Humu-2013-0382

Research Article

Discovery and functional assessment of gene variants in the vascular endothelial growth factor pathway

Laia Paré-Brunet[1], Dylan Glubb[2], Patrick Evans[3], Antoni Berenguer-Llergo[4], Amy S. Etheridge[2], Andrew D. Skol[3], Anna Di Rienzo[5], Shiwei Duan[6], Eric R. Gamazon[3], Federico Innocenti[2*]

[1]Department of Genetics, Hospital de la Santa Creu i Sant Pau. Barcelona, Spain

[2]Eshelman School of Pharmacy, Institute for Pharmacogenomics and Individualized Therapy, Lineberger Comprehensive Cancer Center, School of Medicine, University of North Carolina at Chapel Hill, NC, USA

[3]Department of Medicine, University of Chicago, Chicago, IL, USA

[4]Biomarkers and Susceptibility Unit, Catalan Institute of Oncology (ICO-IDIBELL), L'Hospitalet de Llobregat, Barcelona. CIBER de Epidemiologia y Salud Pública (CIBERESP), Instituto de Salud Carlos III, Spain.

[5]Department of Genetics, University of Chicago, Chicago, IL, USA

[6]School of Medicine, Ningbo University, Zhejiang, China, 315211

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as DOI: 10.1002/humu.22475.

*Corresponding author:

Federico Innocenti, MD, PhD

University of North Carolina at Chapel Hill

1014 Genetic Medicine Bldg

CB 7361, 120 Mason Farm Rd.

Chapel Hill, NC 27599-7361

Tel: 919-966-9422

Fax: 919-966-5863

E-mail: innocent@unc.edu

Research support:

This work was supported by the National Institutes of Health (NIH/NCI K07CA140390-01 and NIH/NIGMS U01 GM61393) and the National Heart, Lung, and Blood Institute (NHLBI) Resequencing and Genotyping (RS&G) Service. LPB is the recipient of a fellowship from the Instituto de Salud Carlos III (FIS 080199) and Beques per a estades de recerca a l'estranger (Be2-2009).

ABSTRACT

Angiogenesis is a host-mediated mechanism in disease pathophysiology. The vascular endothelial growth factor (VEGF) pathway is a major determinant of angiogenesis, and a comprehensive annotation of the functional variation in this pathway is essential to understand the genetic basis of angiogenesis-related diseases. We assessed the allelic heterogeneity of gene expression, population specificity of cis expression quantitative trait loci (eQTLs), and eQTL function in luciferase assays in CEU and YRI HapMap lymphoblastoid cell lines (LCLs) in 23 resequenced genes. Among 356 cis-eQTLs, 155 and 174 were unique to CEU and YRI, respectively, and 27 were shared between CEU and YRI. Two cis-eQTLs provided mechanistic evidence for two GWAS findings. Five eQTLs were tested for function in luciferase assays and the effect of two *KRAS* variants was concordant with the eQTL effect. Two eQTLs found in each of *PRKCE, PIK3C2A,* and *MAP2K6,* could predict 44, 37 and 45% of the variance in gene expression, respectively. This is the first analysis focusing on the pattern of functional genetic variation of the VEGF pathway genes in CEU and YRI populations and providing mechanistic evidence for genetic association studies of diseases for which angiogenesis plays a pathophysiologic role.

Key words: VEGF, angiogenesis, pharmacogenetics, pathway, GWAS, QTL, expression

INTRODUCTION

Angiogenesis is an important biological mechanism for physiological processes such as reproduction, vasculature development and wound healing. However, angiogenesis can become dysregulated, playing a role in many pathologies including autoimmune diseases, macular degeneration, endometriosis, atherosclerosis and cancer (Folkman, 2007). As angiogenesis is a host-mediated biological mechanism, the extent of the angiogenic response to a given stimulus might be partly dependent upon genetic differences in key angiogenesis genes (Folkman, 2007). Thus, germline variation in genes of angiogenesis has the potential to impact these responses and affect the development or course of angiogenesis-dependent diseases. Indeed, germline genetic variants in genes of the vascular endothelial growth factor (VEGF) pathway (Figure 1) have been found to associate with diseases such as cancer, sarcoidosis, coronary heart disease, diabetic retinopathy, arthritis and psoriasis (reviewed in (Rogers and D'Amato, 2011)).

The VEGF-pathway is a key network of proteins which regulate angiogenesis (Hicklin and Ellis, 2005). VEGF is a family of five structurally related proteins: VEGFA, VEGFB, VEGFC, VEGFD and placental growth factor (PGF). These proteins act as ligands, and binding to a VEGF receptor (VEGFR) can lead to signal transduction via downstream proteins. The most important of these ligands in angiogenesis is VEGFA (Olsson et al., 2006), which can be secreted by a variety of cells and binds to two receptors: VEGFR1 (*FLT1*) and VEGFR2 (*KDR*)(Hicklin and Ellis, 2005). The binding of VEGFA to VEGFR2, in particular, leads to a cascade of protein activation through signaling cellular effectors and results in increased endothelial cell survival, proliferation, migration and differentiation (Figure 1) (Bernatchez et al., 2002).

As a consequence of its central role in angiogenesis, the VEGF-pathway has become a target for treatment of angiogenesis-dependent diseases, with several VEGF-pathway

inhibitors in clinical use and more in development for cancer treatment (reviewed in (Teicher, 2011)). Furthermore, germline genetic variations in VEGF-pathway genes have been found to be associated with responses to angiogenesis inhibitors (reviewed in (Teicher, 2011; Schneider et al., 2012)). Nevertheless, complete information on the composition and functional effects of variation in the VEGF-pathway genes is lacking and is needed to rationally examine and interpret the effects of these genes on related clinical phenotypes.

Currently, a major limitation of the association between VEGF-pathway genetic variants and clinical traits is that mechanistic explanations of genetic associations are lacking in a majority of studies (Rogers and D'Amato, 2011). This is a consequence of the limited information about the effects of genetic variation in the VEGF-pathway at the molecular level. This knowledge is essential to support statistical associations and to prioritize SNPs (single nucleotide polymorphisms) for prospective testing of their ability to predict the clinical responses of patients. We have previously characterized the functional genetics of *KDR* (VEGFR2) (Glubb et al., 2011), but a comprehensive analysis of the entire VEGF pathway is not yet available. This study aims to provide a framework and mechanistic basis to address these fundamental gaps in knowledge.

To identify potentially functional genetic variants in the VEGF-pathway, we have examined lymphoblastoid cell lines (LCLs) for VEGF-pathway gene expression quantitative trait loci (eQTL). LCLs have proven a useful model for eQTL studies as a number of LCLs, derived from individuals from several HapMap population groups, have been well characterized by genotyping and gene expression studies (Frazer et al., 2007; Zhang et al., 2008; Zhang et al., 2009). To validate the functional effects of selected candidate variants, we have performed reporter gene assays. The eQTL information on VEGF-genes has also been used to interpret current GWAS results for disease traits.

MATERIALS AND METHODS

DNA samples and candidate genes for resequencing

All of the samples utilized for these studies were contributed with consent to broad data release and to their use in many future studies, including for extensive genotyping and sequencing, gene expression and proteomics studies, and all other types of genetic variation research. The samples include no identifying or phenotypic information. HapMap panel 2 DNA samples from healthy unrelated individuals of the CEPH families (CEU, n=23) and the Yoruba people of Ibadan, Nigeria (YRI, n=24) (Supp. Table S1) were obtained from the Coriell Cell Repository (Camden, NJ) and were used for resequencing. To select and prioritize genes for resequencing, we used the information on the VEGF pathway from PharmGKB (Figure 1). Among 55 genes in the VEGF pathway (see full gene list at (PharmGKB, 2013)), 23 genes were selected based upon their biological importance in endothelial function and VEGF signaling (Table 1).

PCR and sequencing methods

The sequencing was conducted by the NHLBI DNA Resequencing and Genotyping Service (http://rsng.nhlbi.nih.gov/scripts/index.cfm). For each gene, sequencing was performed to the full genomic region of the gene (2 kb 5'-flanking, all introns and exons, and 2 kb 3'-flanking), or to a focused "standard coverage" (2 kb 5'-flanking, exons, evolutionarily conserved rat and mouse non-coding sequences, and 2 kb 3'-flanking) for genes larger than 70 kb. In some genes, customization to specific genomic regions was required. In brief, 5'-M13 tailed-gene specific PCR primers were designed to cover the target region with amplicon sizes ranging from 500-750 bp and with a minimum of 100 bp overlap between adjacent amplicons, where applicable, resulting in double-stranded coverage of all targeted regions. Overlapping amplicons were used to validate gene-specific primer sequences in

independent experiments and rule out the possibility of allele-specific PCR amplifications. All primer sequences were compared to the whole genome assembly hg18 to verify uniqueness against pseudogenes and gene families. Following temperature gradient optimization of small-scale reactions to determine optimal thermal cycling conditions, production level PCR amplifications were performed in 96-well plates in a volume of 7 µl comprising 0.2 µl each of 7 µM forward and reverse primers, 2.8 µl DNA (5 ng/µl), and 0.4 µl elongase enzyme (Invitrogen, Carlsbad, CA) or iProof polymerase (Bio-Rad, Hercules, CA) per well. Following evaluation of PCR products by 1% agarose gel electrophoresis, reactions were diluted four to six fold in ddH_2O to eliminate an extra purification step prior to sequencing.

Sequencing reactions were performed in MJ Tetrad PTC 225 thermal cyclers in 384-well format by using 5% BDT v3.1 sequencing chemistry (ABI, Foster City, CA). Reaction products were precipitated in ethanol with CleanSeq magnetic beads (Agencourt Bioscience, Beverly, MA). Perkin Elmer Minitrak, Multiprobe, and Evolution P3 robots were used to automate liquid handling in the setup of PCR, sequencing reactions and precipitation reactions. Reaction products were air dried and diluted to 30 µl with ddH_2O. Chromatograms were generated from sequence reaction on an Applied Biosystems ABI 3730XL capillary sequencer (ABI, Foster City, CA). Data flow was tracked by using a custom-designed LIMS system.

SNP analysis, quality control, sequencing coverage, and linkage disequilibrium (LD) analysis

All chromatograms were base-called using Phred, assembled into contigs using Phrap, and scanned for SNPs with PolyPhred, version 6.15 to identify polymorphic sites (Stephens et al., 2006). Data quality was monitored and assessed at multiple production checkpoints using

numerous methods. For example, each chromatogram was trimmed to remove low-quality sequence (Phred score <25), resulting in analyzed reads averaging >450 bp with an average Phred quality of 40. Following assembly of all chromatograms onto an initial reference sequence, putative polymorphic sites were selectively reviewed using Consed (Gordon et al., 1998). Individual polymorphic sites in regions with lower quality data, ambiguous base calls, deviations from Hardy-Weinberg equilibrium (HWE) or those identified using laboratory quality control tools were reviewed to eliminate potential false positive positions. This approach generated sequence-based SNP genotypes with accuracy >99.9%. Variations were deposited into a custom postgreSQL database, formatted and submitted to dbSNP for assignment of ss and rs identification numbers.

Sequencing coverage of the VEGF-pathway genes is described in Supp. File S1 according to the GenBank database (http://www.ncbi.nlm.nih.gov/genbank/). *VEGFA* genotype information from the same HapMap individuals used for resequencing was obtained from HapMap data using the default coverage. In all sequenced genes, we performed quality control by comparing our resequencing results to the same genotyped SNPs (when available) in the same individuals in the HapMap (NCBI build 36, dbSNP b126). 99.6% concordance was found between our resequencing and HapMap genotyping (discordant genotypes are listed in Supp. File S2). In addition, we incorporated into our analyses an additional 35 CEU and 36 YRI unrelated individuals of the same panel 2 for whom HapMap genotype data was available. Linkage disequilibrium (LD) analysis was conducted on all SNPs for each gene in each population (CEU and YRI) separately by using Haploview v.4.2 (Smith, 2008). The parameters for the selection of tagging SNPs (tSNPs) were a minor allele frequency (MAF) of ≥ 0.05 and pairwise r^2 of ≥ 0.8.

Comparison with 1000 Genomes Project

To examine the overlap between our data (resequencing and HapMap) and the 1000 Genomes project, we downloaded the 1000 Genomes Project full data for all populations from the consortium's website (http://www.1000genomes.org/announcements/july-2010-data-release-2010-07-20). We ran our data through the UCSC LiftOver (http://genome.ucsc.edu/cgi-bin/hgLiftOver) algorithm to convert our genotype data to build 37 coordinates (used by the 1000 Genomes Project data). We also compared the MAF for each population (CEU and YRI) and data (resequencing and HapMap) with the 1000 Genomes Project. To do so, we used Concordance Correlation Coefficient (CCC) (Lin, 1989) with epiR package (Stevenson et al., 2009), and we obtained the number and percentage of variants exclusive to our data.

mRNA expression data

mRNA expression results were available from the LCLs of 23 CEU and 23 YRI individuals (expression data were not available for YRI individual NA18871) resequenced in this study and from an additional unrelated 35 CEU and 36 YRI HapMap individuals from the same panel 2 (as described in the previous paragraph). Among the VEGF-pathway genes, *NRP1*, *FLT1*, *ITGB5* and *PKRCA* were not interrogated for eQTLs because of their low level of expression (Huang et al., 2007). *KDR* is also expressed at low levels in LCLs, and we have previously characterized the molecular and functional genetics of *KDR* (Glubb et al., 2011). Briefly, LCL mRNA expression data was generated using the Affymetrix GeneChip® Human Exon 1.0 ST Array (Huang et al., 2007). The resulting probe signal intensities were quantile normalized of all transcript clusters (gene-level) expression. Probeset-level expression signals were summarized with the robust multi-array average method (Irizarry et al., 2003). A transcript cluster was considered reliably expressed in LCLs if the $\log_2$-

transformed mRNA expression signals were greater than 5.34 (Huang et al., 2007; Duan et al., 2008). Each transcript cluster includes a set of probesets (exon level) containing all known exons and 5'- and 3'- untranslated regions (UTRs) in the genomes. The probes that hybridize to regions containing SNPs were excluded from the expression analyses (Duan et al., 2008). The Mann Whitney test was used to identify genes differentially expressed between CEU and YRI samples ($p<0.05$).

eQTL analysis

The primary analysis is a cis-based association between each SNP and mRNA levels. Because 11 of the 24 genes significantly differed in their expression levels between CEU and YRI (Supp. Figure S1), the eQTL analysis was conducted for each gene in the two ethnic groups separately. For the eQTL analysis we aimed at increasing our power to detect eQTLs by using both the SNP data from our resequencing (23 CEU and 23 YRI – one YRI sample did not have expression data available) and SNP data in all the available unrelated CEU and YRI individuals (58 CEU and 59 YRI, including the resequenced samples) from the HapMap 2 panel, as some SNPs could be used from the HapMap genotype data even if these loci were not fully covered by the resequencing. Hence, in each ethnic group, three sets of SNPs were used for the eQTL analysis: 1) SNPs found only in the resequenced samples (23 CEU and 23 YRI) and not in the genotyped HapMap samples (58 CEU and 59 YRI), 2) SNPs found in both the genotyped HapMap samples (58 CEU and 59 YRI) and the resequenced samples, and 3) SNPs found only in the HapMap samples (58 CEU and 59 YRI) and not in the sequenced samples, most likely because of the incomplete coverage of the resequencing and gaps in the sequences obtained through resequencing. A quantitative trait association test to determine cis-eQTLs was carried out in R (Purcell et al., 2007), where p-values are calculated using an asymptotic Wald test. Only SNPs with a MAF cutoff of 0.05 in one population were

tested under three genetic models (additive, recessive and dominant). Only models with a minimum of 5 genotypes per group available were considered. For the association between SNPs and gene expression, we used a p-value threshold of 0.05 as a selection feature for guiding the prioritization of downstream analyses aiming at determining functionality through bioinformatics and *in vitro* functional assays. The differences in allele frequencies across populations were measured using F_{ST} (Wright's fixation index).

To identify eQTLs independently associated with gene expression, eQTLs were jointly tested using a multivariate approach as follows: for each gene, those eQTLs showing a $p<0.05$ were included in an automatic selection algorithm taking the null model (no eQTLs included) as the starting point; at each step, the inclusion of each of the eQTLs not yet in the model were evaluated using a F test, and that eQTL showing the minimum p-value was included; this process was repeated until no eQTL showed a F test $p<0.05$. When two or more eQTL SNPs were in LD, one of them was randomly selected for this analysis while the others were excluded. For each gene, only samples without missing genotypes for all SNPs used in the analysis were taken for model fitting in the selection algorithm. In some cases this caused a significant reduction in sample size that could eventually produce over fitting. To avoid this, once the final model was specified, it was refitted using all the samples available for only the SNPs included in it. Only those eQTLs retaining significance in the later model were declared as independent eQTLs. The proportion of overall variance explained by the linear model was computed in each case.

Bioinformatic analyses of eQTL SNPs

The pfSNP browser (http://pfs.nus.edu.sg) was used to evaluate the potential functionality of the eQTL SNPs. pfSNP integrates >40 different algorithms/resources to evaluate the potential functionality of SNPs based on previous published reports, inferred potential functionality

from genetics approaches, as well as predicted potential functionality based on sequence motifs (Wang et al., 2011). This information was supplemented by bioinformatic prediction of microRNA (miRNA) binding sites using the web-based MirSNP tool (http://cmbi.bjmu.edu.cn/mirsnp) (Liu et al., 2012).

Reporter gene analysis of SNP function

As a proof of concept of the overall strategy of eQTL selection and bioinformatics prediction of their functional effects, we identified five SNPs in the 3'UTR of three genes (*FRS2, KRAS* and *GRB2*) for *in vitro* testing. Four of these were cis-eQTL SNPs (rs512283:A>T, rs542403:A>G, rs1137282:A>G and rs1137188:C>T) predicted to be functional by both pfSNP and MirSNP. rs7219:A>G in *GRB2* was also selected because it was predicted to have functional effects according to pfSNP and MirSNP, and displays perfect LD with rs8079197:G>C, an intronic cis-eQTL in CEU (Supp. Figure S2).

DNA regions spanning SNP loci of interest were amplified using primers with engineered restriction enzyme sites (Supp. Table S2). PCR fragments and reporter gene vectors were digested with the corresponding restriction enzymes (New England Biolabs, Ipswich, MA). A 1403 bp region spanning rs512283:A>T and rs542403:A>G (3'UTR of *FRS2*) was amplified and inserted into pmirGLO plasmid (Promega, Madison, WI, USA) using a SalI site in the reverse primer, an XbaI site 184 bp downstream of the forward primer and the corresponding restriction sites in pmirGLO (downstream of the luciferase reporter gene). A 940 bp region spanning rs1137188:C>T (3'UTR of *KRAS*) and a 902 bp region spanning rs1137282:A>G (3'UTR of *KRAS*) were amplified and inserted downstream of the luciferase reporter gene in pMIR-REPORT plasmid (Applied Biosystems) using SacI and HindIII sites for the rs1137188:C>T fragment and SacI and PmeI for the rs1137282:A>G fragment. A 383 bp region spanning rs7219:A>G (3'UTR of *GRB2*) was amplified and

inserted downstream of the luciferase reporter gene in pmirGLO vector using NheI and XhoI sites. Clones were screened by restriction digestion and positive clones were verified by DNA sequencing using the Sanger method. Mutagenesis to create the variant alleles of each SNP was performed using the QuikChange II Site Directed Mutagenesis Kit (Stratagene, La Jolla, CA) according to the manufacturer's protocol.

For the reporter gene assays, HEK-293 cells were obtained from ATCC and cultured with DMEM/F12 50/50 supplemented with 10 % FBS in 5% CO2 incubator at 37°C. HEK-293 were seeded in 24 well plates and cultured to a confluency of ~80%. Cells were transfected with pmirGLO vector constructs or co-transfected with either pMIR-REPORT and the control TK Renilla plasmid (Promega), in triplicate, using Lipofectamine 2000 reagent (Invitrogen, Carlsbad, CA) as per the manufacturer's instructions. After 24 h, cells were washed twice with PBS and lysed with passive lysis buffer. Dual report luciferase assays were performed using the Promega protocol. The luminescence was measured and the Firefly activity was normalized to Renilla. Differences in mean expression were analysed using paired t-tests in GraphPad Prism software (La Jolla, CA).

RESULTS

Gene resequencing

Resequencing encompassed 690,764 bp of DNA of the 23 VEGF-pathway genes (Table 1). Among the regions sequenced, 3,558 genetic variants were identified (Supp. File S3), of which 449 were not found in the dbSNP database build v130. Data have been submitted to the dbSNP database (http://www.ncbi.nlm.nih.gov/SNP/snp_viewTable.cgi?handle=RSG_UW). Among the variants, 693 and 1,708 were uniquely identified in CEU and YRI samples, respectively, and 1,157 variants were found in both populations (Table 2). At a MAF>0.05, 354 and 684

variants were unique to CEU and YRI, respectively, and 924 variants were found in both populations (Table 2). Haplotype-tagging SNPs (tSNPs) were identified in the 24 genes (including *VEGFA*) (Supp. Table S3) at r^2=0.8 and MAF>0.05 thresholds using Haploview 4.2 (Supp. Figure S2).

Comparison of our data with the 1000 Genomes Project showed that 80% of SNPs we identified were also found in the 1000 Genomes Project. For 5% of the remaining SNPs, the chromosomal positions matched those in the 1000 Genomes Project, but the rsnumber SNP identifier has been merged to a new one. The remaining 15% of the SNPs were not found in the 1000 Genomes Project either through matching chromosomal position or SNP identifier. We took all the SNPs in concordance with the 1000 Genomes Project (85%) in the two populations (CEU and YRI) and compared SNP allele frequencies between the 1000 Genomes Project and our two sets of data (resequencing and HapMap) independently. The concordance between the HapMap data and 1000 Genomes Project data (Supp. Figure S3A and B, CCC in CEU: 0.966 and CCC in YRI: 0.954) was higher than that between our resequencing and the 1000 Genomes Project data (Supp. Figure S3C and D, CCC in CEU: 0.923 and CCC in YRI: 0.942).

eQTL and bioinformatics analyses

To identify cis-eQTLs, associations were sought between genotypes of common variants (MAF>0.05) and LCL mRNA levels from each corresponding VEGF-pathway gene. In total, 356 cis-eQTLs were identified in CEU and YRI (Supp. File S4). 155 and 174 cis-eQTLs were detected uniquely in CEU and YRI samples, respectively, and 27 cis-eQTLs were found in both populations (Table 2). Nine (5%) and 20 (10%) of the eQTLs in CEU and YRI, respectively, are located in the 5'UTR region (Supp Table S4). Furthermore, of the 19 genes with eQTLs identified in this study, 15 share no eQTLs between the two sample groups. We

assessed whether the cis-eQTLs unique to CEU and YRI populations, population-specific cis-eQTLs, could be attributed to differential allele frequencies across populations. For the 155 cis-eQTLs identified only in CEU, 24 were rare variants in YRI. Of the remaining 131 cis-eQTLs, 18 CEU cis-eQTLs showed significant differences in allele frequencies (F_{ST} > 0.25) and the allelic effect was in the same direction across both populations for 9. For the 174 cis-eQTLs identified only in YRI, 30 were rare variants in CEU population. The remaining 144 YRI cis-eQTLs showed significant differences in allele frequencies (F_{ST} > 0.25) and the allelic effect was in the same direction across both populations for 2. For the 27 cis-eQTLs found in both populations, only 1 eQTL showed significant differences in allele frequencies which could be indicative of DNA regulatory differences across populations ($p<0.05$) (F_{ST} > 0.25, Supp Table S5) (Supp. Figure S4).

To identify variants with putative functional effects, bioinformatic analyses were performed. Out of all the SNPs identified, 99 are coding, 28 of which are non-synonymous (Supp. Table S6). Twelve of the non-synonymous SNPs were predicted to have possibly damaging effects on protein function either by PolyPhen or SIFT (Supp. Table S6). eQTL SNPs were examined for regulatory function using bioinformatic analyses and 46 SNPs were predicted to have functional effects (Supp. Table S7), of which five were synonymous, 12 were in 3'UTRs, and 26 were intronic.

Luciferase reporter gene assays

eQTL SNPs are more likely to be found near gene start and end sites. As the 3'UTR is a terminal gene region amenable to reporter gene analysis and SNPs in this region can mediate effects on expression through miRNA and transcription factor binding, mRNA splicing and other regulatory mechanisms, we selected five candidate 3'UTR SNPs with predicted functionality (Table 3) from three genes (*FRS2*, *GRB2* and *KRAS*) for eQTL validation. Of

the five candidate SNPs, four showed significant effects ($p<0.05$) on luciferase reporter activity. The minor alleles of rs1137188:C>T and rs1137282:A>G in *KRAS* significantly reduced and increased luciferase activity (Table 3), respectively, consistent with the YRI eQTL findings for these SNPs (Table 3). The minor allele of rs7219:A>G in *GRB2* had no significant effect on luciferase activity (Table 3). rs7219:A>G shares perfect LD with rs8079197:G>C in CEU, which was identified as a *GRB2* CEU eQTL, and impacts the binding sites of three miRNAs, but did not result in a change in luciferase activity (Table 3). The minor alleles of rs512283:A>T and rs542403:A>G in *FRS2* decreased luciferase activity by 11 and 29%, respectively (Table 3). These results were inconsistent with the eQTL findings as the minor alleles of rs512283:A>T and rs542403:A>G were significantly associated with increased *FRS2* expression in CEU (Table 3).

Multivariate analysis of cis-eQTLs

The multivariate analysis showed that three genes had independent cis-eQTLs significantly contributing to the variance in gene expression. In CEU, rs7559522:G>A and rs2053797:C>T (LD r^2=0.027) showed a statistically independent association with *PRKCE* expression and accounted for 44% of its variance (p 0.0008). In YRI, rs7478986:T>C and rs11024158:G>A (LD r^2=0.04) accounted for 45% of *PIK3C2A* expression variance (p 0.0017), while rs16966894:A>G and rs8067307:C>A (LD r^2=0.02) in YRI accounted for 37% of *MAP2K6* expression variance (p 0.0006; Supp. Table S8 and Supp. Figure S2).

VEGF-pathway eQTLs in clinical GWAS

To identify eQTLs that might explain GWAS findings, eQTL SNPs were interrogated using the PheGenI web-based tool (http://www.ncbi.nlm.nih.gov/gap/PheGenI) which searches the NHGRI GWAS catalog data. Two eQTLs significantly associate with clinical phenotypes in

GWAS (Table 4). The variant allele of rs2283792:T>G, a *MAPK1* intronic SNP, associated with decreased expression in LCLs and decreased risk of multiple sclerosis (Sawcer et al., 2011). The variant allele of rs4356203:A>G, an intronic *PIK3C2A* SNP, associated with increased expression in LCLs and increased risk of schizophrenia (Ripke et al., 2011).

DISCUSSION

To interrogate human genome variation and understand its biological and clinical implications can be an overwhelming, daunting task. Human genome variation can be queried through a myriad of repositories, databases and other portals available in the public domain. The collected data pertain to descriptive population allele data, functional annotation, multiple clinical outcomes and endophenotypes, and other layers of information. An integrated and focused analysis from these data is not always possible, and the understanding of the functional and clinical consequences of variants of genes in any given pathway relies on a variety of different experimental approaches.

Because of the central role of angiogenesis in physiology, pathophysiology, and drug treatment, we undertook a resequencing and functional validation project on 23 genes of the VEGF pathway. The 1000 Genomes Project is developing an extensive record of human genetic variation by sequencing the genomes of about 2,000 individuals from LCLs collected from several different populations (www.1000genomes.org). However, data from phase 1 of this study has low coverage at ~2.7-4.6x (Zhang and Dolan, 2010). Thus, resequencing studies are still crucial, not only for the discovery of unreported variants but also for determining their frequency in human populations and reliably assigning genotypes to each sequenced individual. Hence, resequencing of candidate genes coupled with eQTL analyses provides the opportunity to propel the discovery of SNP-phenotype associations. Because of

these premises, we have performed an extensive and focused molecular genetic study of the VEGF-pathway genes.

To our knowledge, this is the first analysis on the genetic variation for rare and common variants in CEU and YRI populations in VEGF-pathway genes which combines SNP discovery with an integrated functional annotation of SNPs using eQTL data, bioinformatics and *in vitro* experiments. Although eQTL data from LCLs have been used as the functional modules to interpret GWAS outcome data and in several other applications, we used the eQTL information derived from the VEGF-pathway genes as a screening tool to prioritize SNPs for downstream analysis of functionality. We realize the limited power to detect eQTLs using the small sample size of resequenced and (HapMap) genotyped samples of this study, as well as the intrinsic limitations of inferring vascular biology through data generated from cells of a different histology, such as the LCLs. Nevertheless, concordance in the effect of LCL eQTLs across tissues that are histologically diverse has been reported in several studies (Dimas et al., 2009; Nica et al., 2011), suggesting eQTLs in LCLs might reflect the basic function of a variant on gene expression. For some genes, we demonstrate that the amount of variation explained by a few eQTLs can be more than 35%, as shown for *PRKCE*, *MAP2K6*, and *PIK3C2A* (Supp. Table S7).

The LCL eQTL information and the bioinformatics inference of functionality should be used to prioritize SNPs for testing. These approaches cannot replace *in vitro* validation of SNP mechanisms at the bench. This has been clearly demonstrated for non-synonymous SNPs (an even simpler proposition compared to regulatory variants with smaller effects), for which prediction of the functional effect can be problematic (Carr et al., 2009). The lack of experimental data to train prediction algorithms for non-coding SNP function and also the context dependence (i.e. tissue, cell type, etc.) of their effects (Pang et al., 2009) contribute to the modest correlation between predicted effects and results from *in vitro* studies. We tested

the concordance between eQTL results and functional *in vitro* effects for a limited number of 3'UTR SNPs in three genes. Out of five SNPs, only two showed concordant effects, and they were located in the 3'UTR of *KRAS*, a key oncogene of the VEGF pathway (Kranenburg, 2005). Interestingly, these two variants are not in LD with rs61764370:T>G in the binding site for the let-7 miRNA, which has been associated with a series of clinical phenotypes related to cancer risk and response to EGFR inhibitors and other cancer drugs (Chin et al., 2008; Graziano et al., 2010; Paranjape et al., 2011; Smits et al., 2011; Ratner et al., 2012; Sebio et al., 2013). These results propose additional *KRAS* variants for testing. Due to the lack of efficient high throughput molecular screens for SNP functionality (Chorley et al., 2008; Glubb and Innocenti, 2011), we propose that only a joint analysis of eQTL, bioinformatics, and *in vitro* confirmation of candidate SNPs could be reliable enough to dictate further molecular and clinical testing of SNPs.

Functional annotation can provide the mechanistic link between the SNP and its clinical association. Through our study, we can provide insights on the clinical relevance of genetic variation in *MAPK1* and *PIK3C2A*. For example, a *MAPK1* SNP, rs2283792:T>G, decreases the risk of multiple sclerosis (Sawcer et al., 2011). In our study, we found this same SNP to be an eQTL of *MAPK1*. A reduction in *MAPK1* expression may provide protection from multiple sclerosis by reducing the activation of Th17 cells (Noubade et al., 2011). Thus, the association of rs2283792:T>G with multiple sclerosis may be explained by our finding that the variant allele of rs2283792:T>G correlates with lower *MAPK1* expression. In the GWAS for multiple sclerosis, no functional analysis of this SNP was provided (Sawcer et al., 2011). For *PIK3C2A*, factors (including SNPs in the gene) that increase gene expression might result in an increased risk of schizophrenia. *PIK3C2A* activates neurosecretory granule exocytosis (Meunier et al., 2005) and increases neuroexcytosis models of schizophrenia in mice (Uriguen et al., 2013), while downregulation of neuroexcoytotic proteins is associated

with antipsychotic drug treatment (Gil-Pisa et al., 2012). Consistent with these reports, we found rs4356203:A>G to associate with increased *PIK3C2A* expression which may lead to greater neuroexocytosis, potentially increasing schizophrenia risk. This is consistent with the finding of a recent GWAS (Ripke et al., 2011) where this SNP was found to increase the risk of schizophrenia, but for which no molecular mechanism was provided.

With specific reference to the VEGF-pathway in oncology, there are large knowledge gaps regarding the effects of these SNPs on molecular and cellular phenotypes of tumor angiogenesis and, in this setting, the interpretation and validation of associations of VEGF-pathway SNPs and patient outcomes is challenging (Schneider et al., 2012; Lambrechts et al., 2013). Associations of VEGF-pathway SNPs in clinical cancer studies tend to fail to replicate if they do not have a mechanistic basis. Compounding these issues is the absence of cellular model systems which allow the interrogation and isolation of the individual effects of SNPs on angiogenic phenotypes (Freedman et al., 2011). In an N-analysis of oncology studies of *VEGFA* SNPs, rs2010963:C>G was found to be significantly associated with a 26% relative improvement in OS across a variety of tumor types, suggesting it could be a putative biomarker (Eng et al., 2012). Genotype-phenotype analyses suggest that the minor allele of rs2010963:C>G is associated with decreased promoter activity and decreased *VEGFA* mRNA expression (Watson et al., 2000; Young et al., 2004; Hussein et al., 2010). Therefore rs2010963:C>G may confer a survival benefit by reducing tumor angiogenesis. rs2010963:C>G has a high degree of LD with rs833068:G>A (r^2=0.96), an eQTL identified in our study of LCLs associated with decreased *VEGFA* expression. This analysis further highlights how eQTL data can be informative of the molecular effects of putative biomarkers.

With this study, we provide a framework for functional inference of common SNPs. Having established methods to "zoom in" variation in any given gene pathway will improve our biological understanding of genome variation, a major roadblock for clinical translation

of the genome data (Green and Guyer, 2011). Leveraging existing eQTL data with the support of functional *in vitro* validation provides essential information to identify genetic candidates for association studies.

ACKNOWLEDGMENTS

The authors would like to acknowledge Sonal Kashyap for her contributions to the in vitro assays, and Jessie Bishop and Anna Sorin for their kind assistance in editing and formatting this paper. The authors have no conflict of interest to disclose.

REFERENCES

Bernatchez PN, Rollin S, Soker S and Sirois MG. 2002. Relative effects of VEGF-A and VEGF-C on endothelial cell proliferation, migration and PAF synthesis: Role of neuropilin-1. J Cell Biochem 85: 629-639.

Carr DF, Whiteley G, Alfirevic A and Pirmohamed M. 2009. Investigation of inter-individual variability of the one-carbon folate pathway: a bioinformatic and genetic review. Pharmacogenomics J. 9: 291-305. doi: 210.1038/tpj.2009.1029.

Chin LJ, Ratner E, Leng S, Zhai R, Nallur S, Babar I, Muller RU, Straka E, Su L, Burki EA, Crowell RE, Patel R, et al. 2008. A SNP in a let-7 microRNA complementary site in the KRAS 3' untranslated region increases non-small cell lung cancer risk. Cancer Res 68: 8535-8540. doi: 8510.1158/0008-5472.CAN-8508-2129.

Chorley BN, Wang X, Campbell MR, Pittman GS, Noureddine MA and Bell DA. 2008. Discovery and verification of functional single nucleotide polymorphisms in regulatory genomic regions: current and developing technologies. Mutat Res 659: 147-157. doi: 110.1016/j.mrrev.2008.1005.1001.

Dimas AS, Deutsch S, Stranger BE, Montgomery SB, Borel C, Attar-Cohen H, Ingle C, Beazley C, Arcelus MG, Sekowska M, Gagnebin M, Nisbett J, et al. 2009. Common Regulatory Variation Impacts Gene Expression in a Cell Type–Dependent Manner. Science 325: 1246-1250.

Duan S, Huang RS, Zhang W, Bleibel WK, Roe CA, Clark TA, Chen TX, Schweitzer AC, Blume JE, Cox NJ and Dolan ME. 2008. Genetic architecture of transcript-level variation in humans. Am J Hum Genet 82: 1101-1113.

Duan S, Zhang W, Bleibel WK, Cox NJ and Dolan ME. 2008. SNPinProbe_1.0: a database for filtering out probes in the Affymetrix GeneChip human exon 1.0 ST array potentially affected by SNPs. Bioinformation 2: 469-470.

Eng L, Azad AK, Habbous S, Pang V, Xu W, Maitland-van der Zee AH, Savas S, Mackay HJ, Amir E and Liu G. 2012. Vascular endothelial growth factor pathway polymorphisms as prognostic and pharmacogenetic factors in cancer: a systematic review and meta-analysis. Clin Cancer Res 18: 4526-4537.

Folkman J. 2007. Angiogenesis: an organizing principle for drug discovery? Nat Rev Drug Discov 6: 273-286.

Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, Pasternak S, Wheeler DA, et al. 2007. A second generation human haplotype map of over 3.1 million SNPs. Nature 449: 851-861.

Freedman ML, Monteiro AN, Gayther SA, Coetzee GA, Risch A, Plass C, Casey G, De Biasi M, Carlson C, Duggan D, James M, Liu P, et al. 2011. Principles for the post-GWAS functional characterization of cancer risk loci. Nat Genet 43: 513-518. doi: 510.1038/ng.1840.

Gil-Pisa I, Munarriz-Cuezva E, Ramos-Miguel A, Urigüen L, Meana JJ and García-Sevilla JA. 2012. Regulation of munc18-1 and syntaxin-1A interactive partners in schizophrenia prefrontal cortex: down-regulation of munc18-1a isoform and 75 kDa SNARE complex after antipsychotic treatment. Int J Neuropsychopharmacol 15: 573-588.

Glubb DM, Cerri E, Giese A, Zhang W, Mirza O, Thompson EE, Chen P, Das S, Jassem J, Rzyman W, Lingen MW, Salgia R, et al. 2011. Novel functional germline variants in the VEGF receptor 2 gene and their effect on gene expression and microvessel density in lung cancer. Clin Cancer Res 17: 5257-5267.

Glubb DM and Innocenti F. 2011. Mechanisms of genetic regulation in gene expression: examples from drug metabolizing enzymes and transporters. Wiley Interdiscip Rev Syst Biol Med. 3: 299-313. doi: 210.1002/wsbm.1125.

Gordon D, Abajian C and Green P. 1998. Consed: a graphical tool for sequence finishing. Genome Res 8: 195-202.

Graziano F, Canestrari E, Loupakis F, Ruzzo A, Galluccio N, Santini D, Rocchi M, Vincenzi B, Salvatore L, Cremolini C, Spoto C, Catalano V, et al. 2010. Genetic modulation of the Let-7 microRNA binding to KRAS 3'-untranslated region and survival of metastatic colorectal cancer patients treated with salvage cetuximab-irinotecan. Pharmacogenomics J. 10: 458-464. doi: 410.1038/tpj.2010.1039.

Green ED and Guyer MS. 2011. Charting a course for genomic medicine from base pairs to bedside. Nature. 470: 204-213. doi: 210.1038/nature09764.

Hicklin DJ and Ellis LM. 2005. Role of the vascular endothelial growth factor pathway in tumor growth and angiogenesis. J Clin Oncol 23: 1011-1027.

Hicklin DJ and Ellis LM. 2005. Role of the vascular endothelial growth factor pathway in tumor growth and angiogenesis. J Clin Oncol 23: 1011-1027.

Huang RS, Duan S, Shukla SJ, Kistner EO, Clark TA, Chen TX, Schweitzer AC, Blume JE and Dolan ME. 2007. Identification of genetic variants contributing to cisplatin-induced cytotoxicity by use of a genomewide approach. Am J Hum Genet 81: 427-437.

Hussein A, Askar E, Elsaeid M and Schaefer F. 2010. Functional polymorphisms in transforming growth factor-beta-1 (TGFbeta-1) and vascular endothelial growth factor (VEGF) genes modify risk of renal parenchymal scarring following childhood urinary tract infection. Nephrol Dial Transplant. 25: 779-785. doi: 710.1093/ndt/gfp1532.

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

Kranenburg O. 2005. The KRAS oncogene: past, present, and future. Biochim Biophys Acta. 1756: 81-82.

Lambrechts D, Lenz HJ, de Haas S, Carmeliet P and Scherer SJ. 2013. Markers of response for the antiangiogenic agent bevacizumab. J Clin Oncol 31: 1219-1230. doi: 1210.1200/JCO.2012.1246.2762.

Lin LI. 1989. A concordance correlation coefficient to evaluate reproducibility. Biometrics. 45: 255-268.

Liu C, Zhang F, Li T, Lu M, Wang L, Yue W and Zhang D. 2012. MirSNP, a database of polymorphisms altering miRNA target sites, identifies miRNA-related SNPs in GWAS SNPs and eQTLs. BMC Genomics 13: 661.

Maitland ML, Lou XJ, Ramirez J, Desai AA, McLeod HL, Weichselbaum RR and Ratain MJ. 2010. Vascular endothelial growth factor pathway. Pharmacogenet and genomics 20: 346-349.

Meunier FA, Osborne SL, Hammond GR, Cooke FT, Parker PJ, Domin J and Schiavo G. 2005. Phosphatidylinositol 3-kinase C2alpha is essential for ATP-dependent priming of neurosecretory granule exocytosis. Mol Biol Cell 16: 4841-4851.

Nica AC, Parts L, Glass D, Nisbet J, Barrett A, Sekowska M, Travers M, Potter S, Grundberg E, Small K, Hedman ÅK, Bataille V, et al. 2011. The Architecture of Gene Regulatory Variation across Multiple Human Tissues: The MuTHER Study. PLoS Genet 7: e1002003.

Noubade R, Krementsov DN, Del Rio R, Thornton T, Nagaleekar V, Saligrama N, Spitzack A, Spach K, Sabio G, Davis RJ, Rincon M and Teuscher C. 2011. Activation of p38 MAPK in CD4 T cells controls IL-17 production and autoimmune encephalomyelitis. Blood 118: 3290-3300. doi: 3210.1182/blood-2011-3202-336552.

Olsson AK, Dimberg A, Kreuger J and Claesson-Welsh L. 2006. VEGF receptor signalling - in control of vascular function. Nat Rev Mol Cell Biol 7: 359-371.

Pang GSY, Wang J, Wang Z and Lee CGL. 2009. Predicting potentially functional SNPs in drug-response genes. Pharmacogenomics 10: 639-653.

Paranjape T, Heneghan H, Lindner R, Keane FK, Hoffman A, Hollestelle A, Dorairaj J, Geyda K, Pelletier C, Nallur S, Martens JW, Hooning MJ, et al. 2011. A 3'-untranslated region KRAS variant and triple-negative breast cancer: a case-control and genetic analysis. Lancet Oncol 12: 377-386. doi: 310.1016/S1470-2045(1011)70044-70044.

PharmGKB. (2013). "VEGF Signaling Pathway." from http://www.pharmgkb.org/do/serve?objId=PA2032&objCls=Pathway#tabview=tab1&subtab=.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ and Sham PC. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81: 559-575.

Ratner ES, Keane FK, Lindner R, Tassi RA, Paranjape T, Glasgow M, Nallur S, Deng Y, Lu L, Steele L, Sand S, Muller RU, et al. 2012. A KRAS variant is a biomarker of poor outcome, platinum chemotherapy resistance and a potential target for therapy in ovarian cancer. Oncogene. 31: 4559-4566. doi: 4510.1038/onc.2011.4539.

Ripke S, Sanders AR, Kendler KS, Levinson DF, Sklar P, Holmans PA, Lin DY, Duan J, Ophoff RA, Andreassen OA, Scolnick E, Cichon S, et al. 2011. Genome-wide association study identifies five new schizophrenia loci. Nat Genet 43: 969-976.

Rogers MS and D'Amato RJ. 2012. Common Polymorphisms in Angiogenesis. Cold Spring Harbor Perspect Med. 2(11). pii: a006510. doi: 10.1101/cshperspect.a006510.

Sawcer S, Hellenthal G, Pirinen M, Spencer CC, Patsopoulos NA, Moutsianas L, Dilthey A, Su Z, Freeman C, Hunt SE, Edkins S, Gray E, et al. 2011. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. Nature 476: 214-219.

Sawcer S, Hellenthal G, Pirinen M, Spencer CC, Patsopoulos NA, Moutsianas L, Dilthey A, Su Z, Freeman C, Hunt SE, Edkins S, Gray E, et al. 2011. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. Nature. 476: 214-219. doi: 210.1038/nature10251.

Schneider BP, Shen F and Miller KD. 2012. Pharmacogenetic biomarkers for the prediction of response to antiangiogenic treatment. Lancet Oncol 13: e427-436. doi: 410.1016/S1470-2045(1012)70275-70279.

Schneider BP, Shen F and Miller KD. 2012. Pharmacogenetic biomarkers for the prediction of response to antiangiogenic treatment. Lancet Oncol 13: e427-436.

Sebio A, Pare L, Paez D, Salazar J, Gonzalez A, Sala N, del Rio E, Martin-Richard M, Tobena M, Barnadas A and Baiget M. 2013. The LCS6 polymorphism in the binding site of let-7 microRNA to the KRAS 3'-untranslated region: its role in the efficacy of anti-EGFR-based therapy in metastatic colorectal cancer patients. Pharmacogenet Genomics 23: 142-147. doi: 110.1097/FPC.1090b1013e32835d32839b32830b.

Smith AV. 2008. Manipulating HapMap Data Using HaploView. CSH Protoc 2008: pdb prot5025.

Smits KM, Paranjape T, Nallur S, Wouters KA, Weijenberg MP, Schouten LJ, van den Brandt PA, Bosman FT, Weidhaas JB and van Engeland M. 2011. A let-7 microRNA SNP in the KRAS 3'UTR is prognostic in early-stage colorectal cancer. Clin Cancer Res 17: 7723-7731. doi: 7710.1158/1078-0432.CCR-7711-0990.

Stephens M, Sloan JS, Robertson PD, Scheet P and Nickerson DA. 2006. Automating sequence-based detection and genotyping of SNPs from diploid samples. Nat Genet 38: 375-381.

Stevenson M, Nunes T, Sanchez J and Thornton R (2009). epiR: Functions for analysing epidemiological data.

Teicher BA. 2011. Antiangiogenic agents and targets: A perspective. Biochem Pharmacol 81: 6-12.

Uriguen L, Gil-Pisa I, Munarriz-Cuezva E, Berrocoso E, Pascau J, Soto-Montenegro ML, Gutierrez-Adan A, Pintado B, Madrigal JLM, Castro E, Sanchez-Blazquez P, Ortega JE, et al. 2013. Behavioral, neurochemical and morphological changes induced by the overexpression of munc18-1a in brain of mice: relevance to schizophrenia. Transl Psychiatry 3: e221.

Wang J, Ronaghi M, Chong SS and Lee CG. 2011. pfSNP: An integrated potentially functional SNP resource that facilitates hypotheses generation through knowledge syntheses. Hum Mutat 32: 19-24.

Watson CJ, Webb NJ, Bottomley MJ and Brenchley PE. 2000. Identification of polymorphisms within the vascular endothelial growth factor (VEGF) gene: correlation with variation in VEGF protein production. Cytokine 12: 1232-1235.

Young HS, Summers AM, Bhushan M, Brenchley PE and Griffiths CE. 2004. Single-nucleotide polymorphisms of vascular endothelial growth factor in psoriasis of early onset. J Invest Dermatol 122: 209-215.

Zhang W and Dolan ME. 2010. Impact of the 1000 genomes project on the next wave of pharmacogenomic discovery. Pharmacogenomics 11: 249-256.

Zhang W, Duan S, Bleibel WK, Wisel SA, Huang RS, Wu X, He L, Clark TA, Chen TX, Schweitzer AC, Blume JE, Dolan ME, et al. 2009. Identification of common genetic

variants that account for transcript isoform variation between human populations. Hum Genet 125: 81-93.

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

Figure 1: The VEGF pathway in the human endothelium, according to www.pharmgkb.org. Copyright PharmGKB (Maitland et al., 2010). Reprinted with permission of PharmGKB and Stanford University.

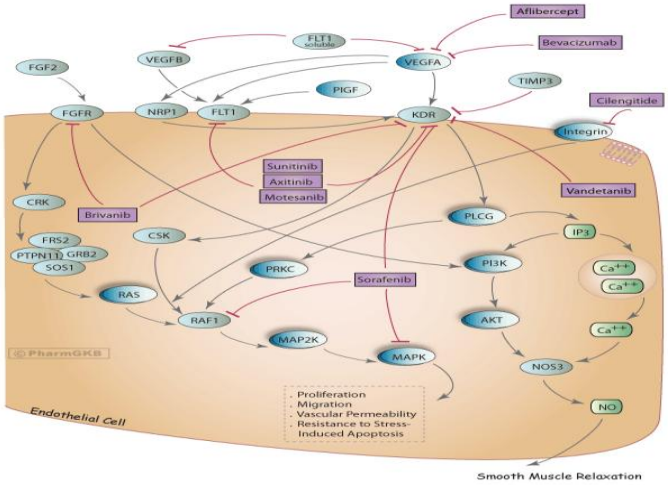

Table 1. VEGF-pathway genes, sequencing coverage and eQTLs

Gene symbol	OMIM accession	Coverage	Chromosome	Gene size (kb)	Sequencing coverage (kb)	CEU eQTLs	YRI eQTLs
AKT1	164730	Full	14q32.32-q32.33	32.2	26.1	0	5
CRK	164762	Standard	17p13	42.1	8.3	3	2
FLT1[*]	165070	Standard	13q12	200.9	26.2	-	-
FRS2	607743	Standard	12q15	117.4	38.8	4	19
GRB2	108355	Standard	17q24-q25	95.7	16.8	3	0
ITGAV	193210	Standard	2q31-q32	98.8	60.1	1	2
ITGB5[*]	147561	Standard	3q21.2	132.2	19.7	-	-
KRAS	190070	Full	12p12.1	53.7	49.1	1	25
MAP2K6	601254	Standard	17q24.3	135.6	17.3	9	10
MAPK1	176948	Standard	22q11.2	116.1	59.1	2	19
MAPK11	602898	Full	22q13.33	14.6	9.3	1	0
MAPK14	600289	Standard	6p21.3-p21.2	91.0	15.2	0	0
MAPK3	601795	Full	16p11.2	26.5	18.9	2	1
NRAS	164790	Full	1p13.2	13.9	11.9	1	1
NRP1[*]	602069	Standard	10p12	161.4	20.0	-	-
PGF	601121	Full	14q24.3	21.8	16.9	2	1
PIK3C2A	603601	Standard	11p15.5-p14	91.2	44.4	27	36
PIK3C2B	602838	Standard	1q32	75.7	45.2	14	21
PIK3R5	611317	Full	17p13.1	40.4	29.6	6	5
PRKCA[*]	176960	Standard	17q22-q24	515.9	35.0	-	-
PRKCE	176975	Standard	2p21	544.1	77.8	101	45
RAF1	164760	Standard	3p25	88.6	37.4	0	7
VEGFA[#]	192240	n/a	6p12	n/a	n/a	5	1
VEGFB	601398	Full	11q13	12.0	7.8	0	1

[*]Low level of mRNA expression in LCLs

[#]*VEGFA* was not resequenced but genotyping data from the HapMap (obtained from the same LCLs used for resequencing) were used for the eQTL analyses

Table 2. Genetic variants identified in CEU and YRI by resequencing

Population	Variants	MAF>0.05	Unique variants in each group		eQTLs
			All	*MAF >0.05*	
CEU	1,850	1,278	693	354	182
YRI	2,865	1,608	1,708	684	201
CEU+YRI	3,558	1,962	-	-	356

Table 3. Reporter gene and bioinformatic analysis of eQTLs

SNP	Gene	Effect of variant allele			
		eQTL effect	*Luciferase activity*	*MirSNP*	*pfSNP*
rs1137188	*KRAS*	↓ YRI	↓10% (p<0.001)	hsa-miR-511 site disrupted	Introduces CDX1, GATA1, GATA2 and GATA3 binding motifs
rs1137282	*KRAS*	↑ YRI	↑53% (p<0.01)	hsa-miR-2681-5p site weakened	Introduces an exon splicing enhancer motif and disrupts an exon splicing silencer motif
rs7219	*GRB2*	↑CEU (rs8079197)	↓3% (p=0.40)	hsa-miR-1288 and hsa-miR-5002-3p sites created; hsa-miR-511 site disrupted	Introduces EVX1 binding motif
rs512283	*FRS2*	↑ CEU	↓11% (p=0.04)	hsa-miR-2909 and hsa-miR-4436b-3p sites created; hsa-miR-29b-1-5p site disrupted	Introduces HNRNPC binding motif
rs542403	*FRS2*	↑ CEU	↓29% (p<0.0001)	No effect	Conserved 3'UTR motif and introduces RFX1 binding motif

For GRB2, rs7219 is in perfect LD with rs8079197, a CEU eQTL.

Table 4. Clinical GWAS correlates of eQTL SNPs

SNP	Gene	Minor allele association		pfSNP
		LCL gene expression	*Clinical phenotype*	
rs2283792	*MAPK1*	↓ YRI	Reduced multiple sclerosis risk	rs5749998 (LD with rs2283792, r^2=0.967): introduces an intronic splicing regulatory element
rs4356203	*PIK3C2A*	↑ CEU	Increased schizophrenia risk	rs214935 (LD with rs4356203, r^2=0.875): introduces a FOXJ2 binding site

NHGRI GWAS catalog data were interrogated for eQTL SNPs using PheGenI (http://www.ncbi.nlm.nih.gov/gap/PheGenI).