Genetic Variants with Significant Association to Age-Related Macular Degeneration (AMD) and their Role in the Regulation of Gene Expression

Dissertation
zur Erlangung des Doktorgrades
der Biomedizinischen Wissenschaften
(Dr. rer. physiol.)
der
Fakultät für Medizin
der Universität Regensburg

vorgelegt von
Tobias Strunz
aus
Marktredwitz

im Jahr
2020

Genetic Variants with Significant Association to Age-Related Macular Degeneration (AMD) and their Role in the Regulation of Gene Expression

Dissertation
zur Erlangung des Doktorgrades
der Biomedizinischen Wissenschaften
(Dr. rer. physiol.)
der
Fakultät für Medizin
der Universität Regensburg

vorgelegt von
Tobias Strunz
aus
Marktredwitz

im Jahr
2020

Dekan:	Prof. Dr. Dirk Hellwig
Betreuer:	Prof. Dr. Bernhard H.F. Weber
Tag der mündlichen Prüfung:	02.12.2020

Parts of this work have already been published in peer-reviewed journals in an open access format:

Strunz T, Grassmann F, Gayán J, Nahkuri S, Souza-Costa D, Maugeais C, Fauser S, Nogoceke E, Weber BHF (2018) A mega-analysis of expression quantitative trait loci (eQTL) provides insight into the regulatory architecture of gene expression variation in liver. Sci Rep 8: 5865.

Strunz T, Lauwen S, Kiel C, den Hollander A, Weber BHF (2020) A transcriptome-wide association study based on 27 tissues identifies 106 genes potentially relevant for disease pathology in age-related macular degeneration. Sci Rep 10: 1584.

Strunz T, Kiel C, Grassmann F, Ratnapriya R, Kwicklis M, Karlstetter M, Fauser S, Swaroop A, Arend N, Langmann T, Wolf A, Weber BHF (2020) A mega-analysis of expression quantitative trait loci in retinal tissue. PLoS Genet 16: e1008934.

Kiel C, Berber P, Karlstetter M, Aslanidis A, Strunz T, Langmann T, Grassmann F, Weber BHF (2020) A Circulating MicroRNA Profile in a Laser-Induced Mouse Model of Choroidal Neovascularization. Int J Mol Sci 21(8): E2689.

Nebauer CA, Kiel C, Strunz T, Stelzl S, Weber BHF (2020) Interaction of age-related macular degeneration (AMD) associated loci influences gene expression in liver. *In preparation*.

Table of contents

Zusammenfassung

Genomweite Assoziationsstudien (GWAS) haben dazu beigetragen eine Vielzahl genetischer Varianten zu identifizieren, die mit dem Risiko komplexer Krankheiten assoziiert sind. Die überhaupt erste erfolgreiche GWAS wurde von Klein et al. im Jahre 2005 durchgeführt und detektierte eine Assoziation genetischer Varianten im Komplement Faktor H (*CFH*) Gen mit der altersabhängigen Makuladegeneration (AMD). AMD ist eine komplexe Netzhauterkrankung und weltweit eine der häufigsten Ursachen für Sehbeeinträchtigungen und Erblindungen. Es wird angenommen, dass sowohl Umweltfaktoren, insbesondere Altern und Rauchen, als auch die genetische Prädisposition das Krankheitsrisiko wesentlich bestimmen. Der Einfluss genetischer Faktoren wurde auf 40 - 71 % geschätzt. Bisher ist nur wenig über die Ätiologie der AMD bekannt, obwohl die aktuellste GWAS von Fritsche et al. (2016) bereits 52 unabhängige Signale in 34 mit AMD-assoziierten Loci aufdecken konnte.

Die meisten der AMD-assoziierten Varianten befinden sich in nicht-kodierenden intergenischen oder intronischen Bereichen des Genoms, wobei eine funktionelle Abklärung eine große Herausforderung darstellt. Solche Varianten könnten sich auf die Regulation der Genexpression auswirken. Aus diesem Grund bestand das Ziel dieser Arbeit darin, die Pathogenese der AMD im Kontext von Effekten auf die Regulation der Genexpression zu betrachten.

In einem ersten Ansatz wurden „expression quantitative trait loci“ (eQTLs) in Lebergewebe untersucht. Dafür wurden Genotyp- und Genexpressionsdaten von vier unabhängigen Studien in einer zusammenführenden Analyse betrachtet. Alle miteinbezogenen Studien und Proben durchliefen ein eigens hierfür entwickelten Datenverarbeitungsprotokoll, das vor allem auf die Identifikation reproduzierbarer Effekte fokussiert war. Insgesamt wurden Daten von 588 Individuen untersucht und es konnten 7.612 Gene gefunden werden, die signifikant (Q-Wert < 0,05) von genetischen Varianten reguliert werden. Bemerkenswerterweise zeigten sich 15 dieser Gene von AMD-assoziierten Varianten beeinflusst und eine vergleichende Analyse ergab, dass diese Gene vor allem in Zusammenhang mit Prozessen des angeborenen Komplementsystems und des Metabolismus von Lipoproteinen stehen.

In einem zweiten Projekt wurden die Daten der „Genotype-Tissue Expression“ (GTEx) Datenbank ausgewertet, um die initialen Untersuchungen auf eine Vielzahl an

Geweben zu erweitern. GTEx beinhaltet Daten zu 48 unterschiedlichen Geweben bzw. Zelltypen, die von bis zu 500 Spendern zur Verfügung stehen. Die eQTL Analyse ermöglichte es, eine neue Hypothese bezüglich genregulatorischer Effekte in einem der am stärksten mit AMD assoziierten Loci aufzustellen. So zeigte sich, dass genetische Varianten innerhalb des *ARMS2-HTRA1* Locus Gene regulieren, die sich an unterschiedlichsten Positionen des Genoms befinden und deren Genprodukte größtenteils an Immunsystem-bezogenen Prozessen teilnehmen. Zusätzlich zu den bioinformatischen Untersuchungen wurden *in vitro* Experimente durchgeführt, um die erarbeitete Hypothese zu valideren. In einer ersten Untersuchung wurde dazu eine Deletion innerhalb des *ARMS2-HTRA1* Locus herbeigeführt und betrachtet, ob dies die Genexpression der vorhergesagten Zielgene beeinflusst. Außerdem wurde in weiteren Experimenten die Genexpression innerhalb des *ARMS2-HTRA1* Locus gezielt verstärkt. Beide Ansätze konnten jedoch in den initialen Experimenten die aufgestellte Hypothese in HEK293T Zellen nicht bestätigen.

In einem weiteren Projekt wurde eine eQTL Analyse von 314 gesunden retinalen Gewebeproben durchgeführt, die von drei unabhängigen Instituten gesammelt wurden. Dabei konnten 9.733 Gene identifiziert werden, die signifikant von genetischen Varianten reguliert werden (Q-Wert < 0,05). Diese zusammenfassende Studie ermöglichte zum ersten Mal eine Analyse der Genexpressionsregulation in ausschließlich gesunden Netzhautproben. Interessanterweise zeigten jedoch nur 7 der 34 AMD-assoziierten Loci eQTL in der Retina, obwohl man davon ausgehen muss, dass dieses Gewebe ein Ort der primären/sekundären Pathologie der AMD ist.

Aus diesem Grund zielte das abschließende Projekt darauf ab, ein zusammenhängendes Bild der Genexpressionsregulation im Lichte der AMD Genetik zu erhalten. Dafür wurde eine transkriptomweite Assoziationsstudie (TWAS) durchgeführt, die die Genotypen von 16.144 AMD Patienten und von 17.832 gesunden Vergleichspersonen aus dem Datensatz des internationalen AMD Genomics Consortium (IAMDGC) miteinschloss. Für alle Proben wurde die individuelle Genexpression in 27 Geweben vorhergesagt und mit dem AMD-Status verglichen. Insgesamt konnten 106 Gene identifiziert werden, die sich in mindestens einem Gewebe mit der AMD assoziiert zeigten. Diese Analyse deckte genregulatorische Effekte in 25 der 34 AMD-assoziierten Loci auf.

Zusammengefasst zeigen die Ergebnisse dieser Arbeit, dass die Regulation der Genexpression ein häufiges Phänomen in AMD-assoziierten Loci darstellt. Die Resultate verdeutlichen eine Beteiligung systemischer Prozesse, wie zum Beispiel des Komplementsystems und der Blut-Lipoproteine, an der AMD Pathogenese. Außerdem konnte die Analyse AMD-assoziierter Gene zeigen, dass diese nicht ausschließlich in der Retina, sondern häufig ubiquitär reguliert werden. So ist es wahrscheinlich, dass die zugrundeliegenden Prozesse der AMD Pathogenese im gesamten Körper ablaufen, wobei es offensichtlich fast ausschließlich zur Expression eines Phänotyps bevorzugt in der Netzhaut kommt.

Summary

Genome-wide association studies (GWAS) have led to the identification of a plethora of risk-associated genetic variants for a multitude of complex diseases. The very first GWAS was performed by Klein et al. in the year 2005 and identified variants in the complement factor H (*CFH*) gene to be associated with age-related macular degeneration (AMD). AMD is a complex eye disease and one of the most common causes of visual impairments and blindness worldwide. It is widely accepted that environmental factors, especially advanced age and smoking, as well as genetic factors contribute substantially to disease risk. Remarkably, the influence of genetics was estimated to be as high as 40-71 %. However, little is known about AMD aetiology, although the latest GWAS performed by Fritsche et al. (2016) revealed 52 independent signals distributed over 34 loci to be associated with AMD.

Most of the AMD-associated variants are located in non-coding intergenic or intronic regions of the genome, where functional annotation presents a major challenge. However, these variants may play an important role in the regulation of gene expression. The aim of this thesis was therefore to examine the pathogenesis of AMD in the context of gene expression regulation.

A first approach investigated expression quantitative trait loci (eQTL) in liver tissue. Thus, genotype and gene expression data from four independent studies were combined to enable a comprehensive analysis. All samples and studies underwent an especially developed data processing protocol, which applied stringent filter to exclusively allow the detection of highly valid associations. Altogether 588 samples were included and 7,612 genetically regulated genes (Q-Value < 0.05) have been identified. Remarkably, 15 of these are influenced by AMD-associated variants and a comparative analysis reinforced the notion that the initial complement system and lipoprotein metabolism play a role in AMD pathogenesis.

In a second project, the Genotype-Tissue Expression (GTEx) database was explored to extend the initial investigations to a variety of tissues. GTEx contains data on 48 different tissues or cell types available from up to 500 donors. The eQTL analysis enabled a new hypothesis regarding gene expression regulatory effects in one of the most significant AMD-associated loci. It was shown that genetic variants within the *ARMS2-HTRA1* locus regulate immune system related genes throughout the whole

genome. In addition to the bioinformatics studies, *in vitro* experiments were conducted to validate the developed hypothesis. First, a large genomic deletion within the *ARMS2-HTRA1* locus was introduced to assess potential consequences on the expression of bioinformatical predicted target genes. In a second approach, gene expression within the locus was enhanced by targeted application of transcription activation factors. Nevertheless, both strategies were not able to confirm the generated hypothesis in HEK293T cells in the initial experiments.

The next project included the comprehensive analysis of eQTL in 314 healthy retinal tissue samples collected from three independent study sites. Altogether, 9,733 genetically regulated genes (Q-value < 0.05) were identified, which allowed insights in gene expression regulation of exclusively healthy retinal tissues for the very first time. Interestingly, only 7 of 34 AMD-associated loci revealed eQTL effects in retina although one must assume that this tissue is a site of the primary/secondary pathology of AMD

Therefore, the last project of this thesis aimed at obtaining a comprehensive view on gene expression regulation in the light of AMD genetics. A transcriptome wide association study (TWAS) was performed, which included the genotypes of 16,144 late-stage AMD cases and 17,832 healthy controls from the International AMD Genomics Consortium (IAMDGC). For all these individuals, gene expression was imputed in 27 tissues and analysed in regard to the respective AMD status. This analysis discovered 106 genes, which expression was found to be associated with AMD genetics in at least one tissue. Regulatory effects on gene expression were identified in 25 of the 34 AMD-associated loci.

Taken together, this work revealed that gene expression regulation is common in AMD-associated loci. The identified genes reinforce the notion that systemic processes like the complement system or blood lipid levels seem to be relevant for AMD pathology. Furthermore, expression of genes associated with AMD is not restricted to retinal tissue, but instead is rather ubiquitous suggesting processes underlying AMD pathology to be of systemic nature, although the pathological phenotype occurs in the eye.

1 Introduction

1.1 Age-related macular degeneration

Age-related macular degeneration (AMD) is one of the most common causes of blindness in industrialised countries. The worldwide prevalence of AMD reaches 8.67 % in the age group of 30 – 97 years. It is further estimated that the number of AMD cases increases from recently around 196 million to 288 million by the year 2040 [1]. The clinical phenotype of AMD manifests in the retina and can be broadly divided into three disease stages progressing from early AMD to intermediate AMD and finally to the late stage forms [2]. In healthy individuals, visual perception is accomplished in the retina by a complex interplay of hierarchically connected cell types, initiated by the photoreceptors, the primary recipients of photons. This process requires a high metabolic activity und needs a well-regulated support system, which comprises the mono-layered retinal pigment epithelium (RPE) and the blood supply, the choroid including the choriocapillaris (**Figure 1 A**).

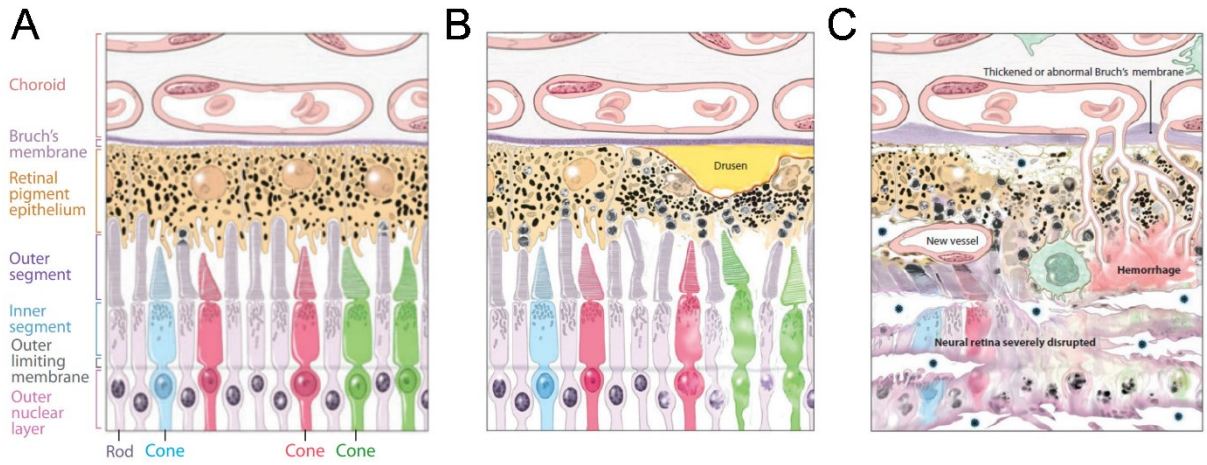

Figure 1: Schematic overview of the human retina and pathological changes caused by AMD. (A) Schematic overview of healthy retinal tissue, supported by the retinal pigment epithelium (RPE) and the chorid. (B) Changes in the retina and Drusen formation caused by early AMD. (C) Schematic changes in a late-stage AMD affected eye. Choroidal neovascularization is characterised by new blood vessels growing from the choroid into the RPE. The following hemorrhages initiate photoreceptor cell death and cause perturbation of the retinal layers. (Figure modified from Swaroop et al. (2009) [3])

Early AMD is accompanied by the formation of extracellular protein-lipid aggregates, known as Drusen, between the RPE and Bruch`s membrane, a five-layered extracellular matrix structure (**Figure 1 B**). The lesions primarily occur around the macula, a region near the centre of the retina, which contains mainly cone photoreceptor cells and is responsible for central, high resolution colour vision. Nevertheless, early AMD is the most common and the least severe form of AMD and

is usually not recognised by the patients. Subsequently, Drusen grow in size and pigmentary abnormalities accumulate, resulting in the progress from the early form to the intermediate AMD, which still only leads to minor visual impairments such as the beginning loss of central vision. Finally, the late-stage AMD lesions present as two distinct forms, which can occur separately or combined, namely geographic atrophy (GA) and choroidal neovascularization (CNV). In eyes affected by GA, Drusen growth continues and severely hinders RPE function, which in-turn causes severe damage to the photoreceptors. GA is slowly progressing over years and progressively impairs vision. In contrast, CNV, is characterised by the formation of new fragile blood vessels growing from the choroid into the RPE (**Figure 1 C**). This leads to rapid loss of vision, caused by bleedings into the retinal and subretinal space. So far, only treatment options for CNV are available through ocular injection of inhibitors targeting the vascular endothelial growth factor (VEGF). However, this treatment exclusively addresses symptoms of the disease but cannot cure the phenotype [4,5].

While the main manifestations of AMD affect the back of the eye, several studies investigated AMD patients in regard to extraocular phenotypes and potential biomarkers. Such studies showed lower complement Factor H (CFH) levels in the serum of AMD patients, which is supposed to result in an increased activation of the innate immune system [6,7]. Furthermore, elevated high-density lipoprotein (HDL) levels were found to be associated with late-stage AMD [8,9].

In general, little is known about AMD aetiology although three main factors seem to be generally accepted as AMD risk contributors: (1) Advanced age, (2) environmental factors, particularly smoking, and (3) genetic predisposition [10–12]. The interplay of environmental risk factors and genetic influences makes AMD to a so-called complex disease.

1.2 The genetics of AMD

Genetic predisposition to AMD was first investigated in the early twenty-first century. Remarkably, a twin study by Seddon et al. (2005) estimated the genetic contribution to AMD to be as high as 71 % [13]. As AMD shows a high prevalence in the general population, it is assumed to be influenced by many common genetic variants together contributing to disease risk [14].

A ground-breaking development in the research of complex diseases was the rise of large-scale genome-wide association studies (GWAS). GWAS investigate genetic variation in hundreds to thousands of individuals and aim to identify statistically significant changes in allele frequencies between a study population and a population of control individuals. The identified genetic variants are then assumed to be associated with the disease or phenotype of interest. GWAS are a hypothesis free approach and are well suited to identify unknown genomic loci. The first successful GWAS was performed by Klein et al. in 2005 and included 96 patients and 50 controls [15]. Remarkably, this study identified a strong association of the *CFH* locus on chromosome 1q31 with AMD and therefore raised the hypothesis of the complement system being involved in AMD pathogenesis. Over time, GWAS steadily increased in sample size and consequently identified variants with smaller effect sizes [16,17]. The most recent GWAS regarding late-stage AMD was conducted by the International AMD Genomics Consortium (IAMDGC) and included 16,144 patients and 17,832 controls [18]. This GWAS identified 52 independent genetic variants at 34 loci associated with AMD at genome wide significance (P-value < 5.0×10^{-08}). Fritsche et al. (2016) validated the findings in the *CFH* locus (**Figure 2 A**) and further demonstrated 7 additional independent hits (IHs) located on chromosome 1q31 - mostly representing rare variants with minor allele frequency (MAF) below 1 %. The 1q31 locus compromises, besides *CFH*, five *CFH*-related genes (*CFHR1* – *CFHR5*). These share high sequence similarities with *CFH* and are thought to compete with CFH for binding the central complement component C3 [19].

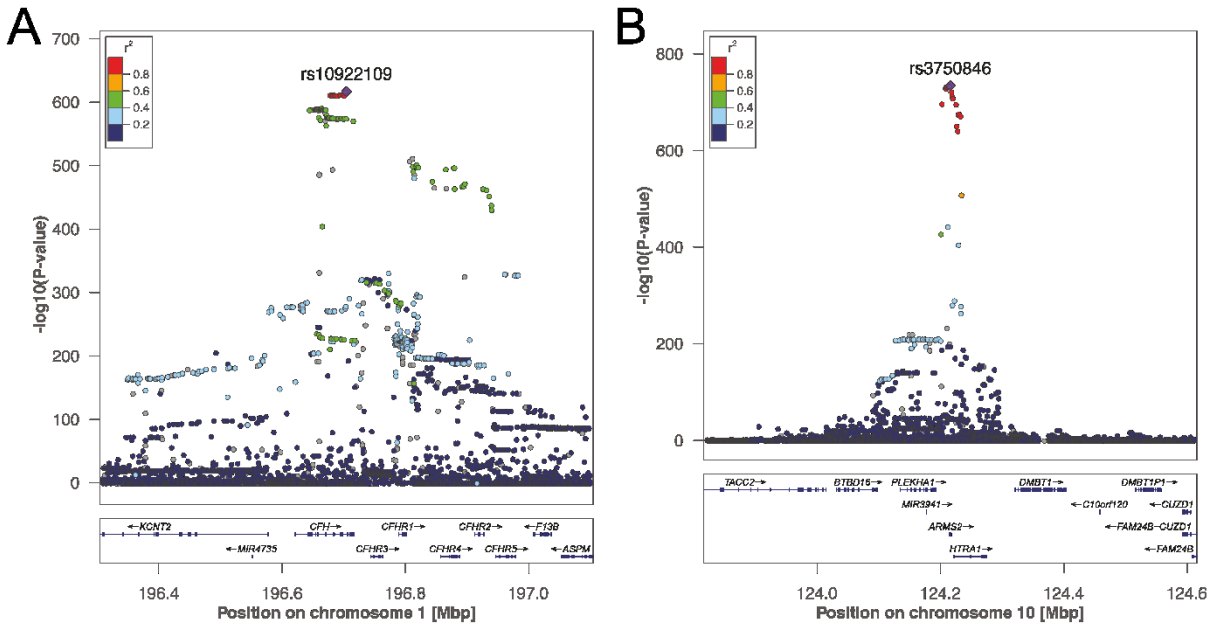

Figure 2: LocusZoom plot of the most significant AMD-associated loci.
Fritsche et al. (2016) conducted a GWAS including 16,144 AMD patients and 17,832 healthy controls. The association signals within the two most signifcant AMD-associated loci were plotted using LocusZoom [20] and the GWAS summary statistics [18]. Each dot represents one genetic variant and is plotted according to its AMD-association displayed by its -log10(P-value). Linkage disequlibrum (LD) with the respective lead variant (purple) is symbolised by a color range from red (R2 = 1) to dark blue (R2 = 0). Genes located within the locus are depicted on the bottom. (A) LocusZoom plot of the *CFH* locus (chromosome 1q31). (B) LocusZoom plot of the *ARMS2-HTRA1* locus (chromosome 10q26). (Figure created using LocusZoom [20] based on the GWAS summary statistics from Fritsche et al. (2016) [18])

The second most significant AMD-associated locus is positioned on chromosome 10q26 and was also identified in 2005 [21]. Since its discovery, the so called *ARMS2-HTRA1* locus was frequently investigated because of its high effect size. An individual carrying one additional C allele of the lead variant rs3750846 has an increased risk of developing AMD by 2.93 times [18]. Remarkably, the C allele is very common in the European population (MAF 20.8 %) and its frequency was found to range around 43.6 % in AMD patients. Despite its large effect size and the strong AMD-association (P-value 6.0 x 10^{-645} in [18]), little is known about the biological mechanisms underlying the GWAS signal at the *ARMS2-HTRA1* locus (**Figure 2 B**). Neither *ARMS2* nor *HTRA1*, the two genes located around rs3750846, were unambiguously shown to contribute in AMD pathogenesis [22–24]. Recently, Grassmann et al. (2017) performed a haplotype analysis based on the IAMDGC data narrowing the association signal to a small region of around 5 kbp, called the “minimal haplotype” [25]. Nevertheless, the detailed mechanisms still remain elusive.

1.3 The GWAS era

After the very first successfully conducted GWAS in 2005 [15] this approach was applied to many other complex diseases. These include *inter alia* neurological diseases, like Alzheimer's disease (AD) [26] or Schizophrenia [27], but also other complex eye diseases, e.g. primary open-angle glaucoma [28] or Myopia [29]. However, GWAS are not restricted to diseases and were applied to a large number of complex phenotypes, including eye colour, height, or blood lipid levels [30–32]. Because of the continuously increasing number of studies, the NHGRI-EBI GWAS Catalog has taken on the task of collecting and storing GWAS results. Remarkably, in September 2018, the repository contained data from 5,687 GWAS comprising 71,673 variant-phenotype associations [33]. The tremendous increase of GWAS loci during the course of time is visualised in **Figure 3**.

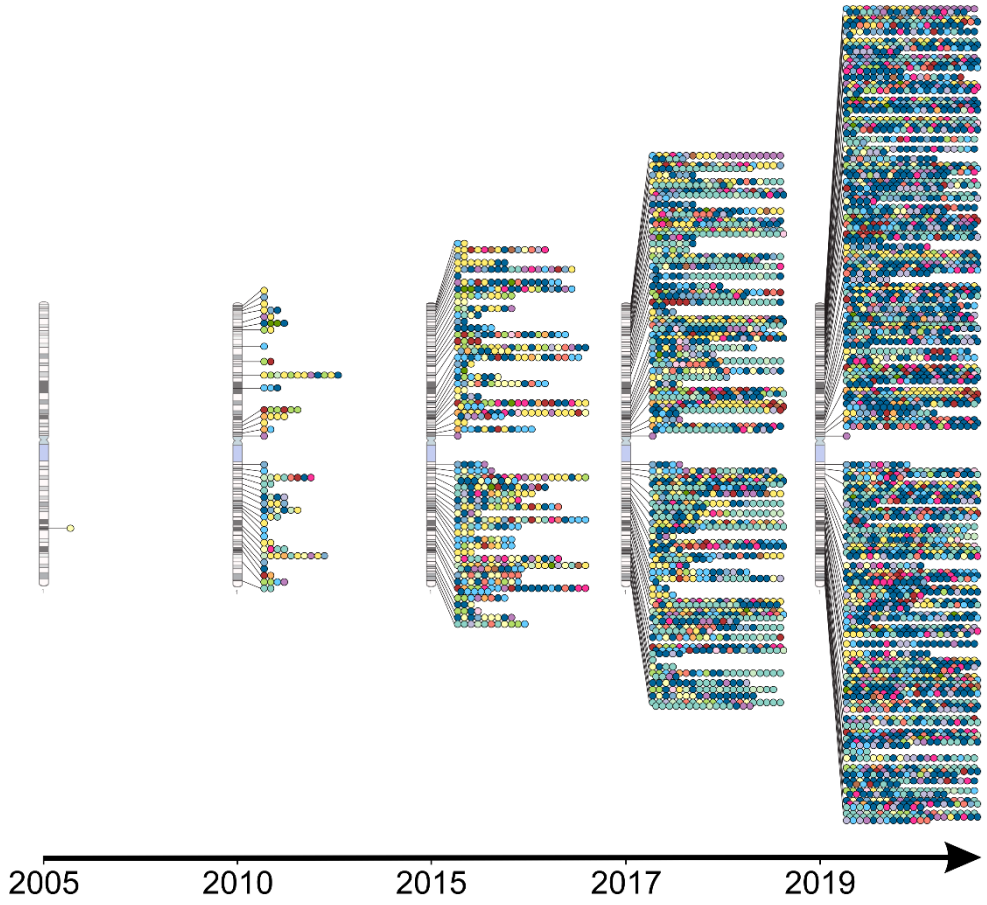

Figure 3: GWAS loci mapped to chromsome 1 during the time period from 2005 to 2019.
The NHGRI-EBI GWAS Catalog collects GWAS results of various complex phenotypes. Shown are the identified GWAS loci on chromosome 1 from 2005 (left) to 2019 (right) at the following time-points: 2005 (fourth quarter), 2010 (first quarter), 2015 (first quarter), 2017 (first quarter), and 2019 (first quarter). Each dot represents one complex phenotype and is colored in respect to predefined groups of potentially related phenotypes. (The plotted data were retrieved from the GWAS catalog online repository [33])

Today, thousands of loci are known to be associated with a multitude of complex phenotypes. In addition, large databases like the UK biobank [34] aim to recruit hundreds of thousands of participants and are likely facilitating the identification of even more GWAS loci. As already mentioned, GWAS aim to identify associated genomic regions but are not suited to draw further conclusions about the underlying biology of the signal. The interpretation of GWAS results is limited by several factors. Due to the extensive linkage disequilibrium (LD) of neighbouring variants in GWAS loci it is usually impossible to classify the signal causing variant (**Figure 2**). Furthermore, GWAS variants are often located in non-coding or intergenic regions of the genome [35,36]. Regarding AMD, altogether 7,218 genome-wide significant variants were identified and statistically fine mapped to a set of 1,345 credible variants [18,37]. Solely 1.9 % of these variants (25 of 1,345) are potentially protein coding and thus modifying the amino acid sequence [18]. Therefore, the associated gene within a GWAS locus frequently remains difficult to determine from the GWAS signal.

Taken together, GWAS are a successful and popular approach to identify genomic regions associated with complex phenotypes. Today, innovative follow up studies are required to enable a deeper understanding of the functional meaning of such association signals.

1.4 Gene expression regulation in GWAS loci

One attractive approach to overcome the above described limitations of GWAS results is to correlate the genotypes of variants, which are associated with disease at genome-wide significance, with mRNA expression in a given tissue using large-scale mRNA expression studies. This type of analysis results in data known as expression Quantitative Trait Loci (eQTL) [38]. eQTL may become evident as local (*cis*) or distant (*trans*) effects (**Figure 4**). Local eQTL implicate that the variant (the so-called eVariant) is located in direct neighbourhood to the affected gene (the so-called eGene) or within the gene body. Local genotype variation possibly affects gene expression by altering transcription factor binding, splicing, DNA methylation or other molecular mechanisms [39]. An altered gene expression usually leads to changes in spatial or temporal transcript levels [40] and thereby possibly influences further genes, located anywhere in the genome. These indirect effects of genomic variants are called distant eQTL and show typically smaller effect sizes than local eQTL (**Figure 4**).

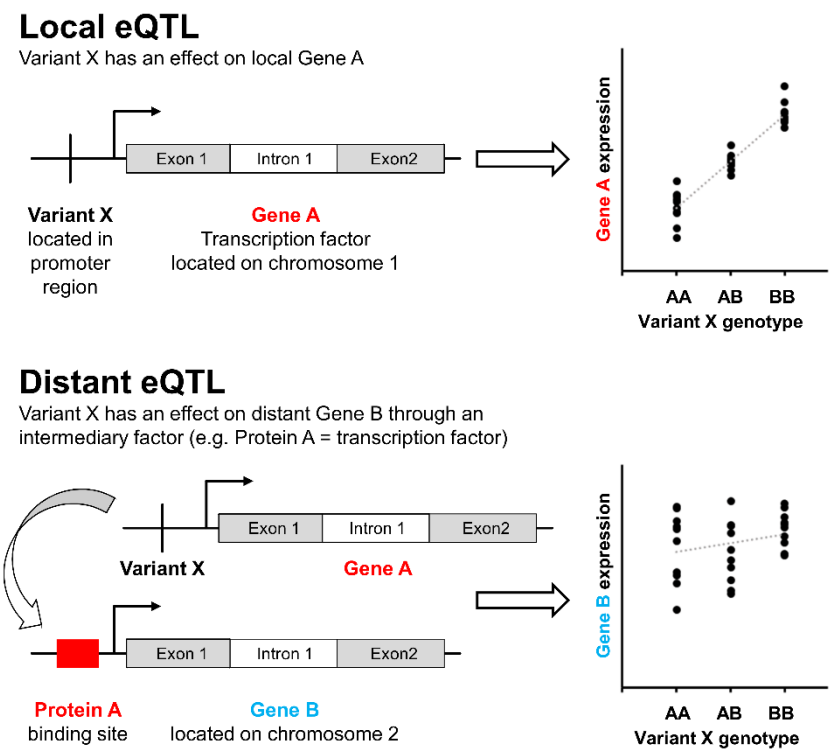

Figure 4: eQTL and their modes of action. Local eQTL variants (eVariants) influence gene expression of nearby genes (eGenes). Distant eQTL effects can be caused if the potentially regulated gene product itself carries out regulatory functions. (Figure modified from Westra et al. (2014) [38])

eQTL studies have proven to be a valuable resource to follow up on GWAS results, since they allow the prioritisation of variants and genes in GWAS loci. Furthermore, eQTL databases are usually covering the whole genome and transcriptome. Their assessment is therefore not restricted to the evaluation of distinct GWAS results and can also be used to find potential commonalities of complex phenotypes or traits. Such pleiotropic effects could reveal pathways contributing to disease aetiology. Nevertheless, eQTL studies are usually based on healthy tissue and do not allow to draw simple implications for pathomechanisms after disease onset.

During the last decade, a large number of studies have investigated eQTL in various tissues [41–44]. The data are usually collected using high throughput platforms, such as genotyping chips to assess the genotypes of the samples and expression microarrays or RNA sequencing (RNA-Seq) to measure the expression of gene transcripts in a given cell type or tissue. Nevertheless, it has become clear that the analysis of single tissue eQTL has limitations, specifically regarding sensitivity and specificity due to a limited statistical power [45]. Furthermore, gene expression may vary between tissues and cell types [46]. Single tissue eQTL studies can miss important signals and correlations. Consequently, combining data from several independent studies can considerably enhance a reproducible outcome of eQTL studies [47,48].

Recently, the integration of more complex models instead of basic linear regression (as shown in **Figure 4**) facilitated a new, comprehensive method to investigate the regulatory influence of genetic variation on gene expression. Transcriptome wide association studies (TWAS) apply a three-step process to identify disease associated genes. First, machine learning algorithms, like ridge regression [49], lasso regression [50], or elastic net [51], are used to determine a set of genetic variants which consistently influence gene expression in a given tissue. Secondly, the corresponding set of genetic variants are extracted from classical GWAS datasets and are used to predict gene expression based on the generated models. This provides a relative expression value per gene for each individual. Finally, predicted gene expression is correlated with each individual's disease status to identify disease-associated genes [52–54]. TWAS have several advantages over classical eQTL studies. Due to the fact that only thousands of genes are investigated instead of millions of genetic variants, less adjustment for multiple testing is required. Additionally, TWAS are an unbiased approach as the machine learning model chooses which variants to use for reproducible gene expression prediction. Nevertheless, TWAS do also not provide information about the biological mechanisms underlying the association signal.

1.5 Genome editing to investigate gene expression regulation

Bioinformatical approaches, like GWAS and eQTL studies, are applied to generate new hypotheses and to provide a higher-level context. Still, such algorithms cannot replace wet lab experiments, which are required to validate findings and to investigate biological models under varying conditions. Although the amount of GWAS studies rapidly increased in the past 15 years, experimental follow up studies were rarely performed [55]. This may in part be attributable to the problematics of interpreting GWAS results as described above. Furthermore, investigating specific genetic variants required extensive technical effort and often resulted in highly artificial model systems. The discovery of the bacterial CRISPR (clustered regularly interspaced short palindromic repeats)/Cas9 (CRISPR-associated protein 9) system changed biological and medical research dramatically [56–58]. Further developments even simplified the multipartite CRISPR/Cas9 complex to require only two components for targeted genome editing: The Cas9 endonuclease protein and a single guide RNA (sgRNA) (**Figure 5 A**) [58]. The 20 nucleotide (nt) long sgRNA sequence can be modified to induce targeted DNA double-strand breaks (DSBs) via the endonuclease activity of

Cas9. sgRNA design further requires the presence of a 3 nt protospacer-adjacent motif (PAM) at the 3 prime end of the target sequence.

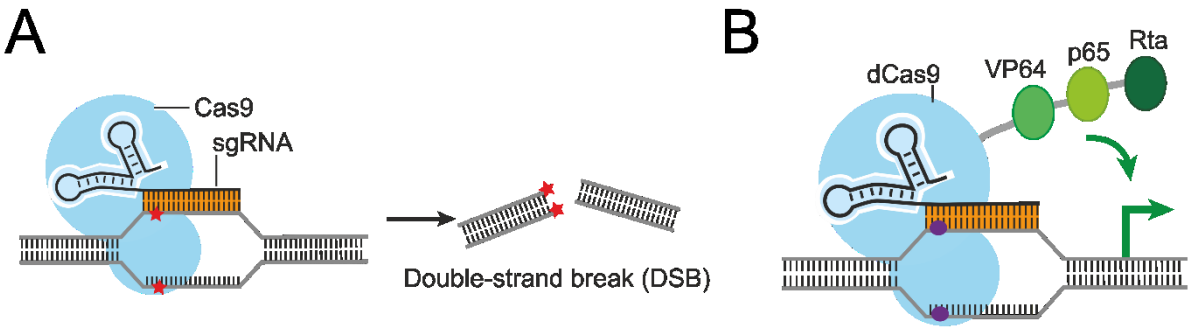

Figure 5: Cas9 mediated genome editing.
(A) The Cas9 endonuclease complex requires a sgRNA to introduce targeted double-strand breaks (DSBs, red stars). (B) Deactivated Cas9 (dCas9) proteins retain their capability to bind DNA, but lost their endonuclease function. The tripartite VPR construct, consisting of the proteins VP64, p65, and Rta, was fused to a dCas9 to enable targeted enhancement of nearby gene expression. (Figure modified from Wang et al. (2016) [59])

Induced DSB are immediately repaired in Eukaryotes by either nonhomologous end joining (NHEJ) or homology-directed (HDR) DNA repair pathways. NHEJ usually leads to small random insertions or deletions at the DSB targeted site, whereas HDR potentially integrates donor DNA sequences by homologous recombination [60–62]. Regarding further experimental investigations of GWAS and eQTL results, both pathways might be valuable depending on the investigated locus and the specific question needed to be addressed. It was further shown that even larger deletions can be introduced with the help of two sgRNAs [63,64]. To facilitate additional usage of DNA-specific targeting, a nuclease-deactivated Cas9 (dCas9) has been engineered. Various effector proteins were fused to dCas9 and have been shown to result in targeted transcriptional activation (**Figure 5 B**) or repression [65,66], and to be capable of modifying epigenetics around the target site [67].

The CRISPR/Cas9 toolbox has been widely applied to address various questions and to generate novel experimental model systems [59]. Still, its implementation, specifically concerning the investigation of GWAS loci and eQTL findings, is under development. Schrode et al. in 2019 were the first to perform an allelic conversion regarding eVariants *in vitro* [68].

1.6 Aim of this study

The IAMDGC identified 52 independent genetic signals in 34 loci to be involved in AMD disease risk [18]. It still remains unclear which variants are indeed causal and exactly

which genes in these loci are affected thus contributing to disease pathology. In general, a genetic predisposition likely exerts a life-time influence, which leads to the question how a genetic variant can contribute to the aetiology of this blinding disease.

This thesis aims to investigate the influence of AMD-associated genetics in the light of gene expression regulation. eQTL databases of various tissues were generated and comprehensively analysed. This process especially included the creation and evaluation of the first eQTL study in healthy retinal tissue to-date. Besides the large-scale bioinformatical studies, one project focused on the experimental assessment of eQTL effects by applying genome editing methods. Finally, a TWAS was performed based on different tissues and the genotypes of over 30,000 AMD patients and controls.

2 Bioinformatical protocols

In this thesis, multiple datasets were collected or generated to calculate eQTL in various tissues. **Table 1** lists all datasets and the respective source. The datasets were initially generated using different platforms and methodological protocols. Therefore, quality control (QC) and data processing was required to jointly analyse genotype and gene expression data. Some datasets were already processed by the respective study site before they were made available. The initial data format and the required processing steps for eQTL calculation are shown in **Table 1**. Altogether three databases were created in this thesis to investigate gene expression regulation in liver tissue, retinal tissue and the Genotype-Tissue Expression (GTEx) project.

Table 1: Overview of analysed eQTL datasets in this thesis

Dataset name	eQTL database	Source	Stored database and accession ID	Genotype data		Gene expression data	
				Received format	Processing before eQTL calculation	Received format	Processing before eQTL calculation
Schadt [69]	Liver	Download	Synapse (syn89614)	Called genotypes (microarray)	Imputation, QC	Gene expression matrix without probe sequences	QC, Normalisation
Schroeder [41]	Liver	Download	GEO (GSE39036, GSE32504)	Called genotypes (microarray)	Imputation, QC	Gene expression matrix and probe sequences	Probe remapping, QC, Normalisation
Innocenti [47]	Liver	Download	GEO (GSE26105, GSE25935)	Called genotypes (microarray)	Imputation, QC	Gene expression matrix and probe sequences	Probe remapping, QC, Normalisation
GTEx version 6 [44]	Liver/GTEx	Download	dbGAP (phs000424.v6.p1)	Called genotypes (microarray)	Imputation, QC	Gene expression matrix of RNA-Seq	QC, Normalisation
GTEx version 7 [44]	GTEx	Download	dbGAP (phs000424.v7.p2)	Called genotypes (WGS)	QC	Gene expression matrix of RNA-Seq	QC, Normalisation
Regensburg	Retina	Data generated in this thesis	-	Raw signal intensities (microarray)	Genotype calling, Imputation, QC	RNA-Seq raw files	Processing of RNA-Seq reads, QC, Normalisation
Cologne	Retina	Provided by Thomas Langmann*	-	Called genotypes (microarray)	Imputation, QC	RNA-Seq raw files	Processing of RNA-Seq reads, QC, Normalisation
NEI [70]	Retina	Provided by Anand Swaroop**	-	Imputed genotypes	QC	RNA-Seq raw files	Processing of RNA-Seq reads, QC, Normalisation

QC = quality control, RNA-Seq = RNA Sequencing; * University Hospital, Cologne, Germany; ** National Eye Institute, Bethesda; USA

2.1 Genotype data processing

2.1.1 Genotype calling

The genotypes of most investigated datasets were detected using microarray platforms and have been made available as hard called genotypes in the VCF format [71] (**Table 1**).

The genotypes of the retinal tissue samples from Regensburg were measured as part of this thesis using an Illumina Custom HumanCoreExome BeadChip. Therefore, genotype calling was necessary before further genotype processing. Hard called genotypes were generated using the Axiome analysis suite version 3.1 based on the "best practice workflow" supplied by the manufacturer.

2.1.2 Quality control before imputation

Before genotype imputation, every dataset underwent several quality control steps regarding the included samples and the genotyped variants. Two datasets, namely Schroeder [41] and Innocenti [47], reported only the zygosity status for each variant encoded as AA, AB and BB. *Biomart* [72] was applied to obtain the according reference and alternative alleles. Additionally, the UCSC *liftover* tool [73] was applied to update genome coordinates to hg19/GRCh37 if required.

For each dataset, a principal component analysis (PCA) was carried out including 30,000 genetic variants of each sample and the corresponding genotype information of the 1000 Genomes Project reference panel (Phase 3, release 20130502) [74]. This analysis was conducted in R (version 3.3.1) [75] using the *snpgdsPCA* function of the *SNPRelate* package [76]. The first two principal components were plotted to determine the ethnicity of each sample. In this thesis, only samples clustering next to the European (EUR) reference individuals were included because haplotype structures can importantly vary between populations. Furthermore, samples were excluded in case of high missing rates (> 5% of genetic variants) and if reported and inferred gender from genotype calling did not match.

To investigate the quality of genetic variants, allele frequencies were calculated and compared to the corresponding allele frequency of the 1000 Genomes Project EUR samples. Alleles were flipped, in case they were given on the opposite strand. Genetic

variants, whose reference allele frequency deviated more than 10% from the reference were excluded from the analysis. Next, *VCFtools* (version 0.1.15) [71] was applied to investigate if variants deviated significantly from Hardy-Weinberg equilibrium (HWE, P-value < 1×10^{-6}) [77]. Only biallelic autosomal variants were kept for further analysis.

2.1.3 Genotype imputation

Before genotype imputation, *SHAPEIT2 (*version 2.r904) was applied to achieve phasing of genotypes with the help of the 1000 Genomes Phase 3 reference panel [78]. *SHAPEIT2* required a two-step protocol: Initially, the *-check* option was used to identify genetic variants, which did not fulfil the manufacturer's criteria. These variants were thereafter excluded from the phasing process. After genotype phasing, IMPUTE2 (version 2.3.2) was utilised with standard options to impute genotypes based on the previously mentioned reference panel [79].

2.1.4 Quality control after imputation

The genotype imputation produced various output files. These files were converted into VCF format with the help of *qctools* (version 1.2, https://www.well.ox.ac.uk/~gav/qctool_v1/#overview accessed February 12th 2017). Furthermore, genotypes were converted into the "estimated allele dosage" format. The VCF files were filtered for low imputation quality (*IMPUTE2* info score) and MAF. The Imputation quality threshold for the liver eQTL database was set to 0.4 and the MAF was at least 5 %. For all other databases imputation quality threshold was 0.3 with a MAF threshold of 1 %. Furthermore, the genomic coordinates of the retina eQTL database were lifted to hg38/GRCh38 by applying the UCSC *liftover* tool.

2.2 Gene expression data processing

2.2.1 Microarray data

The generated eQTL databases in this thesis included three datasets, which measured gene expression via microarray (**Table 27**). Processing of raw data was performed in the respective publication [41,47,69].

The two datasets Schroeder and Innocenti additionally provided the microarray probe sequences. Genome annotation changed with time and therefore array probes were

remapped to an *in silico* mRNA reference database from ensembl [80] using the ReAnnotator pipeline [81]. After remapping, only exome-matching probes showing less than five mismatches were kept. Furthermore, probes which overlapped with a common dbSNP variant (version 142) were removed [82]. Only specific probes measuring one gene were retained. Probes which unambiguously detected gene expression of the same gene, were merged by calculating the mean of all corresponding probes. This value was then weighted by the variance of the respective single probe over all samples.

In contrast, Schadt et al. [69] employed the Agilent Custom 44k array and probe sequences were not available, which made remapping impossible. The provided gene identifier were checked to unanimously match to a gene in the ensemble- or RefSeq- [83] database and were excluded from the analysis if this was not the case. Furthermore, a Shapiro–Wilk test [84] revealed that values above 2 and below -2 were likely outliers and therefore have been set “missing” in the further analysis.

2.2.2 RNA Sequencing (RNA-Seq)

All datasets except the ones mentioned in section 2.2.1 used RNA-Seq to measure gene expression. For the three studies investigating eQTL in retinal tissue, the raw data were available (**Table 32**) and have been analysed with the same protocol to ensure comparability. The RNA-Seq pipeline was based on the protocol of Ratnapriya et al. (2019) [70]. During all steps of the analysis, FastQC (version 0.11.5, https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ accessed January 24th 2018) and MultiQC (version 1.7.dev0) [85] were applied to ensure the correctness of the conducted data processing steps.

First, the raw RNA-Seq reads were trimmed for Illumina adapter sequences and low quality reads were removed with the following options: SLIDING WINDOW 4:5, LEADING 5, TRAILING 5, and MINLEN 25 using Trimmomatic (version 0.39) based on the supplied Illumina TruSeq3 sequences [86]. Afterwards, the Star aligner (version 2.7.1a) [87] was applied to build a human reference genome annotation based on the ensembl version 97 (GRCh38.p13) [80]. Trimmed reads were aligned to this reference using per sample 2-pass mapping and ENCODE standard options. The resulting aligned files were thereafter analysed with the *RSEM* toolbox (version 1.3.1) [88]. To accomplish this, a RSEM reference file was created with the *rsem-prepare-reference*

option and the above mentioned ensembl version 97. *RSEM* then calculated the estimated gene expression per sample using the *rsem-calculate-expression* function with standard parameters and the “forward-prob = 0” option to account for stranded RNA-Seq libraries. Calculation of gene expression counts required RSEM to assume a fragment length distribution, which is done automatically if paired-end reads are supplied. The Regensburg dataset investigated retinal gene expression based on single-end reads and therefore the options *fragment-length-mean* 155.9 and *fragment-length-sd* 56.2 were additionally supplied to the *rsem-calculate-expression* function. Both values have been obtained by calculating the mean fragment length distribution of 30 samples taken randomly from the Cologne and NEI datasets. After gene expression calculation, the *rsem-generate-data-matrix* function created one estimated read count matrix per dataset. The estimated expression counts obtained from RSEM required further normalisation to enable an appropriate comparison of gene expression between samples and datasets. For this reason, the *tmmnorm* function of the *edgeR* package (version 3.16.5) [89] was applied to conduct a trimmed mean of M-values normalisation [90]. The normalised expression matrix was then used by the *cpm* function of *edgeR* to calculate the gene expression in counts per million (CPM).

2.2.3 Data normalisation and quality control

The gene expression matrices of all datasets underwent a uniform data normalisation and quality control protocol in R to allow comparison and combination of data. The applied protocol was independent of the different RNA measurement methods or units. Only expressed genes were kept for data normalisation to remove potential measurement artefacts. A gene was considered to be expressed if the expression value was at least 1 in 10 % of all samples within the dataset. For the GTEx project this threshold was set 0.1 to enable a comparison of results with the original GTEx analysis pipeline. Next, a PCA was performed with the help of the *prcomp* function to identify and to remove potential outlier samples within the dataset. Replicated samples were merged by taking the mean of the gene expression values.

The gene expression matrix was then log2-transformed with an offset of 0.001 (liver and GTEx eQTL datasets) or 1 (retina eQTL datasets). Thereafter, the single gene expression matrices were differently processed according to the three main databases created in this thesis, which purposed the calculation of Liver eQTL, the GTEx database, or the retina eQTL database.

For the calculation of eQTL in liver tissue, only genes were kept which have been expressed in at least two of the four datasets. The expression of genes which has not been directly measured in all datasets was imputed using the K-Nearest-Neighbour method implemented in the *impute.knn* function of the *impute* Bioconductor package [91]. If imputation was not possible, the gene was removed from further analysis. Thereafter, the gene expression matrices of each single dataset were merged into one matrix. The log2 transformed and merged matrix was quantile normalised [92] using the *normalize.quantiles* function of the R package *preprocessCore* (https://github.com/bmbolstad/preprocessCore accessed June 16th 2017). As last normalisation step, an empirical batch correction method called ComBat was performed, which corrected for the different origin of data [93]. The *combat* function is part of the *sva* package in R [94].

The GTEx database was primarily generated based on GTEx v6 (dbGaP: phs000424.v6.p1). During the course of this thesis, the GTEx consortium released v7 (dbGaP: phs000424.v7.p2), which included more samples and tissues. For this reason the gene expression data of the GTEx database was processed twice with slightly different protocols. In version 6 all samples measuring different tissue subtypes, for example “Adipose Subcutaneous” and “Adipose Visceral Omentum”, were merged into higher order tissues (e.g. “Adipose”). This resulted in 28 tissues. Thereafter, the gene expression quality control and normalisation was conducted for each tissue separately. The log2 transformed expression values were quantile normalised and additionally rescaled to a mean of 4 (SD: 1) using the *rescale* function, which is embedded in the *psych* package (https://cran.r-project.org/web/packages/psych/index.html accessed June 16th 2017). Rescaling of gene expression ensured a better comparability of effect sizes between GTEx tissues. Furthermore, a mega-analysis was conducted based on the normalised gene expression matrices of the 28 tissues. For this reason, ComBat was applied to adjust for tissue-specific effects by setting the tissue as covariate. The updated GTEx database (version 7) applied the same data normalisation protocol like version 6 but without merging tissue subtypes. This resulted in 48 different tissues being included in the eQTL analysis.

Three datasets contributed to the retinal eQTL database. Gene expression data were merged into one matrix including exclusively genes, which were expressed in all three

datasets. Afterwards, quantile normalisation and ComBat were employed to normalise the data.

2.3 eQTL analysis

2.3.1 eQTL calculation

In this study, eQTL were calculated based on linear regression models implemented in the *Matrix eQTL* package in R [95]. *Matrix eQTL* required three input files with columns representing samples and rows including the respective data. The files contained (1) genotypes in estimated allele dosage format, (2) normalised gene expression, and (3) covariates. The covariate file comprised information about age, gender, and the first five principal components from the genotype PCA. Furthermore the "*cisDist*" parameter was set to 1 Mbp if local eQTL were investigated. The output of *Matrix eQTL* gave information about several parameters. Besides data about the eVariant and the eGene, it presented the effect size (slope of the linear regression model, beta), the standard error of the effect size (beta-SE) and the P-value of the model. To account for multiple testing, the false discovery rate (FDR, Q-value) was calculated using the *p.adjust* function in R. The results were thereafter filtered for significance according to the given Q-value threshold.

2.3.2 Meta-analysis of eQTL

The meta-analysis approach compromises the eQTL analysis summary statistics of different datasets or tissues and was performed in each database seperately. In this thesis, a random effects model implemented in the function *MiMa* (version 1.4.) [96] was applied to conduct a meta-analysis of *Matrix eQTL* results. It required the beta and the beta-SE of each dataset to estimate the joint beta and standard error as well as the joint P-value. The retrieved P-values were thereafter corrected for multiple testing by applying the FDR.

2.3.3 Mega-analysis of eQTL and conditional eQTL analysis

In this study, a mega-analysis was conducted with each of the three generated databases. The mega-analysis calculates eQTL from the merged genotype and

expression data directly and does not need summary statistics. *Matrix eQTL* was applied after merging the data as described in section 2.3.1.

Furthermore, the enhanced statistical power of the mega-analysis method was utilised to investigate independent eQTL signals for each significant eGene. *Matrix eQTL* was adjusted for the most significant corresponding eVariant per eGene by adding its genotype information to the covariate file. Thereafter, eQTL were re-calculated and remaining significant eVariants were considered to represent an independent signal. The most significant independent eVariant was then also added to the covariate file. This approach was repeated until no additional independent signals were found. The conditional analysis could not be appropriately adjusted for multiple testing. For this reason the P-value threshold for significance of further independent signals was estimated based on the applied Q-value threshold of the respective mega-analysis.

2.4 Transcriptome-wide association study

The TWAS conducted in this thesis was performed to identify AMD-associated genes based on the gene expression regulation of AMD-associated genetic variants. Therefore, the *PrediXcan* algorithm [53] was applied to predict gene expression using genotypes of AMD-cases and healthy controls. The required prediction models have been trained on the data of European individuals within the GTEx v7 release. Model building was performed by Gamazon et al. [53] and the respective files were downloaded from PredictDB (http://predictdb.org/, accessed September 3rd 2018). Gene expression prediction was accomplished based on the genotypes of 33,976 unrelated individuals with European ancestry from the IAMDGC cohort [18]. These included 16,144 late-stage AMD cases, presenting GA and/or CNV, and 17,832 AMD-free controls. Genotypes were transformed into allele dosage format and missing genotypes of single individuals were replaced by the most frequent corresponding genotype. This resulted in 11,722,957 autosomal genetic variants for analysis. Gene expression was predicted for 27 tissues and thereafter the *lm* function was applied in R to calculate the linear regression model of gene expression and AMD status, encoded as 0 (healthy) and 1 (AMD). The analysis model was further adjusted for gender, age and the first two principal components of the genotype PCA performed by Fritsche *et al*. [18]. Multiple testing correction was conducted by calculating the Q-value. Genes with a Q-value smaller than 0.001 were considered to be significantly AMD-associated. Before result evaluation, genes located in the major

histocompatibility complex (MHC) locus (chr6: 28,477,797 - 33,448,354, hg19) were excluded from the analysis.

2.5 Follow-up investigations of eVariants and eGenes

2.5.1 Gene set enrichment analysis with g:Profiler

Gene set enrichment analysis was performed with the help of the web based tool g:Profiler (version r1730_e88_eg35) [97]. The program was used to assign Gene Ontology (GO) biological pathways [98] to all query genes and to perform an enrichment analysis using the “Best per parent” hierarchical filtering. The g:profiler g:SCS method was applied to account for multiple testing and was set to an adjusted P-value threshold of 0.05.

2.5.2 Hierarchical clustering

Clustering of genes based on their expression was performed using the *hclust* function in R. The hierarchical trees were then processed and visualised with the help of the dendextend package [99] in R.

3 Material & Methods: Wet lab experiments

3.1 Material

3.1.1 Escherichia coli (E. coli) strains

Table 2: *E. coli* strains used

Strain	Source
E. coli strain DH5α	Life Technologies, Carlsbad, CA, USA
E. coli strain Stbl3	Life Technologies, Carlsbad, CA, USA

3.1.2 Eukaryotic cell lines

Table 3. Cell lines used and their origin

Cell Line	Organism	Tissue of origin	Source
HEK293T	*Homo sapiens*	Embryonic kidney	ATCC, LGC Standards GmbH, Wesel, Germany

3.1.3 Oligonucleotides for PCR and sequencing reactions

Table 4: Names, sequences and purposes of oligonucleotides used in this thesis

Name	5'-3' Sequence	On-target-score*	Purpose
UP_ARMS2_F_EcoRI	GAA TTC AAT CAG AGG CAA TGG TCT GC	-	Cloning of target region for UP sgRNA testing, Genotyping after ARMS2 locus deletion
UP_ARMS2_R_BamHI	GGA TCC CCT GAT GAA TCA TGG TCG AG		
DOWN_ARMS2_F_EcoRI	GAA TTC TTG ATC ACA TGC CAT GCT TTT	-	Cloning of target region for DOWN sgRNA testing
DOWN_ARMS2_R_BamHI	GGA TCC ACG ATA TTT TAG GTT GAG GAG CA		
UP_ARMS2_sgRNA_1_F	CAC CGG ACA CAA GTG CTA CAA GGC G	86	Cloning of UP sgRNA 1
UP_ARMS2_sgRNA_1_R	AAA CCG CCT TGT AGC ACT TGT GTC C		
UP_ARMS2_sgRNA_2_F	CAC CGG CCC AGG CCT AAT CCA GCG C	83	Cloning of UP sgRNA 2
UP_ARMS2_sgRNA_2_R	AAA CGC GCT GGA TTA GGC CTG GGC C		
UP_ARMS2_sgRNA_3_F	CAC CGA ATT AAC TGA GTG CCA GCG C	83	Cloning of UP sgRNA 3
UP_ARMS2_sgRNA_3_R	AAA CGC GCT GGC ACT CAG TTA ATT C		
UP_ARMS2_sgRNA_4_F	CAC CGG CCA GCG CTG GAT TAG GCC T	81	Cloning of UP sgRNA 4
UP_ARMS2_sgRNA_4_R	AAA CAG GCC TAA TCC AGC GCT GGC C		
UP_ARMS2_sgRNA_5_F	CAC CGG AGG TGA CAG AGC TCT CCG A	77	Cloning of UP sgRNA 5
UP_ARMS2_sgRNA_5_R	AAA CTC GGA GAG CTC TGT CAC CTC C		
DOWN_ARMS2_sgRNA_1_F	CAC CGG ATA CTT AAA AGC CAA CCC C	71	Cloning of DOWN sgRNA 1

DOWN_ARMS2_sgRNA_1_R	AAA CGG GGT TGG CTT TTA AGT ATC C		
DOWN_ARMS2_sgRNA_2_F	CAC CGC ATG CAA CTG ATT TAG GGG A	66	Cloning of DOWN sgRNA 2
DOWN_ARMS2_sgRNA_2_R	AAA CTC CCC TAA ATC AGT TGC ATG C		
DOWN_ARMS2_sgRNA_3_F	CAC CGA TGC AAC TGA TTT AGG GGA A	60	Cloning of DOWN sgRNA 3
DOWN_ARMS2_sgRNA_3_R	AAA CTT CCC CTA AAT CAG TTG CAT C		
DOWN_ARMS2_sgRNA_4_F	CAC CGT GCA GTT AAT GTA ACT CAA T	71	Cloning of DOWN sgRNA 4
DOWN_ARMS2_sgRNA_4_R	AAA CAT TGA GTT ACA TTA ACT GCA C		
DOWN_ARMS2_sgRNA_5_F	CAC CGC ACC TTT GTC CTA TTT TGG A	59	Cloning of DOWN sgRNA 5
DOWN_ARMS2_sgRNA_5_R	AAA CTC CAA AAT AGG ACA AAG GTG C		
UP_ARMS2_F2	TTC AGG CCT CCT TCC TCA AG	-	Genotyping of single clones after minimal haplotype deletion
DOWN_ARMS2_R2	GGA CAA AGG TGA GGA AGT TCA		
YFP-F-AGEI	ACC GGT ACC ATG GTG AGC AAG GGC GAG GA	-	Cloning for px330-GFPo
YFP-R-ECORI	GAA TTC TTA CTT GTA CAG CTC GTC CA		
MID2_ARMS2_F_EcoRI	GAA TTC GAC CTC TGT TGC CTC CTC TG	-	Cloning of target region for MID sgRNA testing
MID2_ARMS2_R_BamHI	GGA TCC TGA CTC CTC TAA CAA CCC GG		
MID_ARMS2_sgRNA_1_F	CAC CGC CAA CTG GGT GGC TTA AAC G	91	Cloning of MID sgRNA 1
MID_ARMS2_sgRNA_1_R	AAA CCG TTT AAG CCA CCC AGT TGG C		
MID_ARMS2_sgRNA_2_F	CAC CGT TCT GTG TAC TGA CAC TAT C	74	Cloning of MID sgRNA 2
MID_ARMS2_sgRNA_2_R	AAA CGA TAG TGT CAG TAC ACA GAA C		
MID_ARMS2_sgRNA_3_F	CAC CGC TGA GAC CAC CCA ACA ATT C	81	Cloning of MID sgRNA 3
MID_ARMS2_sgRNA_3_R	AAA CGA ATT GTT GGG TGG TCT CAG C		
MID_ARMS2_sgRNA_4_F	CAC CGC GTC ACA CAA AAA TGC CCC C	77	Cloning of MID sgRNA 4
MID_ARMS2_sgRNA_4_R	AAA CGG GGG CAT TTT TGT GTG ACG C		
MID_ARMS2_sgRNA_5_F	CAC CGC CTT CCT CTG GTT GAA TAG C	73	Cloning of MID sgRNA 5
MID_ARMS2_sgRNA_5_R	AAA CGC TAT TCA ACC AGA GGA AGG C		
MID_ARMS2_sgRNA_6_F	CAC CGG GCC CCT CAA GCC GGT GAA T	90	Cloning of MID sgRNA 6
MID_ARMS2_sgRNA_6_R	AAA CAT TCA CCG GCT TGA GGG GCC C		
MID_ARMS2_sgRNA_7_F	CAC CGC TCT GGC AGA GCA GGA CTG A	52	Cloning of MID sgRNA 7
MID_ARMS2_sgRNA_7_R	AAA CTC AGT CCT GCT CTG CCA GAG C		
MID_ARMS2_sgRNA_8_F	CAC CGG ATG GCA GCT GGC TTG GCA A	62	Cloning of MID sgRNA 8
MID_ARMS2_sgRNA_8_R	AAA CTT GCC AAG CCA GCT GCC ATC C		
MID_ARMS2_sgRNA_9_F	CAC CGC ACT CTG CGA GAG TCT GTG C	69	Cloning of MID sgRNA 9
MID_ARMS2_sgRNA_9_R	AAA CGC ACA GAC TCT CGC AGA GTG C		

MID_ARMS2_sgRNA_10_F	CAC CGG AAT TGC CTA GGC CTC CCT G	57	Cloning of MID sgRNA 10
MID_ARMS2_sgRNA_10_R	AAA CCA GGG AGG CCT AGG CAA TTC C		
MID_ARMS2_sgRNA_11_F	CAC CGA GAT GGC CTT CTA TAA GCT T	78	Cloning of MID sgRNA 11
MID_ARMS2_sgRNA_11_R	AAA CAA GCT TAT AGA AGG CCA TCT C		
M13F	CGC CAG GGT TTT CCC AGT CAC GAC	-	Vector primer for pGem®-T
M13R	AGC GGA TAA CAA TTT CAC ACA GGA		
MIAT_sgRNA_1_F	CAC CGG CGC CCA TGA AAT TTT AAT G	71	Cloning of MIAT sgRNA 1
MIAT_sgRNA_1_R	AAA CCA TTA AAA TTT CAT GGG CGC C		
MIAT_sgRNA_2_F	CAC CGA TGC GGG AGG CTG AGC GCA C	74	Cloning of MIAT sgRNA 2
MIAT_sgRNA_2_R	AAA CGT GCG CTC AGC CTC CCG CAT C		
MIAT_sgRNA_3_F	CAC CGC ATT AGG CCG CAG AGA GCT C	68	Cloning of MIAT sgRNA 3
MIAT_sgRNA_3_R	AAA CGA GCT CTC TGC GGC CTA ATG C		
MIAT_sgRNA_4_F	CAC CGG CTT CTG CGC CCC TGG TCC G	74	Cloning of MIAT sgRNA 4
MIAT_sgRNA_4_R	AAA CCG GAC CAG GGG CGC AGA AGC C		

* Provided by the *Optimized CRISPR Design*-Tool (http://crispr.mit.edu, accessed February 1st 2018)

3.1.4 Oligonucleotides and corresponding probes used for qRT-PCR

Table 5: Names, sequences and corresponding probe numbers for oligonucleotides used for qRT-PCR

Name	5'-3' Sequence	Gene	Roche Universal Probe Library #
hSDHA-RT-F2	AGC ATC GAA GAG TCA TGC AG	*SDHA*	60
hSDHA-RT-R2	GCT TCC ATC AGC AAA TCT CAA		
huLILRA3_RT_F	TGT GTG GTC TCT ACC CAG TGA	*LILRA3*	7
huLILRA3_RT_R	CAG AGC CAC ACT GGA AGG TC		
huCD300E_RT_F	GGG AGG TGT TGA CCC AAA AT	*CD300E*	66
huCD300E_RT_R	AGG ACC ACG AGC AGG AAG T		
huMUC7_RT_F	TCA ACT GAC AAG TAG TTT GAC CAG A	*MUC7*	69
huMUC7_RT_R	CCA ATC CTT TGA GGA TGG TAA C		
huDEFA5_RT_F	TGA GGC TAC AAC CCA GAA GC	*DEFA5*	60
huDEFA5_RT_R	GCT CTT GCC TGA GAA CCT GA		
huTNFAIP1_RT_F	AGA ACC GGC AAG AAA TCA AG	*TNFAIP1*	41
huTNFAIP1_RT_R	CTG GTA GGA GTC CTT CTT GTC C		
huFCN1_RT_F	GTT CTG GCT GGG GAA TGA C	*FCN1*	38
huFCN1_RT_R	AAC TGG TGG TTG CCC TCA		
huPILRB_RT_F	GGT GGA GGA GAA GGA AAG GT	*PILRB*	7
huPILRB_RT_R	GGG TCT CAC ATC ACG TCC TC		
huC17orf62_RT_F	GCC CTC TCG GGA TGT ACC	*C17orf62*	39
huC17orf62_RT_R	TTC CAG CCC AGG CTA TCA		
huDAZAP1_RT_F	TCG AGG ACG AAC AAT CAG TG	*DAZAP1*	64

huDAZAP1_RT_R	GCT CAG CTC GTT TAA CTT CCA		
huIL6_RT_F	GAT GAG TAC AAA AGT CCT GAT CCA	*IL6*	40
huIL6_RT_R	CTG CAG CCA CTG GTT CTG T		
huNFKB1_RT_F	CCT GGA ACC ACG CCT CTA	*NFKB1*	49
huNFKB1_RT_R	TCA TATG GTT TCC CAT TTA ATA TGT C		
huFLOT2_RT_F	GAC CCT GGA GGG ACA TCT G	*FLOT2*	58
huFLOT2_RT_R	ACT GGT CCC GGT CCT GAT A		
huCYP1A1_RT_F	ACC TTC CCT GAT CCT TGT GA	*CYP1A1*	33
huCYP1A1_RT_R	GAT CTT GGA GGT GGC TGC T		
hHTRA1-RT-F2	AGC AGA CAT CGC ACT CAT CA	*HTRA1*	37
hHTRA1-RT-R2	GAT GGC GAC CAC GAA CTC		
hMIAT_RT_F	AGA ACA CGC TTT ATT ACA GTC TCG	*MIAT*	80
hMIAT_RT_R	CCC GAG GTC CAA AGA GAA GT		
hLOC387715-rt-F2	AGC TCT GCT TAC CAG CCT TCT	*ARMS2*	82
hLOC387715-RT-R	TTG CTG CAG TGT GGA TGA TAG		

3.1.5 Plasmids and expression constructs

Table 6: List of expression constructs, short names, applications, and sources

Vector name	Short name	Application	Source
pGEM®-T	-	Cloning	Promega Corporation, Madison, WI, USA
pCAG-EGxxFP	-	sgRNA test	Addgene, LGC Standards, Teddington, UK
pU6-(BbsI)_CBh-Cas9-T2A-mCherry	px330-mCherry	sgRNA test	Addgene, LGC Standards, Teddington, UK
pSpCas9(BB)-2A-GFP (PX458)	px330-eGFP	sgRNA vector for *ARMS2-HTRA1* haplotype deletion	Addgene, LGC Standards, Teddington, UK
px330_GFPo	px330-GFPo	sgRNA vector for *ARMS2-HTRA1* haplotype expression enhancement	Institute of Human Genetics, University of Regensburg, Germany
SP-dCas9-VPR	dCas9-VPR	Gene expression enhancer	Addgene, LGC Standards, Teddington, UK

3.1.6 Enzymes

Table 7: Enzymes used

Enzyme	Source
AgeI	New England Biolabs, Ipswich, MA, USA
BamHI-HF	New England Biolabs, Ipswich, MA, USA
BpiI	New England Biolabs, Ipswich, MA, USA
EcoRI-HF	New England Biolabs, Ipswich, MA, USA

FastDigest BpiI	Thermo Fisher Scientific, Waltham, MA, USA
GoTaq® DNA Polymerase	Promega Corporation, Madison, WI, USA
House Taq DNA Polymerase	Institute of Human Genetics, University of Regensburg, Germany
Quick CIP	New England Biolabs, Ipswich, MA, USA
RecBCD Exonuclease	New England Biolabs, Ipswich, MA, USA
T4 DNA Ligase	New England Biolabs, Ipswich, MA, USA
T4 PNK Kinase	New England Biolabs, Ipswich, MA, USA
Trypsine	GE Healthcare, Galfont St Giles, GB

3.1.7 Kit systems

Table 8: List of kit systems used

Kit	Source
BigDye Terminator v1.1, v3.1 Cycle Sequencing Kit	Thermo Fisher Scientific, Waltham, MA, USA
Lipofectamine 3000	Thermo Fisher Scientific, Waltham, MA, USA
NucleoSpin® Gel and PCR Clean-up	MACHEREY-NAGEL GmbH & Co. KG, Düren, Germany
NucleoSpin® Plasmid	MACHEREY-NAGEL GmbH & Co. KG, Düren, Germany
NucleoBond® XtraMidi	MACHEREY-NAGEL GmbH & Co. KG, Düren, Germany
Quick Ligation™ Kit	New England Biolabs, Ipswich, MA, USA

3.1.8 Chemicals and cell culture supplements

Table 9: List of chemicals used

Chemical/Reagent	Source
Agarose (Biozym LE)	Biozym Scientific GmbH, Hessisch Oldendorf, Germany
Ampicillin sodium salt	Carl Roth GmbH + Co. KG, Karlsruhe, Germany
Bromphenolblau Natriumsalz	Sigma-Aldrich, St. Louis, MO, USA
4',6-Diamidin-2-phenylindol (DAPI)	Thermo Fisher Scientific, Waltham, MA, USA
Chloroquine	Merck Chemicals GmbH, Schwalbach, Germany
DMEM High Glucose Medium (4,5 g/l)	Thermo Fisher Scientific, Waltham, MA, USA
Dimethyl sulfoxide (DMSO)	VWR International Germany GmbH, Darmstadt, Germany
dNTPs (dATP, dGTP, dCTP, dTTP)	Genaxxon Bioscience, Ulm, Germany
Ethanol ≥ 99,8 p.a	Carl Roth GmbH + Co. KG, Karlsruhe, Germany
Ethidiumbromide	AppliChem GmbH, Darmstadt, Germany
Ethylendiamintetraacetat disodium dihydrate salt (EDTA)	Merck Chemicals GmbH, Schwalbach, Germany
Fetal Bovine Serum Gold (FCS)	Thermo Fisher Scientific, Waltham, MA, USA
Glycerol 87 %	University of Regensburg, Chemical Supplies
Gel Loading Dye Purple (6x)	New England Biolabs, Ipswich, MA, USA
HiDi™ Formamide	Thermo Fisher Scientific, Waltham, MA, USA
Isopropanol	Merck Chemicals GmbH, Schwalbach, Germany
OptiMEM™ Medium	Thermo Fisher Scientific, Waltham, MA, USA
Penicillin (10.000 Units)/Streptomycin (10 mg/ml), (Pen/Strep)	GE Healthcare, Galfont St Giles, GB

Poly-L-Lysine Hydrobromide (0.1 mg/ml)	Sigma-Aldrich, St. Louis, MO, USA

3.1.9 Buffers and solutions

Table 10: Composition of buffers and solutions used

Buffer/Solutions	Composition and amounts
5x TBE	Tris 0,5 M
	Boric acid 0,5 M
	EDTA 10 mM
	H_2O dest.
2x HBS	NaCl 280 mM
	KCl 10 mM
	Na_2HPO_4 1.5 mM
	HEPES 50 mM
	H_2O dest.
LB-Medium	Tryptone 1% w/v
	Yeast extract 0,5% w/v
	NaCL 1% w/v
	H_2O dest. 1 l
LB-Plates	Tryptone 1% w/v
	Yeast extract 0,5% w/v
	NaCL 1% w/v
	Bacto-Agar 15% w/v
	H_2O dest. 1l
SOC-Medium	Tryptone 2 % w/v)
	Yeast extract 0,5 % w/v
	NaCl 10 mM 0,5 g/l
	KCl 2,5 mM 0,2 g/l
	Glucose 20mM 20ml
	H_2O dest. 1 l
HEK29T medium	DMEM High Glucose Medium 89 %
	FCS 10 %
	Pen/Strep 1 %
HEK29T freezing medium	DMEM High Glucose Medium 70 %
	FCS 20 %
	DMSO 10 %

3.2 Methods

In this thesis, a sgRNA mediated CRISPR/Cas9 system was applied to induce DSBs or to enhance gene expression. Before these experiments, sgRNAs were tested for specificity using a two-vector system. One vector included the sgRNA target sequence (pCAG-EGxxFP), whereas the other vector carried the sgRNA- and the Cas9 coding sequence (px330-mCherry). Both vectors required different cloning strategies.

3.2.1 Cloning of pCAG-EGxxFP constructs

3.2.1.1 Polymerase chain reaction (PCR)

The defined sgRNA target sequence was amplified from human genomic DNA conducting a Polymerase chain reaction (PCR). The PCR reaction mix is given in **Table 11** and the respective program in **Table 12**. PCR conditions were adjusted according to primer parameters (given in SnapGene, version 2.8.2) and the required elongation time (1 min/1,000 bp).

Table 11: PCR reaction mix

Component	Volume
5x Green GoTaq® Reaction Buffer	5 µl
Primer forward (10 µM)	1 µl
Primer reverse (10 µM)	1 µl
dNTPs (1.25 mM)	2 µl
human genomic DNA (25 ng/µl)	2 µl
GoTaq® DNA polymerase	0.1 µl
H_2O (Millipore)	13.9 µl

Table 12: Thermocycler program for PCR amplification

Step of the reaction	Temperature	Duration	Cycles
Initial denaturation	95 °C	3 min	
Denaturation	94 °C	30 s	30
Annealing	x °C*	30 s	
Elongation	72 °C	x min	
Final elongation	72 °C	5 min	
Break	4 °C	-	

*x indicates variable temperature and time, adjusted for each sequence to be amplified

3.2.1.2 Agarose gel electrophoresis

PCR products were run on agarose gels to evaluate amplicon size and purity. Agarose gels were generated by heating 1 % (w/v) agarose in TBE buffer until the agarose solved completely. After cooling down the mixture to 37°C, 3 drops of 0.003 % ethidiumbromide solution were added. If necessary, Bromphenolblue loading buffer (5x solution) was added to the samples before loading them onto the gel. 5 µl GeneRuler™ DNA Ladder Mix served as a size standard and gels were run at 220 V for 20 min.

3.2.1.3 Purification of PCR products from agarose gels

PCR products of the correct size were excised from agarose gels and purified using the NucleoSpin® Gel and PCR Clean-up kit according to the manufacturer's instructions. DNA was eluted from columns in 20 µl of Millipore H_2O and stored at -20 °C until further use.

3.2.1.4 Ligation into pGEM®-T

The purified PCR amplicons were ligated into the pGEM®-T vector using the ligation mix given in **Table 13**. The ligation reaction was incubated at 4 °C overnight.

Table 13: pGEM®-T vector ligation mix

Component	Volume
pGEM®-T vector	0.5 µl
PCR fragment	4 µl
T4 DNA Ligase Puffer (2x)	5 µl
T4 DNA Ligase	0.5 µl

3.2.1.5 Heat shock transformation of E. coli

E. coli cells were transformed with plasmid DNA using a heat shock procedure. One 100 µl aliquot of competent *E. coli* cells was thawed on ice for 5 min before half of the ligation mixture was added to the cells. The suspension was mixed by flicking the tube and then incubated on ice for 30 min. Cells were heat shocked at 42 °C for 40 s and placed back on ice for 5 min. 900 µl of SOC medium were added and cells were incubated at 37 °C for 1 to 2 h before plating 200 µl of the suspension on LB plates containing 100 µg/ml ampicillin. Plates were incubated upside down at 37 °C overnight.

3.2.1.6 Plasmid DNA miniprep

Single clones were picked from LB plates and transferred into 5 ml of LB medium containing 100 µg/ml ampicillin. After incubation at 37 °C overnight, DNA isolation was carried out using the NucleoSpin® Plasmid kit according to the manufacturer's instructions. Plasmid DNA was eluted from columns in 40 µl of Millipore H_2O. This procedure was repeated by re-pipetting the eluate into the column, followed by centrifugation for 1 min (8,000 g). DNA concentration was determined using a NanoDrop® ND1000 Spectrophotometer.

3.2.1.7 Sanger sequencing

Sanger sequencing was performed to verify the correctness of clones. For sequencing, the BigDye® Terminator v1.1, v3.1 Cycle Sequencing Kit was used. The required reaction mix and thermocycler program are given in **Table 14** and **Table 15**.

Table 14: Reaction mix for Sanger sequencing

Component	Volume
Plasmid DNA (20 ng/µl)	2 µl
BigDye® Terminator Reaction Mix	0.3 µl
5x BigDye® Terminator Sequencing Buffer	2 µl
Primer (10 µM)	1 µl
H_2O (Millipore)	4.7 µl

Table 15: Thermocycler program for Sanger sequencing

Step of the reaction	Temperature	Duration	Cycles
Initial denaturation	94 °C	2 min	
Denaturation	94 °C	30 s	27
Annealing	58 °C	30 s	
Elongation	60 °C	3 min	
Final elongation	60 °C	5 min	
Break	4 °C	-	

For DNA precipitation, 5 µl EDTA (125 mM) were added followed by an incubation for 10 min at room temperature. Next, 50 µl 100 % Ethanol were added and the sample was centrifuged for at least 15 min at maximum speed. The supernatant was discarded and the sample was washed with 100 µl 70 % Ethanol. After another centrifugation step for 7 min, the supernatant was discarded again. Pellets were suspended in 20 µl of HiDi™ formamide before analysing them with the help of an Abi3130x1 Genetic Analyser. The obtained sequences were evaluated using SnapGene (version 2.8.2).

3.2.1.8 Restriction digestion

The verified DNA sequences were transferred from the pGEM®-T vector into the pCAG-EGxxFP vector. Therefore, the pGEM®-T vector was digested overnight at 37 °C using restriction enzymes (**Table 16**). The digested DNA was run on an agarose gel and fragments of correct size were excised and purified as described in 3.2.1.2 and 3.2.1.3. The DNA fragment was eluted in 20 µl Millipore H_2O and DNA concentration was determined using a NanoDrop® ND1000 Spectrophotometer.

Table 16: Reaction mix for restriction digestion of plasmid DNA

Component	Volume
Plasmid DNA	2-3 µg
Enzyme 1	0.5 µl
Enzyme 2	0.5 µl
10x NEB Endonuclease Buffer*	2.5 µl
H_2O (Millipore)	ad. 25 µl

* Dependent on the enzymes used

3.2.1.9 Ligation into pCAG-EGxxFP vector

The insert DNA and the purified digested pCAG-EGxxFP vector were ligated using the T4 DNA ligase. The required reaction mix is shown in **Table 17**. The ligation was incubated at 14 °C overnight and thereafter transformed into *E. coli*.

Table 17: Reaction mix for ligation of inserts into the pCAG-EGxxFP vector

Component	Volume
Digested pCAG-EGxxFP vector	2 µl
Insert DNA	7 µl
T4 DNA Ligase Puffer (10x)	2 µl
T4 DNA Ligase	1 µl
H_2O (Millipore)	ad. 20 µl

3.2.1.10 Colony PCR

A colony PCR was conducted to identify positively transformed *E.coli* clones. First, single clones were picked and transferred into 8 µl LB medium containing 100 µg/ml ampicillin and incubated at 37 °C for 2 to 4 h. 2 µl of this suspension were used as template for a PCR reaction, which was based on the House Taq DNA polymerase (**Table 18**). The applied thermocycler program is shown in **Table 12**.

Table 18: Reaction mix for colony PCR

Component	Volume
Buffer 10x (15 mM $MgCl_2$)	2.5 µl
Primer forward (10 µM)	1 µl
Primer reverse (10 µM)	1 µl
dNTPs (1.25 mM)	2 µl
E. coli culture	2 µl
House Taq DNA polymerase	0.5 µl
H_2O (Millipore)	16 µl

3.2.1.11 Plasmid DNA "Midi" preparation

Cloned constructs were isolated from 100 ml overnight *E. coli* cultures using the NucleoBond® XtraMidi kit according to the manufacturer's protocol. The DNA pellet was solved in 100 µl of Millipore H_2O. DNA concentration was determined using a NanoDrop® ND1000 Spectrophotometer and adjusted to 1 µg/µl. Plasmid DNA was stored at -20 °C.

3.2.1.12 Preparation of glycerol stocks for long term storage

830 µl of a fresh overnight *E. coli* culture were mixed with 170 µl sterile 87 % glycerol and immediately frozen at -80 °C. Specifications about plasmid constructs were entered into the database for glycerol cultures at the Institute of Human Genetics, Regensburg.

3.2.2 Cloning of sgRNAs

3.2.2.1 Bioinformatical sgRNA design

The UCSC genome browser [100] was used to obtain the DNA sequence of the minimal *ARMS2-HTRA1* haplotype, defined by Grassmann et al. (2017) [25]. The genome browser marked known genomic repeat regions and showed common variant (MAF > 1 %) locations. Next, the *Optimized CRISPR Design-Tool* (http://crispr.mit.edu, accessed February 1st 2018) was applied to identify potential sgRNA candidates and to estimate their on-target score. The sgRNA candidates were filtered for the following criteria: (1) On-target score of at least 50, (2) sgRNA is located outside a genomic repeat region, (3) sgRNA does not overlap a common variant, and (4) no potential off-targets in known genes. If several sgRNAs fulfilled these thresholds, the genomic position was used to manually select candidates. For later cloning processes, two oligonucleotides were designed for each sgRNA by adding a “CACCG” sequence to the 5 prime end of the forward sgRNA sequence (forward primer) and a “C” nucleotide to the 3 prime end of the reverse complement sgRNA sequence (reverse primer). All investigated sgRNAs and the respective on-target-scores are shown in **Table 4**. SnapGene (version 2.8.2) was used to visualise and to proof correct sgRNA design.

3.2.2.2 Cloning of sgRNAs into px330 vectors

All studied sgRNAs were inserted into at least one of the px330 vectors, consisting of px330-mCherry, px330-eGFP, and px330-GFPo. This procedure required multiple steps and used BpiI restriction sites.

First, the px330 vector was digested with BpiI for 30 min at 37 °C (**Table 19**). Thereafter, the reaction was purified using agarose gel electrophoresis and the NucleoSpin® Gel and PCR Clean-up kit as described in 3.2.1.2 and 3.2.1.3.

Table 19: Reaction mix for restriction digestion of the px330 vector

Component	Volume
px330 vector	1 µg
BpiI	1 µl
Quick CIP	1 µl
10x NEB fast digest buffer	2 µl
H_2O (Millipore)	ad. 20 µl

The two corresponding oligonucleotides for each sgRNA were annealed. This was conducted using the reaction mix shown in **Table 20**. Annealing was performed in a Thermocycler, starting with an incubation at 37 °C for 30 min, followed by 95 °C for 5 min and a step-wise ramp down to 25 °C at 5 °C/min.

Table 20: Reaction mix for sgRNA oligonucleotide annealing

Component	Volume
dATP (10 mM)	1 µl
Primer forward (100 µM)	1 µl
Primer reverse (100 µM)	1 µl
10 x T4 Polynucleotide Kinase Reaction Buffer	1 µl
T4 PNK	0.5 µl
H_2O (Millipore)	5.5 µl

Next, the digested px330 vector and the annealed sgRNA oligonucleotides were ligated (**Table 21**) for 10 min at room temperature.

Table 21: Reaction mix for ligation of digested px330 vector and annealed sgRNA

Component	Volume
BpiI digested px330 vector	50 ng
Annealed oligonucleotide duplex (1:200 dilution)	1 µl
2x Quickligation Buffer (Quick Ligation™ Kit)	5 µl
Quick ligase (Quick Ligation™ Kit)	1 µl
H_2O (Millipore)	ad. 11 µl

The ligation reaction was treated with the RecBCD Exonuclease to prevent unwanted recombination products. The respective reaction mix (**Table 22**) was incubated for 30 min at 37 °C.

Table 22: Reaction mix for exonuclease treatment of ligtation reactions

Component	Volume
Ligation reaction mix (**Table 21**)	11 µl
dATP (10 mM)	1.5 µl
NEBuffer™ CutSmart®	1.5 µl
RecBCD Exonuclease	1 µl

After exonuclease treatment, the ligation reaction was transformed into cells of the competent *E.coli* strain Stbl3 as described in 3.2.1.5. Single clones were verified by applying Plasmid DNA miniprep and Sanger sequencing, followed by Plasmid DNA "Midi" preparation, if required.

3.2.3 sgRNA efficiency test

3.2.3.1 Cultivation of HEK293T cells

Human embryonic kidney (HEK293T) cells were cultivated in 10 cm dishes filled with 10 ml cultivation medium (**Table 10**). HEK293T cells were passaged twice a week after reaching about 90 % confluency. Old medium was removed and cells were washed off the dish with fresh medium. HEK293T cells were seeded into a fresh 10 cm dish at a dilution of 1:10.

3.2.3.2 Transfection of HEK293T cells – calcium phosphate method

For sgRNA efficiency tests, HEK293T cells were transfected using the calcium phosphate method [101]. Cells of a confluent 10 cm dish were diluted 1:14 with cultivation medium and seeded on Poly-L-Lysine coated 6-well plates one day before

transfection. Each well on the plate contained 3 ml cultivation medium and was transfected individually. On the day of transfection, the culture medium was changed to HEK293T medium containing 1 µM Chloroquine. After one hour of incubation, the medium was changed back to 2.5 ml HEK293T culture medium. The transfection mix was prepared according to **Table 23** by first mixing DNA with H_2O followed by addition of CaCl2. Thereafter, 250 µl 2x HBS were added to the tube by gently pipetting on the bottom. The resulting two-phases were mixed by gently bubbling air drops into the solution.

Table 23: Transfection mix for calcium phosphate transfection (1 well of 6-well plate)

Component	Volume
pCAG-EGxxFP vector carrying the target sequence	1.5 µg
px330-mCherry vector carrying a sgRNA	1.5 µg
$CaCl_2$ (2 M)	31 µl
H_2O (Millipore)	ad. 250 µl

The mixture was added dropwise to the cells. 7 h after transfection, the medium was changed to HEK293T medium and cells were cultivated for another 48 h. The transfected cells were then transferred onto a black Poly-L-Lysine coated 96-well plate with transparent bottom to enable a standardised fluorescence evaluation. For this reason, the cells were detached from the 6-well plate by changing the medium to 1 ml of a trypsin solution (1x v/v in PBS). After an incubation step of 5 min at 37 °C, 2 ml of HEK293T medium were added. The cell suspension was transferred into a 15 ml falcon tube and centrifuged for 3 min at 1000 g. The supernatant was removed and 4 ml fresh medium were added to the cells. After gently mixing the suspension, 50 µl were added per well on the 96-well plate and thereafter filled up to 100 µl using HEK293T medium. The cells were cultivated for another 24 h at 37 °C.

3.2.3.3 Evaluation of sgRNA efficiency

72 h after transfection, sgRNA efficiency was analysed by measuring fluorescence intensities of transfected cells. Therefore, the culture medium of each well was changed to 100 µl 1 x PBS and the whole plate was transferred into a FLUOstar OPTIMA plate reader. Two fluorescence spectra were recorded: (1) eGFP (excitation: 488 nm, Emission 509 nm) to detect sgRNA efficiency, and (2) mCherry (excitation: 587 nm, Emission 610 nm) to evaluate transfection efficiency. eGFP raw fluorescence

counts were normalised for transfection efficiency and thereafter compared to cells, which were transfected using only pCAG-EGxxFP without px330-mCherry.

Additionally, fluorescence images were taken for documentation purposes concerning the above mentioned channels.

3.2.4 Deletion of the minimal haplotype in the *ARMS2-HTRA1* locus

The CRIPSR/Cas9 system can be applied to induce large genomic deletions. Therefore, two sgRNAs flanking the target region have to be transfected in combination with a Cas9 expression cassette.

3.2.4.1 Transfection of HEK293T cells with Lipofectamine

HEK293T cells were transfected with a combination of one px330-eGFP vector carrying the first sgRNA, which targets the upstream region of the minimal haplotype, and one px330-mCherry vector targeting the downstream region. Lipofectamine 3000 was used according to the manufacturer's protocol for 6-well plates and 1.5 µg of each vector were included in the reaction.

3.2.4.2 FACS sorting and single-cell cultivation

72 h after transfection with Lipofectamine 3000, HEK293T cells were transferred into a 15 ml falcon tube as described in 3.2.3.2 and underwent "Fluorescence activated cell sorting" (FACS). FACS was applied to filter for living cells, which showed an eGFP-, and mCherry fluorescence. Cells, which fulfilled these criteria were transferred onto one well of a Poly-L-Lysine coated 6-well plate and incubated until confluency. During that incubation, half of the medium was exchanged every second day gently by not detaching the cells from the plate. After the transfected cells reached 100 % confluency, one half of the cells was transferred into a new well for further cultivation and the other half was frozen at -80 °C for long term storage using HEK29T freezing medium.

48 h later, the cells were detached from the plate and counted using the CASY TT system. The cells were then diluted in HEK293T cultivation media to an approximate concentration of one cell in 40 µl. 40 µl of this dilution were transferred into one well of a Poly-L-Lysine coated 96-well plate until the whole plate was occupied. The cells were then monitored daily to ensure that exclusively one cell colony arose per well,

otherwise the well was excluded from further analysis. During monitoring, the medium was changed weekly until single clones reached 100 % confluence. Thereafter, cells were split 1:3 on two wells of a six well plate, one for isolation of genomic DNA (gDNA) and one for RNA extraction. The remaining cells were frozen.

3.2.4.3 gDNA isolation

gDNA of HEK293T cells was isolated following the protocol from Lairds et al. (1991) [102].

3.2.5 Measuring gene expression

3.2.5.1 RNA isolation

RNA isolation from mammalian cells was conducted using the Qiagen RNeasy Mini Kit according to the manufacturer's instructions. RNA was eluted two times in 50 µl RNase-free water and RNA concentration was determined using a NanoDrop® ND1000 Spectrophotometer. The RNA was stored at -20 °C for short term and at -80 °C for long term use.

3.2.5.2 cDNA synthesis

For complementary DNA (cDNA) synthesis, 1µg of RNA was diluted in 12.5 µl RNase-free H_2O and mixed with 1 µl of poly(dT) primer (30 nmol). The mixture was then heated to 70 °C for 5 min and thereafter the cDNA synthesis reaction mix (**Table 24**) was added. This reaction was incubated in a thermocycler for 10 min at 25 °C, followed by 42 °C for 1 h and a final step of 70 °C for 15 min.

Table 24: Composition of cDNA synthesis reaction mix

Component	Volume
5x Reaction Buffer for RevertAid™ Reverse Transcriptase	4 µl
dNTPs (1.25 mM)	2 µl
RevertAid™ Reverse Transcriptase	0.5 µl

After cDNA synthesis, 30 µl RNase-free H_2O were added to the reaction volume to dilute the cDNA for further applications. The cDNA was stored at 8 °C for short term use and at -20 °C for long term storage.

3.2.5.3 Quantitative real-time PCR

Quantitative real-time PCR (qRT-PCR) was performed with primers based on the "Universal Probe Library" by Hoffmann-La Roche. The qRT-PCR experiments were conducted in triplicates on 384-well plates using the QuantStudio™ 5 Real-Time PCR System. The reaction mix and the PCR conditions are given in **Table 25** and **Table 26**.

Table 25: Reaction mix for qRT-PCR analysis

Component	Volume
cDNA (20 ng/µl)	2.5 µl
2x TaqMan Gene Expression Master Mix	5 µl
Primer forward (10 µM)	1 µl
Primer reverse (10 µM)	1 µl
Probe	0.125 µl
H_2O (Millipore)	0.375 µl

Table 26: qRT-PCR conditions

Step of the reaction	Temperature	Duration	Cycles
Denaturation	95 °C	40 s	
Annealing	60 °C	60 s	
Elongation	72 °C	2 min	40

The data were analysed using the ΔΔCt-approach and gene expression levels were normalised in regard to the housekeeper gene "succinate dehydrogenase complex flavoprotein subunit A" (*SDHA*).

3.2.6 Targeted enhancement of gene expression

Targeted enhancement of gene expression was performed with the help of the dCas9-VPR vector generated by Chavez et al. (2015) [66]. This approach required two expression constructs: (1) the sgRNA expression cassette and (2) the dCas9-VPR encoding construct. An alternative px330 vector was generated, because the px330 vector family carries the Cas9 expression cassette, which is impedimental for gene expression enhancement. Therefore, the px330-GFPo was created by cutting out the Cas9 expression cassette of a px330-eGFP vector using the restriction enzymes EcoRI-HF and AgeI. The cloning procedure followed the protocols described in 3.2.1. To enhance gene expression, a double-transfection of the px330-GFPo vector including a sgRNA and the dCas9-VPR vector was required. This was performed in

HEK293T cells using Lipofectamine 3000 as described in 3.2.4.1. 72 h after transfection, qRT-PCR was conducted to measure the gene expression of target genes.

4 Results

4.1 A mega-analysis of eQTL in liver tissue

The first project explored the regulatory landscape of gene expression in liver tissue to understand functional consequences of genetic variants associated with complex diseases. In addition, this project should provide the basis for further eQTL studies by elaborating a detailed data analysis protocol. For this reason, publicly available data from four independent studies (**Table 27**) were collected. Each of these studies calculated eQTL in liver tissue and evaluated the results regarding different aspects. In this thesis, the studies were named after their first author in the case of (1) Schadt et al. [69], (2) Schroeder et al. [41], and (3) Innocenti et al. [47] or the respective consortium in case of (4) GTEx v6 [44]. Overall, genotype and gene expression data of a total of 588 individuals were included in the analysis.

Table 27: Study overview of datasets combined in the liver eQTL database

Study	**Schadt** [69]	**Schroeder** [41]	**Innocenti** [47]	**GTEx Start/Mid*** [44]
Sample size after QC	178	149	178	83
Origin of liver tissue	Post-mortem tissue and resections from donor livers	Normal tissue resected during surgery for liver cancer	Post-mortem tissue and resections from donor livers	Post-mortem tissue
RNA array	Agilent Custom 44k	Illumina Human WG-6v2.0	Agilent 4×44k	RNA-seq (Illumina HiSeq 2000)
Genes before QC	40,638	48,701	45,015	56,318
Genes after QC	24,123			
DNA array	Affymetrix 500k; Illumina 650 Y	Illumina HumanHap300	Illumina 610 Quad	Illumina Omni 5M/2.5M*
Variants before QC	449,699	318,237	620,901	2,526,494/ 2,378,075*
Variants after QC	383,719	296,718	545,886	2,389,798/2, 119,410*
Variants merged before imputation**	861,575			
Variants after imputation and QC	6,256,941			

QC = quality control; * GTEx v6 includes two data releases: Start and Mid, which used partially different platforms: Omni 2.5M for the first data release (GTEx start) and Omni 5M for the mid-point release (GTEx mid). ** After quality control, the genotype files of the four studies were merged into a single file and variants, which did not overlap between datasets, were assigned as missing. Variants had to be genotyped in at least 100 samples or were excluded.

The investigated liver eQTL studies used different genotyping and expression profiling platforms (**Table 27**), which demanded a stringent QC to jointly analyse the data. The QC was applied to all included individuals, genotyped variants, and the measured gene expression. A detailed overview of all QC steps is provided in the Bioinformatical protocols section. Briefly, only individuals of European descent with low missing rates of genotype and gene expression data were included. The QC of genotyped variants filtered for variants: (1) measured in all datasets, (2) with allele frequencies comparable to the 1000 Genomes Project reference panel, (3) located on autosomes, (4) with MAF above 5 %, and (5) no significant deviation from HWE. This procedure resulted in 861,575 variants for imputation. The gene expression data underwent a separate QC depending on the data source. 24,123 genes, which were measured in at least two datasets were considered for further data processing.

4.1.1 Elaboration of a data-normalisation protocol

Each of the four studies used distinct platforms and data processing protocols, which required a normalisation pipeline. Normalisation was necessary for genotype and gene expression data. The different genotype files were combined and imputed using the same reference panel. This enabled the analysis of 6,256,941 shared genetic variants. The gene expression data underwent different processing protocols before joint analysis because three studies used microarray platforms, whereas the GTEx data were based on RNA-Seq (**Table 27**). Therefore, gene expression values were merged into one matrix and log2 transformed to evaluate potential cofounder effects by PCA. This analysis showed that samples of the same dataset clustered together and that the range of expression values varied between the studies (**Figure 6 A** and **D**).

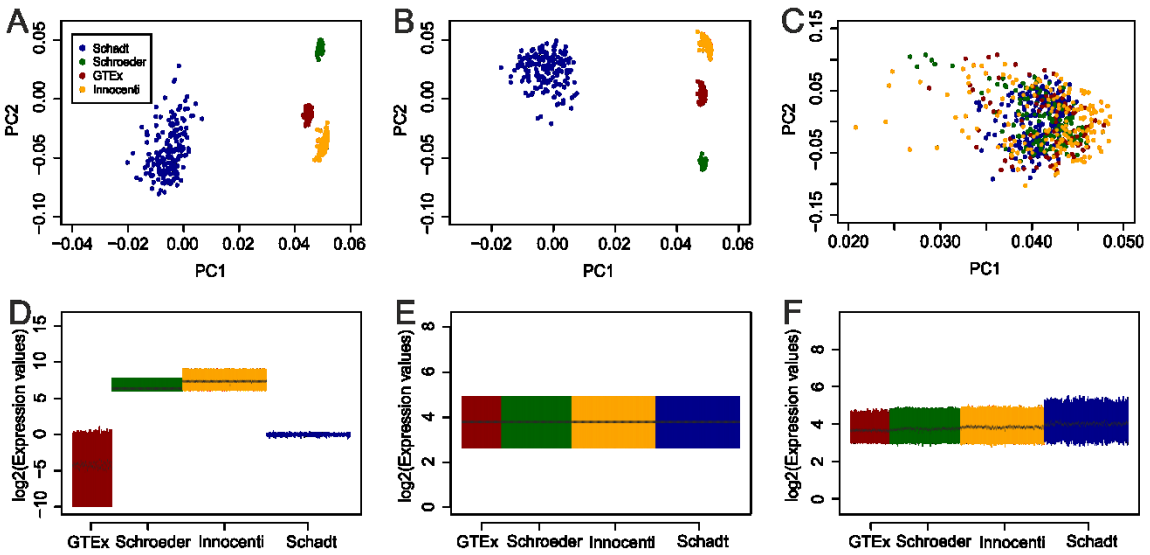

Figure 6: Gene expression data normalisation process.
A PCA was conducted on the merged gene expression data of the four datasets (GTEx, Innocenti, Schadt, Schroeder), at three different consecutive normalisation steps: (A) raw log2 transformed merged data (no normalisation), (B) quantile normalised data and (C) after adjustment for known batch effects using ComBat. In addition, the gene expression values are presented as boxplots at the same stages (D-F). (Figure published in Strunz et al., 2018 [103])

Next, quantile normalisation (QN) was performed to adjust gene expression values in regard to their scale. After QN, the datasets Schroeder, Innocenti, and GTEx converged regarding principal component (PC) 1. In addition, gene expression value ranges showed comparable median values and variability (**Figure 6 B** and **E**). Since QN alone was not sufficient to normalize all studies, an empirical batch correction method called ComBat [93] was applied. After these normalisation steps, clustering of individuals with regard to their original dataset was not apparent to any further extent (**Figure 6 C** and **F**).

4.1.2 Analysis of local eQTL

eQTL calculation was first performed for each of the four studies separately using a linear regression model, which was adjusted for several covariates and included one gene and one variant at a time. Only local eQTL were considered for further analysis by investigating a window of 1 Mbp up- and downstream of the transcription start site or polyadenylation site of a gene locus. Next, mixed effects models were applied to perform a meta-analysis based on the effect sizes and standard errors of each study. These models estimated one joint effect size, standard error and a combined P-value for each eQTL. All P-values were adjusted for multiple testing by calculation of the FDR [104] and Q-values smaller than 0.001 were considered statistically significant. At this

threshold, 101,148 eVariants and 1,313 genes regulated by eQTL were identified (**Table 28**). Remarkably, only 38.5 % (see GTEx Start/Mid) to 60.9 % (see Innocenti) of significant eGenes in the single studies remained significant in the meta-analysis.

Table 28. eQTL results of single datasets and the merged analyses

		Schadt	Schroeder	Innocenti	GTEx Start/Mid	Meta-Analysis	Mega-Analysis
Q-value < 0.05	eQTL	73,999	165,518	122,474	54,639	222,521	444,276
	eVariants (unique)	68,636	154,799	114,635	49,176	205,942	383,213
	eGenes (unique)	1,592	3,453	2,635	1,983	4,811	7,612
	Overlapping eGenes Meta-analysis	802 (50.38 %)	1,578 (45.7 %)	1,379 (52.33 %)	661 (33.33 %)	4,811 (100 %)	4,486 (58.93 %)
	Overlapping eGenes Mega-analysis	1,100 (69.1 %)	2,168 (62.79 %)	1,805 (68.5 %)	1,023 (51.59 %)	4,486 (93.24 %)	7,612 (100 %)
Q-value < 0.001	eQTL	29,546	71,423	52,565	19,802	101,148	202,489
	eVariants (unique)	27,689	69,292	49,594	16,953	95,257	183,872
	eGenes (unique)	363	913	670	387	1,313	1,959
	Overlapping eGenes Meta-analysis	215 (59.23 %)	491 (53.78 %)	408 (60.9 %)	149 (38.5 %)	1,313 (100 %)	1,260 (64.32 %)
	Overlapping eGenes Mega-analysis	288 (79.34 %)	688 (75.36 %)	537 (80.15 %)	207 (53.49 %)	1,260 (95.96 %)	1,959 (100 %)
P-value < 1 x 10^{-6}	Independent Signals	-	-	-	-	-	2,060

Data preparation and QC of the four datasets further allowed to jointly analyse the merged genotype and gene expression data by calculation of eQTL in the entire database. This mega-analysis is known to have a higher statistical power in comparison to the classical meta-analysis approach [48,105]. The mega-analysis yielded 202,489 statistically significant eVariants affecting the expression of 1,959 genes (Q-value < 0.001). Compared to the results from the meta-analysis, the mega-analysis provided a two-fold increase in the number of eVariants and a 1.5-fold increase in the number of differentially regulated genes. Both, mega- and meta-analysis discovered more significant results than any of the four individual studies alone. Furthermore, the overlap of single study results and the mega-analysis is on average 19 % higher (53.5 to 80.15 %) than the overlap observed with the meta-analysis (**Table 28**). Because of these observations, all further evaluations were based on the mega-analysis results. Moreover, the mega-analysis enabled the detection of independent eVariants using a conditional eQTL analysis. Therefore, the eQTL analysis was repeated for each significant eGene after additionally adjusting the linear regression model for the most significant eVariant identified for the respective gene. P-values lower than 1.00 x 10^{-6} were considered significant (corresponding to a Q-value of 0.001 in the primary mega-analysis). The procedure was repeated until no further significant independent eVariants were found. With this approach, 101 additional independent eVariants regulating 93 of the 1,959 liver eGenes were identified. Interestingly, several independent signals would have not been considered significant (Q-value < 0.001) in the primary mega-analysis (**Figure 7**).

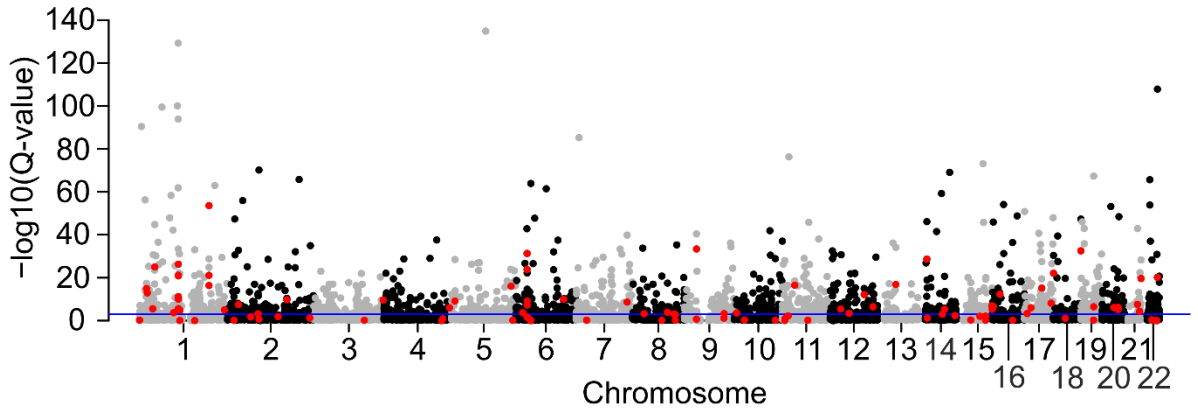

Figure 7: Manhattan plot of the eQTL mega-analysis in liver.
A mega-analysis was conducted including 588 samples of four independent studies detecing eVariants in liver tissue. The Manhattan plot shows the −log10 Q-values of the most significant eVariant for each of the 24,123 analysed autosomal genes. Additionally, 101 independent secondary signals were identified and are highlighted in red. The blue line depicts the threshold for significance 1.00 x 10^{-3}. (Figure published in Strunz et al., 2018 [103])

4.1.3 Characterisation of eVariants in liver tissue

The liver eQTL results were further evaluated to better understand potential molecular mechanisms. First, the most significant eVariant and independent signals for each eGene were plotted in regard to their genomic position (**Figure 8**).

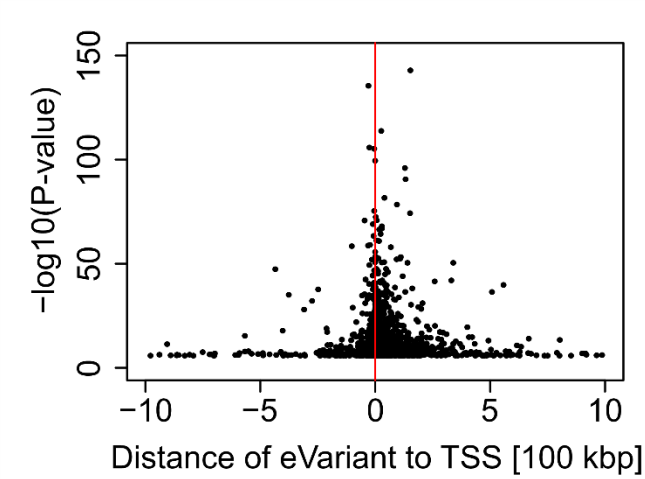

Figure 8: Characterisation of independent eVariants based on their genomic localisation.
The distance to the transcription start site (TSS, red line) is plotted against the -log10 P-values of the most significant eVariant for the respective eGene, including secondary signals (independent hits). Negative/positive distances denote that the variant is located upstream/downstream of the TSS in regard to the direction of transcription. (Figure published in Strunz et al., 2018 [103])

Most of the significant eVariants were located close to the respective TSS. Altogether, 1,599 out of 2,060 independent eVariants were located within 100,000 base pairs around the TSS. Nevertheless, 55 eVariants were located more than 500 kbp away from the regulated eGene.

In a next step, eVariants were further characterised in regard to known DNA features and regulatory elements by searching RegulomeDB [106]. This database applies a seven-level functional scoring system to grade genetic variants. Category one variants affect very likely transcription factor binding and alter gene expression, whereas category 7 variants lack evidence for any functional relevance. Altogether, three groups of variants from the liver eQTL database were evaluated: (1) all unique significant eVariants of the mega-analysis (N = 183,872), (2) the most significant eVariant per eGene and the independent signals (N = 2,060), and (3) a random set of 200,000 genetic variants within 1 Mbp of a gene locus, which served as “control” (**Figure 9 A**). Remarkably, the first set including all eVariants was enriched in RegulomeDB classes one to four (P-values < 6.82×10^{-09}). In addition, the second set of independent signals revealed an even stronger enrichment in classes one to four compared to controls and compared to all eVariants (P-values from 1.72×10^{-04} to 8.27×10^{-11}).

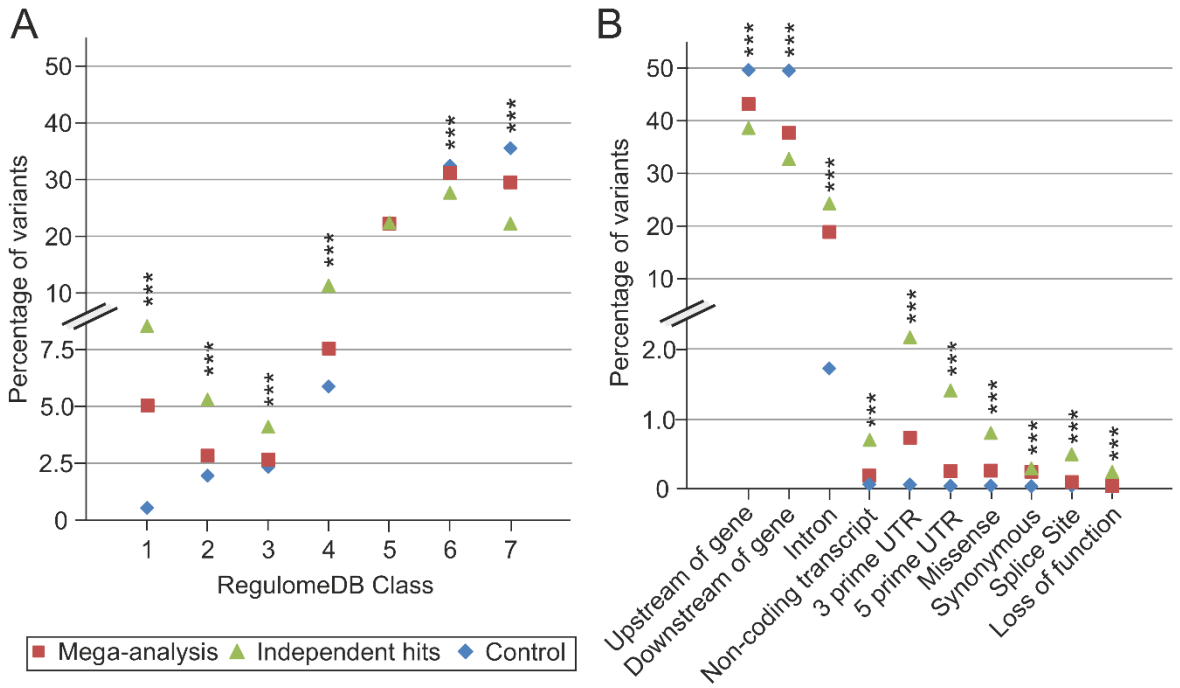

Figure 9: Functional annotations and predicted consequences of local eVariants.
Three sets of variants were evaluated by employing two different databases. Set one (mega-analysis) consists of all significant mega-analysis eVariants (Q-Value < 0.001) while the second group comprises the most significant eVariant and the independent hits for each eGene. Set three (control) includes random variants of the imputed genotype file, which are located next to at least one gene within a distance of a maximum of 1 Mbp. (A) The chart depicts the percentage of variants per variant set categorised into seven groups by RegulomeDB. The seven-level functional score is based on a synthesis of data derived from various sources: category 1 variants are very likely to affect transcription factor binding and are linked to gene expression of a target gene (i.e. are known eVariants); categories 2 and 3 are likely to affect at least transcription factor binding and several other regulatory effects; categories 4-6 show minimal functional indication while category 7 variants lack evidence for any functional relevance.(B) The chart shows the percentage of variants classified into ten classes of consequences according to the Ensembl Variant Effect Predictor (VEP). For variant set one (mega-analysis) and two (independent hits), only the predicted consequence affecting the identified eGene was included. For the control group, one random gene within a variant–gene distance of a maximum of 1 Mbp was chosen. If the variant had different effects on transcripts of the same gene, the most severe effect was selected. *** P-value for difference between groups < 0.001. (Figure published in Strunz et al., 2018 [103])

Besides characterisation of eVariants in regard to transcription factor binding and gene regulation, another database was used to analyse potential molecular mechanisms based on gene structure and variant position. The ensembl variant effect predictor (VEP) [107] rates variants in regard to all surrounding transcripts and classifies them according to potential functional consequences. Control variants were predominantly located upstream (49.22 %) and downstream (49.09 %) of known gene structures. Another 1.63 % of the control variants were found in introns of genes. Less than 0.1 % of the control variants were assigned to functional categories such as missense or untranslated transcript region (UTR). Interestingly, the proportion of intronic variants was significantly larger in both, the mega-analysis variants (19.72 %, $P < 1.00 \times 10^{-150}$) and the independent hit variants (29.17 %, $P < 1.00 \times 10^{-150}$) (**Figure 9 B**). Additionally,

other predicted categories like UTR or coding region variants occurred more often (P-values < 1.72×10^{-07}).

Taken together, these findings indicate that significant eVariants are more often localised within known gene structures and are likely regulatory variants as they are found within regions of transcription factor binding and open chromatin. This is especially the case for the most significant eVariants and independent secondary signals.

4.1.4 Liver eQTL of AMD-associated variants

The liver eQTL database was further used to identify molecular mechanisms, which might be relevant for AMD aetiology. For this reason, the 52 independent AMD-associated variants identified by Fritsche et al. (2016) [18] were investigated in regard to gene expression regulation in liver. 31 of these 52 variants were successfully genotyped or imputed in the liver eQTL database and showed an allele frequency > 5 %. Interestingly, 8 of these variants were associated with gene expression of 15 unique eGenes (Q-value < 0.05, **Table 29**).

Table 29: Liver eVariants overlapping with genome-wide significant AMD-associated variants

IH*	dbSNP ID	CHR	Position [hg19]	Gene Symbol	eQTL Q-Value	Effect size**	SE	Non-risk allele	Risk allele
1.2	rs570618	1	196,657,064	*CFHR1*	4.34E-10	0.711	0.099	G	T
1.1	rs10922109	1	196,704,632	*CFHR4*	1.66E-21	1.118	0.105	A	C
1.1	rs10922109	1	196,704,632	*CFHR1*	2.54E-21	0.992	0.094	A	C
1.1	rs10922109	1	196,704,632	*CFHR3*	2.11E-14	0.923	0.107	A	C
1.1	rs10922109	1	196,704,632	*F13B*	0.012	0.216	0.057	A	C
1.1	rs10922109	1	196,704,632	*CFH*	0.025	0.338	0.095	A	C
1.6	rs61818925	1	196,815,450	*CFHR3*	1.55E-06	0.649	0.113	G	T
1.6	rs61818925	1	196,815,450	*CFHR1*	0.006	0.416	0.103	G	T
1.6	rs61818925	1	196,815,450	*CFHR5*	0.011	-0.371	0.096	G	T
11	rs7803454	7	99,991,548	*PILRB*	5.72E-24	0.251	0.022	C	T
11	rs7803454	7	99,991,548	*PILRA*	1.04E-08	0.372	0.056	C	T
23.1	rs2043085	15	58,680,954	*ALDH1A2*	0.016	0.207	0.056	T	C
23.2	rs2070895	15	58,723,939	*LIPC*	6.88E-07	0.561	0.095	A	G
23.2	rs2070895	15	58,723,939	*ADAM10*	0.021	-0.217	0.06	A	G
24.2	rs17231506	16	56,994,528	*CETP*	0.008	-0.216	0.055	C	T
27	rs6565597	17	79,526,821	*TSPAN10*	2.46E-07	-0.526	0.086	C	T
27	rs6565597	17	79,526,821	*ACTG1*	0.016	0.312	0.084	C	T
27	rs6565597	17	79,526,821	*ANAPC11*	0.036	-0.171	0.05	C	T

CHR: chromosome; SE: standard error of the effect size; * IH: Independent hit according to Fritsche et al. (2016) [18] ** Effect size of a single AMD risk increasing allele

Several of the AMD-associated variants are located in the *CFH* locus (IH 1) and influence gene expression of *CFH* and *CFHR* genes. Particularly, the independent hit variant rs10922109 (independent hit 1.1 in Fritsche et al. 2016 [18]) tags a common deletion of *CFHR1*/*CFHR3*. Since the deletion of both genes is protective against AMD, the risk increasing allele results in an elevated expression of the two genes, which is represented by the respective effect sizes in **Table 29** (rs10922109 - *CFHR1*: 0.992 and rs10922109 - *CFHR3*: 0.923). Besides the *CFH* locus, two other eGenes are well known in AMD-related research: *LIPC* and *CETP*. Both genes are be involved in HDL metabolism and are specifically well characterised in liver tissue.

4.2 Investigation of local eQTL in the GTEx project

Several studies showed that regulation of gene expression is a tissue dependent process [108,109]. The GTEx project measured genotype and gene expression data of various tissues from more than 600 donor individuals. These data were composed using clearly defined sample collection criteria and sample processing steps [44,46]. Furthermore, the GTEx consortium initially performed the tissue-specific analysis of local eQTL and made a curation of their significant results accessible online. Nevertheless, not all of the results are available through their online repository. For this reason, one objective of this thesis was to download the raw data of the GTEx project and to create an openly accessible in-house database at the Institute of Human Genetics Regensburg. This database was generated based on the data processing protocol of the above presented eQTL analysis in liver tissue. The in-house GTEx database was created with GTEx version 6 (v6) and later updated to GTEx version 7 (v7), which included additional samples and used whole genome sequencing instead of genotyping microarrays. **Supplementary Table 1** summarises the information for the 48 tissues of GTEx v7, which were integrated and analysed. The sample size varied from 72 (see “Brain substantia nigra” and “Minor salivary gland”) to 418 (see “Muscle skeletal”) with a mean sample size of 183.6 (SD 94.4) across all tissues. The mean number of expressed genes per tissue was 29,591.9 (SD 3,065.9) (**Figure 10**). Remarkably, in testis (sample size: 197) 42,810 genes were expressed, which equates to 76.2 % of all 56,202 in GENCODE version 19 annotated genes [110].

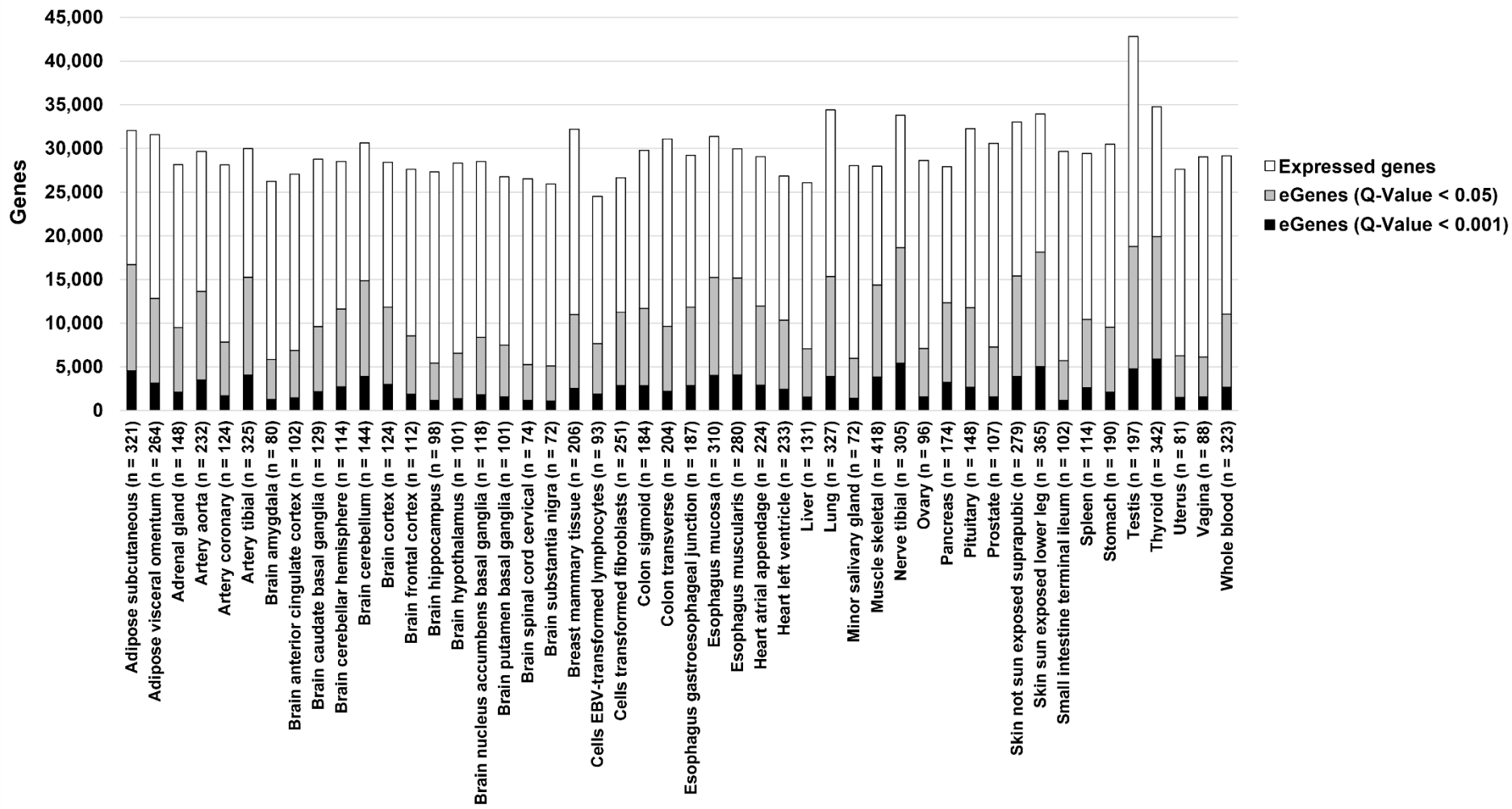

Figure 10: Expressed genes and eGenes of GTEx v7.
GTEx v7 compromises gene expression, genotype, and covariate data of 48 different tissues and cell types. Local eQTL were calculated for each tissue seperately and adjusted for multiple testing (Q-value). The barplot visualises the number of expressed genes per tissue and the identified eGenes using two significance thresholds: Q-value < 0.05 (grey) and Q-value < 0.001 (black). The sample size for each tissue (n) is given in brackets.

The number of eGenes varied widely from 19.4 % (5,741 of 29,667 genes, see “Small intestine terminal ileum”) to 57.17 % (19,890 of 34,789 genes, see “Thyroid”) of all expressed genes in the respective tissue (Q-value < 0.05). A linear regression model showed that the number of expressed genes significantly (P-value: 0.000315, R2: 0.23) correlates with the sample size per tissue (**Figure 11 A**). Remarkably, another analysis revealed an almost linear relationship (P-value: 2.38 x 10^{-19}) with an R2 of 0.83 between the tissue-specific sample size and the number of detected eQTL (**Figure 11 B**).

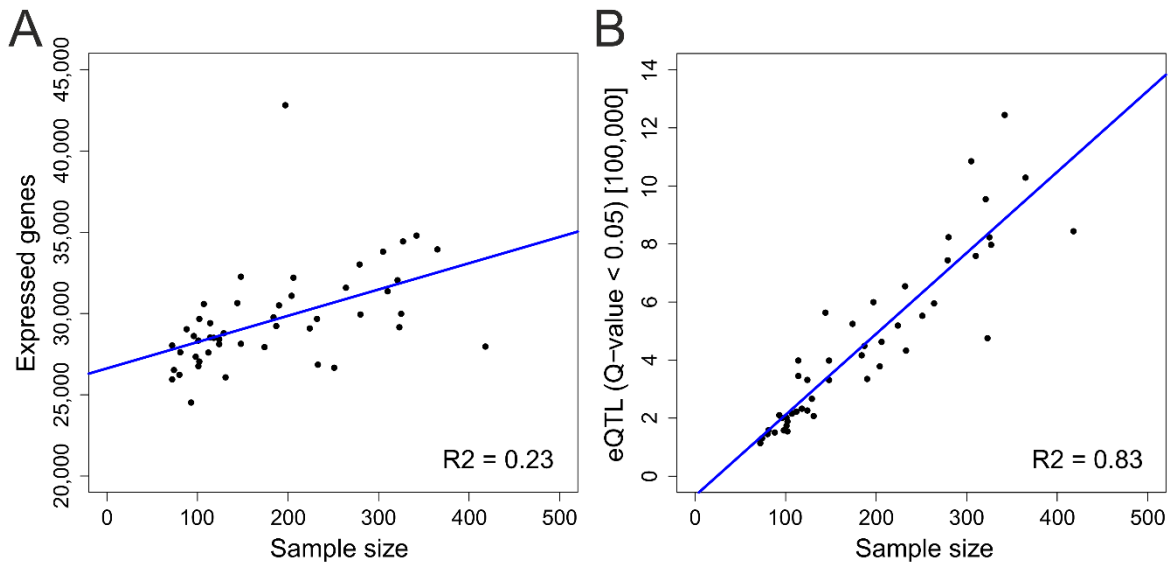

Figure 11: Correlation of sample size and tissue-specific paramters of GTEx v7.
A linear regression model was used to investigate the correlation of the tissue-specific sample size with the respective number of (A) expressed genes and (B) eQTL (Q-value < 0.05). The regression line is depicted in blue and the regression coefficent (R2) for each model is shown in the bottom right corner.

Altogether, the in-house GTEx database included eQTL data regarding 48 tissues and was created as a basis to enable further projects outside the scope of this thesis. These projects included for example the calculation of combinatory effects regarding AMD-associated eVariants and the evaluation of potential pleiotropic effects of eVariants.

4.3 Distant eQTL in the *ARMS2-HTRA1* locus

4.3.1 Distant eQTL calculation

Processing of the GTEx database enabled various further projects besides the calculation of local eQTL. One of these projects aimed at elucidating potential distant eQTL effects of AMD-associated variants and focused on the *ARMS2-HTRA1* locus at 10q26. This locus showed the most significant AMD-association in the European

population (P-value 6.5 x 10^{-735}) and the highest OR (2.81) of all 34 loci identified by Fritsche et al. (2016) [18]. The low P-values and the high LD in the *ARMS2-HTRA1* locus (**Figure 2 B**) initially hindered detailed statistical investigations. Finally, a haplotype analysis of Grassmann et al. (2017) [25] refined the AMD-associated signal to a region of 5,196 bp (chr10:124,210,369-124,215,565, hg19), called the "minimal haplotype". Additionally, the locus contains two variants, which are known to locally regulate the gene expression of *ARMS2* through different mechanisms. rs3750846, the lead variant of the study from Fritsche et al. (2016) [18], co-localises with a deletion of the *ARMS2* gene. The other variant, rs2736911 results in a truncated *ARMS2* protein (R38X). Interestingly, rs2736911 was not found to be associated with AMD [22].

To investigate potential regulatory mechanisms, local and distant eQTL were investigated for the *ARMS2-HTRA1* locus in all GTEx v6 tissues, since GTEx v7 was initially not available. After the eQTL calculation, a meta-analysis jointly evaluated single tissue results. In this analysis, both variants regulate the expression of *ARMS2* (Q-values: rs3750846 1.5 x 10^{-09}, rs2736911 2.8 x 10^{-31}). Altogether the expression of 1,098 respectively 1,120 eGenes was significantly (Q-value < 0.05) associated with rs3750846 or rs2736911. To identify different regulatory effects, the gene lists were filtered to exclude (1) genes regulated by both variants, (2) genes, which expression was correlated with *ARMS2* expression, and (3) genes involved in housekeeping processes. Housekeeping genes were identified by sorting out genes matching the GO processes including the phrases: "ribonucleo" and "metaboli". Filtering was performed to identify the potentially AMD-associated mechanism separated from the shared regulation of *ARMS2*. Interestingly, a gene enrichment analysis showed that the gene list of rs3750846 included mainly immune system related genes, whereas rs2736911 regulates genes involved in cell cycle processes (**Table 30**).

Table 30: Ten most significant gene enrichment analysis results of eGenes associated with rs3750846 or rs2736911

rs3750846 (922 genes)			rs2736911 (962 genes)		
GO term name	**eGenes in GO term**	**Adjusted P-value**	**GO term name**	**eGenes in GO term**	**Adjusted P-value**
Neutrophil mediated immunity	52	6.69E-05	Cell cycle	152	7.06E-11
Myeloid leukocyte activation	59	1.38E-04	Organelle organisation	262	2.71E-09
Myeloid cell activation involved in immune response	53	2.69E-04	Cilium assembly	44	6.33E-06
Neutrophil degranulation	49	4.67E-04	Ciliary basal body docking	21	1.59E-05
Response to stress	228	5.48E-04	Antigen processing and presentation of exogenous antigen	28	4.83E-05
Multi-organism process	159	1.22E-03	Negative regulation of ubiquitin-protein transferase activity	17	8.53E-04
Translational elongation	20	2.08E-02	Cell division	55	3.26E-03
Acute inflammatory response	19	3.51E-02	Intracellular transport	135	3.99E-03
Response to biotic stimulus	68	3.99E-02	Chromosome segregation	37	6.87E-03
Protein folding	26	4.10E-02	Protein deneddylation (removal of the ubiquitin-like protein NEDD8)	6	6.90E-03

Taken together, rs3750846 regulates 922 genes, which expression showed no association with the non AMD-associated variant rs2736911, and which were enriched for immune system related processes. To further narrow down this gene list, a mega-analysis including all GTEx v6 tissues was conducted based on the merged and normalised gene expression files. Furthermore, the mega-analysis was adjusted for tissue donors because some individuals donated multiple organs. After filtering for significant eGenes (Q-value < 0.01), which were not involved in housekeeping processes, rs3750846 regulated the expression of 455 genes. Again, *ARMS2* revealed the most significant result (Q-value 3.7×10^{-12}). The mega-analysis approach facilitated to conduct a conditional analysis, which was adjusted for the expression of the most significant gene and was repeated until none of the primary significant signals (round 0) remained. Interestingly, the adjustment for *ARMS2* expression (round 1) did not affect the significance of any other eGene (**Figure 12**). The most significant gene after adjustment for *ARMS2* was *CD300E* (Q-value 1.3×10^{-12}), which is known to participate in innate immune response [111–113]. Adjustment for *CD300E* resulted in 114, mostly immune related, genes losing significance (arrow, **Figure 12**). The subsequent adjustments for *XKR9* and *KLHDC4* altered the list of eGenes only marginally, whereas *ZNRD1* (round 5) resulted in once more 102 eGenes loosing significance.

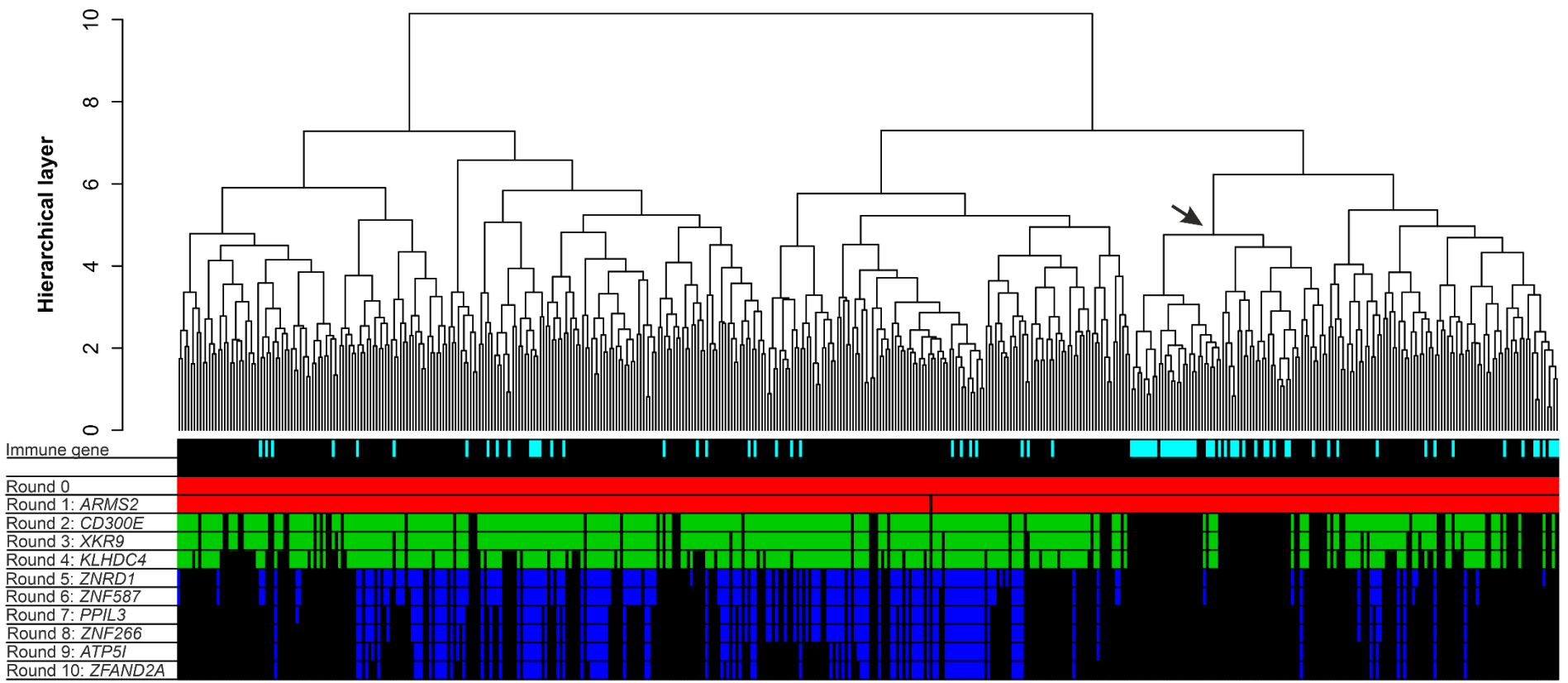

Figure 12: Conditional mega-analysis of rs3750846-associated eGenes in GTEx v6.
Gene expression and genotype files from all GTEx v6 tissues were merged to conduct a mega-analysis regarding rs3750846. The eQTL analysis resulted in 455 genes which were clustered based on their gene expression using the hclust function in R and are shown as dendrogram (top). The bar below the dendrogram visualises if a gene is known to participate in immune system processes ("Immune gene", turquoise). After the primary analysis (round 0), the eQTL calculation was adjusted for the most significant gene and repeated as long as at least one eGene reached significance (Q-value < 0.01, bars from top to bottom). Genes, which lost significance turn black in this schematic figure. The three colors red, green, and blue mark if an adjustment led to noticable changes in the list of significant eGenes, determined by another clustering analysis. The highlighted cluster (arrow) marks immune genes, which lost significance after adjustment for *CD300E* (round 2).

After the conditional mega analysis, the hypothesis emerged suggesting that the strong AMD-association of rs3750846 could be caused by distant effects on gene expression, which are shared by various tissues and cell types. Several parameter were chosen to further evaluate rs3750846-associated eGenes and to finally test the hypothesis *in vitro*. The eGenes were categorised for (1) high absolute effect sizes (> 0.05) in the mega-analysis and (2) for regulation by local eVariants (Q-value < 0.05). If this was the case, the respective local eVariants were explored in the AMD GWAS data as given in Fritsche et al. (2016) [18] for their AMD-association (Q-value < 0.05). This procedure was applied to validate the potential relevance of the eGene in the context of AMD. Furthermore, the eGenes of interest were queried for immune-related GO terms, and if they were shown to be expressed in HEK293T cells. These criteria resulted in 13 potential candidate genes, which fulfilled most aspects (**Table 31**).

Table 31: Manually curated list of potential rs3750846 target genes for experimental validation

Symbol	Strong effect of rs3750846 in mega-analysis (ABS > 0.05)*	Local AMD-associated eVariants**	Immune related	Expressed in HEK293T***
C17orf62	- (-0.01)	+	-	+
CD300E	+ (-0.065)	-	+	+
CYP1A1	+ (0.093)	-	-	+
DAZAP1	- (-0.006)	+	-	+
DEFA5	+ (-0.091)	+	+	+
FCN1	- (-0.045)	+	+	+
FLOT2	- (-0.011)	+	-	+
IL6	+ (-0.063)	-	+	+
LILRA3	+ (-0.1)	+	+	NA
MUC7	+ (-0.127)	+	+	+
NFKB1	- (-0.007)	+	+	+
PILRB	- (0.011)	+	+	NA
TNFAIP1	- (-0.011)	+	+	+

* Effect size of the AMD risk increasing allele, ** Fritsche et al. (2016) [18] Q-value < 0.05 (calculated over all GWAS variants), *** Mean expression of untreated HEK293T cells of three studies [114–116]; NA = gene was not measured or not detected

4.3.2 Genome editing to delete the minimal haplotype in HEK293T cells

After bioinformatical analysis of the 10q26 locus, an experimental approach was chosen to evaluate the hypothesis regarding distant regulatory mechanisms of AMD-associated variants located in the minimal haplotype region. The experiments were designed to experimentally manipulate the *ARMS2-HTRA1* locus using the CRIPSR/Cas9 system [117] (**Figure 13**).

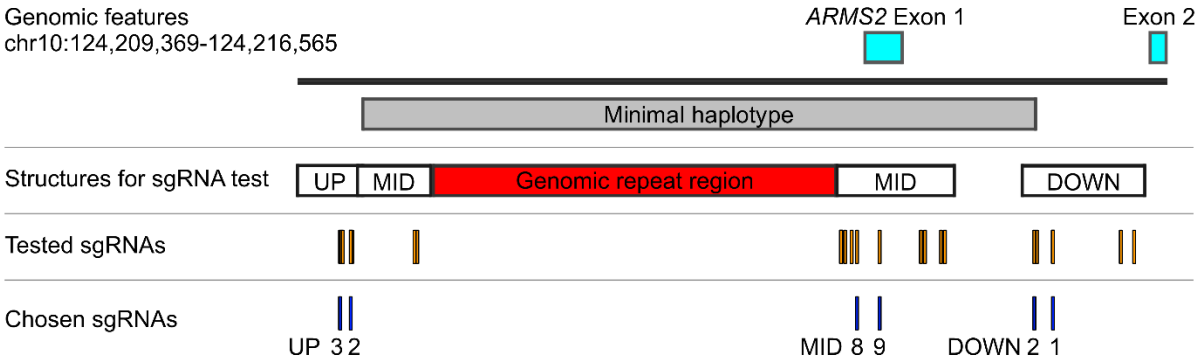

Figure 13: Scaled overview of the genomic region flanking the minimal haplotype.
Grassmann et al. (2017) [25] performed an haplotype analysis of the *ARMS2-HTRA1* locus and identified a 5,196 bp (chr10:124,210,369-124,215,565, hg19) genomic region, which most likely harbours the variants causative for the GWAS signal. Several sgRNAs (orange) were designed upstream (UP), within (MID), and downstream (DOWN) of the minimal haplotype region. After sgRNA specificity testing, six sgRNAs (blue) were chosen for further experiments. No sgRNAs were designed to target the genomic repeat region (red), because these might also bind to other regions in the genome. The figure shows the genomic region chr10:124,209,369-124,216,565 and was scaled to correctly present the positions of all shown elements.

sgRNAs were created to recruit the Cas9 endonuclease and to introduce DSBs at the *ARMS2-HTRA1* locus. Subsequent recombination events are expected to result in a deletion of all or parts of the minimal haplotype region. Five sgRNAs were bioinformatically designed to bind up- (UP) or downstream (DOWN) of the minimal haplotype. These sgRNAs were tested for specificity using the pCAG-EGxxFP system established by Mashiko et al. (2013) [118] (**Figure 14 A**). The pCAG-EGxxFP vector contains an EGFP expression cassette, which is interrupted by the sgRNA target sequence. If the sgRNA specifically binds its target, the Cas9 endonuclease is recruited and introduces a DSB. The subsequent recombination event restores the EGFP cassette and leads to a fluorescence signal, which can be detected via microscopy. The number of positively transfected cells showing green fluorescence serves as quantitative marker for sgRNA specificity. **Figure 14 B** presents a representative set of experiments included in the testing of 5 UP sgRNAs. These were separately cloned into the px330-mCherry vector and transfected into HEK293T cells in combination with the corresponding pCAG-EGxxFP vector.

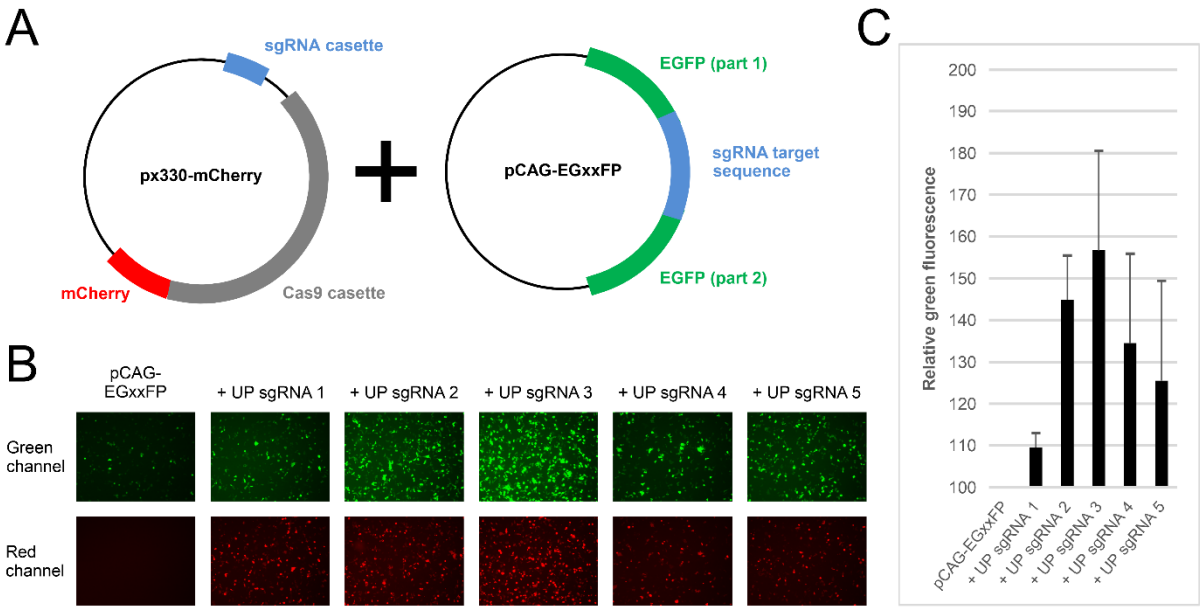

Figure 14: Specificty test of UP sgRNAs.
(A) Schematic overview of the vector set required for the sgRNA specificty test. The px330-mCherry vector carries one sgRNA- (blue) and a Cas9 (grey) expression cassette follwed by a mCherry enconding sequence (red). The pCAG-EGxxFP construct carries an EGFP expression cassette (green) interrupted by the respective sgRNA target sequence (blue). (B) Exemplary set of experiments to test the efficiency of five sgRNAs located upstream (UP) of the minimal haplotype defined by Grassmann et al. (2017) [25]. Each sgRNA was cloned into the px330-mCherry vector and double transfected in combination with the corresponding pCAG-EGxxFP construct. Green flourescence represents sgRNA specificity, whereas red flourescence marks the transfection efficency of px330-mCherry. (C) Quantitative evaluation of three independent UP sgRNA tests using the FLUOstar OPTIMA plate reader. Measurement values were normalised to the green background flourescence of the pCAG-EGxxFP vector (top left in B) and to the mean transfection efficency (red flourescense) per experiment.

After quantitative evaluation of sgRNA specificity, two sgRNAs upstream (UP sgRNA 2 and 3, **Figure 14 C**) and downstream (DOWN sgRNA 1 and 2) were chosen for the targeted deletion of the minimal haplotype (**Figure 13**). Therefore, a combination of one UP (px330-eGFP vector) and one DOWN sgRNA (px330-mCherry vector) was transfected into HEK293T cells. After an incubation time of 72h, FACS sorting was performed to identify cells positively transfected with both constructs. Then, single cells were isolated using a dilution series and seeded onto new plates with a statistical dilution of one cell per well. Two PCR reactions targeting the minimal haplotype region (**Figure 15 A**) enabled the identification of introduced genomic alterations. Altogether, 18 single clones homozygous for the deletion were identified (**Figure 15 B**). Additional 18 clones did not show any recombination events and served as controls, since they underwent the same processing protocol.

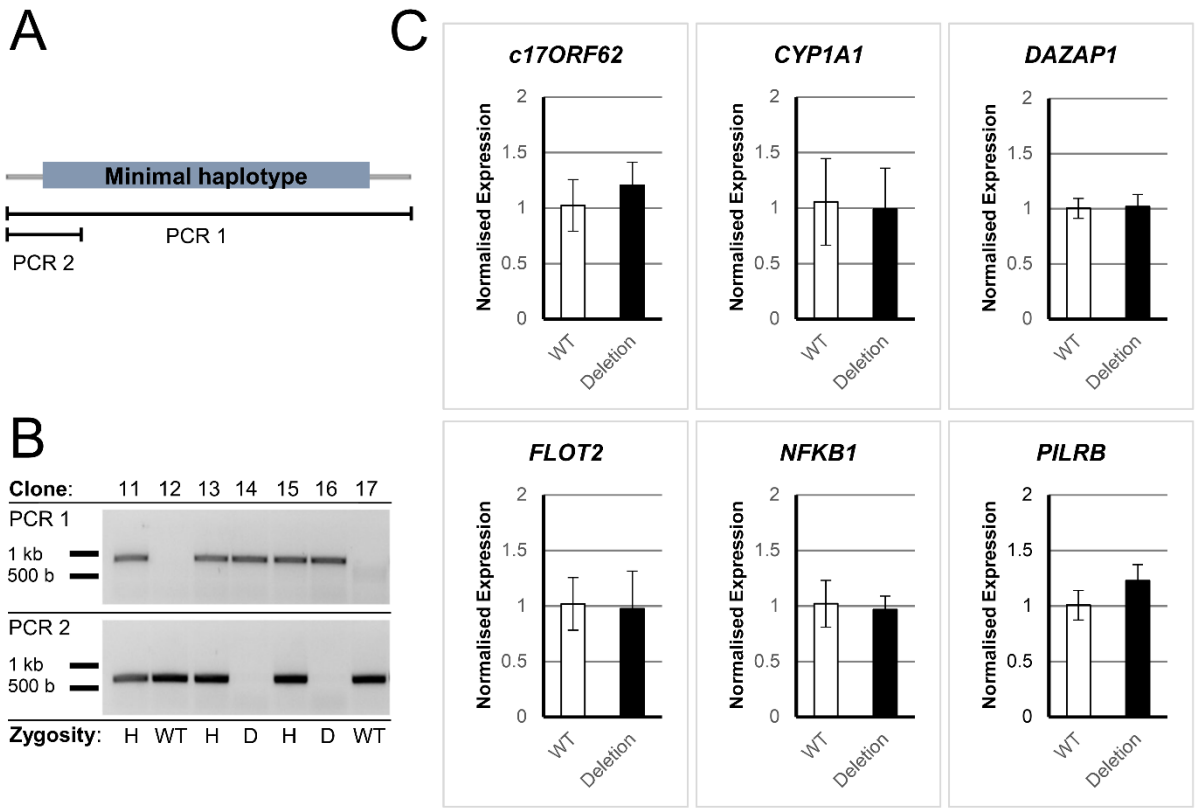

Figure 15: Genotyping and qRT-PCR of HEK239T cells edited in the *ARMS2-HTRA1* locus.
(A) Two PCRs were conducted to genotype HEK293T single clones after genome editing with one sgRNA binding upstream and one sgRNA binding downstream the minimal haplotype region. The regions covered by PCR 1 and 2 are visualised by the black lines above the annotation. The elongation time for both PCRs was 1 min, which is too short to amplify the full minimal haplotype region with PCR 1. Therefore, no amplicon of PCR 1 indicates that no deletion occured. (B) Genoytpe PCR results of seven representative single clones. The zygosity state was determined based on the results of PCR 1 and 2 and is given as: Homozygous for minimal haplotype deletion (D), hemizygous (H), or wild type (WT). The PCRs were replicated indedpently for at least two times to validate genotyping results. (C) qRT-PCR results regarding 6 exemplary target genes (**Table 31**). Shown are the mean values of 7 WT clones and 8 deletion clones. The results were normalised in regard to the respective WT clones

qRT-PCRs regarding the potential target genes (*C17orf62*, *CD300E*, *CYP1A1*, *DAZAP1*, *DEFA5*, *FCN1*, *FLOT2*, *IL6*, *LILRA3*, *MUC7*, *NFKB1*, *PILRB*, and *TNFAIP1*) of the *ARMS2-HTRA1* locus did not reveal any significant differences in gene expression despite the deletion of the minimal haplotype region (**Figure 15 C**). It is important to note that no implications about the potential effect direction are possible because eQTL results were based on the AMD risk allele (**Table 31**) but in this approach the whole minimal haplotype region was deleted.

4.3.3 Enhancing gene expression in the minimal haplotype region

Besides the deletion of the minimal haplotype region, a further approach aimed to enhance its potential influence on gene expressing regulation. Therefore, a protocol published by Chavez et al. (2015) [66] was employed. The workgroup generated the

tripartite activator “VP64-p65-Rta” (VPR), which was fused to a dCas9. Using this construct, targeted enhancement of gene expression is possible without changing the natural chromosomal context. To establish the VPR method at the Institute of Human Genetics Regensburg, the findings of Chavez et al. (2015) were first replicated by targeting the gene *MIAT* with a mixture of the same sgRNAs as published by Chavez et al. (2015). Remarkably, gene expression of *MIAT* was enhanced by a fold change of 113.4 (SD: 14.3) in comparison to a transfection of HEK293T cells, which did not include the MIAT sgRNAs (**Figure 16 A**).

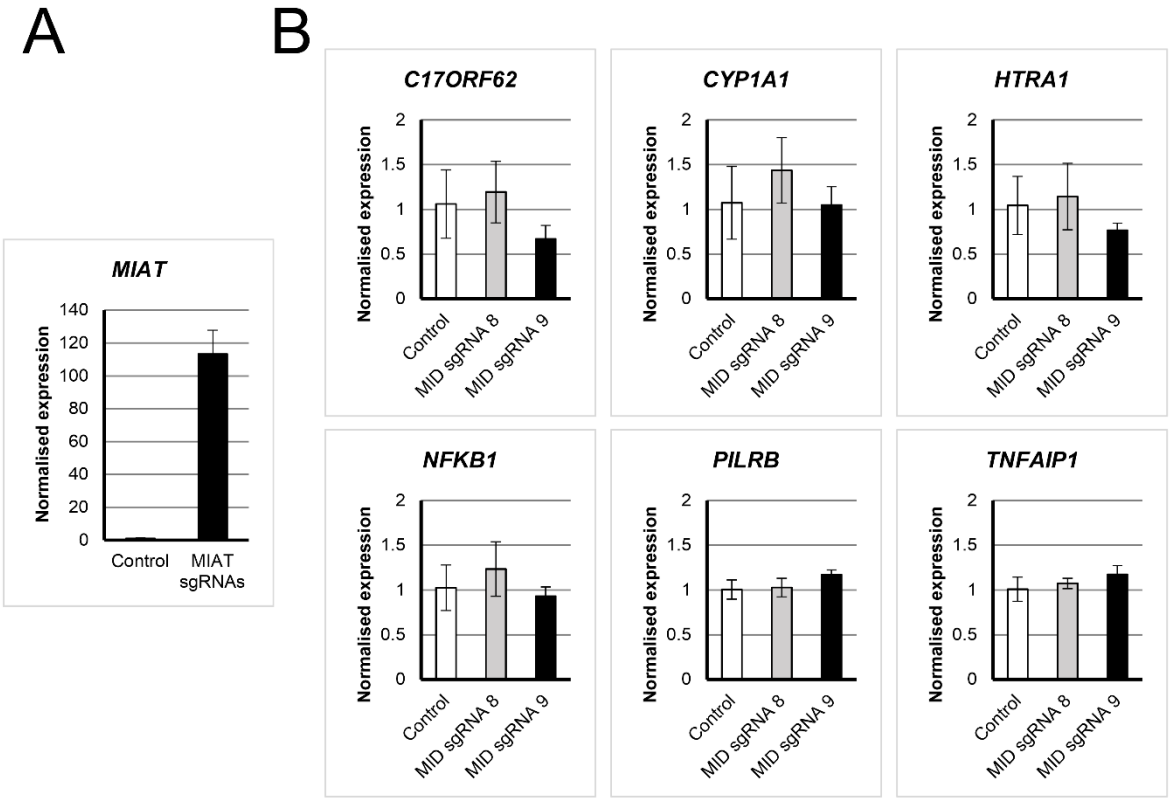

Figure 16: Enhancement of gene expression using dCas9-VPR in HEK293T cells.
(A) qRT-PCR results after double transfection of HEK293T cells (n = 3) with a mixture of four *MIAT* sgRNAs published by Chavez et al. (2015) [66] and the dCas9-VPR vector. (B) Targeted enhancement of gene expression within the *ARMS2-HTRA1* locus was performed with the help of the two sgRNAs MID 8 (n = 6) and MID9 (n = 4). qRT-PCR results of five exemplary bioinfomatically predicted target genes (**Table 31**) and *HTRA1* are shown. qRT-PCRs were normalised in regard to dCas9-VPR transfected HEK293T cells (control, n = 7) without supplying any sgRNA.

Eleven sgRNAs (MID sgRNA 1 to 11) were tested for efficiency following the protocol described above and the two sgRNAs MID 8 and 9 (**Figure 13**) were chosen for targeted enhancement of the *ARMS2-HTRA1* minimal haplotype region. Nevertheless, qRT-PCRs of the bioinformatically predicted target genes did not show any significant changes in gene expression of dCas9-VPR and MID sgRNA transfected cells in comparison with control cells (**Figure 16 B**). The usage of sgRNAs UP 2, UP 3, DOWN

1, and DOWN 2 in combination with dCas9-VPR failed also to reveal an altered expression of target genes.

4.4 RNA sequencing and eQTL analysis of retinal tissue

4.4.1 Study overview of the retinal eQTL database

The liver eQTL database and GTEx did not include eye tissue, which would be a valuable resource for the investigation of ocular diseases and traits. To date, only a single study calculated eQTL in retina, but included over 300 AMD patient eyes in their dataset of a total of 406 samples. Therefore, one aim of the current thesis was to analyse gene expression regulation in 161 healthy retinal samples collected at the Institute of Human Genetics Regensburg. Furthermore, two other collaboration partners, namely the University Hospital in Cologne and the National Eye Institute (NEI), shared their raw RNA-Seq and genotype data to enable an eQTL mega-analysis of healthy retinae. The data processing and QC was performed similar to the mega-analysis in liver tissue. After QC, 314 samples were available for further analysis (**Table 32**).

Table 32: Study, sample, and result summary of the Retina eQTL database

Dataset	Human Genetics Regensburg	University Hospital Cologne	NEI Bethesda [70]
Sample size before QC/ after QC	161 / 144	78 / 76	105 / 94
Mean Age	59.2 (SD: 16.8)	70.1 (SD: 12.6)	74.2 (SD: 9.4)
Gender (M / F)	97 / 47	37 / 39	46 / 48
RNA-Seq library	NEXTFLEX® Rapid Directional RNA-Seq Library Prep Kit	TruSeq® Stranded mRNA Library Preparation Kit	TruSeq® Stranded mRNA Library Preparation Kit
RNA-Seq platform	Illumina HiSeq platform		
RNA-Seq depth	20 m SE	50 - 80 m PE	10 - 20m PE
Read length	83 bp	51 bp	125 bp
Expressed genes (CPM > 1 in 10 % of samples)	18,290	18,971	18,401
Expressed genes overlapping	17,405		
Genotyping Platform	Custom HumanCoreExome BeadChip	Infinium® OmniExpress-24 v1.2 BeadChip	UM_HUNT_Biobank v1.0 chip
Imputed variants after QC	8,686,883		
eVariants (Q-value < 0.05)	869,464		
eVariants (Q-value < 0.05, unique)	600,077		
eVariants regulating several Genes (Q-value < 0.05)	149,078		
eGenes (Q-value <0.05, unique)	9,733		
Independent signals (P-value < 4.0 x 10^{-4})	15,262		
eVariants (Q-value < 0.001)	426,461		
eVariants (Q-value < 0.001, unique)	305,268		
eVariants regulating several Genes (Q-value < 0.001)	69,116		
eGenes (Q-value <0.001, unique)	2,757		
Independent signals (P-value < 3.9 x 10^{-6})	3,082		

PE = Paired-end; QC = quality control; SD = standard deviation; SE = Single-end

RNA-Seq reads were initially analysed separately per individual dataset. A total of 2,412 genes were found to be exclusively expressed (CPM > 1 in at least 10 % of the samples) in only one or two of the three datasets and were subsequently excluded. This left information on a total of 17,405 genes shared between the three datasets which were combined and normalised together. Regarding the genotype data, each dataset was separately imputed, which resulted in 8,686,883 overlapping and quality-controlled variants (**Table 32**).

The merged genotype- and gene expression data were then explored for local eQTL. Local eQTL were calculated by including all variants on the same chromosome that

are located within 1 Mbp up- or downstream of the TSS or polyadenylation site of the respective gene. After adjustment for multiple testing, 869,464 significant eVariants (Q-value < 0.05) were identified, which regulate 9,733 unique eGenes (**Table 32**). Moreover, a conditional analysis revealed 5,529 additional independent (secondary) signals by adjusting for the respective most significant primary eVariant (P-value < 4.0 x 10^{-4}). A more stringent adjustment for multiple testing (Q-value < 0.001) resulted in 2,757 unique eGenes and 325 secondary signals (P-value < 3.9 x 10^{-6}).

4.4.2 Characterisation of gene expression regulation in retina

The primary and secondary signal eVariants were first characterised with respect to their significance and position regarding the corresponding eGenes (Q-value < 0.05) (**Figure 17 A**). Signals were widely distributed around the TSSs of the respective eGenes. Interestingly, highly significant eVariants were observed to be located closer to the TSS in comparison to less significant eVariants. Nevertheless, some eVariants were located several thousand bp away from the respective TSS and showed highly significant P-values. This was especially the case for the eQTL rs577360216 - *MAPK8IP1P2* (P-value: 5.59 x 10^{-117}, TSS distance: +668,829 bp) and rs6075340 - *SIRPB1* (P-value: 5.17 x 10^{-96}, TSS distance: +293,628 bp).

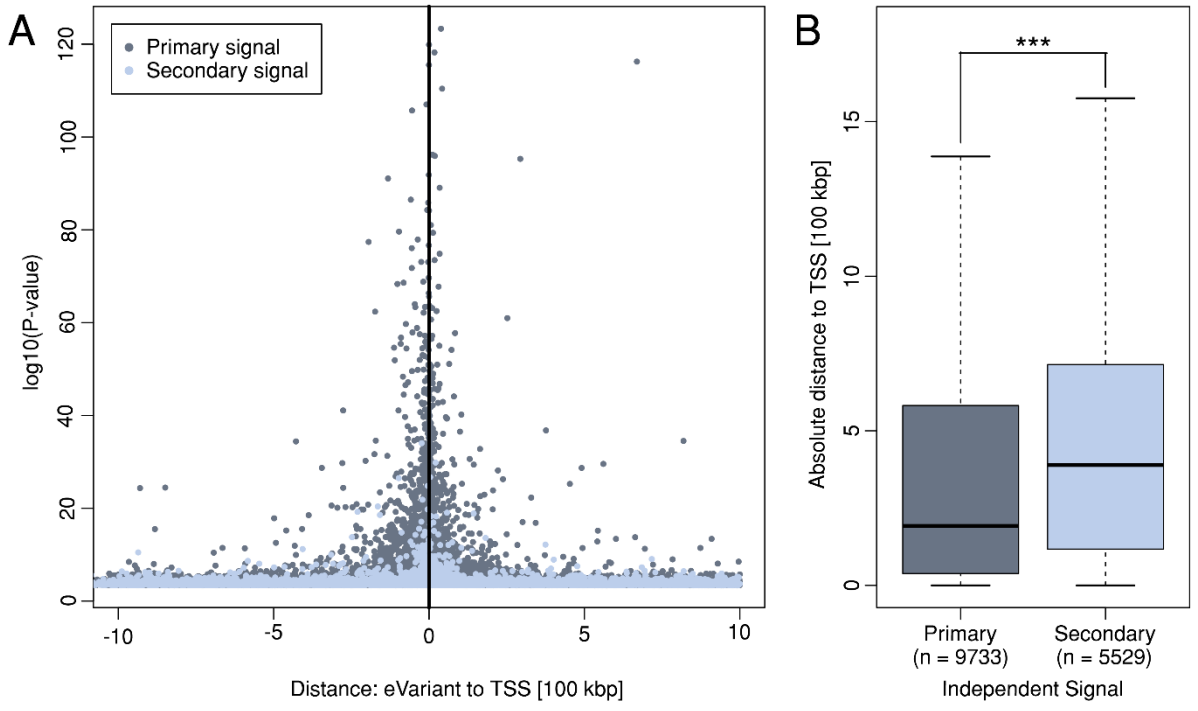

Figure 17: Genomic localisation of eVariants in the retinal eQTL database.
(A) The distance of each eVariant to the TSS of the respective eGene is plotted against the significance of the association (−log10 P-value). Shown are the primary (dark grey) and independent secondary, (light grey) eVariants for each eGene. Negative/positive distances denote that the variant is located

upstream/downstream of the TSS with regard to the direction of transcription. (B) Boxplot of the absolute distance of primary and secondary signals to the TSS. Significance was assessed by a Mann-Whitney-U-Test (P-value = 4.2×10^{-104}). (Figure modified from Strunz et al., 2020 [119]; Note that the shown figure differs from the publication because the data preparation protocol changed during manuscript revision. Details are given in the respective method sections.)

Interestingly, more than half (8,488/15,262) of the independent signals were located downstream of the respective TSSs. Furthermore, primary signals were found to be located significantly closer to the TSS in comparison with secondary signals (**Figure 17 B**, P-value = 4.2×10^{-104}).

149,078 (24.8 %) of the 600,077 unique eVariants (Q-value < 0.05) regulated the expression of more than one eGene. Therefore, the question arose if these highly regulatory active variants are distributed randomly over the genome or if they cluster in so called “regulatory clusters”. To answer this question, the list of eVariants was filtered for (1) a Q-value of 0.001 (305,268 eVariants, **Table 32**) and (2) eVariants regulating at least three genes, resulting in 25,299 variants for further analysis. Thereafter, variants, which were located close to each other (1 Mbp window) were assigned to the same cluster. This analysis revealed 76 regulatory clusters, which are distributed over the whole genome (mean number of clusters per chromosome: 3.45, SD: 2.39) (**Figure 18**). Remarkably, chromosome 7 harbours most clusters (9 of 76), whereas no clusters were found on chromosome 4 and chromosome 13. The cluster size varied widely from 1 bp (clusters 5:122982802-122982802, 10:79629844-79629844, 11:7885630-7885630, 11:49154505-49154505, 16:19584627-19584627), each containing a single eVariant regulating several eGenes to 6,433,565 bp for cluster 6:26678284-33111849 regulating 42 genes.

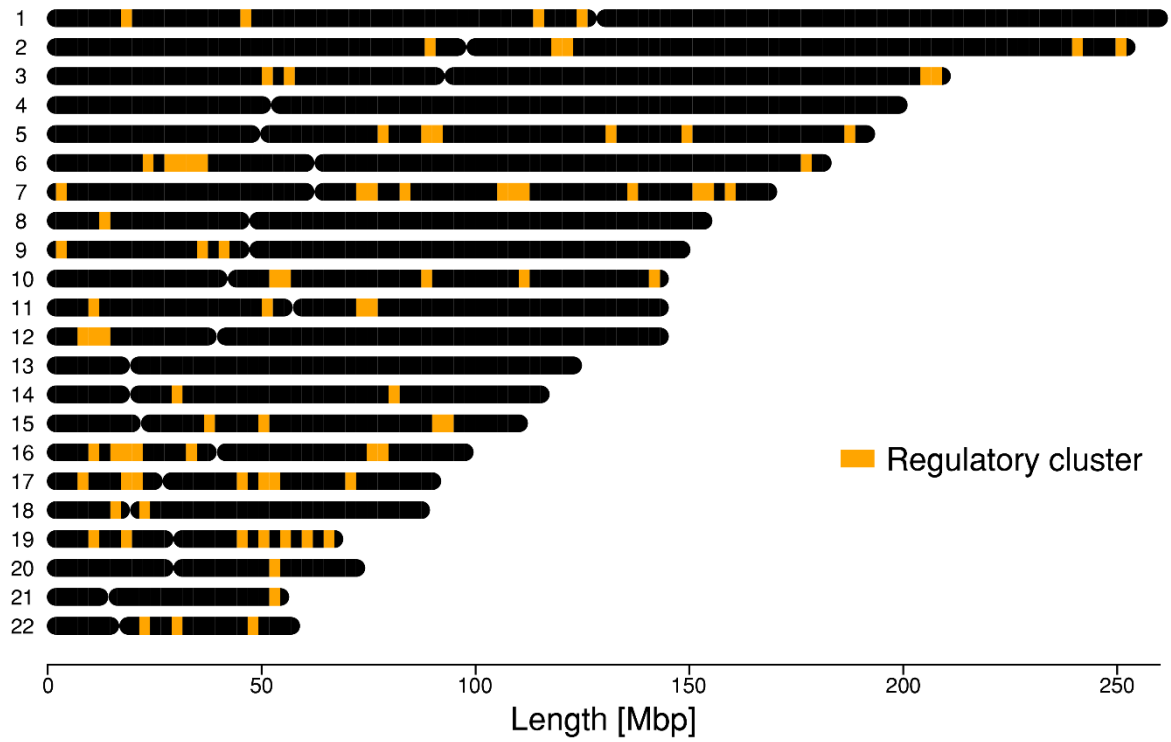

Figure 18: Chromosomal position of regulatory clusters in retinal tissue.
Highly significant eVariants regulating three or more eGenes (Q-Value < 0.001) were combined into 76 regulatory clusters (orange) and mapped onto the human genome (window size 1 Mbp). The plot was generated by using the *chromoMap* package in R [120].

4.4.3 Retinal eQTL and AMD-associated genetic variants

The 52 AMD-associated IHs identified in the AMD GWAS of Fritsche et al. (2016) [18] were investigated in the retinal eQTL database. 41 of these were genotyped or imputed into the dataset and 7 variants regulate the expression of at least one eGene (Q-value < 0.05) (**Table 33**). Altogether, 13 unique eGenes were regulated by AMD-associated variants.

Table 33: Genome-wide significant AMD-associated variants regulating genes in retinal tissue

IH*	dbSNP ID	CHR	Position [hg38]	Gene Symbol	eQTL Q-Value	Beta**	SE	Non-Risk allele	Risk allele
8.3	rs204993	6	32,187,804	*HLA-DQB1*	1.54E-05	-0.484	0.086	A	G
8.3	rs204993	6	32,187,804	*TSBP1-AS1*	1.85E-04	0.190	0.037	A	G
11	rs7803454	7	100,393,925	*PILRA*	4.50E-51	0.850	0.044	C	T
11	rs7803454	7	100,393,925	*PILRB*	7.29E-27	0.785	0.061	C	T
11	rs7803454	7	100,393,925	*STAG3L5P*	1.83E-23	0.557	0.047	C	T
11	rs7803454	7	100,393,925	*ZCWPW1*	3.93E-03	0.155	0.036	C	T
18	rs3750846	10	122,456,049	*BX842242.1*	5.22E-10	0.204	0.027	T	C
19	rs3138141	12	55,721,994	*AC009779.3*	1.91E-03	-0.170	0.037	C	A
24.1	rs5817082	16	56,963,437	*MT3*	1.52E-02	-0.273	0.069	CA	C
24.1	rs5817082	16	56,963,437	*RSPRY1*	2.63E-02	0.082	0.022	CA	C
24.1	rs5817082	16	56,963,437	*GNAO1*	3.00E-02	-0.129	0.034	CA	C
26	rs11080055	17	28,322,698	*TMEM199*	1.28E-02	0.069	0.017	A	C

27	rs6565597	17	81,559,795	*ARL16*	3.96E-02	0.101	0.028	C	T

CHR: chromosome; SE: standard error of the effect size; * IH: Independent hit according to Fritsche et al. 2016 [18] ** Effect size of a single AMD risk increasing allele

4.4.4 Investigation of GWAS variants with regard to different ocular traits

The retina eQTL database facilitates not only the analysis of gene expression regulation in the context of AMD, but may be applied to address various other related questions. Christina Kiel, a researcher at the Institute of Human Genetics, generated a curated list of variants associated with at least one of 82 different traits and diseases (at genome-wide significance, P-value < 5.0×10^{-8}) [121]. The data collection also included variants regarding 12 distinct ocular traits and diseases derived from 16 published GWAS (**Table 34**).

Table 34: Complex eye diseases and traits investigated in the context of retina eQTL

(data kindly provided by Christina Kiel, Institute of Human Genetics, Regensburg [121])

Complex eye disease or trait	PubMed ID	GWAS Variants after QC	Variants included in study	eVariants (Q-Value < 0.05)	eGenes (Q-Value < 0.05)
Age-related macular degeneration	26691988	52	41	7	13
Central corneal thickness	30622277	39	38	3	3
Diabetic retinopathy	26188370, 30178632	3	3	0	0
Intraocular pressure	28073927, 29235454, 29617998, 29785010, 30054594	251	243	32	47
Macular thickness	30535121	135	129	29	45
Myopia	23468642	22	22	3	3
Optic disc - cup area	28073927	24	23	2	2
Optic disc - disc area	28073927	16	16	4	4
Primary angle closure glaucoma	27064256	8	7	1	2
Primary open-angled glaucoma	26752265, 29891935	50	49	4	5
Refractive error	29808027, 23396134	119	98	14	21
Vertical cup-disc ratio	28073927	22	21	1	1

QC = quality control

The number of GWAS variants varied widely from 3 (see “diabetic retinopathy”) to 251 (see “intraocular pressure”). Overall, 690 variants were included in the retinal eQTL database and 100 of these showed an association with at least one eGene (Q-value < 0.05). 125 unique eGenes were identified, since some disease- or trait-associated eVariants regulate multiple genes. Remarkably, 17 of these eGenes are regulated by eVariants associated with multiple different phenotypes (**Figure 19**). For example,

lower expression of the non-annotated protein coding gene *AC009779.3* is potentially associated with increased risk for AMD, refractive error, and increased macular thickness while decreased gene expression of *AC009779.3* is associated with an increased risk of myopia. Furthermore, AMD-associated variants were also found to upregulate the expression of *PILRA*, which expression change is also potentially linked to macular thickness, and to downregulate *HLA-DQB1*, which is downregulated by intraocular pressure-associated variants.

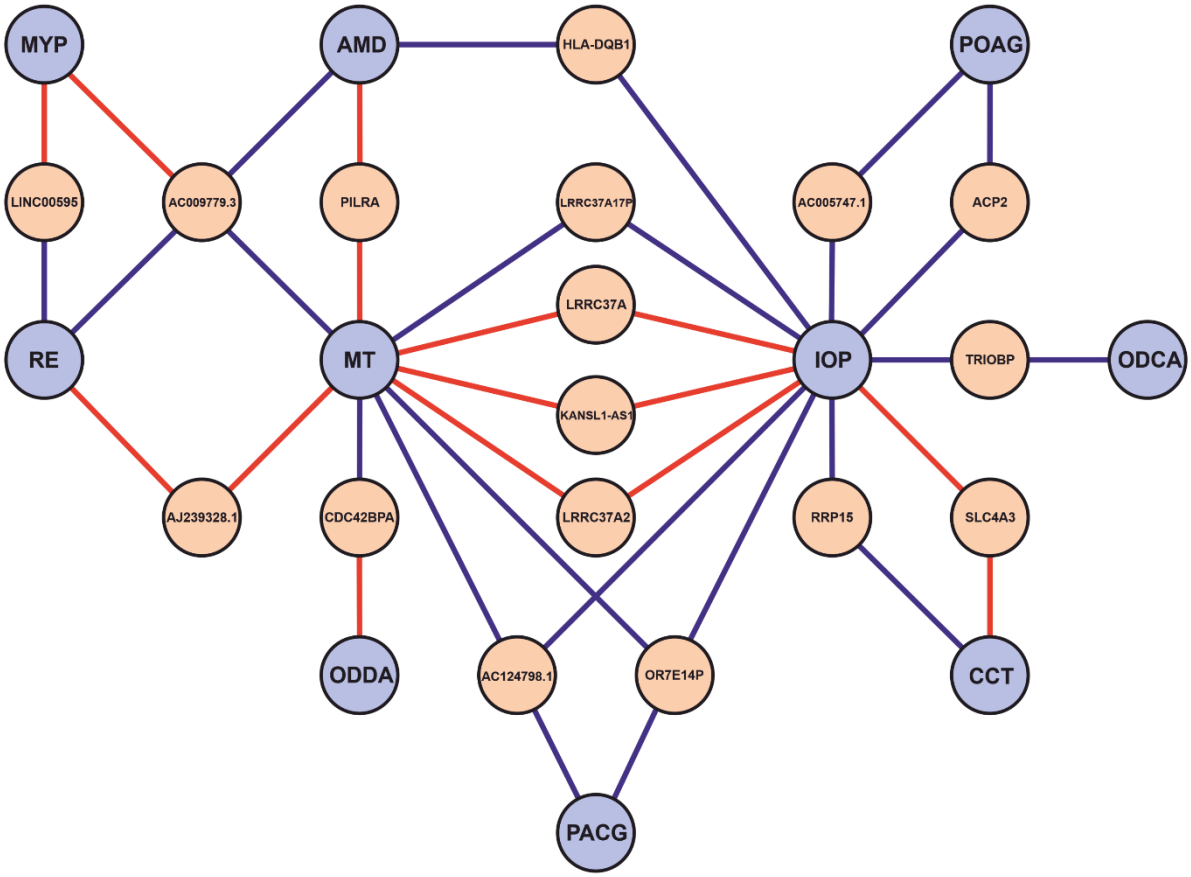

Figure 19: Retinal eGenes regulated by multiple complex eye disease- or trait-associated variants.
17 eGenes (orange) were regulated by genome-wide significant GWAS variants of at least two different complex eye diseases or traits (blue). Connective lines are colored according to the eQTL effect direction of the risk-/trait- increasing allele. Red lines reflect higher gene expression whereas blue lines represent downregulation of expression. AMD = age-related macular degeneration; CCT = central corneal thickness; IOP = intraocular pressure; MT = macular thickness; MYP = myopia; ODCA = optic disc - cup area; ODDA = optic disc - disc area; PACG = primary angle closure glaucoma; POAG = primary open-angled glaucoma; RE = refractive error. (Figure modified from Strunz et al., 2020 [119]; Note that the shown figure differs from the publication because the data preparation protocol changed during manuscript revision. Details are given in the respective method sections)

4.5 TWAS based on AMD genetics and the GTEx project

eQTL analyses are based on linear regression models and usually consider one genetic variant and one gene at a time. Gamazon et al. (2015) proposed a more

complex model, which uses classical machine learning approaches and called it PrediXcan [53]. This algorithm is applied to determine a set of genetic variants which consistently influence gene expression in a given tissue. In a second step, these variants can be extracted from a GWAS dataset to predict the relative gene expression of study participants. Finally, the imputed gene expression is correlated to the individuals' disease status to identify disease-associated genes. The three step procedure is called TWAS and can be applied to identify genetically regulated genes, which are potentially relevant for disease aetiology.

4.5.1 Identification of 106 genes associated with AMD

The PrediXcan algorithm [53] was applied to the full IAMDGC dataset [18], which includes genotype and phenotype data from 16,144 late-stage AMD cases (including clinical diagnoses of GA and/or CNV), and from 17,832 AMD-free controls. The prediction models from 27 tissues were retrieved from PredictDB (http://predictdb.org/, accessed September 3rd 2018) and were implemented into the analysis. These tissues have been chosen because genotype and gene expression data of more than 130 individuals were available for prediction model building. After separate gene expression imputation for each tissue, a linear regression model was applied to identify late-stage AMD-associated genes based on the individual's AMD status. P-values were adjusted for multiple testing using the FDR approach and genes with a Q-value smaller than 0.001 were considered to be significantly associated with AMD. In each tissue, a minimum of 11 (see "Brain Cerebellum" and "Heart Left Ventricle") and up to 28 (see "Adipose Subcutaneous" and "Nerve Tibial") AMD-associated genes (**Figure 20**) were identified (mean 17.63; SD 5.02). Altogether, 106 unique genes were significantly AMD-associated in at least one tissue (**Supplementary Table 2**).

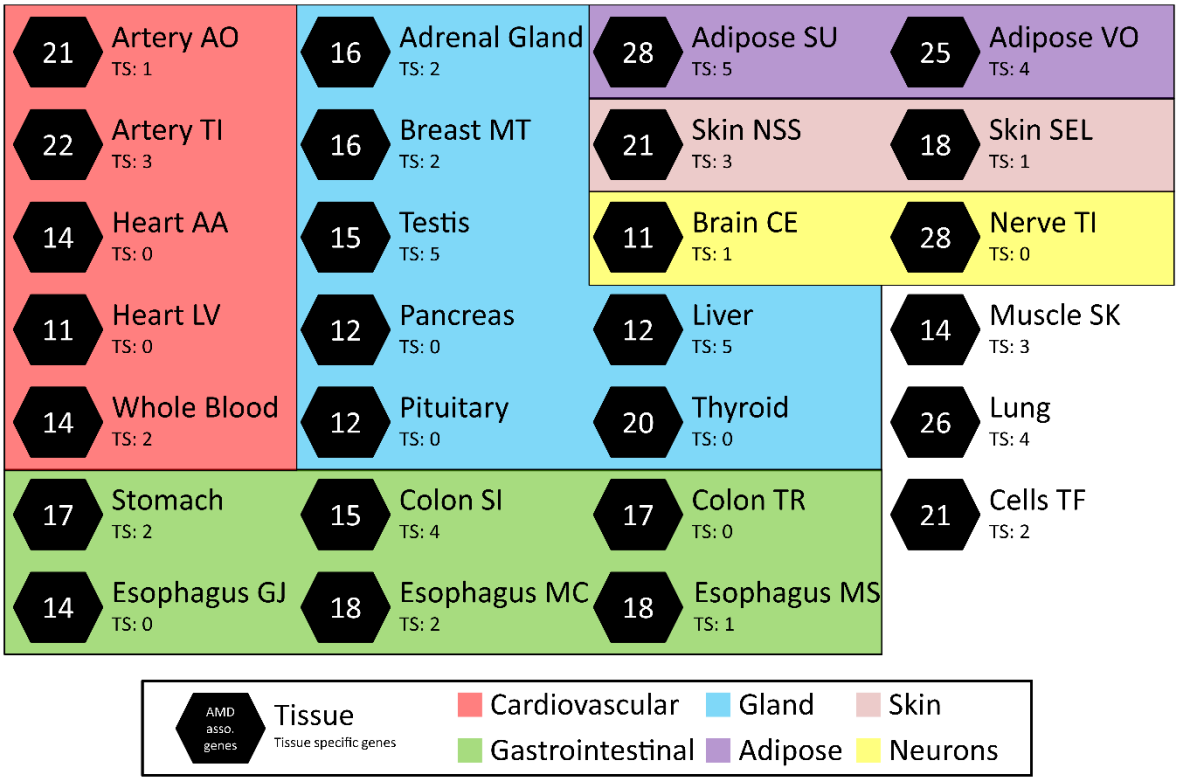

Figure 20: TWAS results for 27 tissues.
A TWAS was conducted based on the genotypes of 16,144 late-stage AMD cases and 17,832 AMD-free controls. Prediction models of 27 tissues were included in the analysis. The schematic overview demonstrates the number of significant AMD-associated genes (Q-value < 0.001) within the respective tissue. If a gene was found exclusively in a single tissue, it was marked as tissue-specific (TS). Tissue classification was performed manually according to main functions or metabolic assignments. Adipose SU: Adipose Subcutaneous; Adipose VO: Adipose Visceral Omentum; Artery AO: Artery Aorta; Artery TI: Artery Tibial; Brain CE: Brain Cerebellum; Breast MT: Breast Mammary Tissue; Cells TF: Cells Transformed fibroblasts; Colon SI: Colon Sigmoid; Colon TR: Colon Transverse; Esophagus GJ: Esophagus Gastroesophageal Junction; Esophagus MC: Esophagus Mucosa; Esophagus MS: Esophagus Muscularis; Heart AA: Heart Atrial Appendage; Heart LV: Heart Left Ventricle; Muscle SK: Muscle Skeletal; Nerve TI: Nerve Tibial; Skin NSS: Skin Not Sun Exposed Suprapubic; Skin SEL: Skin Sun Exposed Lower leg. (Figure published in Strunz et al., 2020 [122])

Of 106 AMD-associated genes, 88 are located in loci known to be AMD-associated with genome-wide significance. 18 additional genes were not located in proximity (window size of 1MB) to any of the 52 independent hits identified by Fritsche et al. (2016), and may denote novel AMD loci [18] (**Figure 21**). The linear regression models also provide an effect size based on the regression slope (beta). Positive effect sizes point to predicted gene expression in healthy tissue being higher in AMD cases than controls. Negative betas are suggestive for decreased gene expression with higher AMD risk. The largest effect sizes ranged from -0.38 (*ARMS2*, see "Testis") to +0.35 (*CFHR1*, see "Liver") (**Supplementary Table 2**). The mean absolute beta across all AMD-associated genes was 0.035 (SD: 0.039).

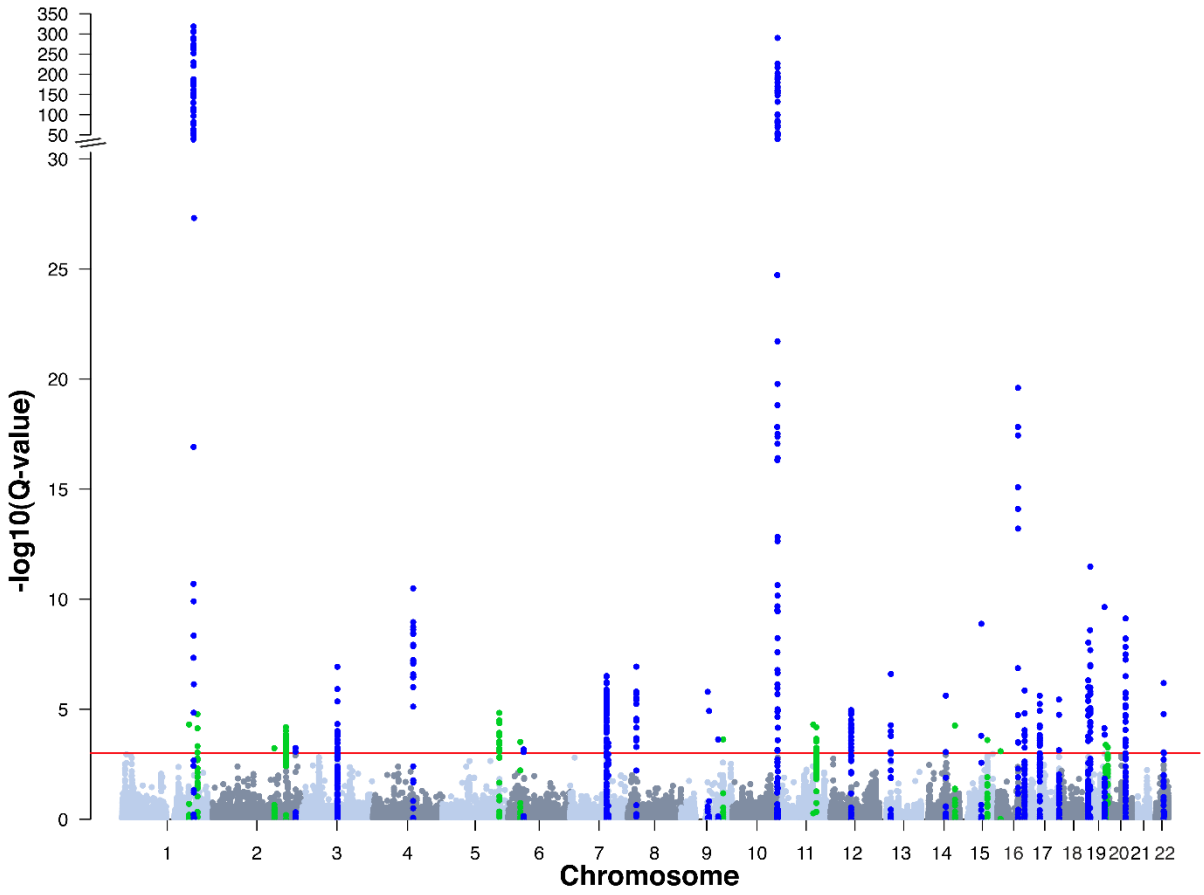

Figure 21: Manhattan plot of the AMD-associated genes in all 27 investigated tissues.
Linear regression models were performed to correlate the predicted gene expression of 27 tissues with AMD and control status. The Manhattan plot shows the −log10 Q-values and the chromosomal position for all predictable genes. Genes, which were significantly AMD-associated (Q-Value < 0.001; red line) in at least one tissue were highlighted in blue, if the gene was located in a known AMD locus, or green if the locus was not genome-wide significant in the GWAS of Fritsche et al. (2016) [18]. (Figure published in Strunz et al., 2020 [122])

Interestingly 54 out of the 106 genes were significantly AMD-associated in more than one of the 27 tissues (**Figure 20** and **Supplementary Table 2**). Remarkably, sixteen genes (*ADAM19*, *ARMS2*, *BTBD16*, *CFH*, *CFHR1*, *CFHR3*, *GPR108*, *PILRA*, *PILRB*, *PLA2G12A*, *PLEKHA1*, *PMS2P1*, *PPIL3*, *RDH5*, *STAG3L5P*, *and TNFRSF10A*) were AMD-associated in over 10 tissues. Furthermore, some genes showed an AMD association of predicted gene expression in almost all analysed tissues. This is especially the case for three genes (*PILRA*, *PILRB*, and *STAG3L5P*) located within the known AMD Locus 11 [18].

4.5.2 Comparison to AMD TWAS of retinal tissue

The study of Ratnapriya et al. (2019) included a TWAS analysis based on retinal eQTL data and the summary statistics of the AMD GWAS from Fritsche et al. (2016) [18,70]. The TWAS comprised data of 406 retinae, which were mainly derived from AMD

patients. Altogether, the TWAS identified 31 significantly AMD-associated genes (Q-value < 0.001, genetic model R2 ≥ 0.01) of which 22 were located outside the MHC locus. These genes were compared to the PrediXcan analysis regarding the 27 GTEx tissues to identify potential retinal-specific effects. 16 of the 22 genes were also found to be AMD-associated in at least one of the 27 GTEx tissues and are therefore unlikely to represent retinal-specific effects. Remarkably, only two genes showed different effect directions in the retinal tissue TWAS compared with other GTEx tissues. One of these genes was *HTRA1*, of which the retinal expression was predicted significantly lower in AMD cases than controls. This was also true for the two tissues “Esophagus Mucosa” and “Esophagus Gastroesophageal Junction”. In contrast, predicted *HTRA1* expression was significantly higher in AMD cases than controls in five GTEx tissues (see “Thyroid”, “Skin Sun Exposed Lower leg”, “Heart Atrial Appendage”, “Pituitary”, and “Testis”). On the other hand, the predicted retinal expression of *PLA2G12A*, located on chromosome 4, was lower in AMD cases compared to controls. The opposite effect direction was observed in all 13 GTEx tissues in which predicted *PLA2G12A* expression was significantly associated with AMD status.

Two of the remaining six genes, exclusively found by Ratnapriya et al. (2019), were not measured in the GTEx dataset: the long non-coding RNA *STAG3L5P-PVRIG2P-PILRB* and the uncharacterised gene *RP11-644F5.10* (ENSG00000258311). Therefore, no conclusions can be drawn. The remaining four genes are expressed in several GTEx tissues, but were not AMD-associated in any of the 27 tissues investigated. Two out of these four genes are the uncharacterised transcripts *PARP12* and *CTA-228A9.3*. Finally, the remaining two genes are the protein coding genes *MEPCE* and RLBP1. The latter encodes the retinaldehyde-binding protein 1, which uses 11-cis-retinaldehyde or 11-cis-retinal as physiologic ligands.

5 Discussion

Publically available GWAS data reveal a plethora of loci and variants which are genome-wide associated with complex diseases and traits. For a number of reasons, functional interpretation of disease-associated genetic variants remains challenging and requires large scale approaches to avoid missing the potential small effects. Most of the GWAS genetic variants are located in non-coding regions of the genome and are common in healthy individuals [33]. Additionally, the extensive LD often hinders the identification of the signal causing variant or the respective gene. Therefore, investigation of gene expression regulation enables to combine statistical methods with the analysis of molecular data. This lays the foundation to generate new hypotheses regarding causal genes in GWAS loci and potentially disease relevant pathways.

Three databases regarding gene expression regulation were generated in this doctoral thesis. First, four different studies investigating gene expression in liver tissue were processed and combined to enable an eQTL mega-analysis. According to the established data processing protocol, gene expression and genotype data of the GTEx project were prepared to build an in-house database, which includes data of 48 different tissues and cells. This database was helpful to support ongoing projects at the Institute of Human Genetics and to generate new hypotheses. In a further project, an eQTL database including 314 retinal tissue samples from three independent study sites was generated and analysed in regard to multiple complex phenotypes. The large datasets were established to enable new insight into the aetiology of AMD, a complex eye disease with a strong genetic background. Besides the identification of gene regulatory functions within AMD-associated loci, a new hypothesis regarding the *ARMS2-HTRA1* locus was generated and evaluated experimentally using genome editing via the CRISPR/Cas9 technology. In a final project of this thesis, machine learning was applied to allow an unbiased investigation into AMD genetics. This analysis resulted in a list of 106 AMD-associated genes potentially involved in various molecular pathways throughout the whole body.

The analysis of gene expression in single tissues revealed that many genes are genetically regulated and that the number of eGenes varies between tissues and databases. For example, 31.6 % of all expressed genes in the liver eQTL database were eGenes (7,612 of 24,123), whereas in the retinal eQTL database this was the case for 55.9 % (9,733 of 17,405).

A more detailed investigation of the liver eQTL database revealed that single studies showed remarkably less eGenes in comparison to the combined analysis. This may be attributable to smaller sample sizes as a correlation of sample size and the number of eQTL has been observed in the GTEx database (R2 = 0.83), but could also be due to the different data processing protocols. The four liver eQTL studies applied either microarrays or RNA-Seq to detect gene expression. The main difference of both techniques consists in the measurement type and the following quantification. Microarrays compare fluorescence signals of single probes with a given reference on the same chip, whereas RNA-Seq quantifies short reads and assembles them to transcripts, which requires a normalisation for each sample on the same flow cell. Independently from the measurement technique, gene expression data need always to be normalised to enable the comparison of different samples, even within the same dataset. This process complicates the evaluation of eQTL and their respective effect size, because an effect size of one dataset is often not comparable to effect sizes in other studies. For example, the eVariant rs7803454 regulates gene expression of *PILRB* in the liver database (effect size: 0.251) and the retinal eQTL database (effect size: 0.785), while it is impossible to make implications whether the effect is stronger in one of the tissues in comparison to the other. Several strategies could be applied to normalise effect sizes: (1) compare effect sizes to known physiological effects, or (2) scale gene expression values to a defined mean and SD. The first approach could be applied based on the eQTL rs10922109 – *CFHR1* (effect size: 0.992, liver eQTL database) as several studies showed that rs10922109 shares a haplotype with the deletion of the genes *CFHR1* and *CFHR3* [123]. However, *CFHR1* and *CFHR3* are not ubiquitously expressed and defining an appropriate physiological effect as reference is challenging. The second approach was applied to compare the different tissues of the GTEx project, since exactly the same data measurement and processing protocol was used for all samples. However, the normalisation processes before eQTL calculation may always influence the comparability of effect sizes between datasets. Nevertheless, the effect direction seems to be a valuable criterion to evaluate eQTL with respect to their potential physiological impact because its algebraic sign is independent of gene expression processing.

Furthermore, the measurement of gene expression in 314 retinal tissue samples originating from three independent study sites revealed that 2,412 genes were exclusively detected in only one or two of the studies. It is important to remark, that

even if comparable measurement methods and the exact same raw data analysis pipeline were applied, hidden batch effects may influence results in single datasets [124,125]. Therefore, data of one study site should always be assessed in comparison with other datasets, to avoid at best the detection of false positive results. Alternatively, false positive findings can be minimised by correcting for multiple testing. The investigation of local eQTL in retina for example required adjustment for over 108.8 million tests. So far, there is no gold standard for this procedure although several different adjustment approaches including Bayesian methods, permutation testing, and FDR calculation, are well accepted [126]. Adjustment for multiple testing gets even more complicated due to small eQTL effect sizes and the high variability of gene expression values between samples. All presented results in this thesis were based on stringent FDR thresholds to minimise detection of false positives, although some effects might remain unnoticed.

As a first take home message, the comparison of effect sizes should always be performed with caution and should rather focus on effect directions, since these are independent of measurement and normalisation methods. Furthermore, combining single eQTL studies with further datasets omits findings caused by hidden confounders as well as batch effects and even enhances the potential to detect more effects because of the higher sample size.

Evaluating the functional impact of eQTL is a highly discussed area facing several potential limitations: (1) mRNA abundance is only partly correlated with protein levels [40], (2) eQTL are frequently measured in post mortem tissue, which might not reflect the *in vivo* situation [127], (3) LD structures complicate the identification of true causal variants [128,129], and (4) the mechanisms underlying the eQTL signals often remain elusive [39]. Addressing these questions requires further methods and model systems. One of the most recent developments in the genome-editing field was the introduction of the CRISPR/Cas9 system, which enables targeted alteration of DNA sequences. In this study two strategies were applied to investigate experimentally gene expression regulation events, identified in the 10q26 (*ARMS2-HTRA1*) locus. First, two sgRNAs combined with a Cas9 endonuclease expression cassette were transfected into HEK293T cells to introduce the genomic deletion of the 5,196 bp “minimal haplotype” region defined by Grassmann et al. (2017) [25]. Thereafter, the deletion was successfully detectable via PCR reactions based on genomic DNA. Nevertheless,

gene expression of the previously bioinformatically predicted target genes showed no difference in modified single cell clones. The second approach aimed to enhance the computationally predicted effects using the dCas9-VPR construct generated by Chavez et al. (2015) [66]. The required protocol was first established in HEK293T cells by replicating the findings of Chavez et al. After generating a 113-fold enhancement of *MIAT* expression, dCas9-VPR was also applied in the minimal haplotype region at 10q26. Again, no alterations in gene expression of the predicted target genes were observed.

The failed replication of the bioinformatical hypothesis may be attributable to various reasons. The immortalised HEK293T cell line was chosen because of its comparably simple handling and the known high transfection efficiency. However, it is derived from embryonic kidney cells and might not reflect the physiological background of the GTEx post mortem samples. It was further seen as a promising model system because the observed eQTL were traceable in many tissues and most of the rs3750846-associated eGenes were known to be expressed in HEK293T cells. Another complication may be caused by the complexity of the minimal haplotype since it contains a 3,105 bp genomic repeat region harbouring multiple short interspersed nuclear elements (SINEs). This area is not specifically targetable by sgRNAs because genome editing might also affect additional loci. Furthermore, other studies previously reported gene expression regulation events caused by SINEs [130–132]. In addition, the minimal haplotype region is poorly covered by databases concerning chromatin conformation and accessibility [133], which could reveal potential mechanisms causing the distant eQTL effects. In general, prediction of the introduced molecular alterations caused by the deletion of the minimal haplotype region is challenging because the effect sizes of the beforehand calculated eQTL cannot be included in the evaluation. eQTL provide information about changes in gene expression based on allelic differences of specific variants. Deleting the whole genomic region around the variant generates a situation which is therefore not covered by eQTL. The affected gene regulation network might be seriously altered, whereby compensatory effects could also occur, especially if important pathways like the complement system are involved [134].

The very first successful *in vitro* CRISPR/Cas9 application investigating local eQTL was published in 2019 by Schrode and colleagues [68]. They altered the eVariant rs4702 in *NGN2* excitatory neurons derived from human induced pluripotent stem cells

and replicated a beforehand identified eQTL in brain tissue [135]. The allelic conversion of rs4702 from AA to GG enabled to further explore the eQTL driving mechanisms in this locus and to assess its functional consequences. Nevertheless, allelic conversion was so far only applied to one specific local eQTL and its success rate might depend on the investigated genomic region and the respective haplotype structure. Another promising approach to explore eQTL *in vitro* and to resolve LD structures is based on cloning short genomic sequences around eVariants in front of a minimal promoter followed by a barcoded open reading frame. The generated constructs are then introduced into cultured cells, which are incubated for several hours. Next, DNA and RNA are isolated and compared to each other. The ratio of both provides information regarding the transcriptional influence of the eVariant. This approach can be further applied considering different alleles and various variants in one locus to resolve LD structures and to accurately identify regulatory DNA motifs. Ulirsch et al. first described this protocol to shed light on GWAS variants of red blood cell traits and called it massively parallel reporter assay [129].

Altogether, developing methods for the functional validation of eQTL is highly relevant because eQTL do often not allow direct implications on the underlying biological mechanisms. Genome editing techniques enable targeted modification of genomic DNA and facilitate the generation of new model systems. Nevertheless, validating distant eQTL remains a complex task, which was not achieved so far. The generated hypothesis regarding the *ARMS2-HTRA1* locus requires further investigations. This might be achieved with the help of other eQTL databases and by refinement of the applied *in vitro* models.

Besides the identification of rs3750846 in the *ARMS2-HTRA2* locus, Fritsche et al. (2016) detected 51 additional AMD-associated IHs distributed over 33 loci. Many of the 18 secondary but independent signal variants in a respective locus showed very low MAFs (< 1 %) and are usually not covered in other studies due to MAF thresholds or unreliable imputation. At first, investigation of potential disease relevant gene expression regulatory events was performed by searching eQTL databases for disease-associated variants. In case of AMD, 31 respectively 41 IHs were covered in the generated liver and retina databases. Eight IHs, distributed over 5 loci, were eVariants in liver and regulated the expression of altogether 15 unique eGenes. In contrast, seven IHs, each positioned in another locus, regulated 13 unique eGenes in

retinal tissue. Compared to retina, 6 AMD-associated variants were exclusively eVariants in liver tissue: rs10922109 (IH 1.1), rs570618 (IH 1.2), rs61818925 (IH 1.6), rs2043085 (IH 23.1), rs2070895 (IH 23.2), and rs17231506 (IH 24.2). These eVariants regulate the expression of 10 eGenes, with 5 eGenes known to be involved in complement activation (*CFH*, *CFHR1*, *CFHR4*, *CFHR3*, and *CFHR5*) and two genes being relevant for HDL metabolism (*LIPC* and *CETP*). Notably, the liver constitutes the main tissue for synthesis of systemic complement factors and blood lipids [136–138]. In contrast, a general interpretation of the five IHs being an eVariant in retinal but not in liver tissue remains complex, since no clearly shared pathways are detectable between the genes *HLA-DQB1*, *TSBP1-AS1*, *BX842242.1*, *AC009779.3*, *MT3*, *RSPRY1*, *GNAO1*, and *TMEM199*. Interestingly, two IHs are eVariants in both databases: rs6565597 (IH 27) regulates three genes in liver (*TSPAN10*, *ACTG1*, and *ANAPC11*) and one in retinal tissue (*ARL16*). The second shared eVariant rs7803454 (IH 11) regulates the genes *PILRA* and *PILRB* with the same effect direction in both organs and two further genes exclusively in retinal tissue: *STAG3L5P* and *ZCWPW1*. PILRA and PILRB proteins are known to function as antagonists within the PTPN6 pathway and have been previously investigated in the context of AD [139,140]. Remarkably, Kikuchi et al. (2019) identified chromatin looping as a key event for gene expression regulation in this locus [141].

In general, it is recommended to investigate gene expression regulation in tissues, which are mechanistically relevant for the disease of interest [54]. AMD is a disease of the posterior pole and it is widely anticipated that the choroid, the RPE, and the retina are mainly involved in pathogenic processes concerning late-stage AMD [142]. Regarding these tissues, to-date solitary expression data of the retina are available in large scale and only 7 of the 52 (13.5 %) AMD-associated IHs were eVariants in the results presented in this thesis. In contrast, a recent study regarding schizophrenia, obviously a brain-related disease, revealed that 51 of 106 (48.1%) schizophrenia-associated GWAS lead variants are eVariants in brain tissue [143]. In consequence, the rarely observed gene expression regulation by AMD-associated variants in retinal tissue raises the hypothesis that the retina is not the primary site of AMD pathology. However, no conclusion can be drawn for the choroid or the RPE since no eQTL data regarding these tissues are available to-date.

Furthermore, gene expression regulation effects occurring in single tissues are difficult to interpret since most proteins are only characterised regarding their general function. Information about potential tissue-specific interaction partners or molecular roles remains elusive. Additionally, proteins often show different tissue- and cell type-specific isoforms, which are again rarely characterised.

In case of AMD, retina-specific regulation of gene expression was only rarely observed in this study. In contrast, many changes in expression were detected in pathways relevant for the many bodily cells or tissues, like the complement and the blood lipid system. For these reasons, an alternative approach was used to elucidate the potential role of AMD-associated variants in AMD aetiology. Instead of investigating tissue-specific eGenes, a TWAS was performed to identify significantly AMD-associated genes in multiple tissues. The usefulness of TWAS was already shown for various complex phenotypes, like pancreatic cancer [144], lung cancer [145], or autism spectrum disorder [146]. In the present study, a TWAS was performed based on the individual genetic background of 16,144 late-stage AMD cases and 17,832 non-AMD controls, a dataset from the IAMDGC. This method represents an unbiased approach since gene expression imputation was not informed about the AMD status. In addition, the analysis was not restricted to AMD-associated IHs, but instead considered all possible local gene expression regulation events. This, in the end, enabled to identify genes associated with AMD genetics, which were not located in significant GWAS loci of previous studies. The TWAS including 27 tissues identified 106 genes, being AMD-associated in at least one tissue. Remarkably, 10 of 15 (66.7 %) eGenes in the liver eQTL database regulated by AMD-associated variants were also identified by the TWAS analysis. Three of these genes (*F13B*, *ALDH1A2*, and *LIPC*) were exclusively AMD-associated in liver tissue. This underscores the validity of the TWAS approach to also cover single eQTL findings. However, it should be mentioned that a small proportion (83 of 588, 14.1 %) of the liver database samples were included in both studies, the liver eQTL mega-analysis and the TWAS.

Nevertheless, the TWAS approach also has limitations, which become particularly apparent in the *ARMS2-HTRA1* locus, since *ARMS2* expression was found to be associated with AMD. As described earlier, several studies point to *ARMS2* expression being potentially not causative for the AMD GWAS signal at this locus, since rs2736911, which results in a truncated *ARMS2* protein, was never found to be

associated with AMD [22,147]. These findings are not recognised by the TWAS because of the extensive LD structure and the highly significant AMD-associations of variants in this locus. The results regarding gene expression regulation should therefore always be evaluated in the context of other studies and experiments. Furthermore, the TWAS did not include RPE or choroid tissue, which might be highly relevant for AMD pathology.

Altogether, 54 genes were AMD-associated in multiple tissues, which points to non-tissue-specific processes and pathways. However, a pathway enrichment analysis of the 54 genes failed to identify prominent processes. Quite the contrary, a large number of AMD-associated genes seem not to exclusively take part in the highly discussed AMD relevant pathways: (1) the complement system, (2) blood lipid levels, or (3) the extracellular matrix, as proposed by other studies [12,18].

It is important to note that the TWAS and all eQTL studies in this thesis were based on healthy tissue and do not allow implications on disease mechanisms after AMD onset. Especially since cell type compositions may change, as occurring in AMD-associated retinal degeneration, which could result in different expression profiles throughout AMD stages. This was already observed for RPE and choroid tissue via single-cell RNA-Seq [148]. Interestingly, Ratnapriya et al. (2019) found no significant difference in gene expression of AMD affected and healthy donor eyes and therefore analysed eQTL in a merged dataset [70]. However, the undetectable differences in gene expression may be contributable to the normalisation methods, which were based on an extensive list of 3,804 “housekeeping” genes [149]. Nevertheless, the 54 AMD-associated genes provide help to generate new hypotheses regarding AMD aetiology and highlight, that individuals with high genetic burden for AMD are expected to show gene expression changes across multiple tissues outside the retina.

In line with the identification of genes associated with AMD genetics in multiple tissues are the discoveries of several studies, which found correlations between the genetic risk of AMD and other complex phenotypes [121,150,151]. This indicates, that genetic variants which contribute to AMD risk potentially have pleiotropic effects. Therefore, a follow up study based on the TWAS results analysed the 106 AMD-associated genes according to a physical overlap of their genomic position with GWAS loci of 82 complex phenotypes [122]. This comparison highlights 50 of 106 (47.2 %) genes that have relevance for AMD aetiology and that potentially affect at least one other phenotype.

Of course, co-localization with a GWAS signal is not a functional evidence as such, but these genes are *a priori* candidate genes to be relevant for disease formation of other phenotypes besides AMD. Altogether, 15 AMD-associated genes are located in loci associated with neurological diseases. 10 genes overlap with GWAS loci of metabolic traits and nine genes with autoimmune diseases [122].

A remarkable observation is that only 2 AMD-associated genes (*RDH5* and *COL4A3*) overlapped with loci of other complex eye diseases and traits [122]. This finding reflects the results of the retinal eQTL database. Only three eGenes of AMD-associated variants are also regulated by GWAS variants of other ocular phenotypes. Kiel et al. (2017) made the observation that genes associated with AMD in general do not overlap with genes relevant for other retinopathies [152]. Taken together, genes which expression is associated with AMD genetics often show an altered expression in various tissues. Furthermore, these genes are frequently located in GWAS loci of other complex phenotypes or traits.

In conclusion, three new comprehensive databases were generated in this thesis to allow the investigation of gene expression regulation based on genetics of complex diseases and traits. The first database represents a meta study of four earlier published datasets from liver tissue and established an up-to-date data processing and normalisation protocol. This enabled the re-analysis of data collected up to ten years ago. The second database represents the largest eQTL study in healthy retinal tissue to-date. Both data repositories identified thousands of regulatory effects and were published in open access journals to enable extensive evaluations regarding diverse hypotheses. Furthermore, a third database including multiple tissues was processed to support recent and future projects at the Institute of Human Genetics Regensburg.

All generated data in this thesis were evaluated in the context of AMD genetics. Taken together, AMD-associated variants have been shown to regulate gene expression of numerous genes. Remarkably, many of these genes are genetically regulated in multiple tissues, which raises the hypothesis that a large part of AMD risk is accompanied by differential gene expression throughout the entire body. Furthermore, AMD-associated genes seem to be also relevant for many other complex phenotypes, which allows to put forward new hypotheses about shared mechanisms in AMD aetiology.

We should be aware, however, that gene expression is only one molecular phenotype of interest to investigate for disease-associated variants. Presently, various new QTL studies are emerging [153]. Moreover, novel model systems and experimental setups are required to validate bioinformatical findings. Especially targeted genome editing opened new avenues to investigate genetically regulated genes and processes.

6 References

1. Wong WL, Su X, Li X, Cheung CMG, Klein R, Cheng C-Y, et al. (2014) Global prevalence of age-related macular degeneration and disease burden projection for 2020 and 2040: a systematic review and meta-analysis. Lancet Glob Heal 2: e106–e116.

2. Ferris FL, Davis MD, Clemons TE, Lee L-Y, Chew EY, Lindblad AS, et al. (2005) A Simplified Severity Scale for Age-Related Macular Degeneration. Arch Ophthalmol 123: 1570.

3. Swaroop A, Chew EY, Bowes Rickman C, Abecasis GR (2009) Unraveling a Multifactorial Late-Onset Disease: From Genetic Susceptibility to Disease Mechanisms for Age-Related Macular Degeneration. Annu Rev Genomics Hum Genet 10: 19–43.

4. Rosenfeld PJ, Brown DM, Heier JS, Boyer DS, Kaiser PK, Chung CY, et al. (2006) Ranibizumab for Neovascular Age-Related Macular Degeneration. N Engl J Med 355: 1419–1431.

5. Brown DM, Kaiser PK, Michels M, Soubrane G, Heier JS, Kim RY, et al. (2006) Ranibizumab versus verteporfin for neovascular age-related macular degeneration. N Engl J Med 355: 1432–1444.

6. Reynolds R, Hartnett ME, Atkinson JP, Giclas PC, Rosner B, Seddon JM (2009) Plasma complement components and activation fragments: Associations with age-related macular degeneration genotypes and phenotypes. Investig Ophthalmol Vis Sci 50: 5818–5827.

7. Ansari M, Mckeigue PM, Skerka C, Hayward C, Rudan I, Vitart V, et al. (2013) Genetic influences on plasma CFH and CFHR1 concentrations and their role in susceptibility to age-related macular degeneration. Hum Mol Genet 22: 4857–4869.

8. Colijn JM, den Hollander AI, Demirkan A, Cougnard-Grégoire A, Verzijden T, Kersten E, et al. (2019) Increased High-Density Lipoprotein Levels Associated with Age-Related Macular Degeneration. Ophthalmology 126: 393–406.

9. Wang Y, Wang M, Zhang X, Zhang Q, Nie J, Zhang M, et al. (2016) The

association between the lipids levels in blood and risk of age-related macular degeneration. Nutrients 8: 663.

10. Chew EY, Clemons T, Sangiovanni JP, Danis R, Domalpally A, McBee W, et al. (2012) The age-related eye disease study 2 (AREDS2): Study design and baseline characteristics (AREDS2 Report Number 1). Ophthalmology 119: 2282–2289.

11. Smith W, Assink J, Klein R, Mitchell P, Klaver CCW, Klein BEK, et al. (2001) Risk factors for age-related macular degeneration. Ophthalmology 108: 697–704.

12. Fritsche LG, Fariss RN, Stambolian D, Abecasis GR, Curcio CA, Swaroop A (2014) Age-Related Macular Degeneration: Genetics and Biology Coming Together. Annu Rev Genomics Hum Genet 15: 151–171.

13. Seddon JM, Cote J, Page WF, Aggen SH, Neale MC (2005) The US twin study of age-related macular degeneration: Relative roles of genetic and environmental influences. Arch Ophthalmol 123: 321–327.

14. Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN (2003) Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. Nat Genet 33: 177–182.

15. Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, et al. (2005) Complement factor H polymorphism in age-related macular degeneration. Science (80-) 308: 385–389.

16. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, et al. (2017) The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res 45: D896–D901.

17. Fritsche LG, Chen W, Schu M, Yaspan BL, Yu Y, Thorleifsson G, et al. (2013) Seven new loci associated with age-related macular degeneration. Nat Genet 45: 433–439.

18. Fritsche LG, Igl W, Bailey JNC, Grassmann F, Sengupta S, Bragg-Gresham JL, 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.

19. Fritsche LG, Lauer N, Hartmann A, Stippa S, Keilhauer CN, Oppermann M, et al. (2010) An imbalance of human complement regulatory proteins CFHR1, CFHR3 and factor H influences risk for age-related macular degeneration (AMD). Hum Mol Genet 19: 4694–4704.

20. Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP, et al. (2010) LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics 26: 2336–2337.

21. Rivera A, Fisher SA, Fritsche LG, Keilhauer CN, Lichtner P, Meitinger T, et al. (2005) Hypothetical LOC387715 is a second major susceptibility gene for age-related macular degeneration, contributing independently of complement factor H to disease risk. Hum Mol Genet 14: 3227–3236.

22. Friedrich U, Myers CA, Fritsche LG, Milenkovich A, Wolf A, Corbo JC, et al. (2011) Risk- and non-risk-associated variants at the 10q26 AMD locus influence ARMS2 mRNA expression but exclude pathogenic effects due to protein deficiency. Hum Mol Genet 20: 1387–1399.

23. Cheng Y, Huang LZ, Li X, Zhou P, Zeng W, Zhang CF (2013) Genetic and Functional Dissection of ARMS2 in Age-Related Macular Degeneration and Polypoidal Choroidal Vasculopathy. PLoS One 8: e53665.

24. Kanda A, Chen W, Othman M, Branham KEH, Brooks M, Khanna R, et al. (2007) A variant of mitochondrial protein LOC387715/ARMS2, not HTRA1, is strongly associated with age-related macular degeneration. Proc Natl Acad Sci U S A 104: 16227–16232.

25. Grassmann F, Heid IM, Weber BHF, Fritsche LG, Igl W, Bailey JN, et al. (2017) Recombinant haplotypes narrow the ARMS2/HTRA1 association signal for age-related macular degeneration. Genetics 205: 919–924.

26. Lambert JC, Ibrahim-Verbaas CA, Harold D, Naj AC, Sims R, Bellenguez C, et al. (2013) Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nat Genet 45: 1452–1458.

27. Ripke S, Neale BM, Corvin A, Walters JTR, Farh KH, Holmans PA, et al. (2014)

Biological insights from 108 schizophrenia-associated genetic loci. Nature 511: 421–427.

28. Bailey JNC, Loomis SJ, Kang JH, Allingham RR, Gharahkhani P, Khor CC, et al. (2016) Genome-wide association analysis identifies TXNRD2, ATXN2 and FOXC1 as susceptibility loci for primary open-angle glaucoma. Nat Genet 48: 189–194.

29. Kiefer AK, Tung JY, Do CB, Hinds DA, Mountain JL, Francke U, et al. (2013) Genome-wide analysis points to roles for extracellular matrix remodeling, the visual cycle, and neuronal development in myopia. PLoS Genet 9: e1003299.

30. Han J, Kraft P, Nan H, Guo Q, Chen C, Qureshi A, et al. (2008) A Genome-Wide Association Study Identifies Novel Alleles Associated with Hair Color and Skin Pigmentation. PLoS Genet 4: e1000074.

31. Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, Rivadeneira F, et al. (2010) Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature 467: 832–838.

32. Willer CJ, Schmidt EM, Sengupta S, Peloso GM, Gustafsson S, Kanoni S, et al. (2013) Discovery and refinement of loci associated with lipid levels. Nat Genet 45: 1274–1285.

33. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. (2019) The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res 47: D1005–D1012.

34. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. (2015) UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLOS Med 12: e1001779.

35. Ward LD, Kellis M (2012) Interpreting noncoding genetic variation in complex traits and human disease. Nat Biotechnol 30: 1095–1106.

36. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, et al. (2012) Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. Science (80-) 337: 1190–1195.

37. Maller JB, McVean G, Byrnes J, Vukcevic D, Palin K, Su Z, et al. (2012) Bayesian refinement of association signals for 14 loci in 3 common diseases. Nat Genet 44: 1294–1301.

38. Westra HJ, Franke L (2014) From genome to function by studying eQTLs. Biochim Biophys Acta - Mol Basis Dis 1842: 1896–1902.

39. Battle A, Montgomery SB (2014) Determining causality and consequence of expression quantitative trait loci. Hum Genet 133: 727–735.

40. Vogel C, Marcotte EM (2012) Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. Nat Rev Genet 13: 227–232.

41. Schröder A, Klein K, Winter S, Schwab M, Bonin M, Zell A, et al. (2013) Genomics of ADME gene expression: mapping expression quantitative trait loci relevant for absorption, distribution, metabolism and excretion of drugs in human liver. Pharmacogenomics J 13: 12–20.

42. Min JL, Taylor JM, Richards JB, Watts T, Pettersson FH, Broxholme J, et al. (2011) The Use of Genome-Wide eQTL Associations in Lymphoblastoid Cell Lines to Identify Novel Genetic Pathways Involved in Complex Traits. PLoS One 6: e22070.

43. Cookson W, Liang L, Abecasis G, Moffatt M, Lathrop M (2009) Mapping complex disease traits with global gene expression. Nat Rev Genet 10: 184–194.

44. Aguet F, Brown AA, Castel SE, Davis JR, He Y, Jo B, et al. (2017) Genetic effects on gene expression across human tissues. Nature 550: 204–213.

45. Kim Y, Xia K, Tao R, Giusti-Rodriguez P, Vladimirov V, van den Oord E, et al. (2014) A meta-analysis of gene expression quantitative trait loci in brain. Transl Psychiatry 4: e459.

46. Ardlie KG, Deluca DS, Segre A V., Sullivan TJ, Young TR, Gelfand ET, et al. (2015) The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science (80-) 348: 648–660.

47. Innocenti F, Cooper GM, Stanaway IB, Gamazon ER, Smith JD, Mirkov S, et al. (2011) Identification, Replication, and Functional Fine-Mapping of Expression Quantitative Trait Loci in Primary Human Liver Tissue. PLoS Genet 7: e1002078.

48. Shrier I, Platt RW, Steele RJ (2007) Mega-trials vs. meta-analysis: precision vs. heterogeneity? Contemp Clin Trials 28: 324–328.

49. Hoerl AE, Kennard RW (1970) Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics 12: 55–67.

50. Tibshirani R (1996) Regression Shrinkage and Selection Via the Lasso. J R Stat Soc Ser B 58: 267–288.

51. Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. J R Stat Soc Ser B (Statistical Methodol 67: 301–320.

52. Gusev A, Ko A, Shi H, Bhatia G, Chung W, Penninx BWJH, et al. (2016) Integrative approaches for large-scale transcriptome-wide association studies. Nat Genet 48: 245–252.

53. Gamazon ER, Wheeler HE, Shah KP, Mozaffari S V., Aquino-Michaels K, Carroll RJ, et al. (2015) A gene-based association method for mapping traits using reference transcriptome data. Nat Genet 47: 1091–1098.

54. Wainberg M, Sinnott-Armstrong N, Mancuso N, Barbeira AN, Knowles DA, Golan D, et al. (2019) Opportunities and challenges for transcriptome-wide association studies. Nat Genet 51: 592–599.

55. Gallagher MD, Chen-Plotkin AS (2018) The Post-GWAS Era: From Association to Function. Am J Hum Genet 102: 717–730.

56. Ishino Y, Shinagawa H, Makino K, Amemura M, Nakatura A (1987) Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isoenzyme conversion in Escherichia coli, and identification of the gene product. J Bacteriol 169: 5429–5433.

57. Gasiunas G, Barrangou R, Horvath P, Siksnys V (2012) Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. Proc Natl Acad Sci 109: E2579–E2586.

58. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E (2012) A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity. Science (80-) 337: 816–821.

59. Wang H, La Russa M, Qi LS (2016) CRISPR/Cas9 in Genome Editing and Beyond. Annu Rev Biochem 85: 227–264.

60. Rouet P, Smih F, Jasin M (1994) Introduction of double-strand breaks into the genome of mouse cells by expression of a rare-cutting endonuclease. Mol Cell Biol 14: 8096–8106.

61. Rouet P, Smih F, Jasin M (1994) Expression of a site-specific endonuclease stimulates homologous recombination in mammalian cells. Proc Natl Acad Sci 91: 6064–6068.

62. Rudin N, Sugarman E, Haber JE (1989) Genetic and physical analysis of double-strand break repair and recombination in Saccharomyces cerevisiae. Genetics 122: 519–534.

63. Chen X, Xu F, Zhu C, Ji J, Zhou X, Feng X, et al. (2014) Dual sgRNA-directed gene knockout using CRISPR/Cas9 technology in Caenorhabditis elegans. Sci Rep 4: 7581.

64. Han J, Zhang J, Chen L, Shen B, Zhou J, Hu B, et al. (2014) Efficient in vivo deletion of a large imprinted lncRNA by CRISPR/Cas9. RNA Biol 11: 829–835.

65. Gilbert LA, Larson MH, Morsut L, Liu Z, Brar GA, Torres SE, et al. (2013) CRISPR-Mediated Modular RNA-Guided Regulation of Transcription in Eukaryotes. Cell 154: 442–451.

66. Chavez A, Scheiman J, Vora S, Pruitt BW, Tuttle M, P R Iyer E, et al. (2015) Highly efficient Cas9-mediated transcriptional programming. Nat Methods 12: 326–328.

67. Hilton IB, D'Ippolito AM, Vockley CM, Thakore PI, Crawford GE, Reddy TE, et al. (2015) Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. Nat Biotechnol 33: 510–517.

68. Schrode N, Ho S-M, Yamamuro K, Dobbyn A, Huckins L, Matos MR, et al. (2019) Synergistic effects of common schizophrenia risk variants. Nat Genet 51: 1475–1485.

69. Schadt EE, Molony C, Chudin E, Hao K, Yang X, Lum PY, et al. (2008) Mapping the Genetic Architecture of Gene Expression in Human Liver. PLoS Biol 6: e107.

70. Ratnapriya R, Sosina OA, Starostik MR, Kwicklis M, Kapphahn RJ, Fritsche LG, et al. (2019) Retinal transcriptome and eQTL analyses identify genes associated with age-related macular degeneration. Nat Genet 51: 606–610.

71. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. (2011) The variant call format and VCFtools. Bioinformatics 27: 2156–2158.

72. Smedley D, Haider S, Ballester B, Holland R, London D, Thorisson G, et al. (2009) BioMart - Biological queries made easy. BMC Genomics 10: 22.

73. Rosenbloom KR, Armstrong J, Barber GP, Casper J, Clawson H, Diekhans M, et al. (2015) The UCSC Genome Browser database: 2015 update. Nucleic Acids Res 43: D670–D681.

74. Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, et al. (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56–65.

75. R Team Core (2017) A language and environment for statistical computing. R Found Stat Comput Vienna, Austria: 2017.

76. Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS (2012) A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics 28: 3326–3328.

77. Wigginton JE, Cutler DJ, Abecasis GR (2005) A Note on Exact Tests of Hardy-Weinberg Equilibrium. Am J Hum Genet 76: 887–893.

78. Delaneau O, Marchini J, Zagury JF (2012) A linear complexity phasing method for thousands of genomes. Nat Methods 9: 179–181.

79. Howie B, Marchini J, Stephens M (2011) Genotype imputation with thousands of genomes. G3 Genes, Genomes, Genet 1: 457–470.

80. Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, et al. (2016) Ensembl 2016. Nucleic Acids Res 44: D710–D716.

81. Arloth J, Bader DM, Röh S, Altmann A (2015) Re-Annotator: Annotation Pipeline for Microarray Probe Sequences. PLoS One 10: e0139516.

82. Ramasamy A, Trabzuni D, Gibbs JR, Dillman A, Hernandez DG, Arepalli S, et

al. (2013) Resolving the polymorphism-in-probe problem is critical for correct interpretation of expression QTL studies. Nucleic Acids Res 41: e88.

83. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. (2016) Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 44: D733–D745.

84. Shapiro SS, Wilk MB (1965) An Analysis of Variance Test for Normality (Complete Samples). Biometrika 52: 591–611.

85. Ewels P, Magnusson M, Lundin S, Käller M (2016) MultiQC: Summarize analysis results for multiple tools and samples in a single report. Bioinformatics 32: 3047–3048.

86. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30: 2114–2120.

87. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. (2013) STAR: Ultrafast universal RNA-seq aligner. Bioinformatics 29: 15–21.

88. Li B, Dewey CN (2011) RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12: 323.

89. Robinson MD, McCarthy DJ, Smyth GK (2009) edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26: 139–140.

90. Robinson MD, Oshlack A (2010) A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol 11: R25.

91. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, et al. (2001) Missing value estimation methods for DNA microarrays. Bioinformatics 17: 520–525.

92. Bolstad BM, Irizarry R., Astrand M, Speed TP (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19: 185–193.

93. Johnson WE, Li C, Rabinovic A (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 8: 118–127.

94. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD (2012) The SVA package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics 28: 882–883.

95. Shabalin AA (2012) Matrix eQTL: ultra fast eQTL analysis via large matrix operations. Bioinformatics 28: 1353–1358.

96. Viechtbauer W (2010) Conducting Meta-Analyses in R with the metafor Package. J Stat Softw 36: 1–48.

97. Reimand J, Arak T, Vilo J (2011) g:Profiler—a web server for functional interpretation of gene lists (2011 update). Nucleic Acids Res 39: W307–W315.

98. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. (2000) Gene Ontology: tool for the unification of biology. Nat Genet 25: 25–29.

99. Galili T (2015) dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. Bioinformatics 31: 3718–3720.

100. Haeussler M, Zweig AS, Tyner C, Speir ML, Rosenbloom KR, Raney BJ, et al. (2019) The UCSC Genome Browser database: 2019 update. Nucleic Acids Res 47: D853–D858.

101. Jordan M, Schallhorn A, Wurm FM (1996) Transfecting Mammalian Cells: Optimization of Critical Parameters Affecting Calcium-Phosphate Precipitate Formation. Nucleic Acids Res 24: 596–601.

102. Laird PW, Zijderveld A, Linders K, Rudnicki MA, Jaenisch R, Berns A (1991) Simplified mammalian DNA isolation procedure. Nucleic Acids Res 19: 4293–4293.

103. Strunz T, Grassmann F, Gayán J, Nahkuri S, Souza-Costa D, Maugeais C, et al. (2018) A mega-analysis of expression quantitative trait loci (eQTL) provides insight into the regulatory architecture of gene expression variation in liver. Sci Rep 8: 5865.

104. Benjamini Y, Hochberg Y (2000) On the Adaptive Control of the False Discovery Rate in Multiple Testing With Independent Statistics. J Educ Behav Stat 25: 60–83.

105. Crowder M (2011) Meta-analysis and Combining Information in Genetics and Genomics edited by Rudy Guerra, Darlene R. Goldstein. Int Stat Rev 79: 134–135.

106. Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, et al. (2012) Annotation of functional variation in personal genomes using RegulomeDB. Genome Res 22: 1790–1797.

107. McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F (2010) Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics 26: 2069–2070.

108. Brown AA, Viñuela A, Delaneau O, Spector TD, Small KS, Dermitzakis ET (2017) Predicting causal variants affecting expression by using whole-genome sequencing and RNA-seq from multiple human tissues. Nat Genet 49: 1747–1751.

109. Heinzen EL, Ge D, Cronin KD, Maia JM, Shianna K V, Gabriel WN, et al. (2008) Tissue-Specific Genetic Control of Splicing: Implications for the Study of Complex Traits. PLoS Biol 6: e1000001.

110. Frankish A, Diekhans M, Ferreira A-M, Johnson R, Jungreis I, Loveland J, et al. (2019) GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res 47: D766–D773.

111. Borrego F (2013) The CD300 molecules: an emerging family of regulators of the immune system. Blood 121: 1951–1960.

112. Brckalo T, Calzetti F, Pérez-Cabezas B, Borràs FE, Cassatella MA, López-Botet M (2010) Functional analysis of the CD300e receptor in human monocytes and myeloid dendritic cells. Eur J Immunol 40: 722–732.

113. Clark GJ, Ju X, Azlan M, Tate C, Ding Y, Hart DNJ (2009) The CD300 molecules regulate monocyte and dendritic cell functions. Immunobiology 214: 730–736.

114. Jin Y, Ratnam K, Chuang PY, Fan Y, Zhong Y, Dai Y, et al. (2012) A systems approach identifies HIPK2 as a key regulator of kidney fibrosis. Nat Med 18: 580–588.

115. Staege MS, Hutter C, Neumann I, Foja S, Hattenhorst UE, Hansen G, et al.

(2004) DNA Microarrays Reveal Relationship of Ewing Family Tumors to Both Endothelial and Fetal Neural Crest-Derived Cells and Define Novel Targets. Cancer Res 64: 8213–8221.

116. Dumaual CM, Steere BA, Walls CD, Wang M, Zhang Z-Y, Randall SK (2013) Integrated Analysis of Global mRNA and Protein Expression Data in HEK293 Cells Overexpressing PRL-1. PLoS One 8: e72977.

117. Liang X, Potter J, Kumar S, Zou Y, Quintanilla R, Sridharan M, et al. (2015) Rapid and highly efficient mammalian cell engineering via Cas9 protein transfection. J Biotechnol 208: 44–53.

118. Mashiko D, Fujihara Y, Satouh Y, Miyata H, Isotani A, Ikawa M (2013) Generation of mutant mice by pronuclear injection of circular plasmid expressing Cas9 and single guided RNA. Sci Rep 3: 3355.

119. Strunz T, Kiel C, Grassmann F, Ratnapriya R, Kwicklis M, Karlstetter M, et al. (2020) A mega-analysis of expression quantitative trait loci in retinal tissue. PLOS Genet 16: e1008934.

120. Anand L (2019) chromoMap: An R package for Interactive Visualization and Annotation of Chromosomes. Cold Spring Harbor Laboratory. 605600 p.

121. Grassmann F, Kiel C, Zimmermann ME, Gorski M, Grassmann V, Stark K, et al. (2017) Genetic pleiotropy between age-related macular degeneration and 16 complex diseases and traits. Genome Med 9: 29.

122. Strunz T, Lauwen S, Kiel C, Fritsche LG, Igl W, Bailey JNC, et al. (2020) A transcriptome-wide association study based on 27 tissues identifies 106 genes potentially relevant for disease pathology in age-related macular degeneration. Sci Rep 10: 1584.

123. Spencer KL, Hauser M a, Olson LM, Schmidt S, Scott WK, Gallins P, et al. (2008) Deletion of CFHR3 and CFHR1 genes in age-related macular degeneration. Hum Mol Genet 17: 971–977.

124. Luo J, Schumacher M, Scherer A, Sanoudou D, Megherbi D, Davison T, et al. (2010) A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data.

Pharmacogenomics J 10: 278–291.

125. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, et al. (2010) Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet 11: 733–739.

126. Huang QQ, Ritchie SC, Brozynska M, Inouye M (2018) Power, false discovery rate and Winner's Curse in eQTL studies. Nucleic Acids Res 46: e133.

127. Ferreira PG, Muñoz-Aguirre M, Reverter F, Sá Godinho CP, Sousa A, Amadoz A, et al. (2018) The effects of death and post-mortem cold ischemia on human tissue transcriptomes. Nat Commun 9: 490.

128. Raychaudhuri S (2011) Mapping Rare and Common Causal Alleles for Complex Human Diseases. Cell 147: 57–69.

129. Ulirsch JC, Nandakumar SK, Wang L, Giani FC, Zhang X, Rogov P, et al. (2016) Systematic Functional Dissection of Common Genetic Variation Affecting Red Blood Cell Traits. Cell 165: 1530–1545.

130. Lunyak V V., Prefontaine GG, Nunez E, Cramer T, Ju B-G, Ohgi KA, et al. (2007) Developmentally Regulated Activation of a SINE B2 Repeat as a Domain Boundary in Organogenesis. Science (80-) 317: 248–251.

131. Schmitz J (2012) SINEs as Driving Forces in Genome Evolution. Repetitive DNA 7: 92–107.

132. Mariner PD, Walters RD, Espinoza CA, Drullinger LF, Wagner SD, Kugel JF, et al. (2008) Human Alu RNA Is a Modular Transacting Repressor of mRNA Transcription during Heat Shock. Mol Cell 29: 499–509.

133. Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, et al. (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489: 57–74.

134. El-Brolosy MA, Stainier DYR (2017) Genetic compensation: A phenomenon in search of mechanisms. PLoS Genet 13: e1006780.

135. Fromer M, Roussos P, Sieberts SK, Johnson JS, Kavanagh DH, Perumal TM, et al. (2016) Gene expression elucidates functional impact of polygenic risk for

schizophrenia. Nat Neurosci 19: 1442–1453.

136. Alper CA, Myron Johnson A, Birtch AG, Moore FD (1969) Human C3: Evidence for the liver as the primary site of synthesis. Science (80-) 163: 286–288.

137. Morgan BP, Gasque P (1997) Extrahepatic complement biosynthesis: Where, when and why? Clin Exp Immunol 107: 1–7.

138. Lewis GF, Rader DJ (2005) New Insights Into the Regulation of HDL Metabolism and Reverse Cholesterol Transport. Circ Res 96: 1221–1232.

139. Mousseau DD, Banville D, L'Abbé D, Bouchard P, Shen S-H (2000) PILRα, a Novel Immunoreceptor Tyrosine-based Inhibitory Motif-bearing Protein, Recruits SHP-1 upon Tyrosine Phosphorylation and Is Paired with the Truncated Counterpart PILRβ. J Biol Chem 275: 4467–4474.

140. Patel T, Brookes KJ, Turton J, Chaudhury S, Guetta-Baranes T, Guerreiro R, et al. (2018) Whole-exome sequencing of the BDR cohort: evidence to support the role of the PILRA gene in Alzheimer's disease. Neuropathol Appl Neurobiol 44: 506–521.

141. Kikuchi M, Hara N, Hasegawa M, Miyashita A, Kuwano R, Ikeuchi T, et al. (2019) Enhancer variants associated with Alzheimer's disease affect gene expression via chromatin looping. BMC Med Genomics 12: 128.

142. Ding X, Patel M, Chan CC (2009) Molecular pathology of age-related macular degeneration. Prog Retin Eye Res 28: 1–18.

143. Jaffe AE, Straub RE, Shin JH, Tao R, Gao Y, Collado-Torres L, et al. (2018) Developmental and genetic regulation of the human cortex transcriptome illuminate schizophrenia pathogenesis. Nat Neurosci 21: 1117–1125.

144. Zhong J, Jermusyk A, Wu L, Hoskins JW, Collins I, Mocci E, et al. (2020) A Transcriptome-Wide Association Study (TWAS) Identifies Novel Candidate Susceptibility Genes for Pancreatic Cancer. JNCI J Natl Cancer Inst: djz246.

145. Bossé Y, Li Z, Xia J, Manem V, Carreras-Torres R, Gabriel A, et al. (2019) Transcriptome-wide association study reveals candidate causal genes for lung cancer. Int J Cancer 146: 1862–1878.

146. Pain O, Pocklington AJ, Holmans PA, Bray NJ, O'Brien HE, Hall LS, et al. (2019) Novel Insight Into the Etiology of Autism Spectrum Disorder Gained by Integrating Expression Data With Genome-wide Association Statistics. Biol Psychiatry 86: 265–273.

147. Yang Z, Tong Z, Chen Y, Zeng J, Lu F, Sun X, et al. (2010) Genetic and functional dissection of HTRA1 and LOC387715 in age-related macular degeneration. PLoS Genet 6: e1000836.

148. Voigt AP, Mulfaul K, Mullin NK, Flamme-Wiese MJ, Giacalone JC, Stone EM, et al. (2019) Single-cell transcriptomics of the human retinal pigment epithelium and choroid in health and macular degeneration. Proc Natl Acad Sci U S A 116: 24100–24107.

149. Eisenberg E, Levanon EY (2013) Human housekeeping genes, revisited. Trends Genet 29: 569–574.

150. Keilhauer CN, Fritsche LG, Guthoff R, Haubitz I, Weber BH (2013) Age-related macular degeneration and coronary heart disease: Evaluation of genetic and environmental associations. Eur J Med Genet 56: 72–79.

151. Kao LT, Wang KH, Lin HC, Tsai MC, Chung SD (2015) Association between psoriasis and neovascular age-related macular degeneration: A population-based study. J Am Acad Dermatol 72: 1090–1092.

152. Kiel C, Lastrucci C, Luthert PJ, Serrano L (2017) Simple and complex retinal dystrophies are associated with profoundly different disease networks. Sci Rep 7: 41835.

153. Zheng Z, Huang D, Wang J, Zhao K, Zhou Y, Guo Z, et al. (2020) QTLbase: an integrative resource for quantitative trait loci across multiple human molecular phenotypes. Nucleic Acids Res 48: D983–D991.

List of abbreviations

Abbreviation	Meaning
AD	Alzheimer's disease
AMD	Age-related macular degeneration
beta-SE	Standard error of the effect size
cDNA	Complementary DNA
CNV	Choroidal neovascularization
CPM	Counts per Million
CRISPR	Clustered regularly interspaced short palindromic repeats
dCas9	Nuclease-deactivated Cas9
DSB	Double-strand break
E. coli	Escherichia coli
eGene	eQTL gene
eQTL	Expression quantitative trait loci
EUR	European
eVariant	eQTL variant
FACS	Fluorescence activated cell sorting
FCS	Fetal bovine serum
GA	Geographic atrophy
gDNA	Genomic DNA
GO	Gene ontology
GTEx	Genotype-Tissue Expression project
GTEx v6	GTEx release 6
GTEx v7	GTEx release 7
GWAS	Genome-wide association study
HDL	High-density lipoprotein
HDR	Homology-directed repair
HWE	Hardy-Weinberg equilibrium
IAMDGC	International AMD Genomics Consortium
IH	Independent hit
indel	Small insertions or deletions
LD	Linkage disequilibrium
LDL	Low-density lipoprotein
MAF	Minor allele frequency
MHC	Major histocompatibility complex
NEI	National Eye Institute
NHEJ	Nonhomologous end joining
nt	Nucleotide
OR	Odds ratio
PAM	Protospacer-adjacent motif
PC	Principal component
PCA	Principal component analysis
PCR	Polymerase chain reaction
Pen/Strep	Penicillin/Streptomycin
QC	Quality control

QN	Quantile normalization
QTL	Quantitative trait locus
qRT-PCR	Quantitative real-time PCR
RNA-Seq	RNA sequencing
RPE	Retinal pigment epithelium
SD	Standard deviation
SINE	Short interspersed nuclear element
TS	Tissue-specific
TSS	Transcription start site
TWAS	Transcriptome-wide association study
UTR	Untranslated transcript region
VEP	Variant Effect Predictor
VPR	VP64-p65-Rta
WT	Wild type

List of figures

List of tables

List of supplementary tables

Acknowledgements

I wish to express my deepest gratitude to my supervisor Prof. Dr. Bernhard Weber for perfectly supporting me in all aspects of my thesis. I really appreciated the opportunity to pursue various projects and that he was always open-minded for new methods and novel ways of result interpretation.

I would like to say a special thank you to Dr. Everson Nogoceke and Dr. Felix Grassmann for their supervision and support.

I also wish to show my gratitude to my mentors Prof. Dr. Rainer Spang and Prof. Michael Rehli for their advice and the critical discussions.

Thanks to Andrea Milenkovic for sharing her laboratory experience and for helping me with the establishment of protocols.

I wish to show my gratitude to Christina Kiel for the many fruitful discussions and her great ideas.

I am also indebted to all the patients and controls that participated in the various studies. None of the projects would have been possible without their willingness to participate.

I am grateful for the support of the Helmut Ecker Stiftung, which enabled my research and allowed to investigate various projects.

Very special thank you to all my colleagues at the Institute of Human Genetics for the positive and constructive atmosphere and the possibility to discuss findings and workflows.

I would like to thank my girlfriend Ann-Kathrin for her encouragement and her support for pursuing my doctoral thesis in Regensburg.

Last but not least, I wish to show my gratitude to my family and friends for their continuous support and encouragement during the last years.

Supplements

Supplementary Table 1: Study and sample summary of the in-house GTEx v7 database

Tissue	Sample size	Expressed genes (RPKM > 1)	Q-value < 0.05				Q-value < 0.001			
			eQTL	eVariant (unique)	eVariant (multiple genes)	eGenes (unique)	eQTL	eVariant (unique)	eVariant (multiple genes)	eGene (unique)
Adipose subcutaneous	321	32,045	954,180	584,487	167,434	16,715	461,840	289,459	79,581	4,567
Adipose visceral omentum	264	31,581	595,155	388,048	97,814	12,853	283,576	183,326	44,112	3,127
Adrenal gland	148	28,134	331,389	232,064	48,876	9,491	145,933	96,500	21,272	2,124
Artery aorta	232	29,666	654,296	426,728	105,377	13,631	305,367	199,792	47,714	3,510
Artery coronary	124	28,114	226,321	159,227	31,389	7,843	107,489	67,125	14,553	1,697
Artery tibial	325	29,980	823,197	536,865	140,427	15,255	393,223	264,563	66,432	4,071
Brain amygdala	80	26,228	145,176	103,764	12,462	5,855	56,194	34,406	4,149	1,303
Brain anterior cingulate cortex	102	27,042	188,474	138,586	20,135	6,898	77,660	51,462	7,277	1,471
Brain caudate basal ganglia	129	28,780	266,517	194,531	31,300	9,606	119,976	76,873	12,115	2,185
Brain cerebellar hemisphere	114	28,521	398,574	256,283	54,530	11,613	167,590	99,206	23,458	2,750
Brain cerebellum	144	30,637	563,368	359,234	87,436	14,881	245,695	147,380	35,925	3,917
Brain cortex	124	28,410	331,768	234,420	43,009	11,836	135,445	91,158	14,818	3,007
Brain frontal cortex	112	27,599	222,449	160,231	27,499	8,569	94,837	61,781	10,581	1,869
Brain hippocampus	98	27,336	157,948	112,752	16,450	5,463	71,681	43,024	7,061	1,173
Brain hypothalamus	101	28,334	173,870	122,671	20,117	6,575	78,546	47,527	8,844	1,376
Brain nucleus accumbens basal ganglia	118	28,500	232,540	162,429	27,222	8,372	96,635	64,121	11,170	1,815
Brain putamen basal ganglia	101	26,761	201,573	141,451	20,410	7,487	92,215	53,811	9,047	1,572
Brain spinal cord cervical	74	26,519	130,781	96,802	12,381	5,272	58,247	34,556	5,167	1,179
Brain substantia nigra	72	25,943	113,778	84,726	10,280	5,143	46,369	27,536	4,707	1,090
Breast mammary tissue	206	32,201	462,591	303,878	74,630	11,010	207,952	136,398	32,067	2,539
Cells EBV-transformed lymphocytes	93	24,521	210,273	157,696	23,797	7,673	80,240	56,250	9,685	1,901

Cells transformed fibroblasts	251	26,660	552,611	375,885	86,698	11,266	254,916	175,691	36,856	2,875
Colon Sigmoid	184	29,760	416,410	282,974	63,590	11,670	183,867	122,002	27,729	2,835
Colon transverse	204	31,085	378,260	256,162	60,334	9,659	177,658	117,144	26,195	2,196
Esophagus gastroesophageal junction	187	29,224	448,797	299,682	70,054	11,832	204,653	134,764	29,459	2,880
Esophagus mucosa	310	31,367	758,704	489,631	122,935	15,218	363,028	235,892	58,529	4,027
Esophagus muscularis	280	29,935	823,304	529,158	140,087	15,161	394,017	258,413	62,865	4,106
Heart atrial appendage	224	29,081	518,888	348,663	82,143	11,972	240,167	160,478	38,948	2,912
Heart left ventricle	233	26,849	432,501	294,186	65,593	10,348	206,252	136,333	31,253	2,416
Liver	131	26,072	207,257	148,804	26,972	7,089	94,855	59,750	13,002	1,560
Lung	327	34,430	797,053	491,156	133,491	15,342	383,640	234,379	64,676	3,937
Minor salivary gland	72	28,031	123,766	90,070	12,902	5,963	51,200	30,727	5,405	1,387
Muscle skeletal	418	27,964	843,838	539,895	143,661	14,397	413,546	268,277	68,225	3,873
Nerve tibial	305	33,801	1,085,095	665,219	191,680	18,647	518,420	327,117	90,081	5,408
Ovary	96	28,610	200,460	139,210	24,185	7,107	85,188	51,142	10,612	1,588
Pancreas	174	27,931	524,816	363,890	81,784	12,348	235,522	159,041	34,454	3,245
Pituitary	148	32,261	398,450	259,858	61,107	11,772	175,493	109,696	25,661	2,687
Prostate	107	30,583	215,608	147,542	28,252	7,306	90,094	56,577	11,928	1,598
Skin not sun exposed Suprapubic	279	33,014	743,789	476,612	123,179	15,395	349,547	224,596	58,678	3,929
Skin sun exposed lower leg	365	33,940	1,028,424	623,890	183,930	18,133	492,538	310,346	87,672	5,037
Small intestine terminal ileum	102	29,667	154,391	108,823	20,377	5,741	62,724	38,319	8,279	1,159
Spleen	114	29,403	345,287	249,183	48,509	10,444	142,080	97,621	18,226	2,592
Stomach	190	30,497	334,985	230,703	46,490	9,547	152,316	100,470	20,160	2,128
Testis	197	42,810	599,548	403,791	94,666	18,773	263,356	180,254	37,793	4,768
Thyroid	342	34,789	1,244,473	737,431	224,902	19,890	605,649	369,950	107,122	5,886
Uterus	81	27,613	158,240	107,643	17,652	6,279	66,922	36,792	7,528	1,514
Vagina	88	29,030	150,503	107,535	14,549	6,135	69,917	40,730	7,561	1,585
Whole blood	323	29,151	475,540	313,432	76,710	11,047	219,796	147,379	33,798	2,666

Supplementary Table 2: Statistically significant AMD-associated genes (Q-Value < 0.001) of the TWAS analysis

Gene	Gene position [hg19]	AMD locus*	Gene expressed in tissues	Predictable tissues**	AMD associated (FDR < 0.001)	Mean beta (SD)	Strongest effect tissue***
C1orf21	1:184356192-184598154	none	27	15	1	-0.028	Liver
KCNT2	1:196194909-196578355	1	27	6	6	-0.052 (0.014)	Nerve Tibial
CFH	1:196621008-196716634	1	27	12	11	-0.052 (0.049)	Nerve Tibial
CFHR3	1:196743925-196763203	1	21	20	20	0.117 (0.055)	Liver
CFHR1	1:196788887-196801319	1	25	15	14	0.105 (0.084)	Liver
CFHR4	1:196819371-196888102	1	2	2	2	0.132 (0.128)	Liver
F13B	1:197008321-197036397	1	3	1	1	0.025	Testis
ASPM	1:197053258-197115824	1	24	2	1	0.036	Skin Not Sun Exposed Suprapubic
ZBTB41	1:197122810-197169672	1	27	5	5	0.03 (0.023)	Brain Cerebellum
RP11.332L8.1	1:197191352-197192385	1	22	1	1	-0.017	Artery Tibial
DENND1B	1:197473878-197744826	1	27	6	1	0.007	Esophagus Mucosa
LHX9	1:197881037-197904608	1	6	2	1	-0.035	Liver
CD55	1:207494853-207534311	none	27	17	3	-0.016 (0.003)	Esophagus Muscularis
CR2	1:207627575-207663240	none	15	3	1	-0.013	Muscle Skeletal
NOSTRIN	2:169643049-169722024	none	27	18	1	-0.015	Esophagus Mucosa
PPIL3	2:201735630-201754026	none	27	27	16	0.037 (0.004)	Adipose Subcutaneous
NDUFB3	2:201936156-201950473	none	27	6	4	0.005 (0.001)	Adipose Subcutaneous
COL4A3	2:228029281-228179508	2	27	7	2	-0.023 (0.011)	Nerve Tibial
TBC1D23	3:99979844-100044095	4	27	13	4	-0.017 (0.013)	Adrenal Gland
NIT2	3:100053545-100075710	4	27	17	5	-0.013 (0.005)	Lung
RP11.114I8.4	3:100080031-100080481	4	27	8	2	0.009 (0.001)	Thyroid
TOMM70A	3:100082275-100120036	4	27	14	2	0.013 (0.012)	Nerve Tibial
TMEM45A	3:100211463-100296288	4	27	11	1	-0.011	Adrenal Gland
CCDC109B	4:110481361-110609784	5	27	3	1	-0.012	Adipose Visceral Omentum

CASP6	4:110609875-110624739	5	27	5	3	0.021 (0.006)	Heart Atrial Appendage
PLA2G12A	4:110631145-110651233	5	27	15	13	0.021 (0.007)	Esophagus Mucosa
CFI	4:110661852-110723335	5	27	2	1	-0.01	Adipose Subcutaneous
ADAM19	5:156822607-157002783	none	27	21	12	-0.013 (0.006)	Adipose Subcutaneous
IP6K3	6:33689444-33714762	none	27	17	1	0.019	Cells Transformed fibroblasts
PPP2R5D	6:42952237-42979831	9	27	7	2	-0.013 (0.004)	Stomach
ZKSCAN1	7:99613204-99639312	11	27	8	1	-0.007	Artery Aorta
STAG3	7:99775186-99818169	11	27	10	1	-0.007	Adipose Subcutaneous
PMS2P1	7:99927805-99939531	11	27	17	14	-0.013 (0.005)	Testis
STAG3L5P	7:99934035-99947781	11	27	27	27	0.039 (0.006)	Artery Tibial
PILRB	7:99949799-99965356	11	27	27	27	0.042 (0.004)	Adipose Subcutaneous
PILRA	7:99971068-99997719	11	27	26	26	0.038 (0.006)	Brain Cerebellum
ZCWPW1	7:99998476-100026415	11	27	9	3	0.016 (0.005)	Nerve Tibial
TSC22D4	7:100060982-100076902	11	27	14	8	0.014 (0.007)	Thyroid
NYAP1	7:100081550-100092422	11	27	7	3	-0.017 (0.007)	Skin Sun Exposed Lower leg
RP11.325F22.5	7:104558007-104567077	10	23	3	1	0.013	Adipose Subcutaneous
RP11.325F22.2	7:104581510-104602781	10	25	10	1	0.003	Adipose Visceral Omentum
TNFRSF10A	8:23048189-23082639	12	27	20	14	-0.018 (0.008)	Cells Transformed fibroblasts
TRPM3	9:73143979-74061751	14	18	6	1	0.022	Testis
RORB	9:77112281-77308093	13	24	4	1	-0.01	Cells Transformed fibroblasts
TGFBR1	9:101866320-101916474	15	27	4	1	0.009	Whole Blood
ZFP37	9:115800660-115819039	none	27	10	1	-0.017	Adipose Subcutaneous
FGFR2	10:123237848-123357972	18	27	5	1	-0.021	Skin Not Sun Exposed Suprapubic
ATE1	10:123499939-123688316	18	27	20	2	0.024 (0.009)	Stomach
TACC2	10:123748709-124014060	18	27	13	1	-0.034	Breast Mammary Tissue
BTBD16	10:124030821-124097677	18	25	23	14	0.02 (0.033)	Brain Cerebellum
PLEKHA1	10:124134212-124191867	18	27	19	18	-0.051 (0.033)	Brain Cerebellum
ARMS2	10:124214169-124216868	18	26	14	14	-0.098 (0.09)	Testis
HTRA1	10:124221041-124274424	18	27	9	7	0.031 (0.068)	Testis
DMBT1	10:124320181-124403252	18	16	3	3	-0.02 (0.008)	Skin Sun Exposed Lower leg

RP11.318C4.2	10:124516210-124558696	18	5	3	2	-0.011 (0.003)	Skin Sun Exposed Lower leg
RP11.107C16.2	10:124578332-124585965	18	6	2	1	-0.016	Skin Sun Exposed Lower leg
RP11.564D11.3	10:124639246-124658230	18	18	3	1	0.012	Brain Cerebellum
IKZF5	10:124750322-124768333	18	27	8	1	-0.031	Stomach
ACADSB	10:124768495-124817827	18	27	6	1	-0.014	Adipose Subcutaneous
RP11.777F6.3	11:87034801-87035401	none	27	2	1	0.007	Testis
CEP57	11:95523129-95565857	none	27	23	3	-0.02 (0.005)	Skin Not Sun Exposed Suprapubic
AP001877.1	11:95556681-95557336	none	27	24	8	-0.016 (0.006)	Nerve Tibial
BLOC1S1	12:56109828-56113871	19	27	12	1	0.006	Muscle Skeletal
RDH5	12:56114151-56118489	19	27	23	17	-0.018 (0.005)	Lung
B3GALTL	13:31774073-31906413	21	27	21	5	0.013 (0.005)	Heart Left Ventricle
PLEKHH1	14:68000018-68056027	22	27	14	2	0.016 (0.007)	Artery Aorta
RIN3	14:92980118-93155339	none	27	13	1	0.018	Colon Sigmoid
ALDH1A2	15:58245622-58790065	23	27	6	1	0.01	Liver
LIPC	15:58702768-58861151	23	24	15	1	0.037	Liver
ULK3	15:75128457-75135538	none	27	17	1	-0.01	Lung
USP7	16:8985951-9058371	none	27	3	1	0.014	Muscle Skeletal
MT1DP	16:56677617-56678698	24	27	4	1	0.014	Lung
HERPUD1	16:56965960-56977798	24	27	8	2	-0.009 (0.004)	Esophagus Mucosa
CETP	16:56995762-57017757	24	27	5	4	-0.017 (0.006)	Colon Transverse
NLRC5	16:57023397-57117443	24	27	9	2	-0.037 (0.019)	Cells Transformed fibroblasts
GPR56	16:57644564-57698944	24	27	2	1	-0.009	Breast Mammary Tissue
BCAR1	16:75262928-75301951	25	27	11	2	-0.011 (0.001)	Brain Cerebellum
CFDP1	16:75327596-75467383	25	27	25	4	-0.01 (0.006)	Esophagus Muscularis
TMEM170A	16:75476952-75499395	25	27	10	2	0.019 (0.001)	Adrenal Gland
TMEM97	17:26646121-26655351	26	27	12	2	0.016 (0.001)	Breast Mammary Tissue
POLDIP2	17:26674036-26684545	26	27	15	3	0.011 (0.01)	Pituitary
TMEM199	17:26684604-26690705	26	27	14	10	0.012 (0.004)	Skin Sun Exposed Lower leg
C17orf70	17:79506911-79520987	27	27	3	1	0.008	Artery Tibial
NPLOC4	17:79523913-79604172	27	27	24	1	0.022	Testis

PDE6G	17:79617489-79630142	27	27	4	1	-0.025	Testis
AC006273.5	19:782755-785080	29	27	2	1	0.007	Skin Not Sun Exposed Suprapubic
MED16	19:867962-893218	29	27	9	3	0.022 (0.004)	Muscle Skeletal
GRIN3B	19:1000418-1009646	29	27	26	2	-0.012 (0.002)	Whole Blood
CNN2	19:1026298-1039068	29	27	13	1	-0.029	Whole Blood
ABCA7	19:1040102-1065568	29	27	24	2	-0.028 (0.01)	Whole Blood
CTC.503J8.6	19:6210390-6212492	28	27	4	1	-0.01	Artery Tibial
GTF2F1	19:6379580-6393992	28	27	15	1	-0.011	Colon Sigmoid
GPR108	19:6729925-6737614	28	27	27	22	0.031 (0.008)	Thyroid
RELB	19:45504688-45541452	30	27	1	1	-0.008	Lung
BLOC1S3	19:45682003-45685059	30	27	4	1	0.009	Esophagus Muscularis
DMPK	19:46272975-46285810	30	27	16	1	0.008	Stomach
FUT2	19:49199228-49209207	none	27	11	1	-0.009	Lung
MAMSTR	19:49215999-49222978	none	27	9	1	0.007	Adrenal Gland
LILRA3	19:54799854-54809952	none	26	23	1	-0.016	Colon Sigmoid
SPATA25	20:44515128-44516274	31	27	5	1	0.013	Adipose Visceral Omentum
NEURL2	20:44517264-44517526	31	27	7	1	0.022	Adipose Visceral Omentum
PLTP	20:44527460-44540794	31	27	23	10	0.017 (0.006)	Adipose Visceral Omentum
SLC12A5	20:44651569-44688784	31	20	10	9	-0.011 (0.014)	Lung
PICK1	22:38452318-38471708	34	27	14	1	-0.015	Colon Sigmoid
BAIAP2L2	22:38480896-38506677	34	27	7	2	-0.023 (0.01)	Esophagus Mucosa
CBY1	22:39052645-39069859	34	27	16	1	-0.009	Liver

* Locus number according to Fritsche et al. (2016) [18]; ** Number of tissues in which gene expression is genetically regulated and imputable according to PredictDB and Gamazon et al. (2015) [53]; *** Tissue which showed the highest absolute beta

Selbstständigkeitserklärung

Ich, Tobias Strunz geboren am 05.06.1991 in Marktredwitz, erkläre hiermit, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe.

Die aus anderen Quellen direkt oder indirekt übernommenen Daten und Konzepte sind unter Angabe der Quelle gekennzeichnet. Insbesondere habe ich nicht die entgeltliche Hilfe von Vermittlungs- bzw. Beratungsdiensten (Promotionsberater oder andere Personen) in Anspruch genommen.

Die Arbeit wurde bisher weder im In- noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt.

Regensburg, 12.12.2020 Tobias Strunz