GRIK5 Genetically Regulated Expression Associated with Eye and Vascular Phenomes: Discovery through Iteration among Biobanks, Electronic Health Records, and Zebrafish

Gokhan Unlu,[1,2,3,9] Eric R. Gamazon,[1,2,4,5,9] Xinzi Qi,[1,2] Daniel S. Levic,[1,2,3,11] Lisa Bastarache,[2,6] Joshua C. Denny,[2,6] Dan M. Roden,[2,6,7] Ilya Mayzus,[8] Max Breyer,[1,2] Xue Zhong,[1,2] Anuar I. Konkashbaev,[1,2] Andrey Rzhetsky,[8] Ela W. Knapik,[1,2,3,10] and Nancy J. Cox[1,2,4,10,*]

Although the use of model systems for studying the mechanism of mutations that have a large effect is common, we highlight here the ways that zebrafish-model-system studies of a gene, *GRIK5*, that contributes to the polygenic liability to develop eye diseases have helped to illuminate a mechanism that implicates vascular biology in eye disease. A gene-expression prediction derived from a reference transcriptome panel applied to BioVU, a large electronic health record (EHR)-linked biobank at Vanderbilt University Medical Center, implicated reduced *GRIK5* expression in diverse eye diseases. We tested the function of *GRIK5* by depletion of its ortholog in zebrafish, and we observed reduced blood vessel numbers and integrity in the eye and increased vascular permeability. Analyses of EHRs in >2.6 million Vanderbilt subjects revealed significant comorbidity of eye and vascular diseases (relative risks 2–15); this comorbidity was confirmed in 150 million individuals from a large insurance claims dataset. Subsequent studies in >60,000 genotyped BioVU participants confirmed the association of reduced genetically predicted expression of *GRIK5* with comorbid vascular and eye diseases. Our studies pioneer an approach that allows a rapid iteration of the discovery of gene-phenotype relationships to the primary genetic mechanism contributing to the pathophysiology of human disease. Our findings also add dimension to the understanding of the biology driven by glutamate receptors such as GRIK5 (also referred to as GLUK5 in protein form) and to mechanisms contributing to human eye diseases.

Introduction

Eye diseases are a diverse set of conditions that are both common and largely consistent with a polygenic genetic architecture. Eye-disease pathophysiologies arise in different cell types and appear to be distinct; however, an understanding of the biological basis of eye diseases has been derived primarily from single-gene approaches or small-scale genetic studies. Unbiased, at-scale, big-data approaches that use electronic health records (EHRs) might generate disease-mechanism hypotheses that can be validated in animal models.

Gene-based tests, as opposed to tests of individual SNPs, offer an alternative approach for the identification of genes for common diseases. PrediXcan (for predicted expression scanning) uses a reference panel to create SNP-based prediction models for transcript expression for each gene in each tissue.[1] Testing for associations with predicted expression improves power relative to individual SNP associations because we are testing a specific feature (in this case imputed, genetically determined gene expression) of genome function; this feature has been shown to drive a substantial fraction of the common variant heritability to common diseases.[2,3] The use of predicted gene expression provides a clearer route to model-organism validation because the association result is at the level of the gene rather than the SNP (as is the case with genome-wide association study [GWAS] results) and has an easy-to-interpret direction of effect.

Recent advances in CRISPR-Cas9 technology have revolutionized the fidelity and speed of the generation of animal models and now allow for precise genome editing with high efficiency and minimal off-target effects in animals and cell lines.[4–7] The rapid development of transparent embryos and genetically modified adults facilitates the phenotypic analysis of variants in intact zebrafish over long periods of time, thereby enabling 4D imaging.

Studies over the last decade have shown that EHR data linked to DNA biobanks are an efficient route to human genetic discovery.[8–13] The use of phenotype algorithms in EHR data allows both the replication of existing genetic associations and the discovery of associations,[14] and it allows for both genome-wide and phenome-wide association studies with results that replicate those from studies conducted with research-quality diagnoses.[15] The broad spectrum of medical phenomes and the much larger sample sizes that are available (and affordable) through the use of EHR phenomics create opportunities for studies that

[1]Department of Medicine, Division of Genetic Medicine, Vanderbilt University Medical Center, Nashville, TN 37232, USA; [2]Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, TN 37232, USA; [3]Department of Cell and Developmental Biology, Vanderbilt University, Nashville, TN 37232, USA; [4]Data Science Institute, Vanderbilt University, Nashville, TN 37232, USA; [5]Clare Hall, University of Cambridge, Cambridge CB3 9AL, UK; [6]Departments of Medicine and Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN 37232, USA; [7]Department of Pharmacology, Vanderbilt University, Nashville, TN 37232, USA; [8]Departments of Medicine and Human Genetics, the University of Chicago, Chicago, IL 60637, USA
[9]These authors contributed equally to this work
[10]These authors contributed equally to this work
[11]Present address: Department of Cell Biology, Duke University Medical Center, Durham, NC 27710, USA
*Correspondence: nancy.j.cox@vanderbilt.edu
https://doi.org/10.1016/j.ajhg.2019.01.017.

would otherwise not be feasible and would lead logically to research questions and discoveries that cannot be considered in research datasets comprised of affected individuals and controls for a single disease.

We summarize here the results of a set of studies beginning with the discovery of the association between reduced genetically determined expression of *GRIK5*, which encodes the glutamate ionotropic receptor kainate-type subunit 5, and a diverse set of eye diseases in BioVU, the EHR-linked DNA biobank at Vanderbilt University Medical Center.[16] Using related, publicly available large-scale GWASs and validation studies in zebrafish, we show that the reduced *GRIK5* expression compromises vascular integrity in an animal model and increases genetic risk for diverse eye diseases in BioVU participants. Our study pioneers a model for discovery and rapid progression to an understanding of physiological mechanisms linking genome variation to human disease by iterating among methods, applied to biobanks and large-scale EHRs, for big-data integration (genome variation by transcriptome measurement) and modern approaches to gene editing in animal model systems.

Material and Methods

DNA Biobank and Electronic Health Records Database

BioVU, the Vanderbilt University biobank linked to EHRs, aggregates ICD-9 (International Classification of Diseases, ninth revision) codes that represent diseases and other phenotypes for use in clinical or genetic studies.[15,17] The phenotype codes ("phecodes") combine related ICD-9 codes to represent clinical traits. These phecodes, first implemented in 2010,[17] are supported by clinical co-occurrence data that identify related phenotypes across the ICD-9 code system. For our study, we used version 1.2, which includes 1,965 hierarchical phecodes and 20,203 ICD-9 codes.[15] The phecode hierarchy includes clinical traits, including "inflammatory bowel disease" as the parent node for "ulcerative colitis" and "Crohn's disease," that are not found in the ICD-9 system. For example, the phecode "other retinal disorders" includes ICD-9 codes for retinal exudates and deposits and retinal nerve fiber bundle defects, as well as other retinal disorders and retinal disorders NEC (not elsewhere classifiable). The phecode for other disorders of the eye includes ICD-9 codes for acute dacryoadenitis, other disorders of the sclera, and other disorders of the eye.

These phecodes have been employed in phenome-wide association studies (PheWASs) for validating known, trait-associated genetic variants, as well as for discovering associations, including those that have been subsequently validated.[15,17–19] However, these PheWASs have focused strictly on SNPs rather than on genes, which are natural, biologically relevant units for evaluating associations with disease and for proposing disease mechanisms. Here we applied PrediXcan[1] to perform gene-level PheWASs of a broad array of traits by using the unprecedented collection of human tissue data (n = 44) made available by the Genotype-Tissue Expression (GTEx) project.[20] We used the GTEx reference resource to impute tissue-level gene expression in the BioVU subjects via the PrediXcan method. We refer to the imputed gene expression level as the *genetically regulated expression* (GReX).[1]

The Synthetic Derivative (SD) is a de-identified and continuously updated EHR image that includes data on >2.6 million subjects.[16] We conducted analyses on data from the SD to test the phenome-level comorbidity between vascular traits and eye diseases. Included in these analyses were several vascular phenotypes and all of the original eye phenotypes that showed association to the GReX[1] for *GRIK5*; the eye phenotypes included retinal detachment and defects, other retinal disorders, cataract, senile cataract, glaucoma, primary open-angle glaucoma, open-angle glaucoma, and other disorders of the eye. To control for potential confounding due to age and sex, we calculated the adjusted relative risk (RR) by using logistic regression[21] with age and sex as covariates. In addition, we performed a stratified analysis, estimated the RR for each stratum (as defined by sex and age group), and pooled the resulting estimates into an adjusted RR. We performed replication of the RR estimates from a stratified (by age and sex) analysis of the MarketScan insurance claims dataset consisting of >150 million subjects. We used the Cochran-Mantel-Haenszel estimator, which takes into account the stratification (by age and sex):

$$\widehat{RR_{CMH}} = \frac{\sum a_i(c_i + d_i)/n_i}{\sum c_i(a_i + b_i)/n_i}$$

where the index i refers to each stratum, and $n_i = a_i + b_i + c_i + d_i$. For each i stratum, a, b, c, and d indexed by i are defined by the following two-by-two table:

		Disease 2	
		Case	Control
Disease 1	Case	a	b
	Control	c	d

PrediXcan Analysis

To evaluate the role a gene might play in the etiology of a disease trait, we utilized the PrediXcan approach.[1] We estimated the genetic component of gene expression in the BioVU samples[15] so we could identify genes associated with a phenotype (despite the lack of directly measured gene expression data). Specifically, from the weights $\widehat{\beta}_j$ derived from the gene expression model on the basis of local genetic variation (+/− 1 Mb of the gene)[1] and the number of effect alleles X_{ij} at the variant j, we estimated the genetically determined component of gene expression as follows:

$$\widehat{G}_i = \sum_j X_{ij}\widehat{\beta}_j$$

A significant association between the estimated genetic component of gene expression and a trait proposes a causal direction of effect (because the germline genetic profile is unlikely to be altered by the trait). We also evaluated whether a significant gene-level association may be "contaminated" by associations with nearby genes; such contamination might result from linkage disequilibrium (LD) between SNP predictors of gene expression and trait causal variants.[22] We note that, in general, the top SNP for a gene is not optimal for predicting the expression of the gene, and multi-SNP models tend to outperform single-SNP models.[1] For example, the number of SNPs in the optimal imputation model, generated from Depression Gene Network (DGN) whole-blood expression data, for *GRIK5* (Table S2) is 15.

eQTL-Based Heritability Estimation and Polygenic Modeling

Polygenic modeling is an approach aimed at relating phenotypic variation to multiple genetic variants simultaneously. As a method for genome-wide analysis, it differs from conventional single-variant tests of association by testing large numbers of loci (potentially in the thousands) for their contribution to the genetic architecture of a phenotype. We sought to estimate SNP-based heritability, defined as the proportion of the phenotypic variance captured by the genetic variants in aggregate, for a broad spectrum of complex human traits. We define the following mixed-effects model:

$$Y = Xb + \Sigma_T g_T + C + e,\ \mathrm{var}(Y) = \Sigma_T A_T \sigma_T^2 + A_C \sigma_C^2 + I\sigma_e^2$$

Here, Y is a phenotype vector, and b is a vector of fixed effects (e.g., principal components [representing ancestry], sex, and probabilistic estimation of expression residuals [PEER] factors [representing hidden or unmeasured variables][23]). A_T is the genetic relationship matrix (GRM) estimated from the SNP set T, and g_T denotes the polygenic component attributable to the set T with a mean of g_T equal to zero and variance equal to $A_T\sigma_T^2$. C is the aggregate genetic effect of the complement set of variants in the genome. The phenotype is modeled as the sum of these genetic effects (random effects), the relevant covariates (fixed effects), and a residual. In the case of a single tissue, g_T and C are assumed to have possibly different distributions of effect sizes. We estimated variances by using restricted maximum likelihood (REML). The heritability attributable to the SNP set T is then calculated as the fraction of the phenotypic variance σ_Y^2:

$$h_T^2 = \sigma_T^2/\sigma_Y^2$$

We utilized both BioVU data and publicly available (independent) GWAS data to test for the existence of a shared polygenic component underlying the eye and vascular phenome comorbidity results. We tested the top associations ($p < 0.05$) from a large-scale GWAS of age-related macular degeneration (AMD)[24] for their association with vascular traits in BioVU. We generated a Q-Q plot from each (vascular trait association) analysis of the top AMD variants; a leftward shift from the diagonal "null" line in a Q-Q plot would indicate enrichment. To show that the observed enrichment was not driven by LD among the tested variants, we applied a polygenic test that takes the LD into account. We used the sum of all the chi-squared 1 df (degrees of freedom) statistics as the overall test statistic and simulated 1,000 multivariate normal-distributed n vectors v (with n meaning the number of tested variants) with a mean of zero and a covariance matrix Σ given by the $n \times n$ pairwise LD matrix:

$$v \sim N(0, \Sigma)$$

$$v \in R^n$$

The proportion of simulated datasets with test statistics that match or exceed the observed test statistic in the actual dataset yields an empirical enrichment p value. More generally, this approach provides a way to test for polygenic enrichment by using only summary statistics.

Initial PrediXcan Studies

Initial PrediXcan studies were conducted in 5,240 BioVU subjects of European descent via prediction equations built from whole blood on the basis of data on >900 subjects from the DGN data set, containing RNA-Seq and genotype array data (Illumina Omni1-Quad),[25] and on data from the heart left ventricle for 190 subjects from GTEx. The BioVU subjects had been genotyped using the Omni1-Quad array, and because the data from these studies had not yet been imputed, we focused on genes for which all SNPs in the prediction equations had been directly genotyped on the Omni1-Quad array; these included 125 genes in whole blood and 298 genes in the heart left ventricle. Summary results for the *GRIK5* GReX (whole blood) associations with phecodes are provided in Table S1.

The institutional review board reviewed the research study and determined that the study does not qualify as "human subject" research per §46.102(f)(2).

Comprehensive PrediXcan Analyses

We used transcriptome data in 44 tissues and whole-genome genotype data (imputed to sequence data from the 1000 Genomes reference panel) derived from the same donors across tissues from the GTEx Consortium[20,26] as a reference resource to build gene-expression prediction models, i.e., SNP predictors and corresponding weights for the gene expression level. For downstream analysis, we are interested primarily in GReX (i.e., imputed gene expression level) rather than the trait-altered component of gene expression or other factors (such as environmental regulators or technical confounders). GReX is then tested for association with a clinical trait as represented by a phecode in an EHR-linked database such as BioVU.[15,27] An observed association between GReX and a clinical trait suggests a causal, easy-to-interpret direction of effect because the clinical trait is not likely to alter the germline genetic profile. This analytic workflow allows us to generate a comprehensive medical-phenome catalog consisting of all associations, including direction of effect, between GReX for each tested gene and each clinical trait in the EHR.

We performed transcriptome-wide association (TWAS) analysis on BioVU by using PrediXcan. This analysis generated a list of genes associated with clinical traits; in particular, we obtained the effect size and the level of significance (p value) for each gene-trait association. As a framework for identifying genes associated with disease traits, we would apply Bonferroni adjustment (adjusted p value < 0.05) on the basis of the number of tissues ($n = 44$) and the number of phecodes ($n = 1{,}965$). The clinical traits we used in this analysis are represented by phecodes,[15] which are algorithmically defined from ICD-9 codes in the EHR database. One of the top findings, on the basis of the level of significance and direction of effect for a set of related clinical traits affecting eye function, was the glutamate ionotropic kainate receptor KA2 encoded by *GRIK5*. The gene is expressed in a variety of tissues, including the brain regions sampled within GTEx.

Assessing Significance for a Set of Related Clinical Traits

We conducted a permutation analysis that preserves both LD for genetic variants and correlations among phecodes. Let Y denote the $N \times Q$ matrix of phecodes, where N is the number of individuals and Q is the number of phecodes. Let G be the $N \times P$ genotype matrix (encoded as the number of effect alleles), where P is the number of genetic variants in the imputation model of a gene. Let $\widehat{\beta}$ be the P-dimensional effect size vector from the imputation model. The rows for the phecode matrix Y were randomly shuffled so that permuted datasets ($n = 10^6$) were generated but the genotype matrix G was left untouched. For each permuted dataset,

logistic regression was performed, as in the actual data, between each phecode (column Y_{*j}) and the N-dimensional vector $G * \widehat{\beta}$ of genetically determined expression. We calculated a statistic T, namely the number of nominally significant associations ($p < 0.05$) between reduced genetically determined expression and the phecodes for eye phenotypes, from each permuted dataset, yielding an empirical distribution. We compared the value T_0 of the statistic for the actual dataset with this empirical distribution and assessed the significance of the multi-trait association as the proportion $\widehat{P}(T \geq T_0)$ of permuted datasets with statistic T that matches or exceeds T_0. We also calculated a combined test statistic by using Fisher's method ($X^2_{2k} \sim -2\sum_{i=1}^{k} \ln(p_i)$) applied to the p values (p_i) for the *GRIK5* associations, and we evaluated its significance by comparing its frequency with an empirical frequency distribution of the maximum combined test statistic across all genes tested.

Zebrafish Studies

Fish Husbandry and Breeding

Zebrafish were raised at 28.5°C in standard laboratory conditions as previously described.[28] All experiments were performed according to an approved animal protocol and guidelines established by the institutional animal care and use committee at Vanderbilt University Medical Center.

Quantitative PCR Analysis

Quantitative real-time PCR (qRT-PCR) was performed as described previously.[29,30] Total RNA was extracted from 10–15 embryos (per sample) at different developmental stages with the TRIzol reagent (Thermo Fisher Scientific). We used 500 ng of total RNA as a template for reverse transcription to make cDNA with Moloney murine leukemia virus (M-MLV) reverse transcriptase (Promega) and poly-T primer. Each PCR reaction was performed with 20 ng of cDNA, SYBR Green Real-Time PCR Master Mix, and 2 μM of each primer. The primer sequences we used in this study are as follows: β-actin: 5′-GACTCAGGATGCGGAAACTG-3′ and 5′-GAAG TCCTGCAAGATCTTCAC-3′; grik5_set#1: 5′-CCACCAGCCTGGA CATCAAT-3′ and 5′-AGCTACGGCCAAATCAGCTT-3′; and grik5_set#2: 5′- CCGTACGGATGGCTGCTATT-3′ and 5′-GACCAC GCCCTTTGGTAGAA-3′. qRT-PCR reactions were run on the CFX96 (Biorad) system. Data were analyzed with the –ΔΔCt method.

Morpholino Knockdown and mRNA Rescue

An antisense morpholino oligonucleotide (MO) (Gene Tools) was designed to target the 5′ UTR of *grik5* (5′-GAGATGCCTTCTGC TGCCTATAGCA-3′) as described previously.[31] 1 nl of MO was injected into one-cell-stage zebrafish embryos at varying concentrations (1 ng, 2 ng, 4 ng, or 6 ng) so that the effective dose could be determined, and 6 ng MO was used throughout this study. As a negative control, a 25-base random-sequence-mixture control oligo (Gene Tools) was injected at the same concentration (6 ng) throughout this study. The phenotypes were evaluated by two screeners, independently, in a double-blinded manner.

We purchased the human *GRIK5* cDNA clone in pENTR223.1 from Plasmid ID, Harvard Medical School (clone ID: HsCD00295568) and subcloned it into pCS2+ by using ClaI and XbaI (NEB) restriction enzyme sites for ligation. We then linearized a ligated construct with NotI (NEB) and used it for an *in vitro* transcription reaction to synthesize mRNA with the mMESSAGE mMACHINE SP6 Transcription Kit (Thermo Fisher Scientific). 250 pg human *GRIK5* mRNA was injected in combination with MO in rescue experiments.

CRISPR-Cas9 Genome Editing

CRISPR-Cas9 target sites within zebrafish *grik5* (GRCz10 assembly) were identified with the CHOPCHOP web tool. With the CHOPCHOP web tool, we investigated all potential (0, 1, 2 and 3) mismatches between off-targets and guide RNAs (gRNAs) to avoid non-specific mutations. gRNAs with *in silico* efficiency scores greater than 0.55 were prioritized. The site g03: GGTGGAC GATGGTCTGTACGGGG, which shows no off-targets, was selected in this study. A cloning-free method for generating gRNA templates was used as previously described.[32] We obtained the RNA template by annealing an 80 nt chimeric gRNA oligo (5′-TTTTGCACCGACTCGGTGCCACTTTTTCAAGTTGATAACGG ACTAGCCTTATTTTAACTTGCTATTTCTAGCTCTAAAAC-3′) with a gene-specific oligo (5′- AATTAATACGACTCACTATAGG[N18] GTTTTAGAGCTAGAAATAGC-3′) that contains a T7 promoter sequence, 18 nt *grik5* gRNA target (5′-TGGACGATGGTCTG TACG-3′), and an overlapping sequence to the chimeric single guide RNA (sgRNA) oligo. Upon the annealing of two oligos, gaps were filled with T4 DNA polymerase (NEB), yielding a double-stranded linear template for *in vitro* transcription. gRNAs were synthesized with the MEGAshortscript T7 transcription kit (Thermo Fisher Scientific).

To generate mutations with the CRISPR-Cas9 system, we injected a mixture of 500 pg purified Cas9 (PNA Bio, Cat No # CP01) and 100–500 pg gRNA into one-cell-stage embryos. Prior to microinjection, gRNA and Cas9 were mixed and incubated on ice for 10 min to form a ribonucleoprotein complex. We included Cas9 -only and gRNA-only injections in the experimental design to control for potential non-specific phenotypes. To generate founders (G_0), we raised embryos injected with Cas9 and *grik5*-targeting gRNA to adulthood.

Direct Sequencing for Mutation Detection

The PCR products of CRISPR targets were cloned into pGEM-T easy (Promega) according to the instruction manual. Plasmids isolated from pGEM-T clones were sequenced with SP6 primer (Genewiz). We used NCBI-NLM's BLAST tool to align mutant sequences with a wild-type (WT) *grik5* sequence to detect deletions, insertions, and substitutions.

Whole-Mount Imaging of Transgenic Zebrafish Embryos

Transgenic Tg(flk1:eGFP) embryos at the 2-days-post-fertilization (dpf) or 3 dpf stage were mounted in low-melting agarose (Sigma A-9414, 1.2% in embryo medium) in glass-bottomed dishes (MatTek Corporation). Images were acquired with a spinning-disk confocal microscope (Nikon) with either a Plan Apo λ 10× or a 20× objective lens and an Andor DU-897 EMCCD camera. Z stacks were acquired for 3D analysis, and volume views were generated with Nikon NIS-Elements software.

Mutant *grik5*^{g03} embryos and their WT and heterozygous siblings were imaged and analyzed in a blinded manner before the genotype was confirmed by sequencing.

Live Imaging of Zebrafish Embryos

Live embryos were anesthetized in tricaine (Sigma) and mounted in 3% methylcellulose (Sigma) on a bridge slide for imaging. Images were acquired with a Stemi 2000-C Stereomicroscope (Zeiss) and an Axiocam HRc camera under transmitted light illumination. For magnified views of hemorrhage phenotypes and detailed localization analysis, embedded embryos were imaged under double spot illumination with an Axioimager.Z1 (EC Plan NEOFLUAR

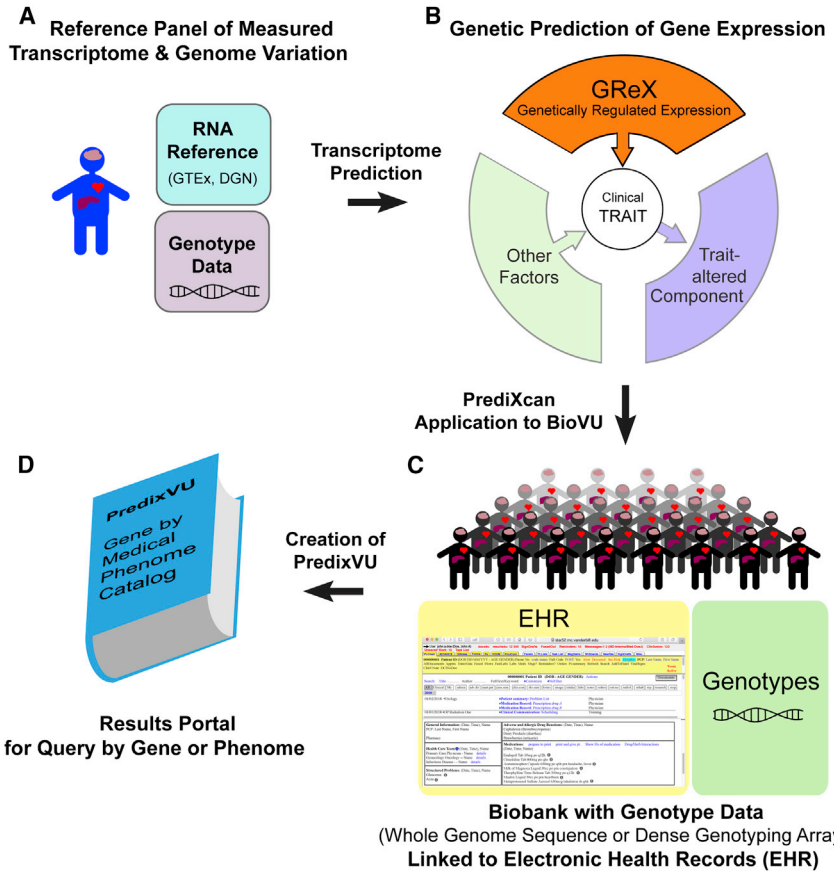

Figure 1. Experimental Pipeline for the Construction of the Medical Phenome Catalog

(A) We utilized transcriptome data in 44 tissues and whole-genome genotype data derived from the same donors from the GTEx Consortium as a reference resource to build gene-expression imputation models.

(B) We used the PrediXcan method and these prediction models to impute tissue-level gene expression in the BioVU subjects. This imputed expression level is referred to as the genetically regulated expression (GReX) to distinguish it from measured expression levels. For downstream analysis, we are interested primarily in GReX rather than the trait-altered component of gene expression or other factors (such as environmental regulators).

(C) Biobank with genotype data linked to electronic health records (EHRs). GReX is tested for association with a clinical trait that is represented by a phecode.

(D) This analytic workflow allows us to generate a comprehensive gene-by-medical-phenome catalog, consisting of all associations, including direction of effect, between the GReX for each tested gene and each clinical trait in the EHR.

5× and 10× objective lenses) equipped with an Axiocam HRc camera.

Nanobead (Microsphere) Injection and Microangiography

0.02-μm-sized carboxylate-modified microspheres (FluoSpheres, Thermo Fisher Scientific F8786) that emitted red fluorescence (580/605) were diluted 1:5 in 0.3× Danieau buffer to 0.4% final concentration. We sonicated the nanobeads with a Heat Systems sonicator three times with 5 min cycles to prevent clumping. We then injected microspheres into the common cardinal veins of 3 dpf transgenic Tg(flk1:eGFP) zebrafish embryos to mark the blood plasma in magenta. Injected, live embryos were lightly anesthetized in tricaine and mounted sagittally in 1.2% low-melting agarose (Sigma, A9414) in coverslip-bottom dishes (MatTek Corp., P35G-1.5-14C) for confocal imaging. Multichannel images were acquired with a spinning-disk confocal microscope (Nikon) with a Plan Apo λ 20× objective lens and an Andor DU-897 EMCCD camera. Z stacks were acquired for 3D analysis, and volume views were generated with Nikon NIS-Elements software.

Results

Construction of PredixVU, a Gene-by-Medical-Phenome Electronic Catalog

We devised a strategy to link genome variation to EHR phenotypes; this project resulted in the construction of a large-scale gene-by-medical-phenome catalog (Figure 1). We used reference data, including measurements of both genome variation and gene expression levels found in such databases as GTEx[26] or the DGN,[25] to build SNP-based prediction or imputation models for measured gene expression levels in up to 44 tissues by using the PrediXcan[1] approach (Figure 1A). We then applied the prediction models to genotyped individuals from BioVU to estimate the level of genetically regulated expression, GReX (see Material and Methods), for each gene in each tissue (Figure 1B). We then tested the associations between the calculated GReX values and DNA variation from subjects in a biobank with 1,145 "phecodes" (phenotype codes, see Material and Methods) associated with these subjects (Figure 1C). The results of these association studies comprise a comprehensive gene-by-medical-phenome electronic catalog we have termed PredixVU (Figure 1D). Here, we present one example of an association we discovered between reduced *GRIK5* expression and the eye phenome, and we test it with *in vivo* loss-of-function approaches in a model organism (zebrafish) and with the analysis of phenome relationships in large-scale EHR data.

Discovery of Association between *GRIK5* GReX and Eye Diseases

Summary results for the initial analysis of GReX (built from the DGN whole blood transcriptome as the reference panel) associations with eye-disease phecodes ($p < 0.05$) in 5,240 BioVU subjects of European descent (determined

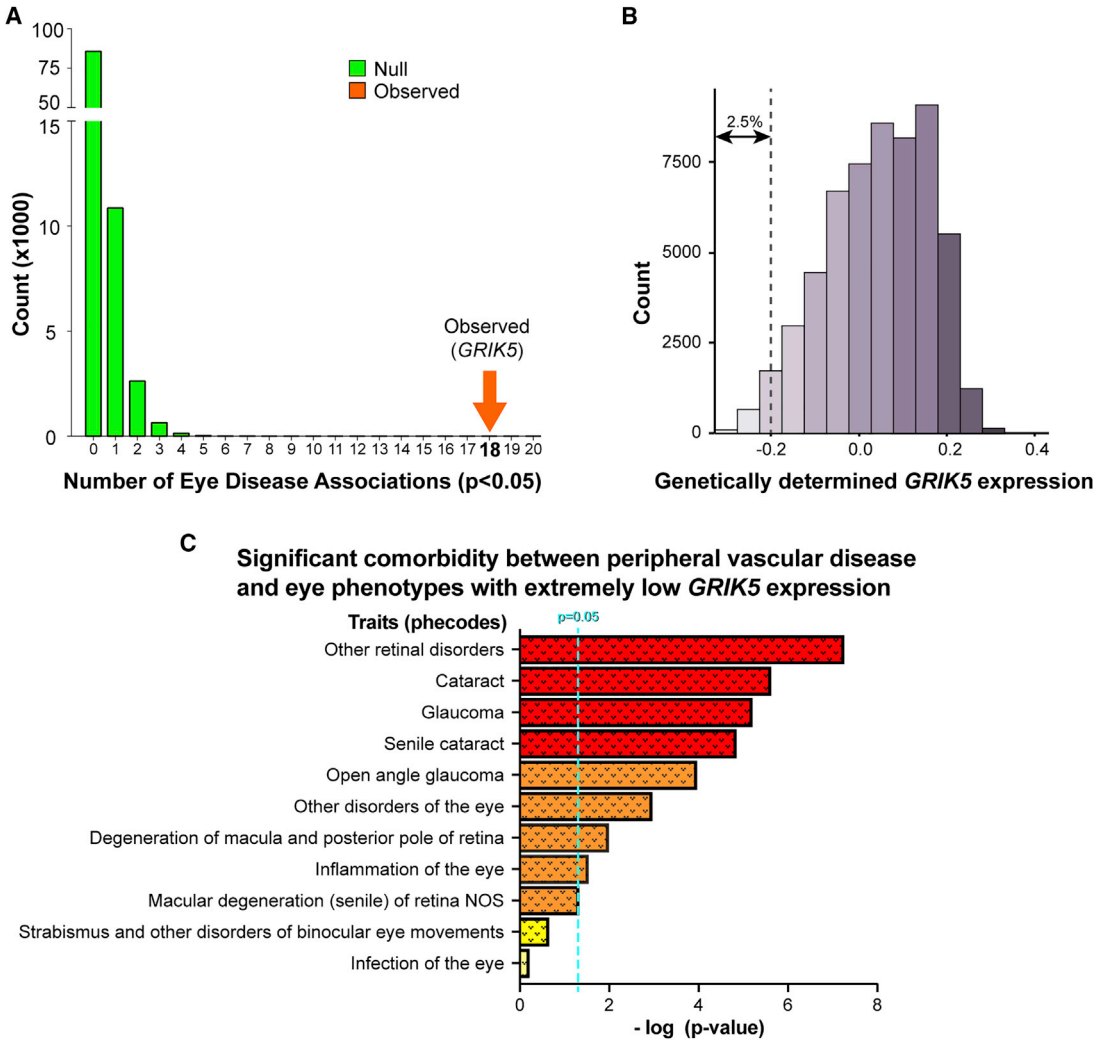

Figure 2. *GRIK5* Genetically Determined Expression and Association with Disease Phenome
(A) On the basis of 100,000 permuted datasets that preserve the pairwise SNP-SNP correlations (i.e., linkage disequilibrium [LD]) and the pairwise trait-trait correlations, the number of associations ($p < 0.05$) that the genetically determined expression of *GRIK5* has with disorders of the eye is significantly greater than expected by chance (empirical $p < 0.001$). The orange arrow shows the observed number in the actual data, and the green bars show the null distribution from the permuted datasets.
(B) A histogram displaying distribution of genetically determined *GRIK5* expression in 60,000 BioVU subjects. The dotted line indicates the cutoff for the bottom 2.5% (with extreme, reduced genetically determined *GRIK5* expression) of the population.
(C) A comorbidity analysis between peripheral vascular disease and eye phenotypes (traits) in the bottom 2.5% of the population from (B). The dotted line shows a $p = 0.05$ cutoff. Yellow = non-significant, orange = nominally significant ($p < 0.05$), and red = significant (Bonferroni-adjusted $p < 0.05$) traits. Only traits present in >20 affected individuals were included.

via a principal-components analysis[33] of genotype data) are illustrated in Figure 2A. Reduced GReX of *GRIK5* showed nominally significant associations with 18 of 55 eye phenotypes (Table S1) that were demonstrated by at least 20 affected individuals and with 16 of 39 eye phenotypes that were demonstrated by at least 50 affected individuals. We considered the possibility of LD-contaminated associations in the locus. None of the 10 most-adjacent genes in each direction showed a similar pattern of eye-phenome association. There were four phenotypes associated at $p < 0.001$ to *GRIK5* and also to one of the 20 adjacent genes; none were eye phenotypes. Using permutation analysis (see Material and Methods) that preserves the correlation among the phecodes and the correlations among genome variants (Table S2) in LD, we found that the observed number of eye-disease associations was significantly greater than expected by chance (permutation $p < 10^{-6}$, Figure 2A and Table S1; in 100,000 permutations, one gene had significant associations with a maximum of six eye phenotypes; no genes showed significant associations with seven or more eye phenotypes). A combined test statistic ($X^2_{2k} \sim -2\sum_{i=1}^{k} \ln(p_i)$) that uses Fisher's method and was applied to the p values for *GRIK5* was highly significant (empirical adjusted $p < 10^{-6}$) on the basis of a comparison with an empirical frequency distribution of the maximum combined test statistic across all genes tested.

In analyses of the largest sample sizes currently available to us for studies on the association of the reduced GReX of *GRIK5* (~60,000 participants), among those with the lowest *GRIK5* expression (bottom 2.5% of the distribution, Figure 2B), there is significantly increased risk of peripheral vascular disease and comorbid eye traits (Figure 2C), including "other retinal disorders" ($p = 5.57 \times 10^{-8}$), cataract ($p = 2.41 \times 10^{-6}$), and glaucoma ($p = 6.26 \times 10^{-6}$) (Figure 2C); this increased risk is consistent with a vascular component to the pathophysiology of these ocular diseases. Notably, we find no associations with infection of the eye ($p = 0.61$) or strabismus and other disorders of binocular eye movements ($p = 0.22$), suggesting a distinct pathophysiology for such eye disorders. These findings led to the fundamental question of the physiological function of *GRIK5* and to how that might relate to the potential comorbidity of eye and vascular diseases.

Reduced Expression of *grik5* in Zebrafish Leads to Eye Phenotypes and Bleeding

GRIK5 encodes a glutamate-binding receptor that has been implicated in the regulation of neurotransmission in the brain and sensory organs by animal model research and association studies.[34–36] However, significantly less is known about its other physiological functions. To examine the role of *grik5* during development, we generated loss-of-function models by using CRISPR-Cas9 genome editing strategies (knockout; KO) and morpholino oligonucleotide-based (MO) protein knockdown (KD) (Figures 3A and 3B). The genetic KO lines of *grik5* offer independent assessment of the phenotypes observed in morphants. We used an approach involving MO because its dose can be titrated (Figure S1), thus allowing testing of the effects of partially reduced protein translation, which in turn better resemble reduced gene expression in humans. We used a dose of 6 ng of the 5′-UTR-targeting MO per embryo. The effectiveness of the *grik5* MO for blocking mRNA translation was confirmed by *in vivo* tests (Figure S2).

Zebrafish *grik5* knockdowns with morpholino oligonucleotides and CRISPR-induced mutations presented with two visible structural defects: smaller eyes (Figure S3) and defects in blood-vessel integrity (Figures 3C and 3E). The small-eye phenotype observed in zebrafish is more likely to be related to the expression of *grik5* in the photoreceptor cells and the outer plexiform layer of the retina, as shown in mouse embryos.[34] Our *in situ* hybridization study also confirmed abundant *grik5* transcripts in the developing zebrafish eye, further corroborating the initial findings (Figure S3). Because *grik5* expression might be directly or indirectly linked to metabotropic activity of the glutamate receptor, it would not be surprising that it results in a small-eye phenotype. The deficits in blood vessel integrity included blood extravasation in various organs at 3 dpf (Figures 3C, 3D, S1, and S4 and Table S3).

Human and zebrafish *GRIK5* and *grik5* gene products are 85% identical and 94% similar, respectively, at the primary sequence level. Taking advantage of this high homology, we tested the specificity of *grik5* MO knockdown by designing a rescue experiment for the bleeding phenotype with human *GRIK5* (*hGRIK5*) mRNA (non-targetable by *grik5* MO). Overexpression of human *GRIK5* in morphants rescued bleeding to levels comparable to those in control-MO and non-injected siblings (Figures 3D and S1B), demonstrating that bleeding defects are specific to *grik5* depletion. Hemorrhage in *grik5* morphants was most often visible in the fourth ventricle of the brain and/or retro-pericardial space.

Concurrently, we generated *grik5* mutations by using guide RNA to target the glutamate binding site (Figures 3A, 3B, 3E, and S5). We sequenced the founder fish carrying germline *grik5* mutations, and we analyzed trans-heterozygous F1 progeny to rule out off-target effects (Figures 3E and 3F). F1 animals displayed local hemorrhage in the brain and/or in the retro-pericardial space in 57 out of 309 analyzed fish (~18%). To ensure that the observed bleeding phenotype was specific to mutants, we sequenced individual F1 embryos in a blinded, unbiased manner. We detected deleterious mutations in embryos with hemorrhage, whereas siblings with normal morphology were either heterozygous or WT homozygous for *grik5* (Figure S5).

We reasoned that the expression of *grik5* would correlate with the onset of the phenotype. We used publicly available data for whole-fish RNA-seq expression,[37] which we confirmed by qPCR (Figures S3A and S3B) at 1–4 dpf and by *in situ* hybridization at 3 dpf (Figures S3D and S3E). Quantitative PCR (qPCR) analysis of whole embryos indicated that *grik5* expression peaked at 3 dpf and localized to the brain and the retina (Figure S3B), as confirmed by an independent set of primers and riboprobes (data not shown). These findings were corroborated by independent RNA-seq data from fluorescence-activated cell sorting (FACS)-sorted Tg(flk1:GFP+ endothelial cells in zebrafish at the same developmental stages (Figure S3C).[38] Collectively, *grik5* is expressed at the time and location that could explain the bleeding (Figures 3C and 3E) and smaller eye phenotypes (Figures S3F and S5) in *grik5* morphants and mutants.

The Comorbidity of Vascular and Eye Disease

The initial and most recent PrediXcan results in BioVU and the zebrafish validation studies together suggest that defects of early and potentially lifelong vascular conditions (e.g., reduced blood-vessel number and integrity, increased vascular permeability) could increase the risk of late-onset eye disease in humans. We therefore hypothesized that eye diseases would be comorbid with vascular diseases. To test this hypothesis, we calculated the relative risk (RR) for eye diseases initially found to be associated with reduced GReX of *GRIK5* in individuals with a set of vascular disease phenotypes that were the most strongly associated with reduced GReX of *GRIK5* in studies on larger numbers of BioVU subjects by utilizing Vanderbilt's SD, which is the

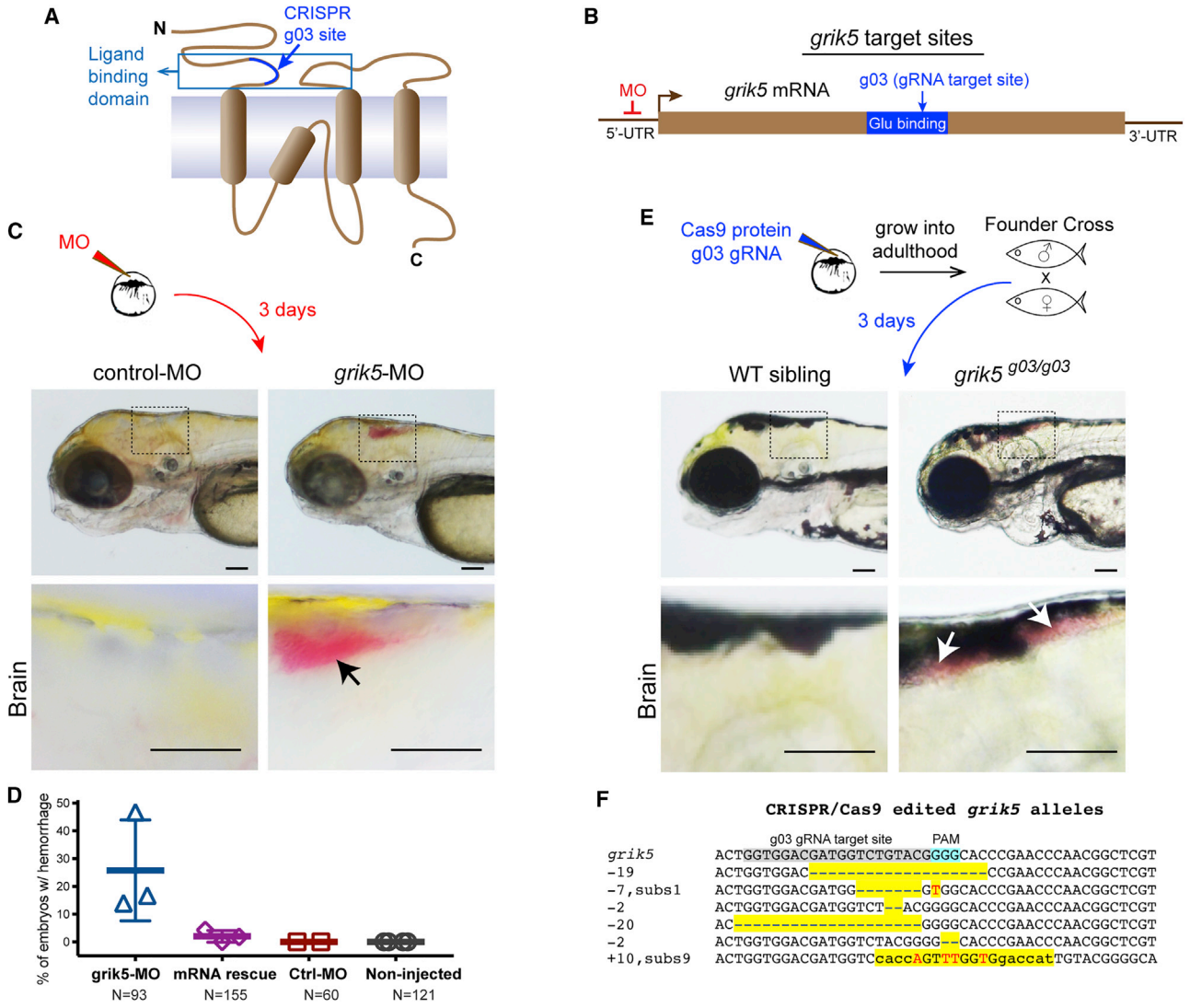

Figure 3. Loss-of-Function of *grik5* in Zebrafish Leads to a Bleeding Phenotype

(A) A model of GRIK5 protein structure shows the predicted glutamate-binding pocket targeted by the guide RNA (gRNA) g03.

(B) The position of the morpholino (MO) and gRNA CRISPR-Cas9 target sites that were used to generate *grik5*-depletion models.

(C) Experimental design and lateral views of head regions in live embryos. The box depicts brain bleeding and is zoomed in below; the arrow points to the 4th ventricle.

(D) A summary of the percentage of embryos with hemorrhage in MO and h*GRIK5* mRNA rescue and control groups. Human *GRIK5* mRNA that is non-targetable by zebrafish *grik5*-MO was used in rescue experiments. N = total number of tested animals. The results of independent experiments are indicated by shapes as data points. Mean and standard deviation bars are indicated on the graph.

(E) Live images of wild-type (WT) and CRISPR-Cas9-edited *grik5*$^{g03/g03}$ mutants. Zoomed-in images of the boxed brain regions are shown below. Arrows mark blood accumulation.

(F) Sequences of *grik5* alleles. Highlighted regions mark the detected mutations in *grik5*^{g03}; dashes (-) = deletions; red color = substitutions; and lowercase = insertions.

Scale bars represent 0.1 mm in panels C and E, which show examples of pigment blocker (PTU)-treated (C) and untreated (E) embryos.

de-identified image of the EHR in >2.6 million subjects. There are substantial increases in the risk for eye disease across the entire spectrum of vascular phenotypes (2- to 15-fold; Figure 4A). The "burns" phenotype is included as a control phenotype that has a broad range of age at diagnosis and, although prior vascular disease might complicate healing from a burn, it is unlikely to have a major impact on whether an individual seeks medical treatment for a burn, and a burn is similarly unlikely to have much impact on whether someone seeks medical treatment for vascular disease. 72% (39/54) of the eye/vascular phenotype pairs have an RR > 4.0. In two of the 39 instances, significant comorbidity had already been reported in the literature.[39] We similarly estimated the RR[21] for these phenotypes in >150 million subjects from the MarketScan insurance claims dataset[40] (Figures 4B, 4C, and S6) by age and sex, confirming the magnitude of the comorbidity for eye and vascular disease (see Material and Methods). Age was not significantly associated with the extent of comorbidity in males and females

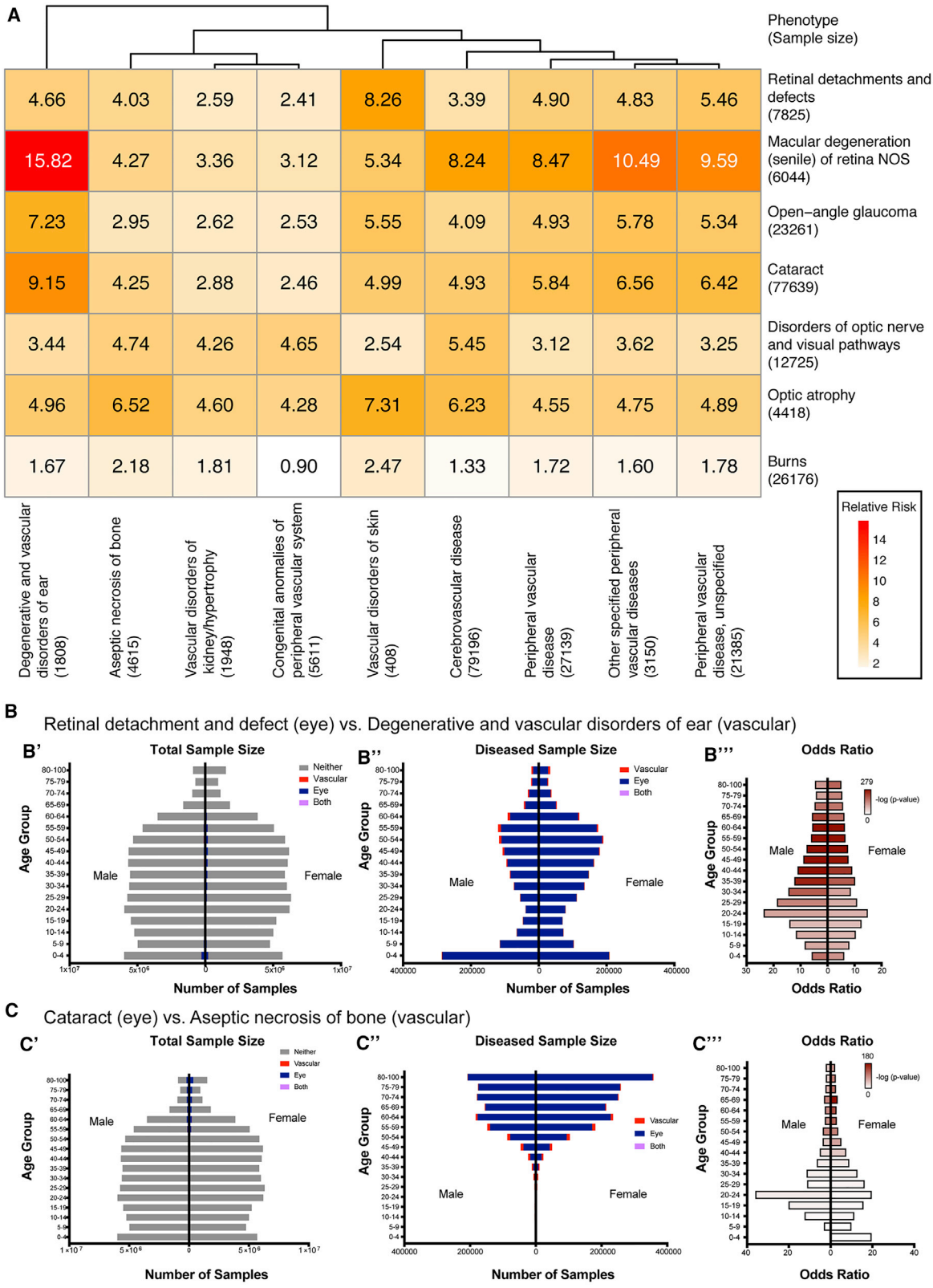

Figure 4. Comorbidity Analysis of Eye and Vascular Disorders

(A) A heatmap shows the relative risk for eye diseases in individuals with a set of vascular disorders. The comorbidity estimates were calculated with Vanderbilt's Synthetic Derivative (SD), the de-identified image of the electronic health records (EHRs) of >2.6 million participants. There is a 2- to 15-fold increase in risk for eye disease across the set of vascular phenotypes. The "burn" is included as a control phenotype. Notably, 72% of the eye-vascular disease pairings have a relative risk (RR) > 4.0.

(legend continued on next page)

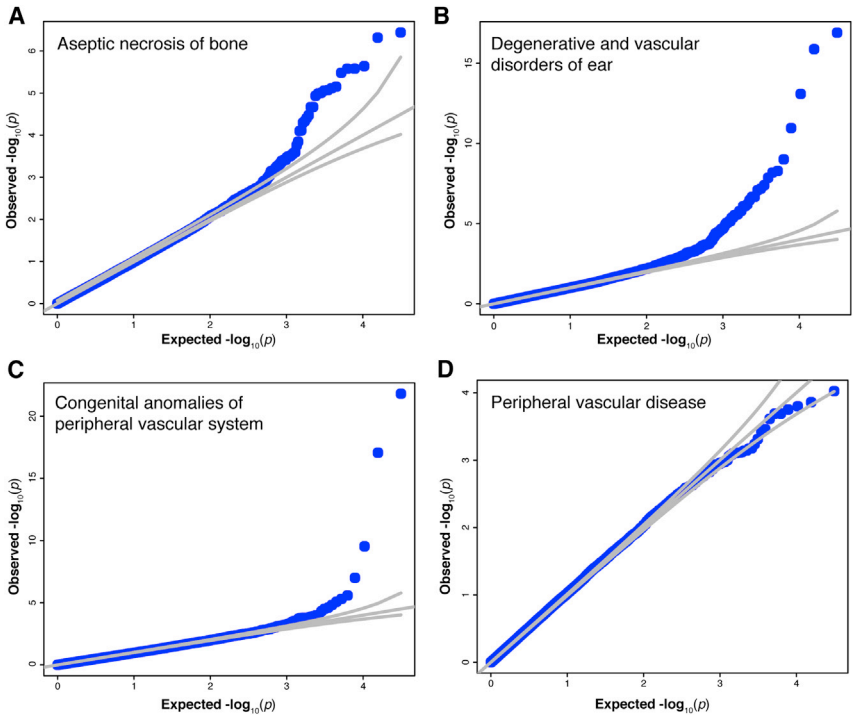

Figure 5. Shared Genetic Architecture Underlying Eye Disease Phenome

(A–D) Q-Q plots show the distribution of association p values for vascular traits and the top age-related macular degeneration (AMD) variants. Using a publicly available genome-wide association study (GWAS) of AMD, we find that top AMD variants ($p < 0.05$) are significantly enriched for a variety of vascular disorders, including aseptic necrosis of bone (A), degenerative and vascular disorders of the ear (B), and congenital anomalies of the peripheral vascular system (C), although they are not enriched for peripheral vascular disease (D). Enrichment for association with vascular traits persisted in comparison with null sets (n = 1000) of SNPs with a matching minor allele frequency (MAF) (generated from bins of length 5% using the 1000 Genomes EUR samples), distance to the nearest gene (on the basis of GENCODE gene annotation data), and number of linkage disequilibrium (LD) partners ($r^2 > 0.50$).

(Kruskal-Wallis test p = 0.38 and p = 0.70, respectively), and no significant difference was observed between males and females (Mann-Whitney U test p = 0.43), demonstrating that this observation is robust and unlikely to be confounded by these demographic variables.

Shared Gene Mechanisms Underlying the Vascular and Eye Phenomes

We sought additional support for the hypothesis that vascular and eye diseases share genetic architecture by integrating publicly available GWAS summary data to further investigate pleiotropic genetic effects for eye diseases and vascular traits. Using data from the recent GWAS meta-analysis (n = 375,000) of migraine risk (a neurological disorder with a vascular etiology),[41] we found that reduced *GRIK5* GReX (from whole-blood DGN transcriptome) was significantly associated with this vascular trait ($p = 9 \times 10^{-8}$). Interestingly, we also found a significant comorbidity (RR = 2.4, Bonferroni-corrected $p = 4.3 \times 10^{-36}$) between migraine and sensory retina dystrophies in the MarketScan dataset. Conversely, we utilized the recent AMD GWAS, the largest GWAS meta-analyses available for eye disease[24] (n > 33,000), to test for associations with vascular traits in BioVU (n = 28,358). Among the top 100 AMD-associated genes, we found significant PrediXcan associations (Benjamini-Hochberg adjusted $p < 0.05$) with "peripheral vascular disease" (Table S4) in BioVU. This extends similar results in the most recent BioVU sample (n = 60,000), in which we observed co-morbidity of a vascular trait (cerebrovascular disease) with "macular degeneration (senile) of retina NOS (not otherwise specified)" ($p = 9.73 \times 10^{-4}$) among the participants with the lowest *GRIK5* expression (in the bottom 2.5%; Figure S7).

We hypothesized the presence of a shared polygenic component between eye diseases and a wide spectrum of vascular traits. We evaluated the top AMD-associated genetic variants from the meta-analysis ($p < 0.05$) for their association with a variety of vascular traits in BioVU and tested the hypothesis with independent base and target studies. Notably, we found among the AMD-associated variants a highly significant enrichment of low p values with vascular traits, including congenital anomalies of the peripheral vascular system, aseptic necrosis of bone, and degenerative and vascular disorders of the ear, though not with peripheral vascular disease. This enrichment is illustrated by the significant departure from the null distribution in the Q-Q plots (Figures 5A–5D, Kolmogorov Smirnov $p < 2.2 \times 10^{-16}$). Using a polygenic empirical test for enrichment (see Material and Methods) that takes into account the LD among variants, we continued to observe a highly significant enrichment (empirical $p < 0.001$) for the same set of vascular traits.

We developed a genetic risk score by using the effect alleles in the AMD dataset and found a significant association with disease status for each vascular trait, including, notably, peripheral vascular disease ($p < 2.2 \times 10^{-16}$ for

(B and C) RR analysis of eye and vascular diseases in the MarketScan dataset.[40] (B) Retinal detachment (eye) and defect versus degenerative and vascular disorders of the ear (vascular). (C) Cataract (eye) versus aseptic necrosis of bone (vascular). (B′ and C′) Total sample sizes represented in each gender and age group. (B″ and C″) Sample sizes include only diseased subjects (excluding subjects with neither eye nor vascular diseases analyzed). (B‴ and C‴) The odds ratios of eye versus vascular disorders displayed in gender and age groups.

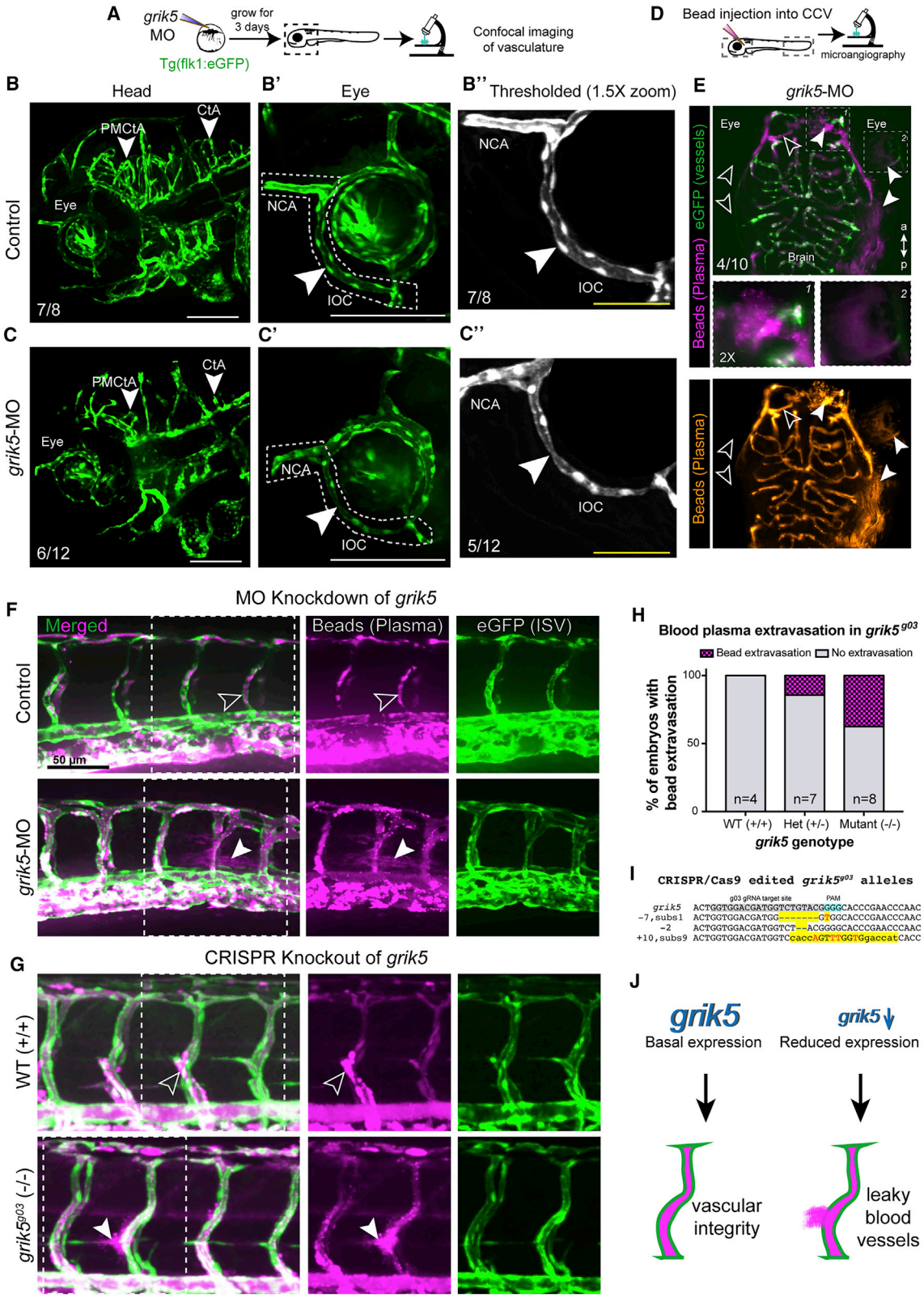

Figure 6. Blood Plasma Extravasation Revealed by Nanobeads in *grik5*-Depleted Embryos

(A–C) The experimental design (A) and live, spinning disk confocal images (B and C) of the lateral head in Tg(flk1:eGFP) show defects in the central arteries (CtA) and posterior mesencephalic central arteries (PMCtA) of the brain in a control (B) and a *grik5*-KD zebrafish (3D reconstruction of confocal stacks) (C). (B′ and C′) Confocal images of the eye vessels. There is thinning of the inner optic circle (IOC) in the *grik5*-KD zebrafish as compared to the control (see arrowheads). A digitized (by ImageJ) depiction of comparable IOC vessel-thinning areas (arrowheads) is demarcated in B″ and C″. The abbreviation NCA = nasal ciliary artery.

(legend continued on next page)

each phenotype), in BioVU. Although the AMD-associated variants were not enriched for the most significant associations with peripheral vascular disease, the significant (Spearman) correlation ($p = 7.1 \times 10^{-9}$) in per-SNP heritability (which is proportional to $\beta^2 * p * (1-p)$, where β is the estimated effect size and p is the minor-allele frequency in controls) between the traits for the top AMD variants contributes to the predictive performance of the genetic-risk score on the vascular trait. Furthermore, the significant comorbidity (as noted above) of the lowest genetically determined *GRIK5* expression (in the bottom 2.5%) of "macular degeneration (senile) of retina NOS" with vascular conditions, including cerebrovascular disease (Figure S7), and peripheral vascular disease (Figure 2C) in the BioVU participants, suggests that genetic regulation of *GRIK5* contributes to the overall effect of the genetic-risk score on these vascular traits.

The availability of the large-scale AMD dataset also allowed us to evaluate how well imputation of gene expression would work when we used non-eye tissue as a reference to analyze ocular traits. Using a recently developed methodology,[42,43] we found that tissue-shared expression quantitative trait loci (eQTLs) (defined as eCAVIAR posterior probability ≥ 0.90 in more than 90% of the GTEx tissues, i.e., 40 or more) explain a greater proportion (2-fold) of the heritability (h^2) and capture a greater proportion of estimated true positive rates (0.11 vs 0.05) of association for AMD than did tissue-specific eQTLs (posterior probability ≥ 0.90 in at most 10% of the tissues, i.e., at most five tissues).

Vascular Quality and Number of Vessels Are Compromised in *grik5*-Depleted Animals

A statistical analysis in BioVU and MarketScan EHR data revealed strong comorbidity between eye and vascular traits. Thus, to investigate the effect of *grik5* depletion in vasculature, we examined *grik5* morphant zebrafish embryos displaying hemorrhage, which typically results from the disruption of a blood vessel and leakage into interstitial spaces. The diminished integrity of vascular endothelium is the common cause of hemorrhage.[44,45] Thus, we reasoned that *GRIK5* expression is required for supporting vascular patterning and integrity. To model reduced *grik5* expression, we injected MO into transgenic Tg(flk1:EGFP) embryos, which express EGFP in vascular endothelial cells,[46] and analyzed the *grik5* morphants' and control embryos' vasculature with confocal imaging and 3D reconstruction (Figure 6A). We observed that central brain arteries (CtA) and posterior mesencephalic central arteries (PMCtA) failed to sprout in severe *grik5* morphants (Figures 6B and 6C). Although the eye vasculature was patterned, the vessels of the inner optic circle (IOC) were dilated and/or constricted throughout (Figures 6B′, 6B″, 6C′, and 6C″ and Video S1), suggesting not that the architectural defects in vasculature were due to general developmental delay but rather that they indicated diminished vascular quality and endothelial integrity. Similar and consistent vascular defects were observed in *grik5* genetic mutants and, to some extent, in heterozygous siblings (Figure S8), corroborating the link between reduced *grik5* expression and compromised vascular quality.

To directly test functional vascular integrity in *grik5*-depleted embryos, we injected 0.02-μm-sized nanobeads into circulation through the common cardinal vein (CCV) at 3 dpf (Figure 6D). We used fluorescent nanobeads to monitor blood flow within green-labeled vasculature of the Tg(flk1:EGFP) transgenic line. Live embryos were imaged via spinning-disk confocal microscopy of the head and the trunk. In control embryos, the nanobeads were confined to GFP+ vessels, whereas in *grik5* morphants the beads extravasated from vessels that appear intact (Figure 6E). In the head, we detected nanobead extravasation in the forebrain, the eye, and the ear of *grik5* morphants (Figure 6E), whereas none of the controls showed extravasation. Similarly, in the trunk, nanobeads extravasated from intersegmental vessels in both morphants and genetic mutants (Figures 6F–6I). In summary, both MO knockdowns and CRISPR-Cas9-mediated knockout strategies produced similar vascular defects in zebrafish embryos, supporting the hypothesis that *grik5* is required for vascular integrity in zebrafish (Figure 6J).

Discussion

Figure 7 summarizes the discovery paradigm that we describe here. The overall pattern of results from PrediXcan investigations conducted in BioVU was notable and unusual. Rather than observing a single significant association between the genetically predicted expression of a gene and a phenotype, we observed that the reduced genetically determined expression of a single gene was

(D) The experimental design for fluorescent nanobead injection (0.02 μm) and microangiography.
(E) A maximum-intensity projection of the *grik5*-KD zebrafish's head (dorsal view) shows plasma leakage around the right anterior cerebral vein (ACeV) (arrowhead) and no leakage on the contralateral site (open arrowhead). Leakage is also detected in the right eye and ear (arrowheads) and absent in the contralateral, control site (open arrowheads). 2× zoomed views of boxed regions 1 and 2 are below; the "beads (plasma)" channel shows sites of leakage. The abbreviation a = anterior, and p = posterior.
(F and G) Maximum-intensity projection images of trunk intersegmental vessels (ISVs) in Tg(flk1:eGFP) embryos (green) injected with fluorescent nanobeads (magenta) in morphants (F) and genetic mutants (G). Microangiography reveals the leakage of blood plasma (nanobeads) from ISVs into the interstitial space (arrow). The open arrowhead marks no extravasation in wild-type (WT) (+/+) siblings.
(H) A graph shows the percentage of embryos with nanobead extravasation.
(I) The genotype of CRISPR-Cas9-edited *grik5*^{s03} alleles analyzed with microangiography.
(J) A schematic summarizing the effect of reduced *grik5* expression on vascular integrity. Insets in panels (B), (C), and ∈ indicate the number of embryos (out of the total analyzed) exhibiting the represented phenotype.

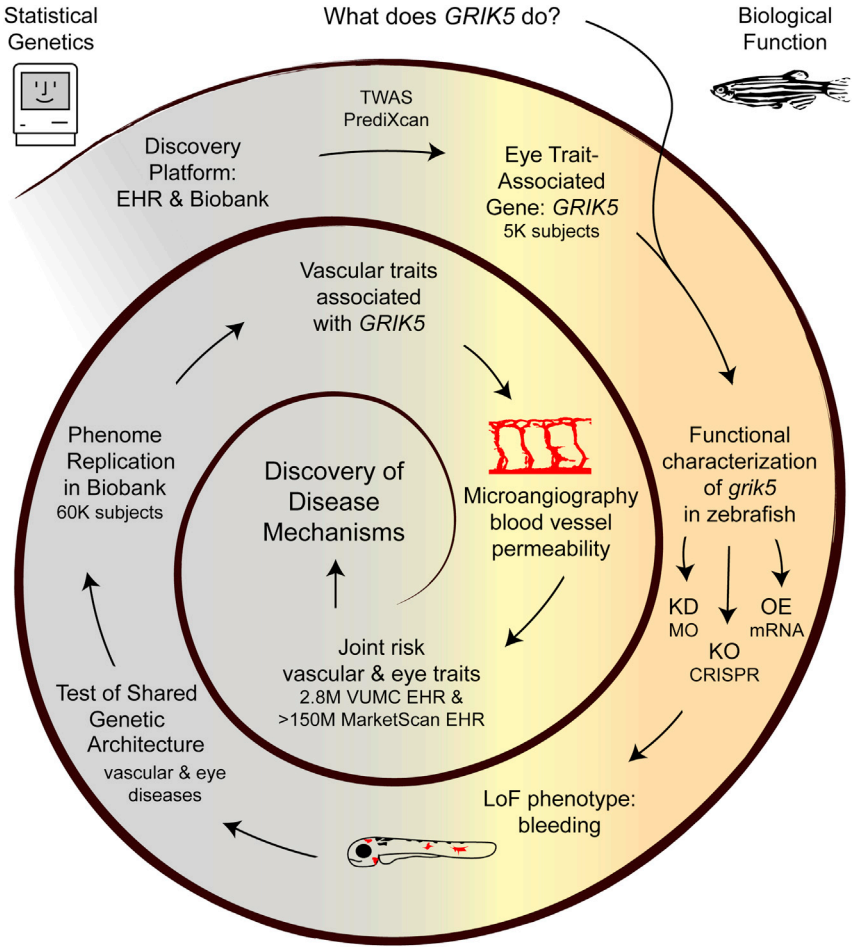

Figure 7. A Strategy for the Identification of Disease Mechanisms via EHR- and Biobank-Based Discovery Platforms and an Animal Model as a Validation Tool.
We used BioVU as an EHR- and biobank-based discovery platform. Through the transcriptome-wide association method PrediXcan, trait-associated genes were identified from BioVU. *GRIK5* displayed significant association with numerous, diverse eye disease traits and was thus selected for functional validation in a zebrafish animal model. Gain- and loss-of-function approaches, such as CRISPR-Cas9-mediated genome editing (KO) and morpholino oligonucleotide-mediated knockdown (KD), led to the discovery of phenotypes in zebrafish. These phenotypes further informed statistical analyses performed with independent phenome datasets (acquired from further biobank genotyping and publicly available genome-wide association studies [GWASs]) in order to replicate eye- and vascular-disease phenotypes implicated in zebrafish. System-level assays in zebrafish provided evidence for disease mechanisms, i.e., vascular traits, associated with *GRIK5* expression in participants from BioVU. Subsequent comorbidity studies in all Vanderbilt electronic health record (EHR) data and in more than 150 million subjects[40] extend the zebrafish findings that eye and vascular diseases are comorbid and have shared genetic architecture. The abbreviation OE = overexpression.

associated with many different eye diseases not typically considered to share a genetic mechanism. The use of a biobank was key to the observation, in that it would otherwise not have been possible to detect the association with so many different eye diseases. It was the unusual nature of this observation that created the impetus to conduct model-system validation studies. The zebrafish was a particularly attractive model system in this case because the zebrafish eye is easily studied throughout early development, and modern gene editing technologies facilitate rapid validation studies.[47–49] The initial results of studies in zebrafish *GRIK5* knockdown and knockout models highlighted the contribution of this gene to normal vascularization of the eye (both during development and over a lifetime of vascular permeability). The evidence for a high expression of *GRIK5* in the developing eye in zebrafish here, and in mice and squirrels as reported by others,[34,50] coupled with our data on the function of *grik5* in vascular patterning, suggested possible contributing mechanisms for the diverse eye diseases we observed to be associated with the GReX of *GRIK5*. We sought support for this hypothesis by looking for shared genetic architecture between eye and vascular disease; we did this through replication of the association between reduced GReX of *GRIK5* and vascular and eye disease in increased sample sizes within BioVU. We also sought support for significant comorbidity between vascular and eye disease in large-scale EHR studies. In these studies, we found that a polygenic risk score built from GWAS meta-analysis results for eye disease (macular degeneration) is significantly associated with vascular disease in BioVU, and further, that vascular and eye diseases are significantly comorbid in not only the >2.6 million subjects in Vanderbilt's SD but also in 150 million subjects from a large-scale insurance database. The ability to iterate among the discoveries in the biobank to the validation and further discoveries in the zebrafish model, then back to the biobank and the larger EHR datasets for comorbidity studies and for studies of shared genetic architecture between eye and vascular phenotypes confirmative of the zebrafish mechanistic discoveries is a key feature of our paradigm for discovery.

Our data support the hypothesis that sub-optimal vascularization of the developing human eye might indeed be expected to increase the risk of a number of different eye diseases and pathologies in adulthood as a part of a polygenic genetic architecture for such diseases. On the basis of the results of these studies, the maintenance of normal vasculature is also an obvious additional mechanism to be investigated as a contributor to late-onset human eye disease.[51] In this regard, the role of the complement system,

which genetic studies have implicated as contributing to the risk of macular degeneration,[24] in vascular permeability has intriguing parallels to the leaky blood vessels characterized in our *grik5* zebrafish model. As shown through the comorbidity studies on EHR data from very large numbers of individuals, macular degeneration is particularly likely to be comorbid with vascular disease across the board. Such observations might increase the likelihood that the contribution of the complement system to vascular permeability[52] might indeed be part of the mechanism by which genetic variation in complement-system genes increases the risk of macular degeneration. The specificity of the eye- and vascular-disease comorbidity in EHRs and the confirmation of the eye/vascular disease comorbidity results in the >150 million individuals from the MarketScan insurance claims database lend further support to the shared genetic component between eye pathologies and vascular traits.

The zebrafish studies provided key information that the zebrafish *grik5* ortholog was highly expressed during early embryonic development (day 3) of the eye, ear, and brain; these results have also been observed in the mouse eye.[34,53,54] Complete or substantial reduction of *grik5* by morpholino- or CRISPR-based editing at this time results in notably reduced numbers of blood vessels and in increased vascular permeability. The fact that many of the comorbid vascular phenotypes are very early onset (congenital anomalies of the vasculature or Mendelian diseases present from birth) and that top SNP signals from the large meta-analysis of macular degeneration had such strong enrichment in vascular diseases that substantially predate macular degeneration are both consistent with the hypotheses that early developmental and/or subtle, lifelong deficits in vascular function can increase the risk of late onset eye diseases.

The variability in expression of *GRIK5* observed in humans is likely to be more modest than the changes we characterized here in zebrafish, and the consequences to the normal vasculature of the eye of reduced expression of *GRIK5* in the ranges predicted for BioVU subjects are likely to be more modest as well. In humans, genetic variation affecting the expression of *GRIK5* would unquestionably be part of the polygenic architecture of vascular biology and diseases, and also of eye diseases, as opposed to the classification of *GRIK5* as a "major gene" capable of causing, by itself, a Mendelian eye or vascular disease, or even a "core gene" in the nomenclature of Boyle et al.[55] Consequently, we did not see, or even expect to see, the equivalent of late-onset human eye disease in zebrafish with reduced expression or knockout of *grik5*. Given the polygenic nature of late-onset eye and vascular disease, it was unclear whether we would observe any phenotype at all, and part of the initial focus was on learning the extent and distribution of *GRIK5* expression in zebrafish.

The observed association of vascular phenotypes with the expression of *grik5* in zebrafish ascribes to this glutamate receptor an important function that is congruent with what is being learned about the multifunctionality of other glutamate receptors. Our findings suggest *grik5* to be linked to the patterning and maintenance of the vasculature. In our study of both knockdown and knockout animals, we have discovered that the *grik5*-deficient embryos fail to pattern blood vessels in the brain; this failure results in fewer, smaller branches. We also found that the diameter of the blood vessels, as observed in the live embryos in a transgenic background, varied and resulted in dilations and constrictions in some regions. Microangiography directly confirmed increased blood-vessel permeability in *grik5*-depleted embryos. The observed vascular functions attributed to *GRIK5* expression generated the hypothesis that vascular biology may contribute to the development of late-onset eye disease. We showed evidence supporting this hypothesis through both shared genetic architecture and comorbidity studies in large-scale EHRs; the findings are consistent with the hypothesis that vascular disease is among the contributing factors to the development of late-onset human eye disease.

Although we are confident that the vascular consequences of reduced expression of *GRIK5* contribute to an increased risk for eye disease, we cannot be sure that this is the only mechanism by which reduced expression of *GRIK5* might contribute to eye disease. *GRIK5* is expressed in endothelial cells (Figure S3C)[38,53,54] but has also been shown to be expressed in mice in the photoreceptor cells of the retina.[56] We observed *GRIK5* to be highly expressed in the eye of 3-day-old zebrafish embryos (Figure S3). Thus, additional early or later biological consequences of reduced expression of *GRIK5* might also contribute to the development of eye disease.

Methodologically, our study presents an efficient framework for discovering gene-disease associations. Indeed, by exploiting the large number of disease phenotypes available in the biobank (1,145 phecodes tested), we find, among those with the lowest *GRIK5* expression (i.e., those closest to exhibiting *GRIK5* knockout), significantly increased risk of vascular traits and eye diseases. Thus, conditioning on genetically determined expression and conducting a phenome-wide pleiotropic scan can discover genes associated with the medical phenome. Our study illustrates the methodology, focusing on the comorbidity of the vascular conditions with eye phenotypes because of the number of eye diseases implicated, but the methodology is broadly applicable to the discovery of associations and corresponding mechanisms.

The research questions asked in biobank-based studies are fundamentally similar to the questions asked in model-system knockout studies—"What does this gene do?" rather than "What genes and variants contribute to this disease?"—and therefore, biobank-based studies have a natural parallel to model system studies, leading to more obvious ways of iterating among human and model system studies. Although the discovery of genes contributing to common human diseases and the discovery of

the mechanisms by which this genetic variation affects risk of disease will remain a challenging endeavor for the foreseeable future, the combination of modern gene editing technologies coupled to large-scale data integration approaches applied to biobanks offers a model for pursuing this goal.

Supplemental Data

Supplemental Data can be found with this article online at https://doi.org/10.1016/j.ajhg.2019.01.017.

Acknowledgments

N.J.C., E.W.K., and J.C.D. gratefully acknowledge financial support from the National Institutes of Health (NIH) (R01MH113362), (N.J.C. [U01HG009086]); D.M.R. would like to thank the NIH (P50 GM115305). J.C.D. and L.B. were supported by R01-LM010685 from the National Library of Medicine. G.U. was supported by Vanderbilt International Scholar Program (VISP) and American Heart Association (AHA) predoctoral fellowship #15PRE22940041. D.S.L. was supported by National Research Service Award (NRSA) F31DE022226 and the T32HD007502 Training Program in Developmental Biology. E.R.G. is grateful to the president and fellows of Clare Hall, University of Cambridge, for a stimulating intellectual home away from home and their generous support during the Lent and Easter terms. We thank the Vanderbilt University Nikon Center of Excellence (NCoE), maintained by Vanderbilt University Cell Imaging Shared Resource (CISR), for technical support with confocal imaging and data analysis, and we thank Cory Guthrie and the Zebrafish Core Facility for fish care.

Declaration of Interests

The authors declare no competing interests.

Received: October 30, 2018
Accepted: January 29, 2019
Published: February 28, 2019

Web Resources

CHOPCHOP, http://chopchop.cbu.uib.no/
Ensembl, http://useast.ensembl.org/index.html
GTEx, https://gtexportal.org
NCBI Gene Portal, https://www.ncbi.nlm.nih.gov/gene/
Phecodes, https://phewascatalog.org/phecodes

References

1. Gamazon, E.R., Wheeler, H.E., Shah, K.P., Mozaffari, S.V., Aquino-Michaels, K., Carroll, R.J., Eyler, A.E., Denny, J.C., Nicolae, D.L., Cox, N.J., Im, H.K.; and GTEx Consortium (2015). A gene-based association method for mapping traits using reference transcriptome data. Nat. Genet. *47*, 1091–1098.
2. Gusev, A., Lee, S.H., Trynka, G., Finucane, H., Vilhjálmsson, B.J., Xu, H., Zang, C., Ripke, S., Bulik-Sullivan, B., Stahl, E., et al.; Schizophrenia Working Group of the Psychiatric Genomics Consortium; SWE-SCZ Consortium; Schizophrenia Working Group of the Psychiatric Genomics Consortium; and SWE-SCZ Consortium (2014). Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. Am. J. Hum. Genet. *95*, 535–552.
3. Gamazon, E.R., Cox, N.J., and Davis, L.K. (2014). Structural architecture of SNP effects on complex traits. Am. J. Hum. Genet. *95*, 477–489.
4. Iyer, V., Boroviak, K., Thomas, M., Doe, B., Riva, L., Ryder, E., and Adams, D.J. (2018). No unexpected CRISPR-Cas9 off-target activity revealed by trio sequencing of gene-edited mice. PLoS Genet. *14*, e1007503.
5. Jao, L.E., Wente, S.R., and Chen, W. (2013). Efficient multiplex biallelic zebrafish genome editing using a CRISPR nuclease system. Proc. Natl. Acad. Sci. USA *110*, 13904–13909.
6. Hwang, W.Y., Fu, Y., Reyon, D., Maeder, M.L., Tsai, S.Q., Sander, J.D., Peterson, R.T., Yeh, J.R., and Joung, J.K. (2013). Efficient genome editing in zebrafish using a CRISPR-Cas system. Nat. Biotechnol. *31*, 227–229.
7. Cong, L., Ran, F.A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P.D., Wu, X., Jiang, W., Marraffini, L.A., and Zhang, F. (2013). Multiplex genome engineering using CRISPR/Cas systems. Science *339*, 819–823.
8. Klarin, D., Emdin, C.A., Natarajan, P., Conrad, M.F., Kathiresan, S.; and INVENT Consortium (2017). Genetic analysis of venous thromboembolism in UK Biobank identifies the ZFPM2 locus and implicates obesity as a causal risk factor. Circ Cardiovasc Genet *10*, e001643.
9. Wei, W.Q., and Denny, J.C. (2015). Extracting research-quality phenotypes from electronic health records to support precision medicine. Genome Med. *7*, 41.
10. Wolford, B.N., Willer, C.J., and Surakka, I. (2018). Electronic health records: The next wave of complex disease genetics. Hum. Mol. Genet. *27* (R1), R14–R21.
11. Kho, A.N., Pacheco, J.A., Peissig, P.L., Rasmussen, L., Newton, K.M., Weston, N., Crane, P.K., Pathak, J., Chute, C.G., Bielinski, S.J., et al. (2011). Electronic medical records for genetic research: Results of the eMERGE consortium. Sci. Transl. Med. *3*, 79re1.
12. McCarty, C.A., Chisholm, R.L., Chute, C.G., Kullo, I.J., Jarvik, G.P., Larson, E.B., Li, R., Masys, D.R., Ritchie, M.D., Roden, D.M., et al.; eMERGE Team (2011). The eMERGE Network: A consortium of biorepositories linked to electronic medical records data for conducting genomic studies. BMC Med. Genomics *4*, 13.
13. Gottesman, O., Kuivaniemi, H., Tromp, G., Faucett, W.A., Li, R., Manolio, T.A., Sanderson, S.C., Kannry, J., Zinberg, R., Basford, M.A., et al.; eMERGE Network (2013). The Electronic Medical Records and Genomics (eMERGE) Network: Past, present, and future. Genet. Med. *15*, 761–771.
14. Denny, J.C., Crawford, D.C., Ritchie, M.D., Bielinski, S.J., Basford, M.A., Bradford, Y., Chai, H.S., Bastarache, L., Zuvich, R., Peissig, P., et al. (2011). Variants near FOXE1 are associated with hypothyroidism and other thyroid conditions: Using electronic medical records for genome- and phenome-wide studies. Am. J. Hum. Genet. *89*, 529–542.
15. Denny, J.C., Bastarache, L., Ritchie, M.D., Carroll, R.J., Zink, R., Mosley, J.D., Field, J.R., Pulley, J.M., Ramirez, A.H., Bowton, E., et al. (2013). Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. Nat. Biotechnol. *31*, 1102–1110.
16. Roden, D.M., Pulley, J.M., Basford, M.A., Bernard, G.R., Clayton, E.W., Balser, J.R., and Masys, D.R. (2008). Development

of a large-scale de-identified DNA biobank to enable personalized medicine. Clin. Pharmacol. Ther. *84*, 362–369.
17. Denny, J.C., Ritchie, M.D., Basford, M.A., Pulley, J.M., Bastarache, L., Brown-Gentry, K., Wang, D., Masys, D.R., Roden, D.M., and Crawford, D.C. (2010). PheWAS: Demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. Bioinformatics *26*, 1205–1210.
18. Hall, M.A., Verma, A., Brown-Gentry, K.D., Goodloe, R., Boston, J., Wilson, S., McClellan, B., Sutcliffe, C., Dilks, H.H., Gillani, N.B., et al. (2014). Detection of pleiotropy through a phenome-wide association study (PheWAS) of epidemiologic data as part of the Environmental Architecture for Genes Linked to Environment (EAGLE) study. PLoS Genet. *10*, e1004678.
19. Pendergrass, S.A., Brown-Gentry, K., Dudek, S., Frase, A., Torstenson, E.S., Goodloe, R., Ambite, J.L., Avery, C.L., Buyske, S., Bůžková, P., et al. (2013). Phenome-wide association study (PheWAS) for detection of pleiotropy within the Population Architecture using Genomics and Epidemiology (PAGE) Network. PLoS Genet. *9*, e1003087.
20. GTEx Consortium (2015). Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science *348*, 648–660.
21. Kleinbaum, D., Kupper, L., Nizam, A., and Muller, K. (1998). Applied Regression Analysis and Other Multivariable Methods (Pacific Grove, CA: Brooks/Cole).
22. Wainberg, M., Sinnott-Armstrong, N., Mancuso, N., Barbeira, A.N., Knowles, D.A., Golan, D., Ermel, R., Ruusalepp, A., Quertermous, T., Hao, K., et al. (2018). Transcriptome-wide association studies: Opportunities and challenges. Biorxiv. https://doi.org/10.1101/206961.
23. Stegle, O., Parts, L., Piipari, M., Winn, J., and Durbin, R. (2012). Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. Nat. Protoc. *7*, 500–507.
24. Fritsche, L.G., Igl, W., Bailey, J.N., Grassmann, F., Sengupta, S., Bragg-Gresham, J.L., Burdon, K.P., Hebbring, S.J., Wen, C., Gorski, M., et al. (2016). A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. Nat. Genet. *48*, 134–143.
25. Battle, A., Mostafavi, S., Zhu, X., Potash, J.B., Weissman, M.M., McCormick, C., Haudenschild, C.D., Beckman, K.B., Shi, J., Mei, R., et al. (2014). Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. Genome Res. *24*, 14–24.
26. GTEx Consortium (2013). The Genotype-Tissue Expression (GTEx) project. Nat. Genet. *45*, 580–585.
27. Flintoft, L. (2014). Disease genetics: Phenome-wide association studies go large. Nat. Rev. Genet. *15*, 2.
28. Montero-Balaguer, M., Lang, M.R., Sachdev, S.W., Knappmeyer, C., Stewart, R.A., De La Guardia, A., Hatzopoulos, A.K., and Knapik, E.W. (2006). The mother superior mutation ablates foxd3 activity in neural crest progenitor cells and depletes neural crest derivatives in zebrafish. Dev. Dyn. *235*, 3199–3212.
29. Melville, D.B., Montero-Balaguer, M., Levic, D.S., Bradley, K., Smith, J.R., Hatzopoulos, A.K., and Knapik, E.W. (2011). The feelgood mutation in zebrafish dysregulates COPII-dependent secretion of select extracellular matrix proteins in skeletal morphogenesis. Dis. Model. Mech. *4*, 763–776.
30. Sarmah, S., Barrallo-Gimeno, A., Melville, D.B., Topczewski, J., Solnica-Krezel, L., and Knapik, E.W. (2010). Sec24D-dependent transport of extracellular matrix proteins is required for zebrafish skeletal morphogenesis. PLoS ONE *5*, e10367.
31. Levic, D.S., Minkel, J.R., Wang, W.D., Rybski, W.M., Melville, D.B., and Knapik, E.W. (2015). Animal model of Sar1b deficiency presents lipid absorption deficits similar to Anderson disease. J. Mol. Med. (Berl.) *93*, 165–176.
32. Varshney, G.K., Pei, W., LaFave, M.C., Idol, J., Xu, L., Gallardo, V., Carrington, B., Bishop, K., Jones, M., Li, M., et al. (2015). High-throughput gene targeting and phenotyping in zebrafish using CRISPR/Cas9. Genome Res. *25*, 1030–1042.
33. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. *38*, 904–909.
34. Haumann, I., Junghans, D., Anstötz, M., and Frotscher, M. (2017). Presynaptic localization of GluK5 in rod photoreceptors suggests a novel function of high affinity glutamate receptors in the mammalian retina. PLoS ONE *12*, e0172967.
35. Young, S.Z., Taylor, M.M., and Bordey, A. (2011). Neurotransmitters couple brain activity to subventricular zone neurogenesis. Eur. J. Neurosci. *33*, 1123–1132.
36. Pin, J.P., and Duvoisin, R. (1995). The metabotropic glutamate receptors: Structure and functions. Neuropharmacology *34*, 1–26.
37. White, R.J., Collins, J.E., Sealy, I.M., Wali, N., Dooley, C.M., Digby, Z., Stemple, D.L., Murphy, D.N., Billis, K., Hourlier, T., et al. (2017). A high-resolution mRNA expression time course of embryonic development in zebrafish. eLife *6*, e30860.
38. Kasper, D.M., Moro, A., Ristori, E., Narayanan, A., Hill-Teran, G., Fleming, E., Moreno-Mateos, M., Vejnar, C.E., Zhang, J., Lee, D., et al. (2017). MicroRNAs establish uniform traits during the architecture of vertebrate embryos. Dev. Cell *40*, 552–565.e5.
39. Yu, X., Lyu, D., Dong, X., He, J., and Yao, K. (2014). Hypertension and risk of cataract: A meta-analysis. PLoS ONE *9*, e114012.
40. Blair, D.R., Lyttle, C.S., Mortensen, J.M., Bearden, C.F., Jensen, A.B., Khiabanian, H., Melamed, R., Rabadan, R., Bernstam, E.V., Brunak, S., et al. (2013). A nondegenerate code of deleterious variants in Mendelian loci contributes to complex disease risk. Cell *155*, 70–80.
41. Gormley, P., Anttila, V., Winsvold, B.S., Palta, P., Esko, T., Pers, T.H., Farh, K.H., Cuenca-Leon, E., Muona, M., Furlotte, N.A., et al.; International Headache Genetics Consortium (2016). Meta-analysis of 375,000 individuals identifies 38 susceptibility loci for migraine. Nat. Genet. *48*, 856–866.
42. Finucane, H.K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.R., Anttila, V., Xu, H., Zang, C., Farh, K., et al.; ReproGen Consortium; Schizophrenia Working Group of the Psychiatric Genomics Consortium; and RACI Consortium (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat. Genet. *47*, 1228–1235.
43. Gamazon, E.R., Segrè, A.V., van de Bunt, M., Wen, X., Xi, H.S., Hormozdiari, F., Ongen, H., Konkashbaev, A., Derks, E.M., Aguet, F., et al.; GTEx Consortium (2018). Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation. Nat. Genet. *50*, 956–967.

44. Montero-Balaguer, M., Swirsding, K., Orsenigo, F., Cotelli, F., Mione, M., and Dejana, E. (2009). Stable vascular connections and remodeling require full expression of VE-cadherin in zebrafish embryos. PLoS ONE *4*, e5772.
45. Butler, M.G., Gore, A.V., and Weinstein, B.M. (2011). Zebrafish as a model for hemorrhagic stroke. Methods Cell Biol. *105*, 137–161.
46. Choi, J., Dong, L., Ahn, J., Dao, D., Hammerschmidt, M., and Chen, J.N. (2007). FoxH1 negatively modulates flk1 gene expression and vascular formation in zebrafish. Dev. Biol. *304*, 735–744.
47. Luderman, L.N., Unlu, G., and Knapik, E.W. (2017). Zebrafish developmental models of skeletal diseases. Curr. Top. Dev. Biol. *124*, 81–124.
48. Vacaru, A.M., Unlu, G., Spitzner, M., Mione, M., Knapik, E.W., and Sadler, K.C. (2014). In vivo cell biology in zebrafish - providing insights into vertebrate development and disease. J. Cell Sci. *127*, 485–495.
49. Unlu, G., Levic, D.S., Melville, D.B., and Knapik, E.W. (2014). Trafficking mechanisms of extracellular matrix macromolecules: Insights from vertebrate development and human diseases. Int. J. Biochem. Cell Biol. *47*, 57–67.
50. Lindstrom, S.H., Ryan, D.G., Shi, J., and DeVries, S.H. (2014). Kainate receptor subunit diversity underlying response diversity in retinal off bipolar cells. J. Physiol. *592*, 1457–1477.
51. Yanagi, M., Kawasaki, R., Wang, J.J., Wong, T.Y., Crowston, J., and Kiuchi, Y. (2011). Vascular risk factors in glaucoma: A review. Clin. Experiment. Ophthalmol. *39*, 252–258.
52. auf dem Keller, U., Prudova, A., Eckhard, U., Fingleton, B., and Overall, C.M. (2013). Systems-level analysis of proteolytic events in increased vascular permeability and complement activation in skin inflammation. Sci. Signal. *6*, rs2.
53. Fukushima, Y., Okada, M., Kataoka, H., Hirashima, M., Yoshida, Y., Mann, F., Gomi, F., Nishida, K., Nishikawa, S., and Uemura, A. (2011). Sema3E-PlexinD1 signaling selectively suppresses disoriented angiogenesis in ischemic retinopathy in mice. J. Clin. Invest. *121*, 1974–1985.
54. Matthaei, M., Hu, J., Meng, H., Lackner, E.M., Eberhart, C.G., Qian, J., Hao, H., and Jun, A.S. (2013). Endothelial cell whole genome expression analysis in a mouse model of early-onset Fuchs' endothelial corneal dystrophy. Invest. Ophthalmol. Vis. Sci. *54*, 1931–1940.
55. Boyle, E.A., Li, Y.I., and Pritchard, J.K. (2017). An expanded view of complex traits: From polygenic to omnigenic. Cell *169*, 1177–1186.
56. Karunakaran, D.K.P., Al Seesi, S., Banday, A.R., Baumgartner, M., Olthof, A., Lemoine, C., Măndoiu, I.I., and Kanadia, R.N. (2016). Network-based bioinformatics analysis of spatio-temporal RNA-Seq data reveals transcriptional programs underpinning normal and aberrant retinal development. BMC Genomics *17* (*Suppl 5*), 495.