DEPARTMENT OF LIFE SCIENCES

FACULTY OF SCIENCES AND TECHNOLOGY
UNIVERSITY OF COIMBRA

Searching for Diabetic Retinopathy Genetic Markers by Whole Exome Sequencing

Thesis submitted to the University of Coimbra as requirement for the attribution of the degree of Master in Biochemistry conducted under the supervision of Dr. Conceição Egas, Assistant Researcher of CNC and co-supervision of Dr. António Portugal, Assistant Professor of the Life Sciences Department, University of Coimbra.

Cristina Maria da Costa e Silva Barroso

2015

Acknowledgements

Com a entrega deste trabalho atingi mais uma etapa!… Outras virão certamente, mas esta já foi conquistada! Para isso, foi fundamental a presença e apoio, expressão de amizade sincera, de muitas pessoas, algumas das quais não posso deixar de reconhecer neste pequeno trecho que agora escrevo!

À Dra. Conceição Egas, que aceitou o desafio de me orientar tão prontamente. Pelo apoio e cuidado mas também pela exigência e rigor que fez com que eu me quisesse superar cada vez mais.

Ao Dr. António Portugal pela confiança absoluta, disponibilidade e prontidão, motivador de minha tranquilidade neste processo.

À minha querida amiga e colega Maria José Simões, pela ajuda e apoio na componente estatística (e não só). Pelas conversas, pelo incentivo e pelas ideias… pela amizade e carinho…

Ao meu colega e companheiro de trabalho Diogo Pinho, por toda a ajuda, jovialidade, força e entusiasmo constante. Pela ajuda incalculável, na parte laboratorial deste trabalho e pelas muitas horas que ainda teremos pela frente…

Ao Hugo Froufe, pela paciência e disponibilidade e por toda a dedicação ao estudo da componente bioinformática.

Aos Susana Carmona e Felipe Santos, pelas pequenas ajudas no dia-a-dia, pelas gargalhadas e pelos almoços.

À Catarina João, Ana Nobre e Isaura Simões, que em tempos diferentes foram incentivadoras da "loucura" que me propunha fazer mas principalmente pelo interesse constante em perceber se eu estava bem e se estava a conseguir… de modo muito especial, à Catarina, pelas horas infindáveis no trabalho!

À minha queridíssima amiga Margarida Catarino!... que palavras para te agradecer o carinho? Não as há! Desde as flores ao brilho no olhar!... Bem hajas, amiga!

Aos meus pais pela vida gerada e a que dão por mim diariamente! Aos meus sogros, pela vida gerada na minha "outra metade"… aos quatro pela ajuda, sempre!

Às minhas irmãs, pelo exemplo. À minha Sarita, peloo que com ela começou "estou a terminar o que contigo pude começar"… Muito obrigada!

Aos meus filhos, Miguel, José, Maria Beatriz, João, Pedro e Tiago, meus maiores tesouros e minha maior conquista! Pela paciência e pela espera constante…Pelo amor, pelo carinho mas também pelas "birras"

Finalmente, a quem mais agradeço e a quem , a par de meus filhos, dedico este trabalho! Ao Nuno, meu companheiro de vida, meu maior e melhor amigo… a outra parte de mim! Por todas as horas de dedicação à nossa família, por todas as abnegações e por tudo fazeres para a minha tranquilidade. Sem ti, não teria conseguido

Os meus sinceros e dedicados agradecimentos a todos os que comigo privaram e trabalham.

A todos, um Bem hajam!

Remissive Index

Table Index

Figure Index

List of Abbreviations and Symbols

ACE	Angiotensin Converting Enzyme
ADAMTS2	ADAM Metallopeptidase with Thrombospondin Type1 Motif2
AGE	Advanced Glycated End Product
AGER	Advanced Glycated End Product Receptor
AKR1B1	Aldo-Keto Reductase family 1
APCDD1L	Adenomatosis Polyposis Coli Down-Regulated 1-Like
APOE	Apolipoprotein E
AR	Aldose Reductase
ASO-PCR	Allele-specific oligonucleotide PCR
BAM	Binary Alignment Map
BRB	blood-retinal barrier
CADD	Combined Annotation Dependent Depletion
CASZ1	Castor Zinc Finger1
DAG	Diacylglycerol
dbSNP	Database of Single Nucleotide Polymorphisms
DME	Diabetic macular edema
DMSO	Dimethyl sulfoxide
DMXL2	Dmx-Like 2
DNA/ADN	Deoxyribonucleic acid/ácido desoxirribonucleico
DNASE1L2	Deoxyribonuclease I-Like 2
dNTP	deoxyribonucleotide triphosphate
DR/RD	Diabetic retinopathy/ Retinopatia Diabética
E2F8	E2F Transcription Factor 8
ECM	Extracelular matrix
EGF	Epidermal Growth Factor
EGFR	Epidermal Growth Factor Receptor
EP300	E1A Binding Protein P300
EPACTS	Efficient and Parallelizable Association Container Toolbox
ESP	Exome Sequencing Project
ETDR	Early Treatment Diabetic Retinopathy
FLAGS	FrequentLy mutAted GeneS
FVM	fibrovascular membrane
GERP	Genomic Evolutionary Rate Profiling
GPR142	G Protein-Coupled Receptor 142
GSTM3	Glutathione S-Transferase Mu 3
GWAS	Genome-wide association studies
HIF1A	Hypoxia Inducible Factor 1
HPRD	Human Protein Reference Database
HWE	Hardy-Weinberg equilibrium
ICAM-1	Intercellular Adhesion Molecule 1
IGV	Integrative Genomics Viewer
IOP1	iron-only hydrogenase-like 1
ISFET	ion-sensitive field-effect transistor
ISP	Ion Sphere Particle

ITGA1	Integrin, Alpha 1
KEGG	Kyoto Encyclopedia of Genes and Genomes
MAF	Minor Allele Frequency
MAML3	MasterMind Like3
MAP	mitogen-activated protein/microtubule-associated protein
MMP1	Matrix Metallopeptidase 1
MNP	multi-nucleotide polymorphisms
MTHFR	Methylenetetrahydrofolate Reductase (NAD(P)H)
NARFL	Nuclear Prelamin A Recognition Factor-Like
NF-KB	NF-kB Transcription Factors
NGS	Next-Generation Sequencing
NICD	Notch IntraCellular Domains
NPDR	Non-proliferative diabetic retinopathy
OR	odds ratios
PDR	Proliferative diabetic retinopathy
PGM	Personal genome Machine
PKC	Protein Kinase C
PLXDC1	Plexin Domain Containing 1
PolyPhen	Polymorphism Phenotyping
PPARGC1A	Peroxisome Proliferator-Activated Receptor Gamma, Coactivator 1 Alpha
PRKCQ	Protein Kinase C, Theta
PXLDC1	Plexin domain-containing 1
Rbcn-3	Rabconnectin3
ROS	Reactive oxygen species
rs	reference SNP
S100A14	S100 Calcium Binding Protein A14
SD	Standard Deviation
SIFT	Sorting Intolerant From Tolerant
SLC2A1	Solute Carrier Family 2 (Facilitated Glucose Transporter), Member 1
SNPs	Single Nucleotide Polymorphisms
SOLiD	Sequencing by Oligo Ligation Detection
SPARC	protein acidic and rich in cysteine
STAB1	Stabilin 1
T2D/DT2	Type 2 Diabetes/Diabetes Tipo 2
TEM7	Tumor Endothelial Marker 7
TGF-β1	Transforming Growth Factor, Beta 1
TNFSF12	Tumor Necrosis Factor (Ligand) Superfamily, Member 12
TNF-α	Tumor Necrosis Factor Alpha
TVC	Torrent variant caller
TWEAK/Fn-14	TNF-like weak inducer of apoptosis/ fibroblast growth factor inducible-14
VCF	Variant call file
VEGFA	Vascular Endothelial Growth Factor A
VEP	Variant Effect Predictor
WES	Whole Exome Sequencing

Abstract

Type 2 Diabetes (T2D) is a debilitating complex disease that affects approximately 8% of the worldwide population. Diabetic retinopathy, the major microvascular complication of this disease is one of the leading causes of adult blindness in T2D patients. Knowledge of the genetic basis underlying this vascular diabetic complication will help understand the pathobiology and ameliorate the standard means of diagnosis, treatment and patient management. We proposed and pursued with this project and with the use of whole exome analysis and next-generation sequencing techniques to identify candidate genetic markers that predispose to the onset of Diabetic Retinopathy. Our study population was a group of 40 patients selected from the Type 2 Diabetes population of diagnosed Portuguese patients, with associated complications, from the Endocrinology Unit of the Hospital Center of the University of Coimbra. The workflow was, exome library preparation and parallel sequencing by Ion Torrent technology. The search for candidate rare variant accumulating genes and common variants was performed by differentiated bioinformatics and statistical approaches. Eleven candidate common variants (rs1035798, rs62357156, rs7125062, rs80067372, rs4434138, rs4234633, rs10794640, rs9907595, rs2296123, rs7483 and rs4698803) in genes AGER, ITGA1, MMP1, TNFSF12, STAB1, NARFL, PLXDC1, PRKCQ, GSTM3 and EGF and 2 rare variant accumulating genes, E2F8 and DMXL2 were considered biologically relevant. These variants were localized in genes involved in mechanisms and pathways related to Diabetic Retinopathy pathogenesis such as Advanced Glycated End (AGE) products trafficking and signaling pathway, fibrovascular membrane formation, EGF-VEGF signaling pathway, VEGFA-dependent angiogenesis, vascular assembly and morphogenesis and the Notch signalling pathway. Validation of the technology, variants and allele frequencies was performed by other sequencing and genotyping methods. This study highlighted several new candidate biomarkers that need to be validated in a larger population before association to Diabetic Retinopathy.

Keywords: Diabetic Retinopathy, Exome Sequencing, Genetic Markers, Rare Variants, Common Variants

This work was held under the COMPETE program and the DoIT-Development and Operation of Translational Research Project, ref: FCOMP-01-0202-FEDER-013853.

Resumo

A Diabetes Tipo 2 (DT2) é uma doença complexa e debilitadora que afeta aproximadamente 8% da população mundial. A Retinopatia Diabética, a maior complicação microvascular desta doença, é uma das principais causas de cegueira em adultos diagnosticados com Diabetes Tipo 2. O conhecimento da base genética subjacente a esta complicação vascular permitirá uma melhor compreensão da doença e consequente aplicação de meios efetivos de diagnóstico, tratamento e gestão de doentes. Foi objetivo deste projeto e com a utilização de sequenciação de exomas com técnicas de última geração, identificar marcadores genéticos candidatos que predispõem para o desenvolvimento da Retinopatia Diabética. A nossa população de estudo de 40 indivíduos foi selecionada a partir de um grupo de doentes Portugueses diagnosticados com Diabetes do Tipo 2, com diferentes complicações associadas, da Unidade de Endocrinologia do Centro Hospitalar da Universidade de Coimbra. A metodologia utilizada foi extração de ADN, seguida de preparação de bibliotecas de exomas e sequenciação massiva paralela pela tecnologia Ion Torrent. A pesquisa de genes que acumulam variantes raras e de variantes comuns foi realizada por diferentes abordagens bioinformáticas e estatísticas. Dois genes que acumulam variantes raras, E2F8 e DMXL2, e onze variantes comuns (rs1035798, rs62357156, rs7125062, rs80067372, rs4434138, rs4234633, rs10794640, rs9907595, rs2296123, rs7483 e rs4698803) nos genes: AGER, ITGA1, MMP1, TNFSF12, STAB1, NARFL, PLXDC1, PRKCQ, GSTM3 e EGF foram considerados biologicamente relevantes. As variantes encontradas localizam-se em genes envolvidos em mecanismos e vias relacionadas com a patogénese da Retinopatia Diabética, tais como a via de sinalização e tráfego de produtos finais de glicação avançados, formação de membrana fibrovascular, via de sinalização EGF-VEGF, angiogénese dependente de VEGFα, assemblagem e morfogénese vascular e a via de sinalização Notch. A validação da tecnologia, variantes e frequências alélicas foi realizada por diferentes métodos de sequenciação e genotipagem. Este estudo destacou alguns biomarcadores candidatos, que carecem de validação numa população maior antes de poderem ser associados à retinopatia diabética.

Palavras Chave: Retinopatia Diabética, Sequenciação de Exomas, Marcadores Genéticos, Variantes Raras, Variantes Comuns

Este trabalho foi efetuado no âmbito do programa COMPETE e do Projeto de Investigação de Desenvolvimento e Operação Translacional –DoIt, ref: FCOMP-01-0202-FEDER-013853.

Chapter 1

Introduction and Literature review

1. Introduction and Literature review

1.1 Introduction

Understanding the genetics of common, complex and debilitating disorders continues to be a challenge and although the recognition of its importance, genetic analysis are difficult due to the complex interaction among multiple susceptibility genes and between genetic and environmental factors (Doria, 2010; van Hoek *et al.,* 2008; Herder & Roden, 2011; Lyssenko & Laakso, 2013; Lyssenko *et al.,* 2005 and Frazer *et al.,* 2009).

Research of the genetic causes of Type 2 Diabetes and its microvascular complication, Diabetic Retinopathy (DR), has been and remains a demanding area of interest (Lohmueller *et al.,* 2013). Some genes have been implicated in the aetiology of the disorder but replication of these findings has been difficult and genetic studies have revealed diverse results (Weedon *et al.,* 2006; Lohmueller *et al.,* 2013; Kuo *et al.,* 2014).

Over the past years, new techniques, such as whole genome analysis through Next-generation sequencing and the availability of large population-based DNA banks have accelerated research and knowledge of the genetics of common diseases. Whole Exome Sequencing being a comprehensive, cost efficient and rapid method for analysing and studying the coding regions of an individual is a useful tool in complex traits genetics (Bonnefond *et al.,* 2010; Johansson *et al.,* 2012; Bamshad *et al.,* 2011; Albrechtsen *et al.,* 2013). It allows for an extensive genetic search and may unravel completely novel genes or variants in individuals with no genetic defect in the known diabetic retinopathy genes *(*Johansson *et al.,* 2012*).*

The ultimate goal of this line of research, as with nearly all research in the genetics of any complex disease, is to improve the understanding of the pathophysiology and disease aetiology so that more effective means of diagnosis, treatment and prevention can be developed (Manolio, 2009; Bamshad *et al.,* 2011).

1.2. Literature Review

1.2.1 Fundamentals of Type 2 Diabetes

Type 2 Diabetes (T2D) is a slow and progressive endocrine disorder (Kahn, 1994) characterized by impaired insulin secretion and variable degrees of insulin resistance (Figure 1) that leads to an elevation in blood glucose, hyperglycaemia, which in time results in debilitating complications and damage to various organs (Anomalies & Brief, 2003; Olokoba *et al.,* 2012).

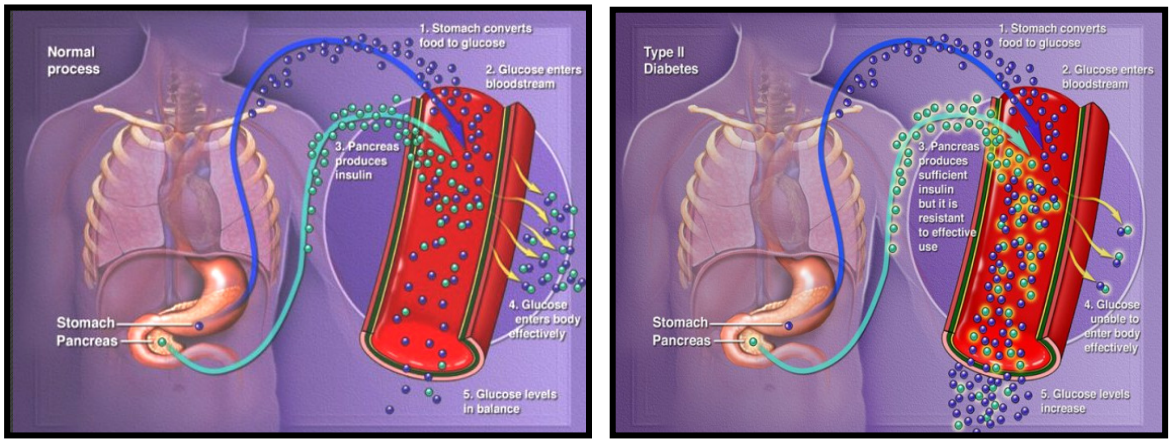


Figure 1. Differences between the normal glucose absorption process in a healthy individual (A) and the insulin resistance condition in a T2D diagnosed individual (B).
In the exemplified case (B), the stomach converts food to glucose which enters the bloodstream. The pancreas produces sufficient insulin but it is resistant to effective use. Thus glucose absorption is inefficient and glucose blood levels are increased. Adapted from http://www.topnews.in/health/scientists-illuminate-cell-pathway-key-insulin-resistance-type-2-diabetes-211013.

T2D is the most common form of diabetes worldwide accounting for 90-95% of all cases (Massi-Benedetti, 2002; American Diabetes Association, 2014; Olokoba *et al.,* 2012). The progression from having a genetic predisposition to T2D and the development of an elevated blood sugar and hence the disease is affected by environmental factors such as being overweight, physical inactivity, age, diet, illness, pregnancy, medication and on how strong the gene traits are causing the disorder in that individual (Kahn, 1994; Reis & Velho, 2002; Olokoba *et al.,* 2012).

1.2.1.1 Epidemiology

This disease has reached epidemic proportions worldwide and although has always been associated to older ages, its prevalence has been increasing in children and adolescents due to the widespread of obesity, particularly central obesity, and unhealthy lifestyles. (Anomalies & Brief, 2003; Bloomgarden, 2004).

Globally in 2013, it was estimated that almost 382 million adults (20-79 years) suffered from T2D for a prevalence of 8.3%, whereas in Europe, almost 52 million adults suffered from this complex disease, with a prevalence of 7.9% (Martinez, 2013) (Figure 2). Of these, 33% (17.2 million) had not been diagnosed and were at a higher risk of developing harmful and costly complications. Half of the individuals that clinically present T2D already present signs of the associated complications (Mathers & Loncar, 2006) and the identification of these individuals is of great importance and interest for health care providers and investigators (Lyssenko & Laakso, 2013).

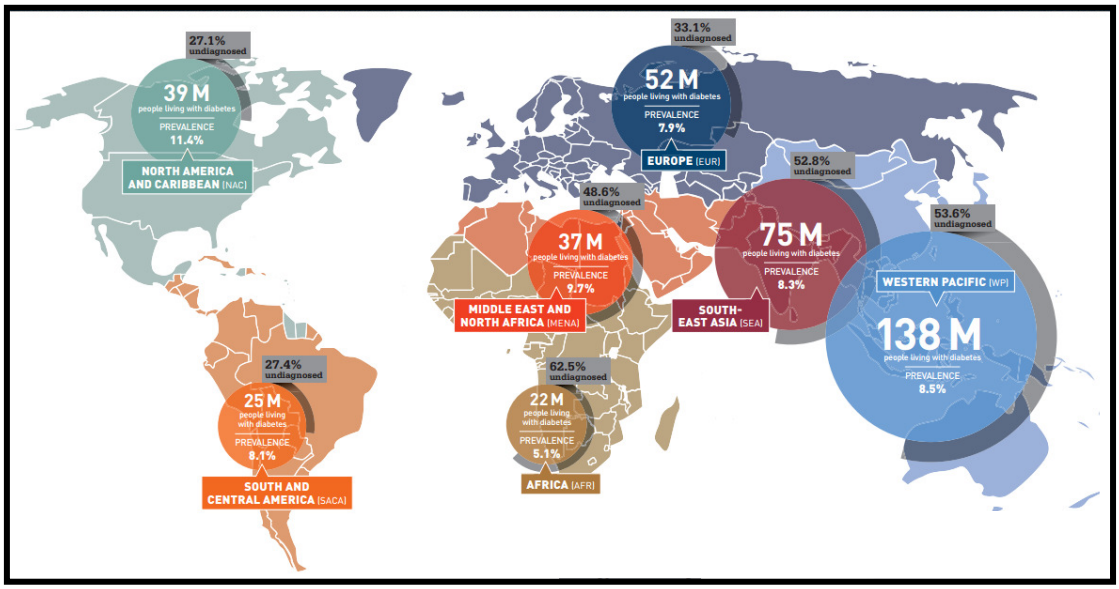


Figure 2. Worldwide Prevalence of Diabetes Type 2.
The percentage of undiagnosed individuals is highest among poorer countries such as Middle East and North Africa, South East Asia, Western Pacific and Africa. Adapted from IDF Diabetes Atlas, Sixth Edition, 2014 Update (Martinez, 2013).

Unless action is taken it is predicted that, by the year 2035, there will be approximately 69 million T2D adults in Europe, with a prevalence of 10.3%. (Martinez, 2013). It is a major health problem all over the world with its greatest impact being in newly industrialized and developing nations and minority groups in developed

countries (Massi-Benedetti 2002; Olokoba *et al.,* 2012). The socio-economic and public health impact of the disease is constantly increasing and has effects on the work force, time taken for treatment, premature morbidity and mortality (Massi-Benedetti, 2002). In 2014, Diabetes caused 537,000 deaths in Europe and 23% of these deaths were in people under the age of 60 (Martinez, 2013).

In 2013 the estimated prevalence of Diabetes in the Portuguese population with ages between 20 and 79 years, corresponding to 7.8 million individuals, was 13,0%, this is, more than 1 million Portuguese people, in this age group, had Diabetes (OND, 2014). The comparative prevalence in Europe was 9.57% (Figure 3).

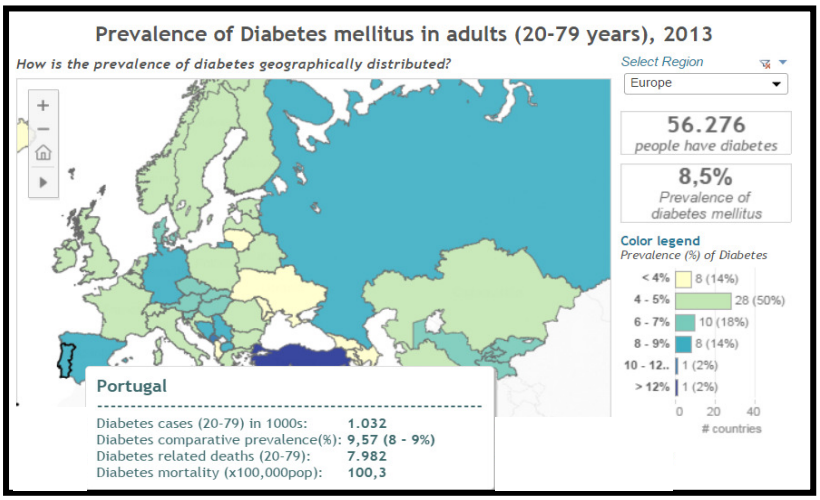


Figure 3. European Prevalence of Diabetes Type 2 in adults (20-79 years).
The number of cases in thousands, comparative prevalence (%), related deaths and mortality (x100 000population) in Portugal is highlighted. Adapted from Prevalence of Diabetes in the World, 2013 (Martinez, 2013). http://healthintelligence.drupalgardens.com/content/prevalence-diabetes-world-2013

Considering only the Portuguese population diagnosed with Diabetes between 20 and 79 years, in 2013, the national health expenses were 962 million Euros, for all individuals. This represented 1% of the Gross National Product and 10% of all health expenses. It is estimated that the national prevalence of Diabetes, in 2035, in Portugal, will be 15.8%. (Figure 4) which will naturally increase the health costs associated.

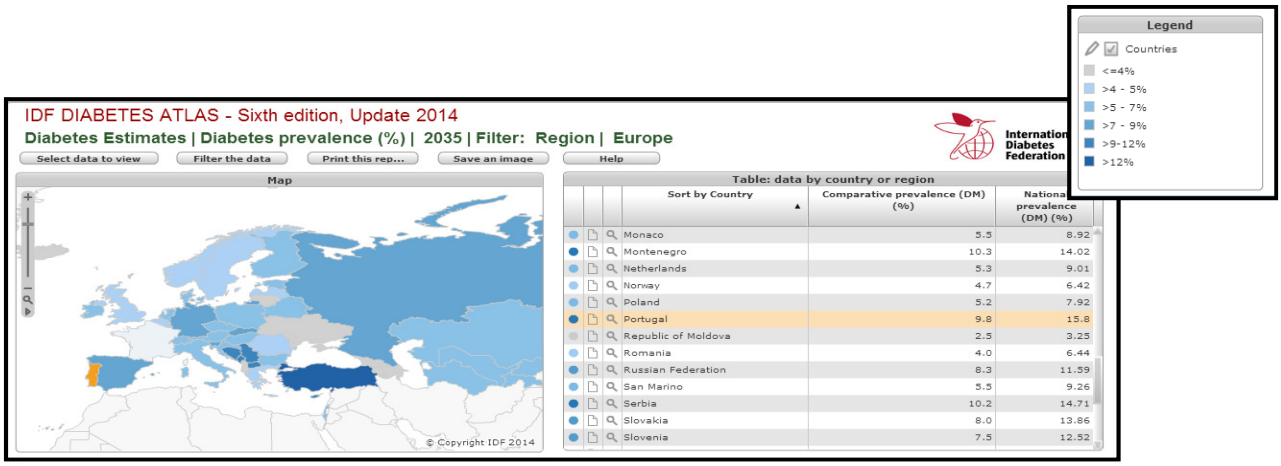


Figure 4. Estimates for the prevalence of Diabetes Type 2 in adults (20-79 years), in Europe for 2035. The comparative prevalence has been calculated by assuming that every country and region has the same age profile (world population) reducing the effect of age differences between countries and regions. Adapted from IDF Diabetes Atlas, Sixth Edition, 2014 Update (Martinez, 2013). http://www.idf.org/atlasmap/atlasmap?indicator=i1&date=2014)

T2D is strongly inherited and proof of the genetic factor cause is evidenced by the high prevalence of the disease in certain racial and ethnic groups such as Hispanics, Asians, Pacific Islanders, African and Native Americans, Southeast Asians and the higher risk inherent to belonging to a family with a history of T2D (Anomalies & Brief, 2003; American Diabetes Association, 2014; Reis & Velho, 2002;Barroso *et al.,* 2003).

1.2.1.2 Pathophysiology of Type 2 Diabetes

Pathogenesis of this complex disease is still incompletely understood. Although insulin resistance is characteristic in T2D individuals, evidence also exists for β-cell dysfunction. High glucose levels may desensitize β-cells (glucose toxicity) and impair insulin secretion. It is however unlikely that glucotoxicity acts alone, and the negative contribution of saturated fatty acids, lipoproteins, leptin and circulating and locally produced cytokines will further burn out the β-cells (Kahn, 1994).

A long time has passed since the idea that insulin solely binds to its receptor leading to the stimulation of glucose transport. In the past decades cellular and molecular biology techniques have greatly enhanced the understanding of the insulin process. Many proteins involved in the insulin action cascade have been identified and cloned at the molecular level giving a new insight into the fascinating and complex process of insulin action in the cell (Figure 5).

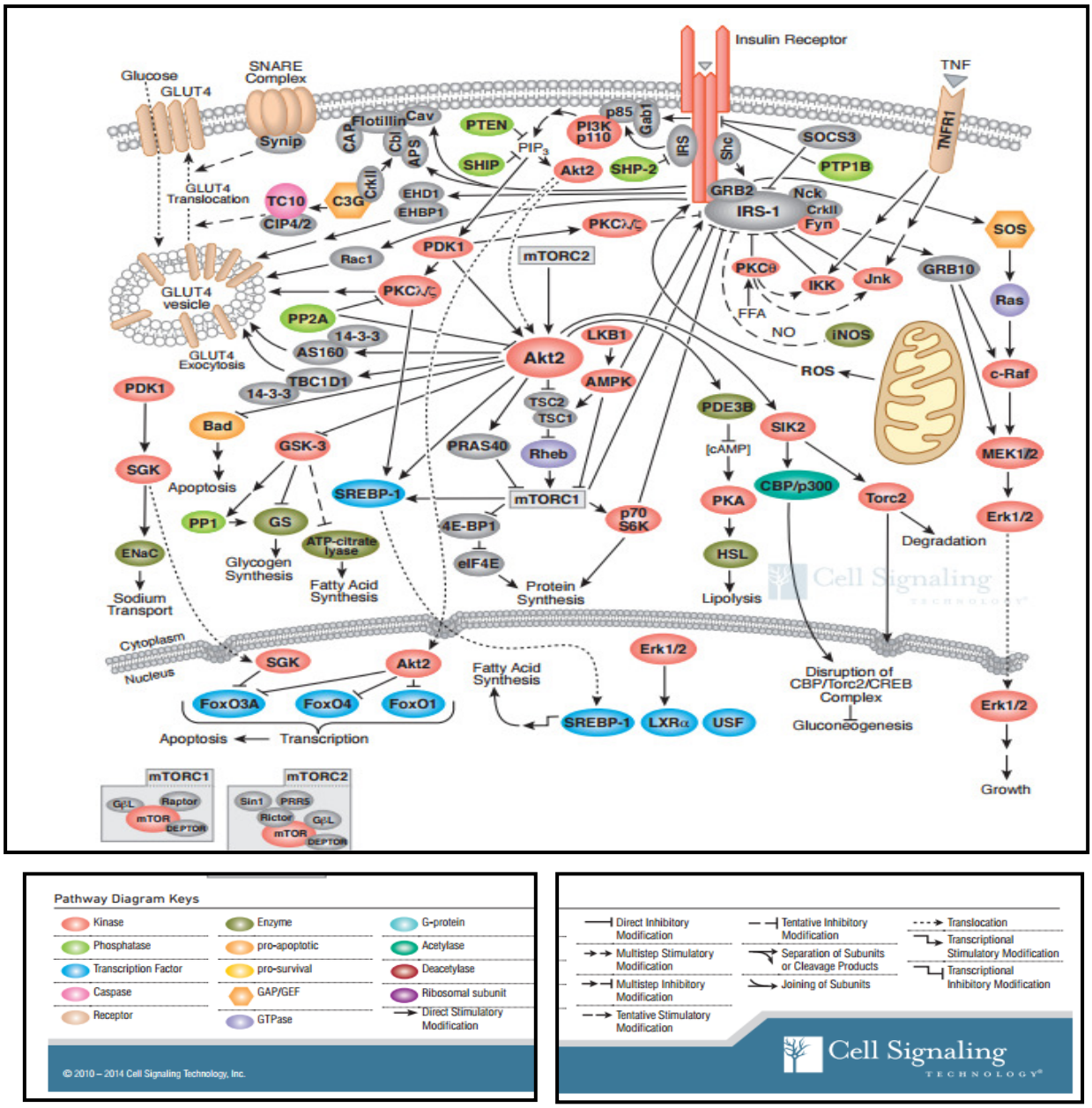


Figure 5. Insulin Receptor Signalling Cascade.
Insulin is the major hormone controlling critical energy functions such as glucose and lipid metabolism. Insulin activates the insulin receptor tyrosine kinase (IR), which phosphorylates and recruits different substrate adaptors such as the IRS family of proteins. Tyrosine phosphorylated IRS then displays binding sites for numerous signaling partners. Among them, PI3K has a major role in insulin function, mainly via the activation of the Akt/PKB and the PKCζ cascades. Activated Akt induces glycogen synthesis through inhibition of GSK-3; protein synthesis via mTOR and downstream elements; and cell survival through inhibition of several pro-apoptotic agents (Bad, FoxO transcription factors, GSK-3, and MST1). Insulin signaling also has growth and mitogenic effects, which are mostly mediated by the Akt cascade as well as by activation of the Ras/MAPK pathway. The insulin signaling pathway inhibits autophagy via the ULK1 kinase, which is inhibited by Akt and mTORC1, and activated by AMPK. Insulin stimulates glucose uptake in muscle and adipocytes via translocation of GLUT4 vesicles to the plasma membrane. GLUT4 translocation involves the PI3K/Akt pathway and IR-mediated phosphorylation of CAP, and formation of the CAP:CBL:CRKII complex. In addition, insulin signaling inhibits gluconeogenesis in the liver, through disruption of CREB/CBP/mTORC2 binding. Insulin signaling induces fatty acid and cholesterol synthesis via the regulation of SREBP transcription factors. Insulin signaling also promotes fatty acid synthesis through activation of USF1 and LXR. A negative feedback signal emanating from Akt/PKB, PKCζ, p70 S6K, and the MAPK cascades results in serine phosphorylation and inactivation of IRS signaling.Adapted from Cell Signaling Technology (**http://www.cellsignal.com/contents/science-cst-pathways-cellular-metabolism/insulin-receptor-signaling-pathway/pathways-irs?Ntt=insulin+receptor&fromPage=search**)

For simplicity, we may think of insulin action occurring at three levels or stages (Kahn, 1994) (Figure 6). Level 1 is composed of the initial events related to receptor tyrosine kinase activity. This includes the insulin receptor itself, the insulin receptor substrate, and the molecules that interact with this substrate. Level 2 refers to the cascade of serine phosphorylation and dephosphorylation reactions centered on the enzyme MAP (mitogen-activated protein/microtubule-associated protein) kinase. Level 3 is the final biological effectors of the insulin cascade. This includes the glucose transport molecules themselves, which reside in an intracellular pool and are translocated to the plasma membrane following insulin stimulation, the enzymes for glycogen and lipid synthesis and the proteins involved in insulin action on gene expression and cell growth (Kahn, 1994).

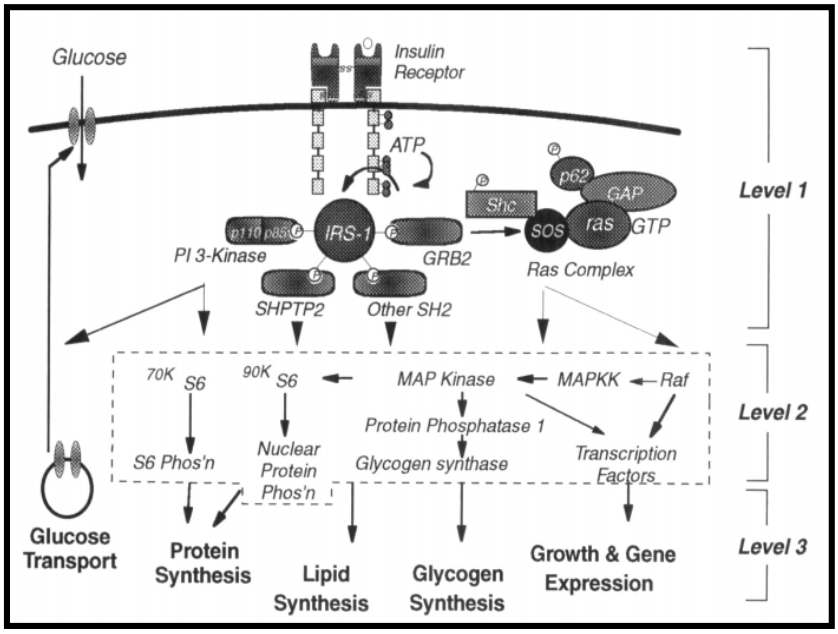


Figure 6. The Three levels of insulin action: molecules and pathways involved in insulin signaling (Kahn, 1994).

1.2.1.3 Associated Complications

It is predicted that by 2030, Diabetes will be the 7th leading cause of worldwide death *(*Mathers & Loncar, 2006*)*, primarily due to complications associated with end organ damage including heart disease, stroke, blindness, kidney disease, peripheral neuropathy and amputations (Anomalies & Brief, 2003; Massi-Benedetti, 2002; American Diabetes Association, 2014).

Poorly controlled hyperglycemia leads to multiple vascular complications that affect small (microvascular) and/or large (macrovascular) vessels (Massi-Benedetti 2002). These vascular complications involve several important organs such as eyes, kidneys and the cardiovascular system *(*Tang *et al.,* 2013) (Figure 7). The mechanisms by which vascular disease develops include: 1) glycosylation of serum and tissue proteins with formation of advanced glycation end products; 2) superoxide production; 3) activation of protein kinase C, a signaling molecule that increases vascular permeability and causes endothelial dysfunction; 4) accelerated hexosamine biosynthetic and polyol pathways leading to sorbitol accumulation within tissues; 5) hypertension and dyslipidemias (abnormal accumulation of lipids in the blood); 6) arterial microthromboses; pro-inflammatory and pro-thrombotic effects and 7) hyperinsulinemia that impairs vascular autoregulation (Massi-Benedetti 2002). Immune dysfunction is another major complication and develops from the direct effects of hyperglycemia on cellular immunity (Kishore, 2013).

The microvascular diseases that may appear are Diabetic Retinopathy, the most common cause of adult blindness characterized initially by retinal capillary microaneurysms and later by macular edema and neovascularisation; Diabetic Nephropathy, a leading cause of chronic renal failure characterized by the thickening of the glomerular basement membrane, mesangial expansion and glomerular sclerosis and Diabetic Neuropathy, resulting from nerve ischemia by the direct effects of hyperglycemia on neurons and intracellular metabolic changes that impair nerve function (Massi-Benedetti 2002).

Large vessel atherosclerosis, a macrovascular disease, manifests by myocardial infarctions, transient ischemic attacks, strokes and peripheral arterial disease (Massi-Benedetti 2002; Kishore, 2013).

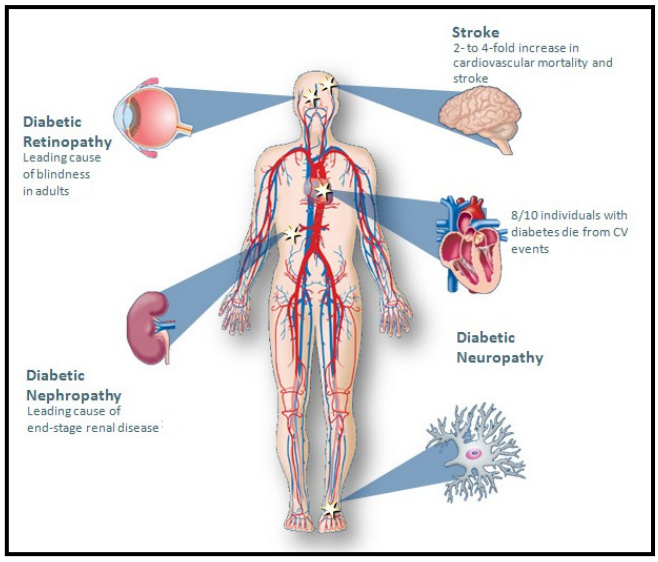


Figure 7. Diabetes complications.
Conditions or pathological processes associated with the disease of diabetes mellitus due to the impaired control of blood glucose level in diabetic patients, pathological processes develop in numerous tissues and organs including the eye, the kidney, the blood vessels and the nerve tissue.
Adapted from http://lookfordiagnosis.com/mesh_info.php?term=Diabetes+Complications&lang=1

Diabetic foot complications such as skin change, ulceration, infection and gangrene are common in T2D individuals and attributable to vascular disease, neuropathy and immunosuppression (Massi-Benedetti 2002; Kishore, 2013).

As for the psychosocial impacts of T2D, depression is a common problem and may precede the development of the disease. Diabetic individuals with untreated depression have poorer glucose control, increased risk of complications and higher health care costs (Anomalies & Brief, 2003). Adequate prevention and management of T2D may require significant behavioural change which can reveal frustrating for patients and health care providers. Yet, the psychosocial and behavioural impacts of knowing the genetic risk for T2D has not been assessed (Anomalies & Brief, 2003).

1.2.2 Fundamentals of Diabetic Retinopathy

Published evidence indicates that major risk factors, such as long term diabetes, poor control of blood glucose and elevated blood pressure are responsible for the onset and progression of diabetic complications (Tang *et al.,* 2013; Kuo *et al.,* 2014). Nonetheless these clinical features have not been consistently identified in different studies and patients cannot be stratified with respect to their risk of developing a microvascular complication based only upon clinical or procedural risk factors. There is

now evidence that genetic factors may explain part of the excessive risk independently of conventional clinical variables (Tang *et al.,* 2013).

Diabetic Retinopathy is a common complications of diabetes, in which the retina becomes progressively damaged, leading to vision loss and blindness (Nawaz, 2010) and the duration of diabetes is probably the strongest predictor for development and progression of retinopathy (Fong *et al.,* 2004, Kuo *et al.,* 2014).

1.2.2.1 Anatomy-physiology of the Human Eye

The retina is a thin layer of light-sensitive tissue that lines the back of the eye. Light rays are focused onto the retina, where they are transmitted to the brain and interpreted as images. The macula is a small area at the centre of the retina that is responsible for pinpoint vision. The surrounding part of the retina, called the peripheral retina, is responsible for peripheral vision (Figure 8).

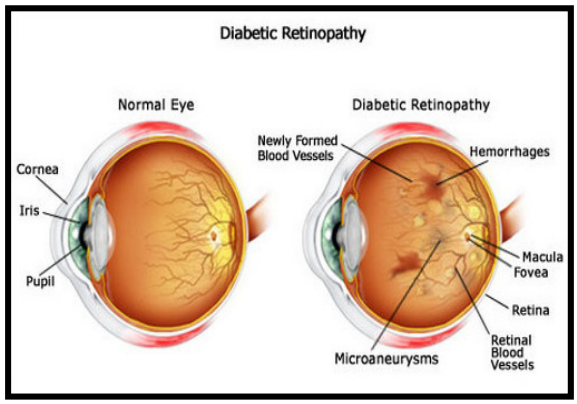


Figure 8. Normal vs Diabetic Retinopathy Eye.
Diabetic Retinopathy affects the blood vessels in the retina, leading to new abnormal blood vessel growth, leakage and bleeding. Adapted from http://www.diabetest1ireland.com/eye-damage.html

Diabetic retinopathy thus occurs when blood vessels in the retina swell and leak fluid or even close off completely. In aggravated cases abnormal new blood vessels grow on the surface of the retina. As the disease progresses various symptoms appear: spots, dots or dark strings floating, floaters (Figure 9); blurred vision; blank or dark areas in the vision field; poor night vision; colours appear washed and vision loss.

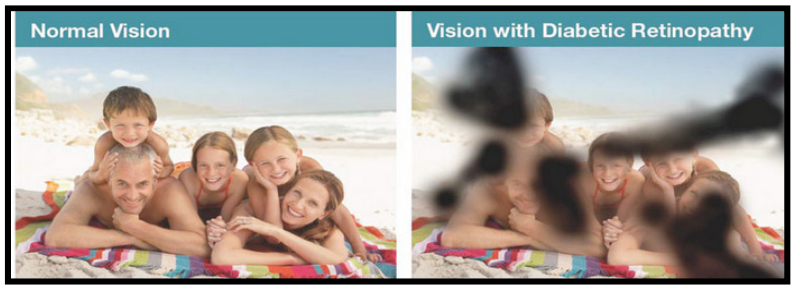


Figure 9. Diabetic Retinopathy Symptoms.
Presence of floaters in the vision field. Adapted from http://eyedoctorwichita.com/eye---vision-problems/diabetic-retinopathy.html

1.2.2.2 Clinical Classification of Diabetic Retinopathy

Diabetic Retinopathy progresses from mild non-proliferative abnormalities, characterized by increased vascular permeability, to moderate and severe non-proliferative diabetic retinopathy (NPDR), characterized by vascular closure. Proliferative diabetic retinopathy (PDR) is characterized by the growth of new blood vessels on the retina and posterior surface of the vitreous. Macular edema, the retinal thickening from leaky blood vessels, can develop at all stages of retinopathy *(Fong et al., 2004)* (Figure 10). Although proliferative retinopathy may lead to loss of vision and blindness, diabetic macular edema (DME) is the main cause of central vision loss. (Ozturk *et al.,* 2009)

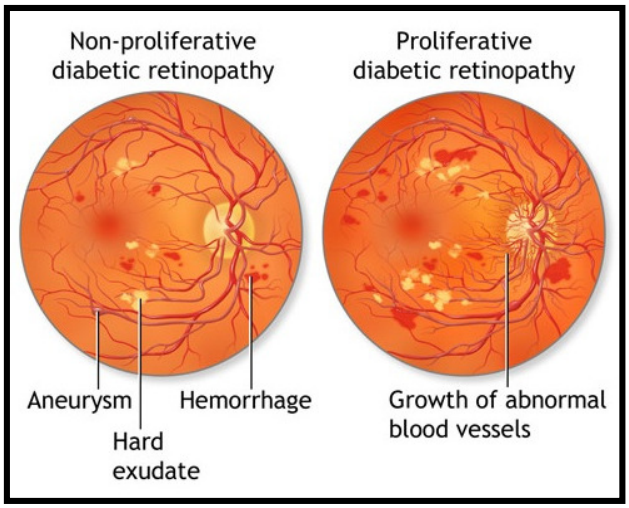


Figure 10. Non-proliferative vs Proliferative Diabetic Retinopathy.
In non-proliferative DR aneurysms, hard exudates and hemorrhages may occur whereas in proliferative DR abnormal blood vessels grow in the area of the retina.
Adapted from https://www.gulfcoasteyecare.com/conditions/diabetic-disease-tampa-fl/

Proliferative Diabetic Retinopathy (PDR) is characterized by vasculopathy associated with abnormal angiogenesis and expansion of extracellular matrix (ECM) resulting in the outgrowth of FibroVascular Membranes (FVM) at the vitreoretinal interface. The formation of this fibrovascular tissue results in severe complications such as vitreous hemorrhage and traction retinal detachment (Abhary *et al.,* 2009; El-Asrar *et al.,* 2013). Angiogenesis is a multistep process requiring the degradation of the basement membranes and Extracellular Cellular Matrix (ECM), endothelial cell migration, endothelial cell proliferation and capillary tube formation.

Vision loss may result from several mechanisms. Central vision may be impaired by macular edema or capillary non-perfusion. New blood vessels of PDR and contraction of the accompanying fibrous tissue can distort the retina and lead to tractional retinal detachment, producing severe and often irreversible vision loss. In addition, the new blood vessels may bleed, adding to the further complication of pre-retinal or vitreous hemorrhage (Fong *et al.,* 2004, Ozturk *et al.,* 2009).

The assessment of Diabetic Retinopathy by a standardized stereoscopic photograph has been proposed to grade the complication and homogenize the phenotype classification. Researchers have been grading DR using the Early Treatment Diabetic Retinopathy Study (ETDRS) severity scale or a similar modification (Kuo *et al.,* 2014).

1.2.2.3 Pathophysiology of Diabetic Retinopathy

Advanced glycation end products (AGEs) are generated by non-enzymatic glycosylation of proteins or lipids after prolonged exposure to glucose (Tamura *et al.,* 2003). AGEs elicit a wide variety of cellular responses including induction of growth factors and cytokines, adhesion molecules activity, oxidant stress, and chemotaxis. These pro-inflammatory responses contribute to the development of pathologies associated with aging, diabetes mellitus, and Alzheimer's disease. (Hammes *et al.,* 2002)

Various studies have indicated that the common pathophysiologic mechanism linking chronic hyperglycemia to vascular pathology in diabetes is the mitochondrial overproduction of reactive oxygen species (ROS) which leads to the increased formation of AGEs, activation of protein kinase C (PKC), activation of the aldose reductase (AR), and deliberation of active nuclear factor kB (NF-kB), mechanisms that have been correlated with the pathogenesis of diabetic microangiopathy (Hammes *et al.,* 2002; van den Oever *et al.,* 2010) (Figure 11).

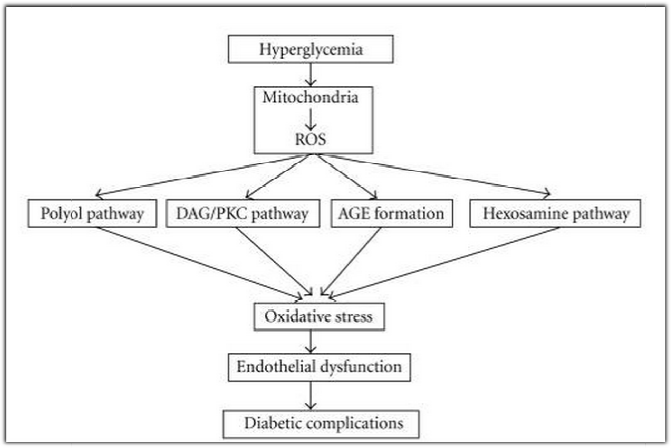


Figure 11. Mechanisms of hyperglycemia which are supposed to cause endothelial dysfunction.
Adapted from van den Oever et al., 2010

Microvascular mural cells, referred to as pericytes, provide vascular stability and control endothelial proliferation (Hammes *et al.,* 2002, Falcão *et al.,* 2010) (Figure 12). Pericyte loss, microaneurysms, vascular basement membrane thickening and acellular-occluded capillary formation are hallmarks of early changes in the retina of diabetic patients (Hammes *et al.,* 2002). With progressive vascular occlusions in the human diabetic eye, the retina responds with either a progressive increase of vascular permeability, leading to macula edema, or the formation of new vessels that finally proliferate into the vitreous (Hammes *et al.,* 2002). The cause of pericyte loss during early diabetic retinopathy is unclear but seems related to the pericytic accumulation of toxic products such as sorbitol or advanced glycation end products (AGEs) (Hammes *et al.,* 2002). Pericyte loss is considered a prerequisite of microaneurysm formation, possibly by local weakening and subsequent outpouching of the capillary wall. As pericytes also control endothelial cell proliferation, pericyte loss may be involved in

the pathogeneis of proliferative diabetic retinopathy (Ozturk *et al.,* 2009; Falcão *et al.,* 2010, Beltramo & Porta, 2013).

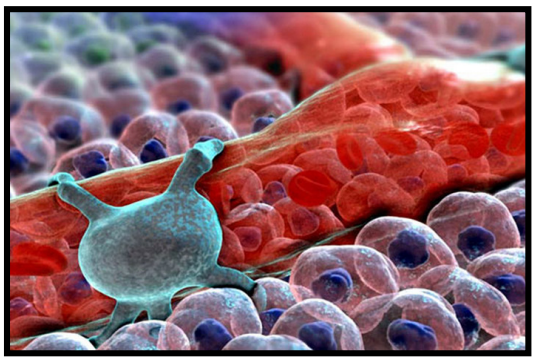


Figure 12. Pericyte on a capillary. Pericytes are contractile cells with a "spider-like" shape found on the outside of the small vasculature.
They play important roles in regulating capillary blood flow and are also responsible for maintaining permeability of the blood-brain barrier. Adapted from http://www.xvivo.net/illustration/pericyte-on-capillary/

The cross-talk between capillary cells as well as the role of capillary pericyte coverage in survival and repair of endothelial cells in the diabetic retina is of utmost importance. (Hammes *et al.,* 2002)

Diabetic maculopathy is characterized by hyperpermeability of retinal blood vessels and subsequent formation of macular edema and hard exudates. The increase in retinal vascular permeability occurs both diffusely and in focal regions (Sugimoto *et al.,* 2013). The blood-retinal barrier (BRB) isolates the retina from the bloodstream, establishing a favorable environment with the regulation of ionic balance, nutrient availability, and blockage of potentially toxic molecules that allows for optimal retinal function (Sugimoto *et al.,* 2013, Falcão *et al.,* 2010). Disruption of this barrier is an important feature of diabetic retinopathy) (Figure 13).

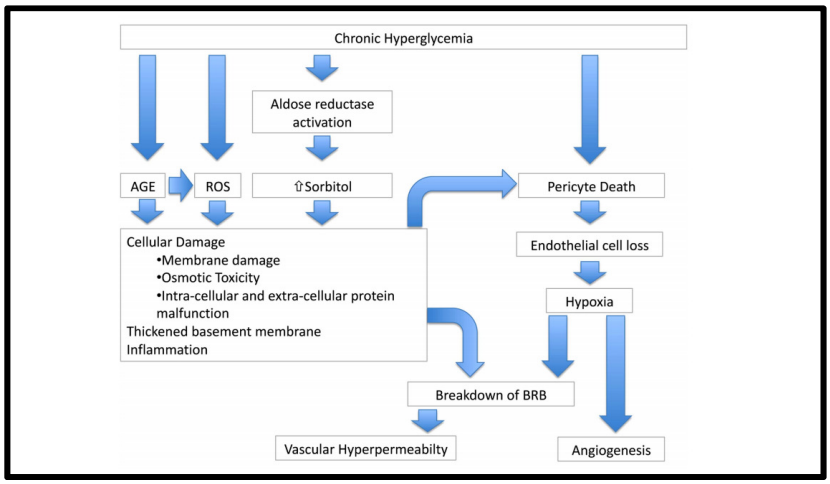


Figure 13. Pathogenesis of diabetic retinopathy. Schematic overview of the effects of chronic hyperglycemia and its implications in cellular damage.
AGE: Advanced Glycation End Products. ROS: Reactive Oxygen Species. BRB: Blood Retinal Barrier. Adapted from Falcão *et al.,* 2010

Diabetic Retinopathy also displays all microscopic signs of inflammation such as vasodilatation, altered flow, fluid exudation and leukocyte migration. Alterations in serum or vitreous levels of many inflammatory cytokines like IL-6, IL-8, IL-10, and VEGF also supports the role of inflammation in DR (Ozturk *et al.,* 2009)

1.2.2.4 Candidate Genes for Diabetic Retinopathy

Several candidate gene association studies have reported promising genes but few of them have been replicated, and the few positive findings show only weak genetic associations (Kuo *et al.,* 2014). However, several genes including *AGE*, *VEGF* and *AKR1B1* among others have been considered to be associated with the risk of developing Diabetic Retinopathy (Tang *et al.,* 2013).

It is established that AGE and its receptor AGER contribute to diabetic complications through a mechanism involving direct tissue damage (Tang *et al.,* 2013) and this effect seems linked to glycosylated hemoglobin levels (Goldin *et al.,* 2006). AKR1B1 is the first and rate limiting enzyme of the polyol pathway and has been linked to diabetes-specific tissue complications in some ethnic groups (Abhary *et al.,* 2009). Other candidate genes such as ACE, MTHFR, SLC2A1 and APOE have been found associated with Diabetic Retinopathy, although not reproducibly, which may be attributed to small sample sizes or methodological limitations (Warpeha & Chakravarthy, 2002;

Abhary *et al.,* 2009; Tang *et al.,* 2013). Vascular endothelial growth factor (VEGF) is linked to the neovascularization and vascular leakage process in proliferative retinopathy and various polymorphisms have been reported in different populations (Warpeha & Chakravarthy, 2002; Abhary *et al.,2009;* El-Asrar *et al.,* 2013; Tang *et al.,* 2013). It is the major angiogenic factor and has been established as a survival factor in retinal capillary endothelium. VEGF and pericytes have complementary roles in promoting endothelial cell survival.

1.2.2.5 Treatment and Management

Despite recent improvements in vitreous surgical techniques, panretinal photocoagulation, and antivascular endothelial growth factor drugs, the prognosis for patients with PDR is still poor, especially for those with advanced PDR at the proliferative stage. It is therefore necessary to develop better diagnosis and treatment techniques based on the exact pathogenesis of fibrovascular membrane formation (Yamaji *et al.,* 2008).

Considerable effort has been invested recently to develop agents that block the formation of new blood vessels. For example, bevacizumab, a selective VEGF inhibitor, was recently found to be effective in the regression of retinal and iris neovascularization secondary to PDR, but because of its cytostatic property, its effect may be limited to established vasculature. Therefore, it has become apparent that targeted destruction of the established vasculature is another avenue for therapeutic opportunities (Yamaji *et al.,* 2008).

1.2.3 Genetics of Complex Diseases

Determining the genetic basis of human diseases is one of the major research areas in medical science (Wang *et al.,* 2013), but despite significant progress of Genome-Wide Association Studies (GWAS) in the identification of a large number of new genetic loci that contribute to complex traits, only a small fraction of the observed heritability is explained by the confirmed, genome wide-significant, common variants

(Rabbani *et al.,* 2014; Manolio, 2009; Koeleman *et al.,* 2013; Frazer *et al.,* 2009; Tang *et al.,* 2013). This has forced reconsideration of the degree of genetic heterogeneity and the role of genetics in the pathogenesis of complex traits (Doria*, 2010;* Frazer *et al.,* 2009). Although advances in the knowledge of the genetic architecture of complex traits has grown over the past years we still have a limited understanding of the number of genetic variants that influence a trait, their allele frequencies, effect sizes and modes of interactions (Frazer *et al.,* 2009).

Recent studies demonstrated that the heritability estimation is improved by using all genome-wide Single Nucleotide Polymorphisms (SNPs) instead of using only the significant variants (Kiezun *et al.,* 2012; Lohmueller *et al.,* 2013; Yi, 2010). Common genetic variation account for a non-negligible but modest proportion of inherited risk, leading to the suggestion that low frequency and rare genetic variants may contribute substantially to the genetic burden underlying common and complex disease (Lohmueller *et al.,* 2013; Marth *et al.,* 2011; Morris *et al.,* 2012; Estrada *et al.,* 2014; Panoutsopoulou *et al.,* 2013; Marian, 2012) (Figure 14). The allelic architecture of complex traits is thus likely to be the combination of multiple common frequency and rare variants (Panoutsopoulou, 2013).

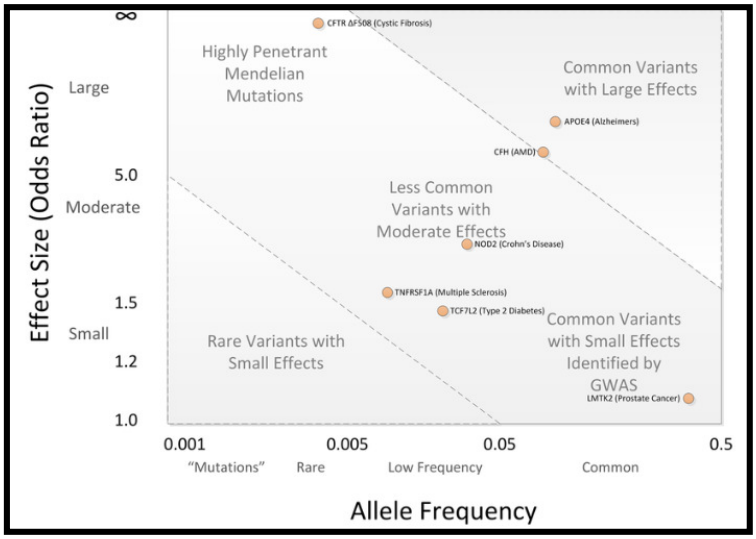


Figure 14 Spectrum of Disease Allele Effects.
Disease associations are often conceptualized in two dimensions: allele frequency and effect size. Highly penetrant alleles for Mendelian disorders are extremely rare with large effect sizes (upper left), while most GWAS findings are associations of common SNPs with small effect sizes (lower right). The bulk of discovered genetic associations lie on the diagonal denoted by the dashed line. Adapted from Bush and Moore, 2012.

The selection of genetic markers investigated in most Genome Wide Association Studies (GWAS) has been based on the "common disease, common variant" hypothesis (Manolio, 2009). These studies are designed to provide a survey of common variations of minor allele frequencies (MAFs) >0.05 (Panoutsopoulou *et al*., 2013). SNP arrays provide a picture of genome-wide polymorphisms in many individuals; but inevitably suffer from ascertainment biases favoring SNPs that are common in the populations for variant discovery (Wang *et al.,* 2013; Rabbani *et al.,* 2014; Panoutsopoulou *et al.,* 2013). Gene sequencing methods provide a more accurate and complete perspective with respect to all polymorphisms in target regions in which low frequency (0.01<MAF<0.05) and rare (MAF<0.01) variants are assessed (Panoutsopoulou *et al.,* 2013). As a result, the field is now shifting toward the study of low frequency variants under the hypothesis of "common disease, rare variant," this is, multiple rare variants with large effect size are in some cases the main determinants of the complex disease genetic risk (Marian, 2012; Eichler, *et al.,* 2010).

Whole Exome Sequencing, an approach in which all exons of protein coding genes are sequenced, shifts this focus from common to rare variants (Koeleman *et al.,* 2013; Ng *et al.,* 2010) and explores the extent to which rare alleles explain the heritability of complex diseases and health related traits. These studies enable the unbiased discovery of coding variations for subsequent association testing for complex traits (Kiezun *et al.,* 2012) but have been underpowered and over 10000 exomes are required to achieve statistical power in order to robustly detect associations. Thus tremendous effort has been devoted to the development of tools for variant analysis in the process of quality control, alignment, variant identification, and downstream association studies (Wang *et al.,* 2013; Wu *et al.,* 2014).

Exome sequencing data contain an abundance of rare coding variation and indicates that a large fraction of this variation is functional. There are many more rare variants than common ones and sequencing additional samples continues to uncover additional rare variants. This relative excess can be attributed to recent population expansion but is also likely to be due to purifying selection (Kiezun *et al.,* 2012). Rare variation is enriched for evolutionary deleterious and thus functional variants and the proportion of non-synonymous variants is higher among rare than among common variants (Kiezun *et al.,* 2012). Common variants are ancient and frequently present in

all human populations whereas rare variants are likely to be population specific, having originated from founder effects 10 to 20 generations ago. Thus, rare variants that are associated with complex phenotypes are likely to have effect sizes larger than those of common variants (Frazer *et al.,* 2009). As it is expected that many rare variants will have a very restricted geographic distribution, matching of case and control ancestries is important (Do *et al.,* 2012).

1.2.3.1 Assessing rare coding variation

Genetic association studies are sensitive and may detect minor susceptibility genes contributing less than 5% of the total genetic contribution to a disease (Tang *et al.,* 2013). The approach used for this type of analysis is based on comparing the frequency of the allele studied in unrelated patients with matched controls. If the allele appears significantly more frequent in patients then in controls, then it is considered to be associated with the disease (Tang *et al.,* 2013). Single nucleotide polymorphisms (SNPs) are the most important genetic markers for genetic association analysis, due to the abundance of SNPs covering the entire human genome at a high density (Tang *et al.,* 2013).

Candidate gene Association Analysis use candidate genes of a known sequence and location that are considered to be involved in the disease pathology. However approaches based on prior hypothesis have a limited power to detect novel genetic variants. Instead, a non-prior hypothesis is a more powerful approach for identifying gene association with a disease by screening the whole human genome or exome (Tang *et al.,* 2013).

1.2.4. Next Generation Sequencing Technologies

In the 1970s, Sanger and colleagues and Maxam and Gilbert developed methods to sequence DNA by chain termination and fragmentation techniques, respectively (van Dijk *et al.,* 2014). A growing demand for increased throughput led to laboratory automation and process parallelization which stimulated the development and

commercialization of next-generation sequencing (NGS) technologies (van Dijk *et al.,* 2014) (Figure 15). These new sequencing methods shared three major improvements: 1) instead of requiring bacterial cloning of DNA fragments they relied on the preparation of NGS libraries in a cell free system; 2) instead of hundreds, thousands to many millions of sequencing reactions were produced in parallel; 3) the sequencing output was directly detected without the need for electrophoresis; base interrogation was performed cyclically and in parallel (van Dijk *et al.,* 2014). The enormous numbers of reads generated by NGS enabled the sequencing of entire genomes at an unprecedented speed. However, a drawback of NGS technologies was their relatively short reads. (van Dijk *et al.,* 2014)

The first NGS technology to be released in 2005 was the pyrosequencing method of 454 Life Sciences, which is now Roche. A year later, the Solexa/Illumina sequencing platform was commercialized. Illumina later acquired Solexa, in 2007. The third technology to be released was Sequencing by Oligo Ligation Detection (SOLiD) by Applied Biosystems, now Life Technologies, in 2007. In 2010, Ion Torrent, Life Technologies released the Personal Genome Machine (PGM) and then the Ion Proton. This system was developed by Jonathan Rothberg, the founder of 454, and resembled the 454 system. An important difference was that the Ion Torrent technology used semiconductor technology and did not rely on the optical detection of incorporated nucleotides using fluorescence and camera scanning. This resulted in higher speed, lower cost, and smaller instrument size.

Figure 15. Next Generation Sequencing Technologies:
Roche, 454 Technology - GS-FLX Titanium and GS-FLX+, Illumina Technology - HiSeq 2000/2500 and MiSeq and Ion Torrent technology - PGM and Ion Proton. Adapted from http://www.macrogen.com/por/service_ngs01.html

Other NGS methods have been developed, such as Qiagen intelligent biosystems sequencing-by-synthesis and a single molecule detection system by Helicos BioSciences. In the latter, the template DNA is not amplified before sequencing, which places this method at the interface between NGS and the so called third-generation sequencing technologies (van Dijk *et al.,* 2014). The leader in the third-generation sequencing field is currently Pacific Biosciences (PacBio). The long reads makes this technology ideal for the completion of de novo genome assemblies. It is based on the detection of natural DNA synthesis by a single DNA polymerase in which incorporation of phosphate labelled nucleotides leads to base specific fluorescence, detected in real time. Other promising technologies are starting to appear. An example is Nanopore sequencing, which is based on the transit of a DNA molecule through a pore while the sequence is read out through the effect on an electric current or optical signal. Nanopore is considered a third-generation technology because it enables the sequencing of single molecules in real time. A major advantage is direct sequencing of DNA or RNA molecules without the need for library preparation or sequencing reagents (van Dijk *et al.,* 2014). It should be noted that the development of NGS has made huge demands on bioinformatics tools for data analysis and management. (van Dijk *et al.,* 2014)

The advent of massively parallel sequencing technologies has transformed the field of human genetics and substantially reduced the cost of sequencing large genomic regions relative to the traditional Sanger sequencing (Wang *et al.,* 2013; Frazer *et al.,* 2009). Researchers are now capable of investigating variants from a wide range of allelic spectrum, including *de novo* mutations, variants that are too rare for inclusion on microarrays and higher-level structural variants. There are two unbiased sequencing approaches for detecting genetic variation within an individual: whole genome sequencing and whole exome sequencing (Gilissen *et al.,* 2011). At this time it is still financially impractical, for most laboratories, to perform whole-genome sequencing of a large number of samples and at a sufficiently high coverage as to present valid large-scale genetic association studies of complex traits, such as Type 2 Diabetes and its complications.

1.2.4.1 Whole Exome Sequencing

As referred, sequencing studies are emerging as a popular approach to test for association of rare coding variants with complex phenotypes under the assumption that multiple rare variants constitute the driving force for the trait of interest (Kiezun *et al.,* 2012; Do *et al.,* 2012)

New technologies such as Exome Sequencing, are needed to identify low frequency (less than 5%) or rare (less than 0.5%) variants having larger effect sizes that could potentially explain part of the "missing heritability" (Lyssenko & Laakso 2013; Manolio 2009). It enables more accurate and complete variant discovery, assuming that the risk variant is exonic and allow for, in theory, the direct association between phenotype and causal variant (Wang *et al.,* 2013). The exome representing approximately 1% of the human genome comprises about 30 million base pairs (Ng *et al.,* 2010) and accounts for about 85% of mutations identified in Mendelian diseases (Rabbani *et al.,* 2014; Ng *et al.,* 2010; Gilissen *et al.,* 2011). Developments in high throughput sequence capture methods have made exome sequencing an attractive and practical approach for the investigation of coding variation (Wang *et al.,* 2013; Ng *et al.,* 2010; Do *et al.,* 2012) as it provides the means to explore the interpretable part of the genome (Kiezun *et al.,* 2012). Alternative strategies can add the regulatory and 3' untranslated regions and other functionally annotated regions of interest such as miRNA genes and various noncoding RNAs (Life Technologies Bulletin, 2012).

1.2.5. Study Objective

The overall aim of this study is to search and identify candidate genetic markers that might explain the excess risk associated to the onset of Diabetic Retinopathy. The genetic factors for Diabetic Retinopathy remain to be established and although promising genes have been reported in various genetic studies, very few have been replicated and even those, show only weak associations to the disease. The "common disease, common variant" hypothesis has been the preferred approach in the past, but

the search for genetic markers underlying the onset of common and complex diseases, and their complications, has been shifting towards the “common disease, rare variant” hypothesis. Although more recent and still in early stages, this has been the main objective of many candidate genetic marker studies: to understand the importance of rare variants or multiple rare variant accumulation and their relation to the onset, development and pathological progression of the complication.

To reach this objective and contribute to the knowledge of which genetic factors may have an important role in this microvascular complication we will perform the Whole Exome Sequencing (WES) of 40 Type 2 Diabetes diagnosed Portuguese patient samples. Of these, 24 patients have Diabetic Retinopathy and 16 have not been diagnosed with Diabetic Retinopathy. The identification of rare and common variants will be pursued by using different bioinformatics and statistical tools in a case-control study. The rare variants will be searched by a gene burden test approach and so the outcome will be a list of biologically relevant genes whereas for the common variants, association tests will be applied.

One of the main goals of this line of research is the ability to perform early diagnosis and apply the correct and appropriate management procedures in order to delay or prevent the outcome of Diabetic Retinopathy. We expect this study will bring forth new genetic markers that, by joining large-scale sequencing data and phenotypic information related to the absence, presence and severity of DR, will improve the treatment and thus visual health of the T2D patients.

Chapter 2

Materials and Methods

2. Materials and Methods

The experimental approach to this study involved Whole Exome Sequencing (WES) of the 40 Type 2 Diabetes (T2D) patients. Various tasks had to be addressed: DNA extraction and quality control, exome library preparation, template amplification and sequencing-by-synthesis with the next-generation sequencing platform Ion Proton. The sequencing, quality filtering, analysis and identification of variants were performed with proprietary Ion Torrent software. Bioinformatics tools, developed at Genoinseq, then filtered, annotated and prioritized the variants. The interpretation of the list of genetic variants, their biological relevance and possible association to Diabetic Retinopathy was further assessed and the variants encountered were validated by genotyping thus concluding the tasks proposed for this study.

2.1 Patients Characterization

Blood samples from 40 Portuguese patients diagnosed with Type 2 Diabetes were used in this study. The samples, collected from the Endocrinology Unit of the Hospital Center of the University of Coimbra, were object of the DIAMARKER subproject: Genetic susceptibility for multi-systemic complications in Diabetes Type 2: New biomarkers for diagnostic and therapeutic monitoring from the Do-IT - Development and Operation of Translational Research project. One of the aims of this project was the development of solutions for diagnosis, prognosis and treatment of cancer, neurodegenerative diseases and diabetes. The object of this thesis was included in the objectives of the overall project to search for genetic markers for Type 2 Diabetes and its complications.

All 40 diabetic patients were characterized for various parameters: age, sex, glycated hemoglobin levels (HbA1c), retinopathy grade measured according to the Early Treatment Diabetic Retinopathy (ETDR) values and diabetes duration (Table I in Appendix). Of these, 24 patients (Cases) were diagnosed with Diabetic Retinopathy (ETDR value ≥ 20) and 16 patients (Controls) without Diabetic Retinopathy (ETDR value < 20) according to the clinical diagnosis of diabetic retinopathy. Twenty individuals not

diagnosed with Type 2 Diabetes, by the same parameters, were selected as second Control group for validation procedures. Moreover, twenty other samples of the Portuguese population, which exomes had been previously sequenced at Genoinseq, were also used for validation of the sequencing technology platform and the frequency of the variants encountered. All participants signed an informed consent.

2.2 DNA Extraction and Quality Control

2.2.1 DNA Extraction

Blood samples contain enzyme inhibitors and common anticoagulants such as heparin and EDTA that can interfere with the sequencing downstream assays. Thus, the DNA isolation procedure has to provide high-quality, inhibitor-free DNA.

DNA was extracted from the blood samples using DNeasy Blood and Tissue Kit (Qiagen, Hilden), with minor modifications to the manufacturer's instructions to ensure the best DNA yield and purity.

Briefly, samples were processed using a lysis buffer containing a detergent, for the disruption of cellular membranes and proteinase K, for the digestion of protein cellular components. Buffering conditions were adjusted to provide optimal DNA-binding and the lysate was loaded onto the DNeasy Mini spin column. During centrifugation, DNA was selectively bound to the DNeasy membrane as contaminants passed through. Remaining contaminants and enzyme inhibitors were removed in two efficient wash steps and the DNA eluted in 100µl Tris-EDTA buffer.

2.2.2 DNA Quality Control

DNA qualification procedures consist of both the quantification of double stranded DNA (dsDNA) and the assessment of its suitability, purity and integrity, for downstream applications. To assess the purity and presence of contaminants, 2µl of each DNA sample was applied to the Nanodrop ND-1000 spectrophotometer (ThermoFischer Scientific, Germany) and the DNA concentration was determined by

measuring the absorbance at 260 nm. The ratio of the readings at 260 nm and 280 nm (A_{260}/A_{280}) on the spectrophotometer provided an estimate of DNA purity with respect to contaminants that absorb UV light, such as protein. Pure DNA has an A_{260}/A_{280} ratio of 1.7–1.9. If, the samples did not reach the minimum acceptance values for A_{260}/A_{280} and A_{260}/A_{230} ratios, the DNA samples were subjected to a standard isopropanol purification method (Table II in Appendix).

The integrity and size of the genomic DNA samples was assessed by running approximately 50ng of DNA against a 1 Kb molecular marker, NZY DNA Ladder III (NzyTech, Lisboa, Portugal) on a 0.8%, TAE 1x, agarose gel in TAE 1x, to which we applied a 90V current for 30 min. The ethidium bromide stained agarose gel was visualized under ultraviolet light with the Molecular Imager Gel Doc XR System (Biorad, California, USA). This procedure is not highly accurate since large DNA molecules migrating through a gel will essentially move together in a size-independent manner however, it provides sufficient information in terms of integrity (DNA size range) and purity as RNA contamination runs as a diffuse smear at the bottom of the gel.

Fluorometry measurements allow the specific quantification of double stranded DNA (dsDNA) by the use of a fluorescent dye. The High Sensitivity PicoGreen Assay (Invitrogen, LifeTechnologies, USA) is highly sensitive and can detect as little as 20 pg of dsDNA in a 200µl assay. DNA standards and samples were mixed with the fluorochrome from the kit, according to standard manufacturers' recommendations and measured on the GeminiEM Microplate fluorometry instrument (Molecular Devices, Sunnyvale, CA). Sample measurements were then compared to the standards to determine DNA concentration. This assay is optimized to minimize the fluorescence contributions of RNA and ssDNA, such that dsDNA can be accurately quantified in the presence of equimolar concentrations of ssDNA and RNA with minimal effect on the quantitative results (Simbolo *et al.,* 2013).

Samples were considered complient if pure (A_{260}/A_{280} and A_{260}/A_{230} ratios above 1.7), integrate, inhibitor free and had a concentration > 10ng/µl.

2.3 Whole Exome Sequencing

Figure 16 illustrates the Ion Proton Exome Sequencing workflow which comprises 4 basic steps: 1) Library Construction; 2) Template Preparation; 3) Sequencing and 4) Data Analysis.

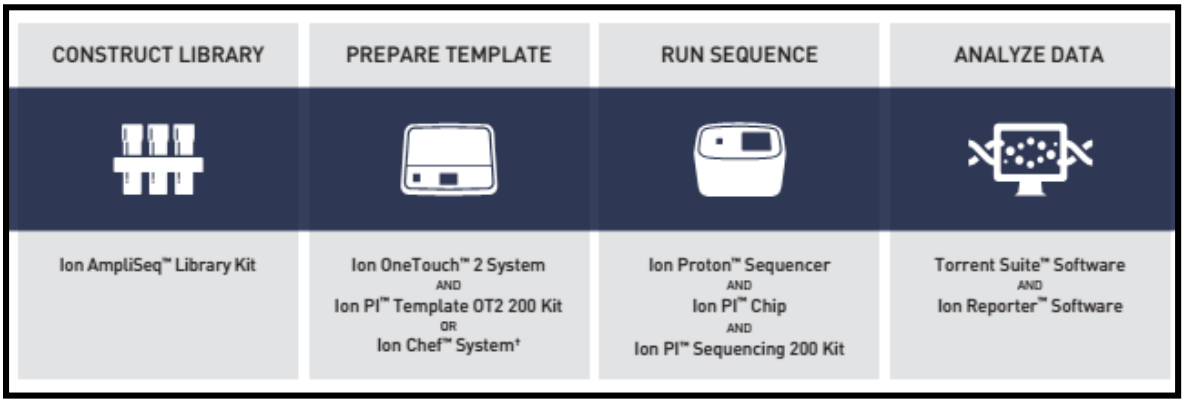


Figure 16. Basic workflow for Ion Proton Exome Sequencing:
Construct Library, Prepare Template, Run Sequence and Analyze Data. Adapted from http://dna.macrogen.com/eng/support/event/event_view.jsp?board_number=20187

2.3.1 Ion AmpliSeq Exome Library Preparation Protocol

Once the high quality DNA sample was purified, an exome library for each sample was prepared from 75ng with the Ion AmpliSeq™ Exome Library Preparation Kit (Life Technologies, Carlsbad, USA), according to the manufacturers' protocols. This procedure amplifies target exonic regions from 10-100 ng of genomic DNA (gDNA). It contains 12 primer pools, in a total of approximately 294,000 primer pairs for ultra-high multiplex PCR enrichment of the exome and is designed to create overlapping amplicons covering large target regions. The amplicons generated were then treated with FuPa Reagent to partially digest the primers and phosphorylate the amplicons. These were then ligated to Ion Adapters with barcodes and finally purified. The concentration of each Ion AmpliSeq Exome library was determined by qPCR with the Ion Library Quantitation Kit (Life technologies, Carlsbad, USA) (Figure 15), according to manufacturers' recommendations. Exome library quality was assessed using the High Sensitivity DNA Kit (Agilent Technologies, Waldbronn, Germany) and the Agilent 2100 Bioanalyzer System (Agilent Technologies, Santa Clara, USA), according to the

recommended manufacturers' protocol. Proper dilutions were prepared for template amplification and Ion Proton Sequencing.

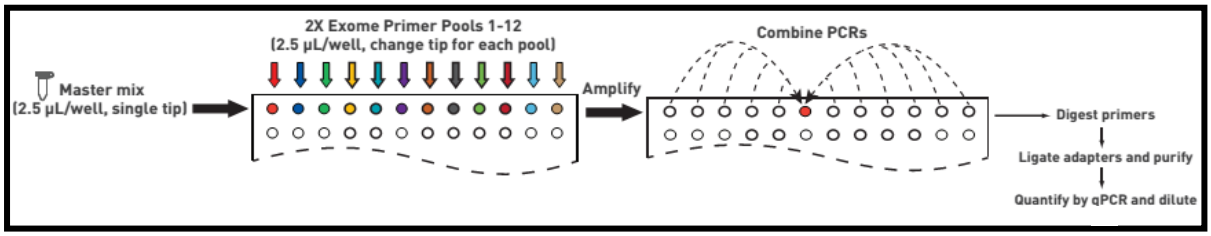


Figure 17. Library Preparation by Ion AmpliSeq Exome Technology.
Adapted from Ion AmpliSeq Exome Library Preparation Quick Guide.

2.3.2 Template Preparation and Exome Sequencing by Ion Torrent Technology

2.3.2.1 Clonal amplification of Templates

For most commercially available next-generation sequencing platforms, the clonal amplification of each DNA fragment by emulsion PCR is necessary in order to generate sufficient copies of sequencing template. The presence of adapter sequences enables selective clonal amplification of the library molecules (Figure 18, A). Furthermore, the adapter sequence also contains a docking site for the platform-specific sequencing primers (Life Technologies, 2012).

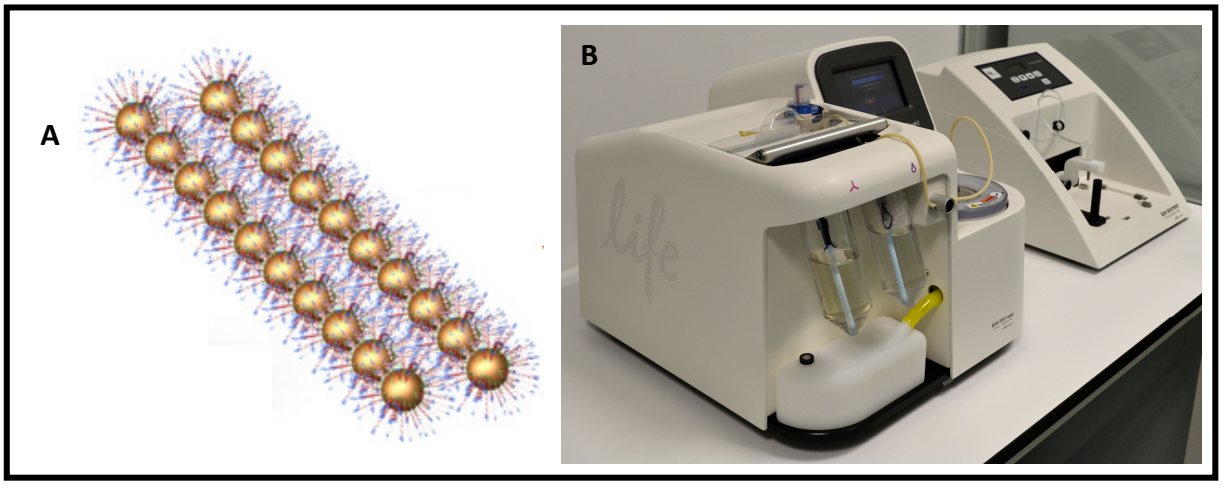


Figure 18. Clonal Amplification of templates.
A. Clonal amplification of each DNA fragment. B. Ion One Touch 2 System. Adapted from Ion PI Template OT2 200Kit v2 protocol.

Template preparation was performed on the Ion OneTouch2 System (Figure 18, B) according to manufacturer's instructions and the Ion PI Template OT2 200Kit v2 protocol.

2.3.2.2 Chip Preparation and Loading

After clonal amplification of the DNA fragments on the Ion Sphere Particles (ISPs), the enrichment of the template-positive ISPs, ion sphere particles that have amplified DNA, and performance of the ISP quality control, the enriched particles were prepared for loading on the PI chip, according to the manufacturers' protocol (Figure 19). The sequencing runs were performed with 2 exome libraries, each with a unique Molecular IDentifier (MID). (Exome vs IonXpress Barcode - Table III in Appendix)

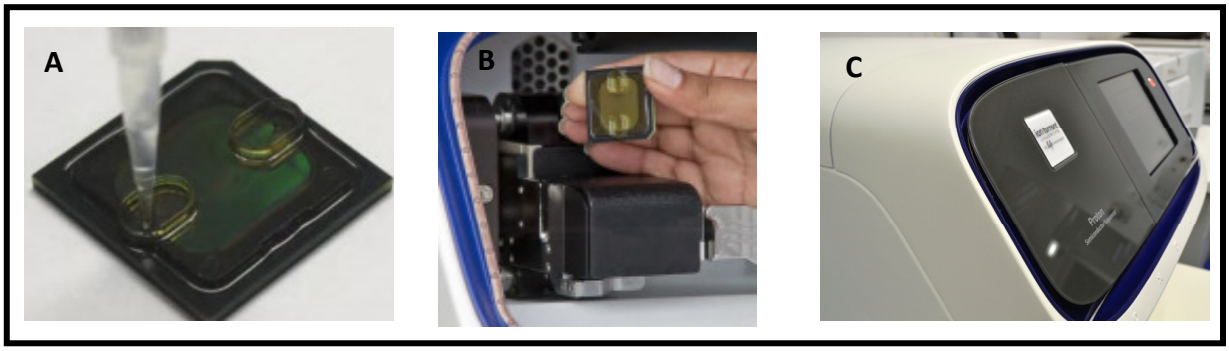


Figure 19. Chip Preparation and loading.
Ion PI chip loading with template-positive ISPs. B. Ion PI chip fitting into chip clamp. C. Ion Proton Sequencer. Adapted from Ion PI Sequencing 200 Kit v2 protocol.

Exomes were sequenced on the Ion Proton System and Ion PI chipv2 following manufacturer's instructions and the Ion PI Sequencing 200Kit v2 protocol.

2.3.2.3 Exome Sequencing

The incorporation of a deoxyribonucleotide triphosphate (dNTP) into a growing DNA strand involves the formation of a covalent bond and the release of pyrophosphate and a positively charged hydrogen ion. Ion semiconductor sequencing

exploits these facts by determining if a hydrogen ion is released upon providing a single species of dNTP to the reaction (Life Technologies, 2012).

Micro wells on a semiconductor chip each containing many copies of one single-stranded template DNA molecule and DNA polymerase are sequentially flooded with unmodified dATPs, dCTPs, dGTPs or dTTPs. The hydrogen ion that is released in the reaction changes the pH of the solution, which is detected by an ion-sensitive field-effect transistor (ISFET) used for measuring ion concentrations in solution. When the H^+ concentration changes, the pH is altered and the current through the transistor changes accordingly (Figure 20).

The series of electrical pulses transmitted from the chip to a computer is translated into a DNA sequence, with no intermediate signal conversion required. Because nucleotide incorporation events are measured directly by electronics, the use of labeled nucleotides and optical measurements are avoided. Signal processing and DNA assembly is then carried out in appropriate software (Life Technologies, 2012).

In this study, twenty sequencing runs were performed for a total of 40 exomes.

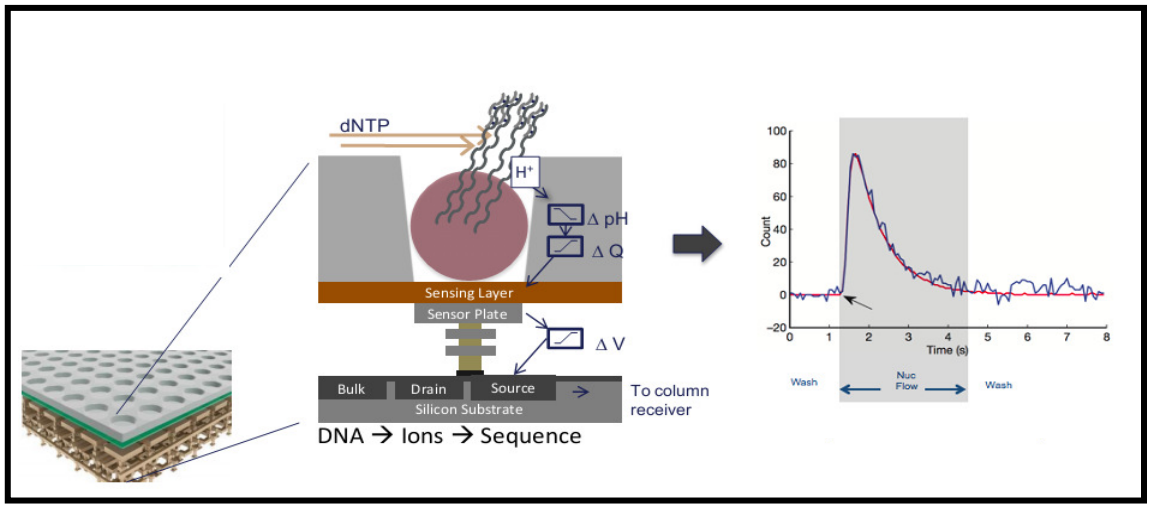


Figure 20. Sensor, well and chip architecture and Data collection.
A simplified drawing of a well, a bead containing DNA template, and the underlying sensor and electronics. Protons (H^+) are released when nucleotides (dNTP) are incorporated on the growing DNA strands, changing the pH of the well (ΔpH). This induces a change in surface potential of the metal-oxide-sensing layer, and a change in potential (ΔV) of the source terminal of the underlying field-effect transistor. Nucleotide incorporation signal from an individual sensor well; the arrow indicates start of incorporation event, with the physical model (red line) and background corrected data (blue line) shown. Adapted from, Rothberg *et al.,* 2011.

2.4 Bioinformatics Analysis

Next-generation sequencing instruments sequence millions of short DNA fragments in parallel. The data generated requires sophisticated computation and bioinformatics tools that become more and more crucial in the analysis and interpretation of sequencing data (Figure 21). Many alignment methods and variant callers have been developed and used to create complex pipelines. For the Ion Proton, the most efficient mapping and variant calling software are the ones developed by the manufacturers, as previously determined at Genoinseq (Life Technologies, 2012).

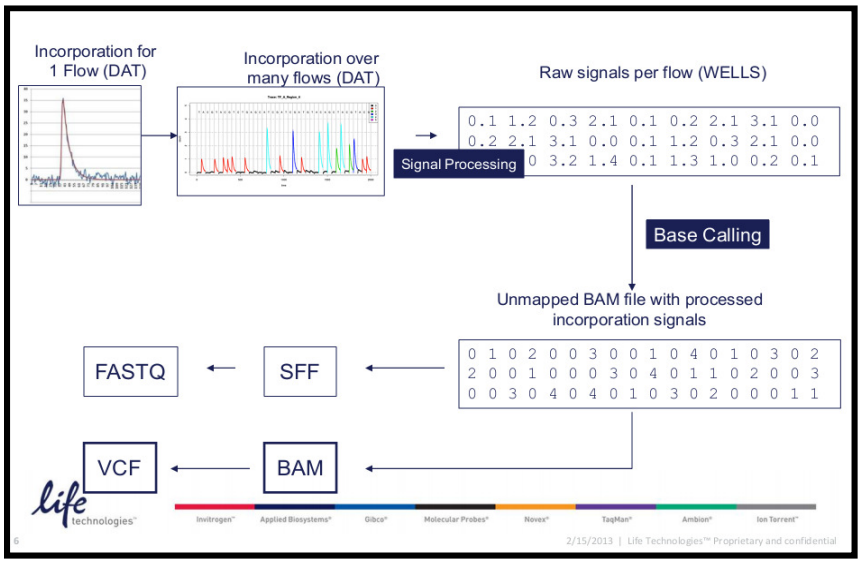


Figure 21. Data Collection, Signal Processing and Base Calling.
Files generated by the Ion Torrent Suite: SFF, FASTQ, BAM and VCF files are shown. Adapted from Life technologies, 2012.

2.4.1 Signal Processing, Base Calling and Mapping of exome sequenced data

After a sequencing run is complete, the main data analysis step of the software included base calling and mapping of the sequence data to a reference genome (Do *et al.,* 2012).

The raw data obtained by Ion Proton Sequencing System was processed on the Ion Proton sequencer and further transferred to the Proton Torrent Server, for read mapping and variant calling. (Figure 22)

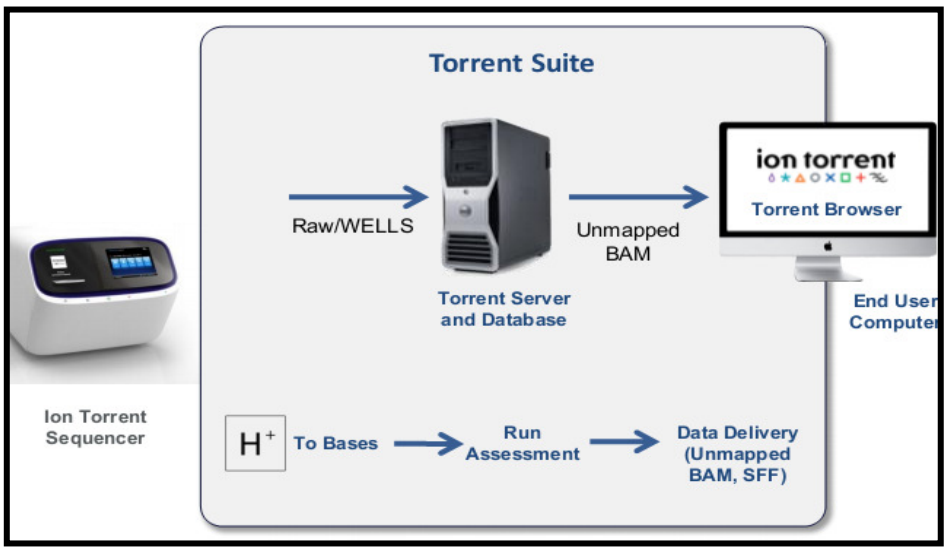


Figure 22. Bioinformatics Processing of sequenced raw data.
Adapted from Life Technologies, 2012.

For each of the 40 sequenced exomes, the Ion Proton adaptor sequences and low quality bases were trimmed using Torrent Suite software (Life Technologies, Carlsbad, CA, USA). The filtered reads were then mapped against the human reference genome hg19, using the TMAP version 4.0.6 (Life Technologies, Carlsbad, CA, USA). The resulting sequence alignment was stored in a Binary Alignment Map (BAM) file (Li *et al.,* 2009).

For sample identification control, the number of reads mapped on the Y chromosome were counted using SAMTools *(Li et al., 2009)* and divided by the number of total mapped reads thereby creating a ratio to determine the sample sex. This informatics sex was then compared to the subject sex. All samples passed this sample identification control.

2.4.2 Variant calling

The Variant Caller identifies variant sites where the aligned sequences deviate from the reference sequences at known positions. Variant calling for homozygous or heterozygous SNPs and small or large insertions and deletions (indels) was the next automated step in the data analysis process and was performed by Torrent Variant

Caller plugin version 4.0 (Life Technologies, Carlsbad, CA, USA), with the optimized parameters for exome sequencing as recommended for Ion AmpliSeq™ Exome Kit (Life Technologies, Carlsbad, CA, USA).

Torrent Variant Caller (TVC) is a genetic variant caller for Ion Torrent™ Sequencing platforms, and is specially optimized to exploit the underlying flow signal information in the statistical model to evaluate variants. It is designed to call single-nucleotide polymorphisms (SNPs), multi-nucleotide polymorphisms (MNPs), insertions, deletions, and block substitutions. Basically it: 1) finds all positions with any evidence for a variant, 2) focuses on these positions and evaluates if there is enough evidence for a SNP or indel call and 3) filters the candidates based on key parameters (strand bias, enough coverage, known error prone position) (Life Technologies, 2012).

Interactive tables of identified variants were able to be visualized and evaluated using Broad Institute's Integrative Genomics Viewer (IGV) (Life Technologies, 2012) (Figure 23). Variants that passed all the set filters were reported to a single output variant call format (VCF) file (Life Technologies, 2012; Danecek *et al.,* 2011). Further steps involved filtering and annotation to reduce variant sites to a set of genes with possible function and activity.

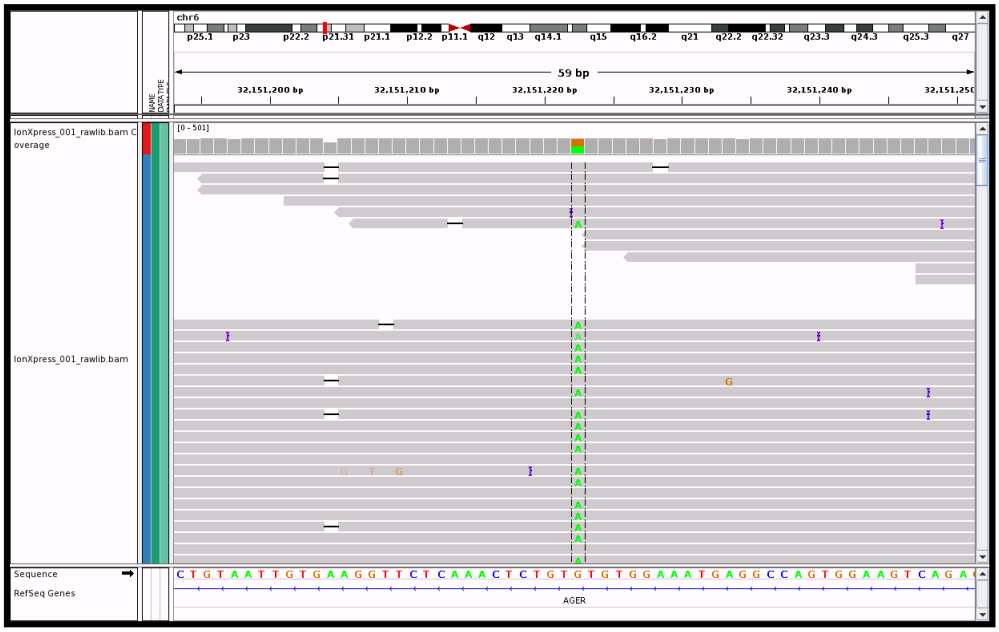


Figure 23. An example of Integrative Genomics Viewer (IGV) interactive table.

In our study, all variants with less than 10X coverage were discarded. In addition, overall variant positions from the 40 exomes were considered and integrated altogether using VCFTools (Danecek *et al., 2011).* For all positions, the homozygous equal to the reference genome positions were further evaluated since no information about these genotypes was reported in the variant calling file generated. A script developed at Genoinseq was used to detect reference genotypes when the altered allele frequencies in the mapping file were lower than 5%. If the altered allele frequency was higher than 5%, the genotype was considered undetermined.

The genotypes were then recalled based on the altered allele frequency using an in-house script. Altered homozygous genotypes were assigned for altered allele frequencies (aaf) higher than 0.8 and heterozygous for frequencies between 0.2 and 0.8.

Functional consequences of variants can be predicted by examining the effects of amino-acid changes using comparative sequence and protein structure analyses. Many computational prediction and conservation methods are available (Kiezun *et al.,* 2012; Do *et al.,* 2012). According to their impact on protein-coding transcripts, annotation tools can identify single nucleotide variants that result in synonymous, missense, nonsense, splice site alterations or read-through alterations. Indels are typically annotated according to whether or not they result in frameshift (Do *et al.,* 2012). These annotation tools also assign each variant a score, based on analysis of protein structure or evolutionary conservation to separate variants with little functional impact from those more likely to damage protein function (Do *et al.,* 2012). Some variants might have more than one annotation resulting from various overlapping transcripts. These annotation conflicts are resolved by focusing on: 1) only canonical transcripts for each gene (RefSeqGene or Ensembl), 2) the longest transcript in each gene or 3) the most deleterious prediction from all available transcripts (Do *et al.,* 2012). The accuracy of those methods is approximately 80% and is likely highest for rare variants. Truly functional variants are most likely deleterious and are kept at low frequencies by purifying selection and so, common variants are most likely neutral and non-functional. Therefore using prediction methods enriches for functional variants and thereby boosts the power of association tests (Kiezun *et al.,* 2012).

In our study and to predict the impact of genetic variation on protein coding transcripts, Variant Effect Predictor (VEP) software was used (McLaren *et al.,* 2010). The transcripts were obtained from Ensemble, version GRCh37 (Cunningham *et al.,* 2015). The VEP results were then used by the Gemini framework (Paila *et al.,* 2013) to annotate the variants against a comprehensive set of genomic annotation files including: dbSNP (Sherry *et al.,* 2001), ENCODE (Dunham *et al.,* 2012), ClinVar (Landrum *et al.,* 2014), 1000 Genomes (McVean *et al.,* 2012), the Exome Sequencing Project (NHLB1), Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa *et al.,* 2014; Kanehisa *et al.,* 2000), Genomic Evolutionary Rate Profiling (GERP) (Cooper *et al.,* 2005), and Human Protein Reference Database (HPRD) (Prasad *et al.,* 2009). Finally, Gemini saved all information related to impact and annotation prediction in a database for further analysis. Furthermore, to detect relatedness between samples the Gemini framework performed pair-wise genetic distance tests.

2.5 Statistic Analysis

The two populations, cases and controls, were characterized for age, sex, disease duration and glycated hemoglobin levels (HbA1c) using the IBM SPSS software, Version 20.0 (Armonk, NY: IBM Corp.). The variants were tested using the Pearson χ^2 test for the qualitative variables (sex and presence vs. absence of DR) and the Fisher exact test for the quantitative variables (age, HbA1c levels and disease duration).

Statistical analyses of both common and rare variants were performed using different approaches. The analyses of rare variants require statistical methods that are fundamentally different from association statistics used for testing common variants. Rare variants have to be combined in a gene (or pathway) for an association test to reach sufficient power (Kiezun *et al.,* 2012) while for common variants, traditional regression-based association tests are normally applied (Kiezun *et al.,* 2012).

2.5.1 Identification of common variants

The common variant approach consisted in the prioritization of common variants using the Pearson X_i^2 test or the Fisher exact test to detect the ones that had significant genotype differences between cases and controls (univariate analysis). The variants with a p-value lower than 0.05 were then tested in a binary logistic regression model for adjustment to external factors like age, sex, disease duration and glycated hemoglobin. Adjusted odds ratios (ORs) and 95% confidence intervals were computed and statistically significant results were considered for $p<0.05$. The univariate analysis and logistic regression were computed with the StatsModels python module (http://statsmodels.sourceforge.net).

2.5.2 Identification of rare variants accumulating genes

Although every analysis of exome sequence data should start with single variant association tests, these are not well powered for rare variants. Because most variants are individually rare, achieving adequate statistical power requires a design with the ability to combine and evaluate groups of variants likely to have an impact on the function of a specific gene and to compare the distribution of these variant groupings to the distribution of the trait of interest (Do *et al.,* 2012).

The approach for rare variants identification aimed to prioritize variants using the EPACTS – Efficient and Parallelizable Association Container Toolbox (Kang *et al.,* 2010), by performing a gene-wise burden test. We used a package that enables the implementation of a variety of burden tests, (http://genome.sph.umich.edu/wiki/EPACTS**)** (Kang *et al.,* 2010; Do *et al.,* 2012).

The Gemini framework was used to select variants that passed the Hardy-Weinberg equilibrium (HWE ≥ 0.001), with a call rate higher than 80%, a minor allele frequency lower than 1% (either in the ESP project or 1000 Genomes project), a minor allele frequency in our population lower than 5% and that were non-synonymous or present in splice site regions and thus predicted to be of medium (MED) or high (HIGH) severity. All indel variants were excluded. The selected variants were saved in the

Variant Call Format (VCF) file. These variants were grouped by gene using EPACTS. A binary (cases vs. controls) and a quantitative gene-wise burden emmaxVT test (cases adjusted to ETDR vs. controls) was then performed to select new candidate genes for the DR complication.

2.6 Genetic Variant Validation

In some rare cases, exome sequencing of a single large sample will be sufficient to demonstrate association of encountered variants with a disease. More often it is necessary to examine the most promising variants in additional samples (Do *et al.,* 2012). Genotyping validation provides an additional measure of call set quality independent of the population genetics statistics (Kiezun *et al.,* 2012) and the genotyping assay should include sites at various allele frequencies, common and rare, but especially at low frequencies (~1%) (Kiezun *et al.,* 2012).

For this study we considered various validation procedures: 1) validation of the sequencing technology by genotyping 8 of the 40 exomes by Illumina microarray with the HumanOmniExpressExome Protocol (Illumina, San Diego, CA, USA); 2) comparison of the MAF for each common or rare variant in 20 sequenced exomes unrelated to the study in question; 3) validation of the common variants, by Sanger sequencing, of 2 homozygous and 2 heterozygous for the altered allele; 4) validation of the rare variants encountered by ASO-PCR genotyping; 5) confirmation that the rare variants encountered are in fact rare in our population by ASO-PCR genotyping of 20 non T2D Portuguese control samples.

The eight samples used for sequencing technology validation were sent to ATLAS Biolabs GmbH at Berlin, Germany, to confirm the results of the Ion Proton platform.

2.6.1 Sanger sequencing

Sanger sequencing is a method of DNA sequencing based on the selective incorporation of chain-terminating dideoxynucleotides by DNA polymerase during in

vitro DNA replication. This method although much less than in decades before, remains in wide use for smaller-scale projects, validation of Next-Generation sequencing results and for obtaining especially long contiguous DNA sequence reads (>800 nucleotides).

To validate the results obtained for candidate common variants, we sequenced 2 homozygous and 2 heterozygous patients for each variant that had not been validated by Illumina microarray technology.

DNA was amplified for each gene region of interest with the forward and reverse primers represented in Table 1 and the purified fragment sent to GATC Biotech (Cologne, Germany) for Sanger sequencing. PCR reactions were performed for each sample using 1x Advantage SA PCR Buffer, 0.2 µM of each PCR primer, 0.2 mM dNTPs (Bioron, Ludwigshafen am Rhein, Germany), 6% DMSO (Roche Diagnostics GmbH, Mannheim, Germany), 1x Advantage 2 Polymerase Mix (Clontech, Mountain View, CA, USA), and 50ng of template DNA in a total volume of 50µl. The PCR conditions involved a 4 min denaturation at 94ºC, followed by 30 cycles of 94ºC for 30s, specific annealing temperature for 45s and 68ºC for 60s and a final extension at 68ºC for 10 min. Negative controls were included for all amplification reactions. Electrophoresis of the PCR products was undertaken on a 1% (w/v) agarose gel and the amplified fragments were purified using AMPure XP beads (Agencourt, Beckman Coulter, USA) according to manufacturer's instructions.

Table 1. rs ID, gene and primer information for common variant validation. The gene sequence ensemble ID, gene direction, primer sequence and fragment size are also represented.

rs ID	Gene	Gene sequence ID	Gene Direction	Primer Name	Primer Sequence 5' - 3'	Fragment Size (bp)
rs62357156	ITGA1	ENSG00000213949.8	Forward Strand (1)	ITGA1_DR_F	*CTCAGGAGAAAGCAGTTGTG*	537
				ITGA1_DR_R	*CACCCATCCAACATGAAGAC*	
rs2296123	PRKCQ	ENSG00000065675.14	Reverse Strand (-1)	PRKCQ_DR_F	*GTGCTCTGTCCTCCTTATAC*	355
				PRKCQ_DR_R	*GGTATTCGTCTTGGCATCTC*	
rs80067372	TNFSF12	ENSG00000239697.10	Forward Strand (1)	TNFSF12_DR_F	*CTTCTTGGGCACATCAGGAG*	463
				TNFSF12_DR_R	*CTCTTCCACGCAGACTCAC*	
rs10794640	NARFL	ENSG00000103245.13	Reverse Strand (-1)	IOP1_DR_F	*GTGAAGGGTGAAGGTCAGTG*	532
				IOP1_DR_R	*CAGGTGAATTCGACCAGTGG*	
rs4434138	STAB1	ENSG00000010327.10	Forward Strand (1)	STAB1_DR_F	*CGTCCTTGCTTCATTCCCTC*	484
rs4234633				STAB1_DR_R	*CATAGAAGGTGGAGAAGTTGG*	

2.6.2 ASO-PCR Genotyping

The increasing need for large-scale genotyping applications of single nucleotide polymorphisms (SNPs) in model and non-model organisms requires the development of low-cost technologies accessible to minimally equipped laboratories. The method used for validation of the rare variants encountered allows efficient discrimination of SNPs by allele-specific PCR with optimized PCR conditions. A common reverse (or forward) primer and two forward (or reverse) allele-specific primers with different tails amplified two allele-specific PCR products, which were visualized by agarose gel electrophoresis. PCR specificity was improved by the introduction of a destabilizing mismatch within the 3' end of the allele-specific primers. This is a simple and inexpensive method for SNP detection *(*Gaudet *et al*., 2009*)* and was applied to validate the true rarity of the candidate rare variants, this is, certify if these variants were in fact rare or if they were a characteristic polymorphism of the Portuguese population.

The ASO-PCR reactions for each rare variant validation had to be individually optimized. The primers used for each genotype determination are listed in Table 2. A pair of control primers were added to each PCR reaction to ensure that all the conditions for amplification were met.

Table 2. rs ID, gene and primer information for common variant validation. The gene sequence ensemble ID, gene direction, primer sequence and fragment size are also represented.

rs ID or Pos.	Gene	Gene sequence ID	Gene Direction	Primer Name	Primer Sequence 5' - 3'	Fragment Size (bp)
rs114516513	DMXL2	ENSG00000104093.13	Reverse Strand (-1)	DMXL2_DR_F_T	*CTGATGAAAGCATTCATAGTGA*	541
				DMXL2_DR_F_C	*CTGATGAAAGCATTCATAGTGG*	
				DMXL2_DR_R	*GTTTAGCCCACTCTCCAACC*	
Chr15: 51743890				DMXL2_DR_F1	*GTGAGAGATGAGGATTCAGG*	454
				DMXL2_DR_R1_C	*AATGTGTCCCAGAGGCAAAC*	
				DMXL2_DR_R1_G	*AATGTGTCCCAGAGGCAAAG*	
Chr15: 51772229				DMXL2_DR_F2	*CTTCAGTAAGGATGGAACTC*	579
				DMXL2_DR_R2_C	*CATGGATGTGATTATTTAAGTAC*	
				DMXL2_DR_R2_G	*CATGGATGTGATTATTTAAGTAG*	
rs141999878	E2F8	ENSG00000129173.12	Reverse Strand (-1)	E2F8_DR_F_C	*CTTTATTAAGAGTTACAGTATAGAG*	431
				E2F8_DR_F_A	*CTTTATTAAGAGTTACAGTATAGAT*	
				E2F8_DR_R	*CTTCTAGGCTTACTATCTGAGG*	
rs77599073				E2F8_DR_F2_C	*GTTTTCTGTTTCAGGCTCCAG*	617
				E2F8_DR_F2_G	*GTTTTCTGTTTCAGGCTCCAC*	
				E2F8_DR_R2	*CCAACCTACCTTGATTGCTC*	
rs793274				E2F8_DR_F3_T	*CAGGATTTACAGCTCCCCAA*	524
				E2F8_DR_F3_C	*CAGGATTTACAGCTCCCCAG*	
				E2F8_DR_R3	*GTCACAGCAACTGATTGTCC*	
Chr11: 19247163				E2F8_DR_F4_C	*CTTCCCCCTTTTCCTTCCAG*	402
				E2F8_DR_F4_G	*CTTCCCCCTTTTCCTTCCAC*	
				E2F8_DR_R4	*GTCACAGCAACTGATTGTCC*	
Chr11: 19258929				E2F8_DR_F5_C	*GTGTCATAAGTTCTTAGCACG*	700
				E2F8_DR_F5_T	*GTGTCATAAGTTCTTAGCACA*	
				E2F8_DR_R5	*CACAGCACATCACTTCACAG*	

Chapter 3

Results and Discussion

3. Results and Discussion

3.1 Patient Characterization

Our study population had forty 40 T2D patients. Twenty-four were diagnosed with Diabetic Retinopathy (ETDR>20) and considered Cases and sixteen had no symptoms of Diabetic Retinopathy (ETDR≤20) and were thus considered Controls. Table 3 presents the characteristics and statistical analysis of the studied population.

Table 3. Characteristics and statistical analysis of the study population. The population was analyzed for sex, age, disease duration and glycated hemoglobin (HbA1C) levels. Differences in the characteristics were evaluated with the Pearson X^2 test for the qualitative variables and the Fisher exact test for the quantitative variables. SD - Standard Deviation .

Patients	DR (n=24)	non-DR (n=16)	t-test	X^2
Female (n, %)	10 (41.7%)	10 (62.5%)		0.197
Male (n, %)	14 (58.3%)	6 (37.5%)		
Age (years ± SD)	63.83 ± 6.91	64.38 ± 8.32	0.824	
HbA1C (% ± SD)	9.45 ± 1.83	9.59 ± 2.94	0.846	
Diabetes duration (years ± SD)	19.21 ± 8.41	12.44 ± 8.03	0.015	

Perhaps the most important step in any exome sequencing study is the choice of samples to sequence (Do *et al.,* 2012). The characteristics of the study population are a key component of robust allelic association studies. The desire is to eliminate all potential confounding as well as competing non-genetic variables in order to amplify the effects of genetic variants. Cases and controls should be matched, as much as possible, for demographics and other characteristics in order to enrich the chance of detecting genetic effects (Marian, 2012).

There were no significant statistical differences between cases and controls for sex, age and HbA1c levels. The only parameter that was statistically significant was diabetes duration ($p<0.05$). It is known that diabetes duration is the parameter that is largely responsible for the development and aggravation of diabetic retinopathy (Tarr *et al.,* 2013; Ola *et al.,* 2012 and Safi *et al.,* 2014). This finding is reflected in our study

population as this covariate sets the difference between presenting, or not, the microvascular complication.

The characterization of all patients is thoroughly presented in Table I (in Appendix).

Various studies have repeatedly shown that uncontrolled diabetes results in complications from the disease (Fong *et al.,* 2004; Droumaguet *et al.,* 2006). Glycated Hemoglobin A_{1c} is formed through the nonenzymatic gycation of hemoglobin. In patients with T2D, it is used as a therapeutic target and shows a good correlation with the risk of developing microvascular complications (Chen *et al.,* 2013). As so, the goal for people with diabetes, is a HbA_{1c} level lower than 7%. (Fong *et al.,* 2004; Droumaguet *et al.,* 2006, Sugimoto *et al.,* 2013, Ozturk *et al.,* 2009). Interestingly, 5 patients: Ex19, Ex27, Ex44, Ex45 and Ex50, suffering with T2D for 23, 19, 33, 20 and 15 years respectively did not have Diabetic Retinopathy and revealed poorly controlled blood glucose levels (HbA_{1c}>7%). On the other hand, Ex12 has had T2D for 6 years, revealed a controlled glucose level (HbA1c=6.3%) but suffers from Diabetic Retinopathy. These observations support the idea that the onset of Diabetic retinopathy may have a genetic cause.

3.2 DNA Extraction and Quality Control

All DNA samples extracted revealed high integrity and the absence of RNA contamination as can be seen in figure 24. The purity determination and quantification results of the all 40 DNA samples are presented in Table II (in Appendix).

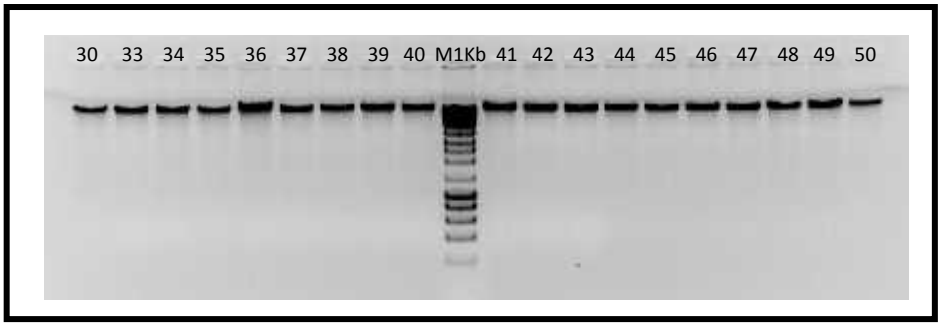


Figure 24. Agarose gel (0.8%) electrophoresis of genomic DNA samples:
Ex30 and Ex33 to Ex50. 1Kb, NZYDNA Molecular Marker Ladder III. The gel was stained with Ethidium bromide and the image presented in inverse.

3.3 Ion AmpliSeq Library Preparation

After Exome Library Preparation, quality control procedures were performed. Figure 25 presents an example of the expected Bioanalyzer profile with libraries ranging from 200 to 350bp. All 40 exomes were controlled and further validated for exome sequencing.

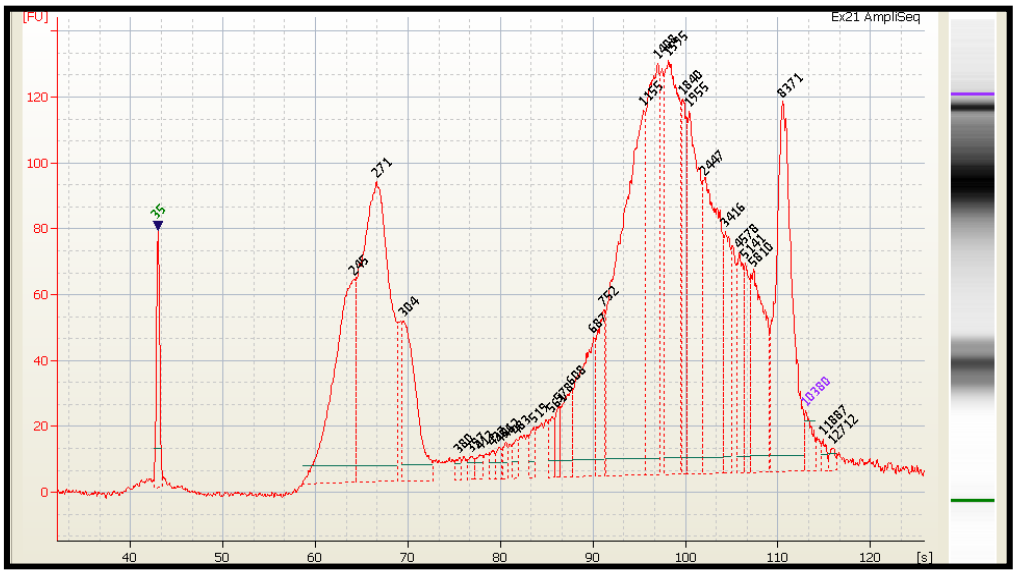


Figure 25. Bioanalyzer profile of the amplified exome library sample Ex21.
The size range was determined on the Agilent Bioanalyzer instrument with the Agilent High Sensitivity DNA kit, according to manufacturer's instructions. The upper and lower marker peaks are identified and correctly assigned (green and purple numbers on the trace). The black numbers on the trace peaks represent the size of the fragments.

3.4 Exome Sequencing Results

Standards for the generation of high-quality exome sequence data are rapidly emerging (Do *et al.,* 2012). A high-quality base coverage of 20x or greater in 80-95% of the protein-coding sequences in each genome, after removal of ambiguously mapped and duplicated reads, should be able to identify the vast majority of protein-coding variants with high specificity but because the efficiency of enrichment protocols exhibit great local variation, sequencing the protein coding region of each individual to an average depth of 60-80x is recommended (Do *et al., 2012).*

3.4.1 Exome Sequencing Metrics

Figure 26 presents the sequencing run report of samples Ex31 and Ex33 in the Ion Proton sequencing platform. In this case, nearly 100 million reads were generated with a mean fragment size of 193bp.

Figure 26. Run Report for Exome Sequencing of Samples Ex31 and Ex33 with Ion Proton.
90,989,101 reads with a mean length of 193bp were generated in a total of 17.5G bases.

The obtained reads were then mapped against the human reference genome hg19. The mean coverage analysis metrics for all 40 samples are presented in Table 2. The average mapped reads, which refer to the number of reads that were mapped to the full reference genome, were 41.6 ± 6.3 million (mean ± standard deviation). Nearly

95% of the reads targeted the aimed region, providing a mean coverage of almost 120X.

Table 4. Coverage Analysis Metrics. The mapped reads refer to the number of reads that mapped the reference genome, the reads on target reflect the percentage of mapped reads that were aligned to the target region, the mean depth is the mean average target base reads coverage and the coverage uniformity is the percentage of target bases covered by at least 0.2x the average base read depth.

Metrics	Values Mean ± Standard Deviation)
Mapped Reads (million ± SD)	41.6 ± 6.3
On Target (% ± SD)	94.5 ± 0.9
Mean Depth (x ± SD)	119 ± 19
Coverage Uniformity (% ± SD)	92.0 ± 1.3

An important consideration for variant discovery is Coverage which typically translates into confidence in variant calling. Greater coverage increases the confidence in putative variants. Another parameter, Coverage Uniformity, is also important for exome sequencing since this minimizes the amount of sequencing needed to achieve a desired coverage threshold of 20x with a significant proportion of exome bases targeted, above 90% *(*Do *et al.,* 2012; Life Technologies, 2012*)*. This value informs whether the exome was uniformly covered or if there was a particular region that was more or less sequenced. The sequencing and coverage metric values we obtained were within the value range predefined by Ion Torrent for high quality exome sequencing with AmpliSeq.

The coverage depth and on target % for all 40 exomes sequenced are represented in Figures A and B in the Appendix.

3.4.2 Variant Caller Metrics

The comparison of the sequenced exomes to the reference genome sequence (hg19) by the Torrent Variant Caller identified an average of 52,698 variants, per sample. Of these, 91.3±2.8% were Single Nucleotide Polymorphisms (SNPs), 3.3±1.8% were insertions, 4.8±1.7% were deletions and 0.3±0.5% were multiple nucleotide polymorphisms (MNPs) (Figure 27).

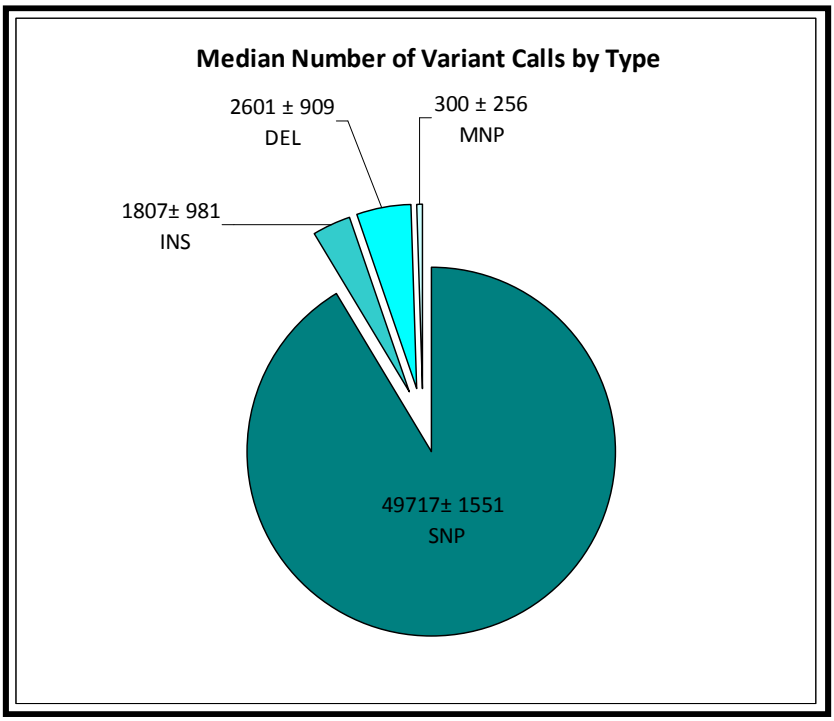


Figure 27. Representation of the median number of variant calls obtained by type.
Single Nucleotide polymorphism (SNP), Insertion (INS), deletion (DEL), Multiple Nucleotide Polymorphism (MNP).

3.5 Candidate Genes obtained by Rare Variant accumulation

One of the advantages of whole exome sequencing is the ability to search the complete coding regions of a genome with the potential of revealing a large number of rare variants. Currently the sequencing error rate is estimated to be about 1%, which is at a similar scale of the frequency of rare variants (Wang *et al.,* 2013*)*. There are several issues that can complicate this variant calling step: 1) the presence of indels, a major source of false positive in variant identification, 2) variable GC content in short reads, an error introduced by library preparation due to PCR artifacts, and 3) variable base quality scores (Wang *et al.,* 2013*)*. Filtering was thus critical for achieving high quality calls *(*Kiezun *et al.,* 2012*)*.

In order to present candidate rare variants with a potential to affect Diabetic Retinopathy onset, a rare variant accumulation approach was used.

3.5.1 Candidate Rare Risk Genes

All rare variants selected and subjected to further gene-wise burden tests passed the Hardy-Weinberg equilibrium (HWE ≥ 0.001), with a call rate higher than 80% and of medium or high severity among other features (presented in Materials and Methods, Statistical Analysis). The Gene-wise burden test approach led to identification of several genes with an accumulation of rare variants. This list was filtered for rare variant accumulating genes with P-value<0.05 and further filtered for rare variants present in T2D individuals diagnosed with DR. Thus our rare variant search was solely done on the "risk" genes (Tables 5 and 6).

Table 5. Genes accumulating rare variants in individuals with Diabetic Retinopathy. This list was obtained by a Binary Test (cases vs controls). PASS_MARKERS are the number of variants encountered in that gene, BURDEN_CNT refers to the number of individuals in which these variants were found. P-values are also represented.

Risk MARKER_ID (chrom: start-end_gene)	PASS_MARKERS	BURDEN_CNT	PVALUE
1:57207874-57257861_C1orf168	4	5	0,0039
22:26164394-26423535_MYO18B	7	8	0,0062
1:215847545-216246585_USH2A	5	5	0,02
15:51743890-51772229_DMXL2	**4**	**4**	**0,023**
8:71025871-71041146_NCOA2	4	6	0,032
16:2287215-2287632_DNASE1L2	**3**	**4**	**0,033**
11:19247163-19258929_E2F8	**5**	**4**	**0,035**
3:195594795-195597000_TNK2	3	3	0,037
14:20711005-20711005_OR11H4	1	2	0,038
19:14029627-14031728_CC2D1A	3	3	0,039
5:37020562-37049406_NIPBL	2	2	0,04
1:9098071-9101726_SLC2A5	2	2	0,042
8:130761778-130774958_GSDMC	2	4	0,043
22:30856091-30857645_SEC14L3	2	2	0,046
6:158910743-158923120_TULP4	3	3	0,048

Table 6. Genes accumulating rare variants in individuals with Diabetic Retinopathy. This list was obtained by a Quantitative Test (cases adjusted to ETDR vs. controls). NUM_PASS_VARS are the number of variants encountered in that gene, NUM_SING_VARS refers to the number of variants present in only one individual (with DR). P-values are also represented.

Risk MARKER_ID (chrom: start-end_gene)	NUM_PASS_VARS	NUM_SING_VARS	PVALUE
17:72365698-72368711_GPR142	**2**	**1**	**0,00023708**
11:64519437-64527310_PYGM	4	3	0,0010121
1:247614602-247615148_OR2B11	3	2	0,0025469
17:63149547-63189687_RGS9	3	3	0,0026079
1:153587865-153587865_S100A14	**1**	**0**	**0,002845**
17:3101591-3101617_OR1A2	2	2	0,002845
4:140640704-140640704_MAML3	**1**	**0**	**0,002845**
5:178555044-178771135_ADAMTS2	**2**	**2**	**0,002845**
22:26164394-26423535_MYO18B	7	6	0,0029763
16:2287215-2287632_DNASE1L2	3	2	0,0052564
2:130931095-130931167_SMPD4	1	0	0,0056899
3:36756863-36780081_DCLK3	2	2	0,0056899
16:14956982-14989425_NOMO1	3	2	0,0063792
13:25254891-25284616_ATP12A	6	5	0,0081569
19:34180082-34262952_CHST8	3	3	0,0097202
10:125506302-125558631_CPXM2	2	2	0,01138
1:247695273-247695756_OR2C3	3	3	0,01138
19:44681046-44681717_ZNF226	3	3	0,01138
22:45128123-45255629_ARHGAP8	2	2	0,01138
22:45128123-45255629_PRR5-ARHGAP8	2	2	0,01138
5:1254510-1254510_TERT	1	0	0,01138
9:27524515-27524731_IFNK	2	2	0,01138
8:130761778-130774958_GSDMC	2	1	0,011678
15:63920891-64041689_HERC1	3	3	0,012565
11:62984850-62997031_SLC22A25	2	1	0,015884
1:47726135-47746113_STIL	3	3	0,015884
20:43723614-43723669_KCNS1	2	2	0,01707
21:45175604-45175847_PDXK	2	2	0,01707
4:983556-983684_SLC26A1	2	2	0,01707
9:75355093-75355093_TMC1	1	0	0,01707
1:10713845-10720270_CASZ1	**3**	**3**	**0,019678**
4:152499205-152570675_FAM160A1	3	3	0,019678
1:215847545-216246585_USH2A	5	5	0,021222
16:772780-775236_CCDC78	2	2	0,025605
17:43481395-43506940_ARHGAP27	2	2	0,025605
22:41537234-41574925_EP300	**2**	**2**	**0,025605**
5:153432606-153433026_MFAP3	2	2	0,025605
9:107554255-107599376_ABCA1	3	3	0,025605
9:128678097-128678097_PBX3	1	0	0,025605
3:124826801-124906167_SLC12A8	4	3	0,027513
11:65144075-65147013_SLC25A45	1	0	0,029161
11:75298327-75298558_MAP6	3	3	0,029161
6:29407955-29408721_OR10C1	5	1	0,030427
10:103899871-103899970_PPRC1	2	2	0,034139
1:171173089-171178090_FMO2	3	3	0,034139
22:21799719-21800202_HIC2	2	2	0,034139
5:140579961-140581006_PCDHB11	2	2	0,034139
6:36104494-36106712_MAPK13	1	0	0,034139
9:35740293-35740293_GBA2	1	0	0,034139
1:246078824-246078824_SMYD3	1	0	0,034851
3:78676694-78763613_ROBO1	3	3	0,034851
1:94467548-94520733_ABCA4	4	4	0,035006
10:89264706-89265292_MINPP1	2	2	0,045519
11:63276396-63276432_LGALS12	2	2	0,045519
12:110418705-110418705_TCHP	1	0	0,045519
12:111748192-111785678_CUX2	2	2	0,045519
1:223177972-223178706_DISP1	2	2	0,045519
12:57648716-57663716_R3HDM2	2	2	0,045519
16:3100094-3108573_MMP25	2	2	0,045519
16:89293934-89294865_ZNF778	2	2	0,045519
17:77758597-77758597_CBX2	1	0	0,045519
19:2217094-2217786_DOT1L	2	2	0,045519
20:57036604-57042655_APCDD1L	**2**	**2**	**0,045519**
22:30188418-30221173_ASCC2	2	2	0,045519
3:50331133-50332327_HYAL3	2	2	0,045519
19:38893807-38899459_FAM98C	3	2	0,0456
12:56075599-56077768_METTL7B	4	4	0,048364
16:10971206-11002927_CIITA	3	3	0,048364
17:7910817-7915471_GUCY2D	3	2	0,048364
2:238483751-238483751_RAB17	1	0	0,048364

All rare variant accumulating genes identified (p-value<0.05) were further evaluated and their function studied by database and literature search. As there were 85 genes that fulfilled our filter criteria and being, in the present study, difficult to correlate rare variants to the complication, our option was to select those with biological relevance and association to mechanisms and pathways related to Diabetic Retinopathy pathogenesis.

From this search, 10 candidate genes aggregated rare risk variants. Three were obtained from the binary test approach (cases *vs* controls) and seven from the quantitative test approach (cases adjusted to ETDR *vs* controls) and were considered potentially interesting for a more detailed analysis.

Genes obtained by the binary test (cases vs controls)

Table 7 presents the results obtained from a published literature review for genes Dmx-Like 2 (DMXL2), also known as Rabconnectin3 (Rbcn-3), E2F Transcription Factor 8 (E2F8) and Deoxyribonuclease I-Like 2 (DNASE1L2). The table details the biological function and possible relevance to the DR pathology.

Table 7. Biological function and possible relevance to DR pathology of genes obtained by the binary test approach. The associated pathway or mechanism and number of rare variants accumulated in the gene are also presented. P-values were obtained by the EPACTS statistical test.

Gene	Biological Function and Relevance	Associated Pathway or Mechanism	P-Value EPACTS	Number of rare variants accumulated
DMXL2	Participates in the regulation of the Notch signaling pathway (Sethi *et al.,* 2010)	Notch Signaling	0,023	5
E2F8	Promoter of sprouting angiogenesis possibly by acting as a transcription activator; associates with HIF1A, recognizes and binds to the VEGFA promoter and activates expression of the VEGFα gene (Weijts *et al.,* 2012)	Angiogenesis	0,035	3
DNASE1L2	Expression is induced by TNFα and IL1β via the NFKB pathway. Possible physiological function under inflammatory conditions. (Shiokawa *et al.,* 2004)	Inflammatory Processes	0,003	3

Genes obtained by the quantitative test (cases adjusted to ETDR levels vs controls)

Table 8 presents the results obtained from a published literature review for genes MasterMind Like3 (MAML3), ADAM Metallopeptidase with Thrombospondin Type1 Motif2 (ADAMTS2), E1A Binding Protein P300 (EP300), Castor Zinc Finger1 (CASZ1), Adenomatosis Polyposis Coli Down-Regulated 1-Like (APCDD1L), S100 Calcium Binding ProteinA14 (S100A14) and G Protein-Coupled Receptor 142 (GPR142). The table details the biological function and possible relevance to the DR pathology.

Table 8. Biological function and possible relevance to DR pathology of genes obtained by the quantitative test approach. The associated pathway or mechanism and number of rare variants accumulated in the gene are also presented. P-values were obtained by the EPACTS statistical test.

Gene	Biological Function and Relevance	Associated Pathway or Mechanism	P-Value EPACTS	Number of rare variants
MAML3	Acts as a transcriptional coactivator of NOTCH proteins. May be important in the Notch signaling pathway (Lin *et al.,* 2002)	Notch Signaling	0,0028	1
ADAMTS2	A metalloproteinase that plays a key role in the processing of fibrillar procollagen precursors. Is able to reduce the proliferation of endothelial cells (Motte *et al.,* 2010).	Angiogenesis	0,028	2
EP300	Important in cell proliferation and differentiation. Has been identified as a co-activator of HIF1A and plays a role in the stimulation of hypoxia-induced genes such as VEGFA (Freedman *et al.,* 2002)		0,025	2
CASZ1	Promotes vascular assembly and morphogenesis through the direct regulation of an EGFL7/RhoA-mediated pathway. SNPs in this gene are associated with blood pressure variation (Charpentier *et al.,* 2013)	Vascular Assembly and Morphogenesis	0,019	3
APCDD1L	Encodes an inhibitor of the Wnt signaling pathway which mediates pathological vascular growth in proliferative retinopathy *(*Zhao *et al.,* 2013)		0,045	2
S100A14	Binds to AGE receptors and stimulates RAGE-dependent signalling cascades, promoting cell proliferation. Regulates cell migration by modulating levels of MMP2, a matrix protease that is under transcriptional control of P53/TP53 (Jin *et al.,* 2011)		0,0028	1
GPR142	Rhodopsin-like receptors are a family of proteins that comprise the largest group of G-protein coupled receptors that includes light receptors. They transduce extracellular signals (Fredriksson *et al.,* 2003)	G-Protein Coupled Receptors	0,0002	3

All variants (from binary test approach and quantitative test approach) were confirmed in the BAM files to assure that there were no false positives. The rs ID, chromosomal position, impact, impact severity, polyphen prediction, scaled CADD value, transcript ID, transcript name and length for the variants are presented in Figures E-G (in the Appendix). These tables also refer the individuals that present the variant, the respective ETDR values and disease duration (years).

The binary test generated a list of rare variant accumulating genes that may be associated to diabetic retinopathy, as can be seen in Figure 28 (A). Rare variants in these genes were found in individuals with ETDR levels ranging from 25 to 90. On the other hand, the quantitative test generated a list of rare variant accumulating genes that are mostly associated to severe diabetic retinopathy, as can be seen in Figure 28 (B). Rare variants in these genes were found in individuals with ETDR values ranging from 43 to 90.

Sample	EX42	EX12	EX24	EX46	EX17	EX43	EX48	EX49	EX7	EX10	EX23
ETDR level	25	35	35	35	43	47	53	53	90	90	90
Approximate Diabetes Duration (years)	17	6	22	13	17	5	9	27	33	21	16
A Variants obtained from binary test	E2F8	DNASE1L2	DMXL2	E2F8			DNASE1L2	E2F8	E2F8		DNASE1L2
	DMXL2			DNASE1L2				E2F8			
								DMXL2			
B Variants obtained from quantitative test					CASZ1	APCDD1L	EP300	CASZ1	GPR142	GPR142	GPR142
							APCDD1L		EP300	MAML3	MAML3
									CASZ1	S100A14	S100A14
										ADAMTS2	ADAMTS2
	variants with CADD scaled >14										

Figure 28. Representation of the rare variant accumulating genes, in T2D individuals with DR, by the binary test approach (A) and the quantitative test approach (B).
Orange cells refer to variants with a scaled CADD value above 14.

The binary test approach (cases *vs* controls) prioritized rare variant accumulating genes that are associated to mild and moderate diabetic retinopathy whereas the quantitative test approach (cases adjusted to ETDR values *vs* controls) prioritized rare variant accumulating genes associated to more severe signs of the complication.

Of all rare variant accumulating genes, 2 genes were considered potential candidate genes (Table 9). *E2F8* and *DMXL2* were the genes that presented a larger number of variant accumulation, some of which with a deleterious effect according to prediction software, and seemed to have, from the published literature research, a possible pathophysiological association to Diabetic Retinopathy and were thus further explored and validated. The deleteriousness of the candidate variants was assessed using Combined Annotation-Dependent Depletion (CADD) scores (Shyr *et al.,* 2014)

Table 9. Candidate Genes for Diabetic Retinopathy. Genes were selected by the identification of rare variants (European Minor Allele Frequency, EUR MAF<0.01) in the group of diabetic retinopathy patients. For gene E2F8 5 variants were identified in 21% of retinopathies (p=0.035). For Rbcn-3 gene 3 variants were found in 13% of case samples (p=0.023). The associated biological mechanisms and the type of variant and respective CADD (Combined Annotation Dependent Depletion) are described.

Gene	Number of rare variants (MAF < 0.01)	Individuals (%)	P-value (EPACTS)	Associated Mechanisms	Type of variant (predicted impact)
E2F8	5	21	0.035	Angiogenesis promotion by activating the transcription of VEGFA	Splice (n.a) (CADD 16,62)
					Missense (benign) (CADD 3,98)
					Missense (benign) (CADD 8,91)
					Missense (deleterious) (CADD 13,21)
					Missense (deleterious) (CADD 35)
DMXL2	3	13	0.023	Functional regulation of the Notch signaling pathway	Missense (benign) (CADD 14,65)
					Missense (deleterious) (CADD 18,03)
					Missense (deleterious) (CADD 17,19)

The exome sequencing technology was validated by genotyping 8 of the 40 individuals by Illumina Microarray under the DoIt project. This service was subcontracted to ATLAS Biolabs GmbH (Berlin, Germany). For this validation, various genotypes obtained, for specific variants, by Ion Torrent Technology, were compared with the genotypes obtained by Illumina microarray. More than 98% of the common variants (assessed by the Ion Torrent sequencing and the Illumina microarray

genotyping) were equal. This procedure validated the next-generation technology used in this study.

DMXL2 gene accumulated 3 rare variants in heterozygosity. One of the variants (rs 114516513) was validated in Ex 49, by the Illumina genotyping microarray. The other two variants (positions 51743890 and 51772229 in chromosome 15), present in Ex42 and Ex24 respectively, which had not rs IDs could not be validated by the same procedure and were validated by ASO-PCR.

E2F8 gene accumulated 5 rare variants in heterozygosity. Three of the variants (rs141999878, rs77599073 and rs793274) were validated in Ex 42, Ex49 and Ex46 respectively, by the Illumina genotyping microarray. The other two variants (positions 19247162 and 19258928 in chromosome 11), present in Ex7 and Ex49 respectively, were validated by ASO-PCR.

We performed ASO-PCR, for all 8 rare variants, in a heterozygote individual, a wild type individual and in 20 non T2D individuals. Furthermore, this variant was also evaluated in 20 samples which exomes had been previously sequenced at Genoinseq in order to check the frequency of this variant in the Portuguese population (Figure 29). The genotype results for all 40 T2D individuals were also considered in the allele frequency determination.

	Candidate Rare Genes					
	DMXL2			**E2F8**		
	T/C	**C/G**	**C/G**	**C/A**	**C/G**	**T/C**
	rs114516513	Chr15: 51743890	Chr15: 51772229	rs141999878	rs77599073	rs793274
Validation Method	Illumina Array			Illumina Array	Illumina Array	Illumina Array
BAM file check	✓	✓	✓	✓	✓	✓
Variant Confirmation	✓			✓	✓	✓
Validation Method	ASO-PCR					
Heterozygote	Ex49	Ex42	Ex24	Ex42	Ex49	Ex46
Wild Type	Ex42	Ex49	Ex49	Ex24	Ex24	Ex24
Allele frequency (24 cases+ 16 controls + 20 previously sequenced exomes + 20 non diabetic individuals)	0.0125	0.0125	0.0125	0.0125	0.0125	0.0125

Figure 29. Candidate rare variant accumulating genes and validation procedures.
The validation methods, variants and results are presented. The samples used for heterozygote and wild type control are also reported. The allele frequency for the rare variants (1 in 80) was 0.0125.

The various validation procedures confirmed the exome sequencing results.

DMXL2 (Rbcn-3) gene relation to Diabetic Retinopathy

Confirmation of the variants by BAM file verification is presented in Figure 30. In all images the comparison between the heterozygote individual (above) and wild type **individual (below) is represented.**

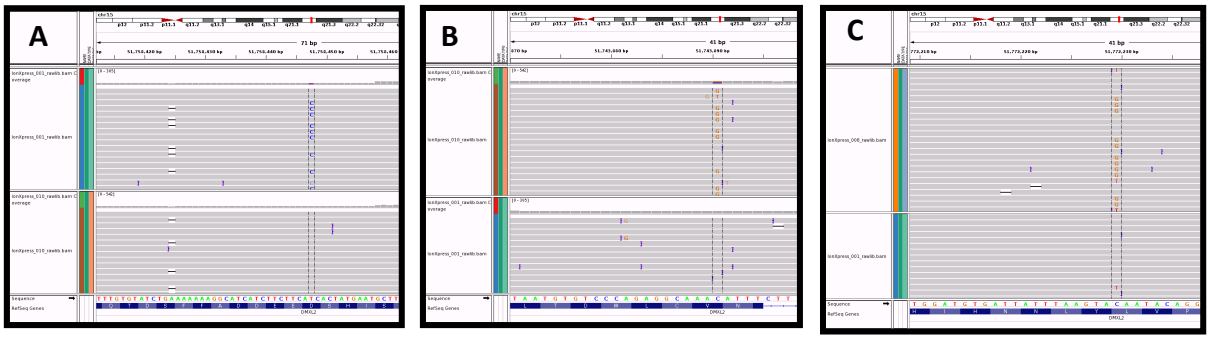


Figure 30. BAM file verification of the 3 rare variants in DMXL2 gene.
A. rs114516513: HT – Ex49 and WT – Ex24. B. 15:51743889: HT – Ex42 and WT – Ex49. C. 15:51772228: HT – Ex24 and WT – Ex49.

An example of the ASO-PCR genotyping for the 3 rare variants in gene DMXL2 is represented in Figure 31. The electrophoresis gel refers to the T2D individual with DR that had the variant and was thus considered our heterozygote, a T2D individual with DR that did not present the variant and was thus our wild type control and 4 non T2D individuals. The 16 other non T2D individual genotypes are presented in the Appendix.

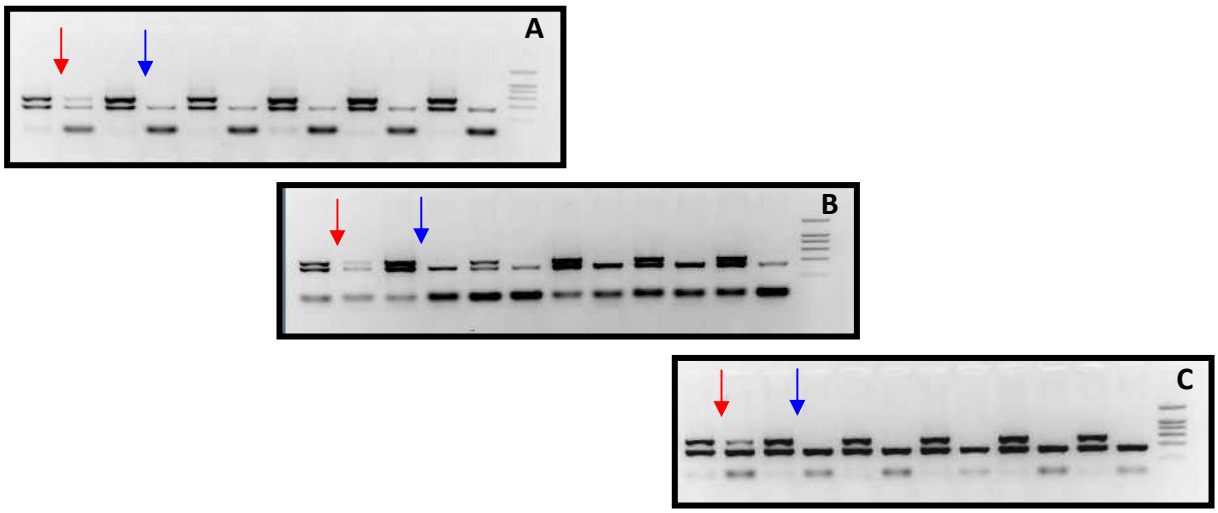


Figure 31. Examples of the electrophoresis gel results from the ASO-PCR genotyping validation of the 3 variants in DMXL2 gene.
The red arrows depict the genotype of the T2D individual with the rare variant and thus our heterozygote control. The blue arrows depict a T2D individual that did not have the variant and thus our wild type homozygous control. The result for 4 control individuals are also represented. All individuals for all variants were wild type. Every two wells correspond to an individual/genotype. A. rs114516513 T/C; B. 15:51743889 C/G; C. 15:51772228 C/G.

The first phases of Diabetic Retinopathy are characterized by the thickening of the basal membrane, loss of inter-endothelial tight junctions and early and selective loss of pericytes accompanied by an increase in the vascular permeability, capillary occlusions, microaneurisms and loss of endothelial cells. Pericytes are responsible for vascular stability and control endothelial proliferation and the loss of these contractile cells is a hallmark of Diabetic Retinopathy (Arboleda-Velasquez *et al.,* 2014).

It has been demonstrated that the Notch signaling pathway has an important role in the survival of the pericytes. For an efficient signaling, Notch membrane receptors need to be cleaved and its intracellular domains, Notch IntraCellular Domains (NICD), internalized to the nucleus where they function as transcription factors for target genes. DMXL2, also known as Rabconnectin-3 (Rbcn3) influences the Notch signaling by its involvement in the acidification of intracellular compartments by the V-ATPase complex. These endocytic compartments modify the activity of γ-secretase, which cleaves NICD, and/or affects the liberation of NICD (Sethi *et al.,* 2010, Roca & Adams, 2007)

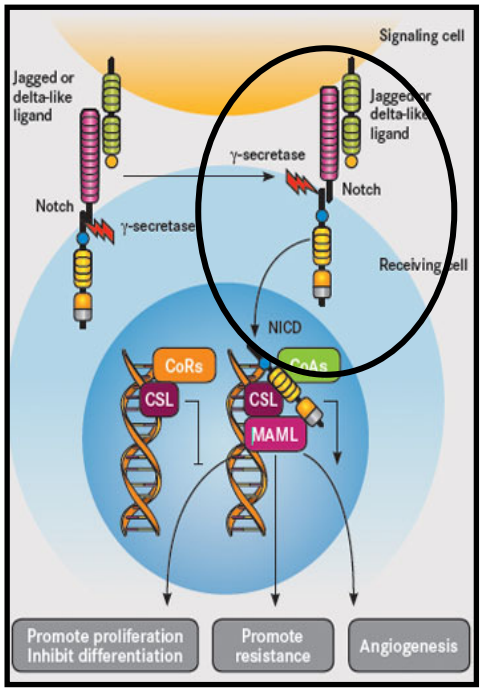


Figure 32. Maturation of the notch receptor involves cleavage at the prospective extracellular side during intracellular trafficking.
This results in a bipartite protein, composed of a large extracellular domain linked to the smaller transmembrane and intracellular domain. Binding of ligand promotes two proteolytic processing events; as a result of proteolysis, the intracellular domain is released and can enter the nucleus to engage other DNA-binding proteins and regulate gene expression. Adapted from http://www.onclive.com/publications/Oncology-live/2013/February-2013/Notch-Signaling-Tackling-a-Complex-Pathway-With-a-New-Generation-of-Agents

E2F8 gene relation to Diabetic Retinopathy

The confirmation of the variants by BAM file verification is shown in Figure 33. In all images the comparison between the heterozygote individual (above) and wild type individual (below) is represented.

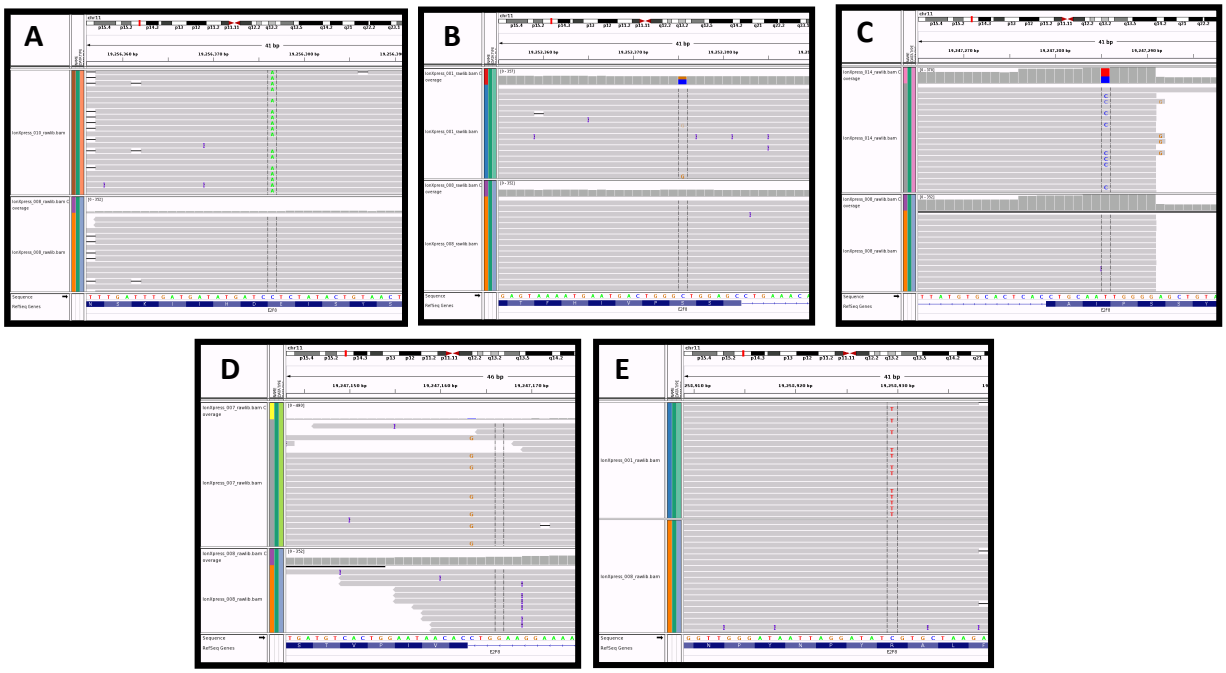


Figure 33. BAM file verification of the 5 rare variants in E2F8 gene.
A. rs141999878: HT – Ex42 and WT – Ex24. B. rs77599073: HT – Ex49 and WT – Ex24. C. rs793274: HT – Ex46 and WT – Ex24. D. 11:19247162: HT – Ex7 and WT – Ex24. E. 11: 19258928: HT –Ex49 and WT –Ex24.

An example of the ASO-PCR genotyping for the 5 rare variants in gene E2F8 is represented in Figure 34. The electrophoresis gel refers to the T2D individual with DR that had the variant and was thus considered our heterozygote, a T2D individual with DR that did not present the variant and was thus our wild type control and 4-6 non T2D individuals. The 14-16 other non T2D individual genotypes are presented in the Appendix.

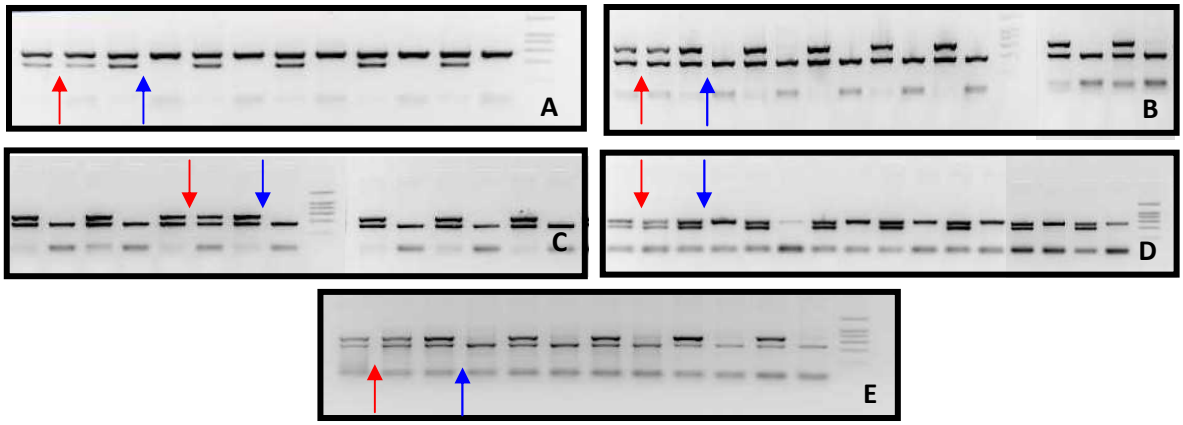


Figure 34. Examples of the electrophoresis gel results from the ASO-PCR genotyping validation of the 5 variants in E2F8 gene.
The red arrows depict the genotype of the T2D individual with the rare variant and thus our heterozygote control. The blue arrows depict a T2D individual that did not have the variant and thus our wild type control. The result for 4 or 5 control individuals are also represented. All individuals for all variants were wild type. Every two wells correspond to an individual/genotype. A. rs141999878 C/A. B. rs77599073 C/G. C. rs793274 T/C. D. 11:19247162 C/G. E. 11: 19258928 C/T.

E2F7 and E2F8 are the most recent members of the E2F transcription factor family and having been reported as essential in the adequate formation of blood vessels, act as promotors of germinative angiogenesis, the extension of new blood vessels by proliferation of the endothelial cells of pre-existent capillaries, by activating the transcription of the Vascular Endothelial Growth Factor A (VEGFA) gene.

E2F7/E2F8 forms a transcriptional complex with the Hypoxia Inducible Factor 1 (HIF1A), recognizes and binds to the VEGFA promoter and activates the expression of the VEGFA gene (Weijts *et al.,* 2012).

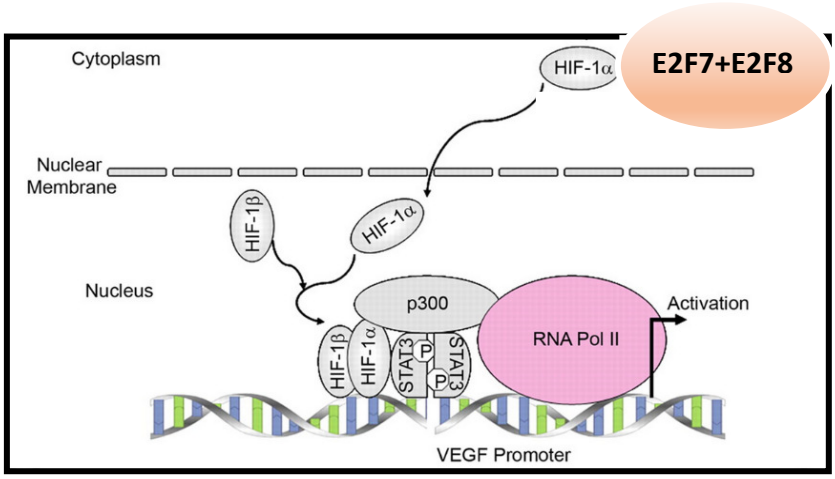


Figure 35. Schematic representation of the activation of the VEGFA promoter by HIF1A.
HIF1A consists of two subunits: an inducibly expressed HIF1 alpha subunit and a constitutively-expressed HIF1 beta subunit. Under hypoxia, HIF1α becomes stable and interacts with coactivators such as p300 to modulate its transcriptional activity. Adapted from Ghosh *et al.,* 2009. It has been proposed that E2F7/F8 forms a transcriptional complex with HIF1A (Weijts *et al.,* 2012).

From what we could gather, E2F8 and DMXL2 seem to be involved in the mechanisms related to Diabetic Retinopathy. However, further studies have to be conducted before considering the association of these genes to the onset, development or aggravation of Diabetic Retinopathy.

In the future, the genotyping of a much larger population and gene functional studies should give a better idea of the relevance of these potential genetic markers.

3.6 Candidate Common Variants

To the list of common variants, after statistical analysis, no other filter was applied as to increase the possibility of encountering common variants that could be biologically relevant to the pathogenesis of the complication. There was a need to define a prioritization procedure as the number of common variants was large.

Genes that had been referred in the literature as possibly associated to the disease and had a p-value<0.05 were checked.

From this approach, 11 common variants (rs1035798 in AGER, rs62357156 in ITGA1, rs7125062 in MMP1, rs80067372 in TNFSF12, rs9907595 in PLXDC1, rs2296123 in PRKCQ and rs7483 in GSTM3, rs4434138 and rs4234633 in STAB1, rs10794640 in NARFL and rs4698803 in EGF were found.

The list of 11 variants that were considered biologically relevant and potentially related to the pathways that have been documented as participants in the development of DR are represented in Table 10.

Six were considered risk variants (rs1035798, rs62357156, rs7125062, rs80067372, rs4434138 and rs4234633) with an Odds Ratio>1 and encountered in T2D patients that were diagnosed with DR and five (rs10794640, rs9907595, rs2296123, rs7483 and rs4698803), with an Odds Ratio<1 were considered protective variants associated to the group of T2D patients without DR.

Table 10. Common genetic variants associated with Diabetic Retinopathy. The frequency of the minor allele identified in the Diabetic Retinopathy patients group (DR), controls and in the European population (1000 Genomes Project) is described for each SNP. The genotypic frequencies between patients with and without DR were analyzed by multinomial logistic regression, adjusted to the following co-variates: age, sex, disease duration and glycated hemoglobin.

ID rs	Gene	Ref	Alt	MAF (DR)	MAF (controls)	MAF (EUR)	P-value	OR (IC 95%)	Associated Mechanisms
rs1035798	AGER	G	A	0.42	0.16	0.23	0.011	22.298 (2.027-245.257)	AGE/AGER signaling pathway
rs62357156	ITGA1	T	A	0.33	0.06	0.17	0.008	23.568 (2.249-246.990)	Cell-cell adhesion and Endothelial Cell migration
rs7125062	MMP1	T	C	0.40	0.09	0.25	0.034	6.336 (1.152-34.837)	VEGFA dependent angiogenesis
rs80067372	TNFSF12	G	A	0.44	0.19	0.23	0.038	5.986 (1.108-32.344)	TWEAK/Fn-14 pathway and pathogenic neovascularisation
rs4434138	STAB1	A	G	0.69	0.33	0.47	0.009	3.898 (1.405-10.817)	Internalization and trafficking of Advanced Glycated End (AGE) products
rs4234633		C	T	0.67	0.31	0.46	0.012	3.769 (1.334-10.649)	
rs10794640	IOP1	G	A	0.04	0.28	0.22	0.007	0.10 (0.000-0.293)	Modulation of HIF1A
rs9907595	PLXDC1	G	A	0.08	0.34	0.23	0.014	0.043 (0.003-0.535)	Fibrovascular membrane formation
rs2296123	PRKCQ	C	G	0.25	0.59	0.36	0.015	0.253 (0.084-0.766)	Pro-inflammatory gene expression
rs7483	GSTM3	C	T	0.21	0.44	0.31	0.033	0.308 (0.104-0.908)	Detoxification of Reactive Oxygen Species (ROS)
rs4698803	EGF	T	A	0.06	0.31	0.20	0.041	0.137 (0.02-0.919)	EGF-VEGF signaling pathway

All common variants listed above were checked in the BAM files and validated as real.

Besides this verification, four common variants were confirmed by the Illumina Microarray: rs7125062 T/C, rs9907595 G/A, rs7483 C/T and rs4698803 T/A and rs1035798 G/A was confirmed by OpenArray Technology (Life Technologies, CA, USA), a study previously performed at Genoinseq.

All other variants were validated by Sanger sequencing of two homozygous and two heterozygous individuals (Figure 36). Once again, all variant results were confirmed.

	Candidate Common Variants										
	AGER	ITGA1	MMP1	TNFSF12	STAB1		IOP1	PLXDC1	PRKCQ	GSTM3	EGF
	G/A	T/A	T/C	G/A	A/G	C/T	G/A	G/A	C/G	C/T	T/A
	rs1035798	rs62357156	rs7125062	rs80067372	rs4434138	rs4234633	rs10794640	rs9907595	rs2296123	rs7843	rs4698803
Validation Method	Open Array		Illumina Array					Illumina Array		Illumina Array	Illumina Array
BAM file check	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Variant Confirmation	✓		✓					✓		✓	✓
Validation Method		Sanger		Sanger	Sanger		Sanger		Sanger		
Homozygote		Ex30		Ex49 and Ex10	Ex42 and Ex46		Ex19		Ex50 and Ex2		
Heterozygote		Ex26 and Ex47		Ex47 and Ex9	Ex37 and Ex41		Ex29 and Ex45		Ex10 and Ex23		
										Protective Variants	
										Risk Variants	

Figure 36. Candidate common variants and validation procedures.
The validation methods, variants and results are presented. The homozygote and heterozygote samples used for Sanger sequencing are also reported. Risk variants are coloured in red and protective variants are coloured green.

3.6.1 Risk Variants

rs1035798 G/A variant in the AGER gene

The splice region variant in AGER was confirmed in the respective BAM file (Figure 37). The information regarding the variant is presented in Table 11.

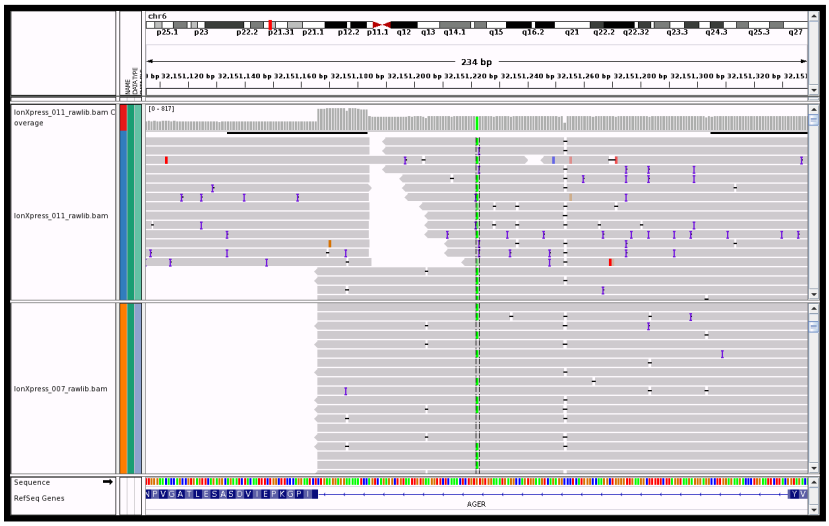


Figure 37. BAM file of sequenced data for rs1035798 G/A variant in AGER gene.
IonXpress_011 refers to Ex11 which is an altered homozygous (A/A) and IonXpress_007 refers to Ex23 which is heterozygous (G/A) for the variant.

Table 11. rs1035798 G7A genotype frequency in our study population (40 individuals). The type of variant, its impact severity and scaled CADD value is also represented.

rs ID	Gene	Impact	Impact severity	CADD scaled	Reference Homozygous (%)	Heterozygous (G/A) (%)	Altered Homozygous (A/A) (%)
rs1035798	AGER	Splice region	Medium	0,01	43%	53%	5%

AGER gene relation to Diabetic Retinopathy

The receptor for Advanced Glycated End products (AGER), encoded by this gene is a member of the cellular immunoglobin receptors. It is a receptor of multiple ligands that besides interacting with AGE, also interacts with molecules implicated in homeostasis and inflammation in certain diseases such as Diabetes and Alzheimer. AGE are non- enzymatic glycosilated proteins that accumulate in vascular tissues and promote inflammation. It has been documented that chronic inflammation alters the micro and macro vasculature leading to damaged organs (van den Oever *et al.,* 2010). The AGE/AGER signaling pathway has an important role in the regulation and expression of TNF-α, oxidative stress and endothelial dysfunction in Type 2 Diabetes. (Falcão *et al.,* 2010)

rs62357156 T/A variant in the ITGA1 gene

The intron variant in *ITGA1* was confirmed in the respective BAM file (Figure 38). The information regarding the variant is presented in Table 12.

Figure 38. BAM file of sequenced data for rs62357156 T/A variant in ITGA1 gene.
IonXpress_014 refers to Ex30 which is an altered homozygous (A/A) and IonXpress_015 refers to Ex47 which is heterozygous (T/A) for the variant.

Table 12. rs62357156 genotype frequency in our study population (40 individuals). The type of variant, its impact severity and scaled CADD value is also represented.

rs ID	Gene	Impact	Impact severity	CADD scaled	Reference Homozygous (%)	Heterozygous (T/A) (%)	Altered Homozygous (A/A) (%)
rs62357156	ITGA1	intron	LOW	4,75	58%	40%	3%

This variant was validated, by Sanger sequencing, of Ex30, a homozygote (A/A) and Ex26 and Ex47, 2 heterozygotes (T/A) (Figure 39).

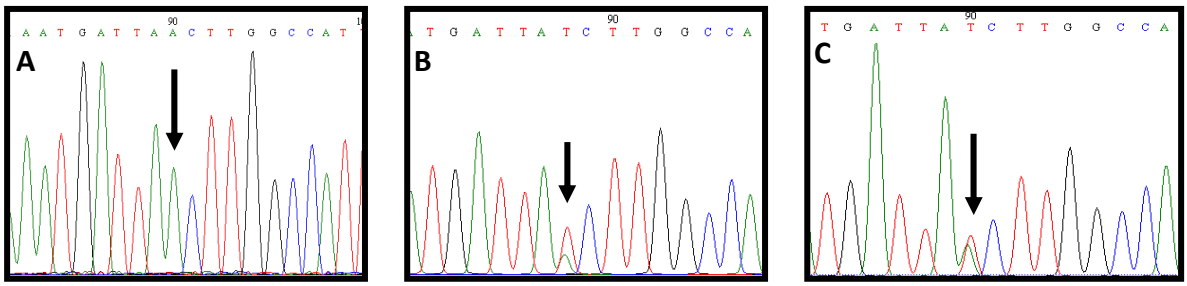


Figure 39. Chromatograms of the sequenced ITGA1 partial gene validating the findings of rs62357156 (T/A) variant.
A. Ex30 (A/A). B. Ex26 (T/A). C. Ex47 (T/A)

ITGA1 gene relation to Diabetic Retinopathy

The ITGA1 gene encodes the α subunit of the integrin receptors. The integrin receptor heterodimerizes with the β subunit to form a cell-surface receptor for collagen and laminin (Figure 40).

It is involved in cell-cell adhesion and may affect the mobility and proliferation of endothelial cells or inhibit their activities. The dysfunction of endothelial cells may promote the development of chronic inflammatory conditions and lead to vascular diseases. ITGA1, mediates physiological angiogenesis and is involved in the regulation of angiogenesis in pathological conditions, hence the possible relation to Diabetic Retinopathy (Park *et al.,* 2014)

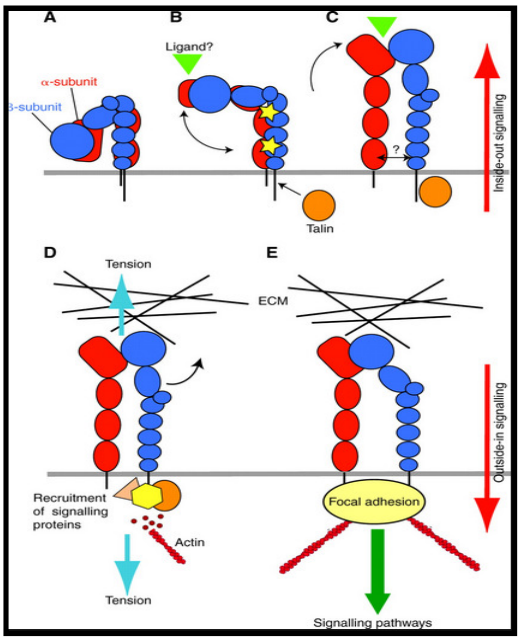


Figure 40. Integrin conformation-function relationships: a model.
A five-component model illustrating conformational changes that are associated with inside-out and outside-in integrin signalling. The α-subunit is in red and the β-subunit in blue. The figure shows the three major conformational states that have been identified so far: inactive (A), primed (B) and ligand bound (C) (ligand is represented by a green triangle), together with possible intermediate conformers. Panels A-C represent conformations that mediate inside-out signalling, and panels D and E, outside-in signalling (the direction is indicated by red arrows). Adapted from Askari *et al.,* 2009.

rs7125062 T/C variant in the MMP1 gene

The intron variant in MMP1 gene was confirmed in the respective BAM file (Figure 41). The information regarding the variant is presented in Table 13.

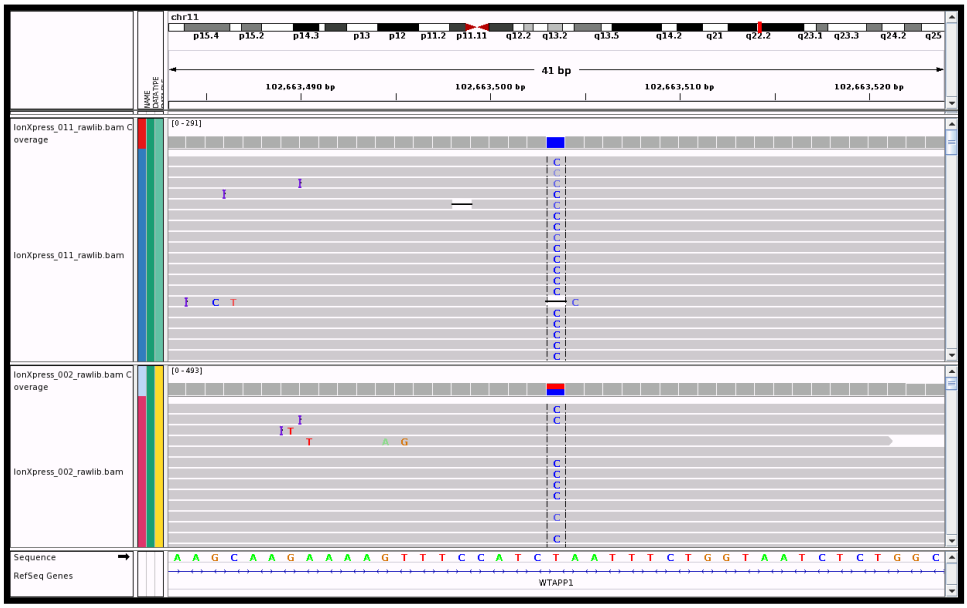


Figure 41. BAM file of sequenced data for rs7125062 T/A variant in MMP1 gene.
IonXpress_011 refers to Ex11 which is an altered homozygous (C/C) and IonXpress_002 refers to Ex34 which is heterozygous (T/C) for the variant.

Table 13. rs7125062 T/C genotype frequency in our study population (40 individuals). The type of variant, its impact severity and scaled CADD value is also represented.

rs ID	Gene	Impact	Impact severity	CADD scaled	Reference Homozygous (%)	Heterozygous (T/C) (%)	Altered Homozygous (C/C) (%)
rs7125062	MMP1	intron	LOW	6,77	55%	35%	10%

MMP1 gene relation to Diabetic Retinopathy

The angiogenic switch from normal to abnormal sprouting of new blood vessels from preexisting ones involves, in part, the proteolytic degradation of basement membranes and extracellular cell (ECM) components by Matrix Metalloproteinases (MMPs) (El-Asrar *et al.,* 2013).

MMPs are a family of zinc ion-binding Ca^{2+}-dependent neutral endopeptidases that act together with other enzymes to degrade most components of the ECM. Under steady state physiologic conditions, the expression of MMPs in most tissues is relatively low but recent studies have indicated that certain MMPs are generally up-regulated in several conditions that accompany angiogenesis and play an important role in the initiation of this process.

MM1 and MMP9 levels are increased dramatically in Proliferative Diabetic Retinopathy (PDR) and may be linked to the progression of the complication (El-Asrar *et al.,* 2013). MMP-1, an interstitial collagenase, acts directly on human microvessel endothelial cells as a pro-angiogenic signaling molecule and induces the expression of different subsets of pro-angiogenic genes (El-Asrar *et al.,* 2013). In addition to removing the physical barriers to new vessel growth, MMPs proteolytically release VEGF from the ECM-associated reservoirs, resulting in increased VEGF bioavailability and triggering the VEGF-driven angiogenic switch (El-Asrar *et al.,* 2013) (Figure 42).

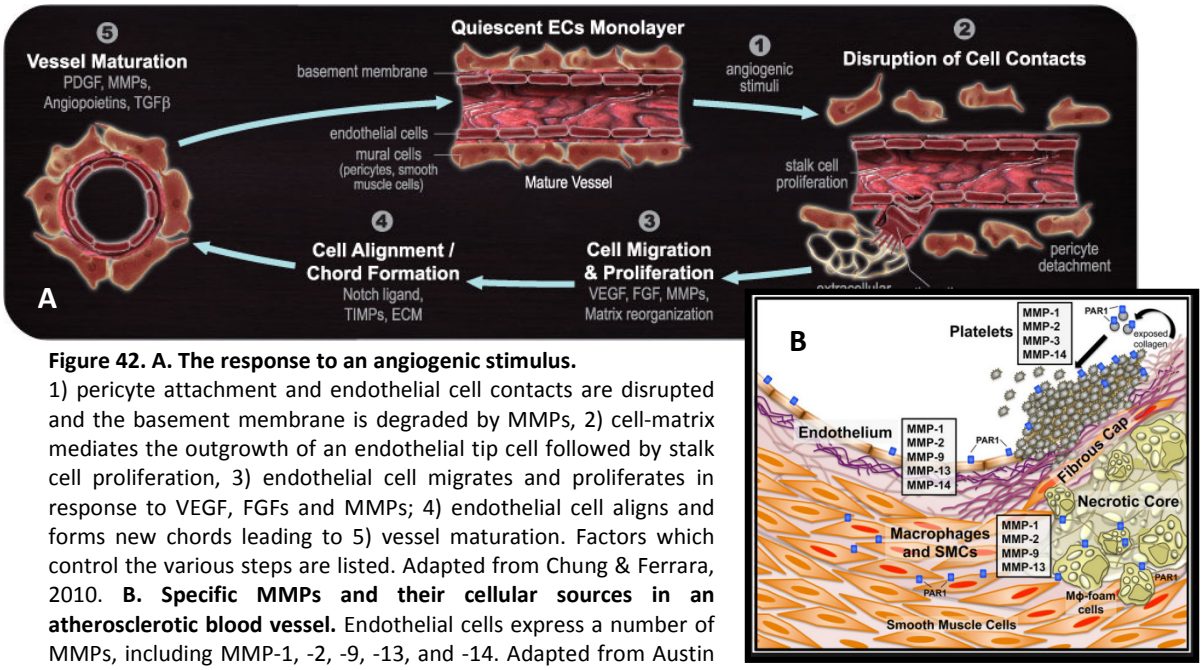


Figure 42. A. The response to an angiogenic stimulus.
1) pericyte attachment and endothelial cell contacts are disrupted and the basement membrane is degraded by MMPs, 2) cell-matrix mediates the outgrowth of an endothelial tip cell followed by stalk cell proliferation, 3) endothelial cell migrates and proliferates in response to VEGF, FGFs and MMPs; 4) endothelial cell aligns and forms new chords leading to 5) vessel maturation. Factors which control the various steps are listed. Adapted from Chung & Ferrara, 2010. **B. Specific MMPs and their cellular sources in an atherosclerotic blood vessel.** Endothelial cells express a number of MMPs, including MMP-1, -2, -9, -13, and -14. Adapted from Austin *et al.,* 2013.

rs80067372 G/A variant in theTNFSF12 gene

The intron variant in TNFSF12 gene was confirmed in the respective BAM file (Figure 43). The information regarding the variant is presented in Table 14.

Figure 43. BAM file of sequenced data for rs80067372 G/A variant in TNFSF12 gene.
IonXpress_008 refers to Ex24 which is an altered homozygous (A/A) and IonXpress_001 refers to Ex33 which is heterozygous (G/A) for the variant.

Table 14. rs80067372 genotype frequency in our study population (40 individuals). The type of variant, its impact severity and scaled CADD value is also represented.

rs ID	Gene	Impact	Impact severity	CADD scaled	Reference Homozygous (%)	Heterozygous (G/A) (%)	Altered Homozygous (A/A) (%)
rs80067372	TNFSF12	intron	LOW	8,21	40%	53%	8%

This variant was validated, by Sanger sequencing, of Ex49 and Ex10, 2 homozygotes (A/A) and Ex47 and Ex9, 2 heterozygotes (G/A) (Figure 44).

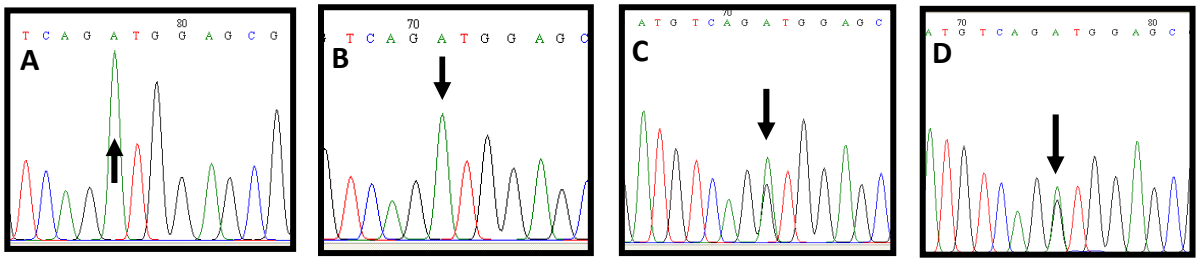


Figure 44. Chromatograms of the sequenced TNFSF12 partial gene validating the findings of rs80067372(G/A) variant.
A. Ex49 (A/A). B. Ex10 (A/A). C. Ex47 (G/A) and D. Ex9 (G/A).

TNFSF12 gene relation to Diabetic Retinopathy

The protein encoded by this gene is a cytokine that belongs to the tumor necrosis factor (TNF) ligand family. TNFSF12, (TWEAK) is a ligand for the receptor FN14/TWEAKR (Figure 45). Existing in the membrane form it is secreted and found to promote proliferation and migration of endothelial cells, and thus acts as a regulator of angiogenesis.

The TWEAK/Fn-14 (TNF-like weak inducer of apoptosis/ fibroblast growth factor inducible-14) pathway is involved in the pathological neovascularization in the retina, associated to Proliferative Diabetic Retinopathy. It seems that Hypoxia inducible factor-1α (*HIF1A*) is likely implicated in the upregulation of Fn14. (Ameri *et al.,* 2014)

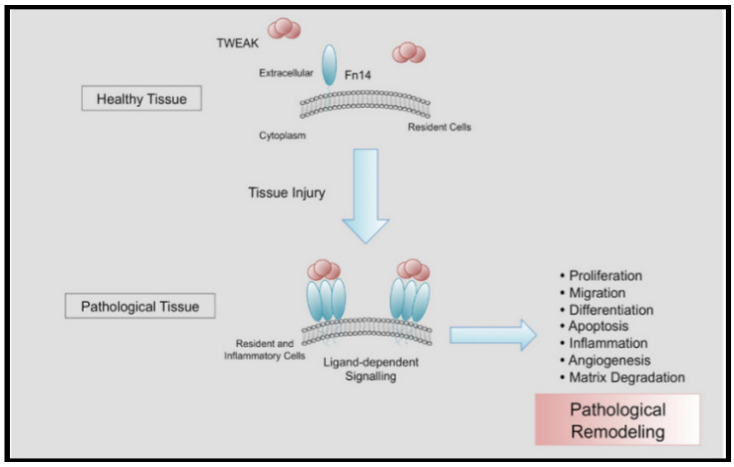


Figure 45. Pathological actions of TWEAK/Fn14 interaction.
Fn14 is almost absent in healthy tissues. After injury, Fn14 is upregulated, facilitating the interaction with its ligand TWEAK and its trimerization. In a context of chronic injury, TWEAK/Fn14 interaction participates in pathological tissue remodeling, promoting proliferation, migration, differentiation, apoptosis, inflammation, angiogenesis, and matrix degradation. Adapted from Ameri *et al,* 2014.

rs4434138 A/G and rs4234633 (C/T) variant in STAB1 gene

The non-synonymous coding and splice region variants in STAB1 gene were confirmed in the respective BAM file (Figure 46). The information regarding the variant is presented in Table 15.

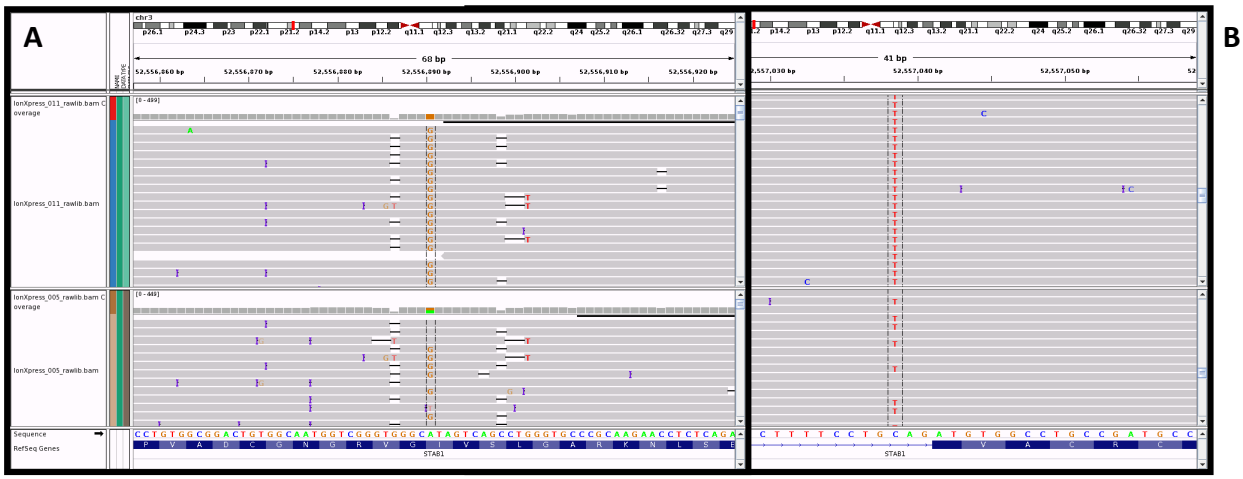


Figure 46. BAM file of sequenced data for rs4434138 A/G variant (A) and rs4234633 C/T (B) in STAB1 gene.
IonXpress_011 refers to Ex43 which is an altered homozygous (G/G) and (T/T) and IonXpress_005 refers to Ex37 which is heterozygous (A/G) and (C/T) for the variants, correspondingly.

Table 15. rs4434138 and rs4234633 genotype frequencies in our study population (40 individuals).
The type of variant, its impact severity and scaled CADD value is also represented.

rs ID	Gene	Impact	Impact severity	CADD scaled	Reference Homozygous (%)	Heterozygous (%)	Altered Homozygous (%)
rs4434138	STAB1	non_syn_coding	MED	11,31	25%	28%	48%
rs4234633		splice_region	MED	8,91	25%	25%	50%

These variants were validated, by Sanger sequencing, of Ex42 and Ex46, 2 homozygotes (G/G) for rs4434138 and (T/T) and (C/T) for rs4234633, respectively and Ex37 and Ex41, 2 heterozygotes (A/G) for rs4434138 and (C/T) for rs4234633. These variants seem to be in linkage in all the samples sequenced except for Ex46 which is homozygous for one variant and heterozygous for the other (Figure 47).

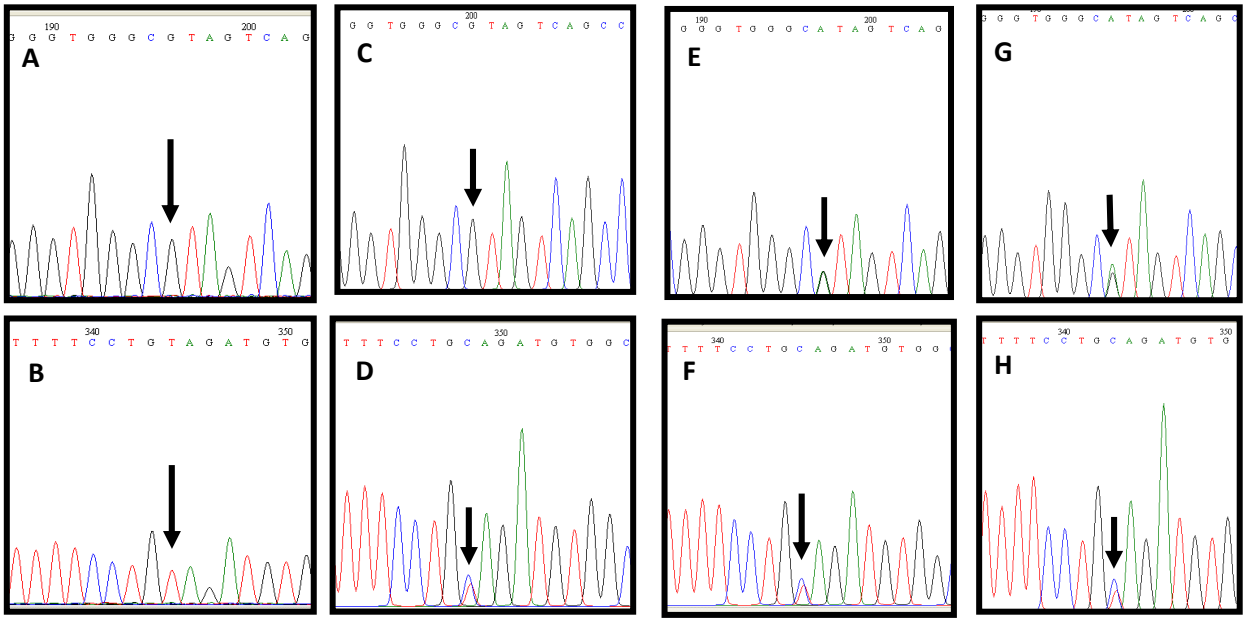


Figure 47. Chromatograms of the sequenced STAB1 partial gene validating the findings of rs4434138 (A/G) and rs4234633 (C/T) variants.
A and B. Ex42 (G/G) and (T/T). C. and D. Ex46 (G/G) and (C/T). E. and F. Ex37 (A/G) and (C/T) and G. and H. Ex41 (A/G) and (C/T).

STAB1 has been classified as a FrequentLy mutAted GeneS (FLAGS) but was considered relevant for further study. The two variants revealed having a Medium severity and appeared in the list of candidate common variants because it was present in one of the groups (cases).

STAB1 gene relation to Diabetic Retinopathy

Stabilin1 (STAB1), also known as Fasciclin EGF-like, laminin-type EGF-like, and Link domain-containing scavenger receptor-1 (FEEL1) is a type-1 transmembrane receptor that mediates endocytis and phagocytic clearance of "unwanted-self" components (Figure 48). It is expressed by the liver and vascular tissues, recognizes Advanced Glycation End products (AGEs) and may have a significant role in its' elimination from circulation as well as in the development of diabetic vascular complications (Tamura *et al.,* 2003; *Kzhyshkowska et al., 2006; Kzhyshkowska et al., 2010*).

The matrix cellular glycoprotein SPARC (secreted protein acidic and rich in cysteine) is a soluble, non-structural component of extracellular matrix (ECM) implicated in developmental processes, tissue remodeling, angiogenesis, and wound healing. It exhibits its multiple biological functions by active modulation of ECM organization, binding to growth factors, and induction of an anti-adhesive state in different cell types (*Kzhyshkowska et al., 2006).* It modulates angiogenesis by multiple mechanisms including the promotion of endothelial cell deadhesion, modulation of vascular ECM, and enhancement of the permeability of endothelial monolayers.

Both SPARC and stabilin-1 are expressed in normal adult tissues and under pathological conditions characterized by exuberant ECM production and/or cellular turnover.

Stabilin-1 is expressed by cells specialized in the clearance of “unwanted self” and in the maintenance of tissue homeostasis and it has been demonstrated that it internalizes and targets SPARC to an endosomal pathway for lysossomal degradation (*Kzhyshkowska et al., 2006).*

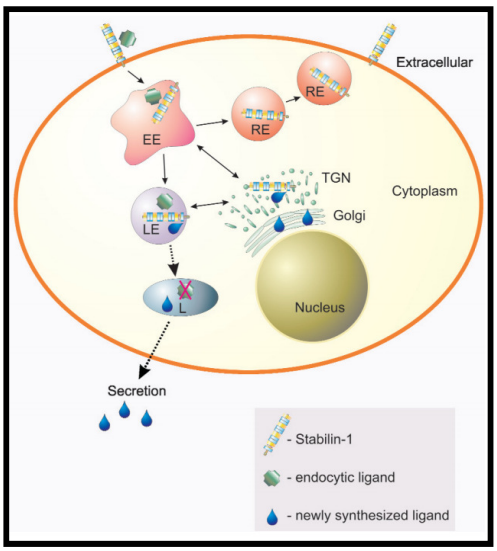


Figure 48. Schematic representation of stabilin-1 trafficking pathways.
Upon binding to surface expressed stabilin-1, ligand is internalized and delivered to early/sorting endosomes (EE). A portion of the ligand-free receptor can recycle back to the cell surface via recycling endosomes (RE). Stabilin-1 delivers its ligand to late endosomes (LE). Endocytic ligands (green) are further degraded in lysosomes. Adapted from *Kzhyshkowska et al., 2006.*

3.4.1 Protective Variants

rs10794640 G/A variant in the NARFL gene (IOP1)

The non-synonymous coding variant in NARFL gene was confirmed in the respective BAM file (Figure 49). The information regarding the variant is presented in Table 16.

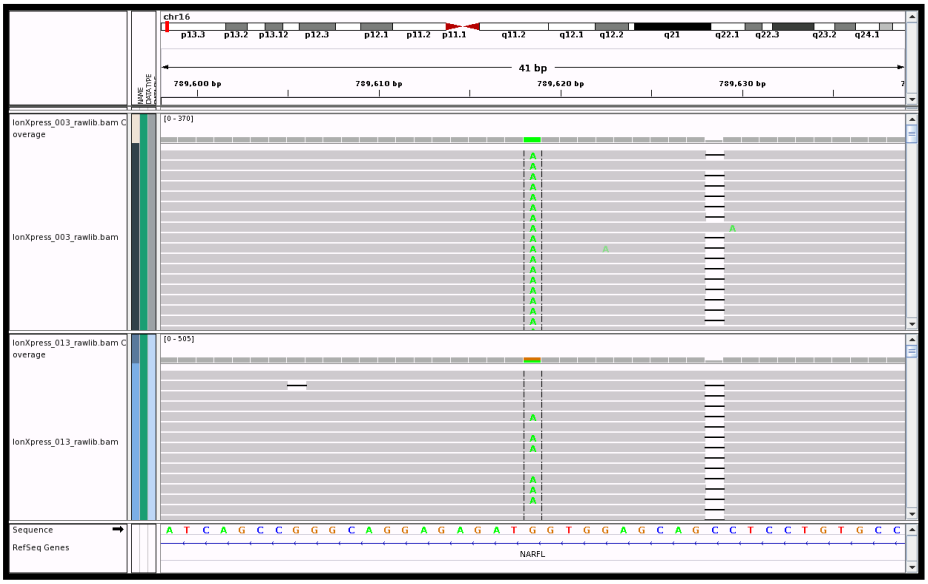


Figure 49. BAM file of sequenced data for rs10794640 G/A variant in IOP1 gene.
IonXpress_003 refers to Ex19 which is an altered homozygous (A/A) and IonXpress_013 refers to Ex29 which is heterozygous (G/A) for the variant.

Table 16. rs10794640 genotype frequency in our study population (40 individuals). The type of variant, its impact severity and scaled CADD value is also represented.

rs ID	Gene	Impact	Impact severity	CADD scaled	Reference Homozygous (%)	Heterozygous (G/A) (%)	Altered Homozygous (A/A) (%)
rs10794640	NARFL	non_syn_coding	MED	1,06	75%	23%	3%

This variant was validated, by Sanger sequencing, of Ex19, a homozygote (A/A) and Ex29 and Ex45, 2 heterozygotes (G/A) for the variant (Figure 50).

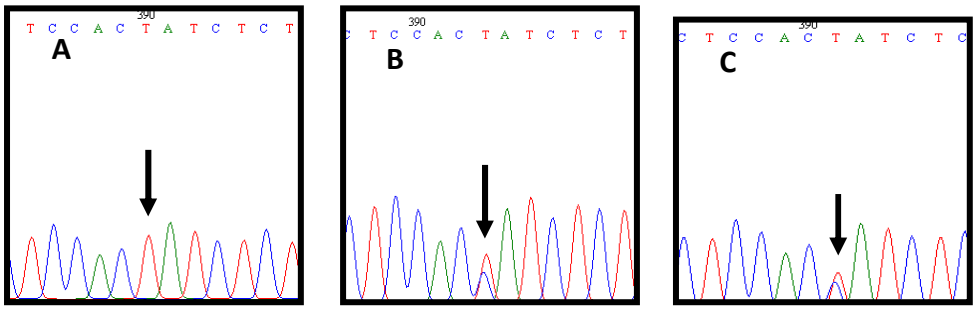


Figure 50. Chromatograms of the sequenced NARFL partial gene validating the findings of rs10794640 (G/A) variant.
A. Ex19 (A/A). B. Ex29 (G/A) and C. Ex45 (G/A). This gene is in reverse strand (-1).

NARFL gene relation to Diabetic Retinopathy.

NARFL (Nuclear Prelamin A Recognition Factor-Like) is a protein coding gene also known as IOP1 (iron-only hydrogenase-like 1). Evidence has shown that IOP1 is a component of the protein network that regulates HIF-1α in cells. (Huang *et al.,* 2007).

Transcription factor HIF-1 (hypoxia-inducible factor-1) plays an important role in cellular response to systemic oxygen levels and a central means by which cells respond to low oxygen tension is through the activation of this transcription factor. It is a heterodimeric complex consisting of α and β subunits, being the α subunits, the primary oxygen-responsive components of the HIF complex.

Under normoxic conditions, HIF-1α is targeted for rapid degradation by the ubiquitin–proteasome pathway. Under hypoxic conditions, this degradation is inhibited, thereby leading to the stabilization and activation of HIF-1α (Figure 51). HIF-1 α heterodimerizes with the β subunit and forms a complex that binds to promoters and enhancers of a multitude of genes involved in cellular, local and systemic responses to hypoxia. These include genes encoding proteins involved in angiogenesis, such as vascular endothelial growth factor (VEGF) (Huang *et al.,* 2007). Studies have demonstrated that NARFL (IOP1) knockdown up-regulates HIF-1α protein levels and augments hypoxia-induced target gene expression. It seems NARFL maintains HIF-1α mRNA levels at low steady-state levels.

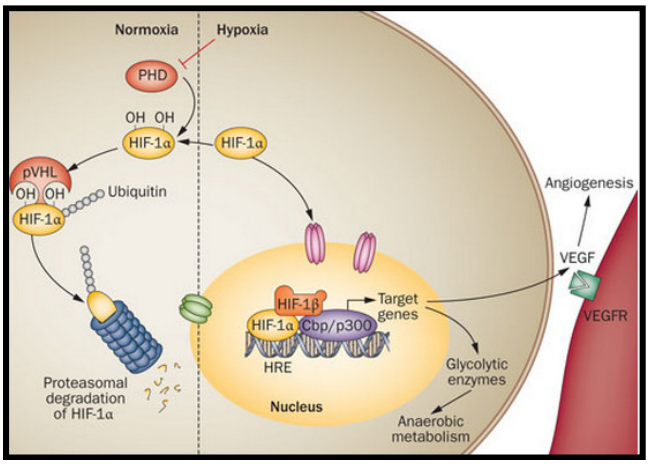


Figure 51. In normoxia, the cellular oxygen sensors (PHDs) hydroxylate HIF-1α, leading to its proteosomal degradation mediated by pVHL, an E3 ubiquitin ligase.
During hypoxia, HIF-1α is not ubiquitinylated or degraded, and acts as a transcription factor that binds to Hypoxia Responsive Elements (HREs) to induce expression of a plethora of target genes, including VEGF, which promotes angiogenesis. Abbreviations: HIF, hypoxia-inducible transcription factor; HRE, hypoxia response element; PHD, prolyl hydroxylase; pVHL, von Hippel–Lindau protein; VEGF, vascular endothelial growth factor; VEGFR, VEGF receptor. Adapted from Maes *et al.,* 2012.

rs9907595 A/G variant in PLXDC1 gene

The intron variant in PXLDC1 gene was confirmed in the respective BAM file (Figure 52). The information regarding the variant is presented in Table 17.

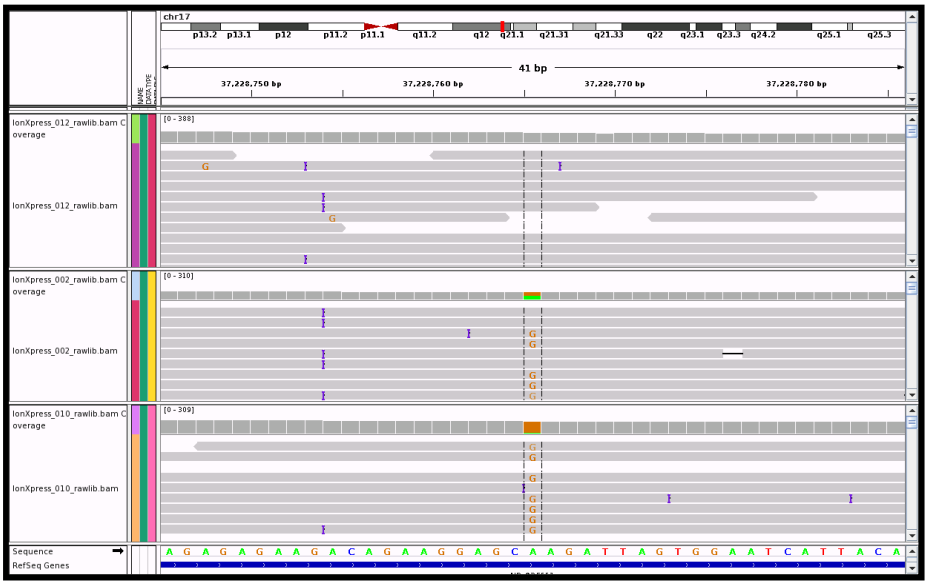


Figure 52. BAM file of sequenced data for rs9907595 A/G variant in PLXDC1 gene.
IonXpress_012 refers to Ex28 which is a homozygous (A/A), IonXpress_002 refers to Ex18 which is heterozygous (A/G) and IonXpress_010 refers to Ex26 which is homozygous (G/G) for the variant. In this case, the A is the minor allele.

Table 17. rs9907595 genotype frequency in our study population (40 individuals). The type of variant, its impact severity and scaled cadd value is also represented. The A allele is the minor allele.

rs ID	Gene	Impact	Impact severity	CADD scaled	Reference Homozygous (%)	Heterozygous (A/G) (%)	Altered Homozygous (G/G) (%)
rs9907595	PLXDC1	intron	LOW	13,94	3%	33%	65%

PLXDC1 gene relation to Diabetic Retinopathy

Much of the retinal damage that characterizes advanced Proliferative Diabetic Retinopathy results from retinal neovascularization (Yamaji *et al.,* 2008; El-Asrar *et al.,* 2013). When the newly formed vessels are associated with fibrous proliferations forming fibrovascular membranes (FVMs), traction retinal detachments can develop, resulting in potentially severe vision loss. Human retinal neovascularization exploits the same genes used by the developing normal endothelial cells and is regulated by a balance between stimulators and inhibitors of angiogenesis, thought to tightly control the normally quiescent capillary vasculature. When this balance is upset, as in PDR, capillary endothelial cells proliferate. (Yamaji *et al.,* 2008).

Plexin domain-containing 1 (PLXDC1), also known as TEM7 (Tumor Endothelial Marker 7), may play a significant role in the proliferation and maintenance of neovascular endothelial cells in the FVMs. It has been demonstrated (Yamaji *et al.,* 2008) that the TEM7 expression is specifically enhanced in neovascular endothelial cells in the FVMs associated with PDR and its' presence further supports the idea that they are markers of neoangiogenesis (Figure 53).

In addition to the biological interest of the TEM7 gene, the presence of a secreted form of TEM7 in FVMs raises a clinically interesting possibility for the development of novel serum markers that can be used to predict retinal angiogenic activity, which may reflect disease severity in patients with PDR. TEM7 may also be a molecular target for new diagnostic and therapeutic strategies for PDR.

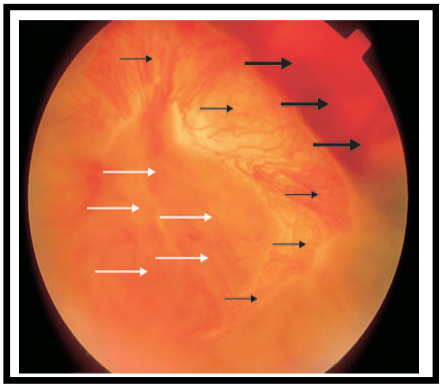


Figure 53. Severe proliferative diabetic retinopathy with vitreous haemorrhage (large black arrows), a huge fibrovascular membrane (small black arrows) that causes traction retinal detachment (white arrows).
Adapted from Gotzaridis *et al.,* 2004.

rs2296123 C/G variant in the PRKCQ gene

The intron variant in PRKCQ gene was confirmed in the respective BAM file (Figure 54). The information regarding the variant is presented in Table 18.

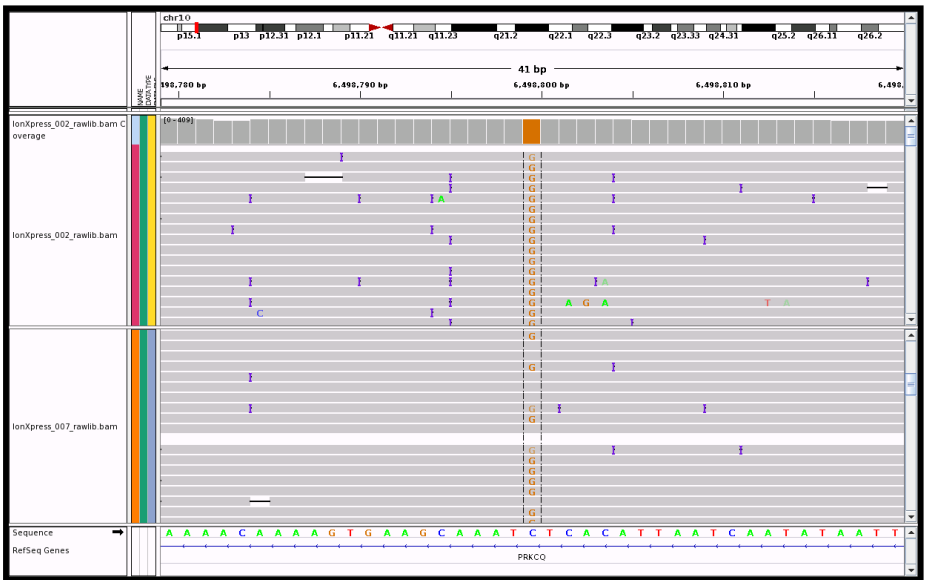


Figure 54. BAM file of sequenced data for rs2296123 C/G variant in PRKCQ gene.
IonXpress_002 refers to Ex50 which is an altered homozygous (G/G) and IonXpress_007 refers to Ex7 which is heterozygous (C/G) for the variant.

Table 18. rs2296123 genotype frequency in our study population (40 individuals). The type of variant, its impact severity and scaled cadd value is also represented.

rs ID	Gene	Impact	Impact severity	CADD scaled	Reference Homozygous (%)	Heterozygous (C/G) (%)	Altered Homozygous (G/G) (%)
rs2296123	PRKCQ	intron	LOW	8,58	40%	45%	15%

This variant was validated, by Sanger sequencing, of Ex50 and Ex2, 2 homozygotes (G/G) and Ex10 and Ex23, 2 heterozygotes (C/G) for the variant (Figure 55).

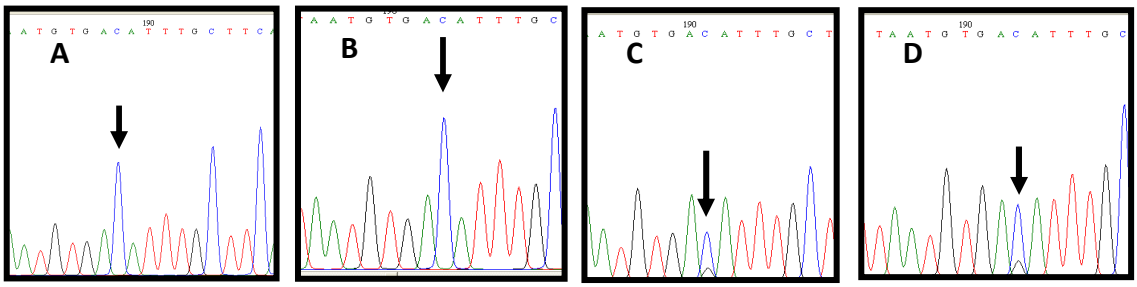


Figure 55. Chromatograms of the sequenced PRKCQ partial gene validating the findings of rs2296123 (C/G) variant.
A. Ex50 (G/G). B. Ex2 (G/G). C. Ex10 (C/G) and D. Ex23 (C/G). This gene is in reverse strand (-1).

PRKCQ gene relation to Diabetic Retinopathy

Protein Kinase C, ζ is encoded by the PRKQ gene and is one of the members of the Protein Kinase C family and reveals being necessary for the activation of T cells. It is a kinase protein that is independent of Ca^{2+} and requires Diacylglycerol (DAG) for the activation of NF-KB and AP-1. It may have an important role in the binding of the T signaling receptors complex with the concomitant activation of transcription factors. The increase in NF-KB induces the expression of pro-inflammatory genes. (Heo *et al.*, 2011)

rs7483 C/T variant in GSTM3 gene

The non-synonymous coding variant in GSTM3 gene was confirmed in the respective BAM file (Figure 56). The information regarding the variant is presented in Table 19.

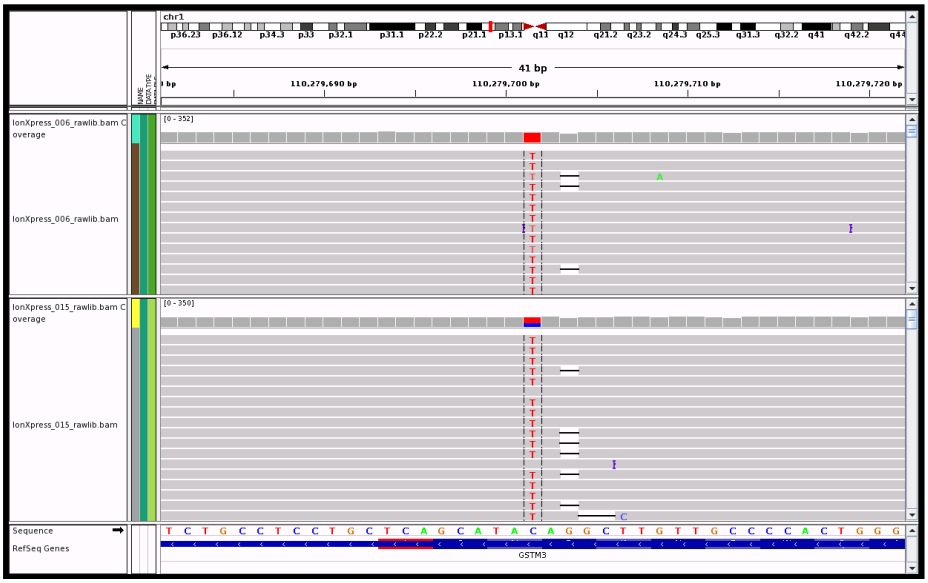


Figure 56. BAM file of sequenced data for rs7843 C/T variant in GSTM3 gene.
IonXpress_006 refers to Ex38 which is an altered homozygous (T/T) and IonXpress_015 refers to Ex47 which is heterozygous (C/T) for the variant.

Table 19. rs7843 genotype frequency in our study population (40 individuals). The type of variant, its impact severity and scaled CADD value is also represented.

rs ID	Gene	Impact	Impact severity	CADD scaled	Reference Homozygous (%)	Heterozygous (C/T) (%)	Altered Homozygous (T/T) (%)
rs7483	GSTM3	non_syn_coding	MED	0,01	53%	35%	13%

GSTM3 gene relation to Diabetic Retinopathy

Oxidative stress is considered as one of the crucial contributors in the pathogenesis of diabetic retinopathy, but oxidative stress appears to be highly interrelated with other biochemical imbalances that lead to structural and functional changes and accelerated loss of capillary cells in the retinal microvasculature and, ultimately, pathological evidence of the disease. (Madsen-Bouterse and Kowluru, 2008). This gene encodes a glutathione S-transferase that belongs to the mu class which functions in the detoxification of electrophilic compounds, including carcinogens, therapeutic

drugs, environmental toxins and products of oxidative stress, by conjugation with glutathione.

rs4698803 A/T variant in the EGF gene

The non-synonymous coding variant in EGF gene was confirmed in the respective BAM file (Figure 57). The information regarding the variant is presented in Table 20.

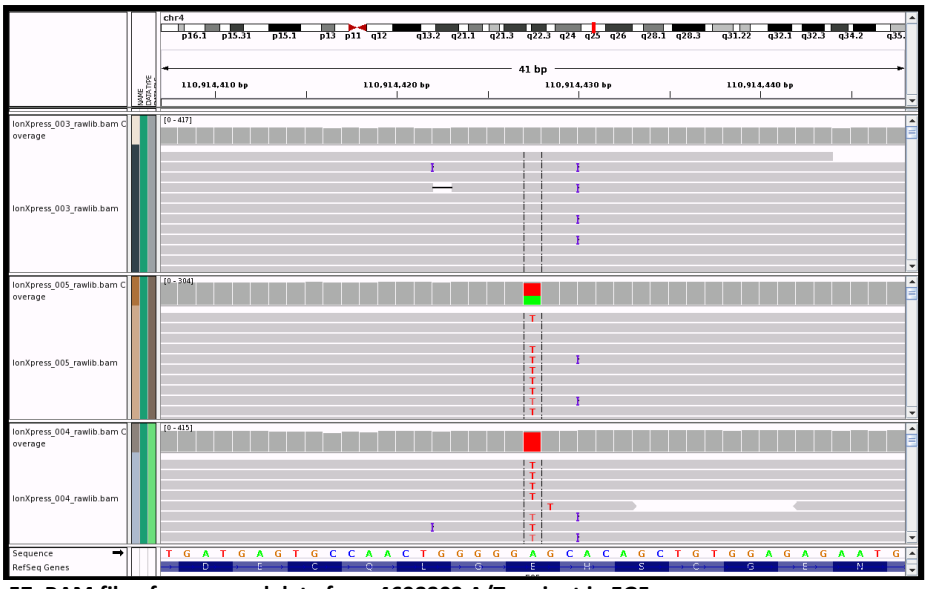


Figure 57. BAM file of sequenced data for rs4698803 A/T variant in EGF gene.
IonXpress_003 refers to Ex19 which is a homozygous (A/A), IonXpress_005 refers to Ex37 which is heterozygous (A/T) and IonXpress_004 refers to Ex4 which is homozygous (T/T) for the variant. In this case, the A is the minor allele.

Table 20. rs4698803 genotype frequency in our study population (40 individuals). The type of variant, its impact severity and scaled CADD value is also represented. The A is the minor allele.

rs ID	Gene	Impact	Impact severity	CADD scaled	Reference Homozygous (%)	Heterozygous (A/T) (%)	Altered Homozygous (T/T) (%)
rs4698803	EGF	non_syn_coding	MED	n.d	8%	18%	75%

EGF gene relation to Diabetic Retinopathy

The EGF gene encodes a member of the epidermal growth factor superfamily. This protein acts a potent mitogenic factor that plays an important role in the growth, proliferation and differentiation of numerous cell types. It binds to a high affinity cell surface receptor, the Epidermal Growth Factor Receptor (EGFR) that after activation by mechanisms of homo/heterodimerization and autophosporylation, initiate a signaling

cascade leading to altered gene expression and augmented cellular proliferation. The EGF signaling pathway is thus activated and can interact with other processes, such as the VEGF pathway (Figure 58). Being closely related they share common downstream signaling pathways. It has been shown that EGF is one of the many growth factors that drive VEGF expression (Tabernero, 2007).

The formation of new vessels begins with hemangioblasts (precursors of vascular endothelial cells) and ends with mature vasculature. It is a complex process that is regulated tightly by proangiogenic and antiangiogenic factors and involves autocrine and paracrine signaling. Thus mediation of VEGF expression is one of the main mechanisms by which tissue vasculature is controlled under normal physiologic conditions. Overexpression of EGFR and dysregulation or increased activity of EGFR signaling pathways are suggested mechanisms whereby the presence of EGFR may confer or promote a malignant phenotype. It is postulated that increased EGFR-mediated signaling may contribute to a cell moving into a state of continuous, unregulated cell proliferation.

Studies have found that treatment with EGF receptor inhibitors reduced retinal vascular leakage in diabetic mice. (Sugimoto *et al.,* 2013). These findings provide unique insight into the role of insulin signaling in mediating retinal effects in diabetes and the role of EGF signaling in the pathogenesis of Diabetic Retinopathy.

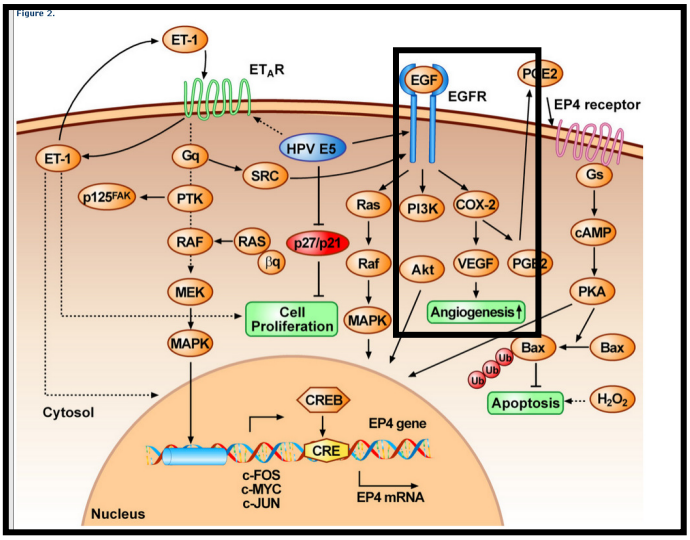


Figure 58. Activation of EGF-R and the downstream Ras-Raf-MAP kinase pathway or PI3K-Akt pathway leads to altered cell proliferation, angiogenesis, and anti-apoptosis.
The last two functions are further enhanced by the E5-induced upregulation of COX-2 expression. COX-2 activates VEGF leading to angiogenesis. Dotted arrow shafts indicate uncertain pathways. Adapted from Venuti *et al.,* 2011.

After variant validation and literature review, a table with all 40 samples, the candidate common variants and associated genotypes was constructed (Figure 59).

		Genes Comuns										
		AGER	ITGA1	MMP1	TNFSF12	STAB1		IOP1	PLXDC1	PRKCQ	GSTM3	EGF
		G/A	T/A	T/C	G/A	A/G	C/T	G/A	G/A	C/G	C/T	T/A
	ETDR value	rs1035798	rs62357156	rs7125062	rs80067372	rs4434138	rs4234633	rs10794640	rs9907595	rs2296123	rs7843	rs4698803
Ex1	w/out DR	G/G	T/T	T/T	G/A	A/G	C/T	G/A	G/G	G/G	C/T	T/T
Ex13	w/out DR	G/G	T/T	T/T	G/G	A/A	C/C	G/A	G/G	C/C	C/C	T/T
Ex19	w/out DR	G/G	T/T	T/T	G/G	A/G	C/T	A/A	A/G	C/G	T/T	A/A
Ex21	w/out DR	G/G	T/T	T/C	G/A	A/A	C/C	G/G	G/G	C/G	C/C	A/T
Ex28	w/out DR	G/A	T/A	T/T	G/G	A/G	C/T	G/G	A/A	G/G	C/C	A/T
Ex35	w/out DR	G/G	T/T	T/T	G/G	A/A	C/C	G/A	G/G	C/G	C/C	T/T
Ex36	w/out DR	G/G	T/T	T/T	G/A	A/A	C/C	G/G	A/G	C/G	C/T	T/T
Ex38	w/out DR	G/A	T/T	T/C	G/G	A/G	C/T	G/G	A/G	C/G	T/T	T/T
Ex44	w/out DR	G/G	T/T	T/T	G/G	A/A	C/C	G/G	G/G	C/G	C/C	T/T
Ex50	w/out DR	G/A	T/T	T/C	G/G	A/G	C/T	G/G	A/G	G/G	C/T	A/A
Ex6	w/out DR	G/G	T/T	T/T	G/G	G/G	T/T	G/G	A/G	C/G	T/T	A/T
Ex16	w/out DR	G/A	T/T	T/T	G/A	A/G	C/T	G/A	A/G	G/G	C/C	T/T
Ex18	w/out DR	G/G	T/T	T/T	G/A	A/G	C/T	G/A	A/G	G/G	C/C	T/T
Ex27	w/out DR	G/A	T/A	T/T	G/A	./.	C/C	G/G	G/G	C/G	T/T	T/T
Ex29	w/out DR	G/G	T/T	T/T	G/G	A/A	C/C	G/A	A/G	C/C	T/T	A/A
Ex45	w/out DR	G/G	T/T	T/T	G/G	A/G	C/T	G/A	A/G	C/G	C/T	A/T
Ex39	ETDR = 20	G/G	T/T	T/C	G/A	A/G	C/T	G/G	G/G	C/G	C/C	T/T
Ex40	ETDR = 20	G/A	T/T	T/C	G/A	A/G	C/T	G/G	G/G	C/C	C/T	T/T
Ex42	ETDR = 25	G/G	T/A	T/T	G/G	G/G	T/T	G/G	G/G	C/G	C/C	T/T
Ex11	ETDR = 35	A/A	T/A	C/C	G/A	G/G	T/T	G/G	G/G	C/C	C/C	T/T
Ex12	ETDR = 35	G/G	T/A	C/C	G/A	G/G	T/T	G/G	G/G	C/C	C/T	T/T
Ex24	ETDR = 35	G/A	T/T	T/T	A/A	A/A	C/C	G/G	G/G	C/C	C/T	A/T
Ex26	ETDR = 35	G/A	T/A	T/T	G/G	A/G	C/T	G/G	G/G	C/G	C/C	T/T
Ex30	ETDR = 35	G/A	A/A	T/C	G/A	G/G	T/T	G/G	G/G	C/G	C/T	T/T
Ex4	ETDR = 35	G/G	T/A	T/C	G/G	A/G	C/T	G/G	G/G	C/C	C/T	T/T
Ex46	ETDR = 35	G/A	T/T	T/T	G/A	G/G	C/T	G/G	G/G	C/C	C/C	T/T
Ex47	ETDR = 35	G/A	T/A	T/T	G/A	A/G	C/T	G/G	G/G	C/C	C/T	T/T
Ex9	ETDR = 35	G/A	T/A	T/C	G/A	A/A	C/C	G/G	G/G	C/C	C/C	T/T
Ex17	ETDR = 43	G/A	T/A	T/C	G/A	G/G	T/T	G/G	A/G	C/C	C/C	T/T
Ex2	ETDR = 43	G/A	T/A	T/T	G/A	A/G	C/T	G/G	G/G	G/G	C/C	T/T
Ex33	ETDR = 43	G/A	T/A	T/T	G/A	G/G	T/T	G/G	A/G	C/G	C/C	T/T
Ex34	ETDR = 43	G/G	T/A	T/C	G/G	A/G	C/T	G/G	G/G	C/C	C/T	T/T
Ex37	ETDR = 43	G/G	T/T	C/C	G/A	A/G	C/T	G/G	A/G	C/G	C/C	A/T
Ex41	ETDR = 43	G/A	T/T	T/T	G/A	A/G	C/T	G/G	G/G	C/C	C/C	T/T
Ex43	ETDR = 47	G/A	T/T	T/T	G/A	G/G	T/T	G/G	A/G	C/G	C/C	T/T
Ex48	ETDR = 53	A/A	T/A	C/C	G/G	G/G	T/T	G/G	G/G	C/C	C/C	T/T
Ex49	ETDR = 53	G/A	T/A	T/C	A/A	G/G	T/T	G/G	G/G	C/C	C/T	T/T
Ex10	ETDR = 90	G/A	T/T	T/C	A/A	G/G	T/T	G/A	G/G	C/G	C/C	T/T
Ex23	ETDR = 90	G/A	T/A	T/C	G/G	A/G	C/T	G/G	G/G	C/G	C/T	T/T
Ex7	ETDR = 90	G/A	T/T	T/C	G/A	A/G	C/T	G/A	G/G	C/G	C/T	A/T

Protective Variants
Risk Variants
Altered Homozygous
Altered Heterozygous

Figure 59. Genotypes of all 40 patients for the 11 common variants grouped by: without DR (ETDR values ≤20) and with DR (ETDR values >20).
The protective variants are highlighted in green and the risk variants in red. The altered homozygous genotypes are in turquoise and the altered heterozygous genotypes in yellow.

The overall genotype distribution is quite clear with the highlight of two major groups: risk variants vs protective variants and T2D with DR vs T2D without DR. There seems to be a larger incidence of homozygotes in the risk variants and the question remains as if it would be possible to create genetic profiles that could explain the excess risk of developing Diabetic Retinopathy or even if there are certain genetic profiles that could explain why some individuals, although diagnosed with T2D for many years, still don't present signs of the microvascular complication. Clearly this would have to be applied to a larger population as well as a thorough study and determination regarding the real implication of these candidate variants as genetic markers for Diabetic Retinopathy.

Chapter 4

Conclusions and Future Work

4. Conclusions and Future Work

This study sought to identify candidate genetic markers that could explain the excess risk associated to the onset of Diabetic Retinopathy, a complication in which diabetes duration and glycemic control, are major contributors to development and progression.

The Ion Torrent technology and Whole Exome Sequencing protocol used revealed being an efficient procedure for the detection of variants with a potential relevance to the complication. In conclusion of all the technical work done in this study, we can clearly state that Whole Exome Sequencing is a valid and complete way of obtaining a large quantity of information in an extended group of genes. This information spans the coding region of the genome and thus increases the possibilities in finding variants, common and rare, that may be considered as candidate without prior knowledge of which genes to look for. The vast majority of variants encountered in the exomes and further studied were real and correctly called. For that, all the filtering procedures and in-house scripts used were essential to guarantee the quality and veracity of the variants identified. The variants identified in the exomes were all validated by other technologies.

The search for rare variants, led to the identification of 10 genes accumulating rare variants: *DMXL2, E2F8, DNASE1L2, MAML3, ADAMTS2, EP300, CASZ1, APCDD1L, S100A14* and *GPR142* that were potentially relevant and related to mechanisms and pathways responsible for the pathogenesis of the complication. Some of these genes seem to be linked to the presence of DR, whereas other genes seem to be linked to the progression of DR and were thus associated to more severe forms of Diabetic Retinopathy. It seems that aggravated DR (ETDR values near 90) is related to more rare variant accumulation and these seem related to specific genes.

The search for common variants led to the identification of 11 variants: rs1035798, rs62357156, rs7125062, rs80067372, rs4434138, rs4234633, rs10794640, rs9907595, rs2296123, rs7483 and rs4698803 in 10 genes: *AGER, ITGA1, MMP1, TNFSF12, STAB1, NARFL, PLXDC1, PRKCQ, GSTM3* and *EGF.* Of these, only *AGER* has been consistently found as associated to Diabetic Retinopathy. All the genes considered relevant are in some

way related to the mechanism and pathways documented in the literature as being associated to DR. In fact, literature review was the criteria for relevance determination of the genes. The mechanisms in which these variants seem to participate or have been documented as being related were: Advanced Glycated End (AGE) products trafficking and signaling pathway, fibrovascular membrane formation, EGF-VEGF signaling pathway, VEGFA-dependent angiogenesis, vascular assembly and morphogenesis and the Notch signalling pathway.

The rare variant accumulating gene approach identified several other genes with an elevated statistical significance, giving our population number, besides the explored E2F8, and DMXL2 genes. However, with the present knowledge it is difficult to associate or even understand the role of these genes and their genetic significance to DR.

This study identified new candidate variants/genes that may reveal useful, in the future, for the understanding of the complication.

As future work we propose that the genes that accumulate rare variants but for which there is still no association to the disease, be studied in detail and this gene wise burden approach be applied to the protective rare variant accumulating genes.

We also propose the extension of this study to a much larger population and that these genes and variants be functionally explored.

The results of this study open the possibility to the determination of genetic profiles that may determine the onset of DR in certain patients, be that of risk variants, which could explain why patients that have T2D for 5 or 6 years and a tight glycemic control, present DR or of protective variants, explaining the cases of T2D patients that have the complex disease for more than 20 years, do not seem to control their glycemic levels and still do not present DR.

This fundaments the notion that DR onset has environmental factors that may delay or even attenuate progression and development but genetic factors are also very preponderant.

Bibliography

Bibliography

Abhary, S., Hewitt, A., Burdon, K. & Craig, J. (2009). A Systematic Meta-Analysis of Genetic Association Studies for Diabetic Retinopathy. Diabetes, Vol 58, pp. 2137-2147.

Albrechtsen A, Grarup N, Li Y, Sparsø T, Tian G, Cao H, Jiang T, Kim SY, Korneliussen T, Li Q, Nie C, Wu R, Skotte L, Morris AP, Ladenvall C, Cauchi S, Stančáková A, Andersen G, Astrup A, Banasik K, Bennett AJ, Bolund L, Charpentier G, Chen Y, Dekker JM, Doney AS, Dorkhan M, Forsen T, Frayling TM, Groves CJ, Gui Y, Hallmans G, Hattersley AT, He K, Hitman GA, Holmkvist J, Huang S, Jiang H, Jin X, Justesen JM, Kristiansen K, Kuusisto J, Lajer M, Lantieri O, Li W, Liang H, Liao Q, Liu X, Ma T, Ma X, Manijak MP, Marre M, Mokrosiński J, Morris AD, Mu B, Nielsen AA, Nijpels G, Nilsson P, Palmer CN, Rayner NW, Renström F, Ribel-Madsen R, Robertson N, Rolandsson O, Rossing P, Schwartz TW; D.E.S.I.R. Study Group, Slagboom PE, Sterner M; DIAGRAM Consortium, Tang M, Tarnow L, Tuomi T, van't Riet E, van Leeuwen N, Varga TV, Vestmar MA, Walker M, Wang B, Wang Y, Wu H, Xi F, Yengo L, Yu C, Zhang X, Zhang J, Zhang Q, Zhang W, Zheng H, Zhou Y, Altshuler D, 't Hart LM, Franks PW, Balkau B, Froguel P, McCarthy MI, Laakso M, Groop L, Christensen C, Brandslund I, Lauritzen T, Witte DR, Linneberg A, Jørgensen T, Hansen T, Wang J, Nielsen R, Pedersen O. (2013). Exome sequencing-driven discovery of coding polymorphisms associated with common metabolic phenotypes. Diabetologia 56:298-310.

Ameri, H., Liu, H., Liu, R., Ha, Y., Paulucci-Holthauzen, A., Hu, S., Motamedi, M., Godley, B., Tilton, R. and Zhan, W. (2014). TWEAK/Fn14 Pathway Is a Novel Mediator of Retinal Neovascularization. Investigative Ophthalmology & Visual Science, Vol. 55, No. 2, 802-813

American Diabetes Association. 2014. Estimates of Diabetes and Its Burden in the Epidemiologic estimation methods. National Diabetes Statistics Report, pp.2009–2012.

Anomalies, M. and Brief, P. (2003). Type 2 Diabetes and Genetic Technology: A Policy Brief; Washington State Department of Health

Arboleda-Velasquez, JF, Primo V, Graham M, James A, Manent J, D'Amore PA (2014). Notch Signaling functions in retinal pericyte survival. Investigative Ophthalmology & Visual Science, 55: 5191-5199 -

Askari, J., Buckley, P., Mould, A. and Humphries, M. (2009). Linking integrin conformation to function. Journal of Cell Science, 122 (2), pp.165-170

Austin, K., Covic, L and Kulipulos, A. (2013). Matrix metalloproteases and PAR1 activation. Blood, Volume 121, Number 3 pp. 431-439

Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA and Shendure J (2011) Exome sequencing as a tool for Mendelian disease gene discovery. Nature Reviews GENETICS; Volume 12; 745-755

Barroso, I., Luan, .J, Middelberg, R., Harding, A,. Franks, P., Jakes, R., Clayton, D., Schafer, A., O'Rahilly, S. and Wareham, N. (2003). Candidate Gene Association Study in Type 2 Diabetes Indicates a Role for Genes Involved in β-Cell Function as Well as Insulin Action. PLoS Biology, Volume 1, Issue 1; pp 441-445

Beltramo, E and Porta, M. (2013). Pericyte loss in Diabetic Retinopathy: Mechanisms and Consequences. Current Medicinal Chemistry, 20, pp 3218-3225

Bloomgarden, Z. (2004). Type 2 Diabetes in the Young: The Evolving Epidemic. Diabetes Care, Volume 27, Number 4; pp. 998-1010.

Bonnefond A, Durand E, Sand O, De Graeve F, Gallina S, Busiah K, Lobbens S, Simon A, Bellanné-Chantelot C, Létourneau L, Scharfmann R, Delplanque J, Sladek R, Polak M, Vaxillaire M and Froguel P. (2010). Molecular Diagnosis of Neonatal Diabetes Mellitus Using Next-Generation Sequencing of the Whole Exome. PLoS ONE; Volume 5, Issue 10, e13630; pg.1-5

Bush, W. and Moore, J. (2012). Chapter 11: Genome-Wide Association Studies. PLOS Computational Biology, Vol 8, Issue 12, e1002822.

Charpentier MS, Christine KS, Amin NM, Dorr KM, Kushner EJ, Bautch VL, Taylor JM, Conlon FL. (2013). CASZ1 promotes vascular assembly and morphogenesis through the direct regulation of an EGFL7/RhoA-mediated pathway. Development Cell. 25(2):132-43.

Chen Y, Cunningham F, Rios D, McLaren WM, Smith J, Pritchard B, Spudich GM, Brent S, Kulesha E, Marin-Garcia P, Smedley D, Birney E, Flicek P. (2010). Ensembl variation resources. BMC Genomics. 11:293.

Chen, P., Ong, R., Tay, W.,Sim, X., Ali, M., Xu, H., Suo, C., Liu, J., Chia, K., Vithana, E., Young, T., Aung, T., Lim, W., Khor, C., Cheng, C., Wong, T., Teo, Y. and Tai, E. (2013). A Study Assessing the Association of Glycated Hemoglobin A1C (HbA1C) Associated Variants with HbA1C, Chronic Kidney Disease and Diabetic Retinopathy in Populations of Asian Ancestry. PLOSOne, Volume 8 Issue 11, e79767

Chung, A. & Ferrara N. (2010). The Extracellular Matrix & Angiogenesis: Role of the Extracellular Matrix in Developing Vessels and Tumor Angiogenesis. www.sabiosciences.com. Pathways, Issue #11.

Cooper, G., Stone, E., Asimenos, G., NISC Comparative Sequencing Program, Green, E., Batzoglou, S., and Sidow, A. (2005). Distribution and intensity of constraint in mammalian genomic sequence. Genome Research, 15(7):901-13

Cunningham, F., M. Ridwan Amode, Daniel Barrell, Kathryn Beal, Konstantinos Billis, Simon Brent, Denise Carvalho-Silva, Peter Clapham, Guy Coates, Stephen Fitzgerald, Laurent Gil, Carlos Garcín Girón, Leo Gordon, Thibaut Hourlier, Sarah E. Hunt, Sophie H. Janacek, Nathan Johnson, Thomas Juettemann, Andreas K. Kähäri, Stephen Keenan, Fergal J. Martin, Thomas Maurel, William McLaren, Daniel N. Murphy, Rishi Nag, Bert Overduin, Anne Parker, Mateus Patricio, Emily Perry, Miguel Pignatelli, Harpreet Singh Riat, Daniel Sheppard, Kieron Taylor, Anja Thormann, Alessandro Vullo, Steven P. Wilder, Amonida Zadissa, Bronwen L. Aken, Ewan Birney, Jennifer Harrow, Rhoda Kinsella, Matthieu Muffato, Magali Ruffier, Stephen M.J. Searle, Giulietta Spudich, Stephen J. Trevanion, Andy Yates, Daniel R. Zerbino and Paul Flicek (2015). Ensembl 2015. Nucleic Acids Research 43 Database issue:D662-D669

Danecek, P., Auton, A., Abecasis;G., Albers, C., Banks, E., DePristo, M., Handsaker, R., Lunter, G., Marth, G., Sherry, S., McVean, G., Durbin, R. and 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. Bioinformatics Applications Note, Vol. 27 no. 15, pages 2156–2158

Do, R., Kathiresan, S. and Abecasis, G. (2012). Exome sequencing and complex disease: practical aspects of rare variant association studies. Human Molecular Genetics, Vol. 21, Review Issue 1; R1-R9.

Doria, A. (2010). Genetics of Diabetes Complications. Current Diabetes Reports, 10 (6): 467-475.

Droumaguet, C., Balkau, B., Simon, D., Caces, E., Tichet, J., Charles, M., Eschwege, E and The DESIR Study group. (2006). Use of HbA1c in Predicting Progression to Diabetes in French Men and Women. Data from an Epidemiological Study on the Insulin Resistance Syndrome (DESIR). Diabetes Care, Volume 29, Number 7

Dunham, I. et al., (2012). An integrated encyclopedia of DNA elements in the human genome. Nature, 2012 Sep 6; 489(7414):57-74.

Eichler, E., Flint, J., Gibson, G., Kong, A., Leal, S., Moore, J and Nadeau, J. (2010). Missing Heritability and Strategies for finding the underlying causes of complex disease. Nat. Rev. Genet. 2010 Jun, 11(6): 446-450

El-Asrar, A., Mohammad, G., Nawaz, M., Siddiquei, M., Van den Eynde, K., Mousa, A., De Hertogh, G. and Opdenakker, G. (2013). Relationship between Vitreous Levels of Matrix Metalloproteinases and Vascular Endothelial Growth Factor in Proliferative Diabetic Retinopathy. PLOS ONE, Volume 8, Issue 12, e 85857

Estrada, K. (2013). Whole-exome sequencing of 4,000 samples identifies rare variants strongly associated with type 2 diabetes risk in Mexicans and Latinos. (Abstract) American Society of Human Genetics, Annual Meeting, Boston, October 22-26

Exome Variant Server, NHLBI GO Exome Sequencing Project (ESP), Seattle, WA

Falcão, M., Falcão-Reis, F. & Rocha-Sousa, A. (2010). Diabetic Retinopathy: Understanding Pathogenic Angiogenesis and Exploring its Tratment Options. the Open Circulation and Vascular Journal, 3, 30-42.

Fong, D., Aiello, L., Gardner, T., King, G., Blankenship, G., Cavallerano, J., Ferris, F., Klein, R. (2004). Retinopathy in Diabetes. Diabetes Care, Vol. 27, Supplement 1, S84-S87.

Frazer, K. Murray, S., Schork, N. & Topol, E. (2009). Human genetic variation and its contribution to complex traits. Nature reviews. Genetics, 10(4), pp.241–251.

Fredriksson R, Höglund PJ, Gloriam DE, Lagerström MC, Schiöth HB. (2003). Seven evolutionarily conserved human rhodopsin G protein-coupled receptors lacking close relatives. FEBS Lett. 2003 Nov 20;554(3):381-8

Freedman SJ, Sun ZY, Poy F, Kung AL, Livingston DM, Wagner G, Eck MJ. (2002). Structural basis for recruitment of CBP/p300 by hypoxia-inducible factor-1 alpha. Proc. Natl. Acad. Sci. U.S.A. 2002; 99(8);5367-72 .

Gaudet, M., Giulia, F., Beritognolo, I and Sabatti, M. (2009). Allele-Specific PCR in SNP Genotyping Single Nucleotide Polymorphisms Methods in Molecular Biology, Volume 578, 2009, pp 415-424

Ghosh, A., Shanafelt, T., Cimmino, A., Taccioli, C., Volinia, S., Liu, C., Calin, G., Crove, C., Chan, D., Giaccia, A., Secreto, C., Wellik, L., Lee, Y., Mukhopadhyay, D and Kay, N. (2009). Aberrant regulation of pVHL levels by microRNA promotes the HIF/VEGF axis in CLL B cells. Blood, 28. Volume 113, number 22, pp.5568-5574.

Gilissen, C., Hoischen, A., Brunner, H. and Veltman, J. (2011). Unlocking Mendelian disease using exome sequencing. Genome Biology, 12:228-238

Goldin A, Beckman JA, Schmidt AM, Creager MA. (2006). Advanced glycation end products: Sparking the development of diabetic vascular injury. Circulation, 114(6), pp.597–605.

Gotzaridis, E., Markou, A. and Gregor, Z. (2004).Management of Diabetic Retinopathy. An Overview. Hormones, 3 (2):92-99

Hammes. H., Lin, J., Renner, O., Shani, M., Lundqvist, A., Betsholtz, C., Brownlee, M and Deutsch, U. (2002). Pericytes and the Pathogenesis of Diabetic Retinopathy. Diabetes, Vol. 51:3107-3112.

Heo, F. , Fujiwara, K. and Abe, J. (2011). Disturbed-flow-mediated vascular reactive oxygen species induce endothelial dysfunction. Circ J. 2011;75(12):2722-30

Huang, J., Song, D., Flores, A., Zhao, Q., Mooney, S., Shaw, L., and Lee, F. (2007). IOP1, a novel hydrogenase-like protein that modulates hypoxia-inducible factor-1α activity. Biochem. J., 401:341-352.

Johansson S, Irgens H, Chudasama KK, Molnes J, Aerts J, Roque FS, Jonassen I, Levy S, Lima K, Knappskog PM, Bell GI, Molven A and Njølstad PR (2012). Exome Sequencing and Genetic Testing for MODY. PLoS ONE; Volume 7, Issue 5, e38050; pp.1-7

Jin, Q., Chen, H., Luo, A, Ding, F and Liu, Z. (2011). S100A14 Stimulates Cell Proliferation and Induces Cell Apptosis at Different Concentrations via Receptor for Advanced Glycation End Products (RAGE). PLOSOne, Vol, 6, Issue 4, e19375.

Kahn, C. (1994). Banting Lecture. Insulin Action, Diabetogenes, and the Cause of Type II Diabetes; DIABETES, vol 43; pg.1066-1084

Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2014). Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res. 42, D199–D205.

Kanehisa, M. and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 28, 27-30.

Kang, H. et al., (2010). Variance component model to account for sample structure in genome-wide association studies. Nature genetics, 42(4), pp.348–354.

Kang, H., Zhan, X. Sim, C. (2012). EPACTS : A flexible and efficient sequence-based genetic analysis software package. Biostatistics Dept, University of Michigan, Ann Arbor, Ann Arbor, MI., meeting 2012

Kiezun, A., Garimella, K., Do, R., Stitziel, N., Neale, B., McLaren, P., Gupta, N., Sklar, P., Sullivan, P., Moran, J., Hultman, C., Lichtenstein, P., Magnusson, P., Lehner, T., Shugart, Y., Price, A., de Bakker, P., Purcel, S and Sunyaev, S. (2012). Exome Sequencing and and the genetic basis of complex traits. Nature Genetics, volume 44 number 6, 623-630

Kishore , P. (2013). "Diabetes Mellitus (DM)" The Merck Manual for Health Care Professionals
http://www.merckmanuals.com/professional/endocrine_and_metabolic_disorders/diabetes_mellitus_and_disorders_of_carbohydrate_metabolism/diabetes_mellitus_dm.html#v988257 (Last full review/revision December 2012, Content last modified October 2013).

Koeleman BPC, Al-Ali A, Van der Laan SW and Asselbergs FW (2013). A Concise History of Genome-Wide Association Studies. Saudi Journal of Medicine & Medical Sciences; Vol.1, Issue 1; pg. 4-10

Kuo, J., Wong, T. and Rotter, J. (2014). Challenges in elucidating the genetics of diabetic retinopathy. JAMA Ophthalmology, 132(1):96-107.

Kzhyshkowska, J., Workman, G., Cardó -Vila, M., Arap, W., Pasqualini, R., Gratchev, A., Krusell, L., Goerdt, S., and Sage, E. (2006). Novel Function of Alternatively Activated Macrophages: Stabilin-1-Mediated Clearance of SPARC. The Journal of Immunology, 176: 5825–5832.

Kzhyshkowska, J. (2010). Multifunctional Receptor Stabilin-1 in Homeostasis and Disease. The Scientific World Journal, 10: 2039-2053.

Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, Maglott DR. (2014). ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 2014 Jan 1;42(1):D980-5.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics Applications Note, Vol. 25 no. 16, pages 2078–2079

Life Technologies Release (2012). Exome sequencing using the Ion Proton System. Proton Exome Sequencing Product Bulletin

Lin, S., Oyama, L., Nagase, T., Harigaya, K. and Kitagawa, M. (2002). Identification of new human mastermind proteins defines a family that consists of positive regulators for Notch signaling.Journal of Biological Chemistry, 277: 50612-50620.

Lohmueller, K., Sparsø T., Li, Q., Andersson, E., Korneliussen, T., Albrechtsen, A., Banasik K., Grarup, N., Hallgrimsdottir, I., Kiil, K., Kilpelainen, T., Krarup, N., Pers T., Sanchez, G., Hu Y., DeGiorgio, M., Jørgensen, T., Sandbæk, A., Lauritzen, T., Brunak S., Kristiansen, K., Li Y., Hansen, T., Wang, J., Nielsen, R. and Pedersen, R. (2013). Whole-Exome Sequencing of 2,000 Danish Individuals and the Role of Rare Coding Variants in Type 2 Diabetes. The American Journal of Human Genetics, 93: 1072-1086.

Lyssenko, V., Almgren, P., Anevsk,i D., Orho-Melander, M., Sjogren, M., Saloranta, C., Tuomi, T., Groop, L. and the Botnia Study Group. (2005). Genetic Prediction of Future Type 2 Diabetes. PLoS MEDICINE; Volume 2, Issue 12, e345; pg.1299-1308

Lyssenko, V. and Laakso, M. (2013). Genetic Screening for the Risk of Type 2 Diabetes: Worthless or valuable?. DIABETES CARE, Volume 36, Supplement 2, S120-S126

Madsen-Bouterse, S and Kowluru, R. (2008). Oxidative stress and diabetic retinopathy: Pathophysiological mechanisms and treatment perspectives. Reviews in Endocrine and Metabolic Disorders, Volume 9, Issue 4,pp.315-327.

Maes, C. Carmeliet, G and Schipani, E. (2012). Hypoxia-driven pathways in bone development, regeneration and disease Nature. Reviews. Rheumatology, 8, pp. 258-266

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA and Visscher PM. (2009) Finding the missing heritability of complex diseases"; Nature. 2009 October 8; 461(7265):747-753. DOI: 10.1038/nature08494

Marian, A. (2012). Molecular Genetic Studies of Complex Phenotypes. Translational Research, 159(2): 64–79. doi:10.1016/j.trsl.2011.08.001

Marth, G., Yu, F., Indap, A., Garimella, K., Gravel, S., Leong, W., Smith, C., Bainbridge, M., Blackwell, T., Bradley, X., Chen, Y., Challis, D., Clarke, L., Ball, E., Cibulskis, K., Cooper, D., Fulton, B., Hartl, C., Koboldt, D., Muzny, D., Smith, R., Sougnez, C., Stewart, C., Ward, A., Yu, J., Xue, Y., Altshuler, D., Bustamante, C., Clark, G., Daly, M., DePristo, M., Flicek, P., Gabriel, S., Mardis, E., Palotie, A., Gibbs, R. and the 1000 Genomes Project. (2011). The functional spectrum of low-frequency coding variation. Genome Biology, 12:R84

Martinez, R.(2013). Prevalence of Diabetes in the World, 2013. Health Intelligence. Nov 18.

Massi-Benedetti, M. (2002). The Cost of Diabetes Type II in Europe: The CODE-2 Study; Diabetologia, 45 (7):S1-S4.

Mathers, C. and Loncar, D. (2006). Projections of Global Mortality and Burden of Disease from 2002 to 2030. PLoS Medicine, November 2006, Volume 3, Issue 11, e442

McLaren, W. et al., (2010). Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics, 26(16), pp.2069–2070.

1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA (2012). An integrated map of genetic variation from 1,092 human genomes. Nature 491, 56–65

Morris, A., Voight, B. And Teslovich, T. (2012). Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. NATURE GENETICS, 2012 September; 44(9):981-990.

Motte, P., Deroanne, C., Colige, A., Lambert, V., Nusgens, B., Lapi, C., Rakic, J., Dubail, J. and Kesteloot, F. (2010). ADAMTS-2 functions as anti-angiogenic and anti-tumoral molecule independently of its catalytic activity. Cell Mol Life Sci. Vol 67 (4213-4232)

Ng, D. (2010). Human Genetics of Diabetic Retinopathy: Current Perspectives. Journal of Ophthalmology, Vol 2010, Article ID 172593.

Ola, M. & Nawaz, M. (2012). Cellular and Molecular Mechanism of Diabetic Retinopathy. Diabetic Retinopathy, Chapter 1. Dr. Mohammad Shamsul Ola (Ed.), ISBN: 978-953-51-0044-7,

Olokoba, A., Obateru, O and Olokoba, L. (2012). Type 2 Diabetes Mellitus: A review of Current Trends. Oman Medical Journal, Vol.27, No. 4:269-273.

Organização Nacional da Diabetes. (2014). Diabetes: Factos e Números 2014. Relatório Anual do Observatório Nacional da Diabetes

Ozturk, B., Bozkurt, B., Kerimoglu, H., Okka, M., Kamis, U. and Gunduz, K. (2009). Effect of serum cytokines and VEGF levels on diabetic retinopathy and macular thickness. Molecular Vision, 15:1906-1914

Paila, U., Chapman, B., Kirchner, R., Quinlan, A. (2013). GEMINI: Integrative Exploration of Genetic Variation and Genome Annotations. PLoS Comput Biol 9(7): e1003153.

Panoutsopoulou, K., Tachmazidou, I. & Zeggini, E., (2013). In search of low-frequency and rare variants affecting complex traits. Human Molecular Genetics, 22(R1), pp.16–21.

Park,J., Park,Y., Jung, W., Lee, J. and Lee, J. (2014). Microarray Analysis for Genes Associated with Angiogenesis in Diabetic OLETF Keratocytes . J Korean Med Sci. 2014 Feb; 29(2): 265–271.

Prasad, T., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., Balakrishnan, L., Marimuthu, A., Banerjee, S., Somanathan, D., Sebastian, A., Rani, S., Ray, S., Kishore, C., Kanth, S., Ahmed, M., Kashyap, M., Mohmood, R., Ramachandra, Y., Krishna, V., Rahiman, A., Mohan, S., Ranganathan, P., Ramabadran, S., Chaerkady, R. and Pandey, A. (2009). Human Protein Reference Database - 2009 update. Nucleic Acids Research. 37, D767-D772

Rabbani, B., Tekin, M. and Mahdieh, N. (2014) The promise of whole-exome sequencing in medical genetics. Review. Journal of Human Genetics, 59: 5–15

Reis, A. And Velho, G. (2002). Bases Genéticas do Diabetes Mellitus Tipo 2. Arquivos Brasileiros de Endocrinolagia e Metabologia, Vol 46, nº4; pg. 426-432

Roca, C and Adams, R. (2007). Regulation of vascular morphogenesis by Notch siganling. Genes and Development, 21:2511-2524.

Rothberg, J. et al, (2011). An integrated semiconductor device enabling non-optical genome sequencing. Nature 475, 348–352

Safi, S., Qvist, R., Kumar, S., Batumalaie, K and Ismail, I. (2014). Molecular Mechanisms of Diabetic Retinopathy, General Preventive Strategies and Novel Therapeutic Targets. BioMedical Research International, Volume 2014, Article ID 801269.

Sethi, N., Yan, Y., Quek, D., Schupbach, T and Kang, Y. (2010). Rabconnectin-3 is a functional regulator of Mammalian Notch Signaling. Journal of Biological Chemistry, Vol 285, no. 45:34757-34764

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K.(2001). dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001 Jan 1;29 (1):308-11.

Shiokawa, D., Matsushita, T., Kobayashi, T, Matsumoto, Y & Tanuma, S. (2004). Characterization of the human DNASE1L2 gene and the molecular mechanism of its transcriptional activation induced by inflammatory cytokines. Genomics, 84: 95-105.

Shyr C, Tarailo-Graovac M, Gottlieb M, Lee J, van Karnebeek C, Wasserman W. (2014). FLAGS, frequently mutated genes in public exomes. BMC Medical Genomics, 7:64.

Simbolo, M., Gottardi, M., Corbo, V., Fassan, M., Mafficini, A., Malpeli, G., Lawlor, R. and Scarpa, A. (2013). DNA Qualification Workflow for Next Generation Sequencing of Histopathological Samples. PLOS One, Volume 8, Issue 6: e62692

Sugimoto, M., Cutler, A., Shen, B., Moss, S., Iyengar, S., Klein, R., Folkman, J. amd Anand-Apte, B. (2013).Inhibition of EGF Signaling Protects the Diabetic Retina from Insulin-Induced Vascular Leakage The American Journal of Pathology, Vol 183, No. 3 pp.987-995

Tabernero, J. (2007). The Role of VEGF and EGFR Inhibition: Implications for Combining Anti–VEGF and Anti–EGFR Agents. Molecular Cancer Research, Vol 5(3)pp203-217.

Tamura, Y., Adachi, H., Osuga, J., Ohashi, K., Yahagi, N., Sekiya, M., Okazaki, H., Tomita, S., Iizuka, Y., Shimano, H., Nagai, R., Kimura, S., Tsujimoto, M. and Ishibashi, S. (2003). FEEL-1 and FEEL-2 Are Endocytic Receptors for Advanced Glycation End Products. The Journal Of Biological Chemistry, Vol. 278, No. 15, Issue of April 11, pp. 12613–12617, 2003

Tang, Z., Fang, Z and Zhou, L. (2013). Human Genetics of Diabetic Vascular Complications. Journal of Genetics, Vol 92, No.3.

Tarr, J., Kaul, K., Chopra, M., Kohner, E. and Chibber, R. (2013). Pathophysiology of Diabetic Retinopathy. ISRN Ophthalmology, Article ID 343560.

van den Oever, I., Raterman, H., Nurmohamed, M & Simsek, S. (2010). Endothelial Dysfunction, Inflammation and Apoptosis in Diabetes Mellitus. Mediators of Inflammation, Volume 2010,Article ID 792393.

van Dijk, E., Auger, H., Jaszczyszyn, Y . and Thermes, C. (2014). Ten Years of Next-Generation Sequencing Technology. Trends in Genetics, Vol. 30, No. 9, pp. 418-426

Venuti, A., Paolini, F., Nasir, L., Corteggio, A., Roperto, S., Campo, M. and Borzacchiello, G. (2011). Papillomavirus E5; the smallest oncoprotein with many functions. Molecular Cancer, 10:140

Wang, Z., Liu, X., Yang, B. Z. and Gelernter, J. (2013). The role and challenges of exome sequencing in studies of human diseases. Review Article. Frontiers in Genetics, Vol 4, Article 160.

Weedon, M., McCarthy, M., Hitman, G., Walker, M,. Groves, C., Zeggini, E., Rayner, N., Shields, B., Owen, K., Hattersley, A. and Frayling, T. (2006). Combining Information from Common Type 2 Diabetes Risk Polymorphisms Improves Disease Prediction. PLoS MEDICINE; Volume 3, Issue 10, e374; pg. 1877-1882

Warpeha, K. & Chakravarthy, U. (2003). Molecular genetics of microvascular disease in diabetic retinopatht. Eye, 17, 305-311.

Weijts BG1, Bakker WJ, Cornelissen PW, Liang KH, Schaftenaar FH, Westendorp B, de Wolf CA, Paciejewska M, Scheele CL, Kent L, Leone G, Schulte-Merker S, de Bruin A. (2012). E2F7 and E2F8 promote angiogenesis through transcriptional activation of VEGFA in cooperation with HIF1A. The EMBO Journal, 31, 3871-3884.

Wu, J., Li, Y. & Jiang, R. (2014). Integrating Multiple Genomic Data to Predict Disease-Causing Nonsynonymous Single Nucleotide Variants in Exome Sequencing Studies. PLoS Genetics, 10(3), e1004237.

Yamaji, Y., Yoshida, S., Ishikawa, K., Sengoku, A., Sato, K., Yoshida, A., Kuwahara, R., Ohuchida, K., Oki, E., Enaida, H., Fujisawa, K., Kono, T. and Ishibashi, T. (2008). TEM7 (PLXDC1) in Neovascular Endothelial Cells of Fibrovascular Membranes from Patients with Proliferative Diabetic Retinopathy. Investigative Ophthalmology & Visual Science, July 2008, Vol. 49, No. 7: 3151-3157.

Yi, N. (2010). Statistical Analysis of Genetic Interactions. Genet Res (Camb).December; 92(5-6):443-459.

Zhao, L., Patel, S., Pei, J and Zhang, K. (2013). Antagonizing Wnt pathway in Diabetic Retinopathy. Diabetes, Vol. 62 pp. 3993-3995.

APPENDiX

Table I: Patient Characterization. Information for sex, actual age at the time of the study, approximate Diabetes Duration (years), the presence or absence of DR, the ETDR level and glycated haemoglobin % is represented.

Exome ID	Sex (M=1; F=0)	Actual Age (years)	Approximate Diabetes Duration (years)	Diabetic Retinopathy (1=DR; 0=nonDR)	ETDR level	Glycated Hemoglobin (%)
EX1	1	64	9	0	12	10,2
EX2	1	70	23	1	43	9,3
EX4	1	72	14	1	35	11
EX6	0	58	13	0	14	8,5
EX7	0	70	33	1	90	7,1
EX9	1	48	12	1	35	9,9
EX10	1	61	21	1	90	12,6
EX11	1	65	12	1	35	7,8
EX12	1	56	6	1	35	6,3
EX13	0	50	3	0	12	17,3
EX16	1	74	5	0	15	5,2
EX17	1	68	17	1	43	8,6
EX18	1	68	10	0	15	6,6
EX19	1	70	23	0	12	10,3
EX21	0	58	9	0	12	14,3
EX23	0	63	16	1	90	6,5
EX24	0	65	22	1	35	10,8
EX26	0	53	29	1	35	8,4
EX27	0	77	19	0	15	9,8
EX28	0	73	4	0	12	6,9
EX29	1	62	13	0	15	7,6
EX30	1	66	22	1	35	11,2
EX33	1	71	13	1	43	8,3
EX34	0	65	24	1	43	8,9
EX35	0	52	9	0	12	7,7
EX36	0	65	9	0	12	9,3
EX37	1	64	22	1	43	7,7
EX38	0	74	5	0	12	10,5
EX39	1	65	23	1	47	8,5
EX40	0	54	17	1	20	11,5
EX41	1	76	41	1	43	9,7
EX42	0	58	17	1	25	12,3
EX43	1	70	5	1	47	9
EX44	0	57	33	0	12	11
EX45	1	70	20	0	15	9,3
EX46	0	62	13	1	35	7,7
EX47	0	65	23	1	35	10,9
EX48	1	55	9	1	53	10,8
EX49	0	70	27	1	53	11,9
EX50	0	58	15	0	12	9

Table II. Quality Control Results after DNA extraction. Y - Yes

Sample	Fluorescence Values (ng/uL)	Absorvance values 260/280	Absorvance values 260/230	Purification Procedure (Isopropanol Precipitation)
Ex1	25,1	1,71	1,25	**Y**
Ex2	39,3	1,85	2,20	
Ex4	23,6	1,75	1,56	**Y**
Ex6	30,8	1,75	1,91	
Ex7	32,3	1,81	1,71	
Ex9	36,0	1,78	1,65	**Y**
Ex10	63,1	7,83	2,06	
Ex11	6,5	1,63	0,95	**Y**
Ex12	22,3	1,73	1,04	**Y**
Ex13	43,9	1,84	2,21	
Ex16	63,0	1,79	1,89	
Ex17	56,3	1,80	2,10	
Ex18	30,2	1,85	1,73	
Ex19	38,6	1,79	1,84	
Ex21	29,3	1,66	1,09	**Y**
Ex23	63,9	1,78	1,85	
Ex24	36,1	1,85	2,09	
Ex26	65,8	1,73	1,59	**Y**
Ex27	48,5	1,79	1,67	**Y**
Ex28	49,8	1,79	1,57	**Y**
Ex29	53,8	1,76	1,43	**Y**
Ex30	61,3	1,79	1,61	**Y**
Ex33	39,4	1,81	1,24	**Y**
Ex34	58,8	1,77	1,51	**Y**
Ex35	72,6	1,80	1,68	**Y**
Ex36	42,2	1,80	1,97	
Ex37	35,6	1,78	1,60	**Y**
Ex38	75,8	1,81	1,87	
Ex39	31,6	1,79	1,96	
Ex40	41,3	1,77	1,40	**Y**
Ex41	48,8	1,84	2,13	
Ex42	52,7	1,78	1,92	
Ex43	67,5	1,79	2,12	
Ex44	46,7	1,79	1,97	
Ex45	56,6	1,83	1,99	
Ex46	40,9	1,75	2,10	
Ex47	75,2	1,79	2,22	
Ex48	40,2	1,80	1,79	
Ex49	54,9	1,76	1,78	
Ex50	26,3	1,77	1,39	**Y**

Table III. IonXpress barcode (MID) attribution to samples.

Sample	IonXpress Barcode
EX1	001
EX2	002
EX4	004
EX6	006
EX7	007
EX9	009
EX10	010
EX11	011
EX12	012
EX13	013
EX16	016
EX17	001
EX18	002
EX19	003
EX21	005
EX23	007
EX24	008
EX26	010
EX27	011
EX28	012
EX29	013
EX30	014
EX33	001
EX34	002
EX35	003
EX36	004
EX37	005
EX38	006
EX39	007
EX40	008
EX41	009
EX42	010
EX43	011
EX44	012
EX45	013
EX46	014
EX47	015
EX48	016
EX49	001
EX50	002

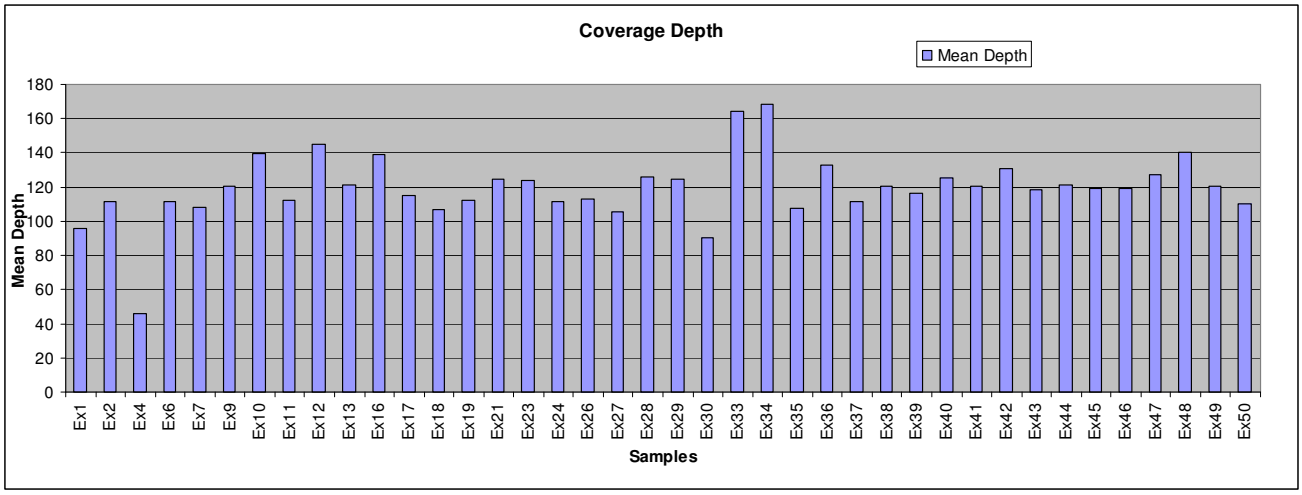


Figure A. Coverage Depth for all 40 exomes sequenced

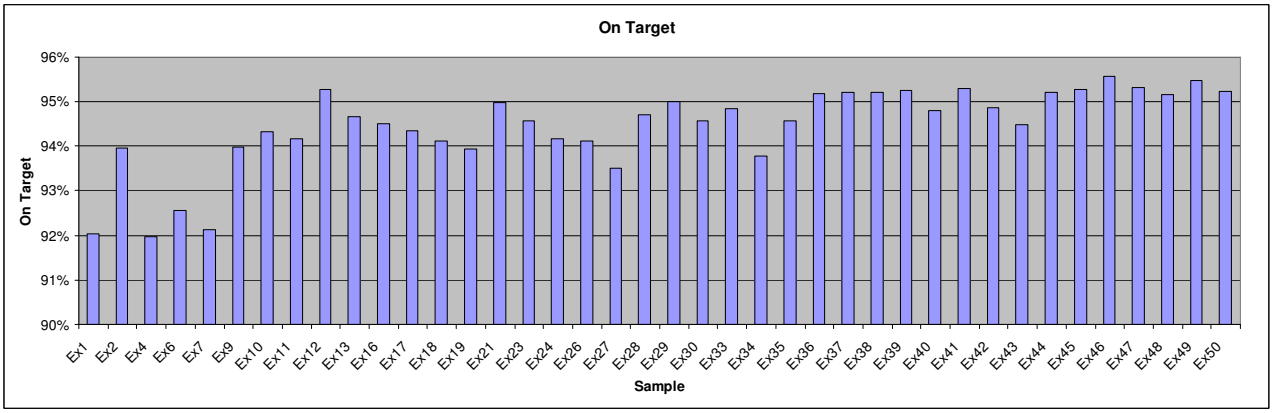


Figure B. On Target (%) for all 40 exomes sequenced.

MARKER_ID	NUM_PASS_VARS	NUM_SING_VARS	PVALUE
17:72365698-72368711_GPR142	2	1	0,00023708
20:47864230-47887168_ZNFX1	3	3	0,00094832
6:157488325-157525120_ARID1B	2	1	0,00094832
11:64519437-64527310_PYGM	4	3	0,0010121
14:105169564-105185146_INF2	5	5	0,0025023
1:247614602-247615148_OR2B11	3	2	0,0025469
17:63149547-63189687_RGS9	3	3	0,0026079
10:74899499-74908110_ECD	2	2	0,002845
1:153587865-153587865_S100A14	1	0	0,002845
17:3101591-3101617_OR1A2	2	2	0,002845
17:39412071-39412125_KRTAP9-9	2	2	0,002845
17:39780752-39780752_KRT17	1	0	0,002845
17:60814250-60837337_MARCH10	2	2	0,002845
4:140640704-140640704_MAML3	1	0	0,002845
5:178555044-178771135_ADAMTS2	2	2	0,002845
22:26164394-26423535_MYO18B	7	6	0,0029763
6:7373434-7387238_CAGE1	3	3	0,0037933
16:2287215-2287632_DNASE1L2	3	2	0,0052564
10:73964269-73970502_ASCC1	2	2	0,0056899
1:86249980-86591175_COL24A1	2	2	0,0056899
2:130931095-130931167_SMPD4	1	0	0,0056899
2:70900443-70933491_ADD2	2	2	0,0056899
3:36756863-36780081_DCLK3	2	2	0,0056899
16:14956982-14989425_NOMO1	3	2	0,0063792
1:201169532-201185638_IGFN1	3	2	0,0065298
11:56143113-56143943_OR8U1	2	0	0,007993
15:45389453-45399036_DUOX2	4	4	0,007993
13:25254891-25284616_ATP12A	6	5	0,0081569
19:15563529-15563934_RASAL3	2	1	0,0097202
19:20228996-20229780_ZNF90	4	4	0,0097202
19:34180082-34262952_CHST8	3	3	0,0097202
7:1586617-1590576_TMEM184A	2	1	0,0097202
10:125506302-125558631_CPXM2	2	2	0,01138
1:247695273-247695756_OR2C3	3	3	0,01138
19:44681046-44681717_ZNF226	3	3	0,01138
21:44480591-44480591_CBS	1	0	0,01138
22:45128123-45255629_ARHGAP8	2	2	0,01138
22:45128123-45255629_PRR5-ARHGAP8	2	2	0,01138
2:73172214-73188370_SFXN5	3	3	0,01138
5:1254510-1254510_TERT	1	0	0,01138
9:27524515-27524731_IFNK	2	2	0,01138
9:94171786-94172507_NFIL3	2	2	0,01138
8:130761778-130774958_GSDMC	2	1	0,011678
15:63920891-64041689_HERC1	3	3	0,012565
10:116207642-116251624_ABLIM1	3	2	0,015884
11:62984850-62997031_SLC22A25	2	1	0,015884
1:47726135-47746113_STIL	3	3	0,015884
7:150164346-150174507_GIMAP8	2	1	0,015884
11:62652950-62655818_SLC3A2	2	2	0,01707
19:19308064-19310047_RFXANK	1	0	0,01707
20:43723614-43723669_KCNS1	2	2	0,01707
20:57767960-57828134_ZNF831	2	2	0,01707
21:45175604-45175847_PDXK	2	2	0,01707
22:21344733-21351614_LZTR1	2	2	0,01707
4:983556-983684_SLC26A1	2	2	0,01707
9:75355093-75355093_TMC1	1	0	0,01707
6:12122645-12164308_HIVEP1	4	4	0,019427
1:10713845-10720270_CASZ1	3	3	0,019678
4:152499205-152570675_FAM160A1	3	3	0,019678
1:215847545-216246585_USH2A	5	5	0,021222
10:55600221-55782754_PCDH15	3	2	0,022678
22:18300278-18371970_MICAL3	3	3	0,024182
12:110924375-110924375_FAM216A	1	0	0,025605
12:20522514-20769270_PDE3A	2	2	0,025605
1:222798106-222802261_MIA3	2	2	0,025605
1:23847537-23857060_E2F2	2	2	0,025605
14:69256441-69256441_ZFP36L1	1	0	0,025605
16:772780-775236_CCDC78	2	2	0,025605
16:8895679-8898702_PMM2	2	2	0,025605
17:43481395-43506940_ARHGAP27	2	2	0,025605
19:35616111-35624994_LGI4	2	2	0,025605
22:41537234-41574925_EP300	2	2	0,025605
2:33482545-33585796_LTBP1	2	2	0,025605
2:73303267-73315220_RAB11FIP5	2	2	0,025605
3:142395072-142402915_PLS1	2	2	0,025605

MARKER_ID	NUM_PASS_VARS	NUM_SING_VARS	PVALUE
5:484771-492142_SLC9A3	2	2	0,025605
6:160461723-160482649_IGF2R	2	2	0,025605
9:107554255-107599376_ABCA1	3	3	0,025605
9:128678097-128678097_PBX3	1	0	0,025605
20:55027391-55033420_CASS4	3	2	0,026363
3:124826801-124906167_SLC12A8	4	3	0,027513
11:65144075-65147013_SLC25A45	1	0	0,029161
11:75298327-75298558_MAP6	3	3	0,029161
6:29407955-29408721_OR10C1	5	1	0,030427
10:103899871-103899970_PPRC1	2	2	0,034139
1:171173089-171178090_FMO2	3	3	0,034139
13:28794482-28844967_PAN3	2	2	0,034139
15:70957020-70970463_UACA	2	2	0,034139
17:19451355-19470502_SLC47A1	2	2	0,034139
18:34156497-34205521_FHOD3	2	2	0,034139
22:21799719-21800202_HIC2	2	2	0,034139
3:46563126-46574357_LRRC2	2	2	0,034139
4:2252847-2259728_MXD4	2	2	0,034139
5:140579961-140581006_PCDHB11	2	2	0,034139
6:36104494-36106712_MAPK13	1	0	0,034139
9:35740293-35740293_GBA2	1	0	0,034139
10:100182151-100202971_HPS1	3	3	0,034851
1:246078824-246078824_SMYD3	1	0	0,034851
1:3679886-3680362_CCDC27	2	1	0,034851
3:78676694-78763613_ROBO1	3	3	0,034851
9:73151752-73167838_TRPM3	1	0	0,034851
15:42106923-42116750_MAPKBP1	3	2	0,035006
1:94467548-94520733_ABCA4	4	4	0,035006
3:111603901-111672861_PHLDB2	4	2	0,036848
2:186653502-186678720_FSIP2	6	6	0,04101
19:20002603-20003486_ZNF253	2	1	0,041252
19:54973482-54974538_LENG9	3	3	0,041252
2:29287863-29297091_C2orf71	3	2	0,041252
7:123302011-123303052_LMOD2	3	3	0,041252
6:160958935-161020526_LPA	7	4	0,044196
10:82330044-82348429_SH2D4B	2	2	0,045519
10:89264706-89265292_MINPP1	2	2	0,045519
1:145301716-145304656_RP11-458D21.5	1	0	0,045519
11:63276396-63276432_LGALS12	2	2	0,045519
11:6976984-6977031_ZNF215	2	2	0,045519
12:110418705-110418705_TCHP	1	0	0,045519
12:111748192-111785678_CUX2	2	2	0,045519
1:223177972-223178706_DISP1	2	2	0,045519
12:57648716-57663716_R3HDM2	2	2	0,045519
16:113057-113713_RHBDF1	2	2	0,045519
16:3100094-3108573_MMP25	2	2	0,045519
16:89293934-89294865_ZNF778	2	2	0,045519
17:67081193-67111579_ABCA6	2	2	0,045519
17:77758597-77758597_CBX2	1	0	0,045519
19:2217094-2217786_DOT1L	2	2	0,045519
1:9657118-9673046_TMEM201	2	2	0,045519
20:24944523-24952091_APMAP	2	2	0,045519
20:57036604-57042655_APCDD1L	2	2	0,045519
21:40752359-40765192_WRB	2	2	0,045519
22:30188418-30221173_ASCC2	2	2	0,045519
22:31673116-31674324_LIMK2	2	1	0,045519
2:97270095-97270095_KANSL3	1	0	0,045519
3:121828181-121828181_CD86	1	0	0,045519
3:50331133-50332327_HYAL3	2	2	0,045519
6:27858249-27860534_HIST1H3J	2	2	0,045519
6:28093526-28097586_ZSCAN16	2	2	0,045519
9:99797981-99797981_CTSL2	1	0	0,045519
6:51640639-51923404_PKHD1	6	6	0,045542
1:231830492-231906713_DISC1	3	2	0,0456
19:38893807-38899459_FAM98C	3	2	0,0456
12:56075599-56077768_METTL7B	4	4	0,048364
14:24619811-24629557_RNF31	3	3	0,048364
15:72122642-72338423_MYO9A	3	3	0,048364
16:10971206-11002927_CIITA	3	3	0,048364
17:7910817-7915471_GUCY2D	3	2	0,048364
20:3652365-3655681_ADAM33	4	4	0,048364
2:163241287-163361057_KCNH7	3	3	0,048364
2:238483751-238483751_RAB17	1	0	0,048364
3:185906117-185990072_DGKG	3	3	0,048364

Figure C. Lists of rare variant accumulating genes from the quantitative test approach

Risk MARKER_ID (chrom: start-end_gene)	PASS MARKERS	BURDEN CNT	PVALUE
20:47864230-47887168_ZNFX1	3	3	0,0013
6:7373434-7387238_CAGE1	3	3	0,0014
14:105169564-105185146_INF2	5	5	0,002
17:60814250-60837337_MARCH10	2	2	0,0026
6:157488325-157525120_ARID1B	2	3	0,0029
10:74899499-74908110_ECD	2	2	0,003
17:39412071-39412125_KRTAP9-9	2	2	0,0034
17:39780752-39780752_KRT17	1	2	0,0035
1:57207874-57257861_C1orf168	4	5	0,0039
21:44480591-44480591_CBS	1	2	0,0055
22:26164394-26423535_MYO18B	7	8	0,0062
2:70900443-70933491_ADD2	2	2	0,0062
14:69256441-69256441_ZFP36L1	1	2	0,0064
1:86249980-86591175_COL24A1	2	2	0,0065
20:3652365-3655681_ADAM33	4	3	0,0065
2:33482545-33585796_LTBP1	2	2	0,0067
7:150164346-150174507_GIMAP8	2	3	0,0068
1:222798106-222802261_MIA3	2	2	0,0069
10:73964269-73970502_ASCC1	2	2	0,0072
2:73172214-73188370_SFXN5	3	2	0,0077
6:12122645-12164308_HIVEP1	4	4	0,0079
12:110924375-110924375_FAM216A	1	2	0,0085
1:201169532-201185638_IGFN1	3	4	0,009
19:19308064-19310047_RFXANK	1	2	0,0107
11:62652950-62655818_SLC3A2	2	2	0,0114
3:41795902-41949348_ULK4	3	2	0,012
9:94171786-94172507_NFIL3	2	2	0,013
19:15563529-15563934_RASAL3	2	3	0,014
8:23160918-23225775_LOXL2	2	2	0,014
17:39871696-39872026_GAST	3	2	0,0141
14:74823701-74824956_VRTN	2	2	0,018
2:163241287-163361057_KCNH7	3	3	0,018
19:38572678-38684261_SIPA1L3	2	1	0,019
6:28093526-28097586_ZSCAN16	2	2	0,019
9:131038429-131038643_SWI5	3	1	0,019
1:215847545-216246585_USH2A	5	5	0,02
22:21344733-21351614_LZTR1	2	2	0,02
6:27858249-27860534_HIST1H3J	2	2	0,02
11:2437185-2439468_TRPM5	2	1	0,021
13:114083328-114083333_ADPRHL1	2	1	0,021
5:13830154-13894796_DNAH5	3	3	0,021
5:55250727-55264100_IL6ST	2	1	0,021
9:34485237-34514388_DNAI1	3	3	0,021
11:73717247-73718063_UCP3	2	1	0,022
2:219871207-219892344_CCDC108	3	3	0,022
2:88826005-88828867_C2orf51	2	2	0,022
15:51743890-51772229_DMXL2	4	4	0,023
2:73303267-73315220_RAB11FIP5	2	2	0,023
5:484771-492142_SLC9A3	2	2	0,023
X:153578465-153593616_FLNA	2	1	0,024
15:70957020-70970463_UACA	2	2	0,025
3:142395072-142402915_PLS1	2	2	0,026
9:135139790-135203803_SETX	3	2	0,026
20:57767960-57828134_ZNF831	2	2	0,027
3:33038788-33055721_GLB1	2	2	0,027
9:135102244-135102327_NTNG2	1	2	0,027
17:59949672-59988892_INTS2	3	3	0,028
20:24944523-24952091_APMAP	2	2	0,028
2:236708131-236877255_AGAP1	2	2	0,028
19:20002603-20003486_ZNF253	2	3	0,029
5:134910364-134910382_CXCL14	2	2	0,029
11:64604225-64607024_CDC42BPG	1	2	0,03
3:129694696-129695938_TRH	2	3	0,031
3:195481111-195518170_MUC4	15	13	0,031
14:31097427-31119819_SCFD1	2	2	0,032

Risk MARKER_ID (chrom: start-end_gene)	PASS MARKERS	BURDEN CNT	PVALUE
8:71025871-71041146_NCOA2	4	6	0,032
10:116207642-116251624_ABLIM1	3	3	0,033
12:20522514-20769270_PDE3A	2	2	0,033
16:2287215-2287632_DNASE1L2	3	4	0,033
16:68914537-69056827_TMCO7	2	2	0,033
18:34156497-34205521_FHOD3	2	2	0,033
19:35616111-35624994_LGI4	2	2	0,033
14:94754804-94756669_SERPINA10	2	2	0,034
17:19451355-19470502_SLC47A1	2	2	0,034
2:86373287-86378518_IMMT	1	2	0,034
11:19247163-19258929_E2F8	5	4	0,035
22:26068297-26083642_ADRBK2	2	2	0,035
3:46563126-46574357_LRRC2	2	2	0,035
14:94953705-94964127_SERPINA12	1	2	0,037
17:76420087-76571085_DNAH17	12	12	0,037
2:202557702-202557756_MPP4	2	2	0,037
3:195594795-195597000_TNK2	3	3	0,037
5:176002294-176016152_CDHR2	4	4	0,037
13:28794482-28844967_PAN3	2	2	0,038
14:20711005-20711005_OR11H4	1	2	0,038
15:45389453-45399036_DUOX2	4	4	0,038
7:1586617-1590576_TMEM184A	2	3	0,038
11:56143113-56143943_OR8U1	2	4	0,039
17:39724414-39728051_KRT9	2	2	0,039
17:4535035-4542145_ALOX15	2	3	0,039
19:14029627-14031728_CC2D1A	3	3	0,039
4:2252847-2259728_MXD4	2	2	0,039
11:6976984-6977031_ZNF215	2	2	0,04
14:91927783-91975846_SMEK1	2	2	0,04
5:37020562-37049406_NIPBL	2	2	0,04
15:83680313-83680333_C15orf40	2	3	0,041
16:84189349-84203460_DNAAF1	2	2	0,041
9:117835900-117853210_TNC	3	3	0,041
1:901922-909428_PLEKHN1	2	3	0,042
1:9098071-9101726_SLC2A5	2	2	0,042
21:40752359-40765192_WRB	2	2	0,043
4:89408212-89425756_HERC5	2	2	0,043
8:130761778-130774958_GSDMC	2	4	0,043
9:135275596-135277136_TTF1	2	3	0,044
9:84606620-84609152_FAM75D1	5	5	0,044
6:160461723-160482649_IGF2R	2	2	0,045
10:120801896-120833251_EIF3A	2	4	0,046
12:25267613-25267613_CASC1	1	3	0,046
14:93398717-93399011_CHGA	2	3	0,046
16:64981952-65016003_CDH11	3	3	0,046
22:30856091-30857645_SEC14L3	2	2	0,046
19:12574827-12574979_CTD-2192J16.17	2	2	0,047
19:36213572-36224674_MLL4	2	2	0,047
19:11488877-11492737_EPOR	2	3	0,048
6:158910743-158923120_TULP4	3	3	0,048
10:88213499-88277502_WAPAL	2	2	0,05
16:8895679-8898702_PMM2	2	2	0,05
5:73048875-73197045_ARHGEF28	3	4	0,05

Figure D. List of rare variant accumulating genes from the binary test approach.

DNASE1L2

rs ID	impact	impact severity	polyphen pred	CADD scaled	Transcript ID	Transcript Name	Transcript length	Samples	ETDR Values	Disease Duration (years)
rs160550	splice_region	MED	None	0,12	ENST00000569184	DNASE1L2-005	762	Ex12, Ex46	35,35	6,13
rs200934792	non_syn_coding	MED	probably_damaging	15,21	ENST00000569184	DNASE1L2-005	762	23	90	16
rs200149634	non_syn_coding	MED	probably_damaging	14,44	ENST00000569184	DNASE1L2-005	762	48	53	9

DMXL2 =RC3

rs ID	impact	impact severity	polyphen pred	CADD scaled	Transcript ID	Transcript Name	Transcript length	Exomes	ETDR Values	Disease Duration (years)
None	non_syn_coding	MED	benign	14,65	ENST00000449909	DMXL2-003	7465	42	25	17
rs114516513	non_syn_coding	MED	possibly_damaging	18,03	ENST00000449909	DMXL2-003	7465	49	53	27
None	non_syn_coding	MED	possibly_damaging	17,19	ENST00000449909	DMXL2-003	7465	24	35	22

E2F8

rs ID	impact	impact severity	polyphen pred	CADD scaled	Transcript ID	Transcript Name	Transcript length	Samples	ETDR Values	Disease Duration (years)
None	splice_acceptor	HIGH	None	16,62	ENST00000527884	E2F8-001	3432	Ex7	90	33
rs793274	non_syn_coding	MED	benign	3,98	ENST00000527884	E2F8-001	3432	Ex46	35	13
rs77599073	non_syn_coding	MED	benign	8,91	ENST00000527884	E2F8-001	3432	Ex49	53	27
rs141999878	non_syn_coding	MED	probably_damaging	13,21	ENST00000527884	E2F8-001	3432	Ex42	25	17
None	non_syn_coding	MED	probably_damaging	35	ENST00000532666	E2F8-004	719	Ex49	53	27

CADD > 10

Figure E. Information regarding rare variants accumulated in genes: DNASE1L2, DMXL2 and E2F8, considered biologically relevant to Diabetic Retinopathy. Cells highlighted in red refer to variants with a CADD score >10.

MAML3

rs ID	impact	impact_severity	polyphen_pred	cadd_scaled	Transcript ID	Transcript Name	Transcript length	Exomes	ETDR Values	Duration (years)
rs4863506	non_syn_coding	MED	benign	3.52	ENST00000509479	MAML3-001	6844	23,10	90, 90	16, 21

S100A14

rs ID	impact	impact_severity	polyphen_pred	cadd_scaled	Transcript ID	Transcript Name	Transcript length	Exomes	ETDR Values	Duration (years)
rs151066419	splice_region	MED	None	4.06	ENST00000368701	S100A14-002	1080	23,10	90, 90	16, 21

CASZ1

rs ID	impact	impact_severity	polyphen_pred	cadd_scaled	Transcript ID	Transcript Name	Transcript length	Exomes	ETDR Values	Duration (years)
rs149552670	non_syn_coding	MED	benign	3.6	ENST00000344008	CASZ1-001	4405	7	90	33
rs61736955	non_syn_coding	MED	probably_damaging	25.7	ENST00000344008	CASZ1-001	4405	49	53	27
rs3748686	non_syn_coding	MED	probably_damaging	16.61	ENST00000344008	CASZ1-001	4405	17	43	17

EP300

rs ID	impact	impact_severity	polyphen_pred	cadd_scaled	Transcript ID	Transcript Name	Transcript length	Exomes	ETDR Values	Duration (years)
rs140187237	non_syn_coding	MED	possibly_damaging	11.12	ENST00000263253	EP300-001	9585	48	53	9
None	non_syn_coding	MED	unknown	7.58	ENST00000263253	EP300-001	9585	7	90	33

CADD > 10

Figure F. Information regarding rare variants accumulated in genes: MAML3, S100A14, CASZ1 and EP300, considered biologically relevant to Diabetic Retinopathy. Cells highlighted in red refer to variants with a CADD score >10.

APCDD1L

rs ID	impact	impact_severity	polyphen_pred	cadd_scaled	Transcript ID	Transcript Name	Transcript length	Exomes	ETDR Values	Duration
None	non_syn_coding	MED	benign	3.66	ENST00000439429	APCDD1L-201	2040	43	47	5
rs7265902	non_syn_coding	MED	possibly_damaging	21.2	ENST00000439429	APCDD1L-201	2040	48	53	9

GPR142

rs ID	impact	impact_severity	polyphen_pred	cadd_scaled	Transcript ID	Transcript Name	Transcript length	Exomes	ETDR Values	Disease Duration (years)
rs145323799	non_syn_coding	MED	benign	17.45	ENST00000335666	GPR142-002	1437	10	90	21
rs140397567	non_syn_coding	MED	benign	8.99	ENST00000335666	GPR142-002	1437	23.7	90, 90	16, 33

ADAMTS2

rs ID	impact	impact_severity	polyphen_pred	cadd_scaled	Transcript ID	Transcript Name	Transcript length	Exomes	ETDR Values	Disease Duration (years)
rs76488852	non_syn_coding	MED	benign	0.04	ENST00000251582	ADAMTS2-001	6754	23	ETDR = 90	16
rs146222244	non_syn_coding	MED	benign	10.79	ENST00000251582	ADAMTS2-001	6754	23	ETDR = 90	21

CADD > 10

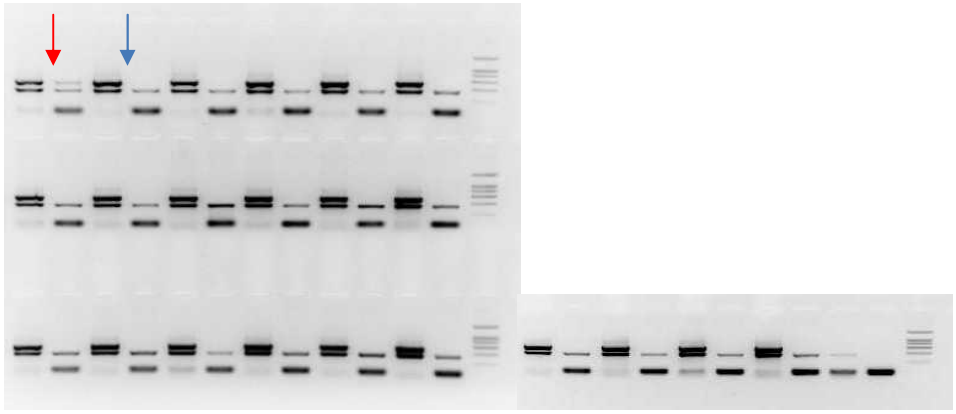
Figure G. Information regarding rare variants accumulated in genes: APCDD1L, GPR142 and ADAMTS2, considered biologically relevant to Diabetic Retinopathy. Cells highlighted in red refer to variants with a CADD score >10.

ASO-PCR Genotyping Results

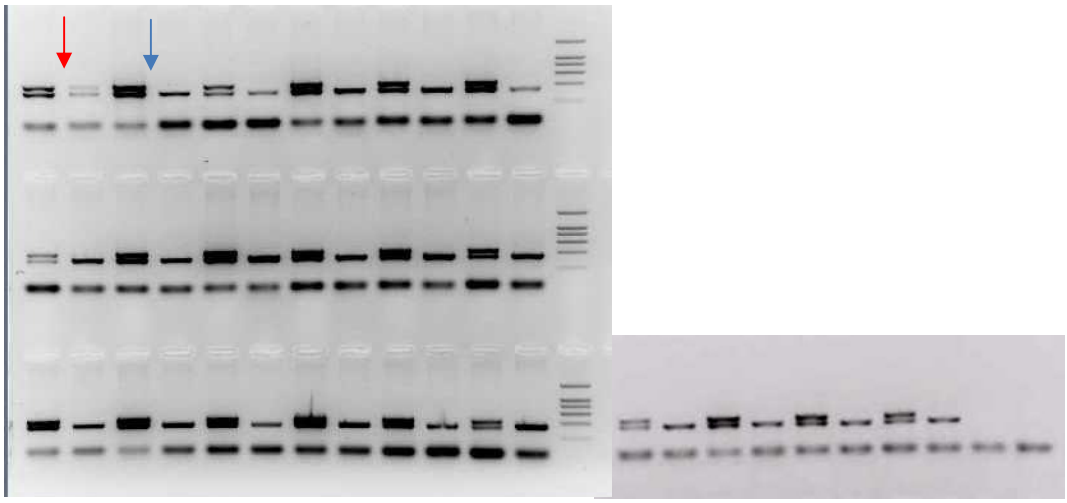
Electrophoresis gel results from the ASO-PCR genotyping validation of the 3 variants in DMXL2 and E2F8 genes.

The red arrows depict the genotype of the T2D individual with the rare variant and thus our heterozygote control. The blue arrows depict a T2D individual that did not have the variant and thus our wild type control. The result for all 20 control individuals are represented. Every two wells correspond to an individual/genotype. X means that the ASO-PCR genotyping of that sample was not efficient and would have to be repeated. A negative control for each ASO-PCR amplification was prepared.

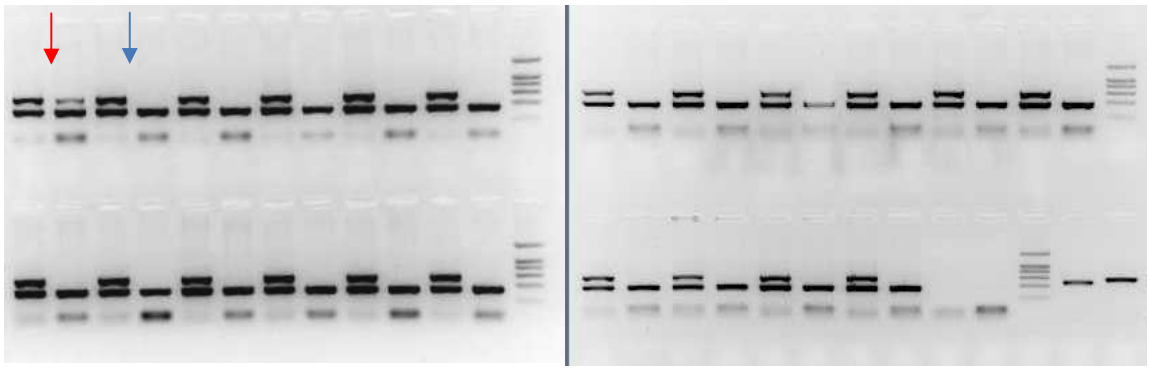
DMXL2_rs rs114516513 T/C



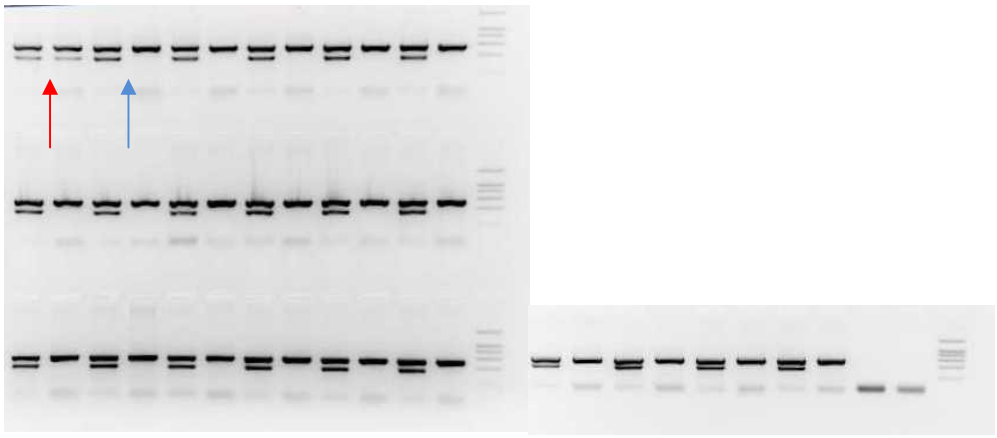
DMXL2_ 15: 51743889 C/G



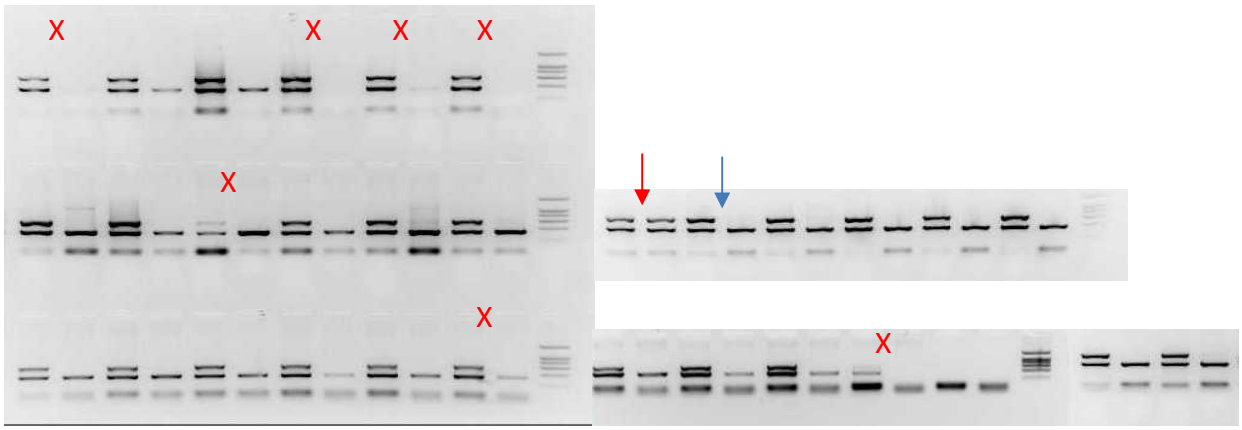
DMXL2_ 15: 51772228 C/G



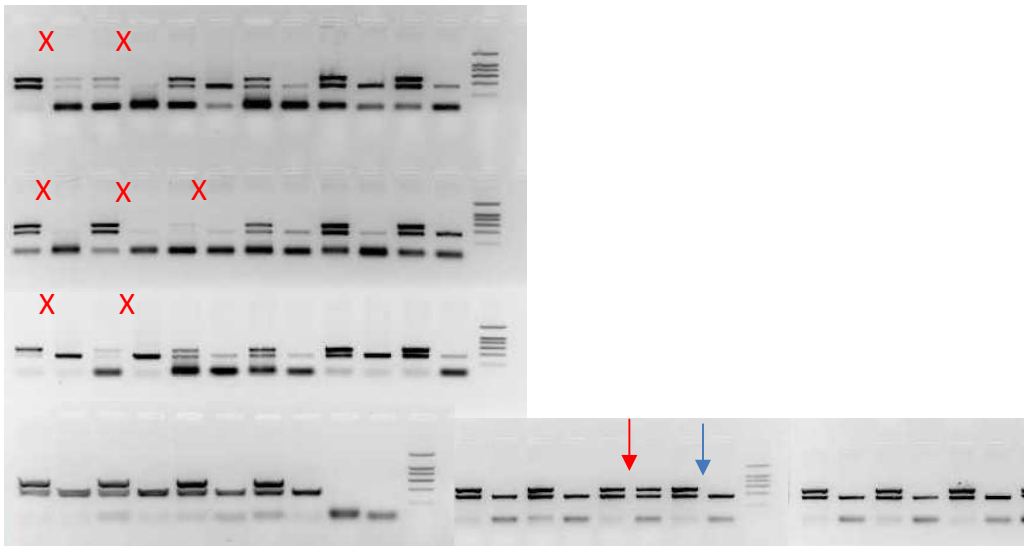
E2F8_ rs141999878 C/A:



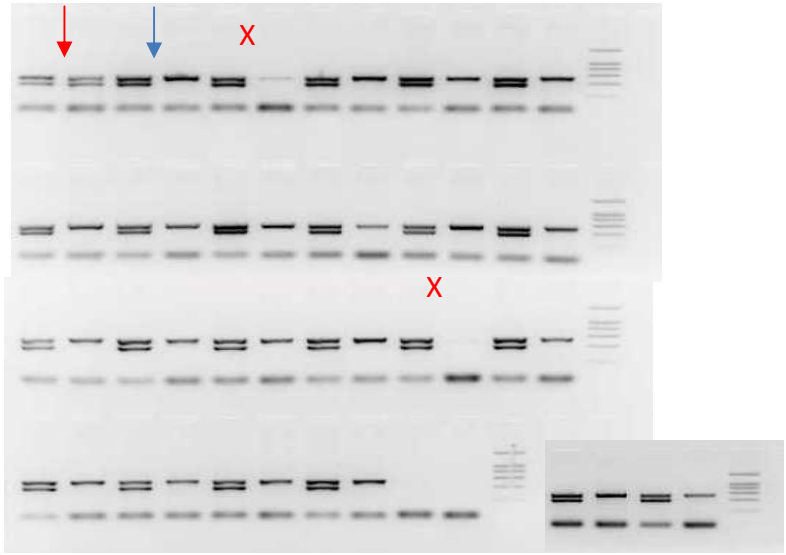
E2F8_ rs77599073 C/G



E2F8_ rs793274 T/C



E2F8_ 11: 19247163 C/G



E2F8_ 11:19258929 C/T