted between the two populations. The W-Y approach (n = 10,000 permutations) was then used to compute simultaneous *P* values that control the overall or family-wise error rate. In addition, the W-Y approach (n = 10,000 permutations) was applied on the unrelated CEU or YRI samples to detect potential differential probesets between males and females. The probesets with a significant permutation-adjusted *P* value ($P_c < 0.01$) were chosen for further analyses. The permutation-adjusted one-sided *P* values were calculated using the software Permax 2.2, http://biowww.dfci.harvard.edu/~gray/permax.html, which has an implementation of the W-Y approach and is provided as a contributory library by Robert Gray in the R statistical package (R Development Core Team 2005). The annotations for the differentially spliced probesets including gene symbol, cytoband and whether the probeset overlaps coding regions were retrieved from the Affymetrix NetAffx Analysis Center.

Biological process and pathway analyses

We used the DAVID (Database for Annotation, Visualization and Integrated Discovery) (Dennis et al. 2003; Huang da et al. 2007) (http://david.abcc.ncifcrf.gov) to identify enriched Gene Ontology (GO) (Ashburner et al. 2000) (http://www.geneontology.org) or PANTHER (Thomas et al. 2003) (http://www.pantherdb.org/pathway) biological processes as well as known pathways such as those in the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al. 2004) (http://www.genome.jp/kegg), Biocarta (http://www.biocarta.com) and PANTHER (Thomas et al. 2003) among the genes that showed differential transcript isoform variation between the CEU and YRI samples. The analysis set of 7,701 uniquely-annotated transcript clusters were used as the background list. Biological processes that were overrepresented relative to the background were selected (5 hits or more, Fisher's exact test P_c<0.50 after Benjamini-Hochberg, BH correction) (Benjamini and Hochberg 1995; Huang da et al. 2007). The same criteria were applied to identify enriched pathways. In addition, DAVID was also used to check if there were any genes with known AS events among our identified genes with differentially spliced probesets between the two populations. We further examined if these identified genes were involved in any Mendelian diseases as annotated in the Online Mendelian Inheritance in Man (OMIM) database (McKusick 1998) (http://www.ncbi.nlm. nih.gov/Omim).

Genotypic data for the HapMap samples

SNP genotypes were downloaded from the International HapMap Project website (http://www.hapmap.org) (Thorisson et al. 2005) (release 22 March 2008). To reduce the effect of possible genotyping errors, we excluded the SNPs with Mendelian allele transmission errors on 22 autosomes in the CEU and YRI samples, respectively. Thus, our final genotypic dataset was comprised of about 1.57 million SNPs for the two populations.

F_{st} values

F_{st}, a metric representation of the effect of population subdivision, was estimated according to Wright's approximate formula $F_{st} = (H_T - H_S)/H_T$, where H_T represents expected heterozygosity per locus of the total population and H_S represents expected heterozygosity of a subpopulation (Wright 1950). An F_{st} value was calculated for each SNP of interest using allele frequencies estimated from the unrelated individuals for each population.

Cluster analysis

For the differentially spliced probesets, the Pearson correlation coefficients of the SI values were computed for the 176 samples to represent pairwise similarity. The probesets were then grouped by a hierarchical clustering algorithm (Eisen et al. 1998) using the average linkage method, which was implemented in the MeV:MultiExperiment Viewer (Saeed et al. 2003) (http://www.tm4.org).

Identifying common genetic variants correlated with AS patterns

The SI values of the differential probesets were evaluated for association with SNP genotype using the QTDT software (Abecasis et al. 2000a, b). The association study was carried out in the combined CEU and YRI data with gender and population as covariates (QTDT $P<3.18 \times 10^{-8}$, $P_c < 0.05$ after Bonferroni correction by ~1.57 million common SNPs in both the HapMap CEU and YRI populations). The incomplete trios were also used in the QTDT analysis. We defined a probeset as locally-regulated if the SI was associated with a SNP(s) within 2.5 Mb on the same chromosome, while a probeset was distantly-regulated if the SI was associated with SNP(s) on different chromosome(s) or more than 2.5 Mb away on the same chromosome.

Validation of transcript isoform variation between populations

Total RNA from 53 unrelated CEU and 48 unrelated YRI cell lines was extracted using the RNeasy Mini kit (Qiagen Inc., Valencia, CA) following the manufacturer's protocol. RNA quality assessment and quantification were conducted using the optical spectrometry 260/280 nm ratio. Subsequently, mRNA was reversely transcribed to cDNA using Applied Biosystems High Capacity Reverse Transcription kit (Applied Biosystems, Foster City, CA). Reverse transcription reactions were prepared to yield a final cDNA concentration of 50 ng/uL. Primers used for quantitative RT-PCR (Supplemental Table 3) were designed using Primer3 software (Rozen and Skaletsky 2000). Expression measurements were performed on the Applied Biosystems 7500 Real-Time PCR system. Total reaction was carried out in 25-μL volume which consisted of 12.5 μL ABI SYBR Universal mix (Applied Biosystems, Foster, CA), 1.5-μL primers along with 10-μL diluted cDNA. The thermocyler parameters were: 50°C for 2 min, 95°C for 10 min, and 40 cycles of 95°C for 15 s/60°C for 1 min. Each cycle threshold (C_t) value obtained for each probeset of interest was quantified into relative expression levels using the relative standard curve method (Applied Biosystems 2004). Each standard curve was created using a mixture of cDNA of known concentration from all samples being tested. Each experiment was conducted in duplicate for samples from both populations. A ratio comparing the relative quantity of the probeset of interest relative to the quantity of the constitutive exon was compared with the splicing index values from the expression arrays to determine replication of findings.

Results

Detecting differentially spliced probesets between populations

We compared the SI values of 102,729 probesets (belonging to 7,701 uniquely-annotated transcript clusters with reliable expression in LCLs). Using the W-Y approach (Westfall and Young 1993) that adjusts for the trio structure in these samples, 782 probesets within 570 transcript clusters had significantly different SI values ($P_c < 0.01$, permutation-adjusted), indicating variations in AS events between the CEU and YRI samples. Among the 782 probsets, we found that 397 probesets had significantly lower SI values in the CEU samples, while 385 probesets had significantly lower SI values in the YRI samples. Figure 1 shows the genomic distribution of these differentially spliced probesets. No chromosomes were overrepresented or underrepresented in terms of the number of differentially spliced probesets ($P_c < 0.05$ after BH correction). In addition, 514 out of the 782 differential probesets were in coding regions

and the remaining 268 probesets were in untranslated regions (UTRs). The details of these 782 probesets are presented in Supplemental Table 1. The 782 probesets could be grouped into two distinguishable clusters representing the populations based on the splicing index values (Fig. 2).

Biological process and pathway analyses

Three GO biological processes ("response to stimulus", "regulation of cellular process" and "transcription") and four PANTHER biological processes ("nucleoside, nucleotide and nucleic acid metabolism", "asymmetric protein localization", "cell proliferation and differentiation" and "cell structure and motility") were found to be enriched in the 570 uniquely-annotated transcript clusters ($P_c < 0.50$ after BH correction) (Table 1). At the same significance level, one KEGG pathway ("antigen processing and presentation") was found to be enriched among these transcript clusters ($P_c < 0.50$ after BH correction) (Table 1). Among the 570 differentially spliced genes, 80 are linked to certain diseases as maintained in the OMIM database (Supplemental Table 1), though no individual disease was enriched ($P_c < 0.50$ after BH correction). These diseases include, for example, type I diabetes and certain types of cancer. In contrast, the term "immune (disease)" (41 genes, $P = 0.0075$, $P_c = 0.13$ after BH correction) was enriched among these genes by searching the "GENETIC_ASSOCIATION_DB_DISEASE_ CLASS" database, which compiles ~9,000 associations from the literatures by DAVID (Dennis et al. 2003;Huang da et al. 2007). Notably, among the 570 differentially spliced genes we identified, 171 genes are known to have alternative products (Supplemental Table 1) by searching the Protein Information Resource (PIR) (McGarvey et al. 2000) through DAVID (Dennis et al. 2003;Huang da et al. 2007). The category of "alternative products" was enriched relative to the analysis set of 7,701 genes ($P = 0.027$, $P_c = 0.093$ after BH correction).

Identifying common genetic variants that associate with differentially spliced probesets

Association with ~2 million common HapMap (International HapMap Consortium 2003, 2005) SNPs (minor allele frequency ≥5% in the unrelated parents of each population) using the QTDT software (Abecasis et al. 2000a, b) was evaluated in both the CEU and YRI samples with population and gender as covariates. We identified 2,393 local SNPs that were correlated with the SI values of 97 differentially spliced probesets in 85 transcript clusters. In addition, 419 distant SNPs were found to be correlated with the SI values of 152 probesets in 124 transcript clusters. Details for these associated SNPs are listed in Supplemental Table 2. Among them, both local and distant SNPs were identified for 36 differentially spliced probesets in 34 transcript clusters. Table 2, Fig. 3 and 4 show some representative local SNP/SI relationships with relatively higher F_{st} values ($F_{st} > 0.15$). Supplemental Table 2 lists the details for all significant SNP/SI relationships ($P_c < 0.05$ after Bonferroni correction).

Validation of transcript isoform variation between populations

From the probesets that were differentially spliced (Supplemental Table 1), we randomly chose 3 internal exons: PS3764493 (*MTMR4*), PS3303658 (*MRPL43*) and PS3476020 (*MPHOSPH9*) to experimentally validate. In addition, we included PS3527423 (*PARP2*) as the positive control, which was previously shown to be differentially spliced in the unrelated CEU samples (Kwan et al. 2008). Using the unrelated CEU cell lines, we confirmed the within-population variation of probeset PS3527423 (*PARP2*) demonstrated by Kwan et al. (2008) (Supplemental Fig. 2). Quantitative Real-Time PCR showed a difference in the ratio of isoforms between the two populations for probesets PS3764493 (*MTMR4*) and PS3303658 (*MRPL43*) (Supplemental Fig. 3). The quantitative Real-Time PCR results for PS3764493 (*MTMR4*) and PS3303658 (*MRPL43*) were consistent with the trend of SI values calculated from the exon array data (Supplemental Table 1).

Discussion

The Affymetrix GeneChip® Human Exon 1.0 ST Array was utilized to measure probeset (exon-level) expression in EBV (Epstein-Barr Virus)-transformed LCLs derived from 176 apparently healthy individuals (CEU, 87 cell lines; YRI, 89 cell lines) (Zhang et al. 2008a). Transcript cluster (gene-level) expressions were computed by summarizing signals from RefSeq-supported (Pruitt et al. 2007) exons (core set) within each transcript cluster. Our first goal was to identify probesets in transcript clusters with evidence for between-population transcript isoform variation. We compared the splicing index values (Affymetrix Inc. 2006; Gardina et al. 2006) of 7,701 uniquely-annotated transcript clusters (containing 102,729 probesets). Because non-expressed exons are known to introduce false positive results in the SI calculation, particularly in the presence of gene expression level changes (Affymetrix Inc. 2006), we limited our analyses to transcript clusters and probesets with reliable expression in the two populations as a whole. To identify meaningful AS events, we only focused on transcript clusters with a minimum of three expressed probesets. A significantly lower SI value in a population indicates that the particular probeset (exon-level) may be skipped in an AS isoform or that the respective transcription isoform has a lower relative ratio among all isoforms. The proportion of expressed genes (~50%) we defined is comparable to previous observations in LCLs (Cheung et al. 2003; Spielman et al. 2007), though a precise profiling of expressed genes in these samples has not been investigated experimentally.

Using the permutation-based W-Y approach (Westfall and Young 1993), we identified 782 probesets within 570 transcript clusters that showed differential SI values between the two populations (Fig. 1). The advantages of the W-Y approach include that (1) it considers dependence between genes when testing expression; (2) it allows the cluster-level permutation, thus taking into account the parents-child trio structure of the CEU and YRI samples. Although differential gene expression between males and females has been detected in a panel of CEPH LCLs (Zhang et al. 2007), no probesets (at $P_c < 0.05$, permutation-adjusted) were found to show gender-specific differences in either CEU or YRI samples, suggesting transcript isoform variation may not commonly contribute to gender-specific gene expression. Using RT-PCR, two of the three randomly-chosen exons (67%) from the 782 probesets could be validated for population differences in abundance of respective transcript isoforms (Supplemental Fig. 3), though a more comprehensive validation would be necessary to provide a more accurate estimation of the current findings. In addition, among the 570 transcript clusters containing differentially spliced probesets, approximately a third (171 genes) are known to have AS events or alternative products (literature-based evidence) as maintained in the PIR database (McGarvey et al. 2000) (Supplemental Table 1). Our list of differentially spliced genes between the two populations was found to overrep-resent the category of “alternative products” relative to the analysis set of 7,701 genes ($P = 0.027$, $P_c = 0.093$ after BH correction), indicating that many of the identified genes have known alternatively spliced transcript isoforms. Another interesting question would be whether the population differences in transcript isoform variation are mainly regulatory in nature at the level of RNA expression or due to changes at the protein level. We classified the 782 differentially spliced probesets based on their locations in the gene structure. More were located in coding regions (514 probesets) than UTRs (268 probesets) ($P < 2.2 \times 10^{-16}$, binomial test), suggesting that the majority of these population differences are potentially at the protein level.

Since the disruption of specific AS events has been implicated in several human genetic diseases (Faustino and Cooper 2003), we searched the OMIM database to see if any of the differentially spliced genes are involved in human diseases. Among the diseases found (Supplemental Table 1), FSGS (glomerulosclerosis, focal segmental, 1) is known to be more common in African Americans than Europeans (Sorof et al. 1998). We found that one probeset (PS3832645) of the causal gene *ACTN4* (actin, alpha 4) showed significantly lower SI values

in CEU, indicating possible skipping in these samples (Supplemental Table 1). Another interesting example is TIDM (type I diabetes mellitus). It has been known that fewer African American children develop type 1 diabetes (also known as juvenile onset diabetes) than white children (Diabetes Epidemiology Research International Study Group 1988). We found that two probesets of *OAS1* (2′,5′-oligoadenylate synthetase 1), which has been implicated in TIDM showed significantly lower SI values in the CEU (PS3432462, PS3432463) and YRI (PS3432451,PS3432457, PS3432458), separately, suggesting different transcript isoforms could play a role in the racial disparity of this disease (Supplemental Table 1). Interestingly, Tessier et al. recently confirmed the association of TIDM with a splicing alteration in OAS1 (Tessier et al. 2006).

Furthermore, using the DAVID (Dennis et al. 2003) web application, three GO biological processes, four PANTHER biological processes, and one KEGG pathway were found to be enriched among the 570 differentially spliced genes relative to the background (Table 1). Notably, both the enriched GO term “response to stimulus” and the enriched KEGG pathway “antigen processing and presentation” are related to immune response. We previously found that transcript clusters (gene-level) differentially expressed between the CEU and YRI samples were enriched in immune response genes (Zhang and Dolan 2008a; Zhang et al. 2008a). It has been reported that African Americans may be more susceptible to infection by certain bacteria than Caucasians (Noble and Miller 1980) and some genetic polymorphisms that may lead to different antimicrobial response (Jordan et al. 2005). Our finding that the immune response-related genes were enriched among the differentially spliced genes suggests that AS or transcript isoform variation could be a critical mechanism in defining the racial differences in the infectious diseases. Another enriched GO term is “transcription”, which includes lower level processes required for the maturation of mRNA such as “mRNA splicing via spliceosome”. In contrast, the PANTHER biological process “nucleoside, nucleotide and nucleic acid metabolism” was also enriched, suggesting that the splicing of these transcriptionrelated genes including those spliceosome-related genes (e.g., splicing factors *SFPQ* and *SFRS5*, Supplemental Table 1) could potentially be involved in the regulation of transcript isoform variation between human populations. However, since a large proportion of genes have no pathway annotation and the validation of pathways in the databases is often not rigorously performed, interpretation of these results warrants some caution.

Previous studies have shown that common genetic variants account for the population differences in gene expression (Spielman et al. 2007; Storey et al. 2007; Stranger et al. 2007; Zhang et al. 2008a, b) and transcript isoform variation within the unrelated CEU samples (Hull et al. 2007; Kwan et al. 2007, 2008). We tried to investigate if the differences in allele frequency of common genetic variants contribute to the observed differences in transcript isoform variation between the CEU and YRI samples. To identify genetic variants that account for this variation, we carried out a genome-wide eQTL analysis by associating the HapMap genotypic data (International HapMap Consortium 2003, 2005) on ~1.57 million SNP markers with the SI values of the 782 differentially spliced probesets using the QTDT software, which has the advantages of conducting the powerful total association analyses using the entire panel of samples while correcting for internal correlations among all the members (Abecasis et al. 2000a, b). A probeset associated with SNP(s) within 2.5 Mb on the same chromosome was defined as locally-regulated, while a probeset associated with SNP(s) on different chromosome (s) or more than 2.5 Mb away on the same chromosome was defined as distantly-regulated. By combining the CEU and YRI data and using population identity as a covariate, the QTDT analysis after Bonferroni correction provided us a list of SNPs whose associations with differential SI values of probesets were the most striking, suggesting that the allele frequency differences of these associated common genetic variants account for a substantial fraction of the differences in transcript isoform variation between the two populations. Notably, many of the locally associated SNPs and some distantly associated SNPs are in linkage disequilibrium

(LD) (Supplemental Table 2). For example, two local SNPs, rs2791650 and rs2791648 associated with a probeset of *FRAP1* are in complete LD (Supplemental Table 2). The allele-frequency-driven transcript isoform variation difference between the CEU and YRI samples is further illustrated in Figs. 3 and 4, which show some examples of the contribution of local genetic variants to the observed differences in transcript isoform variation. Because of the existence of both local and distant SNPs, our findings also suggest that a complete network of regulation of differential AS patterns could potentially be the result of interactions among various local and distant genetic elements. On one hand, our findings suggest that approximately 30% of the differentially spliced genes could be accounted for by the allele frequency differences of either local or distant single SNPs, an observation similar to what we observed for gene-level expression differences between these two populations (Zhang et al. 2008a). On the other hand, our findings suggest that the remainder could be due to other mechanisms such as DNA methylation or controlled by multiple SNPs.

In this study, we present the first comprehensive view of the transcript isoform variation and its regulation by genetic variants between individuals of European and African ancestry. Our results suggest that although between one-third and two-thirds of all human genes could undergo alternative splicing (Sorek et al. 2004), the proportion of genes with differential AS between human populations could be much lower (~8% based on our estimate at $P_C<0.01$). A number of biological processes such as those involving immune response and mRNA synthesis were found to be enriched in the differentially spliced genes between the CEU and YRI samples. Our results suggest that genetic variation of DNA sequence contributes to a substantial fraction of the population-level transcript isoform variation, though some other non-genetic factors could also potentially influence the observed differences between populations. Technically, although the reproducibility of the exon arrays is generally high (Affymetrix Inc. 2007; Kwan et al. 2007), one limitation of this work is that technical replicates were not available for these samples (Zhang et al. 2008a), thus limiting our focus to only sets of genes that are differentially spliced between populations. For a more comprehensive view of the AS patterns, one would need to consider inter-individual and inter-population variation together. Finally, in addition to the intrinsic limitations of using the HapMap samples (e.g., one tissue type), there are other challenges and confounding factors (such as capturing unknown SNPs, YRI samples collected decades after CEU) that might be considered in future studies (Zhang and Dolan 2008b, c) to help us better utilize this tremendous resource to yield new insights into the alternative splicing process in humans.

Data availability

Gene expression data deposited in Gene Expression Omnibus (GEO): GSE9703.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This Pharmacogenetics of Anticancer Agents Research (PAAR) Group (http://www.pharmacogenetics.org) study was supported by NIH/NIGMS grants U01 GM61393 and U01 GM61374. We are grateful to Dr. Jeong-Ah Kang for maintaining cell lines, Cheryl A. Roe for reviewing the manuscript and Drs. James Fackenthal and Emily Kistner for helpful discussion. T.A.C., T.X.C., A.C.S., and J.E.B. are employees of Affymetrix, Inc.

References

Abecasis GR, Cardon LR, Cookson WO. A general test of association for quantitative traits in nuclear families. Am J Hum Genet 2000a;66:279–292. [PubMed: 10631157]

Abecasis GR, Cookson WO, Cardon LR. Pedigree tests of transmission disequilibrium. Eur J Hum Genet 2000b;8:545–551. [PubMed: 10909856]

Affymetrix Inc.. Affymetrix Technical Note. 2006. Identifying and validating alternative splicing events.

Affymetrix Inc.. Affymetrix GeneChip Gene and Exon Array Whitepaper Collection. 2007. Human Gene 1.0 ST Array Performance.

Alberts R, Terpstra P, Li Y, Breitling R, Nap JP, Jansen RC. Sequence polymorphisms cause many false cis eQTLs. PLoS ONE 2007;2:e622. [PubMed: 17637838]

Applied Biosystems. Technical Note. 2004. Guide to performing relative qualification of gene expression using Real-Time quantitative PCR.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000;25:25–29. [PubMed: 10802651]

Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B 1995;57:289–300.

Brinkman BM. Splice variants as cancer biomarkers. Clin Biochem 2004;37:584–594. [PubMed: 15234240]

Cheung VG, Conlin LK, Weber TM, Arcaro M, Jen KY, Morley M, Spielman RS. Natural variation in human gene expression assessed in lymphoblastoid cells. Nat Genet 2003;33:422–425. [PubMed: 12567189]

Dennis G Jr, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol 2003;4:P3. [PubMed: 12734009]

Diabetes Epidemiology Research International Study Group. Geographic patterns of childhood insulin-dependent diabetes mellitus. Diabetes Epidemiology Research International Group. Diabetes 1988;37:1113–1119. [PubMed: 3391346]

Duan S, Huang RS, Zhang W, Bleibel WK, Roe CA, Clark TA, Chen TX, Schweitzer AC, Blume JE, Cox NJ, Dolan ME. Genetic architecture of transcript-level variation in humans. Am J Hum Genet 2008a;82:1101–13. [PubMed: 18439551]

Duan S, Zhang W, Bleibel WK, Cox NJ, Dolan ME. SNPinProbe_1.0: a database for filtering out probes in the Affymetrix GeneChip®g.0 ST array potentially affected by SNPs. Bioinformation 2008b; 2:469–470. [PubMed: 18841244]

Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA 1998;95:14863–14868. [PubMed: 9843981]

Faustino NA, Cooper TA. Pre-mRNA splicing and human disease. Genes Dev 2003;17:419–437. [PubMed: 12600935]

Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, Pasternak S, Wheeler DA, Willis TD, Yu F, Yang H, Zeng C, Gao Y, Hu H, Hu W, Li C, Lin W, Liu S, Pan H, Tang X, Wang J, Wang W, Yu J, Zhang B, Zhang Q, Zhao H, Zhao H, Zhou J, Gabriel SB, Barry R, Blumenstiel B, Camargo A, Defelice M, Faggart M, Goyette M, Gupta S, Moore J, Nguyen H, Onofrio RC, Parkin M, Roy J, Stahl E, Winchester E, Ziaugra L, Altshuler D, Shen Y, Yao Z, Huang W, Chu X, He Y, Jin L, Liu Y, Shen Y, Sun W, Wang H, Wang Y, Wang Y, Xiong X, Xu L, Waye MM, Tsui SK, Xue H, Wong JT, Galver LM, Fan JB, Gunderson K, Murray SS, Oliphant AR, Chee MS, Montpetit A, Chagnon F, Ferretti V, Leboeuf M, Olivier JF, Phillips MS, Roumy S, Sallee C, Verner A, Hudson TJ, Kwok PY, Cai D, Koboldt DC, Miller RD, Pawlikowska L, Taillon-Miller P, Xiao M, Tsui LC, Mak W, Song YQ, Tam PK, Nakamura Y, Kawaguchi T, Kitamoto T, Morizono T, Nagashima A, Ohnishi Y, et al. A second generation human haplotype map of over 3.1 million SNPs. Nature 2007;449:851–861. [PubMed: 17943122]

Gardina PJ, Clark TA, Shimada B, Staples MK, Yang Q, Veitch J, Schweitzer A, Awad T, Sugnet C, Dee S, Davies C, Williams A, Turpaz Y. Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array. BMC Genomics 2006;7:325. [PubMed: 17192196]

Gilad Y, Rifkin SA, Bertone P, Gerstein M, White KP. Multispecies microarrays reveal the effect of sequence divergence on gene expression profiles. Genome Res 2005;15:674–680. [PubMed: 15867429]

Huang da W, Sherman BT, Tan Q, Collins JR, Alvord WG, Roayaei J, Stephens R, Baseler MW, Lane HC, Lempicki RA. The DAVID gene functional classification tool: a novel biological module-centric algorithm to functionally analyze large gene lists. Genome Biol 2007;8:R183. [PubMed: 17784955]

Huang RS, Kistner EO, Bleibel WK, Shukla SJ, Dolan ME. Effect of population and gender on chemotherapeutic agentinduced cytotoxicity. Mol Cancer Ther 2007;6:31–36. [PubMed: 17237264]

Hull J, Campino S, Rowlands K, Chan MS, Copley RR, Taylor MS, Rockett K, Elvidge G, Keating B, Knight J, Kwiatkowski D. Identification of common genetic variation that modulates alternative splicing. PLoS Genet 2007;3:e99. [PubMed: 17571926]

International HapMap Consortium. The International HapMap Project. Nature 2003;426:789–796. [PubMed: 14685227]

International HapMap Consortium. A haplotype map of the human genome. Nature 2005;437:1299–1320. [PubMed: 16255080]

Ioannidis JP, Ntzani EE, Trikalinos TA. `Racial' differences in genetic effects for complex diseases. Nat Genet 2004;36:1312–1318. [PubMed: 15543147]

Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 2003;4:249–264. [PubMed: 12925520]

Jordan WJ, Eskdale J, Lennon GP, Pestoff R, Wu L, Fine DH, Gallagher G. A non-conservative, coding single-nucleotide polymorphism in the N-terminal region of lactoferrin is associated with aggressive periodontitis in an African-American, but not a Caucasian population. Genes Immun 2005;6:632–635. [PubMed: 16208406]

Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for deciphering the genome. Nucleic Acids Res 2004;32:D277–D280. [PubMed: 14681412]

Kurian AK, Cardarelli KM. Racial and ethnic differences in cardiovascular disease risk factors: a systematic review. Ethn Dis 2007;17:143–152. [PubMed: 17274224]

Kwan T, Benovoy D, Dias C, Gurd S, Serre D, Zuzan H, Clark TA, Schweitzer A, Staples MK, Wang H, Blume JE, Hudson TJ, Sladek R, Majewski J. Heritability of alternative splicing in the human genome. Genome Res 2007;17:1210–1218. [PubMed: 17671095]

Kwan T, Benovoy D, Dias C, Gurd S, Provencher C, Beaulieu P, Hudson TJ, Sladek R, Majewski J. Genome-wide analysis of transcript isoform variation in humans. Nat Genet 2008;40:225–231. [PubMed: 18193047]

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, et al. Initial sequencing and analysis of the human genome. Nature 2001;409:860–921. [PubMed: 11237011]

McGarvey PB, Huang H, Barker WC, Orcutt BC, Garavelli JS, Srinivasarao GY, Yeh LS, Xiao C, Wu CH. PIR: a new resource for bioinformatics. Bioinformatics 2000;16:290–291. [PubMed: 10869023]

McKusick, VA. A catalog of human genes and genetic disorders. Vol. 12th edn.. Johns Hopkins University Press; Baltimore: 1998. Mendelian inheritance in man.

Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, Spielman RS, Cheung VG. Genetic analysis of genome-wide variation in human gene expression. Nature 2004;430:743–747. [PubMed: 15269782]

Noble RC, Miller BR. Auxotypes and antimicrobial susceptibilities of *Neisseria gonorrhoeae* in black and white patients. Br J Vener Dis 1980;56:26–30. [PubMed: 6768418]

Novoyatleva T, Tang Y, Rafalska I, Stamm S. Pre-mRNA missplicing as a cause of human disease. Prog Mol Subcell Biol 2006;44:27–46. [PubMed: 17076263]

Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 2007;35:D61–D65. [PubMed: 17130148]

R Development Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing; Vienna: 2005.

Rozen S, Skaletsky H. Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol 2000;132:365–386. [PubMed: 10547847]

Saeed AI, Sharov V, White J, Li J, Liang W, Bhagabati N, Braisted J, Klapa M, Currier T, Thiagarajan M, Sturn A, Snuffin M, Rezantsev A, Popov D, Ryltsov A, Kostukovich E, Borisovsky I, Liu Z, Vinsavich A, Trush V, Quackenbush J. TM4: a free, open-source system for microarray data management and analysis. Biotechniques 2003;34:374–378. [PubMed: 12613259]

Sorek R, Shamir R, Ast G. How prevalent is functional alternative splicing in the human genome? Trends Genet 2004;20:68–71. [PubMed: 14746986]

Sorof JM, Hawkins EP, Brewer ED, Boydstun II, Kale AS, Powell DR. Age and ethnicity affect the risk and outcome of focal segmental glomerulosclerosis. Pediatr Nephrol 1998;12:764–768. [PubMed: 9874323]

Spielman RS, Bastone LA, Burdick JT, Morley M, Ewens WJ, Cheung VG. Common genetic variants account for differences in gene expression among ethnic groups. Nat Genet 2007;39:226–231. [PubMed: 17206142]

Storey JD, Madeoy J, Strout JL, Wurfel M, Ronald J, Akey JM. Gene-expression variation within and among human populations. Am J Hum Genet 2007;80:502–509. [PubMed: 17273971]

Stranger BE, Forrest MS, Clark AG, Minichiello MJ, Deutsch S, Lyle R, Hunt S, Kahl B, Antonarakis SE, Tavare S, Deloukas P, Dermitzakis ET. Genome-wide associations of gene expression variation in humans. PLoS Genet 2005;1:e78. [PubMed: 16362079]

Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, Beazley C, Ingle CE, Dunning M, Flicek P, Koller D, Montgomery S, Tavare S, Deloukas P, Dermitzakis ET. Population genomics of human gene expression. Nat Genet 2007;39:1217–1224. [PubMed: 17873874]

Tessier MC, Qu HQ, Frechette R, Bacot F, Grabs R, Taback SP, Lawson ML, Kirsch SE, Hudson TJ, Polychronakos C. Type 1 diabetes and the OAS gene cluster: association with splicing polymorphism or haplotype? J Med Genet 2006;43:129–132. [PubMed: 16014697]

Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A. PANTHER: a library of protein families and subfamilies indexed by function. Genome Res 2003;13:2129–2141. [PubMed: 12952881]

Thorisson GA, Smith AV, Krishnan L, Stein LD. The International HapMap Project Web site. Genome Res 2005;15:1592–1593. [PubMed: 16251469]

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, et al. The sequence of the human genome. Science 2001;291:1304–1351. [PubMed: 11181995]

Westfall, PH.; Young, SS. Resampling-based multiple testing: examples and methods for p-value adjustment. Wiley Publishers; New York: 1993.

Wright S. Genetical structure of populations. Nature 1950;166:247–249. [PubMed: 15439261]

Zhang W, Dolan ME. Ancestry-related differences in gene expression: findings may enhance understanding of health disparities between populations. Pharmacogenomics 2008a;9:489–492. [PubMed: 18466094]

Zhang W, Dolan ME. Beyond the HapMap genotypic data: prospects of deep resequencing projects. Curr Bioinform 2008b;3

Zhang W, Dolan ME. On the challenges of the HapMap resource. Bioinformation 2008c;2:238–239. [PubMed: 18317571]

Zhang W, Bleibel WK, Roe CA, Cox NJ, Dolan M Eileen. Gender-specific differences in expression in human lymphoblastoid cell lines. Pharmacogenet Genomics 2007;17:447–450. [PubMed: 17502836]

Zhang W, Duan S, Kistner EO, Bleibel WK, Huang RS, Clark TA, Chen TX, Schweitzer AC, Blume JE, Cox NJ, Dolan ME. Evaluation of genetic variation contributing to differences in gene expression between populations. Am J Hum Genet 2008a;82:631–640. [PubMed: 18313023]

Zhang W, Ratain MJ, Dolan ME. The HapMap resource is providing new insights into ourselves and its application to pharmacogenomics. Bioinform Biol Insights 2008b;2:15–23. [PubMed: 18392109]

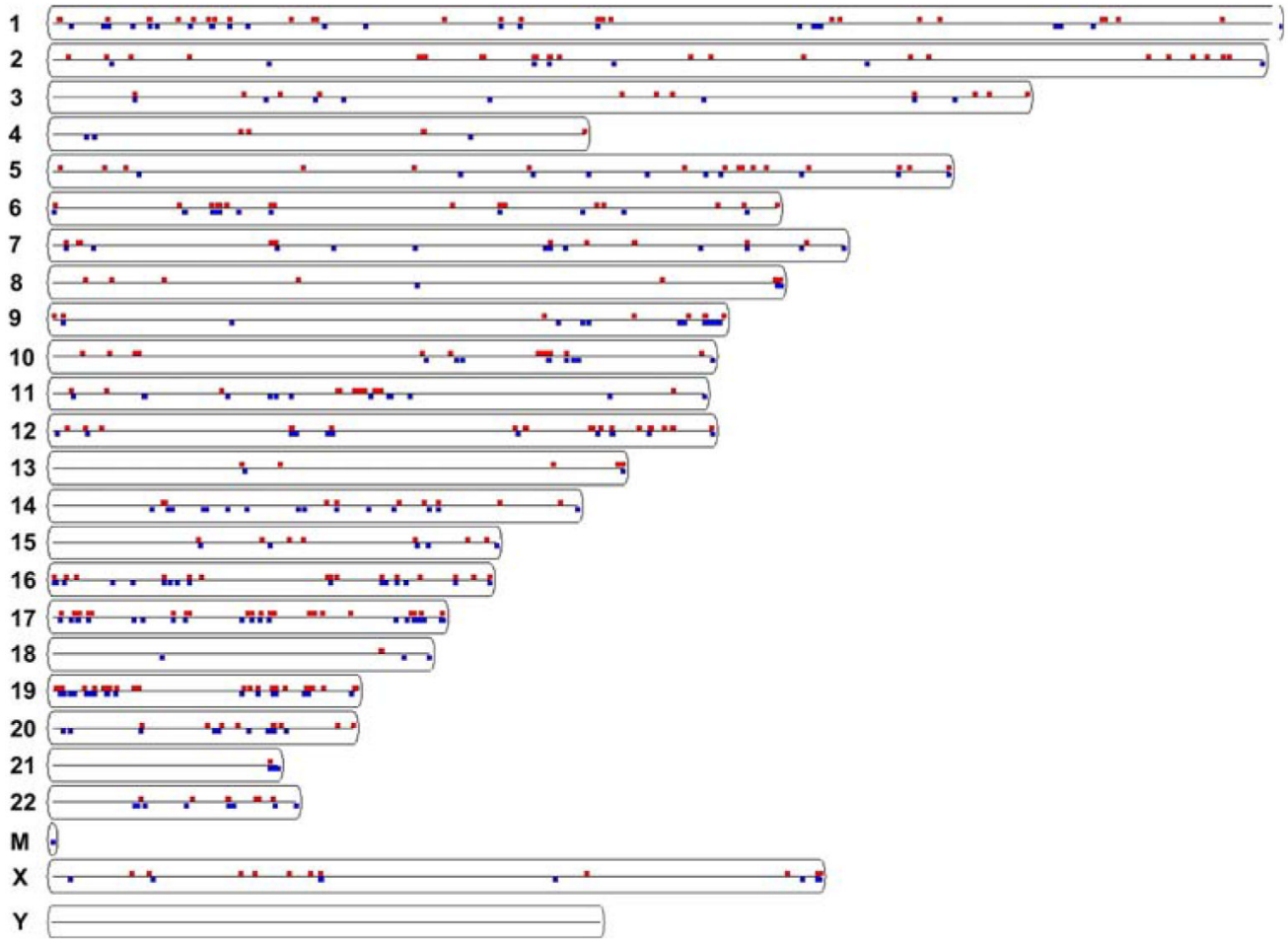

Fig. 1.

Genomic distribution of the differentially spliced probesets between the CEU and YRI samples. 397 probesets had significantly lower SI values in the CEU samples (*top ticks* along chromosomes), while 385 probesets had significantly lower SI values in the YRI samples (*bottom ticks* along chromosomes)

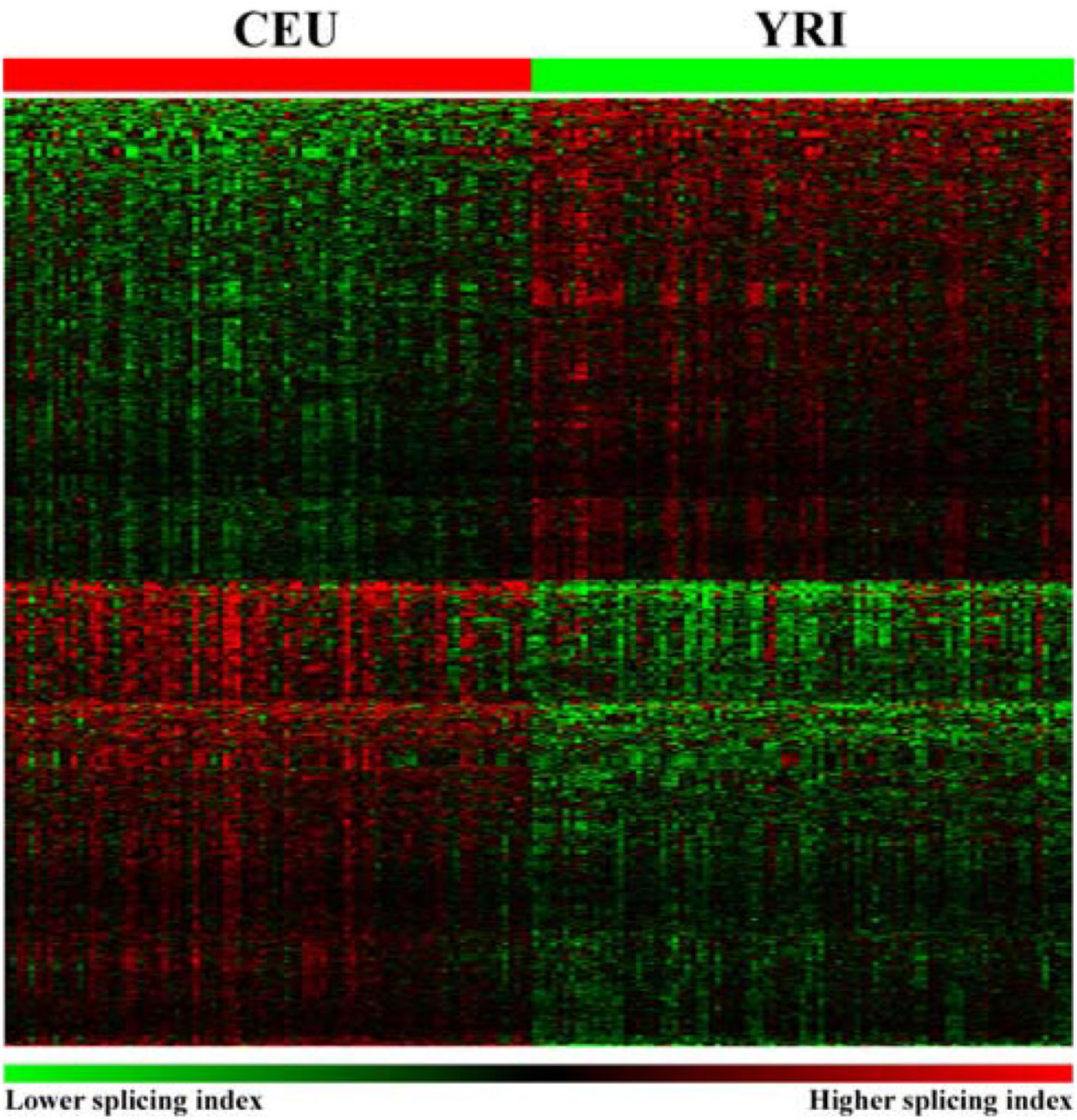

Fig. 2.
Cluster analysis of the differentially spliced probesets. The 782 differentially spliced probesets were grouped into two clusters representing the two populations based on their splicing index values. The columns are cell lines and the rows are probesets

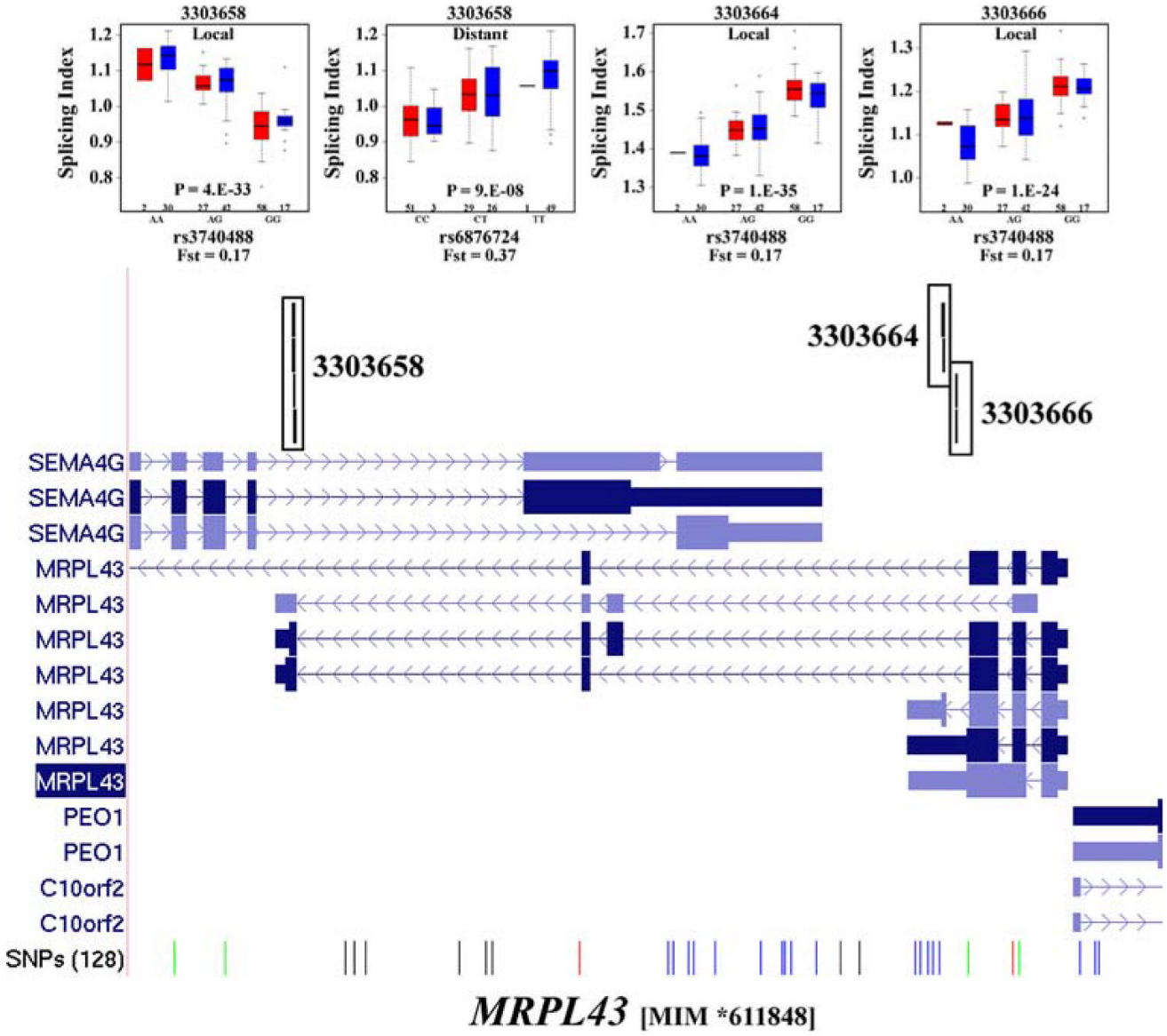

Fig. 3.

Common genetic variants account for transcript isoform variation of *MRPL43* between populations. Probesets PS3303658, PS3303664 and PS3303666 of MRPL43 were differentially spliced between the CEU and YRI samples. PS3303658 had a lower splicing index in the CEU samples. PS3303664 and PS303666 had a lower splicing index in the YRI samples (Supplemental Table 1). The *box plots* show some local SNPs associated with the splicing index under an additive model

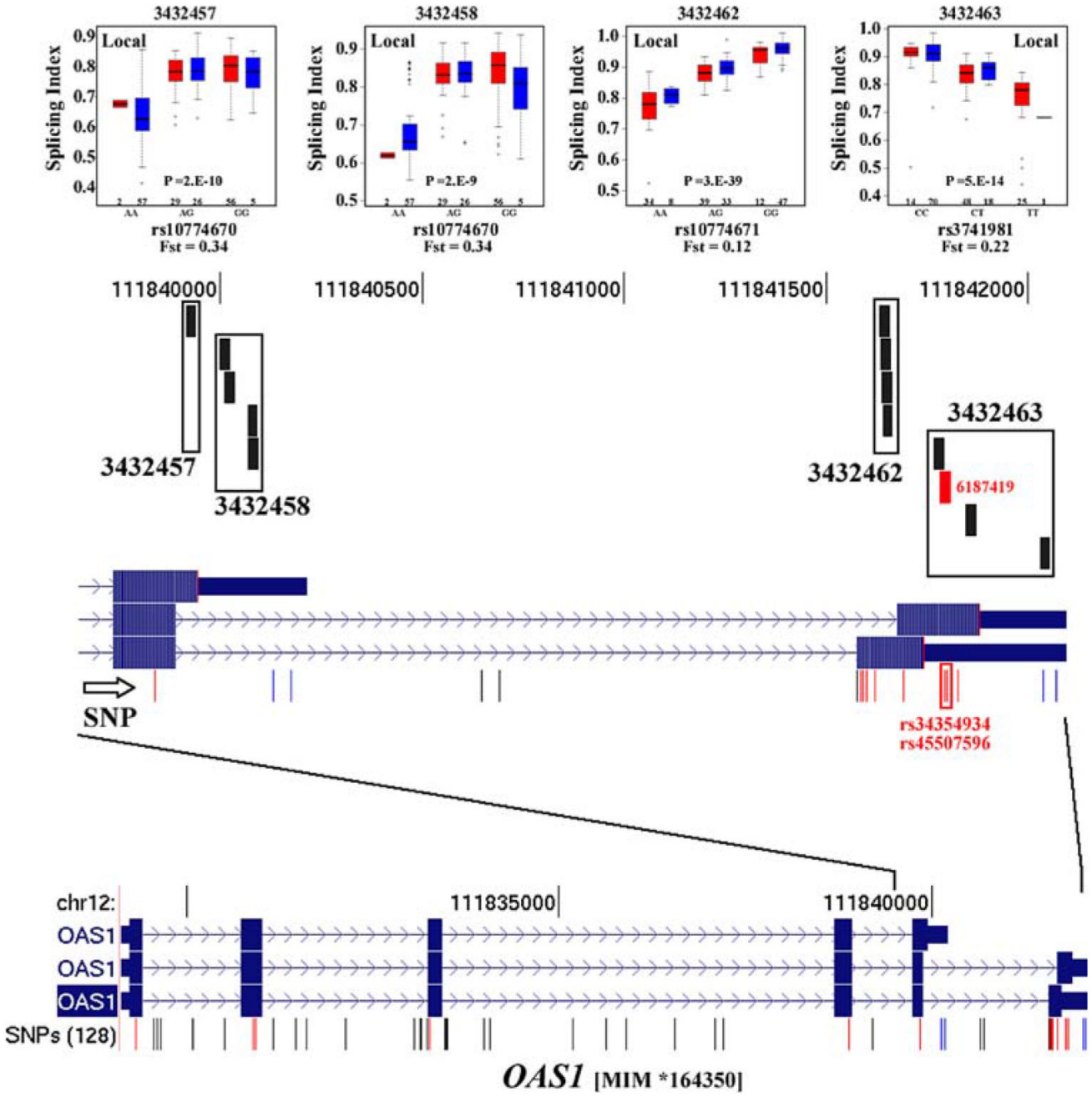

Fig. 4.

Common genetic variants account for transcript isoform variation of *OAS1* between populations. Probesets PS3432457, PS3432458, PS3432462 and PS3432463 and PS3432463 of *OAS1* were differentially spliced between the CEU and YRI samples. PS3432462 had a lower splicing index in the CEU samples. PS3432457 and PS3432458 had a lower splicing index in the YRI samples (Supplemental Table 1). The *box plots* show some local SNPs associated with the splicing index under an additive model. The *red bar* in PS3432463 indicates a probe potentially affected by two known SNPs (rs3435934 and rs45507596) in dbSNP v129. This probe was filtered before summarizing the probeset intensity

Table 1

Enriched biological processes and pathways among the genes with differentially spliced probesets

Category	Database ID	Term	Gene counts	*P*	P_c [a] (×10^{-1})
GO biological process	0050896	Response to stimulus	91	4.00×10^{-02}	1.90
	0050794	Regulation of cellular process	188	3.90×10^{-02}	4.70
	0006350	Transcription	120	3.70×10^{-02}	5.00
PANTHER biological process	BP00031	Nucleoside, nucleotide and nucleic acid metabolism	161	5.50×10^{-04}	1.10
	BP00139	Asymmetric protein localization	8	1.70×10^{-03}	1.70
	BP00224	Cell proliferation and differentiation	51	3.00×10^{-03}	1.90
	BP00285	Cell structure and motility	57	1.10×10^{-03}	4.50
KEGG pathway	Hsa04612	Antigen processing and presentation	10	3.70×10^{-03}	5.00

[a] After BH correction

Table 2

Representative local SNPs associated with differential SI values ($P_c < 0.05$ after Bonferroni correction)

Lower SI	Affymetrix PS ID	Affymetrix PS ID	Chromosome	Gene symbol	SNP	SNP-gene symbol	*P*	F_{st}
CEU	2367743	2367757	1	PRDX6	rs12092383	ANKRD45	2.E-08	0.43
	2396537	2396584	1	FRAP1	rs1318348	EXOSC10	3.E-14	0.48
	2398073	2398103	1	FBXO42	rs17419150	SPATA21	5.E-24	0.16
	2404521	2404524	1	PEF1	rs10914464	COL16A1	1.E-11	0.45
	2423264	2423272	1	TMED5	rs797680	CCDC18	2.E-53	0.40
	2440327	2440329	1	SLAMF1	rs977019	CD 84	1.E-40	0.28
	2540317	2540337	2	PDIA6	rs1686482	ATP6V1C2,PDIA6	4.E-28	0.25
	2541230	2541373	2	NAG	rs2031011	NAG	7.E-27	0.32
	2663551	2663554	3	NUP210	rs877511	NUP210	1.E-18	0.22
	2779897	2779932	4	MANBA	rs223326	CISD2	4.E-22	0.15
	2863964	2864000	5	ARSB	rs10462560	ARSB	3.E-10	0.25
	2899340	2899343	6	BTN2A2	rs9467745	BTN2A2	8.E-12	0.24
	2902559	2902572	6	CSNK2B	rs2844463	BAT3	7.E-11	0.40
	2927722	2927732	6	HEBP2	rs6570232	HEBP2	3.E-22	0.27
	2950629	2950643	6	TAPBP	rs1800838	TAPBP	8.E-09	0.21
	3126087	3126102	8	ASAH1	rs208024	PCM1	7.E-11	0.22
	3157901	3157977	8	PLEC1	rs7014582	PLEC1	1.E-16	0.29
	3221916	3221949	9	AKNA	rs10739408	AKNA	1.E-31	0.22
	3281068	3281092	10	LOC643475	rs10828316	PIP4K2A	2.E-20	0.16
	3301218	3301223	10	PDLIM1	rs11188246	PDLIM1	4.E-19	0.21
	3303652	3303658	10	MRPL43	rs3740488	C10orf2	4.E-33	0.17
	3333711	3333716	11	SLC3A2	rs3763851	NXF1,STX5	1.E-14	0.35
	3340589	3340610	11	SERPINH1	rs646474	SERPINH1	1.E-10	0.26
	3394123	3394161	11	HYOU1	rs592190	HMBS	2.E-35	0.20
	3432438	3432462	12	OAS1	rs10744785	OAS1	5.E-10	0.26
	3457824	3457837	12	TIMELESS	rs3809125	MIP	5.E-09	0.18
	3634071	3634080	15	TSPAN3	rs16968623	PSTPIP1	2.E-08	0.23
	3704896	3704899	16	PCOLN3	rs447735	C16orf55	2.E-28	0.19
	3707642	3707692	17	RABEP1	rs2641263	C17orf87,LOC100130950	7.E-10	0.32
	3740479	3740522	17	PRPF8	rs11078563	PRPF8	2.E-08	0.21

Lower SI	Affymetrix PS ID	Affymetrix PS ID	Chromosome	Gene symbol	SNP	SNP-gene symbol	*P*	F_{st}
	3742627	3742635	17	UNQ5783	rs2585281	C17orf87	8.E-49	0.38
	3744800	3744805	17	STX8	rs9909240	NTN1	5.E-10	0.16
	3766651	3766657	17	ERN1	rs77684	ERN1	1.E-09	0.59
	3770743	3770754	17	GRB2	rs4542691	GRB2	2.E-08	0.49
	3812426	3812475	18	RTTN	rs11876150	RTTN	1.E-08	0.76
	3899173	3899201	20	RRBP1	rs2236250	RRBP1	1.E-09	0.38
	2391647	2391652	1	SSU72	rs3766169	SSU72	9.E-23	0.50
	2560122	2560131	2	GCS1	rs1047911	CCDC142,MRPL53	6.E-19	0.58
YRI	2670784	2670807	3	SEC22C	rs663673	NKTR	1.E-21	0.22
	2795819	2795829	4	DCTD	rs7677967	DCTD	2.E-15	0.26
	2868131	2868133	5	ARTS-1	rs27039	ERAP1	3.E-13	0.16
	2871923	2871928	5	ATG12	rs1058600	ATG12	7.E-28	0.28
	2888519	2888546	5	RAP80	rs2940531	UIMC1	2.E-34	0.21
	2950329	2950333	6	HLA-DPA1	rs10214910	HLA-DPA1	5.E-14	0.53
	2954771	2954792	6	GTPBP2	rs9472084	POLH	9.E-18	0.31
	2963929	2963964	6	RNGTT	rs2756369	RNGTT	2.E-10	0.37
	3037304	3043037	7	MGC12966	rs3750040	MGC12966	2.E-17	0.27
	3067080	3067144	7	COG5	rs2299421	COG5	5.E-11	0.34
	3190558	3190619	9	SPTAN1	rs3737308	SPTAN1	6.E-12	0.36
	3224650	3224806	9	DENND1A	rs2479104	DENND1A	2.E-20	0.48
	3279575	3279625	10	RSU1	rs7910261	RSU1	3.E-15	0.45
	3289631	3289645	10	CSTF2T	rs2292828	CSTF2T,PRKG1	2.E-14	0.31
	3294242	3294253	10	ECD	rs12258241	ANXA7	3.E-16	0.28
	3322717	3322759	11	GTF2H1	rs4150581	GTF2H1	6.E-09	0.38
	3393200	3393213	11	PCSK7	rs10790175	PAFAH1B2	2.E-11	0.34
	3474697	3474728	12	SPPL3	rs7977343	UNQ1887	7.E-11	0.20
	3558071	3558081	14	RABGGTA	rs729421	RABGGTA	3.E-12	0.23
	3566304	3566310	14	EXOC5	rs7141198	C14orf108	1.E-10	0.72
	3678395	367839	16	N-PAC	rs9923349	UBN1	8.E-16	0.21
	3759077	3759083	17	CGI-69	rs11654436	RUNDC3A	1.E-57	0.20
	3821015	3821042	19	LDLR	rs6413504	LDLR	1.E-14	0.16

Lower SI	Affymetrix PS ID	Affymetrix PS ID	Chromosome	Gene symbol	SNP	SNP-gene symbol	*P*	F_{st}
	3838118	3838136	19	RUVBL2	rs1062708	RUVBL2	6.E-12	0.17
	3959631	3959764	22	EIF3S7	rs9610529	EIF3D	1.E-19	0.28
	3962494	3962500	22	POLDIP3	rs137124	CYB5R3	6.E-10	0.58
	3965393	3965401	22	ALG12	rs1321	ALG12	3.E-11	0.47