Università di Siena

Dipartimento di Medicina Molecolare e dello Sviluppo

Dottorato di Ricerca in Medicina Molecolare

Ciclo XXXIV

Coordinatore: Prof. Vincenzo Sorrentino

"DEVELOPMENT OF MODELS TO UNDERSTAND THE COMPLEXITY IN CARDIOVASCULAR RESEARCH AND DIAGNOSTICS"

Settore scientifico disciplinare: PATOLOGIA CLINICA MED05

DOTTORANDO:

Dott. Samuele Suraci

TUTOR:

Prof.ssa Betti Giusti

Anno Accademico 2020/2021

Index

Aim of the PhD Project

The large part of cardiovascular diseases derive from both heritable and environmental contributions. The advent of high throughput sequencing technology enlightened that either common multifactorial or rare cardiovascular diseases, classically thought to be monogenic, result from the contribution of more genes that cause or modulate the phenotype. The aim of this PhD project was to get an insight into the complexity of atherothrombotic diseases taking advance from the opportunities offered by omics technologies.

The first part of this work is about the development and application of bioinformatics pipelines for the mutational analysis of data derived from high throughput sequencing technology Illumina and their optimization for diagnostic purpose. The mutational analysis regarded gene panels implicated in Marfan Syndrome and related disorders, Von Willebrand Disease and Familial Dyslipidaemia.

The second part regards the development and application of a bioinformatics pipeline to evaluate the global RNA expression profiles from cerebral thrombi, obtained during thrombectomy treatment and from peripheral venous blood in patients with acute ischaemic stroke, using Affymetrix technology.

The third part concerns reverse-genomic and functional characterisation of single nucleotide variants in *DPP3* gene. DPP3 is a cytosolic enzyme involved in the degradation of cardiovascular mediators. Circulating DPP3 is emerging as biomarker of clinical outcome in patients suffering of acute and chronic cardiovascular diseases (i.e., acute heart failure, cardiogenic shock and aneurysmal subarachnoid haemorrhage). Databases of data deriving from thousands of genome-wide association studies were explored to search mutations in *DPP3* gene associated with atherothrombotic phenotypes. For the variants presenting a significative association, a model of the mutated protein was built through a bioinformatic software. The proteins presenting most promising mutations were then heterology expressed in E. coli and an initial biochemical characterization of their enzymatic activity was done.

Introduction to Atherothrombosis

Atherothrombosis occurs when a thrombus forms over an unstable atherosclerotic plaque. From the clinical point of view, the disease is a single pathologic entity that affects different vascular territories (Viles-Gonzalez *et al.,* 2004): it is a diffuse process that starts early in childhood and progresses asymptomatically through adult life affecting multiple vascular beds. Later in life, it is clinically manifested as coronary artery disease (CAD), stroke, transient ischemic attack (TIA), renal and peripheral arterial disease.

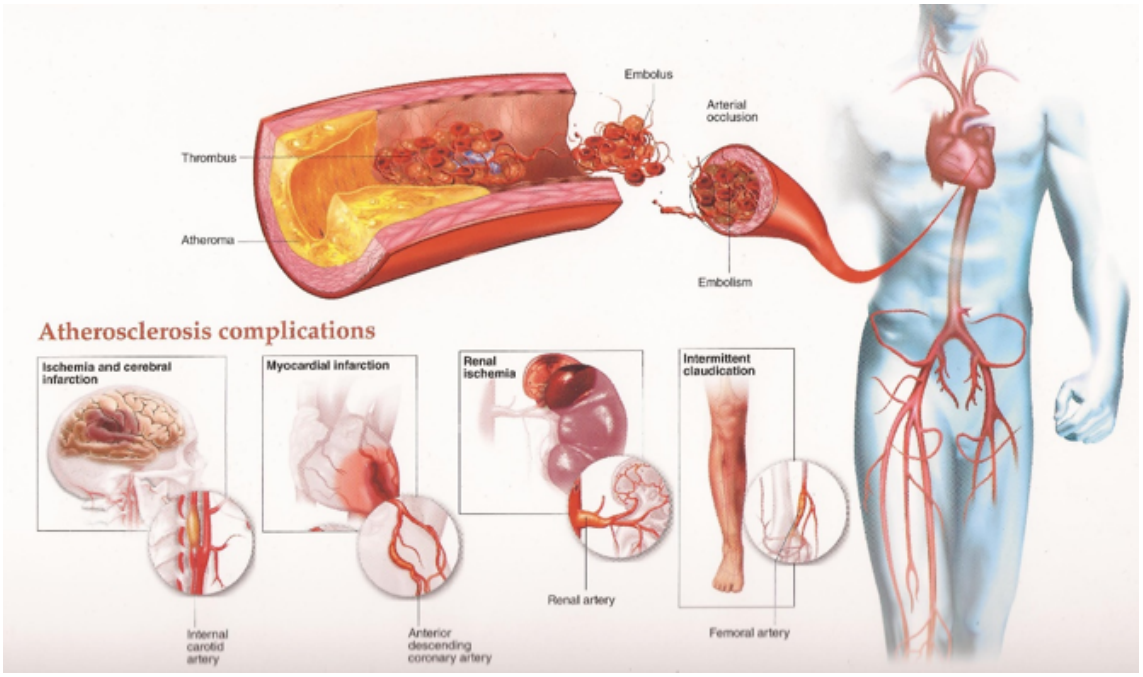


Figure 1: *Vascular territories affected by atherothrombosis. (Free stock image from the web)*

Epidemiology and impact of lifestyle on atherothrombosis

One person dies every 40 seconds in the United States from cardiovascular disease (CVD) (Mozaffarian *et al.,* 2016). More than 25 million people in the US have at least one clinical manifestation of Atherothrombotic vascular disease (AVD). About 80% of deaths from cardiovascular events occur as a result of a stroke or a heart attack. One third of these deaths occur prematurely in people under 75 years of age. Cardiovascular disease represents the 31% of general mortality in the world. Atherothrombosis is the main cause of mortality due to cardiovascular diseases (CVD). Approximately 75% of the cases of heart attack and approximately 90% of strokes associated with carotid arteries atherosclerosis are caused by thrombosis (Thim *et al.*, 2008). CVD is a big financial burden for healthcare systems. In 2009, CVD-related costs totaled 106 billion euro, which was approximately 9% of the total expenditure on health care in the European Union (Ab *et al.*, 2011). There

exists global trend towards the increase in the incidence of lifestyle diseases and a decrease in cases of premature death as compared to years on disability. In the context of lost years of life and life years on disability, ischaemic heart disease (IHD) and stroke are, respectively, in the first and third place in the world (Murray *et al.*, 2012). About 85–90% of strokes are of ischaemic aetiology (Berry *et al.*, 2012). The risk of death due to CVD during lifetime is approximately 30% (Liu *et al.*, 2012). Among diabetics, most of whom die due to CVD, 8 of 10 deaths are due to atherothrombosis. According to the findings of the Global Burden of Disease Study from 2010 onwards, adjusted for age, the mortality due to CVD has fallen approximately 20% during the last 80 years of the twentieth century. The success of the reduction of mortality due to CVD is associated with the development of methods of treatment and better organisation of healthcare, as well as preventive activities, including non-pharmacological interventions (Gu *et al.*, 1998). To the above, one can also add changes in tobacco legislation, which can lead to a 15% reduction in the risk of hospitalisation and a 16% reduction in mortality from coronary heart disease and stroke (Tan and Glantz, 2012). Not less important are the lifestyle changes, including eating habits. It has been demonstrated that appropriate physical activity and dietary intervention can contribute to approximately 35% reduction in the risk of death already accepted with adjustment of pharmacologic medication (Suzuki *et al.*, 2012). Proper diet contributes to the reduction of cardiovascular events (CVE) in patients after 55 years of age diagnosed with diabetes or a history of CVE irrespective of the use of drugs in secondary prevention (Dehnghan *et al.*, 2012). In terms of cardiovascular risk reduction, the introduction of statin therapy was the pharmacological milestone. This has proven effective in reducing CVE and mortality due to CVD (Alberico *et al.*, 2016).

Their use in low-risk populations decreased by approximately 30% the relative risk in this population. In addition, for patients intolerant of statins or for those who despite optimal therapy fail to achieve their therapeutic goal, Ezetimibe or Evolucumab can be currently used, new drugs of proven efficacy and safety of therapy (Cannon *et al.*, 2015). Ezetimibe connects with Niemann-Pick C1-like 1 (NPC1L1) proteins preventing absorption of cholesterol from the gastrointestinal tract. Used together with simvastatin, it significantly reduced the risk of mortality compared to statin monotherapy. Evolocumab is a monoclonal antibody interacting with enzyme PCSK9 (pro-protein convertase subtilisin kexin type-9) and significantly lowering LDL-cholesterol and total cholesterol and reducing CVE rate in combination with a statin as compared to statin monotherapy (Cannon *et al.*, 2015).

Very important element of therapy is patient's compliance. Adherence of the patient affects the effectiveness of the therapy. The review of approximately 20 studies involving

a total of over 375,000 patients showed only 42–61% of adherence to treatment in patients receiving cardiovascular drugs as primary prevention and 62–76% adherence in secondary prevention (Naderi *et al.*, 2012). There are some differences in CVD mortality among different races. Black people have a higher risk of death from coronary heart disease and 2–4 times higher risk of ischaemic stroke than white people. The Asian race and the people of the Pacific Islands are at the highest risk for hemorrhagic stroke (WHO, 2011). Appropriate prevention would reduce the CVD cases by 80% (Liu *et al.*, 2012; NICE 2010). Unfortunately, there are still inequalities between countries. About 80% of deaths from CVD take place in countries with low-to-moderate prosperity, in which the frequency of multiple risk factors, especially obesity and diabetes mellitus (DM), tends to increase significantly (Finucane *et al.*, 2011). Interestingly, despite the general decline in the consumption of tobacco products, there currently exists three times higher risk for smoking in women because of the trend to start the habit at a younger age.

Among patients with CAD, the most common manifestations of atherothrombosis are myocardial infarctions with ST segment elevation (STEMI) and non-ST segment elevation myocardial infarction (NSTEMI). In-hospital mortality in STEMI varies, according to a variety of records, around 6–14%. Despite the development of pharmacotherapy and the invasive therapy, the mortality rate in 6 months after STEMI is still approximately 12%. Over the last decade, the proportion of STEMI has been reduced as compared to NSTEMI. Although, in the early years, the population of patients with NSTEMI acute coronary syndrome is characterised by lower mortality; after about 2 years, it is similar as in patients with STEMI (Roffi *et al.*, 2016).

Pathophysiology of the atherosclerosis formation process

In spite of a common pathophysiologic pathway, atherosclerotic lesions are very heterogeneous and the “high-risk plaque” of each vascular bed has unique characteristics. Insights into the disease have advanced beyond the notion of progressive occlusion of the coronary artery into the recognition that plaque disruption and superimposed thrombus formation are the leading causes of acute coronary syndromes and cardiovascular death.

Histologically, these rupture-prone (also called vulnerable or high-risk) lesions consist of a large core of extracellular lipid, a dense accumulation of macrophages, reduced numbers of vascular smooth muscle cells, and a thin fibrous cap. Hence, is not surprising that these plaques are less stable and have a greater propensity to rupture than the fibrous, collagen-rich plaques. Plaque disruption usually occurs at the weakest point (“shoulder”), where the cap is often thinnest and most heavily infiltrated with inflammatory cells (van der Wal

et al., 1994). Once the plaque is disrupted, the highly thrombogenic, lipid- rich core, with abundant tissue factor, is exposed to the bloodstream, triggering the formation of a superimposed thrombus that leads to vessel occlusion and subsequent ischaemic symptoms distal to it (Fuster, 1999).

In contrast with most high-risk coronary plaques, high-risk carotid plaques are considerably more stenotic. They are not lipid-rich but rather heterogeneous and very fibrous. Plaque disruption is often caused by an intramural haematoma or dissection that probably is related to the systolic stroke of blood against the resistance they offer (Glagov *et al.*, 1988). Although lipid accumulation in the carotid arteries is quite diffuse, has been reported the presence of ruptured lipid-rich plaques in patients with TIA and stroke (Yuan *et al.*, 2002). In addition, also the so-called "cryptogenic strokes" have an atherothrombotic origin. The source of emboli is usually a carotid or aortic thrombus (Cohen and Amarenco, 2002). Similarly, high-risk plaques of the lower extremities appear to be very stenotic and fibrotic (Ouriel, 2001). Available evidence suggests that in peripheral artery disease (PAD), plaque stenosis associated with hyper thrombogenicity of the blood seem to be major contributors to acute ischaemic syndromes (sudden ischaemic pain, gangrene). This is suggested by the high prevalence of known causes of a hyper thrombotic state of the blood, such as diabetes, cigarette smoking, and dyslipidaemia (McDermott *et al.*, 2001), in PAD patients (Sambola *et al.*, 2003). Conversely, high-risk plaques in the thoracic aorta frequently contain a high proportion of extracellular lipids and are characterised by a shift toward greater macrophage con-tent relative to smooth muscle cells in the cap. At autopsy, aortic plaques from persons who died of ischaemic heart disease often have ulceration and mural thrombosis (Davies and Wolf, 1993). Aortic plaque characterisation by magnetic resonance imaging (MRI) has confirmed their lipid-rich composition (Fayad *et al.*, 2000).

Pathophysiological studies have shown that the most common cause of formation of a blood clot is rupture of the fibrous cap, which separates the contents of the plaque from the blood (Hansson *et al.*, 2015; Nilsson *et al.*, 2017; Virmani *et al.*, 2006; Fuster *et al.*, 2005). Other mechanisms are damage of endothelial cells known as erosion on the surface of atherosclerotic plaque (plaque erosion) consisting of 30–35% of the acute coronary syndrome (ACS) and of approximately 2–7% endovascular calcifications (Virmani *et al.*, 2006; Fuster *et al.* 2005). Inflammatory changes ongoing in the atherosclerotic plaque cause a loss of stability making it vulnerable to rupture, the so-called unstable atherosclerotic plaque (vulnerable plaque). Unstable plaque is characterised by a thin fibrous cap (thin cap fibroatheroma—TCFA) covering the big necrotic core around which revolves

the inflammatory process and positive remodelling of the artery (Hansson *et al.*, 2015; Nilsson *et al.*, 2017; Virmani *et al.*, 2006; Fuster *et al.*, 2005). A similar transformation in the atherosclerotic plaque has also been observed in the internal carotid artery (ICA). This location is responsible for TIA and strokes (Hansson *et al.*, 2015). Positive remodelling of the artery proves that narrowing of its lumen does not have to be relevant, and it may not exceed 50–70% (Hansson *et al.*, 2015; Virmani *et al.*, 2006).

Endothelial dysfunction

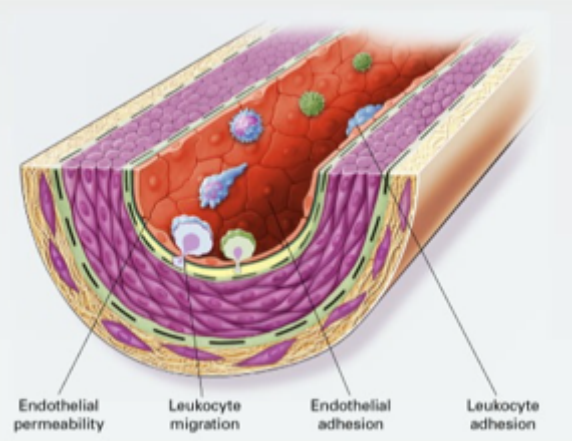


Figure 2*:Scheme of the Endothelial tissue and its physiological role. (Free stock image from the web)*

Endothelial dysfunction is a systemic, reversible disorder considered the earliest pathologic process of atherothrombosis (Behrendt *et al.*, 2002; Weiss *et al.*, 2002). Endocrine endothelial function, which consists in the synthesis and secretion of NO and PGI, is a prerequisite for the preservation of its integrity and correct relationship between it and flowing blood. Known atherosclerotic risk factors may interfere with this function. of encouraging the penetration of lipoproteins, monocytes and lymphocytes into the vessel wall. Inflammation in connexion with the accumulation of cholesterol, principally oxygen-modified LDL cholesterol (oxy-LDL), causes the proliferation of monocytes from peripheral blood, which is then converted into macrophages. It also stimulates the recruitment of myofibroblasts producing proteoglycans, the main substrate of extracellular matrix. Cytokines produced by macrophages infiltrates would lead to the degradation of the connective matrix tissue and smooth muscle cell necrosis and, consequently, to fibrous cap rupture. Gradual increase of the volume of the emerging plaque leads to abnormal blood flow, which in turn causes the oscillating shear stress and intensifies the atherothrombotic mechanisms stimulating further plaque growth (Ketelhuth and Hansson, 2016; Narula *et al.*, 2013). The resultant atherosclerotic plaque is made of fibrous cap of smooth muscle cells and connective tissue. The cap separates the lumen of vessels from the contents in which necrotic core is surrounded by inflammatory infiltrates containing macrophages, foam cells and lymphocytes. Angiogenesis plays an important role in the pathophysiology of plaque instability and plaque rupture. New blood vessels rarely penetrate from the main lumen, but more often from the vasa vasorum (Kumamoto *et al.*, 1995). They lack the cells constituting the vessel wall which

are fragile and porous, so that they become a source of local extravasation plasma protein and blood cells. Such bleedings in the plaque are frequent, may increase the volume of necrotic core and cause sudden progression of artery stenosis (Kolodgie *et al.*, 2003).

Role of inflammation in the process of the formation of atherosclerosis

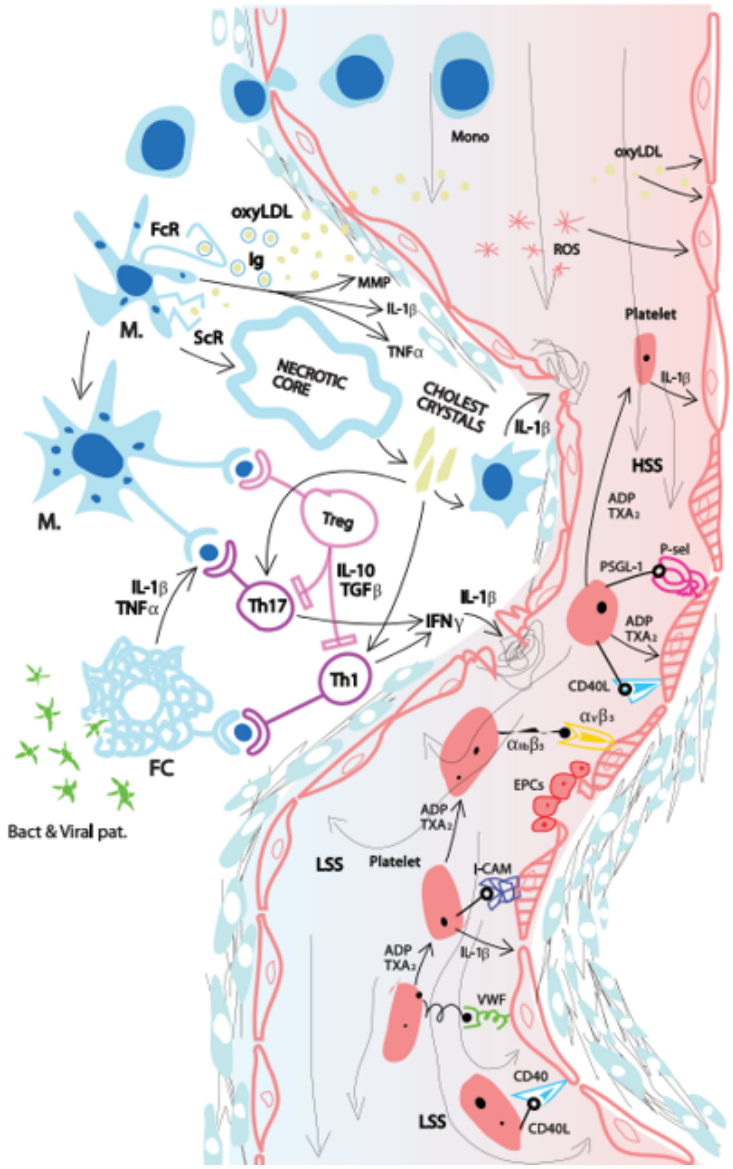


Figure 3: *Mechanism of plaque vulnerability and erosion (from Dziedzic et al., 2018)*

Currently, a greater role in initiation of changes is attributed to lymphocytes and mutual relations between them and macrophages, which cause varying degrees of inflammation activity (Figure 3). The main antigen that initiates and maintains inflammation in the vessel wall is oxy-LDL (Li and Ley, 2015). It is toxic to the vessel wall cells causing the immune system to try eliminating it. The presence of oxy-LDL derived antigens on the surface of the dendritic cells, macrophages and different types of lymphocytes regulates the activity of inflammation (Groom *et al.*, 2012). Antigen can also be protein of bacterial cells, viral, heat shock protein 60 or β2glycoprotein 1 (Jawień, 2008). It has been shown that circulating T lymphocytes, CD4+ and CD8+, differentiate preferentially in the direction of Th1, Th2 and Th17, which promotes the transformation from stable atherosclerotic plaque into unstable vulnerable plaque (Han *et al.*, 2007; Groom *et al.*, 2012). Arising Th cells produce different cytokines, including interleukin 2 (IL-2), which controls immune processes by influencing the maturation of lymphocytes Treg demonstrating immunosuppressive reactivity (Li and Ley, 2015; Han *et al.*, 2007; Groom *et al.*, 2012). It is now suspected that cell response (Th1, Th17) and its mediators: tumour

necrosis factor-α (TNFα), interferon-γ (INFγ) and interleukins (IL)—IL-1β, IL-12, IL-17, IL-18 are responsible for promoting the development of atherosclerosis, whereas humoral immune response (Th2) and its mediators: IL-2 IL-4, IL-5, IL-10, IL-13 have an inhibitory effect on this process [3]. Propagators of the ongoing inflammation are increased levels of proinflammatory cytokines (i.e. TNF-α, IL-6); soluble adhesion-intercellular adhesion molecule (ICAM), vascular cell adhesion molecule (VCAM), l-selectin and so-called acute-phase proteins—inflammation C-reactive protein (CRP), amyloid A and fibrinogen (Beręsewicz *et al.*, 2001).

The mechanism of lymphocytes penetration into the arterial wall is not exactly known. Probably this is done with the participation of chemokines and l-selectin (Li and Ley, 2015). Because most chemokine receptors are found on different cell types, research to clarify the mechanisms of lymphocytes homing in the atherosclerotic plaque is mostly inconclusive. Th1 and Th17 cells produce large amounts of IFNγ that activates inflammatory processes and expresses the transcription factor β (T-bet). These factors play a decisive role in the destabilization of atherosclerotic plaque. IFNγ activates APCs and macrophages, re-duces collagen synthesis and increases production of cytokines degrading extracellular matrix. An important role in the degradation of the extracellular matrix is also played by matrix metalloproteinases (MMP) (Friese *et al.*, 2009; Shah *et al.*, 1995). A strong factor in boosting the activity of MMP-9 (an enzyme that breaks down collagen type IV, which is component of the fibrous cap) is TNFα (Bahar-Shany *et al.*, 2010).

Treg cells inhibit the inflammatory reactions by IL production such as IL-10 and similarly acting transforming growth factor beta (TGFβ) (Cochain and Zernecke, 2017). They reinforce simultaneously the fibrous cap by stimulating the proliferation of smooth muscle cells and the production of collagen (Flego *et al.*, 2016). Functional balance between these T-cell types provides the stability of atherosclerotic plaque. The excess oxy-LDL in the blood penetrating into plaque results in the formation of the complexes with immunoglobulin. These complexes combine with Fc receptors (FcR) on the surface of macrophages and stimulate the secretory activity for MMP, IL-1β and TNFα. At the same time, the presentation of the oxy-LDL on the surface of macrophages stimulates Treg lymphocytes, which slow the inflammation down by reducing the activity of the Th lymphocytes. Such control system works well in people with low levels of risk factors. The binding of oxy-LDL to scavenger receptor located on macrophages, which is not subject to mentioned feedback, results in overloading of these cells leading to their death, releasing of free cholesterol and increasing volume of necrotic core (Moore and Tabas, 2011). The

release of such antigens triggers the mechanisms of the vicious circle by stimulation of the Th1 and Th17 instead of Treg lymphocytes (Ketelhuth and Hans-son, 2016).

Lately, attention is focused on the role of receptor programmed target death protein-1 (PD-1) presented on the naive CD8+ cells. It bears the responsibility for the so-called immune exhaustion observed in chronic inflammatory states (tbc, HIV) and cancer. There are suggestions that chronic stimulation of TCR by oxy-LDL leads to increased presentation of PD-1. The presence of PD-1 correlates inversely with the level of IL-2 produced by Th2. This can interfere with the CD8+ cell differentiation in the direction of Treg and potentiate their apoptosis leading to competitive advantage mechanisms acting as pro-inflammatory and destabilizing factors in the plaque (Zidar *et al.*, 2016).

Probably within a plaque, there are three subtypes of macrophages. The most common are classically activated macrophages M1 induced by INFγ or Th1 and Th17 lymphocytes cytokines. The second group are macrophages M2 induced by cytokines of helper lymphocytes Th2 (IL-4 and IL-13). They produce the anti-inflammatory acting cytokines IL-10 and TGFβ (Chen *et al.*, 2016). Probably, there is a third group of macrophages presenting CD163+, activated by haemoglobin, which do not produce pro-inflammatory cytokines and have reduced ability to produce inducible nitric oxide synthase (iNOS). A decrease in the levels of intracellular iron ions within the macrophage probably plays a leading role in the transcription of genes protecting these security cells from the accumulation of lipids. This is done by increase in the levels of ferroportin-1 leading to reduction of free radicals (-OH) production as result of iron ions accumulation and depletion. One of the key regulators of atherosclerotic plaque stability may prove to be hepcidin, responsible for ferroprotein-1 degradation, resulting in the accumulation of iron ions, the accumulation of intracellular lipids and apoptosis of macrophages. Hepcidin blockage inhibits the development of atherosclerosis by regulating ATP-binding protein subfamily G (Chen *et al.*, 2016). A significant role in the weakening of the fibrous cap, consequently causing it to rupture, is played by T-cells CD4 + CD28nul. They produce a significant amount of INFγ and TNFα, strongly stimulating macrophages. They also have cytotoxic properties in relation to fibrous cap, smooth muscle cells and are apoptosis resistant. These cells are presenting cytotoxic immunoglobulin (killer immunoglobulin) on their cell membrane that acts as cytotoxic receptors (Ig-like receptors). They also produce cytolytic enzymes against the endothelial cells that directly kill them, such as perforins, granzyme A and granzyme B, which are usually present in killer T cells (KTC) and natural killer cells (NK) (Liuzzo *et al.*, 2007; Meeuwsen *et al.*, 2017). It has been shown that the number of these cells in the circulation is an important prognostic for occurrence and course of ACS

(Liuzzo *et al.*, 2007). Production of pro-inflammatory proteins, such as IL-1β, chemokine (C-C motif) ligand 2 (CCL2), chemokine (C-C motif) ligand 3 (CCL3), E-selectin (SELE), ICAM-1, MMP-3 and the MMP-9, involved in the process of destabilising atherosclerotic plaque denotes a genetic profile connected with polymorphism of many genes. Polymorphism of this plays an important role in the susceptibility to risk factors for atherosclerosis and to changes in already existing atherosclerotic plaque. What's more, single nucleotide polymorphisms located within the regions of functional genes for these proteins may affect their concentration and activity causing further clinical implications (Biscetti *et al.*, 2015). By examining the mechanisms leading to the development of atherosclerosis, it was shown that in these processes, beyond the stimulated endothelial cell and cells of the immune system, also vascular smooth muscle cells (VSMC) are involved. VSMC function is not just limited to the production of extracellular matrix in the vessel wall. It has been shown that in response to a stimulus, these cells may change the type of produced extracellular matrix and thus affect the lipid content in the vessel wall and the multiplication of other cells. Under specific conditions, they can also take over the function of other cells, for example macrophages, and due to the expression of the relevant receptors acquire absorption capacity of fat by mimicking foam cells. While taking over some functions of endothelial cells, they can produce cell adhesion molecules, VCAM-1 or ICAM-1. In addition, being a component of atherosclerotic plaque, they can also produce cytokines—platelet-derived growth factor (PDGF), TGFβ, IFN and monocyte chemoattractant protein 1 (MCP-1) (Doran *et al.*, 2008). Under the influence of these cytokine, extracellular matrix degradation occurs into fi-brous cap. It is weakened further due to result of the apoptosis of smooth muscle cells and cell death due to primary necrosis (Crisby *et al.*, 1997). Contact of the flowing blood with the content of the ruptured plaque activates processes of coagulation, which can occur rapidly. A large amount of tissue factor (TF) liberated by inflammation tissue activates plasma factor VII, which runs the enzymatic coagulation cascade. TF forms a complex with factor VII, activating it to active form (VIIa). The complexes TF/VIIa activate factors IX and X, leading to thrombin generation. The consequence of this cascade of activation is rapid formation of the intravascular thrombus (Kaplan and Jackson, 2011).

Erosion on the surface of the atherosclerotic plaque as a cause of ACS and stroke

Epidemiological studies have shown that myocardial infarction may occur in people with normal levels of LDL cholesterol. In addition, as demonstrated by pathomorphological and clinical studies using optical coherence tomography, 30-40% of patients with

vascular thrombosis atherosclerotic plague have no inflammatory features (Falk *et al.*, 2013). The morphology of such plaques is completely different from the above, subjected to the inflammatory changes. The blood clot formed on its surface is in direct contact with the intima at a place completely devoid of the endothelium. Fibrous cap is well demarcated and includes numerous smooth muscle cells and an extensive connective tissue forming an extracellular matrix (Virmani *et al.*, 2000). The interior of the well-demarcated plaque contains few macrophages and lymphocytes. As well, the profile of patients with ACS, who have been found to have this type of plaque, differed from the profile of patients who suffered vulnerable plaque. In available reports, these patients were younger, 80% of these were premenopausal women, and frequent tobacco smokers (Falk *et al.*, 2013). The mass of the plaque, which was the basis of thrombosis, was less than in the case of plaque rupture and often it was nonconcentric (Virmani *et al.*, 2006). In contrast to the inflammatory plaques that show positive remodelling, arteries affected by erosion are characterised by negative one. Demonstrated characteristics suggest a different mechanism in formation of a blood thrombus, which, like in the case of plaque rupture, can cause both the closure of an artery and peripheral embolism, more often associated with such morphology of the intravascular changes (Falk *et al.*, 2013; Hannsson *et al.*, 2015). However, the mechanism of the formation of this type of inter arterial thrombosis has not been fully understood. It is suspected that a decisive role in its formation plays abnormal blood flow due to arterial plaque stenosis, which causes changes in shear stress, endothelial dysfunction that covers plaque affecting its anti-inflammatory and anti-thrombotic signals of the endothelium (Hannsson *et al.*, 2015). Laminar flow disorders more often occur in places of bifurcation and in the folds of the arteries. The correct endocrine function of the endothelium creates normal shear stress, which is the force of friction between the flowing blood and cellular layer covering the interior surface of the vessel. Both low and improperly high shear stress interfere with endothelial functions and can produce prothrombotic state. In the case of atherosclerotic narrowing of the artery, both pre- and post- stenosis flow are slowed down whereas at the apex of the plaque, the flow is accelerated. This creates conditions conducive to damaged endothelium. Low shear stress and turbulent blood flow facilitate the accumulation of lipids, the recruitment of inflammatory cells and increased expression of adhesion molecules and proteases (Chen *et al.*, 2016). The correct vascular flow induces the enzyme systems that prevent the expression of pro-inflammatory and pro-thrombotic genes and at the same time promote the layout security by activating the endothelial NO synthase. Simultaneously, high shear stress stimulates the synthesis of several types of microRNAs that interact with Krüppel-like factor 2

(KLF2) and nuclear factor erythroid cell-specific 2-related factor2 (Nrf2), which support the interaction of anti-inflammatory and anti-thrombotic pathways (Nilsson, 2017). Accelerated flow of blood within the largest narrowing and supra-physiologically high shear stress suppresses these systems, which prevent inflammation and activate the prothrombotic processes, and may be the reason for damage to the endothelial cells and the activation of inflammation with further consequences of thrombosis (Nilsson, 2017). These biomechanical force fluctuations associated with shear stress are particularly apparent in people with hypertension and their effects are intensified under the influence of other atherosclerotic risk factors, such as hypercholesterolemia, advanced glycation end-products in diabetes, tobacco smoking, vasoactive amines and immune complexes. These factors in terms of alternating shear stress can lead to endothelial dysfunction (Fuster *et al.*, 2005). In areas of damaged vascular endothelium and high shear stress, there are abnormal interactions between thrombocytes and endothelial cells. Another suggested mechanism that can coexist with described above is increased tendency of endothelial cells covering the atherosclerotic plaque to apoptosis. It contains a hyaluronan molecules in its structure, identical to that found in Gram+ bacteria, which can result in the recognition of these cells as foreign by the immune system, leading to their destruction, and thus initiating thrombotic processes (Hannsson *et al.*, 2015). Activation of endothelial cells, arising not only under the influence of flow disorders but also under the influence of reactive oxygen species (ROS), and circulating lipoprotein leads to the expression of surface adhesion proteins like P-i L-selectin and von Willebrand factor (VWF), which support the mutual relationships between endothelial cells and platelets (Kaplan and Jackson, 2011). Endothelial cells present selectin on their surface, such as selectin-P, which stimulate platelets to produce glycoproteins, including glycoprotein-1. Complex selectin-P-glycoprotein-1 (PSGL-1) allows platelet adhesion and rolling but not tight binding of endothelial cells. Glycoprotein surface receptors GPVI and αvβ3 bind vascular walls collagen, which activates the channels for intracellular calcium flux, causing the activation of receptors αIIbβ3 and release of ADP and thromboxane A2 (TxA2). This initiates a platelet thrombus protruding into artery lumen and is essential for the later stages of thrombus formation. The combination of glycoprotein platelet receptors αIIbβ3 with VWF and fibrinogen stabilises the platelet clot, which in clinical practice becomes the target of preventing thrombolysis by pharmacological interventions. Impregnation of so-formed platelets conglomerate on the inner surface of the artery by fibrin finally decides the formation of a stable thrombus (Kaplan and Jackson, 2011).

Mutual activation of endothelium and platelets largely depends on the IL-1β, accumulated in granules of platelets and activated by mRNA already several hours after thrombin stimulation or adhesion-dependent integrins. Stimulation of platelets by IL-1β in-duces secretion of IL-6 and IL-8 and the expression of surface adhesion molecules-ICAM-1, αvβ3 and hemotactic monocyte's protein-1 (MCP-1). Due to these mechanisms, platelets are capable of recruiting monocytes and neutrophils from blood and then causing them to migrate and participate in the above-described pathophysiology of atherosclerotic plaque vulnerability. Presented mechanisms ensure the presence of activated thrombocytes in the centre of the pathophysiology of the process, not only with the thrombus formation but also as an important part in the activation and maintenance of the inflammatory process. Human platelets are capable of producing all types of toll-like receptors (TLR). It has been shown that higher TLR expression in women may be responsible for differences in cardiovascular risk profile, as well as the tendency for a higher incidence of ACS in the superficial thrombosis mechanism, without active inflammatory features in the atherosclerotic plaque (Chen *et al.*, 2016). Damage to the endothelial cells activates endothelial progenitor stem cells (EPCs) de-rived from bone marrow, which proliferate at the place of damage and may prevent the described processes that leading to the thrombus formation. These repair mechanisms are defective in patients with diabetes, characterised by a general weakness of repair capacity of damaged tissues (Virami *et al.*, 2016).

High-risk blood

Two thirds of ACSs are caused by the disruption of a high-risk atherothrombotic plaque with superimposed thrombus formation. In one third of ACSs, particularly in sudden coronary death, there is no rupture of a high-risk atherothrombotic plaque but only a superficial erosion of a markedly stenotic and fibrotic lesion (Virmani *et al.*, 2000). Thrombus formation in such cases may depend on a hyper thrombogenic state triggered by systemic factors. Indeed, several cardiovascular risk factors, including elevated LDL cholesterol, cigarette smoking, and hyperglycaemia, have been associated with increased blood thrombogenicity (Rauch *et al.*, 2001). Circulating tissue factor has been associated with increased blood thrombogenicity in patients with unstable angina (Kaiita *et al.*, 1997) and chronic coronary artery disease. Blood levels of tissue factor have also been shown to predict outcome in patients with unstable angina (Soejima *et al.*, 1999). Several lines of evidence support the hypothesis that circulating apoptotic cells and cellular microparticles may play a significant role in blood thrombogenicity. Patients with ACS have elevated levels of circulating tissue factor, and there is evidence that acute thrombosis may be

initiated by membrane-bound circulating tissue factor originating from activated or injured cells (Giesen *et al.*, 1999). It is believed that a major source of blood-borne tissue factor could be the circulating microparticles, which are endowed with potent procoagulant potential, attributable to the presence of phosphatidylserine on their surface (Mallat and Tedgui, 2001). A significant increase in the number of circulating endothelial cells, some of them apoptotic, has also been reported in patients with ACS (Mutin *et al.*, 1999). The circulating procoagulant microparticles may also contribute to the blood thrombogenicity of patients with hyperlipidaemia or high blood glucose concentrations; these vascular risk factors are known to be responsible for increased apoptotic activity in vitro (Dimmeler *et al.*, 1997). Elevated LDL cholesterol levels have been found to increase blood thrombogenicity and the growth of thrombi under defined rheology conditions (Dangas *et al.*, 1999). Reducing LDL cholesterol levels with statins has been shown to decrease thrombus growth by approximately 20% (Rauch *et al.*, 2000a). The extent to which this antithrombotic effect contributes to the reduction of total vascular events, including death, coronary events, and stroke, is a matter of debate (Fuster, 1999). Diabetic patients, especially those with poorly controlled diabetes, have increased blood thrombogenicity (Osende *et al.*, 2001) Platelets from patients with diabetes have been shown to have increased reactivity and hyperaggregability and expose a variety of activation-dependent adhesion proteins. Abnormal platelet function is reflected by in-creased platelet consumption and increased accumulation of platelets on the altered vessel wall. The increased procoagulant activity in diabetes is also attributed to leukocytes, which may, in part, activate the tissue factor pathway and contribute to the high blood thrombogenicity. (Rauch *et al.*, 2000b). Fibrinogen concentration was found to be associated with increased blood thrombogenicity (Koenig, 2003). Several of the classic risk factors have been shown to modulate fibrinogen levels. Fibrinogen levels tend to be higher in patients with diabetes, hypertension, obesity, smoking habit, and sedentary lifestyles (Meade *et al.*, 1987). However, further clinical trials are needed before it can be determined whether fibrinogen is directly involved in the pathogenesis of atherothrombosis or is merely a marker of the degree of vascular damage. As previously described, lipid-rich atherosclerotic plaques contain tissue factor associated with macrophages within the lesion (Toschi *et al.*, 1997), which may account, in large part, for the high thrombogenicity of these lesions. In addition, specific inhibition of the tissue factor pathway by its physiologic inhibitor, tissue factor pathway inhibitor (TFPI), significantly reduces plaque thrombogenicity (Badimon *et al.*, 1999).

Genetic of atherothrombotic diseases

The large part of cardiovascular diseases are polygenic and derive from both heritable and environmental contributions (Goldstein and Brown, 2009).

Approaches to identifying the genetic causes of polygenic cardiovascular diseases (and other polygenic diseases) before completion of the draft sequence of the human genome were largely unsuccessful. A decade later, hundreds of loci associated with many cardiovascular diseases and traits have been identified. The advent of high throughput sequencing technology enlightened the fact that both common multifactorial and rare cardiovascular diseases, classically thought to be monogenic, result from the contribution of more genes that cause or modulate the phenotype (Lee *et al.*, 2006).

In the last three decades, genetic studies have evolved in parallel with progress in the knowledge of the genome and its variation and the techniques of DNA analysis allowing to study comprehensively the association between genetic variants and diseases (Figure 4). Until the advent of next generation sequencing (NGS) the main approaches utilised were represented by genetic epidemiology studies.

There are three types of genetic epidemiology studies aimed at assessing the association between family history and the onset of stroke: linkage, the candidate-gene approach, and genome-wide association studies (GWAS) (Markus HS, 2012).

Linkage analysis studies rely on the assumption that a genetic variant causing a disease phenotype is coinherited with the disease. Linkage studies are based upon families in which there are multiple cases with a disease. Genetic variants, used as 'markers' of the different areas of the genome (genetic loci), are genotyped in affected and unaffected family members and the coinheritance of variants with the disease is statistically tested. Markers inherited together with the disease pinpoint an area of the genome that contains variants that are likely to be implicated in the development of the disease. Genes encompassed in the locus are then sequenced in order to identify specific disease-causing gene mutations (Hirschhorn, 2005).The approach of linkage studies has proved to be successful in monogenic diseases for the identification of gene(s) that harbour disease-causing mutations but was by far less successful in the identification of genetic loci that contribute to the occurrence of common, complex diseases.

Candidate gene studies compare the frequency of selected SNPs, usually in the frame of retrospective case–control studies carried out in subjects with and without a disease. The SNPs are selected a priori, based on their localization in genes which encode proteins with a known function in a biological pathway that is putatively implicated in the pathophysiology of the disease. For instance, the frequency of SNPs located in genes implicated

in lipoprotein metabolism or blood coagulation is compared in cases with myocardial infarction (MI) or stroke and in healthy controls.

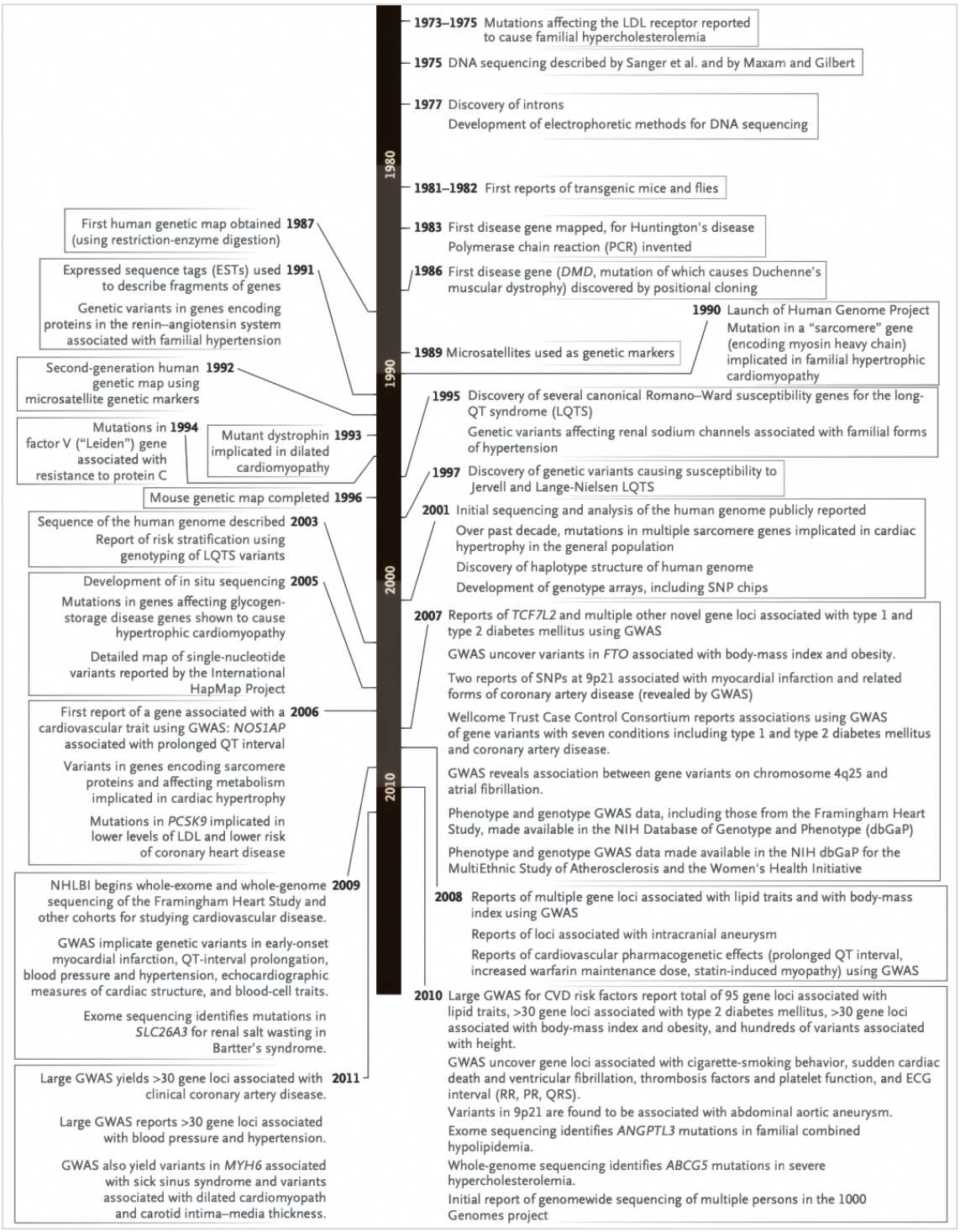


Figure 4:*Timeline of the first 35 years of genetic and genomic research in cardiovascular medicine (from O'Donnell and Nabel, 2011).*

Genome-wide association studies have become possible after the publication of the International Haplotype Map Project (HapMap), and thanks to the development of array-based platforms that enable to investigate up to 1 million SNPs in cases and controls of a certain disease (or other phenotypic traits).

These studies are characterized by a multi-stage procedure (McCarthy *et al.*, 2008):

1- A complex disease, with a high prevalence in the population (GWAS typically need thousands of participants) is chosen, provided there is already evidence for a genetic basis for the condition/disease object of the study.

2- Cases and controls are recruited, possibly with bias-free modalities that guard against the selection of individuals not representative of the reference population. More rarely, GWAS are conducted in the frame of prospective, population-based cohort studies.

3- Genetic variants are then sequenced both in cases and controls. Genotyping is performed using arrays. Information on additional (hundreds of thousands) SNPs, the so called 'imputed SNPs', is obtained indirectly by using HapMap data on variants that are inherited together in a certain population owing to linkage disequilibrium.

Statistical analysis is then performed to identify variants associated with the disease. A variant is considered to be associated with a disease when its prevalence is significantly greater in cases than controls. Since testing for the association of approximately 1 million variants exposes to the risk of false positive findings owing to multiple testing, the conventional statistical significance threshold used to define a true association in GWAS is 5×10^{-8} (corresponding to 0.05 after adjustment for 1 million independent tests). In GWAS, few variants do reach such stringent threshold, even when more than 4000 individuals are investigated. Usually, the first genome-wide screen allows to identify a number of variants that show suggestive association with the disease and that subsequently need retesting in other cohorts to confirm or exclude association (replication stage). GWAS identify regions of the genome (loci) rather than variants of specific genes. Indeed, a specific variant identified by GWAS may simply represent the signal of one or more hidden variant in linkage disequilibrium with that variant. A large number of genetic loci have been identified in association with MI, coronary artery disease (CAD), thrombotic events, lipid traits, and other circulating biomarkers related to atherothrombotic disease (Rader, 2015). Genome-wide association studies (GWAS) have identified ≈50 discrete genetic loci that are significantly associated with MI/CAD (Deloukas *et al.*, 2013). The first GWAS locus discovered to be associated with MI/ CAD was a locus on chromosome 9p21.3 spanning an ≈60 kilo base region, which has come to be called the chromosome 9p21.3 coronary heart disease (CHD)–Associated Region (C9CAR). This locus remains the most robustly associated with CAD, accounts for ≈20% of the attributable risk for CAD, and is also associated with other vascular traits, such as aneurysms. However, the mechanism linking this locus to CAD remains to be elucidated.

One of the earliest novel loci found to be associated with MI/CAD was the *CXCL12* gene locus encoding the chemokine CXCL12, or stomal-derived factor 1. Variation at the *CXCL12* locus is also associated with endothelial progenitor cell numbers. There has been intense interest in understanding the functional biology underlying these robust genetic associations. Several articles have addressed the biology of CXCL12 and its major receptor CXCR4 in atherosclerosis and vascular biology (Michineau *et al.*, 2014; Mehta *et al.*, 2014). One of the most interesting new GWAS loci associated with atherosclerotic cardiovascular disease is HDAC9 that has been associated with both CAD1 and ischaemic stroke (Bellenguez *et al.*, 2012). Histone deacetylases (HDACs) modulate gene expression by deacetylation of histone and non histone proteins. Thus, this genetic finding suggests the potential of a direct link between epigenetic modification by HDAC9 and atherosclerotic vascular disease (Kaluza *et al.*, 2013).

Lysophosphatidic acid (LPA) and sphingosine-1-phosphate are bioactive lipids that act through G-protein–coupled receptor–mediated pathways and have increasingly been implicated in vascular biology and atherosclerosis. Lipid phosphate phosphatase 3 (LPP3), encoded by the *PPAP2B* gene, dephosphorylates and inactivates the signalling actions of LPA and sphingosine-1-phosphate (Smyth *et al.*, 2014). The *PPAP2B* locus has been identified as genome-wide significantly associated with CAD independent of other traditional risk factors. Data using expression quantitative trait locus analysis suggest that the allele associated with increased CAD risk is associated with reduced leukocyte expression of *PPAP2B* (Rader, 2015).

Plasma lipoprotein(a) (Lp(a)) levels are highly genetically determined and have been associated with increased risk of CVD in multiple observational studies. Lp(a) has emerged in the past few years as a poster child for Mendelian randomisation, in that plasma levels are strongly associated with incident cardiovascular events and genetic variants at the *LPA* gene locus specifically associated with plasma Lp(a) levels are them-selves strongly associated with cardiovascular events. Nevertheless, important questions remain unanswered. One key question has been whether Lp(a) levels are associated with incident cardiovascular events in people with overt CAD at the time of baseline measurement. The results of the Long-Term Intervention with Pravastatin in Ischemic Disease (LIPID) trial suggests that elevated Lp(a) could be causal for promoting progression and destabilisation of coronary plaques even in patients who already have overt CAD (Nestel *et al.*, 2013).

Genetic variation in the *LPA* gene results in a highly polymorphic protein that includes multiple copies of the kringle domain and is a major factor affecting plasma Lp(a) levels. However, some of the variations in Lp(a) levels attributed to the *LPA* gene are

independent of kringle copy number variation. Has been addressed the relationship of a relatively common (minor allele frequency, 3%) null allele of *LPA* (rs41272114) with Lp(a) levels and prevalent CAD in the Precocious Coronary Artery Disease (PROCARDIS) study. This single-nucleotide polymorphism results in alternative splicing and premature truncation, generating an apoA protein that cannot covalently bind to apoB to form a mature Lp(a) particle. Carriers of the null allele were found to have significantly lower Lp(a) levels; importantly, they also had a significantly reduced risk of CAD. This finding adds to the growing body of data supporting a causal role for Lp(a) in CAD and refines our understanding of this relationship by indicating that it extends beyond genetic factors that influence Lp(a) levels solely through isoform size (Kyriakou *et al.*, 2014).

Tissue-type plasminogen activator (tPA) is a glycoprotein enzyme made by endothelial cells that cleaves plasminogen to form plasmin, itself an active enzyme that lyses fibrin-containing clots. Circulating tPA is mostly found to be associated with its inhibitor plasminogen activator inhibitor-1 as part of an inactive complex. Plasma tPA levels are, somewhat counterintuitively given that tPA promotes clot lysis, positively associated with the risk of incident cardiovascular events. Factors regulating plasma levels of tPA are incompletely understood, but one source of variation is, as with most circulating proteins, genetic in nature. The meta-analysis of 14 studies comprehending genome-wide genotype data and plasma tPA quantification, involving a total of ≈27 000 subjects revealed that one genome-wide significant locus included the PLAT gene, which encodes tPA itself. The other 2 loci included the genes *STXBP5* (which encodes the protein syntaxin-binding protein 5) and *STX2* (which encodes the protein syntaxin 2), both are novel findings. Each locus harboured a strong expression quantitative trait locus for the respective gene but not for other genes at the loci, suggesting that these may be the causal genes (Huang *et al.*, 2014; Chasman *et al.*, 2014).

Syntaxins are members of a family of membrane-integrated soluble N-ethylmaleimide-sensitive factor attachment protein receptor proteins that participate in exocytosis. Provocatively, siRNA studies in vascular endothelial cells revealed that silencing of *STXBP5* decreased tPA release, whereas silencing of *STX2* increased the tPA release. Previously reported GWAS for von Willebrand factor had also identified the *STXBP5* and *STX2* loci as genome-wide significantly associated with levels of von Willebrand factor. Syntaxin-4, another member of the syntaxin family, is required for the release of von Willebrand factor from intracellular endothelial Weibel-Palade vesicles. Combined, these data suggest a broad role for syntaxins in the release of circulating hemostatic factors by endothelial cells (Rader, 2015).

CD14 is a glycosylphosphatidylinositol-anchored membrane glycoprotein expressed on neutrophils, monocytes, and macrophages. On binding of many proinflammatory ligands, it participates in the activation of intracellular proinflammatory signalling pathways. As with many cell surface receptors, CD14 can be enzymatically cleaved to generate a soluble form (sCD14), and this cleavage is induced by inflammatory stimuli, leading to increased sCD14 levels in the setting of acute and chronic inflammatory conditions. Thus, sCD14 is a circulating biomarker of potential interest in the setting of atherothrombotic disease. Reiner *et al.* measured the baseline levels of sCD14 in the Cardiovascular Health Study (CHS) involving >5000 subjects aged >65 years. Plasma levels of sCD14 were higher in people of European ethnicity and female sex and were positively correlated with smoking, hypertension, diabetes mellitus, and other inflammatory biomarkers (C-reactive protein, interleukin-6, and fibrinogen). They were associated with ankle-brachial index and carotid intimal-medial thickness and strongly predicted incident cardiovascular events and all-cause mortality (Reiner *et al.*, 2013). A genome-wide association analysis of sCD14 levels identified 2-genome-wide significant loci. One was the CD14 structural locus on chromosome 5q21, including a novel African ancestry–specific allele of CD14 associated with lower sCD14. The second locus included the gene PIGC, which encodes an enzyme required for the first step in glycosylphasphotidylinositol-anchored biosynthesis. A missense variant of PIGC, Pro266Ser, was noted for the first time to be significantly associated with higher plasma sCD14 levels. This finding suggests that defective glycosylphosphatidylinositol-anchored synthesis may result in increased release of sCD14 from cells into the circulation. In women with HIV or hepatitis C virus, plasma sCD14 levels were positively correlated with carotid intimal-medial thickness and atherosclerotic lesions. Whether sCD14 is a mediator of atherosclerotic disease or simply a biomarker remains to be definitively established (Shaked *et al.*, 2014).

The protein C (PC) pathway is critical in preventing inappropriate blood coagulation. Circulating protein C is activated on the surface of vascular endothelial cells after binding to its endothelial PC receptor (EPCR) and being presented to the thrombin-thrombomodulin complex. Activated PC, in conjunction with its cofactor protein S, reduces thrombin generation by degrading the coagulation cofactors Va and VIIIa. Genetic variation in this pathway can influence the risk of venous thromboembolism (VTE).

In cultured endothelial cells, the missense variant (Ala455Val) in the *THBD* gene encoding thrombomodulin was associated with increased cellular thrombomodulin, reduced soluble thrombomodulin in media, and increased PC activation. Subjects carrying the A455V allele had reduced soluble thrombomodulin levels, increased circulating activated

PC levels, and importantly reduced VTE risk. These results established that a missense variant in thrombomodulin modulates its activity, the generation of activated PC, and the risk of VTE (Navarro *et al.*, 2013).

In a follow-up study about common haplotypes H1 and H3 in *PROCR*, the gene encoding the EPCR, for their association with EPCR expression and risk of VTE, studies in cultured endothelial cells showed that the H1 haplotype was associated with increased membrane bound EPCR, increased PC activation, and reduced soluble EPCR in the media. Subjects carrying the H1 haplotype were found to have increased plasma activated PC levels, reduced plasma soluble EPCR levels, and most notably, reduced VTE risk. In contrast, the H3 haplotype was associated with reduced membrane bound EPCR, reduced PC activation, and increased soluble EPCR in media. Subjects with the H3H3 genotype had reduced plasma activated PC levels, increased plasma soluble EPCR levels, and an increased VTE risk. These results indicate that genetic variation at the *PROCR* locus influences the activity of the EPCR and thus the risk of VTE (Medina et., 2014).

Wu *et al.* performed targeted gene sequencing of exon 3 of *PROC* (encoding PC) and exons 2 and 3 of *PROCR* (encoding EPCR) in 653 patients with VTE and 627 healthy controls. Three subjects were found to have private mutations in PROC that affected the protein sequence (Arg-1Cys, Arg9Cys, and Val34Met), and all had decreased synthesis, with 2 also showing reduced ability to be activated. Two subjects had private missense mutations in *PROCR* (Arg96Cys and Val170Leu) that demonstrated reduced affinity for fluorescently labelled PC. Although the overall numbers were small, these findings suggest that private mutations in *PROC* or *PROCR* that impair PC–EPCR interactions may be associated with an increased risk of VTE (Wu *et al.*, 2013).

Elevated triglyceride levels are associated with increased CAD risk, but the causal nature of this relationship has been uncertain (Langsted *et al.*, 2011). A gain-of-function variant in the *LPL* gene, S447X, is associated with reduced triglyceride and reduced risk of cardiovascular disease (Varbo *et al.*, 2013) consistent with a protective effect of lipoprotein lipase (LPL) in not only reducing triglyceride levels but also reducing the risk of CVD. Common genetic variants that influence triglyceride levels are significantly associated with CAD risk even after adjusting for their effects on other lipid traits (Do *et al.*, 2013). Loss-of-function mutations in *APOC3* that reduce plasma levels of apoC-III (an inhibitor of LPL) are associated with lower triglyceride and decreased risk of coronary calcification30 and clinical CAD. In contrast, loss-of-function mutations in *APOA5* (which encodes apoA-V, an activator of LPL) are associated with elevated triglycerides (Jørgensen *et al.*, 2014) and in some cases with increased CAD risk. Remarkably, in a hypothesis-

free exome sequencing experiment in people with early MI compared with older controls without MI identified a significant enrichment of rare *APOA5* mutations in early MI cases (Do *et al.*, 2015). These findings establish that disruption of apoA-V protein function increases the risk of cardiovascular disease and makes it imperative to better understand the normal physiology of apoA-V and the effect of structural mutations on its function.

Angiopoietin-like proteins (ANGPTLs) are secreted proteins characterised by key structural motifs, and several of which play roles in triglyceride metabolism. ANGPTL3 and ANGPTL4 are the 2 members of the family that have been most extensively studied, and for which human genetic data exist, supporting a causal role in modulating the metabolism of triglyceride-rich lipoproteins. ANGPTL3 reversibly inhibits and ANGPTL4 irreversibly inhibits LPL. In mice, over-expression of ANGPTL3 or ANGPTL4 causes elevated triglycerides and deletion of ANGPTL3 or ANGPTL4 results in a decrease in triglyceride levels. Common variants at the *ANGPTL3* and *ANGPTL4* loci are associated with triglyceride levels (Teslovic *et al.*, 2010) and non-synonymous loss-of-function variants in both proteins are associated with lower triglyceride levels (Romeo *et al.*, 2007). Exome sequencing in a family with decreased triglyceride, LDL cholesterol (LDL-C), and high-density lipoprotein cholesterol identified loss-of-function mutations in *ANGPTL3*. Mehta *et al.* determined plasma ANGPTL3 and ANGPTL4 levels in 1770 subjects and assessed their association with lipids and metabolic traits. Plasma ANGPTL3 levels were positively associated with LDL-C and high-density lipoprotein cholesterol levels but not triglyceride levels. In contrast, plasma ANGPTL4 levels were negatively associated with LDL-C and high-density lipoprotein cholesterol and positively associated with triglycerides. In addition, ANGPTL4, but not ANGPTL3, levels were positively associated with fasting blood glucose and metabolic syndrome (Mehta *et al.*, 2014(b)). Thus, although ANGPTL3 and ANGPTL4 both inhibit LPL, their in vivo physiology is complex and additional studies of their plasma levels incorporating genetic data are needed. One approach is to take advantage of individuals with loss-of-function mutations for further deep phenotyping. Robciuc *et al.* recruited homozygotes and heterozygotes with the S17X loss-of-function mutation in *ANGPTL3* and age- and sex- matched non-carrier controls. Postheparin plasma LPL mass and activity were significantly higher, and plasma free fatty acid, insulin, and glucose were significantly lower in S17X homozygotes when compared with S17X heterozygotes and controls. No changes in hepatic lipase or endothelial lipase activities were noted, even in homozygotes (Robciuc *et al.*, 2013). These results suggest that ANGPTL3 may influence insulin sensitivity and glucose metabolism, in addition to its role in lipid metabolism.

Early detection with non-invasive imaging technology

Several imaging platforms are used for molecular imaging. Some are already being used in current clinical practice and contributing to clinical decision-making, whereas others are currently at advanced development stages (Wang and Peter, 2017; Wildgruber *et al.*, 2013). Although each of these imaging modalities has its own strengths and weaknesses, they are often complementary to one another and hence lead to the development of multimodality and hybrid imaging.

X-ray–based Computed tomography (CT) has been used for imaging of anatomic structures based on its short scanning time and its high spatial resolution. Invasive coronary angiography, based on the use of a radio-opaque contrast agent and X-rays, is the current gold standard for the diagnosis of coronary artery disease. Recently, coronary CT angiograms have become increasingly used as a gatekeeper for further invasive diagnostic procedures, particularly coronary angiograms (Shaw *et al.*, 2012).

Magnetic resonance imaging (MRI) incorporates high spatial and temporal resolution, thereby providing excellent soft tissue contrast and functional imaging capabilities with no requirement of radiation (Wang and Peter, 2017).

Positron emission tomography (PET) provides high sensitivity with relatively limited spatial resolution. PET is often applied in a hybrid approach with CT imaging. Most recently, the hybrid approach of PET and MRI has become highly attractive based on technical advances in the development of positron detectors that can be used in MRI scanners (Wang and Peter, 2017).

Single-photon emission computed tomography (SPECT) technology has similar properties to PET, but it comes with slightly lower cost and higher general availability. The tracers for SPECT also have a longer half-life when compared with those of PET ((Wang and Peter, 2017).

Ultrasound imaging is a low-cost modality, with scanners generally available in hospitals and many outpatient settings. Newer generations of ultrasound scanners are also highly portable compared with most other imaging technologies, which enables imaging to be performed at the bedside, in emergency situations, or outside of hospitals. Ultrasound is real-time, has high temporal resolution, and does not involve ionising radiation; however, ultrasound has restricted depth penetration and is operator dependent. The prototypical contrast agent for ultrasound imaging is micro-bubbles, which are already Food and Drug Administration approved (Wang and Peter, 2017).

In addition to these established modalities, other highly promising imaging platforms are currently being used in preclinical small-animal imaging but are still under development

for molecular imaging in patients. The most promising is optical imaging, which offers low cost, obviates radiation, and is highly versatile based on its simultaneous multispectral recording and high resolution. But this modality is so far limited by the low depth penetration of light through tissues, and, therefore, for cardiovascular applications, the accessibility of arteries deep in the tissue/body is a major challenge, at least with the currently available technology. This type of application would require invasive (catheter based) approaches. Nevertheless, both fluorescence molecular tomography and bioluminescence imaging provide efficient, low-cost imaging which is fast and sensitive. However, further technical advances are necessary to overcome the current restriction to low depth penetration.

Most recently, photo-acoustic imaging has developed into a highly promising molecular imaging technology. In this modality, short pulses of light are absorbed in the tissue, creating ultrasonic waves that are received by ultrasound transducers, thereby converting light into sound (Wang and Peter, 2017).

Biomarkers of atherothrombosis

In recent years, a number of biomarkers have been proposed as significant predictors of atherosclerosis and its thrombotic complications.

C-reactive protein (CRP) is the best characterised of the currently available inflammatory biomarkers and has emerged as a potential marker for cardiovascular risk (Ridker, 2003). Composed of 5 23-kDa subunits, CRP is a circulating pentraxin that plays a major role in the human innate immune response (du Clos, 200). Although generally considered to be an acute-phase reactant, CRP is also produced in smooth muscle cells within human coronary arteries and is expressed preferentially in diseased vessels (Jabs *et al.*, 2003; Calabro *et al.*, 2003). CRP may directly affect expression of adhesion molecules, impact fibrinolysis, and alter endothelial dysfunction (Szmitko *et al.*, 2003). Clinically, CRP can be measured with several standardized, validated, and inexpensive high-sensitivity assays (Ledue and Rifai, 2003). More than 20 prospective, epidemiologic studies demonstrate that hsCRP is an independent predictor of risk of myocardial infarction (MI), stroke, peripheral arterial disease, and sudden cardiac death, even in apparently healthy individuals (Torres and Ridker, 2003).hsCRP is surprisingly specific for the prediction of vascular events, and elevated levels do not predict non cardiovascular mortality or the development of classical inflammatory disorders (Tice *et al.*, 2003).

The American Heart Association and the Centers for Disease Control and Prevention issued clinical guidelines for the use of hsCRP and suggested that evaluation be considered

for those deemed by global risk prediction to be at "intermediate risk."30 Levels of hsCRP of < 1 mg/L, 1 to 3 mg/L, and > 3 mg/L should be interpreted as lower, moderate, and higher vascular risk, respectively. Screening for hsCRP should be performed at the discretion of the physician as a part of global risk evaluation, not as a replacement for LDL and high-density lipoprotein cholesterol testing. The relation- ship between hsCRP and risk appears linear across a full range of values, so individuals with hsCRP levels > 10 mg/L may be at even higher levels of risk than those with levels between 3 and 5 mg/L. Any clinical use of hsCRP, however, is best limited to those at "intermediate risk," that is, individuals with anticipated 10-year event rates between 6% and 20%. Several studies additionally show that hsCRP levels > 3 mg/L also predict recurrent coronary events, thrombotic complications after angioplasty, poor outcome in acute ischemia, and complications after coronary bypass surgery (de Winter *et al.*, 2003). In acute myocardial ischemia, hsCRP levels predict poor outcome when troponin levels are normal, suggesting that an enhanced inflammatory response is a factor in determining subsequent plaque rupture (Lindhal *et al.*, 2000). Clinical data suggest that individuals with elevated hsCRP levels may be more likely to benefit from aggressive interventions (Lindmark *et al.*, 2001).

Plasma fibrinogen is an important acute-phase reactant and multiple epidemiologic studies demonstrate that baseline fibrinogen levels predict future risk of MI and stroke (Kannel *et al.*, 1987). Fibrinogen also plays a major role in hemostasis and traditionally has been classified among novel hemostatic and thrombotic risk factors.

Several other markers show clinical promise. These include alternative acute-phase reactants such as serum amyloid A, the inflammatory cytokines interleukin-6 (IL-6), IL-18, matrix metalloproteinase-9 (MMP-9), tumour necrosis factor alpha, the leukocyte adhesion molecules ICAM-1, VCAM-1, P-selectin, and soluble CD40 ligand, lipoprotein-associated phospholipase A2 and biomarkers of leukocyte activation, including myeloperoxidase (Ridker *et al.*, 2004).

Antithrombotic approaches

Treatment of atherothrombotic patients include the management of cardiovascular risk factors, antiplatelet therapy and anticoagulants.

The aims of the therapy are, firstly, to prevent the occurrence of acute ischemic events through inhibition of platelet thrombus formation and, secondly, to protect distal tissues to prevent venous thromboembolism (VTE).

Antithrombotic therapy has reduced the relative risk of cardiovascular events by up to 25%. Several landmark trials have established the efficacy of aspirin in atherothrombosis.

Remarkably, the ISIS-2 study found that the effect of aspirin in acute myocardial infarction (MI) was comparable to the effect of a fibrinolytic agent (streptokinase). A meta-analysis by the Antithrombotic Trialists Collaboration suggests that the use of aspirin should be expanded to populations such as those with diabetes, peripheral arterial disease, carotid disease, and end-stage renal disease. They also concluded that there is no additional benefit by using chronic aspirin doses higher than 75 mg. The Thrombosis Prevention Trial (TPT) demonstrated a 20% relative reduction in the combined endpoint of coronary death and nonfatal MI with a dose of aspirin of 75 mg.

In the antiplatelet armamentarium, clopidogrel represents a critical advance and several clinical trials have been carried out with this drug. A daily 75 mg dose of clopidogrel was compared with a daily 325 mg dose of aspirin in patients with cardiovascular disease in the CAPRIE (Clopidogrel versus Aspirin in Patients at Risk of Ischemic Events) trial. After an average of 1.9 years follow-up, the data demonstrated a statistically significant 8.7% relative risk reduction in the composite endpoint of MI, ischemic stroke, and vascular death. This is noteworthy when one takes into account that aspirin, which itself has a marked effect compared with placebo, was used as an active control.

Clopidogrel for the Reduction of Events During Observation (CREDO) was a multicenter, double-blind study of patients with stable and unstable angina who were undergoing percutaneous coronary intervention. The trial demonstrated the safety and efficacy of clopidogrel treatment before the procedure, and the beneficial effect of prolonged (1 year) versus short-term (1 month) antiplatelet therapy (Steinhubl *et al.*, 2002).

The combination of aspirin and clopidogrel has a synergistic effect in preventing thrombus formation. The CURE68 (Clopidogrel in Unstable angina to prevent Recurrent Events) trial tested the efficacy of this combination compared with aspirin alone. The results showed a 20% relative risk reduction of the composite endpoint of nonfatal MI, stroke, and cardiovascular death in the combination group. Patients assigned to the dual anti-platelet treatment had higher rates of major bleeding, but no increase in life-threatening bleeding. A subgroup analysis of patients who underwent percutaneous coronary intervention (PCI) during the CURE trial, PCI- CURE (Yusuf *et al.*, 2001), demonstrated that pre-treatment (mean 1⁄4 10 days) with clopidogrel and aspirin before percutaneous coronary intervention, as well as long-term treatment (mean 1⁄4 8 months), was useful in reducing ischaemic events.

Bleeding is the major side effect of antithrombotic therapy. The risk of major bleeding is ≈1.8 fold higher with dual anti-platelet therapy (DAPT) with aspirin plus clopidogrel than with aspirin alone (Bowry *et al.*, 2008). Likewise, the risk of bleeding increases at least 2

fold when aspirin is used in combination with an anticoagulant (Rothberg *et al.*, 2005). Although the direct oral anticoagulants (DOACs), which include dabigatran, apixaban, edoxaban, and rivaroxaban, are associated with less major bleeding than vitamin K antagonists such as warfarin, the risk of bleeding almost doubles when the DOACs are administered in combination with aspirin (Dans *et al.*, 2013; Shah *et al.*, 2016; Alexander *et al.*, 2014; Xu *et al.*, 2016). Therefore, there remains a need for safer anticoagulant therapy.

Introduction to Stroke

Stroke is one of the main causes of death and major reasons for long-term disability worldwide, causing a high economic burden to both society and individual patients (Mozaffarian *et al.*, 2016; Heydari *et al.*, 2019). A stroke results from sudden decrease of blood flow to the brain which causes rapid loss of function. Its symptoms, including hemiparesis, vomiting, drowsiness, and loss of consciousness, often go unrecognized until the acute treatment window has passed. There are two major categories of brain damage in stroke patients called ischemic stroke and haemorrhagic stroke.

Ischaemic stroke is a heterogeneous disorder with more than 100 pathologies implicated in its pathogenesis. It represents the most common type and is responsible for about 85% of all strokes, characterized by a deficiency of blood flow depriving brain tissue of needed power and oxygen. It occurs as a result of thrombotic obstruction within a brain blood vessel, usually the middle cerebral artery (MCA) (Deb *et al.*, 2010; Ramos *et al.*, 2017) (Figure 5).

Ischemic stroke

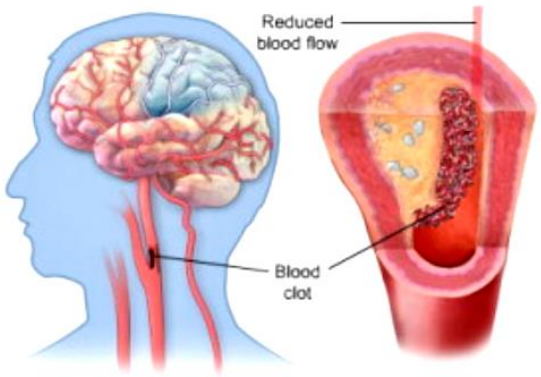


Figure 5:*Ischemic stroke (https://ctrnd.med.ufl.edu/research/stroke/strokebackground/)*

Ischemic stroke may manifest in the form of thrombotic and embolic subtypes. Thrombotic strokes occur as a result of a clot forming in a cervical or intracranial vessel, most commonly as a result of atherosclerosis, and can be further subdivided into lacunar and non-lacunar strokes. Embolic strokes occur when a plaque or clot moves away from a more proximal source and is deposited in an intracranial vessel, occluding the artery and disrupting the blood supply to the brain. The source of most emboli is atherosclerotic plaque in cervical vessels or clot in the heart from atrial fibrillation (Young and Schaefer, 2016).

Regardless of the cause of acute stroke, the primary event that occurs (85–90%) in majority of patients is represented by a compromised vascular supply to the brain (Deb *et al.*, 2010). In particular, the irreversibly affected brain area is called "ischemic core", while, the term "penumbra" describes those brain portions only partially injured with the potential to recover (Deb *et al.*, 2010). Those areas in fact contain metabolically active cells, and preserving these cells is the main purpose of therapeutic agents (Aliaga *et al.*, 2010; Kumar *et al.*, 2010) (Figure 6). During the ischemic insult, the neurovascular Unit

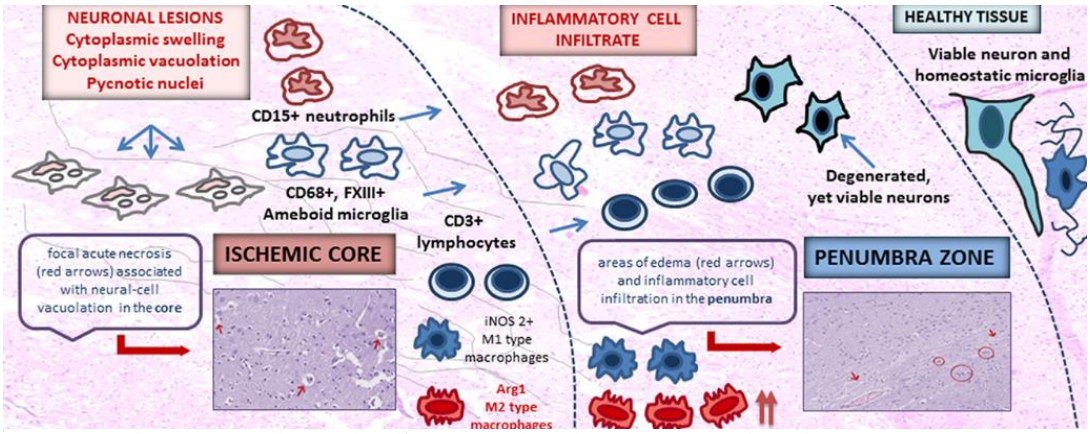


Figure 6: *Ischemic core and penumbra (from Horváth et al, J Neuroimmunol 2018)*

(NVU), a dynamic terminal structure consisting of microvessels (endothelial cells, basal lamina matrix, astrocyte end-feet, pericytes), astrocytes, neurons and their axons, and other supporting cells (microglia and oligodendroglia) participates in the reperfusion battleground between ischemic core and the tissue having the potential to recover (Del Zoppo GJ, 2010) (Figure 7). An important function of the NVU is to form the blood-brain barrier (BBB); this term refers to a complex of cells that separates the brain interstitium from the luminal contents of the cerebral vasculature and is essential for the homeostatic maintenance of ionic balance, nutrient transport, and for preventing the passage of harmful molecules into brain parenchyma (Hawkins

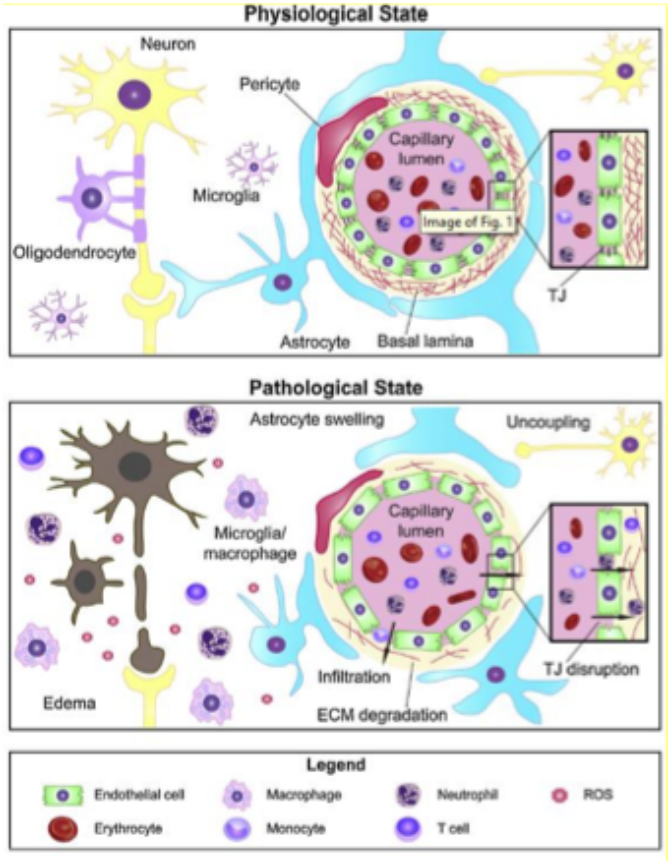


Figure 7: *Physiological and pathological state of NVU (from Leak et al., Prog Neurobiol. 2014)*

and Davis, 2005). It also regulates the bidirectional passage of substances between neurons and their supplying microvessels with the participation of the intervening astrocytes. BBB dissolution starts 2 h after the onset of ischemia and is rapidly followed by an increase in BBB permeability. There are two crucial NVU phases which are important to discriminate: a healthy NVU with an intact BBB and a pathological NVU in which the BBB is severely compromised (Dirnagl, 2012; Maki *et al.*, 2013) (Figure 7).

Oxidative stress is an early stimulus for BBB injury and may trigger release of MMPs [a family of enzymes that proteolytically degrade various components of the extracellular matrix] by neurons, glia, astrocytes, pericytes (Figure 8) resulting in BBB damage through digestion of the endothelial basal lamina (Sumii and Lo, 2002). Oxidative stress occurs when the production of free radicals overcomes the endogenous capacity of cellular antioxidant defenses. In this early phase, the BBB opening could be reversible.

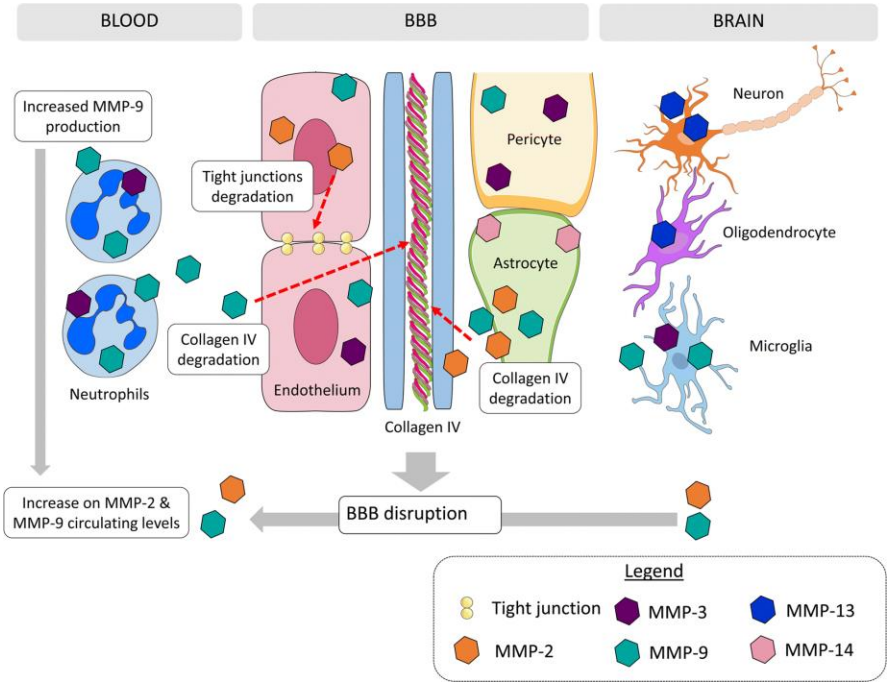


Figure 8 *Role of MMP after ischemic stroke. MMP-2 and MMP-14 (MT1-MMP) are the main constitutive enzymes of the brain, expressed mainly in astrocytes. MMP-3 and MMP-9 are brain-inducible MMPs, mainly produced by resident microglia and infiltrating neutrophils. Furthermore, MMP-13 is also involved in the wound cascade, being produced mainly by neurons and oligodendrocytes. After the ischemic event, there is an increase in MMPs both in the blood and in the brain, being the most studied MMP-9 and MMP-2. These two proteins are thought to be responsible for the degradation of collagen IV, an important component of the basal lamina, which eventually leads to the BEE breaking. (from Montaner et al., Cell Mol Life Sci. 2019)*

After the early BBB opening, there is a second phase of severe BBB injury within 24–72 h after infarction (Yang and Rosenberg, 2011). This phase is more complicated and results in greater tissue damage through leukocyte infiltration and marked release of proteases such as MMPs from neutrophils transmigrated to the ischemic brain. Disruption of the BBB allows leakage of blood components into the brain parenchyma. Extravasation of high molecular weight molecules is followed by water due to osmosis. The disruption of BBB leads to an expansion of brain infarction, vasogenic edema, which may cause

secondary damage through intracranial hypertension and hemorrhagic transformation. Additionally, extravasation of red blood cells leads to hemorrhagic transformation of the infarcted area.

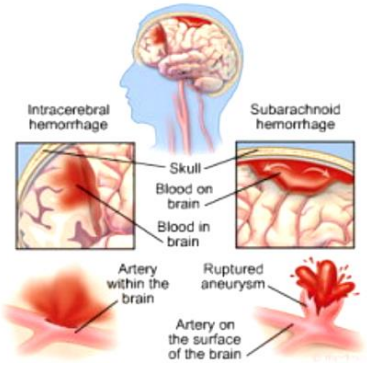


Figure 9: *Hemorrhagic stroke (https://ctrnd.med.ufl.edu/research/stroke/stroke- background/)*

Haemorrhagic stroke

Haemorrhagic stroke, which accounts for the remaining 15% of cases, is characterized by a higher mortality rate (responsible for 30% of all stroke deaths) and it occurs when a blood vessel in the brain weakens and ruptures causing bleeding (Deb *et al.*, 2010; Ramos *et al.*, 2017) (Figure 9). Intracerebral haemorrhage (ICH), as a kind of hemorrhagic stroke, is usually caused by rupture of a blood vessel in the brain that is degenerated due to long-standing hypertension. Its dangerous effects are the result of: 1) hypoxia due to disrupted vascular supply; 2) irritant effect of released blood on brain parenchyma and vasculature; and 3) raised ICP due to continuous bleeding, which may further restrict cerebral blood flow. Therefore, the hemorrhagic stroke is more harmful than the ischemic stroke. There are two types of hemorrhagic stroke: one resulting from intracerebral haemorrhage (developing over 30–90 min) secondary to hypertension, cerebral amyloid angiopathy, or degenerative arterial disease; and the other due to subarachnoid haemorrhages caused by rupture of an aneurysm (Escudero Augusto *et al.*, 2008). Focal neurological symptoms, vomiting and drowsiness are common and headache may be present. Most sub-arachnoid haemorrhages manifest suddenly with intense headache, vomiting and neurological deficit and altered consciousness may occur in about 50% of patients. In the United States, 8–10 million people (3% prevalence) might have an aneurysm, and bleeding occurs in only 30.000 people per year. The main risk factors are advanced age, heavy alcohol consumption and hypertension. Cocaine abuse is an important cause of cerebral haemorrhage in young people (Easton *et al.*, 2001).

Genetic of Stroke

Ischemic Stroke is a complex of heterogeneous disease caused by a combination of environmental and genetic factors and has a heritable component (Guo JM *et al.*, 2010). That ischemic stroke could have originated at least in part genetically has been suspected for many years and a positive family history of cerebrovascular disease is commonly considered a risk factor for stroke.

Linkage analysis studies have identified several genes associated with monogenic stroke, such as the *NOTCH3* gene causing "cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy" (CADASIL) (Dichgans *et al.*, 2019).

Common variants at *HDAC9* (encoding histone deacetylase 9), which conferred a 40% increased risk for large artery stroke per copy of the risk allele are the result of the first stroke GWAS that included 3548 cases and 5972 controls (Bellenguez *et al.*, 2012).

In initial studies which attempted to replicate GWAS, two variants (*PITX2* and *ZFHX3*), which were initially associated with atrial fibrillation, have both been shown to be independent risk factors for ischaemic stroke, especially those thought to be cardioembolic (Gretarsdottir *et al.*, 2008; Gudbjartsson *et al.*, 2009).

A variant at chromosome 9p21 which was originally associated with myocardial infarction and coronary artery disease (Helgadottir *et al.*, 2007), was found to be associated also with large artery stroke (Gschwendtner *et al.*, 2009). Subsequently, studies of larger sample size led to the identification of at least 35 loci with substantial correlations to stroke (Biffi A *et al.*, 2010; Woo *et al.*, 2014; Malik *et al.*, 2018).

The combinations of different polymorphisms predisposing to stroke with modifiable risk factors can have a synergistic effect on the overall risk of stroke, in particular in young individuals. A study suggested that the risk of stroke in patients under the age of 45 increased with the number of polymorphisms simultaneously present in the same patient and this risk was even higher if the patient was also a smoker or hypertensive (Pezzini *et al.*, 2005)

The role of genetic factors in cerebral stroke can be direct or mediated. In the first case the genetic alterations may be linked to stroke onset themselves while, in other cases, they could induce stroke alongside classic or new risk factors.

In 1989, a prospective study of 1805 stroke patients suggested that in addition to the classic risk factors for stroke, that occurs in 50% of cases, other elements as the genetic factors might be involved in ischemic stroke (Sacco *et al.*, 1989). However, the results from various studies on family history (Jousilahti *et al.*, 1997), such as twins (Bak *et al.*, 2002) and candidate genes (Flobmann E *et al.*, 2004) studies, on the involvement of genetic background in stroke are controversial (Razvi et al 2006). Stroke may be the outcome of single gene disorders or more commonly, a polygenic multifactorial disease. Mutations in several candidate genes have been found to be associated with stroke (Francis *et al.*, 2007). However, the monogenic disorders that cause stroke must be distinguished from polygenic and multifactorial forms. Advances in genomic technologies, sequencing costs lowering, biobanking, and data sharing, have collectively accelerated genetic discovery.

Advancement in sequencing technology has facilitated the discovery of single-gene disorders associated with stroke beyond classic syndromes (Dichgans *et al.*, 2019).
Currently, at least 50 stroke-related monogenic conditions are known. They are rare conditions in which the candidate gene confers a high risk of disease to the carrier of the mutation. The genes most likely related to stroke are those underlying the dominant autosomal amyloid angiopathies (*APP*, *CST3* and *BRI* genes) and CADASIL (autosomal dominant cerebral arteriopathy with subcortical infarcts and leukoencephalopathy; *NOTCH 3* gene). In such situations, tools are already available for molecular diagnosis and genetic counselling for clinical practice. Further studies are instead needed to better clarify the pathways that link genotype and phenotype and to develop therapeutic approaches. Potential candidates for genetic screening are patients with stroke in young age with no classic risk factors and an important family history (Tournier-Lasserve, 2002). Icelandic Studies have led to the identification of two "stroke genes" that confer a substantial risk for ischemic stroke. Both genes encode enzymes, “phosphodiesterase 4D (*PDE4D*)” and "arachidonate 5-lipoxygenase-activating protein (*FLAP*)" (Gretarsdottir *et al.*, 2002). Ischemic stroke with a polygenic genetic background is sustained by different variants and each gene confers a small relative risk. A crucial role in increasing the risk of the disease is the synergy or effect that is obtained when multiple genes interact with each other in an additive or multiplicative manner (Deb *et al.*, 2010). In most cases, stroke presents as multifactorial disorder or complex trait, for which it is not possible to demonstrate inheritance pattern due to high prevalence and phenotypic and genetic heterogeneity. Numerous alleles that confer a low risk of disease combine with very complex additive and/or multiplicative manner, leading to increased risk of the disease (Deb P *et al.*, 2010). A study identified significant genetic associations between premature ischaemic stroke and haplotypes in genes involved in methionine metabolism, suggesting their possible contribution to genetic susceptibility for early-onset ischemic stroke (Giusti *et al.*, 2010). These discoveries, along with the expanding availability of other omics data and rare genetic variants, are fundamental for a better understanding of the pathophysiological and genetic bases of the stroke, also aiming to address new therapeutic targets (e.g. proprotein convertase subtilisin/kexin type 9, PCSK9) (Khera and Kathiresan 2017; Malik *et al.*, 2018).

Therapeutic treatment of stroke

Alterations in synaptic function and vasculature have been shown to correlate with behavioral improvement after stroke as the brain remaps to compensate for damaged

networks (Winship and Murphy, 2009). Because of the complexity of the restorative processes that occur after the initial ischemic damage, a single mechanistic pathway will likely not be sufficient to greatly improve functional outcomes. For this reason, strategies such as cell therapies, stimulation, or mild hypothermia that affect several of these pathways, or a combination of therapeutic approaches, may prove to be the most promising for clinical translation. The goal of therapeutic approaches for stroke, aimed to early reperfusion therapy, is to minimize neurologic impairment, long-term disability, and stroke-related mortality. Currently, the systemic and endovascular reperfusion therapies represent the mainstay for achieving rescue of the ischemic penumbra.

Intravenous (IV) or intra-arterial (IA) thrombolysis with the recombinant form of tissue plasminogen activator (rtPA) has been approved by FDA in 1995. Intra-arterial (IA) mechanical thrombectomy for removal of blood clots was approved in 2015 (Lansberg *et al.*, 2012; Moussaddy *et al.*, 2018). Other therapeutic strategies available for the treatment of ischemic stroke are neuroprotection and neurorecovery (Amemori *et al.*, 2013). Therapeutic angiogenesis also seems to represent a promising tool to improve the prognosis of cerebral ischemia (Benedek *et al.*, 2019). After the acute phase, focused physical rehabilitation of the injured area is the primary current therapy that is proven to be effective (Veerbeek *et al.*, 2014). The extent of neurologic recovery is still limited and novel approaches to augment or enhance the body's endogenous regenerative abilities are required (George and Steinberg 2015). Several large clinical trials observed a complete recanalization, in particular when it is achieved with MT combined with rtPA (Berkhemer *et al.*, 2015; Goyal *et al.*, 2015; Saver *et al.*, 2015). However, two significant limitations remain: first, a substantial number of patients is not eligible for recanalization mainly because of substantial extension of irreversible infarction at the time of presentation and/or because of late presentation outside the accepted treatment windows of time; second, extensive infarction and absence of clinical improvement may occur during or after treatment also in cases where the recanalization result was proven to be complete and stable (Berkhemer *et al.*, 2015). IV-tPA is approved for recanalization in patients presenting ischemic stroke within 4.5h of symptoms onset. IV-tPA is given in a total dose of 0.9 mg/kg, with 10% given as a bolus over 1 min, and the rest over 1h. All patients eligible for IV-tPA should receive IV-tPA (Powers *et al.*, 2018). rt-PA a fibrin specific activator for the conversion of plasminogen to plasmin, stimulates thrombolysis and rescues ischemic brain by restoring blood flow. However, emerging data suggest that under some conditions (if reaching extravascular space), could be potentially neurotoxic. This effect is thought to be caused by the increase in cerebrovascular permeability through various factors such as ischemic

reperfusion injury and the activation of MMPs, but the detailed mechanisms are unknown (Suzuki *et al.*, 2016). According to the classical theory, rt-PA increases BBB permeability via degradations of basement membrane mediated by low density lipoprotein receptor associated protein-1 (LRP-1) stimulation, and MMPs induction and activation. Another possible explanation is that rtPA might enhance BBB permeability by activating the vascular endothelial growth factor (VEGF) system that determines both the dissociation of endothelial cell junctions and endothelial endocytosis, and causes a subsequent increase in vessel permeability. The tPA or Alteplase is produced in vitro with the recombinant DNA technique, commercially available as a single and glycosylated chain which was first introduced in 1985. Recombinant tPA (rtPA), a thrombolytic agent, is a serine protease that is found naturally in vascular endothelium, which catalyzes the production of plasmin, an active proteolytic enzyme, by cleaving the arginine-valine bond of plasminogen (Korninger and Collen, 1981). It has been reported to have high thrombus specificity with proteolytic activity increased by 400-fold within a thrombus. This results in accelerated fibrinolysis. This agent reaches the thrombus within seconds and has an active catalytic half-life of 5-10 min. This is due to its neutralization by a binding protein, the tissue plasminogen inhibitor (tPI), and clearance of the complex by the liver. Among different variations of rtPA molecules, alteplase is the only currently approved thrombolytic agent for the treatment of acute stroke. The use of alteplase in the time-sensitive treatment of acute ischemic stroke was initially proven in the National Institute of Neurological Disorders and Stroke (NINDS) trial, published in 1995. That trial, which included 624 patients, showed a 30% increase in good outcome at three months (minimal or no disability) (NINDS trial, 1995). It was however associated with a 6.4% risk of symptomatic intracranial hemorrhage (sICH), defined as a newly detected hemorrhage on follow-up CT associated with a decline in neurological status within 36 h, half of which were fatal (Hacke *et al.*, 2004; Gonzalez, 2006) (Graph 1). Since then, other studies have been conducted, and the European Cooperative Acute Stroke Study III (ECASSIII) was the main study for testing and proving the safety and efficacy of Alteplase in a 4.5 h window from symptom onset, even with modest results (Hacke W *et al.*, 1998). Nevertheless, that benefit was later confirmed in a meta-analysis that pooled results of all thrombolysis studies (OR1.26, 95%CI, 1.05-1.51) (Millan *et al.*, 2017; Al-Ajlan *et al.*, 2016). The International Stroke Trial (IST-III), included 3035 patients. The study, combined with the prior studies, confirmed the efficacy of thrombolysis and observed benefit of the therapy in patients over 80 years old (IST-3 collaborative group, 2012). The first experiments indicated rtPA as a possible therapeutic agent in acute stroke and were performed on rabbit models of

embolic stroke (Zivin et al, 1985). rTPA is a white, insoluble water powder, whose solubility is increased thanks to the addition of Arginine, one of the elements taken into consideration to justify the toxic effect of the thrombolytic (Harston *et al.*, 2010). Alteplase is rapidly eliminated from the blood compartment and is also mainly metabolized in the liver (plasma clearance 550-680 ml / min). The relevant plasma t 1⁄2 α half-life is about 4-5 minutes; after 20 minutes less than 10% of the initial quantity is present in the plasma. Alteplase, when administered intravenously remains relatively inactive in the circulatory system but once bound to fibrin, it becomes active, inducing the conversion of plasminogen to plasmin with dissolution of the fibrin clot. This occurs by splitting a particular peptide bond, Arg560-Val561 of zymogen plasminogen, to form the active plasmin enzyme, whose function is to hydrolyse the Arg-Lys bonds, splitting the fibrin polymer into soluble peptides. In order to better understand the complexity of the Alteplase effect, it is useful to begin to reconsider the pleiotropic role of endogenous tPA. In addition to the lysis effect of the clot, the endogenous tPA is involved in the remodeling of the extracellular matrix, a fundamental element in the development of the central and peripheral nervous system. In the adult brain, it would seem to play a role in dynamic remodeling at dendritic and synaptic level (Kaur *et al.*, 2004).

Intra-arterial thrombolysis and thrombectomy

Intra-arterial delivery of a thrombolytic agent, proximal to an occlusion, has been described since the 1980s. Intra-arterial therapy, performed using a microcatheter, has the advantage of administering the thrombolytic agent directly at the level of the occluding thrombus (Nesbit *et al.*, 1996). Several specialized centers for stroke management use programs to treat AIS patients with endovascular therapy. This approach is generally limited to occlusion of a large cerebral artery accessible to a microcatheter within 3-5 hours. The main advantage at the time over intravenous therapy was the ability to visualize simultaneously the occlusion through angiography. Over the following decade, however, CT angiography has become available. Five major randomized controlled trials assessed the aforementioned mechanism of thrombolytic agents delivery (PROACT-II, MELT, EMS-bridging trial, IMS and SYNTHESIS).

Endovascular options include mechanical thrombectomy, chemical thrombolysis or even a combination of the two mechanisms. Mechanical thrombectomy has recently demonstrated efficacy even 24 h after the onset of a cerebrovascular event (Montaner *et al.*, 2019). The effectiveness and subsequent approval of thrombus retrieval techniques (Moussaddy *et al.*, 2018) has further marginalized the role of intra-arterial thrombolysis

which is currently used in limited circumstances and is uniquely considered in cases of distal non-retrievable occlusions in which intravenous thrombolytics are contra-indicated. Stent retrieval allowed for rapid flow restoration, as soon as the stent was deployed and even before thrombus retrieval, potentially further improving associated clinical outcomes because of faster restoration of blood flow. The Mechanical Embolus Removal in Cerebral Ischemia, designed by University of California, Los Angeles in 2001, better known as MERCI Retriever is the only device approved by the FDA for endovascular treatment of stroke patients (Smith *et al.*, 2005). Two Randomized trials, SWIFT and TREVO-2, observed the superiority of these devices over their predecessors, with higher recanalization rates [86% TREVO, 89% Solitaire, 60-67% MERCI)] and clinical outcomes (90 days functional independence, 40% TREVO, 36% Solitaire, 22-29% MERCI). Additionally, these new devices have better safety profiles, as intracerebral hemorrhages resulted to be reduced by half for Solitaire compared to MERCI, and vessel perforation was 10 times higher in the MERCI group compared to the TREVO group.

Recent randomized clinical trials (RCTs), DAWN and DEFUSE 3 demonstrated improved functional outcomes after endovascular thrombectomy in those treated up to 24 h after selection with perfusion imaging increasing the number of patients eligible for thrombectomy (Wang *et al.*, 2018). Since TM is only applicable in approximately 5% - 10% of patients with ischemic stroke, rtPA therapy is still the treatment of choice for most patients with AIS (Jauch *et al.*, 2013).

Endovascular therapy has emerged as a promising new therapy for stroke patients, although further progress in this field is imperative to deliver further improvements to patient outcomes.

Role of neuroimaging

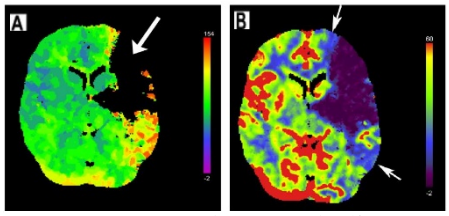


Figure 10*: Computed tomography (from Birenbaum D, West J Emerg Med. 2011)*

Neuroimaging (Figure 10) is an important method for in the diagnosis of stroke. The utilization of various imaging modalities has been shown to be crucial in identifying which patients may benefit from these therapies. An accurate and practical understanding of imaging is essential for the management of the acute stroke patient. Among the neuroimaging techniques the most used are CT and MRI that affect treatment selection and patient outcomes. Imaging is crucial in identifying patients presenting with acute stroke symptoms who may benefit from IV-tPA and intra-arterial therapies (Young and Schaefer, 2016).

With the increasing use of multimodal CT imaging in the setting of AIS, various imaging predictors of HT have been preliminarily studied. BBB disruption is a key- phenomenon of tissue injury after reperfusion, which can be exacerbated by the revascularization treatments of acute phase. CT perfusion (CTP) can provide information about the extension of ischemic core, the impairment of BBB and the status of collateral circulation. Hence it has been proposed as a pre-treatment imaging technique for evaluating the risk of HT in acute stroke patients (Aviv *et al.*, 2009; Hom *et al.*, 2011). Three small studies just focusing on the role of CTP, have reported quite high sensitivity and specificity to predict HT or BE after AIS (Aviv *et al.*, 2009; Hom *et al.*, 2011; Lin *et al.*, 2012). A relationship between MMP9 and BBB disruption assessed by CTP has not been proven yet and scanty clinical data exist on MRI. In 41 patients evaluated for AIS, baseline MMP-9 has been proved to be a significant predictor of BBB disruption, as assessed using gadolinium enhancement of cerebrospinal fluid on fluid-attenuated inversion recovery (FLAIR) MRI at 24 hours from symptoms onset (Barr *et al.*, 2010). In a larger cohort of 180 acute stroke patients, T2 FLAIR hyperintensities, possible expression of vasogenic edema, turned out to be associated with both baseline MMP-9 level and risk of hemorrhage (Jha *et al.*, 2014). Multimodal Magnetic Resonance (MR) imaging techniques such as diffusion weighted imaging (DWI) and perfusion weighted imaging (PWI) or both have been studied as predictors of HT (Singer *et al.*, 2008; Campbell *et al.*, 2010). Similar results have been found with CT perfusion, which has been demonstrated to be a reliable predictor of HT (Souza *et al.*, 2012). Compared to DWI and PWI in MRI, CT perfusion offers a time sensitive and widely practicable assessment of cerebral hemodynamics and parenchymal viability, thus playing a key role in acute stroke management.

Development and application of bioinformatic pipelines for the mutational analysis of data derived from high productivity sequencing technology Illumina and workflow optimization for diagnostic purpose.

Introduction to next-generation sequencing

For the past 30 years, Sanger method (sanger *et al.*, 1977) has been the dominant approach and gold standard for DNA sequencing. The first sequencing project of model organism was the whole-genome sequencing of bacterium Haemophilus influenzae Rd (Fleischmann *et al.*, 1995). The genome of the first eukaryotic organism, Saccharomyces cerevisiae, was sequenced in the next year through the collaboration of 19 countries (Goffeau *et al.*, 1996). At the turn of the century, the flowering plant Arabidopsis thaliana was sequenced as the first plant genome project (Arabidopsis Genome Initiative, 2000). Most notably, the Human Genome Project (HGP) started in 1990 and completed in 2003 through the collaboration of 20 institutions and genome centres in 6 countries (International Human Genome Sequencing Consortium, 2004). One of the consequences of these early genome projects is the development of the genomic biotechnologies, which enables the production of high-throughput data at increasingly lower cost.

The commercial launch of the first massively parallel pyrosequencing platform in 2005 marked the beginning of the new era of high-throughput genomic analysis now referred to as next-generation sequencing (NGS). With the advent of NGS, the time and resources needed to sequence an entire human genome have fallen from years to days, and from several billion to a few thousand dollars (Hegele *et al.*, 2015).

The next-generation sequencing (NGS) based approaches has enabled genome projects like the the 1000 Genomes Project (1000 Genomes Project Consortium *et al.*, 2010). More recently, these technologies have been used in larger population-based genomic projects such as the 100,000 Genomes Project (100,000 Genomes Project Pilot Investigators, 2021) in the United Kingdom and the GenomeAsia 100K (GenomeAsia100K Consortium, 2019). The goal of these later two genome projects is to sequence 100,000 individuals in the United Kingdom and Asia, respectively, in order to understand the health and population structure and relation in these populations. NGS platforms share a common technological feature massively parallel sequencing of clonally amplified or single DNA molecules that are spatially separated in a flow cell. In NGS, sequencing is

performed by repeated cycles of polymerase-mediated nucleotide extensions or, in one format, by iterative cycles of oligonucleotide ligation. As a massively parallel process, NGS generates hundreds of megabases to gigabases of nucleotide-sequence output in a single instrument run, depending on the platform.

Next generation sequencing (NGS) has had a substantial impact on basic genomics research in terms of large-scale analysis and feasibility allowing us to start defining the characteristics of entire genomes and delineate differences between them, gaining a deeper understanding of the full spectrum of genetic variants, defining its role in phenotypic variability and in the pathogenesis of complex traits.

Besides the improved throughput and decreased cost, Wole genome sequencing (WGS) has the advantage of unbiasedness in the survey of genetic variants across genome compared to other high-throughput methods such as genotyping microarrays.

The applications of NGS are not limited to WGS. Exome sequencing and targeted sequencing are also popular approaches for investigating focused genomic regions. By focusing on certain genomic regions, this approach allows higher sequencing depth and larger sample size, which is valuable in characterize rare genetic variants.

These innovative techniques allow to carry out studies of various kinds in a single experiment, among which the simultaneous characterization of genomes, the identification of balanced and unbalanced chromosomal rearrangements, deletions and copy number variations (CNV). Specifically, NGS techniques allow sequencing of: GENOMIC DNA (whole genome, exome: only the part of DNA transcribed in RNA, exons, targeted genes, amplicons: only PCR products) TRANSCRIPTOME (total RNA, mRNA, small RNA: <30 nt) EPIGENOME (ChIP-Seq: chromatin immunoprecipitation sequencing, DNA or RNA to which specific proteins are bound, methyl-Seq: study of DNA methylation pattern, epigenetics).

Challenges posed by the Big Data in genomics

Big Data is used to describe the high-throughput data generated by the NGS and related technologies. Big Data creates unique challenges and opportunities characterized by the 5Vs, i.e., Volume, Velocity, Variety, Veracity, and Value.

Data integration

With current technologies in genomics, data at different layers are available, such as primary DNA sequencing data, DNA methylation data, gene expression data and environmental factors. The goal is of data integration to relate these different types of data to the responses such as disease status, disease progression, and response to treatment.

Primary DNA sequences contain all the genetic information—blueprint for life. In order to realize the information encoded in the DNA sequences, they need to be expressed into RNAs and proteins, which has biological functions. All the cells in a human contain the same DNA sequences, yet different cell types have different gene expressions, hence the different morphology and functions of different cell types. In the human population, genetic variations (mutations) lead to variations in gene expression, which then potentially lead to diseases.

DNA methylation is a biological process by which methyl groups are added to the DNA molecule. Methylation can change the activity of a DNA segment without changing the sequence. Levels of DNA methylation are determined partly by genetic variation and partly by environmental factors such as smoking and diet. The main effect of DNA methylation was thought to be on gene expression. The hyper-methylation in gene promoter regions has been shown to be inversely related to gene expression. DNA methylation is the result of some proteins (enzymes) in the methylation pathway, whose gene expression levels will affect DNA methylation levels.

Gene expression is affected by DNA sequences, DNA methylation, and environmental factors. Variation in gene expression will eventually lead to different responses (phenotypes), including diseases.

With all these inter-connected parts, it is highly likely that variations at DNA level is reflected as variations at DNA methylation and gene expression levels and the signal may be amplified at successive levels. Therefore, integrated analysis of the various types of genomic data has the potential of maximizing statistical power by combining information across data types (Ritchie *et al.*, 2015). Current methods for integrated analysis of genomic data could be roughly put into two categories, multi-stage analysis and simultaneous analysis. In multi-stage analysis, the analysis consists of multiple steps. For example, step 1 of the analysis could be the association of sequence variation with the phenotype; genetic variants passed through the first step are then used to filter gene expressions. The expression values of the corresponding genes are tested for association with the phenotype in step 2. The specific statistical methods used for the association will depend on the outcome variable. For continuous outcome variable, linear regression is a commonly used approach. For categorical outcome variable, common methods are based on generalized linear models such as logistic regression. The advantage of the regression based methods is that potential cofounding effects can be adjusted by including covariates in the models. Common covariates in biomedical research include age, sex, race, disease stage, and medications. The multiple stage analysis approach has been successfully applied in recent

studies to investigate the genetic basis of drug induced toxicity (Huang *et al.*, 2007, 2008). The disadvantage comes from the arbitrary in the selection criteria usually with P-value cutoffs at each stage. The over-stringency at early stages could lead to missed true signals and therefore overall low statistical power. The optimal strategy for setting the selection criteria has yet to be established. In simultaneous analysis, genomic data of different types are combined in one meta-data set for analysis. The advantage of this approach is potentially multivariate methods could be applied and there is no loss of information since all the data are combined. The disadvantage is that the corresponding model will be more complex with the different data types.

The most straightforward approach for simultaneous analysis is to concatenate various types of genomic data into one big matrix by sample ID. Appropriate statistical methods considering the heterogeneity of the data types can then be applied to the combined data for the analysis. One example of such an approach is a Bayesian integrative model to study the joint effect of genetic variants (SNPs) and gene expressions on a continuous gemcitabine-treatment responses in cancer cell lines (Fridley *et al.*, 2012). The model first specifies the direct effect of SNPs and gene expressions on the response variable with a linear model that includes both SNPs and gene expressions as predictors. Next, the model specifies the effect of SNPs on genes expressions using a linear framework, assuming the gene expressions follows a Normal distribution. Lastly, this approach performs Bayesian variable selection using stochastic search variable selection (SSVS) (George and McCulloch, 1993; Mukhopadhyay *et al.*, 2010) through model averaging and shrinkage of SNP effect toward zero. The prior distribution of the SNP effect is a mixture of two Normal distributions, both centered at 0 but with different variances, to represent the cases of inclusion or exclusion of the SNP in the final model. Another example is the method proposed by Mankoo *et al.* in 2001 to perform an integrative analysis of DNA copy number variation, DNA methylation, miRNA and gene expression on time to event (survival time) in ovarian cancer. This method first performs variable selection using least absolute shrinkage and selection operator (LASSO) from the full model with all the different types of independent variables. The selected variables are then used in the Cox regression model to predict the survival time. Because this type of simultaneous analysis combines all the variables, this will increase the number of independent variables substantially and some types of data reduction methods such as the variable selection method in the two examples would be necessary for further statistical analysis.

One of the difficulties of the concatenation-based method is that different types of genomic data often have very large difference in scales, which can create biases in statistical

inference when combined directly. To overcome this problem, several methods have been proposed to transform the data to proper scale before combining them. One example is the graph-based integration approach (Kim *et al.*, 2012) to predict cancer outcomes in brain and ovarian tumours using copy number variation, DNA methylation, miRNA and gene expression data. In this approach an individual graph is generated for each types of genomic data through Graph-based semi-supervised learning (Zhou *et al.*, 2004), in which a node represents a sample and an edge connecting two nodes represents the relationship of the two samples, determined by a Gaussian function of the Euclidean distance between the two samples. The multiple graphs generated from each type of genomic data are then combined through linear combination to generate the final graph for the prediction of cancer outcomes.

In some cases where different types of genomic data are generated from different set of subjects, it is possible to perform the analysis of each data type separately to generate one prediction/classification model for each data type, then perform the integration of the models. An example is the study of driver mutations of melanoma using chromosome copy number and gene expression data (Akavia *et al.*, 2010). In this study, a Bayesian network is constructed using each data type. The resulting Bayesian networks are then combined with a Bayesian scoring function maximizing the overall joint probability of the data and the model structure.

High dimensionality

Big data in genomics is characterized by its high dimensionality, which refers both to the sample size and number of variables and their structures. The pure volume of the data brings challenges in data storage and computation. The data volume can be on the order of terabytes for just the raw data of each sample. For the different types of genomic data, it is a good practice to keep the raw data, often in the image file format so that more sophisticated base calling algorithm can be applied later when available for improved accuracy. Data can be stored locally with hard drive arrays and backed up in other more permanent storage media. It is also a good practice to deposit the data into public databases for easy sharing in the scientific communities, such as the Gene Expression Omnibus at the National Center for Biotechnology Information (NCBI) for functional genomics data (Barrett *et al.*, 2013). Cloud storage is another option where the data can be stored and maintained at a central location accessible by all the research communities.

Big Data is characterized by its large number of variables. The traditional algorithms could become instable with the large number of variables in the big data of genomics.

The large number of variables also contributes to false positive findings due to multiplicity of statistical testing. Data heterogeneity is also a challenge for big data with the increasing popular international collaborations in order to achieve a large sample size, where data were collected from diverse laboratories and time points. While data heterogeneity is a challenge for big data analysis, it also provides unique opportunities for understanding the unique and common features of each subgroup due to its large sample sizes. For example, one of the popular approach for inferring population structure using genetic data is the Bayesian clustering method STRUCTURE (Pritchard *et al.*, 2000), which can assign proportional ancestry to several populations for admixed individuals. STRUCTURE uses Markov Chain Monte Carlo (MCMC) algorithm. It begins by a random assignment of individuals to a K pre-determined populations. Each population has a dis- tinct genetic allele frequencies from other populations. Genetic allele frequencies are then estimated in each population from the individual genotypes and individuals are reassigned based on the updated frequency estimates. This iterative process is repeated many times until convergence, typically comprising 100,000 iterations. Upon convergence, we can obtain the final allele frequency estimates in each population and the assign each individual to a particular population according to the posterior allele frequency estimates.

This method has tremendous impact on the research in human genetics, evolutionary genetics and ecology. However, this method is limited with the number of genetic markers and sample size due to computational cost with the MCMC algorithm. This is apparently a limitation for Big Data in genomics. There are several recent works focusing on overcoming this limitations with likelihood-based methods (Alexander *et al.*, 2009; Tang *et al.*, 2005) and assumptions on variations (Raj *et al.*, 2014).

Computing infrastructure

The large volumes of Big Data in genomics make the computation infeasible with traditional computing facility. It could take months to finish alignment and annotation of NGS reads for studies with large samples using desktop computers. One solution to this problem is to use the high-performance computing facilities such as computer clusters. The idea is to split the big computing job into small jobs and distribute them to each computing node in the cluster. The result is the highly parallel computing so that the big job can be finished fast (Almasi and Gottlieb, 1989). Most of the computing clusters are of the Beowulf type, where generally identical computers are connected to the header computer in a local-area network. Besides its improved computing power, Beowulf clusters are highly

scalable and relatively easy to maintain, which make them especially appealing to Big Data computing needs.

Cloud computing is another potential solution to the challenge in computing facilities. In cloud computing, major computing companies provide services to end users with computing platform, storage, software and CPU times. Current cloud computing services include AWS (Amazon Web Service), Microsoft Azure, Google Cloud Platform, VMware Cloud service, and IBM Cloud. The salient feature of cloud computing is its elasticity and scalability. Users can buy the right service according to the size of the project. The service is available anytime and anywhere with internet collection. It is also maintenance-free and users can assume the platforms are well-maintained with the most recent software packages. Having the data stored at a central database hosted by cloud service providers removes the need of data transfers in separate local databases and thus could save on the time of data transfer, which is usually a bottleneck for Big Data computing.

Dimension reduction

Big Data poses challenges in computing and analysis due to its high dimensionality. One solution to this challenge is to use the statistical techniques for dimension reduction. In WGS, the data could be represented by a n x d matrix, where n is the number of subjects and d is the number of genetic variants. Each entry in the matrix is the respective genotype or genetic score for the subject at the genetic variant. Because of the large number of genetic variants, it is generally infeasible to use the data matrix directly in the standard statistical analysis. The idea of dimension reduction is to reduce the data dimension through linear or non-linear transformation while keeping as much information in the original data matrix as possible.

One common dimension reduction method is principal component analysis (PCA). This is a linear transformation method where it first calculates the eigenvectors of the sample covariance (correlation) matrix. The principal components (first k eigenvectors with the largest eigenvalues) are used to construct a k dimension subspace spanned by the principal components. The original data matrix is then projected to this subspace to obtain a data matrix with n x k dimension. The reduced data matrix retains a large fraction of the variance in the original data matrix. This approach has been shown to be effective to adjust for population stratification in genome-wide association studies (Price *et al.*, 2006). With large sample size for genomic studies, direct application of PCA may not be feasible. New methods are needed for efficient dimension reduction with Big Data. One potential method is random projection based on the Johnson-Lindenstrauss lemma, which

projects the original data matrix to a subspace that preserves the distance between data points (Johnson and Lindenstrauss, 1984). This method is powerful and computational simple in that its complexity increases linearly with sample size.

Data security

Genomic data is special in that given enough genetic marker information, it can uniquely identify an individual, much like the fingerprints. Indeed, genetic information has long been used in forensics for individual identification (Jeffreys *et al.*, 1985a,b). Genomic data contains critical information for life. With the availability of Big Data in genomics, it is possible to make predictions of many individual characteristics including major disease risks from the data. Therefore, data security could be a big concern for Big Data in genomics.

Genomic data should be considered as protected health information and be handled according to the regulations of HIPAA (Health Insurance Portability and Accountability Act) in the United States. Common security practices should be implemented such as password protection, data/disk encryption, secure storage, secure transmission, and regular checking of data integrity with checksum analysis. Cloud computing offers convenient access of data and computing services at the same host with continuous support, and hence very popular for Big Data analysis. However, data security is a concern for cloud computing because the data are hosted externally of the investigator's institution. The cloud computing service providers need to address these concerns by providing corresponding security measures such as controlled access and secure data transfer.

Data analysis workflow for Clinical Next Generation Sequencing

A new field of NGS technology development is linked to diagnostic applications and to the identification of phenotypical severity biomarkers or pharmacogenomics for personalized therapies (Wesolowska A *et al.*, 2011). Implicit in clinical adoption of this technology is the need for bioinformatics to process and aid in the interpretation of the massive amount of data generated by the sequencing instruments (Gullapalli *et al.*, 2012). Bioinformatics is a recently defined discipline that develops and applies advanced computational tools to analyse and interpret high dimensional biological data.

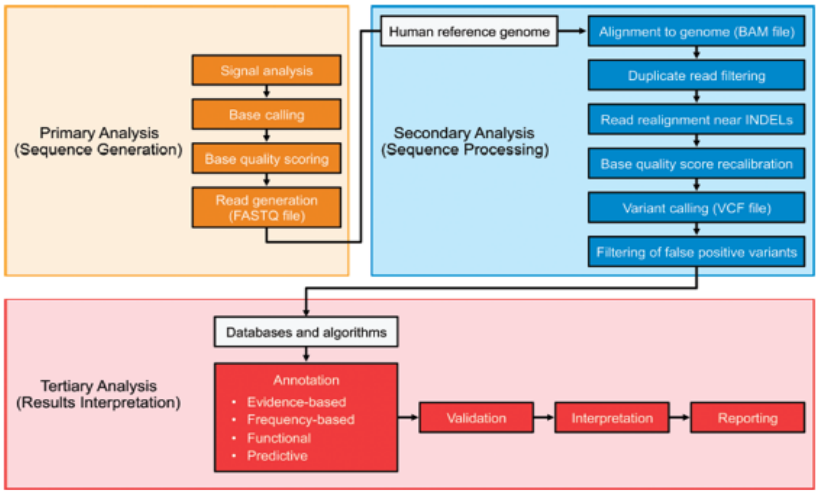


Figure 11: *Flow diagram illustrating the major components of a clinical NGS analytical pipeline (Oliver et al., 2015)*

NGS-based bioinformatics analytics are designed to convert signals to data, data to interpretable information, and information into actionable knowledge. This process essentially consists of three general steps (Figure 11). In brief, primary analysis consists of processing raw sequencing instrument signals into nucleotide base and short-read data. Secondary analysis involves the alignment to a reference sequence or de novo assembly of the NGS nucleotide reads and subsequent variant detection, and tertiary bioinformatics analyses provide context to the information generated during an NGS experiment by associating the sample-specific genomic profile with disparate descriptive annotations.

Sequence generation

Primary analysis software is provided by all major sequencing vendor companies and usually is installed on the hardware systems supporting the sequencing instruments.

To date, there has been limited development of independent primary analysis software programs.

Primary analysis consists in converting the raw signals generated by the sequencing instruments into nucleotide bases with associated quality scores. Short nucleotide sequences or “reads” varying from dozens to hundreds of base pairs for each fragment are combined together, often in a form of a FASTQ file. In some instances, the primary analysis also includes demultiplexing of multiple samples indexed and pooled into a single sequencing run.

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

Figure12: *Example of a string from a FASTQ file.*

FASTQ format (figure 12) is a text-based format for storing both a biological sequence (usually nucleotide sequence) and its corresponding quality scores. Both the sequence letter and quality score are each encoded with a single ASCII character for brevity. A FASTQ file normally uses four lines per sequence. Line 1 begins with a '@' character and

is followed by a sequence identifier and an optional description (like a FASTA title line). Line 2 is the raw sequence letters. Line 3 begins with a '+' character and is optionally followed by the same sequence identifier (and any description) again. Line 4 encodes the quality values for the sequence in Line 2 and must contain the same number of symbols as letters in the sequence.

Alignment and variant detection

The first step of the sequence processing analysis step is the mapping of short nucleotide reads to a reference genome. De novo assembly of a genome is also possible but currently less common in human applications owing to the existence of a high-quality reference genome and the more experimental nature of genome assembly methods (Ulahannan *et al.*, 2013). Each of the millions of short reads must be compared to the 3 billion possible positions within the human genome in order to be aligned correctly to their appropriate location. This step is thus computationally intense and time consuming (Day-Williams A and Zeggini, 2011). There are various software programs, some commercially available and others freely available to the scientific community, that can be used to perform sequencing read alignment. Various programs differ in speed and accuracy. Most alignment algorithms use an indexing method in order to more rapidly narrow down potential alignment locations within the reference genome with ungapped alignment, although other algorithms allow for gapped alignment.

The Bowtie algorithm is both ultrafast and memory efficient due to it (Langmead *et al.*, 2009) use of an index built with the Burrows-Wheeler transformation (Ferragina and Manzini, 2000). It claims a small memory footprint – about 1.3 gigabytes for the entire human genome. Bowtie makes some compromises to provide its speed and memory usage. It does not guarantee the highest quality read mapping if no exact match exists. Additionally, it may fail to align some reads with valid mappings when configured for maximum speed. Bowtie2 allows for analysis of gapped reads, which may result either from true insertions or deletions, or from sequencing errors. The newer adaptions utilize full-text minute indices and hardware-accelerated dynamic programming algorithms to optimize both speed and accuracy (Langmead and Salzberg, 2012).

BWA (Burrows-Wheeler Alignment) can be considered as “MAQ version 2”. Whereas MAQ (Mapping and Assembly with Qualities) uses a hash-based index to search the genome, BWA uses an index built with the Burrows-Wheeler transformation that allows for much faster searching than its predecessor. Like its predecessor, BWA reports a

meaningful quality score for the mapping that can be used to discard mappings that are not well supported due to e.g. a high number of mismatches (Li H and Durbin, 2009).
SOAP was developed for use in gapped and ungapped alignment of short reads using a seed strategy for either single-read or pair-end reads and can also be applied to small RNA and mRNA tag sequences (Li R *et al.*, 2008). SOAP version 2 (Li R *et al.*, 2009) reduced memory usage and increased speed using an index based on the Burrows- Wheeler transformation (BWT). SOAP3 is a GPU (graphics processing unit) version of the compressed full-text index-based SOAP2, which allows for a speed improvement (Liu *et al.*, 2012).
The mr- and mrsFAST tools are notable in that they report all mappings of a read to a genome rather than a single "best" mapping. The ability to report all possible reference genome locations is useful in the detection of copy number variants (Bailey *et al.*, 2002). Indeed, these algorithms are developed primarily for applications that involve detection of structural variants. mr- and mrsFAST use a seed-and-extend method for alignment and create hash table indices for the reference genome.
Novoalign is a proprietary product of Novocraft (Novocraft, 2010) that uses a hashing strategy like that of MAQ (Li *et al.*, 2008). It has become quite popular in recent publications due to its accuracy claims, and it allows up to 8 mismatches per read for single end mapping.

Once reads have been aligned to the genome, several refinement steps are often performed (DePristo *et al.*, 2011). These steps routinely include flagging or filtering of duplicate reads likely to be PCR artifacts, and realignment, which leverages a collective view of reads around putative insertion/deletion (indel) sites to minimize erroneous alignment of read ends. The resulting sequence alignment is stored in a SAM (sequence alignment/map) or BAM (binary alignment/map) file (Li H *et al.*, 2009b).
Each tool is able to report a meaningful quality score for the mapping (Mapping Quality, MQ) that can be used to discard mappings that are not well supported due to e.g. a high number of mismatches. In a probabilistic view, each read alignment is an estimate of the true alignment
and is therefore also a random variable. It can be wrong. If the mapping quality of a read alignment is Q, the probability Pe that the alignment is wrong can be calculated with:

$$P_e = -10 \times log_{10}(\frac{Q}{100})$$

Given 1000, for example, read alignments with mapping quality being 30, one of them will be wrong in average.

Variant Calling

After alignment of the short reads to the reference genome, the next step in the bioinformatics process is variant calling. The computational challenges in SNP (variant) calling are due to the issues in identifying "true" variants versus alignment and/or sequencing errors. Quality scores allocated by the sequencing software will often be recalibrated on the basis of alignment data, before proceeding to the variant calling stage. Variant calling involves the comparison of the sequenced reads to their point of alignment on the human genome to determine areas that differ on the basis of statistical modelling techniques that aim to distinguish genuine genomic variations from errors (Nielsen *et al.*, 2011). These variants may be responsible for disease, or they may simply be genomic noise without any functional effect. Variant call format (VCF) is the standardized generic format for storing sequence variation including SNPs, indels, larger structural variants and annotations (Danecek *et al.*, 2011). In general, specialized programs are selected dependent on the class of variant being investigated.

The Genome Analysis ToolKit (GATK), developed by the Broad Institute, is one of the most popular methods for variant calling using aligned reads. It is designed in a modular way and is based on the MapReduce functional programming approach (McKenna *et al.*, 2010). The package has been used for projects such as the Cancer Genome Atlas (Cancer Genome Atlas Network, 2012) and the 1000 Genomes Project (1000 Genomes Project Consortium, 2010) that have covered analyses of HLA typing, multiple sequence realignment, quality score recalibration, multiple sample SNP genotyping and indel discovery and genotyping (McKenna *et al.*, 2010).

Developed by the Beijing Genome Institute, SOAPsnp is an open source algorithm (http://soap.genomics.org.cn/) that requires access to a high-quality variant database using SOAP alignment results as an input (Li R *et al.*, 2008). It can be used for consensus calling and SNP detection for the Illumina Genome Analyzer platform and utilizes the phred-like quality score to calculate the likelihood of each genotype based on the alignment results and sequencing quality scores. Building upon the speed of the alignment algorithm Bowtie and using SOAPsnp for SNP calling, an open source cloud-computing tool called Crossbow (Langmead *et al.*, 2009) was developed to perform both alignment and SNP calling.

VarScan (http:// genome.wustl.edu/tools/cancer-genomics/) is an open source tool developed by the Genome Institute at Washington University in St. Louis for short read variant detection of SNPs and indels that is compatible with multiple sequencing platforms and aligner algorithms such as Bowtie and Novoalign (Koboldt *et al.*, 2009). It can detect

variants at 1% frequency, which can be useful for pooled samples; VarScan permits analysis of individual samples as well. VarScan2 (Koboldt *et al.*, 2012) includes some improvements upon VarScan, such as the ability to analysis tumor-normal sample pairs for somatic mutations, LOH (loss of heterozygosity) and CNAs (copy number alterations). This program reads tumor and normal sample Samtools pileup or mpileup output simultaneously for pairwise comparisons of base calling and normalized sequence depth at each position.

Filtering

Variant calling errors are common, as NGS technologies are inherently less accurate than traditional sequencing methods and, therefore, artifacts occur with greater regularity (Voelkerding *et al.*, 2009). This problem is partially corrected for by increasing sequencing depth (i.e., sequencing each base position multiple times). The use of high-depth sequencing is particularly powerful in panel-based approaches in which the query region is small and great depths can be attained. In comparison, exome and genome sequencing efforts are complicated by the increased target region size and issues such as variable capture or sequencing efficiency, which collectively introduce regions of insufficient sequence depth and increase validation burden (Yu *et al.*, 2012). Repetitive genomic regions and pseudogenes introduce alignment ambiguities due to the relatively short read lengths generated by most NGS technologies, and this represents another source of error (Treangen *et al.*, 2012). Erroneous variant calls inevitably occur, and thus filtering or confidence-based prioritization of variant calls is a key component of the secondary analysis workflow. Prioritization is often preferred to removal of candidate variants to avoid the incorrect and irreversible filtering of a genuine variant call. The filtering or prioritization process can involve computational or human efforts, including visual inspection of variant alignments, and can be based on empirical cutoffs or more advanced statistical approaches. Criteria used to assess the quality of variant calls varies but examples include the frequency with which a variant allele is observed in a sample, the base quality of the variant alleles as predicted by the sequencing instrument, and the ability of a read containing a variant allele to map uniquely to a single location on the human reference genome.

Annotation

Following detection, variants must be annotated to determine their biological significance and enable functional prioritization and downstream interpretation. This characterization is generally achieved using a combination of biological annotation sources including

frequency, structural, prediction, or evidence-based data. Comparison of an individual's genome to the current human reference sequence will produce many variant calls that essentially represent benign interindividual human variation. Population frequency-based annotations are often a core component of the analysis because variants that are common in the general population are unlikely to have biological relevance in the context of a clinical assay. Rare nonsynonymous SNPs are SNPs that cause amino acid substitution (AAS) in the coding region, which potentially affect the function of the protein coded and could contribute to disease. Unlike nonsense and frameshift mutations, which often result in a loss of protein function, pinpointing disease-causal variants among numerous SNVs has become one of the major challenges due to the lack of genetic information. For instance, ∼1,300 loci are shown to be associated with ∼200 diseases by GWASs but only a few of these loci have been identified as disease-causing variants (Lander, 2011). Exome sequencing enables the identification of more novel genetic variants than previously possible, but it still requires computational and experimental approaches to predict whether a variant is deleterious. To this end, several approaches have been developed to identify rare nonsynonymous SNPs that cause amino acid substitution (AAS) in the coding region. The major principle of the protein-sequence-based methods to predict deleteriousness in the coding sequence is based on comparative genomics and functional genomics. Comparative sequencing analysis assumes that amino acid residues that are critical for protein function should be conserved among species and homologous proteins; therefore, mutations in highly conserved sites are more likely to result in more deleterious effect. Other modalities to predict disease-causing variants include protein biochemistry, such as amino acid charge, the presence of a binding site, and structure information of protein. SNVs that are predicted to alter protein feature (such as polarity and hydropathy) and structure (binding ability and alteration of secondary/tertiary structure) have a higher probability of being deleterious. Although the majority of research has focused on protein altering variants, noncoding variants constitute a large portion of human genetic variation. Results obtained from GWAS indicate that ∼88% of trait-associated weak effect variants are found in noncoding regions, demonstrating the importance of functional annotation of both coding and noncoding variants (Hindorff *et al.*, 2009). Further selection of causal variants can be based on existing annotation or predicted functional effect.

Many programs exist to examine relevant variants by referencing previously known information about their biological functions and inferring potential effects based on their genomic context.

The Ensembl Variant Effect Predictor (VEP) software provides tools and methods for a systematic approach to annotate and prioritize variants in both large-scale sequencing projects and smaller analysis studies. By automating annotation in a standard manner and reducing the time required for manual review, it helps manage many of the common challenges associated with analysis of SNVs, short insertions–deletions, copy number variants, and structural variants. The VEP annotates variants using a wide range of reference data, including transcripts, regulatory regions, frequencies from previously observed variants, citations, clinical significance information, and predictions of biophysical consequences of variants.

The quality, quantity, and stability of variant annotation obtained depends on the choice of transcript set used. As such, the VEP allows flexibility of transcript choice. To effectively manage large numbers of variant annotations and transcript isoforms, the VEP provides several methods to prioritize results and reduce the number of variants needing manual review. A selection of these filters is available and VEP also supports building of custom filters. Uniquely, the VEP algorithm can be expanded to perform additional calculations via plugins and can analyse custom, potentially private, data.

snpEFF is an open source, Java-based program that rapidly categorizes SNP, indel, and MNP variants in genomic sequences as having either high, medium, low or modifier functional effects (Cingolani *et al.*, 2012). Variant annotation is based on genomic location (intron, exon, untranslated region, upstream, downstream, splice site, intergenic region) and predicted coding effect (synonymous/nonsynonymous amino acid replacement, gain/loss of start/stop site, frameshift mutations). The program may find several different functions for a single variant due to competing predictions based on alternative transcripts. snpEFF uses a VCF input and output style. Currently snpEFF does not support structural variants but there are plans to incorporate such support soon. snpEFF is compatible with GATK and Galaxy, which are popular variant-calling toolkits. The program currently supports 260 genome versions and can be used with custom genomes and annotations.

The ANNOVAR software tool (http:// www.openbioinformatics.org/annovar/) utilizes up to date information to rapidly functionally annotate genetic variants called from sequencing data (wang *et al.*, 2010). ANNOVAR works on a number of diverse genomes including hg18, hg19, mouse, worm, fiy, and yeast. The annotation system allows the user fiexibility in the set of genomic regions that are queried. Annotations can be gene-based (users can select the gene definition systemfi RefSeq, UCSC, ENSEMBL, GENCODE, etc.), region-based (transcription factor binding sites, DNAse I hypersensitivity sites,

ENCODEmethylation sites, segmental duplication sites, DGV sites, etc.), filter-based (e.g., using only variants reported in dbSNP, or only variants with MAF > 1%), or based on any of many other user-driven functionalities.

Materials and methods

Genomic Isolation from Blood Samples

The genomic isolation has been made from blood samples, collected through venepuncture, into a vacutainer tube containing an anticoagulant (EDTA) to stop it from clotting, by using the GeneCatcher technology. GeneCatcher™ technology has been designed and optimised for scalable and automated genomic DNA isolation from blood samples using the Tecan® Freedom EVO® liquid handling platform (Tecan Group Ltd, Switzerland) (Figure 13). It requires no centrifugation or filtration steps and uses reagents that will not clog lines or cause vapor pressure build-up. Furthermore, a magnetic rack (separator) allows processing of several samples simultaneously during the magnetic bead-separation steps. In order to obtain the best results with the GeneCatcher™ gDNA 0.3-1 mL Blood Kit, we used the 24-well Magnetic Separator. These magnetic separators are compatible with the volumes used in the kit protocols and provide effective magnetic strength from neodymium magnets, which are aligned with plate wells or tubes, respectively.

Figure 13: *Tecan Freedom EVO platform*

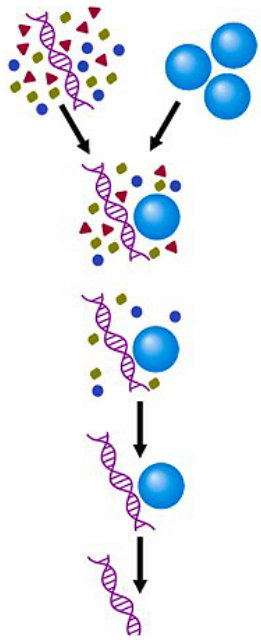


Figure 14: Gene CatcherTM technology workflow

The GeneCatcher™ gDNA Blood Kits allow rapid and efficient extraction of genomic DNA (gDNA) from human blood (Figure 14), including archived or poorly stored blood samples. Genomic DNA is extracted from blood samples using the cost-effective, user-friendly magnetic bead-based Technology. The novel three-step clean-up process of GeneCatcher™ technology is based on DNA capture, DNA purification and DNA elution. In the first step cells are lysed and crude DNA is captured on magnetic beads, leaving most of the cell debris and protein behind in solution. In the second step proteins are digested and then washed away to leave pure intact DNA. In the final step the pure DNA is eluted into an appropriate volume, in order to be used in downstream protocols.

Quantitation and quality assessment of the DNA

NanoDrop One

For quantitative and qualitative evaluation of DNA samples, we used NanoDrop One Spectrophotometer (Figure 15). Thermo Scientific NanoDropTM Spectrophotometers include the absorbance of all molecules in the sample that absorb at the wave-length of interest permitting us to measure the concentration of the genomic DNA. Since nucleotides, RNA, ssDNA, and dsDNA all absorb at 260 nm, they will contribute to the total absorbance of the sample. Therefore, to ensure accurate results, nucleic acid samples will require purification prior to measurement. The NanoDrop One will accurately measure DNA samples up to 3700 ng/ul without dilution. To do this, the instrument automatically detects the high concentration and utilizes the 0.2 mm pathlength to calculate the absorbance. To obtain a quality assessment, it is necessary to analyze the following data: 260/280: ratio of sample absorbance at 260 and 280 nm. The ratio of absorbance at 260 and 280 nm is used to assess the purity of DNA and RNA. A ratio of ~1.8 is generally accepted as "pure" for DNA. If the ratio is 5-2 times appreciably lower in either case, it may indicate the presence of protein, phenol or other contaminants that absorb strongly at or near 280 nm. 260/230: ratio of sample absorbance at 260 and 230 nm. This is a secondary measure of nucleic acid purity. The 260/230 values for "pure" nucleic acid are often higher than the respective 260/280 values. They are commonly in the range of 1.8-2.2. If the ratio is appreciably lower, this may indicate the presence of co-purified contaminants.

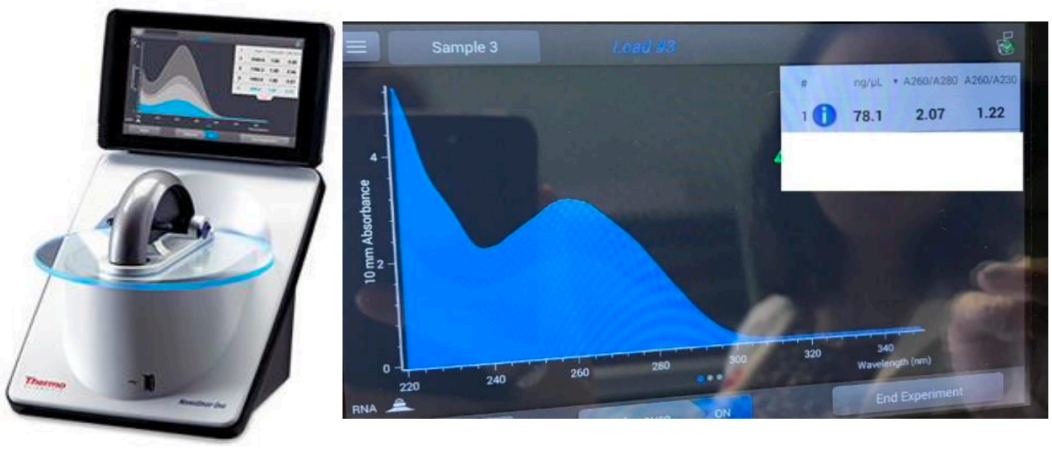


***Figure 15**: Evaluation of quantitative and qualitative parameter of RNA*

Quant-iTTM PicoGreen® dsDNA Assay

The major disadvantages of the absorbance method are the large relative contribution of nucleotides, single-stranded nucleic acids and proteins to the signal, the interference caused by contaminants commonly found in nucleic acid preparations, the inability to distinguish between DNA and RNA, and the relative insensitivity of the assay. Quant-iTTM PicoGreen® dsDNA reagent is an ultra-sensitive fluorescent nucleic acid stain for quantitating double-stranded DNA (dsDNA) in solution. Detecting and quantitating small amounts of DNA is extremely important in a wide variety of biological applications. These include standard molecular biology techniques, such as synthesizing cDNA for library production and purifying DNA fragments for subcloning, as well as diagnostic techniques, such as quantitating DNA amplification products and detecting DNA molecules in drug preparations. The Quant-iTTM PicoGreen® reagent has recently been used to quantitate PCR amplification yields in a method for direct cycle sequencing of PCR products. Quant-iTTM PicoGreen® dsDNA reagent enables to quantitate as little as 25 pg/mL of dsDNA (50 pg dsDNA in a 2 mL assay volume) with a standard spectrofluorometer and fluorescein excitation and emission wavelengths. It is also developed to minimize the fluorescence contribution of RNA and single-stranded DNA. PicoGreen assay was used to the quantitation of: - genomic DNA before NGS libraries preparation; - targeted libraries before NGS reactions.

Illumina sequencing technology

Illumina, born in 2006 after the acquisition of Solexa, developed the most widespread NGS technology in both diagnostic and research fields due to its reduced costs and rapid timing. The underlying concept of Illumina technology is similar to that of capillary electrophoretic sequencing, in which DNA polymerase catalyses the incorporation of fluorescence-labeled deoxynucleotide triphosphates into a DNA template during the repetition of synthesis cycles. During each cycle, when incorporated, the nucleotides are identified by fluorescence emission. The Illumina sequencing chemistry called "sequencing by synthesis" (SBS) (Figure 16) is based on reversible chain terminators that allows to identify the nucleotide sequence according to which single base is incorporated in the templates. Illumina high throughput analysis (HTA) workflow includes 4 steps: library preparation, cluster generation, sequencing and data analysis. After library preparation, which can be performed by different protocols, it is loaded into a flow cell, consisting of an optically transparent "slide" consisting of 8 compartments to which the anchoring oligonucleotides are linked to the library adapters. Under limit dilution conditions, the single-stranded

DNA template linked to the adapters is captured and immobilized by hybridization with the anchor oligonucleotides. Each fragment is then amplified to form different clonal clusters through a bridge-PCR amplification (Figure 16). Multiple amplification cycles convert the single DNA template molecule into a "cluster" of folded, clonally amplified fragments; each consisting of approximately 1000 amplicons. Up to 50 million of different clusters can be generated within a single flow cell. When the cluster formation is complete, the templates are ready for sequencing: the "clusters" are denatured, and a subsequent reaction of chemical cleavage and washing allows to obtain only the "sense" ("forward") filaments, since only these are sequenced the antisense filaments are denatured and removed. The sequencing of the "sense" filaments takes place through the hybridization of a "primer" complementary to the adaptive sequence opposite to that which still the filament to the flow cell: the free end of a fragment of DNA bends to bridge and hybridises to a close adapter primer with a complementary sequence. Subsequently, thanks to the action of DNA polymerase all 4 dNTPs linked to the reversible terminator and labelled with different fluorochromes, are incorporated into the DNA fragments of the clusters basing on the complementarity of each strand sequence, during each sequencing cycle. After incorporation, the excess reagents are removed with a wash to allow the polymerase adding the next complementary nucleotide (Figure 16). At each incorporation cycle, after a wash, the fluorescence relative to each cluster is detected and recorded by a CCD-camera. Fluorescence information is converted into the corresponding sequences and analysed (Voelkerding *et al.*, 2009).

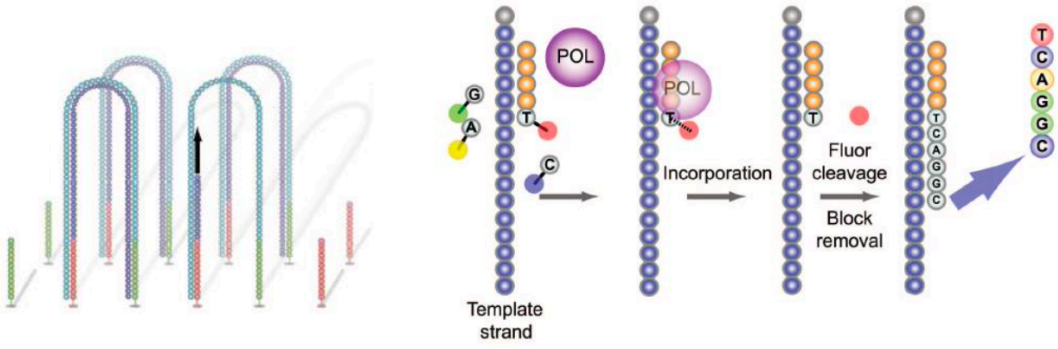


Figure 16:*Cluster generation by bridge PCR and sequencing by reversible dye terminators characterising Illumina sequencing technology.*

Genomic libraries preparation

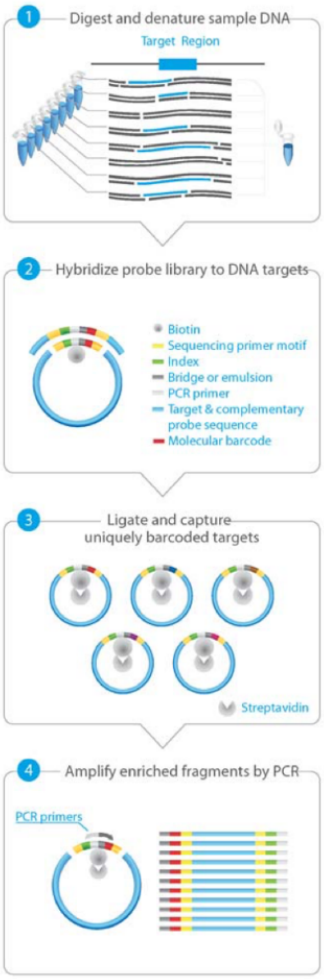


***Figure 17**: Overall HaloPlex HS target-enriched sequencing sample preparation workflow.*

The customized capture arrays used for the mutational analysis of the gene panels were designed to capture all coding exons, and flanking intron sequences of the genes in the panel, using SureDesign Agilent online tool (https://earray.chem.agilent.com/suredesign/).

The different libraries were prepared according to the kit manual.

The library preparation kit and the platform used for each gene panel is summarised in table 1.

The overall sample preparation workflows are presented in figure 17 and figure 18. Quality and characteristics (length and concentration) of libraries were detected by Agilent Bioanalyzer 2100 (High Sensitivity Kit).

***Table 1**:Llibrary preparation kits and platform used*

Mutational Panel	Library Preparation kit	Platform Used
Marfan Syndrome (Table 2)	SureSelect^{QXT} (Protocol, Version C1, December 2016. Agilent Tecnhologies.)	MiSeq® (Illumina) MiSeq® System Guide (Material # 20000262 Document # 15027617 v01. Illumina) MiSeq Reagent Kit v3 (Illumina)
Von Willebrand Factor (Table 4)	HaloPlex^{HS} Target Enrichment System (Protocol, Version C1, December 2016. Agilent Tecnhologies.)	
Dyslipidaemia (Table 3)	HaloPlex^{HS} Target Enrichment System (Protocol, Version C1, December 2016. Agilent Tecnhologies.)	

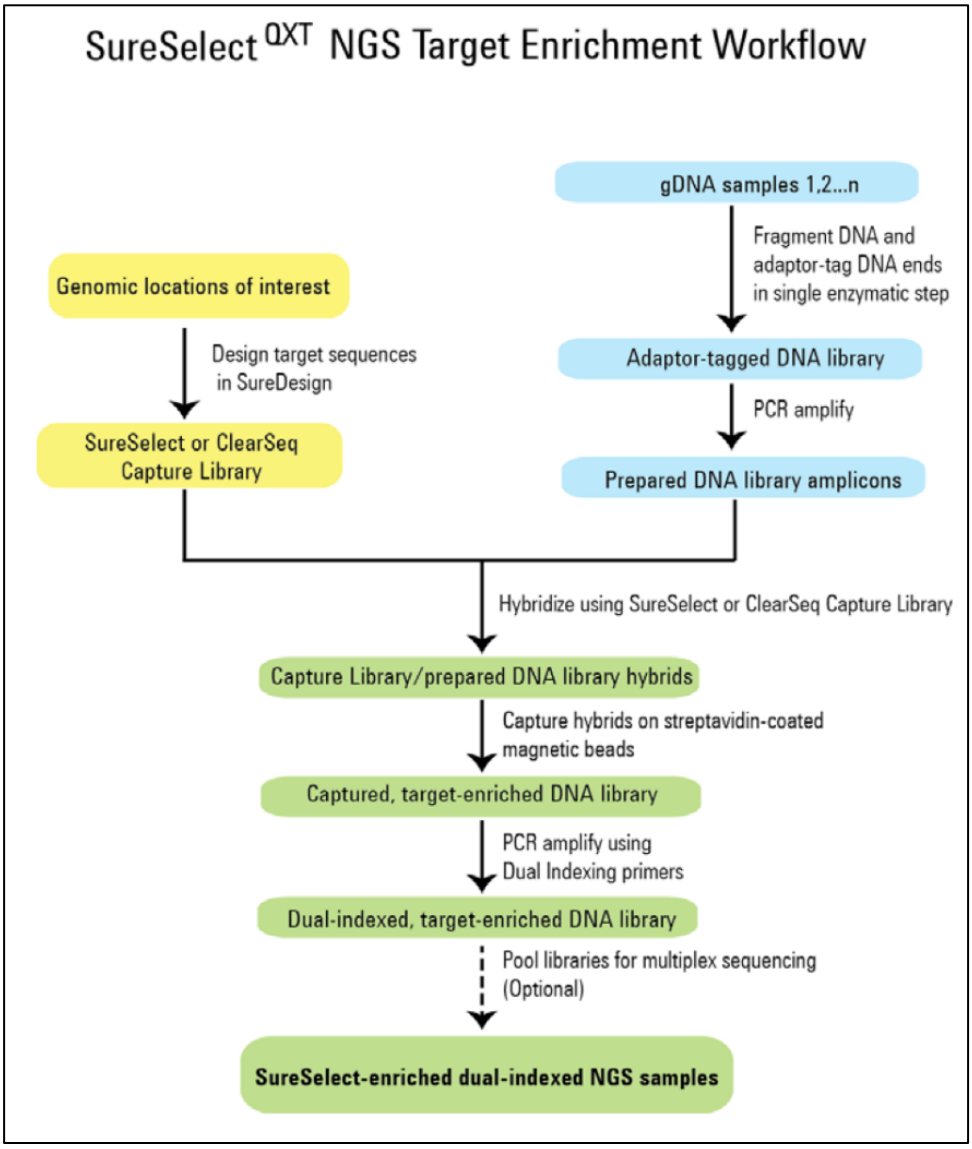


Figure 18*:Overall SureSelect qxt target-enriched sequencing sample preparation workflow.*

Table 2. Marfan syndrome and related disorders 97 genes panel

Gene	Name	Cytogenetic location	OMIM
COL11A1	COLLAGEN, TYPE XI, ALPHA-1	1p21.1	* 120280
MTHFR	5,10-METHYLENETETRAHYDROFOLATE REDUCTASE	1p36.22	* 607093
PLOD1	PROCOLLAGEN-LYSINE, 2-OXOGLUTARATE 5-DIOXYGENASE	1p36.22	* 153454
B3GALT6	BETA-1,3-GALACTOSYLTRANSFERASE 6	1p36.33	* 615291
ADAMTSL4	ADAMTS-LIKE 4	1q21.2	* 610113
MFAP2	MICROFIBRILLAR-ASSOCIATED PROTEIN 2	1p36.13	* 156790
PTGS2	PROSTAGLANDIN-ENDOPEROXIDE SYNTHASE 2	1q31.1	* 600262
SKI	V-SKI AVIAN SARCOMA VIRAL ONCOGENE HOMOLOG	1p36.33-p36.32	* 164780
TGFB2	TRANSFORMING GROWTH FACTOR, BETA-2	1q41	* 190220
CAPN2	CALPAIN 2	1q41	* 114230
AGT	ANGIOTENSINOGEN	1q42.2	+ 106150

MTR	5-METHYLTETRAHYDROFOLATE-HOMOCYSTEINE S-METHYL-TRANSFERASE	1q43	* 156570
TGFBR3	TRANSFORMING GROWTH FACTOR-BETA RECEPTOR, TYPE III;	1p33-p32	* 600742
TNXB	TENASCIN XB	6p21.33-p21.32	* 600985
COL11A2	COLLAGEN, TYPE XI, ALPHA-2	6p21.32	* 120290
COL9A1	COLLAGEN, TYPE IX, ALPHA-1	6q13	* 120210
DSE	DERMATAN SULFATE EPIMERASE	6q22.1	* 605942
FKBP14	FK506-BINDING PROTEIN 14	7p14.3	* 614505
FGF8	FIBROBLAST GROWTH FACTOR 8	10q24.32	* 600483
RET	REARRANGED DURING TRANSFECTION PROTOONCOGENE	10q11.21	+ 164761
ACTA2	ACTIN, ALPHA-2, SMOOTH MUSCLE, AORTA	10q23.31	* 102620
B3GAT3	BETA-1,3-GLUCURONYLTRANSFERASE 3	11q12.3	* 606374
LTBP3	LATENT TRANSFORMING GROWTH FACTOR-BETA-BINDING PROTEIN 3	11q13.1	* 602090
EFEMP2/fbln4	EGF-CONTAINING FIBULIN-LIKE EXTRACELLULAR MATRIX PRO-TEIN 2	11q13.1	* 604633
LRP5	LOW DENSITY LIPOPROTEIN RECEPTOR-RELATED PROTEIN 5	11q13.2	* 603506
CCND1	CYCLIN D1	11q13.3	* 168461
LRP6	LOW DENSITY LIPOPROTEIN RECEPTOR-RELATED PROTEIN 6	12p13.2	* 603507
COL2A1	COLLAGEN, TYPE II, ALPHA-1	12q13.11	+ 120140
LRP1	LOW DENSITY LIPOPROTEIN RECEPTOR-RELATED PROTEIN 1	12q13.3	* 107770
DCN	DECORIN	12q21.33	* 125255
LTBP2	LATENT TRANSFORMING GROWTH FACTOR-BETA-BINDING PROTEIN 2	14q24.3	* 602091
TGFB3	TRANSFORMING GROWTH FACTOR, BETA-3	14q24.3	* 190230
FBLN5	FIBULIN 5	14q32.12	* 604580
ADAMTS17	A DISINTEGRIN-LIKE AND METALLOPROTEINASE WITH THROM-BOSPONDIN TYPE 1 MOTIF	15q26.3	* 607511
CHST14	CARBOHYDRATE SULFOTRANSFERASE 14	15q15.1	* 608429
FBN1	FIBRILLIN 1	15q21.1	* 134797
SMAD3	MOTHERS AGAINST DECAPENTAPLEGIC, DROSOPHILA, HOMO-LOG OF, 3	15q22.33	* 603109
MYH11	MYOSIN, HEAVY CHAIN 11, SMOOTH MUSCLE	16p13.11	* 160745
ABCC6	ATP-BINDING CASSETTE, SUBFAMILY C, MEMBER 6	16p13.11	* 603234
MAPK3	MITOGEN-ACTIVATED PROTEIN KINASE 3	16p11.2	* 601795
PDIA2	PROTEIN DISULFIDE ISOMERASE, FAMILY A, MEMBER 2	16p13.3	* 608012
AXIN1	AXIS INHIBITOR 1	16p13.3	* 603816
MMP2	MATRIX METALLOPROTEINASE 2	16q12.2	* 120360
CRYBA1	CRYSTALLIN, BETA-A1	17q11.2	* 123610
COL1A1	CRYSTALLIN, BETA-A1	17q21.33	* 123610
ACE	ANGIOTENSIN I-CONVERTING ENZYME	17q23.3	+ 106180
KCNJ2	POTASSIUM CHANNEL, INWARDLY RECTIFYING, SUBFAMILY J, MEMBER 2	17q24.3	* 600681

EMILIN2	ELASTIN MICROFIBRIL INTERFACER 2	18p11.32	* 608928
SMAD2	MOTHERS AGAINST DECAPENTAPLEGIC, DROSOPHILA, HOMOLOG OF, 2	18q21.1	* 601366
SMAD4	MOTHERS AGAINST DECAPENTAPLEGIC, DROSOPHILA, HOMOLOG OF, 4	18q21.2	* 600993
LTBP4	LATENT TRANSFORMING GROWTH FACTOR-BETA-BINDING PROTEIN 4	19q13.2	* 604710
TGFB1	TRANSFORMING GROWTH FACTOR, BETA-1	19q13.2	* 190180
ADAMTS10	A DISINTEGRIN-LIKE AND METALLOPROTEINASE WITH THROMBOSPONDIN TYPE 1 MOTIF, 10	19p13.2	* 608990
MMADHC	MMADHC GENE	2q23.2	* 611935
ACVR1	ACTIVIN A RECEPTOR, TYPE I	2q24.1	* 102576
COL3A1	COLLAGEN, TYPE III, ALPHA-1	2q32.2	* 120180
COL5A2	COLLAGEN, TYPE V, ALPHA-2	2q32.2	* 120190
FN1	FIBRONECTIN 1	2q35	* 135600
COL6A3	COLLAGEN, TYPE VI, ALPHA-3	2q37.3	* 120250
EMILIN1	ELASTIN MICROFIBRIL INTERFACER 1	2p23.3	* 130660
LTBP1	LATENT TRANSFORMING GROWTH FACTOR-BETA-BINDING PROTEIN 1	2p22.3	* 150390
JAG1	JAGGED 1	20p12.2	+ 601920
EMILIN3	ELASTIN MICROFIBRIL INTERFACER 3	20q12	* 608929
MMP9	MATRIX METALLOPROTEINASE 9	20q13.12	* 120361
SLC2A10	SOLUTE CARRIER FAMILY 2 (FACILITATED GLUCOSE TRANSPORTER), MEMBER 10	20q13.12	* 606145
GATA5	GATA-BINDING PROTEIN 5	20q13.33	* 611496
CBS	CYSTATHIONINE BETA-SYNTHASE	21q22.3	* 613381
COL6A1	COLLAGEN, TYPE VI, ALPHA-1	21q22.3	* 120220
COL6A2	COLLAGEN, TYPE VI, ALPHA-2	21q22.3	* 120240
UFD1L	UBIQUITIN FUSION DEGRADATION 1-LIKE	22q11.21	* 601754
MAPK1	MITOGEN-ACTIVATED PROTEIN KINASE 1	22q11.22	* 176948
VHL	VHL GENE	3p25.3	* 608537
ZPLD1	ZONA PELLUCIDA-LIKE DOMAIN-CONTAINING PROTEIN 1	3q12.3	* 615915
MYLK	MYOSIN LIGHT CHAIN KINASE	3q21.1	* 600922
AGTR1	ANGIOTENSIN RECEPTOR 1	3q24	* 106165
PDCD10	PROGRAMMED CELL DEATH 10	3q26.1	* 609118
TGFBR2	TRANSFORMING GROWTH FACTOR-BETA RECEPTOR, TYPE II	3p24.1	* 190182
FBN2	FIBRILLIN 2	5q23.3	* 612570
NKX2-5	NK2 HOMEOBOX 5	5q35.1	* 600584
B4GALT7	BETA-1,4-GALACTOSYLTRANSFERASE 7	5q35.3	* 604327
ADAMTS2	A DISINTEGRIN-LIKE AND METALLOPROTEINASE WITH THROMBOSPONDIN TYPE 1 MOTIF, 2	5q35.3	* 604539
AGGF1	ANGIOGENIC FACTOR WITH G-PATCH AND FHA DOMAINS 1	5q13.3	* 608464
MTRR	METHIONINE SYNTHASE REDUCTASE	5p15.31	* 602568

NOS3	NITRIC OXIDE SYNTHASE 3	7q36.1	+ 163729
HOXA1	HOMEOBOX A1	7p15.2	* 142955
CCM2	CCM2 GENE	7p13	* 607929
KRIT1	KREV INTERACTION TRAPPED	7q21.2	* 604214
COL1A2	COLLAGEN, TYPE I, ALPHA-2	7q21.3	* 120160
TGFBR1	TRANSFORMING GROWTH FACTOR-BETA RECEPTOR, TYPE I	9q22.33	* 190181
PTGS1	PROSTAGLANDIN-ENDOPEROXIDE SYNTHASE 1	9q33.2	* 176805
ENG	ENDOGLIN	9q34.11	* 131195
COL5A1	COLLAGEN, TYPE V, ALPHA-1	9q34.3	* 120215
NOTCH1	NOTCH, DROSOPHILA, HOMOLOG OF, 1	9q34.3	* 190198
GNAQ	GUANINE NUCLEOTIDE-BINDING PROTEIN, Q POLYPEPTIDE	9q21.2	* 600998
AGTR2	ANGIOTENSIN II RECEPTOR, TYPE 2;	Xq23	* 300034
FLNA	FILAMIN A	Xq28	* 300017
ELN	ELASTIN	7q11.23	* 130160

Table 3. Familial dyslipidemia 57 genes panel

Gene	Name	Cytogenetic location	OMIM
LDLRAP1	LOW DENSITY LIPOPROTEIN RECEPTOR ADAPTOR PROTEIN 1	1p36.11	* 605747
PCSK9	PROPROTEIN CONVERTASE, SUBTILISIN/KEXIN-TYPE, 9	1p32.3	* 607786
ANGPTL3	ANGIOPOIETIN-LIKE 3	1p31.3	* 604774
CELSR2	CADHERIN EGF LAG SEVEN-PASS G-TYPE RECEPTOR 2	1p13.3	* 604265
APOA2	APOLIPOPROTEIN A-II DEFICIENCY, INCLUDED	1q23.3	+ 107670
APOB	APOLIPOPROTEIN B	2p24.1	* 107730
GCKR	GLUCOKINASE REGULATORY PROTEIN	2p23.3	* 600842
ABCG5	ATP-BINDING CASSETTE, SUBFAMILY G, MEMBER 5	2p21	* 605459
ABCG8	ATP-BINDING CASSETTE, SUBFAMILY G, MEMBER 8	2p21	* 605460
INSIG2	INSULIN-INDUCED GENE 2	2q14.1-q14.2	* 608660
ITIH4	INTER-ALPHA-TRYPSIN INHIBITOR, HEAVY CHAIN 4	3p21.1	* 600564
STAP1	SIGNAL TRANSDUCING ADAPTOR FAMILY MEMBER 1	4q13.2	* 604298
ABCG2	ATP-BINDING CASSETTE, SUBFAMILY G, MEMBER 2	4q22.1	* 603756
MTTP	MICROSOMAL TRIGLYCERIDE TRANSFER PROTEIN	4q23	* 157147
DAB2	DAB ADAPTOR PROTEIN 2	5p13.1	* 601236
GHR	GROWTH HORMONE RECEPTOR	5p13-p12	* 600946
HMGCR	3-HYDROXY-3-METHYLGLUTARYL-CoA REDUCTASE	5q13.3	+ 142910
SAR1B	SECRETION-ASSOCIATED RAS-RELATED GTPase 1B	5q31.1	* 607690
MYLIP	MYOSIN REGULATORY LIGHT CHAIN-INTERACTING PROTEIN	6p22.3	* 610082
HFE	HOMEOSTATIC IRON REGULATOR	6p22.2	* 613609
BTN2A1	BUTYROPHILIN, SUBFAMILY 2, MEMBER A1	6p22.2	* 613590

SLC22A1	SOLUTE CARRIER FAMILY 22 (ORGANIC CATION TRANSPORTER), MEMBER 1	6q25.3	* 602607
LPA	LIPOPROTEIN(a)	6q25-q26	* 152200
PPP1R17	PROTEIN PHOSPHATASE 1 REGULATORY SUBUNIT 17	7p14.3	* 604088
NPC1L1	NPC1-LIKE 1	7p13	* 608010
ABCB1	ATP-BINDING CASSETTE, SUBFAMILY B, MEMBER 1	7q21.12	* 171050
PON1	PARAOXONASE 1	7q21.3	+ 168820
LPL	LIPOPROTEIN LIPASE	8p21.3	* 609708
EPHX2	EPOXIDE HYDROLASE 2, CYTOSOLIC	8p21.2-p21.1	* 132811
GPIHBP1	GLYCOSYLPHOSPHATIDYLINOSITOL-ANCHORED HIGH DENSITY LIPOPROTEIN-BINDING PROTEIN 1	8q24.3	* 612757
DGAT1	DIACYLGLYCEROL O-ACYLTRANSFERASE 1	8q24.3	* 604900
ABCA1	ATP-BINDING CASSETTE, SUBFAMILY A, MEMBER 1	9q31.1	* 600046
CH25H	CHOLESTEROL 25-HYDROXYLASE	10q23.31	* 604551
OSBPL5	OXYSTEROL-BINDING PROTEIN-LIKE PROTEIN 5	11p15.4	* 606733
APOA5	APOLIPOPROTEIN A-V	11q23.3	* 606368
APOA4	APOLIPOPROTEIN A-IV	11q23.3	* 107690
APOC3	APOLIPOPROTEIN C-III	11q23.3	* 107720
APOA1	APOLIPOPROTEIN A-I	11q23.3	* 107680
ST3GAL4	ST3 BETA-GALACTOSIDE ALPHA-2,3-SIALYLTRANS-FERASE 4	11q24.2	* 104240
SLCO1B1	SOLUTE CARRIER ORGANIC ANION TRANSPORTER FAMILY, MEMBER 1B1	12p12.1	* 604843
GPD1	GLYCEROL-3-PHOSPHATE DEHYDROGENASE 1	12q13.12	* 138420
LRP1	LOW DENSITY LIPOPROTEIN RECEPTOR-RELATED PROTEIN 1	12q13.3	* 107770
SCARB1	SCAVENGER RECEPTOR CLASS B, MEMBER 1	12q24.31	* 601040
NYNRIN	NYN DOMAIN AND RETROVIRAL INTEGRASE CON-TAINING	14q12	
NPC2	EPIDIDYMAL SECRETORY PROTEIN	14q24.3	* 601015
LIPC	LIPASE, HEPATIC	15q21.3	* 151670
LMF1	LIPASE MATURATION FACTOR 1	16p13.3	* 611761
CETP	CHOLESTERYL ESTER TRANSFER PROTEIN, PLASMA	16q13	* 118470
LCAT	LECITHIN:CHOLESTEROL ACYLTRANSFERASE	16q22.1	* 606967
SREBF1	STEROL REGULATORY ELEMENT-BINDING TRAN-SCRIPTION FACTOR 1	17p11.2	* 184756
NPC1	NPC1 GENE	18q11.2	* 607623
CREB3L3	cAMP RESPONSE ELEMENT-BINDING PROTEIN 3-LIKE 3	19p13.3	* 611998
LDLR	LOW DENSITY LIPOPROTEIN RECEPTOR	19p13.2	* 606945
APOE	APOLIPOPROTEIN E	19q13.32	* 107741
APOC2	APOLIPOPROTEIN C-II	19q13.32	* 608083
LIPI	LIPASE I	21q11.2	* 609252
SREBF2	STEROL REGULATORY ELEMENT-BINDING TRAN-SCRIPTION FACTOR 2	22q13.2	* 600481

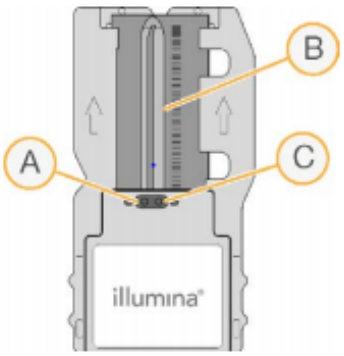
Table 4. von Willebrand Disease 10 genes panel

Gene	Name	Cytogenetic location	OMIM
ITGB3	INTEGRIN, BETA-3	17q21.31	* 173470
VWF	VON WILLEBRAND FACTOR	12p13.31	* 613160
ADAMTS13	A DISINTEGRIN-LIKE AND METALLOPROTEASE WITH THROMBOSPONDIN TYPE 1 MOTIF, 13	9q34.2	* 604134
F8	COAGULATION FACTOR VIII	Xq28	* 300841
GP1BA	GLYCOPROTEIN Ib, PLATELET, ALPHA POLYPEPTIDE	17p13.2	* 606672
GP1BB	GLYCOPROTEIN Ib, PLATELET, BETA POLYPEPTIDE	22q11.21	* 138720
GP5	GLYCOPROTEIN V, PLATELET	3q29	* 173511
GP9	GLYCOPROTEIN IX, PLATELET	3q21.3	* 173515
ITGA2B	INTEGRIN, ALPHA-2B	17q21.31	* 607759
P2RY12	PURINERGIC RECEPTOR P2Y, G PROTEIN-COUPLED, 12	3q25.1	* 600515

Illumina MiSeq® Platform

Figure 19: *Illumina MiSeq® platform*

The Illumina MiSeq® system combines proven sequencing by synthesis (SBS) technology with a revolutionary workflow that enables you to go from DNA to analyzed data in as little as 8 hours. The MiSeq integrates cluster generation, sequencing, and data analysis on a single instrument. A specially designed single-use prefilled reagent cartridge provides reagents for cluster generation and sequencing, including paired-end sequencing reagents and indexing reagents. To perform a run on the MiSeq, you need a single-use MiSeq Reagent Kit, which is available in different types and sizes. Each type of MiSeq Reagent Kit includes a kit-specific flow cell type and all reagents required for performing a run. The flow cell (Figure 20), PR2 bottle, and reagent cartridge (Figure 21) provided in the kit use radio-frequency identification (RFID) for accurate consumable tracking and compatibility. The MiSeq flow cell is a single-use glass-based substrate on which clusters are generated and the sequencing reaction is performed. Reagents enter the flow cell through the inlet port, pass through the single-lane imaging area, and then exit the flow cell through the outlet port. Waste exiting the flow cell is delivered to the waste bottle. Libraries are loaded onto the reagent cartridge before setting up the run, and then automatically transferred to the flow cell after the run begins.



A Outlet Port
B Imaging Area
C Inlet Port

Figure 20:. *Illumina MiSeq flow cell®*

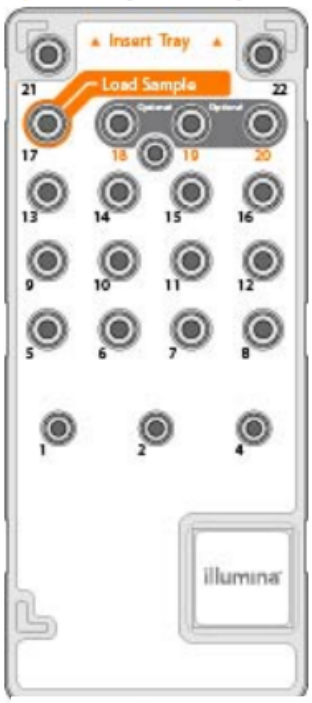


Figure 21:. *Illumina MiSeq® reagent cartridge*

Bioinformatic analysis of NGS data

The analytical pipeline was developed, implemented, and validated for data analysis of targeted sequencing for diagnostic purposes.

Fastq files quality was checked with FASTQC.

```
fastqc \
-o $ExperimentFolder/Fastq/QC \
-t 4 \
--nogroup $i
```

Adapters and quality trimming were performed using SurecallTrimmer.

```
java -jar /home/sam/bioinfo/AGeNT/SurecallTrimmer_v4.0.1.jar \
-fq1 ${i}_R1_001.fastq.gz \
-fq2 ${i}_R2_001.fastq.gz \
-$Tech \
-out_loc $ExperimentFolder/Fastq/TrimmedFastq
```

Trimmed reads were aligned to the human reference genome (Human GRCh37/hg19) using BWA-MEM.

```
bwa mem \
-M \
-t $Threads \
-C \
-R "@RG\tID:${i}\tLB:$LB\tPL:illumina\tPU:unknown\tSM:${i}" \
$Reference \
${i}_R1_001_trimmed.fastq.gz \
${i}_R2_001_trimmed.fastq.gz \
>$ExperimentFolder/SamFiles/${i}.sam
```

Sam files were converted to bam files using LocatIt.

```
java -jar /home/sam/bioinfo/AGeNT/LocatIt_v4.0.1.jar \
-i \
-o $ExperimentFolder/Intermediate_BamFiles/${g}.bam \
-X $ExperimentFolder \
-U \
$ExperimentFolder/SamFiles/${g}.sam
```

Intermediate bam files were sorted and indexed using samtools.

```
samtools sort \
$g \
```

```
samtools index \
$i
```

Bam files quality was evaluated with Qualimap.

```
unset DISPLAY
qualimap bamqc \
-bam $i \
-gff $FileBedQC \
-outdir $ExperimentFolder/BamFiles/QC \
-outfile `basename $i`.pdf
```

Variant calling was performed using GATK4 HaplotypeCaller in GVCF mode.

```
gatk HaplotypeCaller \
-R $Reference \
-I $f \
-L $FileBedVC \
-O $ExperimentFolder/GvcfFiles/`basename $f`.g.vcf \
-ERC GVCF
```

GVCF files were joined using gatk CombineGVCFs

```
gatk CombineGVCFs \
    -R $Reference \
    -O $ExperimentFolder/GvcfFiles/$FILE_OUT.g.vcf.gz \
    --variant $ExperimentFolder/GvcfFiles/gvcfs.list
```

and genotyped using GenotypeGVCFs tool.

```
gatk GenotypeGVCFs \
    -R $Reference \
    -V $ExperimentFolder/GvcfFiles/$FILE_OUT.g.vcf.gz \
    -O $ExperimentFolder/$FILE_OUT.vcf.gz
```

Gatk VariantFiltration tool was used to mark low quality variants.

```
gatk VariantFiltration \
    -V $ExperimentFolder/$FILE_OUT.vcf.gz \
    -filter "QD < 2.0" --filter-name "QD2" \
    -filter "QUAL < 30.0" --filter-name "QUAL30" \
    -filter "SOR > 3.0" --filter-name "SOR3" \
    -filter "FS > 60.0" --filter-name "FS60" \
    -filter "MQ < 40.0" --filter-name "MQ40" \
    -filter "MQRankSum < -12.5" --filter-name "MQRankSum-12.5" \
    -filter "ReadPosRankSum < -8.0" --filter-name "ReadPosRankSum-8" \
    -O $ExperimentFolder/$FILE_OUT.filtered_SNP.vcf
```

Variants were annotated using VEP 99 and several plugins.

```
vep \
    -i $ExperimentFolder/$FILE_OUT.filtered_SNP.vcf \
    -o $ExperimentFolder/$FILE_OUT.filtered_SNP.vep.vcf \
    --fasta /home/sam/bioinfo/database/humandb/Homo_sapiens.GRCh37.dna.primary_assemb
    --vcf \
    --cache \
    --offline \
    --everything \
    --check_existing \
    --total_length \
    --allele_number \
    --no_progress \
    --xref_refseq \
    --assembly GRCh37 \
    --pick \
    --custom /home/sam/bioinfo/database/humandb/gnomad.genomes.r2.0.1.sites.noVEP.vcf
    --plugin CADD,/home/sam/bioinfo/database/humandb/whole_genome_SNVs.tsv.gz,/home/s
    --plugin SpliceRegion \
    --plugin dbscSNV,/home/sam/bioinfo/database/humandb/dbscSNV1.1_GRCh37.txt.gz \
    --plugin MaxEntScan,/home/sam/bioinfo/database/humandb/MaxEntScan \
    --plugin GeneSplicer,/home/sam/miniconda3/bin/genesplicer,/home/sam/bioinfo/datab
    --plugin dbNSFP,/home/sam/bioinfo/database/humandb/dbNSFP4.1a_grch37.gz,FATHMM_pr
    --fork $Threads \
    --fields Location,Allele,Gene,Feature,Feature_type,cDNA_position,STRAND,VARIANT_C
```

Due to the diagnostic purpose of the analysis, it is mandatory to reach at least a coverage of 30× in the 99% of the genes in the panel. Regions presenting a lower coverage need to be sequenced again trough Sanger technology. A custom script was developed in R to check the coverage. The script, for each gene, computes the coverage of each region corresponding to a row in the .bed experimental design file (usually corresponding to each exon of the gene). It returns as output the boxplot in figure 22.

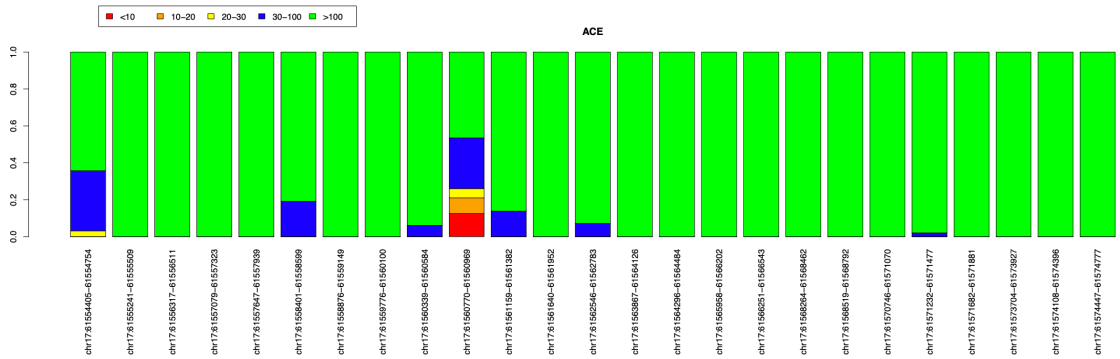


Figure 22: *Boxplot of the coverage in ACE gene. Each bar represents an exon.*

When a poorly covered region is detected, it automatically creates also a boxplot of the base per base coverage in that region, allowing the operator to know exactly which region needs to be repeated by Sanger (Figure 23).

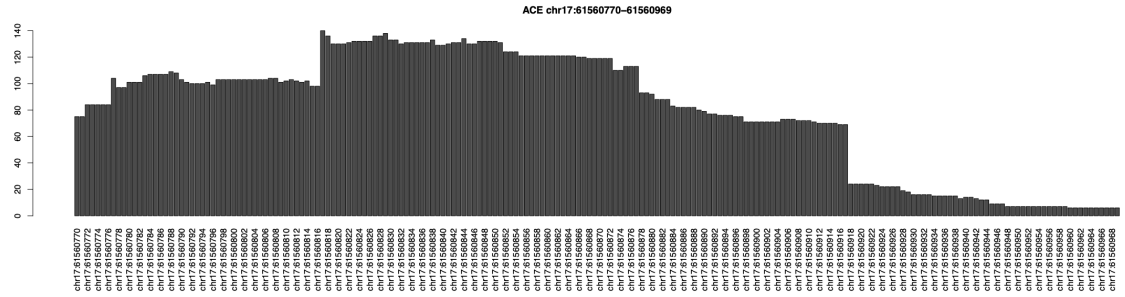


Figure 23: *Boxplot of the base per base coverage in the region detected to present a low coverage in ACE gene.*

Ninety-nine percent of targeted regions were covered. Variants were filtered according to the phred quality score (Q) ≥30 and a minimum coverage depth of 30×. Variants called following guidelines suggested by the Broad Institute, commonly accepted as standard, and identified according to (a) MAF <0.01; (b) the potential pathogenetic/modulatory role, according to variant classification recommendation (Richards *et al.*, 2015), literature genotype–phenotype association data and/or biological plausibility; (c) in silico predictor tools (CAD, SIFT1; PROVEAN2 ; PolyPhen-23 ; MutationTaster4 ; FATHMM5 ; Human Splicing Finder6 ; NetGene27); (d) type of genetic variants; (e) localization (exonic, splicing regions variants); and (f) allele balance >0.2 were validated trough Sanger sequencing.

Validation by Sanger Sequencing

Specific flanking intronic primer pairs for the selected NGS variants to be validated were designed using the Primer3 algorithm8, one of the most widely used primer designing tools (Kumar and Chordia, 2015). Primer sequences were checked for the presence of single-nucleotide polymorphism (SNPs) on their complementary DNA strands (visual inspection of sequences/SNP database)9. The resulting amplicons were checked for sequence similarity throughout the human genome using the Primer-BLAST tool10. The PCR amplicons were then purified and Sanger sequenced. PCR was performed in a final volume of 25 µl using FastStartTM Taq DNA Polymerase, 5 U/µl Kit (Roche), 1 µl of dNTPs (2.5 mM), 0.5 µl of each primer with a concentration of 10 pmol/µl (Eurofins Scientific, Luxembourg), and 2 µl genomic DNA in a concentration of approximately 50 ng/µl. Amplicons were purified using an Exonuclease I 20 U/µl/FastAP Thermosensitive Alkaline Phosphatase 1 U/µl (Thermo Fisher Scientific, United States) mixture. Purified PCR products were sequenced using the BigDye Terminator Kit v1.1 (Thermo Fisher Scientific, United States) and ABI 3500Dx Sequencer (Applied Biosystems, Foster City, CA, United States).

Results

Among the total of 945 variants identified by NGS and selected for Sanger validation, 942 (99,7%) were confirmed. The mean coverage of experimental sessions was 173× for the SureSelect approach and 1100× for the Haloplex approach with an at least 98% of analysable target bases; all variants met the phred-scaled quality score Q ≥30. Three out of 945 variants (0.3%) showed a discrepancy between the NGS datum and the subsequent validation. Two variants were in the LTBP2 gene while the third one involved the TGFB1 gene. All variants' discrepancies were related to their heterozygous/homozygous state. General characteristics of mutations are reported in Table 5. The depth of coverage for the three loci ranged from 173× to 199× (Table 5). All 3 variants were called as heterozygous and presented with balanced reads containing the wild-type or mutant allele (percentages of mutant on total alleles range from 45 to 54%).

Table 5*:Discrepant genetic variants analyzed in the study.*

	Gene	Variant description	dbSNP	MAF_EU (ExAC)	Chromosomic position (reads number per allele)	QUAL*
P1	LTBP2 NM_000428	c.3979C > T p.Arg1327Cys	rs758023418	0.0000088	chr14_74973455_C/T (0/1:107,92)	3197.33
P2	LTBP2 NM_000428	c.4203G > A p.Thr1401=	rs150977380	0.00428	chr14_74971852_C/T (0/1:96,103)	4613.20
P3	TGFB1 NM_000660	c.169C > A p.Leu57Met	rs1203938760	0.0000649	chr19_41858781_G/T (0/1:96,77)	2060.33

In patient 1 (P1), the first variant showing a discrepancy between the NGS call and the subsequent Sanger validation involved the LTBP2 gene (latent transforming growth factor-beta-binding protein 2, OMIM∗ 602091). The missense LTBP2 mutation, NM_000428.3:c.3979C > T (p.Arg1327Cys, rs758023418), had a MAF of 0.0000088 in the ExAC database for the European population. Primers were designed as previously described. The Sanger electropherogram in Figure 24A only revealed the presence of the mutated base (T at 3979 locus). The primer-binding regions were therefore checked for the presence of SNPs on their complementary DNA sequences and found negative. A second primer pair, external to the first ones, was designed and used for the direct sequencing of the new amplicons. The resulting electropherograms (Figure 24) showed the point mutation in the heterozygous state (as called by NGS) as well as a 2-bp deletion within the forward primer-binding DNA region (NM_000428.3:c.3908-98_3908-97delAG, rs149267227, Figure 24C). This intronic variant had a MAF in the European population of 0.0029.

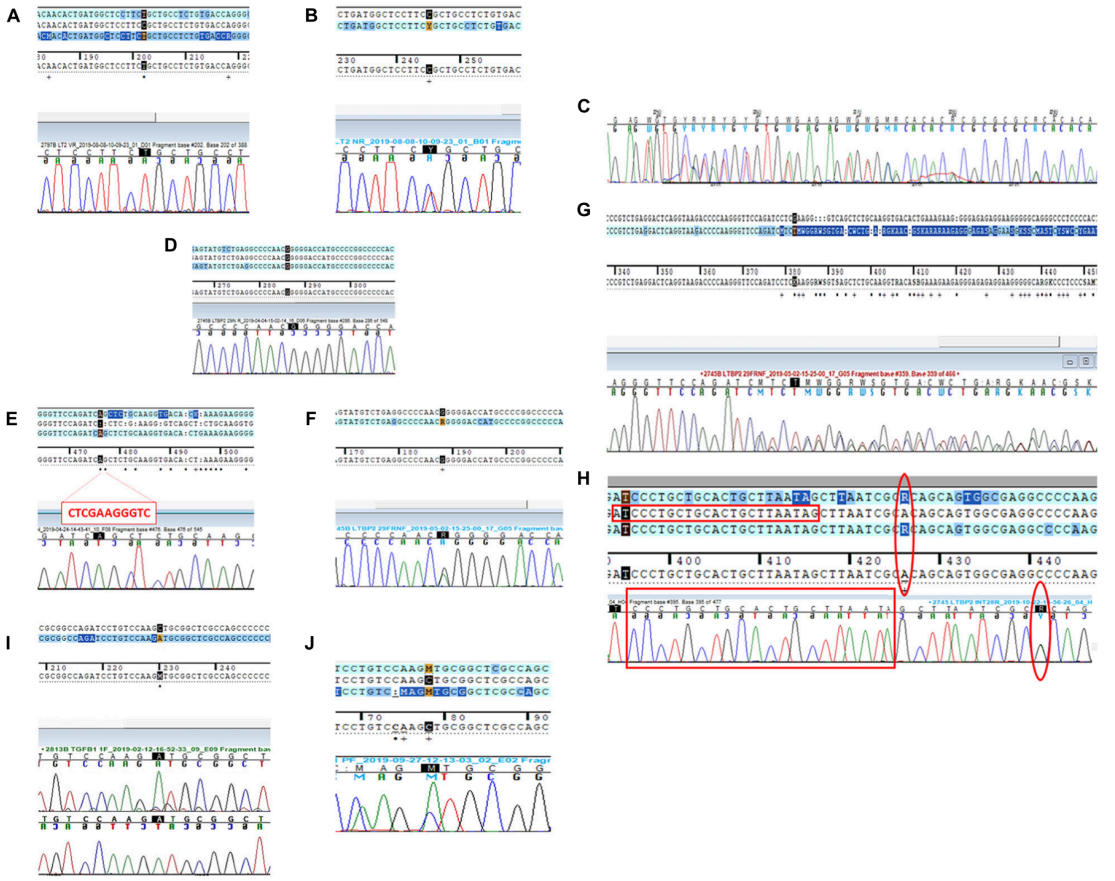


Figure 24: *Direct sequencing electropherograms of the genetic variants in (A) LTBP2 (P1), original primers; (B) LTBP2 (P1), new primers; (C) LTBP2 deletion (P1), new primers; (D) LTBP2 point mutation (P2), original primers; (E) LTBP2 deletion (P2), original primers; (F) LTBP2 point mutation (P2), second F primer + original R primer; (G) LTBP2 deletion (P2), second F primer + original R primer; (H) LTBP2 (P2) intronic primers (the empty red rectangles indicate the original F primer sequence; the empty red circles indicate the rs11846588 SNP proximal to the 3′ primer end); (I) TGFB1 (P3), original primers; and (J) TGFB1 (P3), new primers.*

In patient 2 (P2), LTBP2 second variant (NM_000428.3:c.4203G > A) was a synonymous mutation (p.Thr1401=) with a MAF of 0.00428 in the European ExAC population. Primers were constructed, and the electropherogram, deriving from the direct sequencing, did not reveal the point mutation (only the wild-type base was present) (Figure 24D); in addition, an 11-bp deletion was identified at the homozygous state (Figure 24E). The combination of a second forward primer and the original reverse one generated amplicons whose electropherograms revealed both the mutation and the deletion in the heterozygous state, as they were called by NGS (Figures 24F, G).

A third intronic primer pair was designed in order to analyse the sequence of the original forward primer. The resulting electropherogram was negative although it actually revealed the presence of an SNP [NM_000428.3:c.4178- 224A > G, rs11846588, MAF_EU (ExAC) = 0.16] proximal to the 3′ primer end which was not previously detected (Figure 24H).

Patient 3 (P3) The last analysed variant was a missense mutation (NM_000660.6:c.169C > A, p.Leu57Met, rs1203938760) involving the TGFB1 gene (transforming growth factor-beta-1, OMIM∗190180) with a MAF in the European ExAC population of 0.0000649. We proceeded with the Sanger validation ultimately obtaining an electropherogram showing the variant in the homozygous state (Figure 1I). Once again, we tested the primer pair used for the initial amplification for their capacity to bind to SNPs and they resulted negative. We therefore checked the experimental steps (initial amplification or direct sequencing) in order to unequivocally identify the phase in which the allelic loss took place. A restriction fragment length polymorphism (RFLP) procedure was performed using AluI enzyme cutting at a specific cleavage DNA site (5′ . . .AG ↓CT. . .3′), which was identified by the NEBcutter V2.011 online tool. The digestion reaction produces two different bands profiles when visualized on a 3% agarose gel according to the presence of the mutation at the heterozygous or homozygous state (Figure 25A). The results of the AluI digestion performed on our PCR product along with a negative control are shown in Figure 25B. Our sample digestion profile demonstrated the mutation to be in the heterozygous state, as the enzyme was able to recognize three cleavage sites on the wild-type sequence and two on the mutated one (51, 82, 171, 192, and 243 bp). A new, internal primer pair was able to confirm the NGS call (Figure 24J).

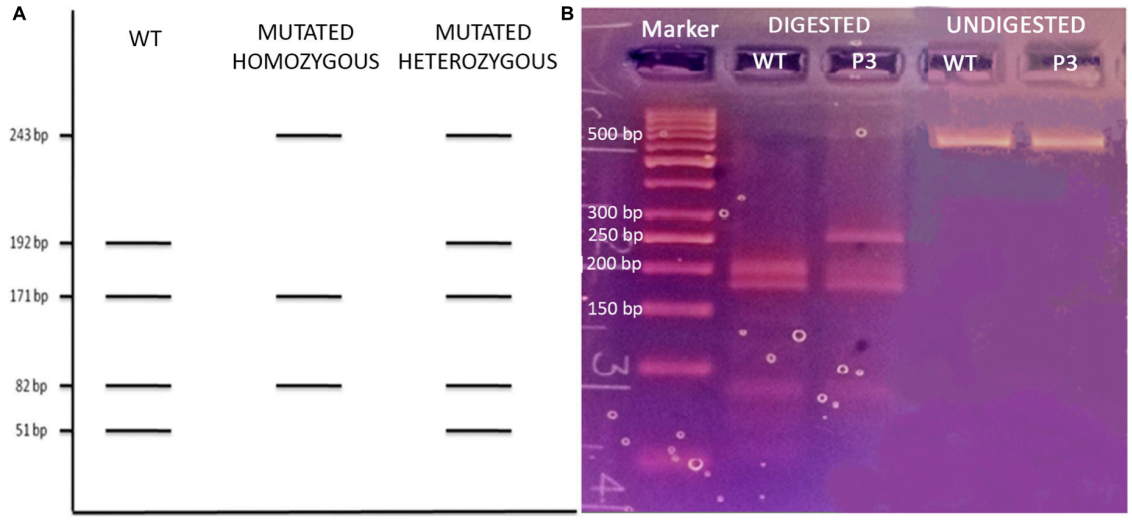


Figure 25: *(A) Schematic representation of the AluI enzyme digestion on a wild-type, homozygous, and heterozygous mutated TGFB1 fragment; (B) 3% agarose gel electrophoresis of the AluI enzyme digested and undigested P3 and WT control.*

Discussion

In this study, we showed that high-quality NGS data robustness could not benefit from Sanger validation due to errors occurring in the technology.

The two last decades witnessed the rapid and impressive advancements of high-throughput sequencing techniques, in terms of both experimental workflows and data-processing pipelines dedicated to the management of the large volumes of data generated by the various platforms. These technologies provided indeed unprecedented opportunities for the study and characterization of variants at the DNA and RNA level. Genetic information is, in fact, investigated at such a level of precision that the supporting role of these technologies for the study of complex hereditary phenotypes has been implemented in both research and diagnostics, also for the capability of different platforms to analyse and interrogate large gene panels, exomes, and genomes in times and costs that are progressively decreasing. Despite the attempt made by the European Guidelines (2016) (Matthijs *et al.*, 2016) to provide the most useful indications to laboratories for the evaluation and validation of variants identified by NGS, these technologies remain strictly dependent on computational tools and bioinformaticians for the highly complex data analysis, whose quality parameters may vary between laboratories as well as the pipelines used for alignment, variant calling, filtering, and annotation of variants. Current NGS guidelines do not define quality parameters or concrete guidance for confirmatory analysis. Therefore, Sanger sequencing is still considered the gold standard for the validation of NGS genetic variants and an essential step in the diagnostic routine. This kind of approach, however, raised a question about the actual cost-effectiveness of using very powerful platforms generating increasing quality data at progressively decreasing costs and the need to apply a "time- consuming" and not error-free [i.e., Allelic dropouts (ADOs)] technique, whose cost does not decrease as quickly over time, to validate the results.

Allelic dropout represents a significant cause of genotyping errors, potentially determining substantial negative consequences as it might lead to misdiagnosis of genetic diseases and false-negative/positive results depending on the allele that drops out (mutant or wild-type). ADO during PCR can be caused by a variety of mechanisms, and several factors influencing its rate have been described. Most ADO mechanisms are determined by the presence of a single-nucleotide variant (SNV) situated inside the primer-binding sequences on the targeted DNA, the SNV causing the failure of the primer-template annealing and the consequent amplification error. Concomitant presence of a differential allelic methylation and G-quadruplex motifs in some regions of the genome, DNA degradation

leading to PCR refractory breaks, imperfect PCR conditions preventing DNA template accessibility, and presence of both homopolymer tracts and pseudogenes were also described as potential determinants of ADO or preferential amplification (Piyamongkol *et al.*, 2003; Stevens *et al.*, 2017).

In the current study, 942 out of 945 (99.7%) high-quality NGS variants identified in 218 subjects were validated by Sanger sequencing. Our data are in keeping with previous studies, suggesting that Sanger sequencing may not represent a necessary step to validate NGS variants when dealing with data meeting high-quality scores and an adequate depth of coverage. In fact, several previous studies evaluating data from different NGS platforms and approaches (targeted, exome, or whole-genome sequencing) identified almost 100% Sanger validation rate on a total of 14,495 variants (McCourt *et al.*, 2013; Sikkema-Raddatz *et al.*, 2013; Strom *et al.*, 2014; Baudhuin *et al.*, 2015; Beck *et al.*, 2016; Zheng *et al.*, 2019). In these studies, variants not validated by Sanger sequencing did not match adequate quality scores (Strom *et al.*, 2014; Beck *et al.*, 2016; Zheng *et al.*, 2019). Both ours and literature data move in the direction of a limited usefulness of Sanger validation for NGS-derived variants associated with robust quality scores, suggesting a re-consideration of its application in routine diagnostics that should be limited to validation of a specific clinical phenotype-associated variant, quality assurance, and risk-avoidance purposes. Beyond the previous issue, whether an accurate NGS approach is used, our data demonstrated that potential well-known failures affecting Sanger technology could further reduce its utility and instead determine higher costs and delay in analysis conclusion and laboratory report. In fact, among 945 variants, we identified three discrepant variants. Despite their high-quality parameters, namely, a balanced read number and high depths of coverage, Sanger validation failed in confirming NGS datum: in all three cases, NGS attributed a heterozygous state and Sanger sequencing, a mutated homozygous state in P1 and P3, and a wild-type homozygous state in P2. In the first case (LTBP2 gene variant), the discrepancy was due to the presence of a small deletion in the DNA region binding the forward primer of the original pair. This deletion, potentially preventing the correct primer annealing during the initial amplification phase, had a frequency below 1% in the European ExAC population. This is of note as this rarity could presumably allow some of the online-automated primer-designing programs (which refer to dbSNP databases to omit common variants) as well as a manual approach to miss this kind of information during the primer construction and its inclusion in their sequence. This aspect is similar for private mutation of the patient in analysis. In fact, these variants evade the “masking” phase of the primer construction process where common variants are averted. As concerns

the further discrepant variant at the LTBP2 locus, we speculated that this discrepancy might be the result of an alternative ADO phenomenon. In particular, the presence of SNV outside primer sequences (non-primer-binding-site SNVs) was demonstrated to promote a hairpin formation of the PCR products, this secondary structure preventing further amplification and extension failures (Lam and Mak, 2013). Lam and coworkers were in fact able to demonstrate that a heterozygous NGS deletion (FAH, NM_000137.1:c.1035_1037del) resulted in homozygosity at a first Sanger sequencing run due to a non-primer-binding-site SNV (FAH:c.961-35C, rs2043691) forming a strong hairpin structure and leading to amplification failure of the wild-type allele. Similarly, in our case, a non-primer-binding site SNP located outside the 3′ end of the original F primer (Figure 24H) was identified, which can be presumably held accountable due to the discrepancy we observed. Regarding the third case, we were instead able to demonstrate that ADO had not occurred during the first amplification cycle via traditional PCR but exclusively within the direct sequencing reactions. Even though Sanger sequencing by using the original primers pair showed a mutated homozygous state, RFLP procedure with the same primers showed the presence of both alleles of TGFB1 mutation. The heterozygous state was confirmed at a second Sanger validation run with new primers. Our data do not define the fine mechanisms behind the ADO we observed in case number 3 during Sanger sequencing reaction; nevertheless, the previously mentioned mechanisms (apart from those involving SNV inside or proximal to the primer binding sequences) could be a potential explanation for the discrepancies we observed. These phenomena could also be impacted by intrinsic characteristics of Sanger technology which uses chain-terminating di-deoxynucleotides under suboptimal PCR conditions. Direct sequencing in fact employs a multiplex PCR ensuring the amplification of several distinct DNA fragments in a single reaction. This strategy necessarily operates under less stringent PCR conditions which might adversely affect amplification of individual loci or lead to secondary structure formation, thus invalidating the synthesis of the newly generated DNA strands (Piyamongkol *et al.*, 2003).

In conclusion, our data suggest the chance of Sanger approach errors that go beyond the presence of common variants. Actually, both studies of Beck *et al.* (2016) and Zheng *et al.* (2019) had also encountered this kind of criticism: in the first case, authors had to re-sequence 19 discrepant variants, 17 of which were validated after a second sequencing run as the first orthogonal Sanger validations were themselves incorrect. 17 of the NGS variants would have been considered false positives if a single round of Sanger sequencing were used as validation criteria. Zheng *et al.* (2019) were instead able to validate 98

high-quality NGS variants by mass spectrometry but not by Sanger sequencing, being some of those variants characterized by the presence of a homopolymer at the 100-bp flanking sequence or within pseudogenes. The authors suggested that if such practice were used in a clinical setting, more positive NGS variants would be discarded or incorrectly designated, when compared to using the NGS data directly. In fact, PCR-based amplification remains susceptible to ADO due to different mechanisms, such as private variants within primer-binding sites or secondary structure formation, potentially determining false-positive/negative results, heavily impacting genetic diagnosis of several diseases in the clinical setting. In addition to that, non-primer-binding-site SNVs have been demonstrated to have the ability to interfere with the PCR as well, making the primer designing process more laborious and time-consuming. Some genomic regions are also extremely difficult to amplify and might not yield high-quality Sanger results even after multiple attempts, thus possibly rendering the Sanger validation of a high-quality NGS variant not an adequate support. NGS still remains susceptible to errors, but, in our experience and according to those of many laboratories around the world, variants above the technology-dependent quality threshold are confirmed by Sanger in almost 100% of cases. Our study demonstrated that in case of discrepancy between a high-quality NGS variant and the subsequent Sanger validation, NGS call should not be a priori assumed to represent the source of the error. On the other hand, NGS approaches [targeted/whole-exome (WES)/genome sequencing (WGS)] exhibit some limitations, as they are characterized by a lack of uniformity in sequencing depth. In addition, due to capture of probe design matching the reference sequencing, preferentially enrichment of reference allele might represent a further NGS bias. Moreover, a further limitation of the NGS approach might be represented by the possible presence of false negatives (although reduced with the progressive improvement in experimental/computational analysis procedures) and in turn lack of information concerning genetic variants clinically relevant, even though standardized criteria are usually adopted (Sims *et al.*, 2014; LaDuca *et al.*, 2017; Dunn *et al.*, 2018). Moreover, certain types of variants (copy number variations, large genomic rearrangements) remain difficult to detect by NGS, but experimental workflows, bioinformatic tools, and open- source software are being developed and constantly updated in order to overcome that issue (Shen and Seshan, 2016; Anwar *et al.*, 2020; Han *et al.*, 2020).

Development and application of a bioinformatics pipeline to evaluate the global RNA expression profiles from cerebral thrombi, obtained during thrombectomy treatment and from peripheral venous blood in patients with acute ischemic stroke, using Affymetrix technology.

Introduction to microarrays

Microarrays are tools that allow the identification and quantification of the mRNA transcripts present in the cells. The number of molecules of mRNA, coming from the transcription of a given gene, can be considered as an approximation to the level of expression of that gene. A microarray consists of a solid surface on which strands of polynucleotide called probes have been attached or synthesized in fixed positions. Two types of expression microarrays are the most popular. One of the main differences among them relies on how these probes are put on the slide.

- Spotted or cDNA microarrays take their name because probes are synthesized apart and printed mechanically on the slide. The term cDNA is used because the probe is a complimentary copy of the original sequence and each probe represents one gene.
- In oligonucleotide chips, where main representatives are Genechip and Affymetrix(c), the name of the commercial brand that manufactures them, the probes are directly synthesized on the surface. The term oligonucleotide refers to the fact that the synthesis process allows to create only small fragments so that a gene is not represented by one probe but by as a set of them (a probe set).

To start a microarray experiment RNA is extracted from the subject cells. After this, some of its molecules are substituted by others containing a fluorescent dye. The resulting labelled transcripts are called targets. Once the samples are prepared, they are deposited over the array and left inside a hybridization chamber for some hours. The labelled targets bind by hybridization to the probes on the array with which they share sufficient sequence complementarity. After this time the array is washed to eliminate the targets that have not hybridized (Lipshutz *et al.*, 1999).

The way in which the previous step is performed is the second important difference between the two types of chips: In spotted microarrays cDNAs from two tissues of interest, labelled with fuorescent dyes of different color (usually red and green), are hybridized to a single chip. The two targets are said to compete to hybridize with the probes.

The Affymetrix system hybridizes only one sample per chip (Figure 26). This requires more slides per experiment and does not enjoy the advantage of using competitive hybridization, however it simplifies experimental design and is based on a much more sensitive technology. At this point each probe on the microarray may be bound to a certain quantity of labelled target that should be proportional to the level of expression of the gene represented by that probe. To determine the amount of sample hybridized the microarray is illuminated by a laser light that causes the labelled molecules to emit fluorescence (proportionally to their quantity). This fluorescence is captured by a scanner yielding an image that consists in a grid of shined spots, corresponding each one to a probe. Finally, this image will be transformed into numbers and will be the basis of the analysis. Gene variants and gene expression of an individual characterize the susceptibility to a disease, or its complications, or the subject's response to pharmacological treatments. Microarray technology represents an important resource for the researcher and the clinician to understand the molecular bases of diseases and, increasingly, for a more established diagnostic and prognostic therapy and for the application of new personalized therapeutic strategies (Auer *et al.*, 2009).

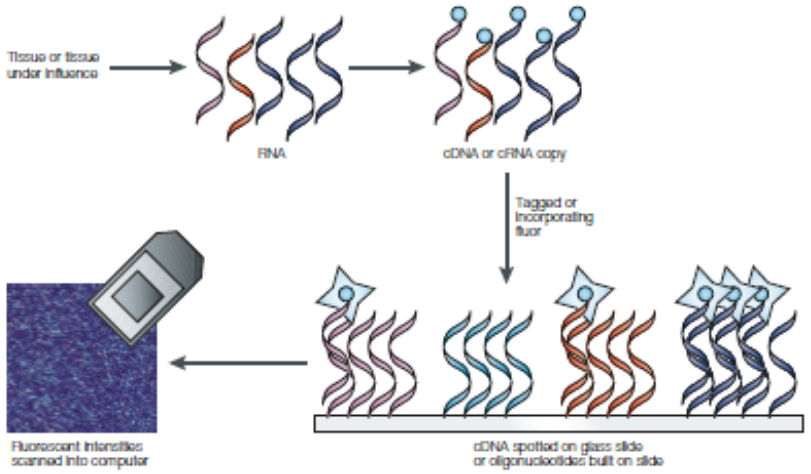


Figure 26: *Schematic experimental process using a microarray (from Buttle A, Nat Rev Drug Dis-cov. 2002)*

Microarray Analysis Process

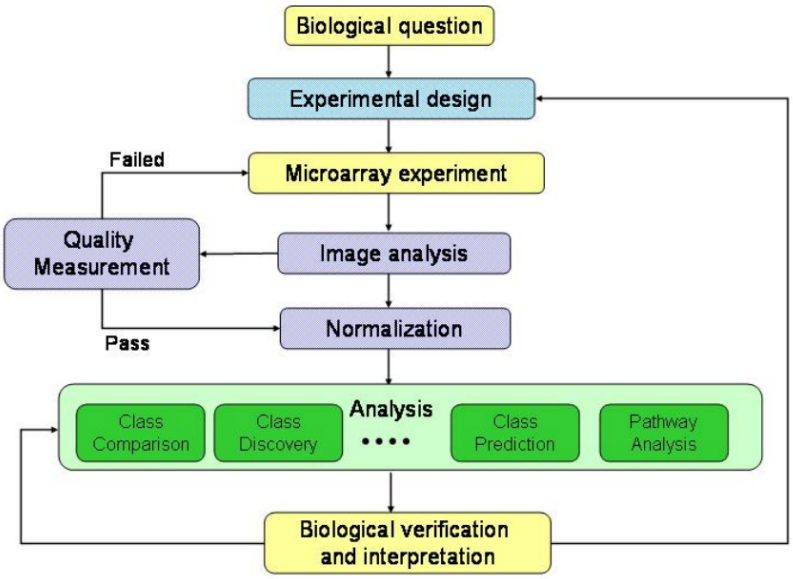


Figure 27*: Workflow of microarray analysis.*

Quality control

A microarray experiment produces a set of images that are transformed into numerical values representing absolute (single channel) or relative (two channel) intensities.

As in any statistical analysis, and particularly in image analysis, the first step is to check the quality of the data. High throughput data have an additional difficulty: the huge data matrices obtained make it virtually impossible to detect most problems by visual inspection. Specific quality control procedures have been developed to address this issue.

The goal of the quality control step is to determine if the whole process has worked well enough so that the data can be considered reliable. There are no standard methods for microarray QC. Most quality controls are based on images and plots.

Background Correction and Normalization

Once the quality of the data has been assessed it is still necessary to make some pre-processing before the analysis:

The *background adjustment* remove signal due to non-specific hybridization, that is signal emitted by other things than sample hybridized to probe.

The *normalization* of the data correct for systematic biases due to causes such as different dye absorption, spatial heterogeneity in the chip or others.

In Affymetrix arrays, it is also necessary to summarize the different signals obtained from all the probes representing one gene in a unique value.

The goal of the microarray production process is to obtain an intensity value which can be considered proportional to the level of expression. This is based on determining how much hybridization has been produced between the sample and the targets. It is known that a part of the observed signal is due to non-specific binding, that is, a small quantity of the sample may combine to non-complementary chains. Some of the signal may also be due to non-biological sources. Altogether there is a need to estimate and remove the signal due to specific (real) hybridization from that due to any other reasons, generically called background. Different methods have been developed as alternatives and several comparisons have been published (Ritchie *et al.* 2017). A general conclusion of these studies is that model based methods are those performing best at removing background. Three commonly used methods are: normexp (Smyth *et al.*, 2005) for two channel arrays, VSN (Hueber *et al.*, 2002) for both types of arrays and RMA (Irizarry *et al.*, 2003) for oligonucleotide chips. The last two methods combine background correction and normalization in the same process. Normalization is a key point in the microarray analysis process and much effort has been devoted to developing and test different methods (Quackenbush, 2002). One reason for such abundance is that there are different technical artifacts that must be corrected for, and not every method can deal with all of them. In general, normalization methods are based on the principle that most genes in the array are either not expressed or equally expressed in any condition. Only a small number of genes should show changes of expression between conditions. The most used method for one channel arrays like Affymetrix chips is RMA (Robust Multichip Average). It consists of three steps: a background adjustment based on a probe level model, a quantile normalization and, finally, a summarization integrating the values of all probes corresponding to one gene. A conceptually simpler approach is the one proposed by the manufacturer of the chips: the MAS5 algorithm. Some studies comparing both (and other) methods (Bolstad *et al.*, 2003; Hill *et al.*, 2001) concluded the superiority of the RMA method, although, this is not a closed discussion yet.

Statistical analysis of microarray data

The steps described in the previous section are preparatory for data analysis. The output of this initial process is the gene expression matrix, whose rows (1000-50000) represent the genes and whose columns represent the samples (from 2 to several hundreds).

Class Discovery

Clustering, also known as class discovery, is the most popular method currently used in the first step of gene expression matrix analysis. Clustering, much like Principal

Components Analysis (see pg. 50), reduces the dimensionality of the system and by this, allows easier management of the data set. Clustering techniques can be applied to construct classifications of arrays (experimental conditions), genes or both together. When they are applied to cluster the genes they can help to identify groups of coregulated genes, to identify spatial or temporal expression patterns, to reduce redundancy in prediction models. If they are used to cluster samples they will be useful to identify new biological classes or to detect experimental artifacts.

Hierarchical clustering has been mainly used to find a partition of the samples more than of the genes because there are much less samples than genes so that, with genes, the resulting dendrogram is often difficult to interpret. Eisen *et al.*, 1998 is the classical reference on using hierarchical clustering with microarray data. A popular display, related to this method, is a colour image plot called heatmap (Gentleman *et al.* 2005) which consists of a rectangular array of coloured blocks, with the colour of each block representing the expression level of one gene on one array. Typically, in a heatmap (Figure 28), shades of red are used to represent degrees of increasing expression, and shades of blue are use to represent degrees of decreasing expression. Each column of boxes represents an array and each row of boxes corresponds to a gene. Heatmaps display intensities and can be used independently of clustering. However it is very common to perform a hierarchical clustering of samples and/or genes and to sort the columns and/or rows according to the resulting dendrogram to emphasize the presence of groups. The k-means method (Kaufman & Rosseeuw, 1990) is also very popular although it has the disadvantage that it does require specification of a number of clusters and an initial partitioning, what makes the final results to be very sensitive to these choices. In this case the researcher may try different cluster numbers (k) and then pick up the k number that fits best the data. In addition, the resulting groups may change between successive runs because of different initial clusters. K-means and hierarchical clustering share another problem, which is more difficult to overcome, that the produced clustering may be hard

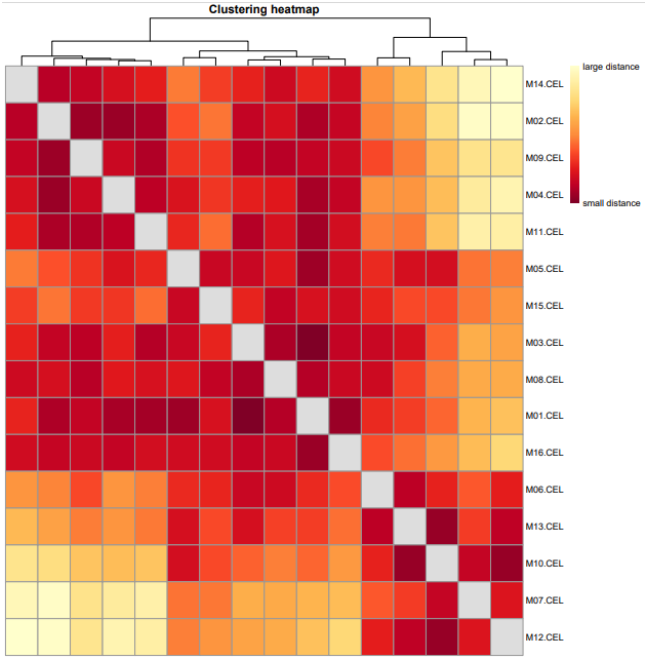


Figure 28. *Example of heatmap.*

to interpret: the order of the genes within a given cluster and the order in which the clusters are plotted do not convey useful biological information. This implies that clusters that are plotted near each other may be less similar than clusters that are plotted far apart. Another important application of class discovery is to perform it on all the genes, not only the ones that have been selected. One can cluster the initial (normalized) dataset to discover patterns, probably due to some systematic (block) effect. There can be multiple sources of systematic variation: production batch, technician, biological source (cell lines) etc. Clustering samples with all the genes followed by an appropriate visualization can help discovering the existence of these effects. After detecting such unexpected effects, it is possible to include them into the model used for detecting differentially expressed genes so that they can be estimated and eventually removed.

Class Comparison

The class comparison problem can be defined as the selection of genes whose expression is significantly different between conditions. These are called differentially expressed genes. There are many models and methods available for the analysis. Some are based on parametric models whereas other rely on nonparametric approaches in order to overcome the difficulties associated with distributional assumptions.

Model based methods use analysis of the variance models ANOVA to capture the main sources of variability in the experiment. In this case a single model is used for all the genes simultaneously. Global tests, despite their name, analyse each gene separately, using a common model which can be parametrical or not. The SAM method (Tusher *et al.*, 2001) is the most popular non-parametric approach, while the limma method (Gordon *et al.*, 2005) is the most common parametric approach using linear models and empirical bayes.

Model-based methods

Wu et al proposed an analysis of variance model specified in two stages for two color microarrays where the expressions are treated separately (that is, it relies on absolute expression values). The firsters-stage model is as follows:

$$Y_{ijgr} = \mu + A_i + D_j + AD_{ij} + r_{ijgr},$$

where the indices track the (A)rray (i), the (D)ye, (j), the gene (g) and the (r)eplicated measurement (r). The first stage generates the term rijgr which, in a second stage, is mod-

$$r_{ijgr} = G + TG_{ij} + DG_j + AG_i + \epsilon_{ijr},$$

elled in terms of gene specific effects as:

where G is the average intensity associated with a particular gene, AG_i is the effect of the array on that gene, DG_j is the effect of the dye on that gene and ijr is the residual. T G_{ij} is called the "treatment by gene" term and is the main interest in the analysis which captures variations in the expression levels of a gene across samples. It must be noted that this approach does not need a previous normalization to account for dye or array effect, because this is already done by the corresponding dye or array terms. The gene specific model can be modified for Affymetrix data by removing the DG and AG terms because there is no dye factor (one-color) and the array effects become part of the residual error term. In practice what a user will do is to fit model 6 to the data and call differentially expressed those genes where the interaction term TG is significative (Wu *et al.*, 2003).

Global tests

One of the main practical differences between model-based and global methods lies in the way that normalization is done. Model-based methods do it implicitly when the model is fitted whereas global tests require a previous normalization step. If one considers one gene at a time a microarray experiment can be seen as simply an experiment so that a reasonable way to analyze it is to use a standard linear model approach. This is however considered inefficient due mainly to two common problems in this type of experiments: firtrst, sample sizes very small, which complicate variance estimation; second, the variances themselves may be very variable between the genes. These facts altogether may yield nonstable variance estimates, which at their time induce high variability in F-like test statistics. To deal with this problem a commonly accepted strategy is variance shrinkage which consist of relying on improved variance estimates, S˜, where this improvement comes from borrowing information from all the genes in the array. The test statistics used by the SAM (Tusher *et al.*, 2001) or the limma methods (Gordon *et al.*, 2005) use different versions of variance shrinkage.

$$t = \frac{\overline{X}}{\hat{\sigma}_n} \approx \frac{\overline{X}}{\tilde{S}},$$

Where

$$\tilde{S}_{SAM} = c_0 + \hat{\sigma}_n$$
$$\tilde{S}_{limma} = \sqrt{\frac{d_0\hat{\sigma}_0^2 + d\hat{\sigma}_n^2}{d + d_0}}$$

where σˆn is the usual standard error estimate (with d degrees of freedom) for each gene (subindex omitted). In SAM c0, is estimated from the data using a permutation method. In limma d0 and s0 are unknown and are estimated from the data using an empirical bayes approach.

Approaches to deal with false positives

The analysis of microarrays on a gene-by-gene basis involves multiple testing. Testing thousands of genes is likely to produce hundreds of false positives if no correction is applied.

One approach is to control the family-wise error rate (FWER), which is the probability of accumulating one or more false positive errors over a number of statistical tests (Dudoit *et al.*, 2003). The simplest FWER procedure is the Bonferroni correction. FWER criteria may be too restrictive because control of false positives implies a considerable increase of false negatives. In practice, however, many biologists seem willing to accept that some errors will occur, as long as this allows findings to be made. For example, a researcher might consider acceptable a small proportion of errors (say 10%20%) between her findings. In this case, the researcher is expressing interest in controlling the false discovery rate (FDR), which is the proportion of false positives among all the genes initially identified as being differentially expressed. Unlike a significance level which is determined before looking at the data, FDR is a post data measure of confidence. It uses information available in the data to estimate the proportion of false positive results that have occurred. If one obtains a list of differentially expressed genes where the FDR is controlled at, say, the 20%, one will expect that a 20% of these genes will represent false positive results. This represents a less restrictive approach than controlling the FWER. The decision of controlling FDR or FWER depends on the goals of the experiment. If the objective is gene fishing allowing a certain number of false positives is reasonable and FDR is preferred. If instead one is working with a shorter list which one wishes to verify if some specific genes are expressed, then FWER is the appropriate criteria.

However one chooses to compute the significance values (p-values) of the genes, it is interesting to compare the size of the fold change to the statistical significance level. The volcano plot (Figure 29) arranges genes along dimensions of biological and statistical significance. The first (horizontal) dimension is the fold change between the two groups

(on a log scale, so that up and down regulation appear symmetric), and the second (vertical) axis represents the p-value from the moderated test on a negative log scale, so smaller p-values appear higher up. The first axis indicates biological impact of the change; the second indicates the statistical evidence, or reliability of the change. This allows the researcher to make judgements about the most promising candidates for follow-up studies, by trading off both these criteria by eye.

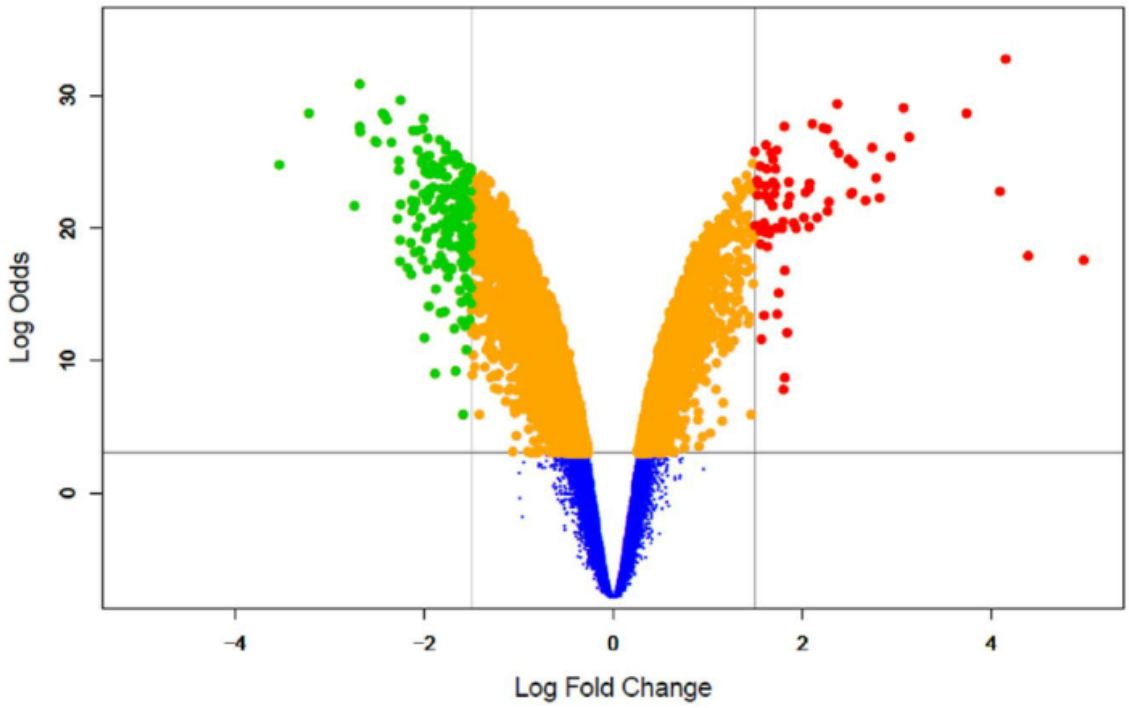


Figure 29. *Example of volvano plot for differential gene expression*

Pathway Analysis and Biological interpretation

A typical microarray experiment is one who looks for genes differentially expressed between two or more conditions. This mean genes which behave differently in one condition than in another. Such an experiment will result very often in long lists of genes which have been selected using some criteria to assign them statistical significance.

A common approach to biological interpretation is to reprocess the list trying to relate the genes it contains with one or more functional annotation databases such as the Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG) or others. There are many methods and models to do this (Khatri and Drăghici, 2005; Shedden, 2004). Two of the most used are Gene Enrichment Analysis and Gene Set Enrichment Analysis. Gene Enrichment Analysis (GE) aims at establishing if a given category, representing for example a biological process (GO) or a pathway (KEGG), appears more (enriched) or less (impoverished) often in the list of selected genes than in the (gene) population from where

they have been obtained, i.e., the array, the genome, or simply the genes which were selected for testing. The significance of this potential enrichment/impoverishment is established using a hypergeometric test. The Gene Set Enrichment Analysis (GSEA) method differs from the previous in that it requires, besides the list of genes, a numerical variable to rank them, usually the p-value of a test for differential expression. Starting from the ranked list a cumulative (enrichment) score based on the presence or absence of each gene in a selected category or `gene set is computed. A Kolmogorov Smirnov test is used to compare the distribution of the scores in the category with the empirical distribution of the numerical variable in the gene list in order to decide if the gene set is over represented at the top or bottom of the gene list. In both methods the tests are performed one category or one gene set at a time, followed by a multiple testing adjustment.

Material and Methods

RNA isolation from thrombus and peripheral blood venous

The RNA was extracted from thrombus and from peripheral venous blood using the PAXgene miRNA kit (Qiagen) according to the manual. Until the time of extraction, the thrombus were collected in RNA later (Figure 33A) and the peripheral blood venous stabilized in PAXGene Blood RNA tube (Figure 33B), both reagents that immediately stabilize the intracellular RNA to preserve the gene expression profile, and then stored at -80°C. This step is essential as RNA is an extremely delicate molecule subject to degradation.

.

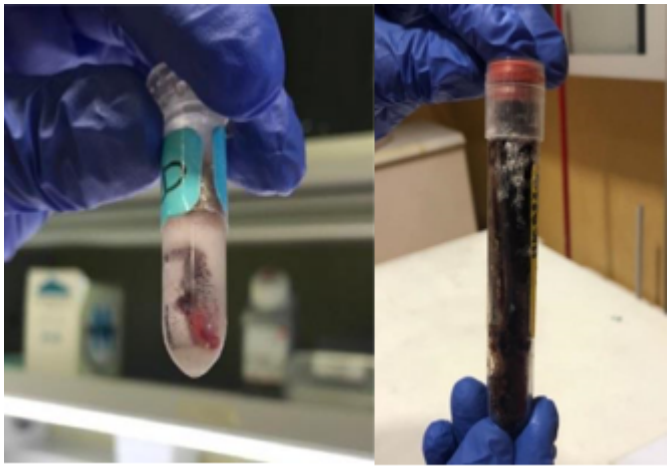


Figure 30: *Thrombi taken by thrombectomy and peripheral venous blood collected in PAXGene Tube.*

Quantitation and quality assessment of the total RNA

For quantitative and qualitative evaluation of RNA samples, we used NanoDrop One Spectrophotometer (see pg.61).

The presence of RNA degradation was assessed by evaluation of 28S and 18S band sharpness after denaturing electrophoresis electrophoretic run on a 1.2% agarose gel.

In order to evaluate the integrity of the extracted RNA, we also evaluated the RNA Integrity number (RIN) using Agilent 2000 Bioanalyzer.

To process samples, we used Agilent RNA 6000 Pico Kit in according to manufacturer. Agilent RNA kits contain chips and reagents designed for analysis of RNA fragments. Each RNA Pico chip contains an interconnected set of microchannels is used for separation of nucleic acid fragments based on their size as they are driven through it electrophoretically; RNA 6000 Pico kit is complementary to the RNA 6000 Nano kit and is suitable

for all applications where the amount of RNA (or cDNA) is limited, e.g. for biopsy samples, samples from microdissection experiments. The computation of the RIN is part of data analysis for total RNA samples.

For this study, the evaluation of RNA quality with the methods described above represented a crucial step for proceeding with the analysis of gene expression profile. RNA is a molecule that is easily subject to degradation. Furthermore, literature data described the difficulties in extracting RNA from a critical tissue such as the thrombus for obtaining an adequate RIN for subsequent analyses. (Popova *et al.*, 2008; Fraser *et al.*, 2019). Therefore, the first part of the study concerns the evaluation of the quality and the quantity of the material obtained from thrombectomy compared with the material obtained from peripheral venous blood. We performed a quantitative and qualitative analysis using Agilent Technology. The average RNA yield extracted from the thrombus in 40 samples was 8234 ng ± 1291, while the average RIN yield extracted from peripheral venous blood of the 37 samples was 2574 ng ±2156.

The average integrity of the RNA extracted from the thrombus of the 40 patients resulted in a RIN index of 5.5 ± 0.5, while the peripheral venous blood resulted in a RIN index of 9.05 ± 0.24. The average RIN of the thrombus extracts was lower ($p < 0.001$) than that of the RNA extracted from peripheral venous blood of the same patients, demonstrating the criticality of the treated material.

GeneChipTM Human Transcriptome Array 2.0

After the extraction, quantitative and qualitative analysis, 100 ng of the RNA were used to process the Affymetrix GeneChip technology and Affymetrix GeneChip® Human Transcriptome 2.0 Arrays (Figure 31) using the GeneChip® WT Plus Reagent Kit, according to manufacturer's instructions [GeneChip® WT PLUS Reagent Kit Manual Target Preparation for GeneChip® Whole Transcript (WT) Expression Arrays P/N 703174 Rev. 2, 2013].

The WT PLUS Reagent Kit enables you to prepare RNA samples for whole transcriptome expression analysis with GeneChipTM Whole Transcript (WT) Expression Arrays. The kit generates amplified and biotinylated sense-strand DNA targets from total RNA without the need for an up-front selection or enrichment step for mRNA. The kit is optimized for use with GeneChipTM Sense Target (ST) Arrays. The WT PLUS Reagent Kit uses a reverse transcription priming method that primes the entire length of RNA, including both poly(A) and non-poly(A) mRNA to provide complete and unbiased coverage of the transcriptome. The kit is comprised of reagents and a protocol for preparing hybridization-ready targets from 50 to 500 ng of total RNA. WT PLUS Reagent is optimized to work with a wide range of samples including tissues, cells, cell lines, and whole blood. The total RNA samples can be used directly without removal of ribosomal or globin RNA prior to target preparation with WT PLUS Reagent.

A supplied set of poly-A RNA controls is designed specifically to provide exogenous positive controls to monitor the entire target preparation. It should be added to the RNA prior to the First-Strand cDNA Synthesis step. Each eukaryotic GeneChipTM probe array contains probe sets for several B. subtilis genes that are absent in eukaryotic samples (lys, phe, thr, and dap). These poly-A RNA controls are in vitro synthesized, and the polyadenylated transcripts for the B. subtilis genes are premixed at staggered concentrations. The concentrated Poly-A Control Stock can be diluted with the Poly-A Control Dil Buffer and spiked directly into RNA samples to achieve the final concentrations. For 100 ng of total starting RNA three serial dilutions are made; 1:20 1:50 1:50 1:5o. 2 µl of the final dilution is then added to the RNA sample.

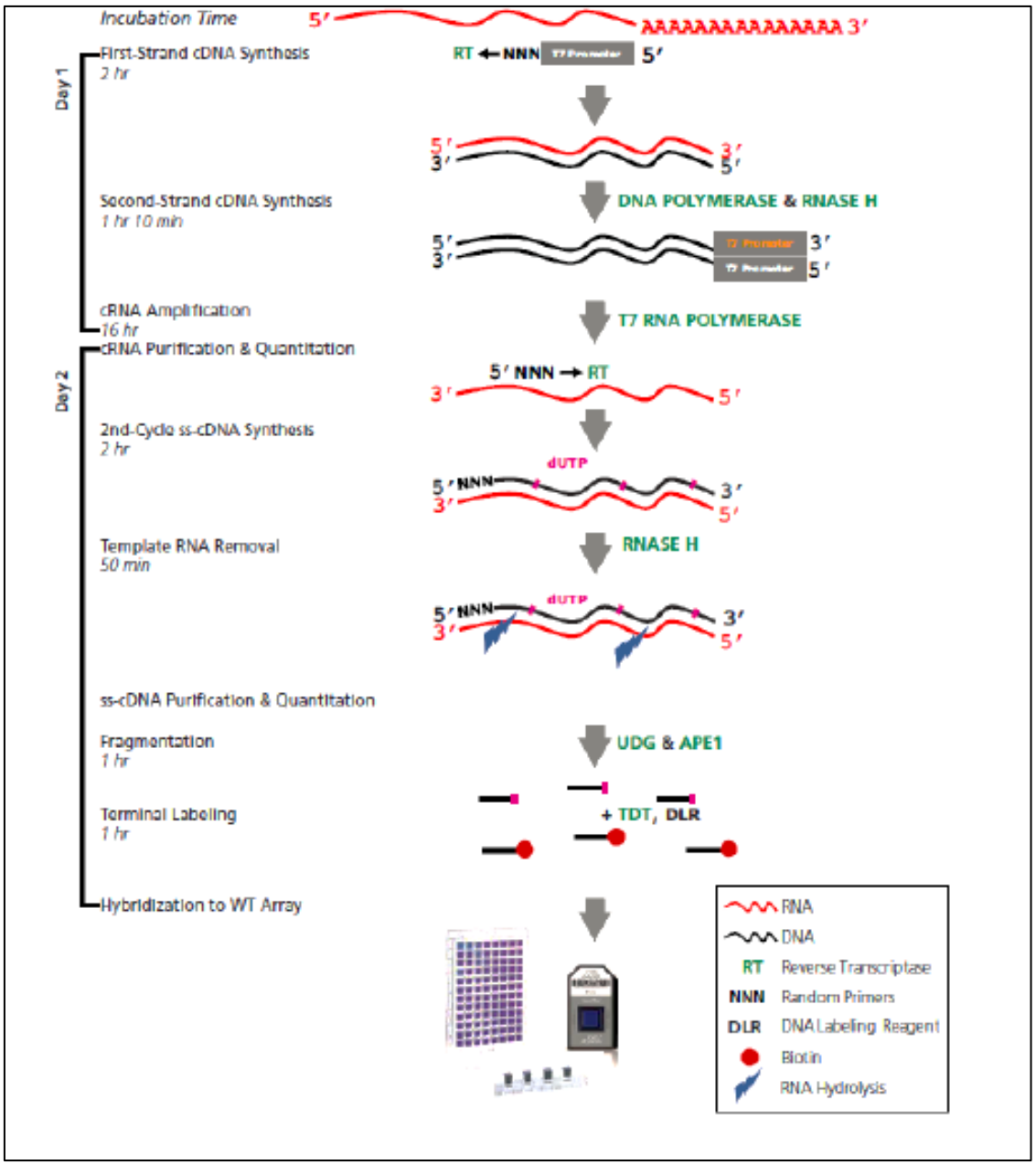


Figure 31: *the GeneChip® WT Plus workflow.*

After scanning we performed quality control for gene expression analyses using arrayQualityMetrics package from Bioconductor in R software environment. The package produces a comprehensive HTML report of quality metrics about a microarray dataset. The quality metrics can be used to assess the relative quality of different arrays within a dataset and to diagnose batch effects, and thus the quality of the overall dataset.

The data were normalized using R Oligo package (www.bioconductor.org) with the Robust Multiarray Averaging (RMA) algorithm. In order to identify differentially expressed genes, at statistically significant level, in RNA obtained by thrombi and peripheral blood samples, we applied a t-statistic variant approach. We used the significance analysis of microarrays (SAM) method (Tusher *et al.*, 2001), in which the t-statistic has a constant

value added to the standard deviation. We performed SAM analysis with Siggenes package (http://www.bioconductor.org) (Gentleman *et al.*, 2004). The minimization of the False Discovery Rate (FDR) allowed us to identify differentially expressed probe sets with a different delta. All the analysis was written in the freely available statistical language R. Subsequently, Gene Ontology (GO) enrichment analysis was performed in order to analyse the involvement of differentially expressed genes in different biological functional groups, all the genes present on the microarray were annotated for their role in biological processes. In our study we used one of the three ontologies produced by the GO consortium, the biological process ontology (Giusti *et al.*, 2009). The term ‘‘biological process’’ should be interpreted as a biological function to which the gene product contributes. Briefly, given a set of genes and one ontology, we first found the set of all unique GO terms within the ontology that were associated with one or more of the genes of interest. Next, we determined how many of the differentially expressed genes were annotated at each term and how many of the genes that were assessed (all the genes represented on the microarray) were annotated at the term. The test evaluated if there were more genes of interest at the term than would be expected by chance alone.

Results

Gene Expression Profiling Cohort

We evaluated the gene expression profiles of 40 RNA obtained from AIS patients' thrombi and 37 RNA from their venous peripheral blood for a total of 52 patients. Nineteen subjects had both thrombus and venous peripheral blood profiles. In table 6 demographic and clinical characteristics of 52 RNA profiled patients were reported.

Table 6. *Demographic and Clinical Characteristics of Gene Expression Profiling Cohort*

Features	Acute Ichaemic Stroke N=52
Age, years mean±SD	76.5±13.4
Sex, male N (%)	24 (46.0)
Hypertension, N (%)	36 (70.0)
Diabetes, N (%)	5 (9.0)
Dyslipidemia, N (%)	17 (33.0)
Current smoking, N (%)	38 (76.3)
Baseline NIHSS, median (IQR)	18.5 (15-24)
Heart failure, N (%)	8 (15.4)
Atrial Fibrillation N (%)	9 (37.0)
Systemic thrombolysis, N (%)	22 (41.6)

Picture of probe sets expressed in thrombi and venous peripheral blood

After data processing and application of the filtering criteria, the average of analysable probe sets numbered 440,085 in thrombi and 602,874 in venous peripheral blood samples. In thrombi and peripheral blood samples, among all probe sets, 378,476 and 515,048 had an associated identifier symbol, and 20,343 and 20,902 were unique symbols, respectively. Looking to their intersection, 20341 symbols were common to RNA from the different type of specimens,562 were unique symbols in venous peripheral blood, whereas the 3 symbols described below, were unique in thrombi:

-MTRNR2L5: MT-RNR2 Like 5 is a protein coding gene expressed in particular in brain and cortex, which plays a role as a neuroprotective and antiapoptotic factor. Diseases associated with MTRNR2L5 include asthenopia and facial paralysis.

-LINC01028: Long Intergenic Non-Protein Coding RNA 1028 is an RNA Gene, and it is affiliated with the lncRNA class.

-MIR124-3: Among numerous miRNAs, microRNA-124 (miR-124) is most abundantly expressed in the central nervous system (CNS). In the case it is downregulated it abolishes

neuronal differentiation, whereas the overexpression leads to acquisition of neuronal identity.

The Gene Ontology enrichment analysis of the 562 unique symbols present only in venous peripheral blood profiling resulted in the significant enrichment of 17 terms (Table 7).

Concerning the Gene Ontology enrichment analysis of the 20,341 symbols common to RNA from the two different type of specimens, the significant terms numbered 383. In table 8, the first 30 significant terms are reported. As showed among the first 30 ones, terms correlated with axonogenesis, regulation of synaptic signalling and transmission, regulation of neuron development are strongly represented together with biological processes implicated in regulation of cell adhesion and leukocyte differentiation.

Table 7. *Significant Biological Process Gene Ontology terms in gene enrichment analysis of 562 unique symbols present only in venous peripheral blood*

ID	Description	Gene Ratio	Bg Ratio	p-adjusted
GO:0050907	detection of chemical stimulus involved in sensory perception	103/315	477/18670	2.843732e-83
GO:0007608	sensory perception of smell	97/315	454/18670	6.999631e-78
GO:0035278	miRNA mediated inhibition of translation	15/315	93/18670	1.336044e-08
GO:0040033	negative regulation of translation, ncRNA-mediated	15/315	93/18670	1.336044e-08
GO:0045974	regulation of translation, ncRNA-mediated	15/315	93/18670	1.336044e-08
GO:0007195	adenylate cyclase-inhibiting dopamine receptor signaling pathway	5/315	11/18670	1.288743e-04
GO:0034249	negative regulation of cellular amide metabolic process	16/315	229/18670	3.774447e-04
GO:0048148	behavioral response to cocaine	5/315	20/18670	3.053379e-03
GO:0042742	defense response to bacterium	17/315	330/18670	8.173796e-03
GO:0043535	regulation of blood vessel endothelial cell migration	11/315	156/18670	9.905364e-03
GO:0001963	synaptic transmission, dopaminergic	5/315	31/18670	1.596603e-02
GO:0043534	blood vessel endothelial cell migration	11/315	180/18670	2.284555e-02
GO:0014046	dopamine secretion	5/315	40/18670	3.932584e-02
GO:0014059	regulation of dopamine secretion	5/315	40/18670	3.932584e-02
GO:0030336	negative regulation of cell migration	15/315	334/18670	4.121822e-02
GO:1904995	negative regulation of leukocyte adhesion to vascular endothelial cell	3/315	11/18670	4.576392e-02
GO:0060544	regulation of necroptotic process	4/315	26/18670	4.981147e-02

Table 8. *Thirty most significant Gene Ontology terms in the enrichment analysis of 20,341 common symbols in thrombi and venous peripheral blood.*

ego@result.ID	ego@result.Description	ego@result.GeneRatio	ego@result.BgRatio	ego@result.p.adjust
GO:0099177	regulation of trans-synaptic signaling	415/15251	437/18670	3.591789e-13
GO:0050804	modulation of chemical synaptic transmission	414/15251	436/18670	3.591789e-13
GO:0022604	regulation of cell morphogenesis	455/15251	484/18670	2.424564e-12
GO:0007409	Axonogenesis	439/15251	468/18670	1.828649e-11
GO:0023061	signal release	433/15251	462/18670	3.442934e-11
GO:0010975	regulation of neuron projection development	465/15251	499/18670	6.139191e-11
GO:0050808	synapse organization	384/15251	408/18670	1.087817e-10
GO:0001655	urogenital system development	314/15251	330/18670	1.977604e-10
GO:0060562	epithelial tube morphogenesis	306/15251	322/18670	5.169800e-10
GO:0006732	coenzyme metabolic process	378/15251	403/18670	5.169800e-10
GO:0016569	covalent chromatin modification	439/15251	474/18670	2.596074e-09
GO:0016570	histone modification	421/15251	454/18670	3.632954e-09
GO:0150063	visual system development	343/15251	366/18670	5.701277e-09
GO:0045785	positive regulation of cell adhesion	375/15251	403/18670	1.051581e-08
GO:0015711	organic anion transport	444/15251	482/18670	1.261490e-08
GO:0031346	positive regulation of cell projection organization	357/15251	383/18670	1.484910e-08
GO:1990778	protein localization to cell periphery	293/15251	311/18670	2.085965e-08
GO:0009914	hormone transport	302/15251	322/18670	4.729259e-08
GO:0051961	negative regulation of nervous system development	295/15251	315/18670	1.178331e-07
GO:0048545	response to steroid hormone	356/15251	385/18670	1.850173e-07
GO:0001101	response to acid chemical	319/15251	343/18670	2.047918e-07
GO:0032386	regulation of intracellular transport	389/15251	423/18670	2.084496e-07

GO:1902107	positive regulation of leukocyte differentiation	141/15251	144/18670	2.202414e-07
GO:0010038	response to metal ion	337/15251	364/18670	2.919965e-07
GO:0042391	regulation of membrane potential	398/15251	434/18670	2.956608e-07
GO:0009314	response to radiation	410/15251	448/18670	3.412854e-07
GO:0048871	multicellular organismal homeostasis	442/15251	485/18670	3.649068e-07
GO:0051260	protein homooligomerization	325/15251	351/18670	4.277203e-07
GO:0001503	Ossification	366/15251	398/18670	4.598993e-07
GO:0007265	Ras protein signal transduction	409/15251	448/18670	6.665269e-07

SAM analysis of gene expression in thrombi and peripheral blood according to gender and reperfusion therapy

the SAM analysis of gene expression in thrombi and peripheral blood was performed between males and females, and between patients treated with MT only or with the combination of systemic thrombolysis and MT. Concerning analysis according to gender, at FDR <0.01%, we observed 5 genes differentially expressed (Table 9) in thrombi: *XIST* (X-inactive specific transcript) an RNA gene on the X chromosome that acts as a major effector of the X inactivation process and 4 genes localized on chromosome Y (*DDX3Y. EIF1AY, UTY, KDM5D*). No significant differences in expression were observed in genes on autosomes.

Table 9. *Differentially expressed genes in thrombi of females with respect to males.*

Symbol	D value	$\log_2$ Fold Change
DDX3Y	-9.03	-4.29
EIF1AY	-8.09	-3.49
XIST	7.29	3.85
UTY	-6.80	-1.70
KDM5D	-4.70	-1.26

In venous peripheral blood, we observed 34 genes differentially expressed with a FDR <20%, and 14 out of 34 with FDR<1% in bold (Table 10). The large majority of genes were localized on sexual chromosomes except those marked with (°) in Table 10.

Table 10. *Differentially expressed genes in venous peripheral blood of females with respect to males.*

Symbol	D value	$\log_2$ Fold Change
XIST	33.65	6.62
UTY	-28.50	-6.49
KDM5D	-20.33	-4.25
USP9Y	-18.23	-4.22
EIF1AY	-17.09	-4.31
TXLNGY	-15.77	-2.33
PRKY	-15.59	-2.08
DDX3Y	-15.18	-2.38

TTTY15	-12.68	-2.07
RPS4Y2	-10.35	-2.33
TTTY10	-10.30	-2.36
TTTY14	-8.36	-1.26
ZFY	-8.13	-1.86
KDM5C	7.46	1.24
RPS4X	7.08	1.45
DDX3X	6.82	1.15
BCORP1	-5.85	-1.56
PCNT°	5.66	0.99
FAM230J°	-5.37	-2.36
ARHGEF39°	4.56	1.14
ADGRG2	-4.49	-0.74
FNBP1L°	4.48	1.08
CEACAM16°	4.40	1.07
AKAP6°	-4.34	-0.55
ASTN2°	4.24	0.83
TPCN1°	-4.20	-0.92
TMSB4Y	-4.19	-0.54
ANOS1	-4.13	-0.53
CCDC57°	-4.01	-0.53
CR2°	-3.99	-1.06
SRSF6°	-3.98	-0.41
CD79B°	-3.97	-0.45
PRKX	-3.96	-0.72
FCRL2°	-3.95	-0.64

Concerning analysis according to treatment, profiles of AIS patients treated with IVT with recombinant tissue plasminogen activator (rt-PA) + mechanical thrombectomy (MT) versus patients treated only with MT in thrombi and peripheral blood, no significant differences in gene expression were observed in genes on autosomes.

SAM analysis of gene expression in thrombi according to different clinical outcomes

By using the SAM method to assess the expression of the 440,085 called probe sets according to the 3-month modified Rankin Scale (mRS) (patients with score 3-6 versus score 0-2), we observed 5,713 probe sets (transcripts, miRNAs, lncRNAs) differentially expressed with a FDR<3.7% (Figure 32).

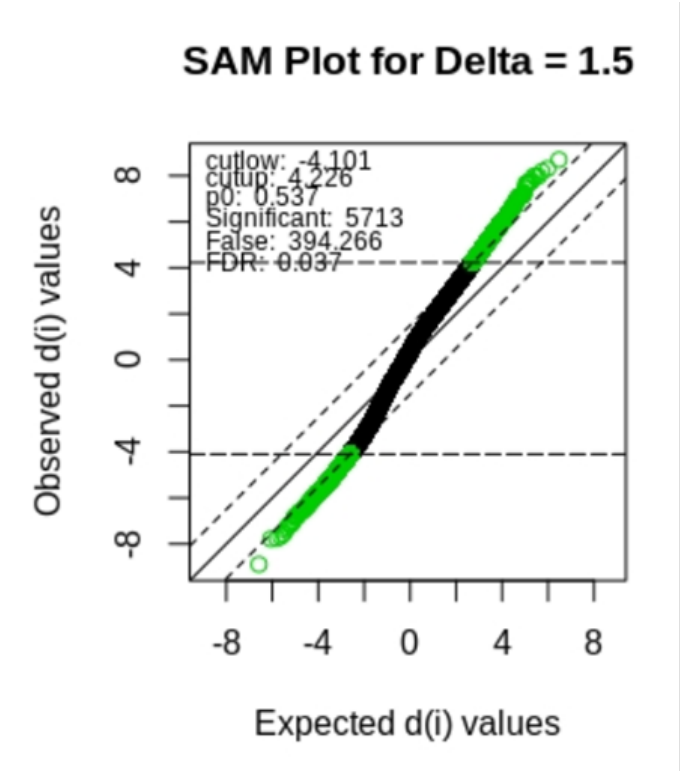


Figure 32. *SAM plot of probe sets according to 3-month modified Rankin Scale*

According to GO analysis adjusted by using the FDR multiple testing correction, we observed 221 significant biological processes (terms) associated with genes differentially expressed in patients with poor outcome evaluated by mRS. In Table 11, the first 40 most significantly GO terms are reported. For the first 20 terms, the Network Analysis representing their interactions is reported as NetPlot (Figure 33). Among significant terms, those associated with regulation of neutrophil mediated immunity and activation play a crucial role (Figure 33).

Table 11. *List of the 40 most significant enriched Biological Process Gene Ontology terms of known probe sets emerged from SAM analysis according to 3-month modified Rankin Scale (mRS)*

ID	Description	Gene Ratio	Bg Ratio	p-value adjusted
GO:0006401	RNA catabolic process	111/2648	397/18670	1.612741e-09
GO:0006402	mRNA catabolic process	104/2648	364/18670	1.612741e-09
GO:0051169	nuclear transport	98/2648	346/18670	7.786698e-09
GO:0006913	nucleocytoplasmic transport	97/2648	343/18670	7.786698e-09
GO:0016569	covalent chromatin modification	123/2648	474/18670	7.786698e-09
GO:0016570	histone modification	118/2648	454/18670	1.566696e-08
GO:0034976	response to endoplasmic reticulum stress	81/2648	285/18670	1.811697e-07
GO:0008380	RNA splicing	117/2648	469/18670	1.811697e-07
GO:0048193	Golgi vesicle transport	96/2648	368/18670	5.507478e-07
GO:0050657	nucleic acid transport	60/2648	193/18670	5.702299e-07
GO:0050658	RNA transport	60/2648	193/18670	5.702299e-07
GO:0010256	endomembrane system organization	109/2648	438/18670	5.702299e-07
GO:0035966	response to topologically incorrect protein	61/2648	199/18670	6.699227e-07
GO:0002446	neutrophil mediated immunity	120/2648	499/18670	6.929071e-07
GO:0051236	establishment of RNA localization	60/2648	196/18670	8.548218e-07
GO:0006611	protein export from nucleus	56/2648	179/18670	1.084500e-06
GO:0043312	neutrophil degranulation	116/2648	485/18670	1.294750e-06
GO:0042119	neutrophil activation	118/2648	498/18670	1.608381e-06
GO:0002283	neutrophil activation involved in immune response	116/2648	488/18670	1.714187e-06
GO:0016050	vesicle organization	85/2648	325/18670	1.719578e-06
GO:0031503	protein-containing complex localization	76/2648	281/18670	2.088766e-06

GO:0009896	positive regulation of catabolic process	103/2648	423/18670	2.676854e-06
GO:0033044	regulation of chromosome organization	87/2648	342/18670	4.115366e-06
GO:0071166	ribonucleoprotein complex localization	43/2648	128/18670	4.147391e-06
GO:0042176	regulation of protein catabolic process	94/2648	381/18670	5.322855e-06
GO:0035967	cellular response to topologically incorrect protein	50/2648	161/18670	5.462472e-06
GO:0019080	viral gene expression	56/2648	191/18670	7.511071e-06
GO:0022604	regulation of cell morphogenesis	112/2648	484/18670	9.647718e-06
GO:0002028	regulation of sodium ion transport	32/2648	85/18670	1.027720e-05
GO:0090150	establishment of protein localization to membrane	83/2648	332/18670	1.362316e-05
GO:0032386	regulation of intracellular transport	100/2648	423/18670	1.367034e-05
GO:0046777	protein autophosphorylation	64/2648	235/18670	1.412909e-05
GO:0043161	proteasome-mediated ubiquitin-dependent protein catabolic process	99/2648	419/18670	1.560352e-05
GO:0031331	positive regulation of cellular catabolic process	88/2648	361/18670	1.703241e-05
GO:0006413	translational initiation	55/2648	193/18670	1.964529e-05
GO:0019083	viral transcription	51/2648	177/18670	3.585522e-05
GO:0034504	protein localization to nucleus	68/2648	262/18670	3.585522e-05
GO:0030968	endoplasmic reticulum unfolded protein response	39/2648	121/18670	3.665806e-05
GO:1902305	regulation of sodium ion transmembrane transport	25/2648	62/18670	4.164529e-05
GO:0098876	vesicle-mediated transport to the plasma membrane	33/2648	96/18670	5.267398e-05

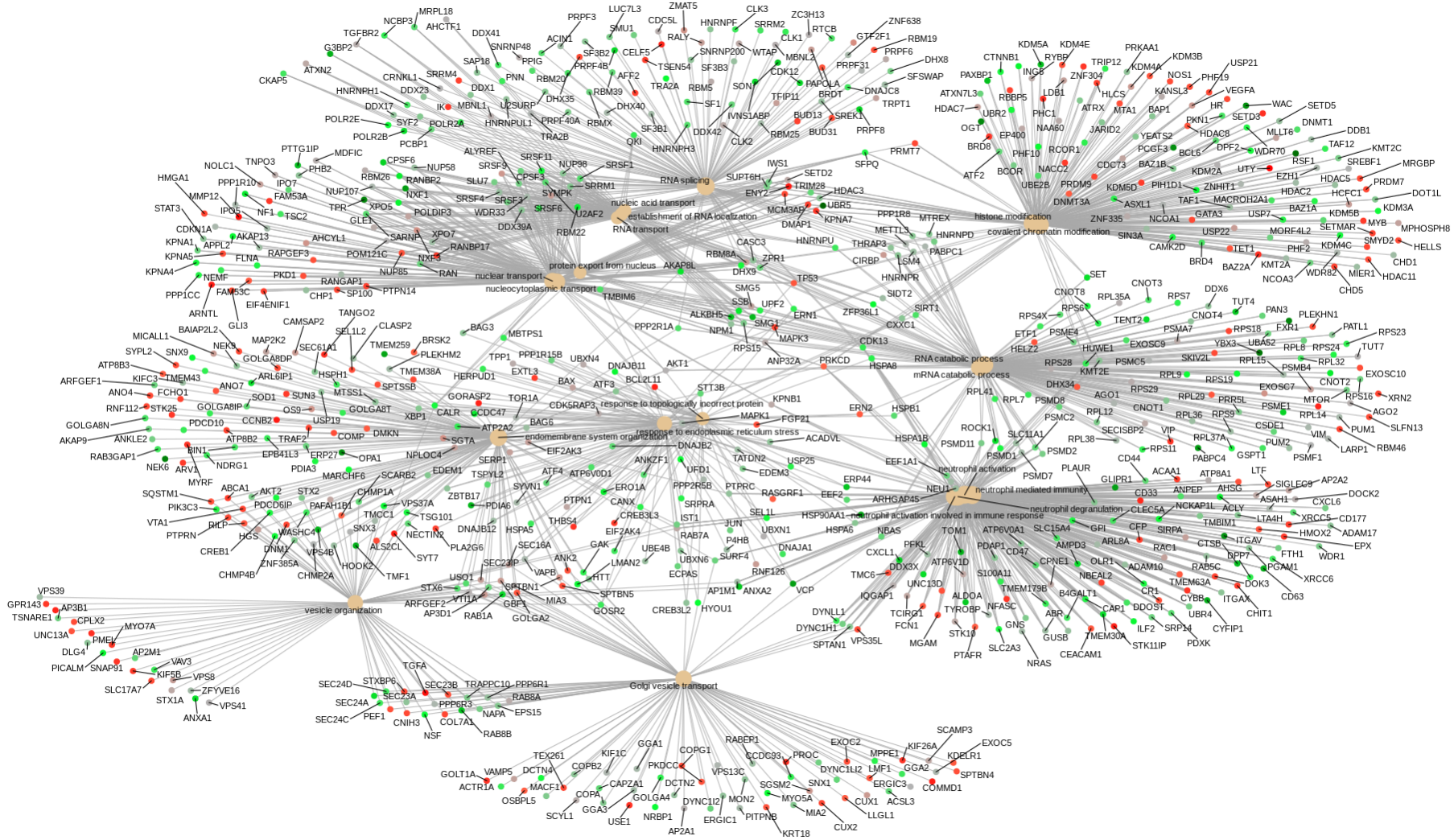


Figure 33: *NetPlot first 20 Biological Process Gene Ontology terms of probe sets differentially expressed according to mRS*

In thrombi, SAM analysis did not identify any significant difference according to the other hard clinical endpoints: primary-endpoint [relevant haemorrhagic transformation - defined as haemorrhagic infarction type two or any type of parenchymal haemorrhage (PH1 and PH2) - at 24 hours Computed Tomography; death; 24 hours edema; and any symptomatic intracranial haemorrhage (ICH).

SAM analysis of gene expression in venous peripheral blood according to different clinical outcomes

In venous peripheral blood, SAM analysis did not identify any significant difference according to the hard endpoints: mRS and any symptomatic intracranial hemorrhage (ICH). The SAM analysis identified:

- 2 probe sets differentially expressed according to primary-endpoint (Figure 34) at FDR 11%. The 2 probe sets represent 1 gene *RNF165* with a D value of 5.36 and a $\log_2$ fold change of 0.96. The gene *RNF165 (*Arkadia-like; Arkadia2; Ark2C) is expressed specifically in the nervous system;
- 298 probe sets according to 24 hours edema (Figure 35) at FDR <24%. In Table 12 , the significantly Biological Process GO terms are reported. The Network Analysis representing the interactions among significant GO terms is reported as NetPlot (Figure 36). Among significant terms, those associated with regulation and activation of transcriptomes of cells in general play a crucial role (Table 12, Figure 36).
- 1 probe set differentially expressed according to death (Figure 36) at FDR 5%. The probe set represent the gene *AADACL3* with a D value of -6.92 and a $\log_2$ fold change of -0.82. Poor information is available for the gene *AADACL3 (*arylacetamide deacetylase like 3) coding a lipolytic enzyme.

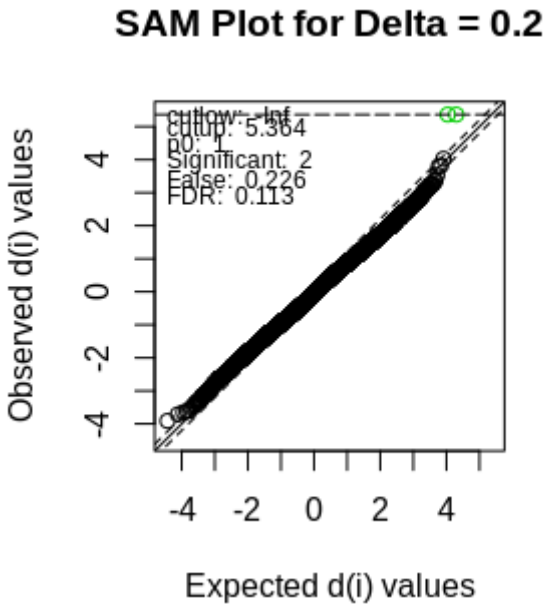


Figure 34. *SAM plot of probe sets according to primary-endpoint in peripheral blood*

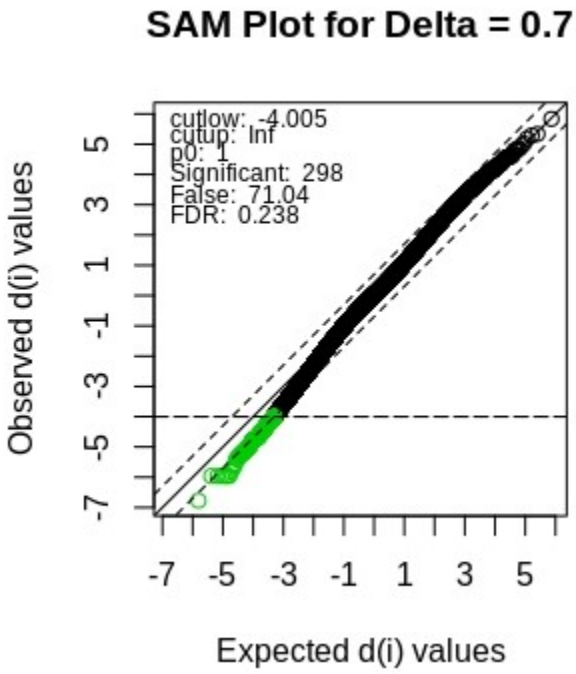


Figure 35. *SAM plot of probe sets according to 24 hours edema in peripheral blood*

Table 12. *List of significant enriched Biological Process Gene Ontology terms of known probe sets emerged from SAM analysis according to 24 hours edema*

ID	Description	Gene Ratio	Bg Ratio	p-value adjusted
GO:0042254	ribosome biogenesis	19/192	297/18670	5.800355e-07
GO:0006403	RNA localization	13/192	230/18670	5.973933e-04
GO:0022618	ribonucleoprotein complex assembly	14/192	277/18670	6.386989e-04
GO:0071826	ribonucleoprotein complex subunit organization	14/192	291/18670	7.707898e-04
GO:0032200	telomere organization	11/192	175/18670	7.707898e-04
GO:0034470	ncRNA processing	16/192	384/18670	7.831351e-04
GO:0000723	telomere maintenance	10/192	162/18670	1.596011e-03
GO:0000054	ribosomal subunit export from nucleus	4/192	13/18670	1.596011e-03
GO:0033750	ribosome localization	4/192	13/18670	1.596011e-03
GO:0000956	nuclear-transcribed mRNA catabolic process	11/192	207/18670	1.785134e-03
GO:0051973	positive regulation of telomerase activity	5/192	36/18670	3.852806e-03
GO:0015931	nucleobase-containing compound transport	11/192	241/18670	4.523707e-03
GO:0006575	cellular modified amino acid metabolic process	10/192	202/18670	4.907958e-03
GO:0032069	regulation of nuclease activity	4/192	22/18670	5.807071e-03
GO:0000075	cell cycle checkpoint	10/192	216/18670	6.545985e-03
GO:0006611	protein export from nucleus	9/192	179/18670	7.672669e-03
GO:0006278	RNA-dependent DNA biosynthetic process	6/192	77/18670	9.801583e-03
GO:0006413	translational initiation	9/192	193/18670	1.104581e-02
GO:0060249	anatomical structure homeostasis	14/192	437/18670	1.104581e-02
GO:0071166	ribonucleoprotein complex localization	7/192	128/18670	1.836405e-02
GO:0032212	positive regulation of telomere maintenance via telomerase	4/192	34/18670	1.962368e-02
GO:0009119	ribonucleoside metabolic process	6/192	98/18670	2.403717e-02

GO:0061684	chaperone-mediated autophagy	3/192	16/18670	2.413464e-02
GO:0051298	centrosome duplication	5/192	67/18670	2.663384e-02
GO:1901657	glycosyl compound metabolic process	7/192	147/18670	3.437259e-02
GO:0051131	chaperone-mediated protein complex assembly	3/192	19/18670	3.669931e-02
GO:1900034	regulation of cellular response to heat	5/192	79/18670	4.982673e-02

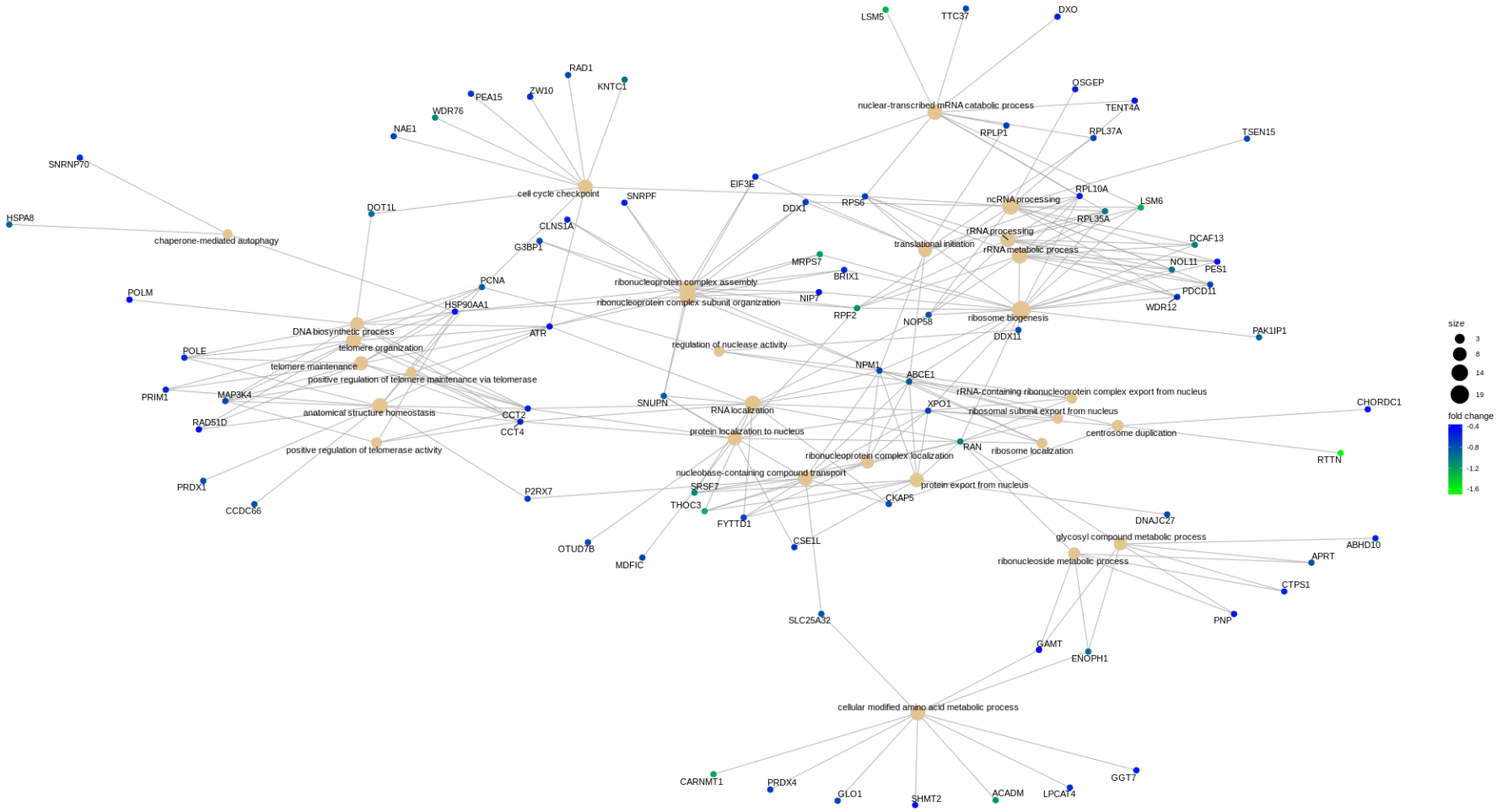


Figure 36. *NetPlot of Biological Process Gene Ontology terms of probe sets differentially expressed according to 24h edema in peripheral* blood

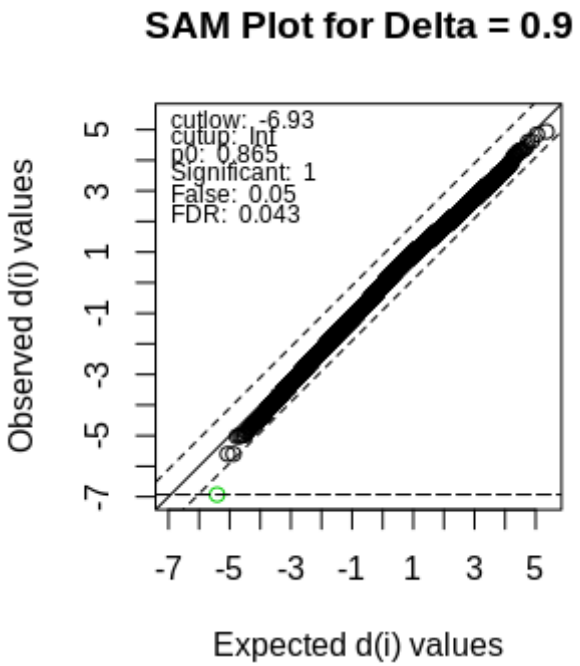


***Figure 37**:. SAM plot of probe sets according to death in peripheral blood.*

Discussion

Mechanical thrombectomy offers an interesting window to study changes occurring during ischemia. This opportunity has fostered our group as well as others to focus attention on the evaluation of cerebral thrombi characteristics in order to reach a better knowledge on pathophysiological mechanisms, novel determinants of poor outcomes and further potential targets of intervention.

Large part of current available data has focused on histopathology of the thrombus variety (Marder *et al.*, 2006; Niesten *et al.*, 2014; Boeckh-Behrens et a., 2016), but they did not evaluate other aspects of thrombus, in particular global gene expression profiling, or the molecular milieu of the thrombus.

Nevertheless, these works represent an important source of information for our research, in particular, and for knowledge on stroke pathophysiological mechanisms, in general. In fact, Niesten and coworkers (2014) evaluated red blood cell, platelet, and fibrin compositions of 22 subjects after thrombectomy, assessing relationships with stroke subtype. Simons and coworkers (2015) evaluated correlations between hyperdense artery sign and thrombus composition finding that the radiographic sign was associated with early phase clot pathology in 40 patients (Simons *et al.*, 2015). Dargazanli and coworkers (2016)

evaluated the presence of CD3+ T cells in the thrombi removed during thrombectomy, in relationship to stroke subtype (Dargazanli *et al.*, 2016).

In our study, we decided to integrate data on clinical, imaging and biochemical circulating parameters including RNA extracted from venous peripheral blood with those obtained from gene expression profiling of RNA obtained from intracranial thrombus collected during endovascular procedure. At the best of our knowledge, only one study is applying a similar approach in human. Fraser and coworkers (2019) developed a tissue banking protocol using mechanical thrombectomy to capture thrombus along with arterial blood proximal and distal to it. They reported that their protocol provides adequate amounts of tissue from distal blood, proximal blood, and thrombi for gene expression and proteomics analyses (Fraser *et al.*, 2019).

The first difficulty we had to confront with was RNA quantity and quality as concerns RNA extracted from thrombi. In fact, whereas mean RIN values for peripheral blood indicated high-quality RNA with minimal degradation, those of the thrombi were relatively low indicating a low integrity of RNA. This result was consistent with small amount of RNA in the thrombus and the known presence of RNases due to necrotic and apoptotic processes. Nevertheless, unlike Fraser and coworkers (2019), the quantity and quality of RNA from thrombi allowed us to evaluate gene expression profiles.

Interestingly, even if the number of probe sets expressed in thrombi were significantly lower than those identified in venous peripheral blood (440,085 vs 602,874), the number of unique symbols analyzable in samples from the two different specimens were comparable (20,343 vs 20,902). This issue could be ascribable, at least in part, to the low quality of RNA from thrombi, even if attenuated by the probes design of the HTA 2.0 GeneChip (Affymetrix).

Looking to the intersection of data from the different specimens, 20,341 symbols were common. As evidenced by looking to the first 30 Biological Process Gene Ontology terms resulted by the analysis of common features, the terms correlated with axonogenesis, regulation of synaptic signaling and transmission, regulation of neuron development are strongly represented together with biological processes implicated in regulation of cell adhesion and leukocyte differentiation and functions. These data suggest that in venous peripheral blood are present transcripts, miRNA and lncRNA implicated in biological processes that contribute to the response to damage, and to restore full functionality of damaged brain structures.

Concerning symbols peculiar of thrombi, they are 3 (1 lncRNA and 1 protein coding gene and 1 microRNAs):

-Long Intergenic Non-Protein Coding RNA 1028 (LINC01028) is an RNA Gene localized on chromosome 17, and it is affiliated with the lncRNA class. Expression in normal human tissues from GTEx of LINC01028 gene was found in brain, cortex, cerebellum, and whole blood but not in white blood cells;

-MT-RNR2 Like 5 (MTRNR2L5) is a protein coding gene expressed in particular in brain and cortex, which plays a role as a neuroprotective and antiapoptotic factor. Diseases associated with MTRNR2L5 include asthenopia and facial paralysis;

-MicroRNA-124 (MIR124-3) is most abundantly expressed in the central nervous system (CNS). In the case it is downregulated it abolishes neuronal differentiation, whereas the overexpression leads to acquisition of neuronal identity (Sun *et al.*, 2015; Yang *et al.*, 2017). MicroRNAs (miRNAs), a family of non-coding RNAs of 20–25 nt, play pivotal roles during this remodeling process by regulating target genes at post-transcriptional level (Liu *et al.*, 2013). Several studies have demonstrated significant alterations in the cerebral "miRNA-ome" following ischemia (Dharap *et al.*, 2009; Jeyaseelan *et al.*, 2008). These reports suggest that miRNAs may act as innovative gene therapeutic candidates contributing to neurogenesis, angiogenesis, and neural plasticity (Sun *et al.*, 2015; Yang *et al.*, 2017). Among numerous miRNAs, microRNA-124 (miR-124) is most abundantly expressed in the central nervous system (CNS) (Mishima *et al.*, 2007). It has been reported that attenuation of endogenous miR-124 in neural progenitor cells in subventricular zone can abolish neuronal differentiation, whereas overexpression leads to acquisition of neuronal identity (Åkerblom *et al.*, 2012; Cheng *et al.*, 2009). After ischemia, the expression of miR-124 was increased in ischemic penumbra. Exogenous miR-124, by using agomir or liposomated mimic, could reduce infarct area (Sun *et al.*, 2013; Hamzei Taj *et al.*, 2016). Recently, a study showed that viral vector expressing miR-124 yielded increased angioneurogenesis and significant neuroprotection against stroke (Doeppner *et al.*, 2013). Studies above revealed the neuroprotective and neurorestorative potential of miR-124 (Yang *et al.*, 2017). A recent study provides strong evidence that serum brain-derived neurotrophic factor (BDNF) and the BDNF-regulatory miR-124 may serve as molecular markers for AIS (Wang *et al.*, 2019). The BDNF transcript and BDNFAS (BDNF antisense RNA) transcripts were expressed both in thrombi and peripheral blood in our study. BDNF is a pro-survival factor induced by cortical neurons that is necessary

for survival of striatal neurons in the brain (Zuccato *et al.*, 2001). Liu *et al.* (2005) identified 7 BDNFAS splice variants generated from the BDNF (113505) antisense strand. These transcripts were predicted to be noncoding. RT-PCR showed that the majority of BDNFOS transcripts were expressed in different regions of the human brain, with lower expression in peripheral tissues (Liu *et al.*, 2005). Wang and coworkers (2019) observed that the BDNF level of AIS patients is very low compared with that of controls and, in contrast, a very high serum level of miR124 in AIS patients relative to healthy individuals. The Authors revealed a negative correlation between NIHSS score and BDNF level, whereas a positive correlation was observed between NIHSS score and miR-124. In addition, the relationship between serum BDNF and miR-124 was negative in AIS patients. They hypothesized that serum BDNF and the BDNF-regulatory miR-124 may serve as molecular markers for (Wang *et al.*, 2019). Our data might contribute to furtherly clarify the source of transcripts and therefore the possibility to use this mechanism not only as marker of disease but as target of intervention.

Concerning the 562 symbols peculiar of venous peripheral blood, resulted in the significant enrichment of 17 biological process terms.

Interestingly, the 562 peculiar transcripts among 17 biological processes are particularly involved in:

- regulation of blood vessel endothelial cell migration and negative regulation of leukocyte adhesion to vascular endothelial cell;
- miRNA mediated inhibition of translation and regulation of translation ncRNA-mediated;
- regulation of dopamine secretion and dopaminergic synaptic transmission;
- defence response to bacterium;
- regulation of necroptotic process.

These data strongly suggest that transcripts peculiarly found expressed in peripheral blood are crucial to regulate mechanisms that sustain the development and response to ischemia damage.

Surprisingly, a large part of the 562 transcripts are involved in detection of chemical stimulus involved in sensory perception and sensory perception of smell. On the other hand, previous data demonstrated an important contribution of blood vessel to modulation of smell (Lucero, 2013; Ferrer *et al.*, 2016). This issue will certainly deserve further attention

to investigate the regulation mechanisms of the sensory perception of smell and their further consequences.

Stroke affects both men and women; however, the incidence rates and outcomes differ between the 2 genders. Age-specific stroke rates are higher in men, but women experience more frequent stroke events because of their long-life expectancy, and high stroke incidence at older ages (Hiraga, 2017). After stroke, women have poorer functional outcomes and lower quality of life than men. In our study, we tested whether genetic expression profiles in AIS patients according to gender allowed us to revealed differences.

Concerning analysis according to gender, we observed 5 genes differentially expressed in thrombi: *XIST* (X-inactive specific transcript) an RNA gene on the X chromosome that acts as a major effector of the X inactivation process and 4 genes localized on chromosome Y (*DDX3Y. EIF1AY, UTY, KDM5D*). No significant differences in expression of genes on autosomes were observed.

Similarly, in peripheral blood we found several genes involved in gender differences, but we found also transcripts with differential expression in females respect to males localized on autosomes. Few data exist on these genes and stroke. A previous work suggests that ischemic brain tissue-targeted and selective inhibition of alternative complement pathway provide self-limiting inhibition of complement activation and reduces acute injury while maintaining complement-dependent recovery mechanisms into the subacute phase after stroke (Alawieh *et al.*, 2015). Therefore, the complement receptor 2 (CR2) reduced expression in females observed in our cohort of AIS patients suggests the need to deepen the topic with prospective data in the cohort.

Nevertheless, extending the study to a greater number of subjects in the cohort, we should work to evaluate whether the differential expression observed in females with respect to males are or not specifically associated to peculiar aspect of the AIS and its outcomes according to gender.

Due to the possible differential treatment combination in our AIS patients, we would test whether there were differential profiles according to treatment. Profiles of AIS patients treated with intravenous thrombolysis with recombinant tissue plasminogen activator (rt-PA) + mechanical thrombectomy (MT) versus patients treated only with MT in thrombi and peripheral blood did not show statistically significant differences in gene expression. This datum is in some way surprising, but as observed in similar situations, we should take in mind that the analysis is performed in an acute phase in which there is a strong

stress of the system that might overcome the possibility to evaluate possible difference induced by different treatments.

Looking to profiles obtained in thrombi, SAM analysis showed 5,713 probe sets (transcripts, miRNAs, lncRNAs) differentially expressed according to the 3-month modified Rankin Scale (mRS) (patients with score 3-6 versus score 0-2). According to GO analysis adjusted by using the FDR multiple testing correction, we observed 221 significant biological processes (terms) associated with genes differentially expressed in patients with poor outcome evaluated by mRS. Among the first most significantly GO terms, those associated with regulation of neutrophil mediated immunity and activation play a crucial role.

After ischemic stroke, the integrity of the blood-brain barrier is compromised. Peripheral immune cells, including neutrophils, T cells, B cells, dendritic cells, and macrophages, infiltrate into the ischemic brain tissue and play an important role in regulating the progression of ischemic brain injury (Jian *et al.*, 2019).

Microglia in central nervous system (CNS) and peripheral immune cells, including blood-derived monocytes/macrophages, neutrophils, and lymphocytes, are recruited into the ischemic cerebral hemisphere which induce inflammatory response (Anrather and Iadecola, 2016; Rayasam *et al.*, 2018). The inflammatory cascade in brain tissue could accelerate, expand or delay, alleviate ischemic brain injury (Serbina *et al.*, 2008).

Neutrophils are the first leukocytes subset to appear in the ischemic brain, which are detected within the first hour. These neutrophils remain in the cerebral microvessels, from where they damage the BBB by releasing reactive oxidative species (ROS) and proteolytic enzymes (Malone *et al.*, 2019). Neutrophils penetrate the CNS parenchyma mainly following the more damaging second opening of the BBB, resulting in severe endothelial damage, destruction of adjacent blood vessels, and in some cases hemorrhagic transformation (Perez-de-Puig *et al.*, 2015).

Ischemic stroke starts in the blood vessels, where arterial occlusion results in hypoxia, ROS production, and coagulation cascade. In addition, ischemia also impacts the brain parenchyma. Hypoperfusion causes deprivation of glucose and oxygen, leading to a series of interconnected events (acidosis, oxidative stress, excitotoxicity, and inflammation), eventually causing neuronal cell death. The dying and dead neurons release danger-associated molecular patterns (DAMPs) which result in the activation of microglia. The release of chemokines and cytokines (TNF-α, IL-1β, IL-6) from microglia generates an

inflammatory environment featuring activated leukocytes, and the increased expression of adhesion molecules on endothelial cell. Neutrophils enter the brain as early as 1 h after stroke and increase blood-brain barrier permeability by secreting matrix metalloproteinases (MMPs), further aggravating ischemic injury. T cells have a damaging effect in this acute phase of stroke. Th17 cells and γδT cell further increase neutrophilic activity and aggravate the acute ischemic through the production of IL-17. B cells produce antibodies against brain-derived molecules, resulting in further neuronal damage in 4–7 weeks following stroke onset, possibly leading to clinical stroke sequelae such as dementia (Jian *et al.*, 2019).

On the other hand, these data are supported by the observation of the association with different outcomes in our cohort of metalloproteinases and chemokines/cytokines produced by neutrophils (see comment to data in chapter 5).

Our data confirming and enlarging the role of neutrophils mediated immunity and activation as determinants/markers of poor outcome and disability in our cohort open significant further work to improve knowledge on suggested mechanisms by the present work.

In thrombi, SAM analysis did not identify any significant difference according to the other hard endpoints: primary-endpoint, 24 hours edema and any ICH.

In profiles obtained in peripheral blood, SAM analysis did not identify any significant difference according to the hard endpoints: mRS and any symptomatic intracranial haemorrhage (ICH). Whereas SAM analysis identified:

-the increased expression of gene *RNF165* in patients positive for the primary-endpoint. The gene *RNF165 (*Arkadia-like; Arkadia2; Ark2C) is expressed specifically in the nervous system. Its loss in mice causes motor innervation defects that originate during development and lead to wasting and death before weaning;

- 298 probe sets according to 24 hours edema. Among significant enriched biological process gene ontology terms, those associated with regulation and activation of transcriptomes of cells in general play a crucial role;

- the reduced expression in dead AIS patients of the gene AADACL3 (arylacetamide deacetylase like 3) coding a lipolytic enzyme for which poor information is available.

Preliminary data on validation by real time PCR (miR-124, RNF165, AADACL3) confirmed the results obtained by Affymetrix approach (data not shown). Nevertheless, definite information will be available as soon as we will confirm the data on these and further transcripts on the whole cohort.

In conclusion, MT has become a standard treatment for emergent large vessel occlusion stroke and multiple clinical trials demonstrated the benefit of endovascular recanalization of the occluded vessel in the anterior circulation. However, newer trials continue to expand the inclusion criteria of this treatment (Goyal *et al.*, 2016; Berkhemer *et al.*, 2015; Saver *et al.*, 2015; Campbell *et al.*, 2015; Jovin *et al.*, 2015; Nogueira *et al.*, 2018; Albers *et al.*, 2018). Therefore, results of the study obtained on cerebral thrombi may be influenced by the clinical setting to which the mechanical thrombectomy procedure is reserved, and not completely extendable to the entire population of stroke patients. Nevertheless, using a definite protocol, a prospective collection of data, and an adequate number of patients assuring statistically powered data, Reperfusion Injury after ischemic Stroke Study (RISKS) can integrate substantially scanty clinical information about biological factors involved in reperfusion injury after cerebral ischaemia. Moreover, RISKS Study combines advanced neuroimaging techniques for the study of blood–brain barrier disruption with analyses of circulating biomarkers as potential predictors of reperfusion injury after acute phase interventions.

A limitation of the study is the lack of a control group of patients who had a stroke not treated with revascularization therapies. Another limitation is that study will include patients with acute ischaemic stroke treated with intravenous thrombolysis, endovascular treatment or both. This heterogeneity might influence levels of circulating biomarkers at 24 hours. A further limitation is the lack of standardization for the assessment of recanalization.

Identification through a reverse-genetic approach of genetic variants in *DPP3* gene associated to phenotype linked to atherothrombotic diseases and their functional characterisation.

Introduction

Dipeptidyl Peptidase 3 (DPP3, EC 3.4.14.4) has been associated with several pathophysiological process, including blood pressure regulation, pain signalling, and cancer cell defence against oxidative stress. It is emerging as biomarker of clinical outcome in patients suffering of acute and chronic cardiovascular diseases like acute heart failure (Boorsma *et al.*, 2021), cardiogenic shock (Iborra-Egea *et al.*, 2020) and aneurismal subarachnoid haemorrage (Neumaier *et al.*, 2021).

DPP3 is a zinc-dependent aminopeptidase. It is the sole member of the M49 family of metallopeptidases. The protein was purified for the first time from bovine pituitary (Ellis and Nuenke, 1967) and named so being the third dipeptidyl peptidase to be identified. Subsequently, it has been purified from many other tissues and organisms ranging from lower to higher eukaryotes. By virtue of its affinity for different types of bioactive peptides and synthetic substrates, it was named enkephalinase B (Lee and Snyder, 1982), red cell angiotensinase (Abramic *et al.*, 1988), dipeptidyl aminopeptidase III (Swanson *et al.*, 1978) and dipeptidyl arylamidase III (Ellis and Nuenke, 1967)).

Metallopeptidases historically were regarded to play mundane roles in processes such as the terminal stages of protein turnover and digestion. However, studies over the last few decades insinuated their role in oligopeptide N-terminal processing and degradation of bioactive peptides (Vazeux *et al.*, 1996), cell cycle regulation (Constan *et al.*, 1995), protein maturation (Selvakumar *et al.*, 2009), viral infection (Soderberg *et al.*, 1993), hormone processing (Pham *et al.*, 2007) and defence from oxidative stress (Liu *et al.*, 2007) changing the old notion. Today they are known to be implicated in virtually every aspect of life, creating a great deal of medical as well as physiological interest.

After identification and initial kinetic characterization of DPP3 ((Ellis and Nuenke, 1967)), further progress remained imperceptible for a long time (Smyth and O'Cuinn, 1994) probably due to the unavailability of its structural information, specific inhibitor and its ability to cleave tetrapeptides to decapeptides irrespective of their amino acid composition and sequence. Its cytosolic localization, and hence a suspected role in terminal

stages of protein turnover (Abramic *et al.*, 1988), a process underestimated earlier, probably further slowed the pace. In the meantime, however, it was implicated in pathophysiological processes such as inflammation (Hashimoto *et al.*, 2000), pain modulation (Sato *et al.*, 2003) and blood pressure regulation (Vanderberg *et al.*, 1985). A strong correlation between elevation of DPP3 activity and histological aggressiveness of human ovarian carcinoma (Simaga *et al.*, 1998) revitalized the field and since then several studies have been undertaken to identify its physiological substrate(s) using synthetic peptides of biological significance.

Structure of DPP3

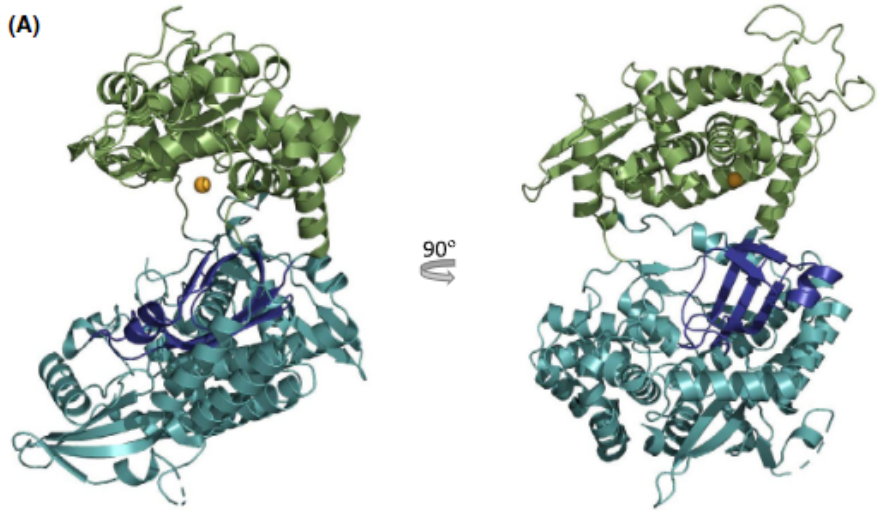


Figure 38*: DPP3 structure. (from Malovan et a., 2022)*

Recently, crystal structures of DPP3 of yeast (Baral *et al.*, 2008) and human (Figure 38) origin (PDB: 3fvy (Dobrovetsky *et al.*, 2009) have been resolved and a selective inhibitor (Yamamoto *et al.*, 2000) has been designed. These findings elucidated substrate(s) characteristics and the mechanism of their hydrolysis which may eventually help in identification of the physiological substrates and hence its plausible role in cellular physiology. Human DPP3 (hDPP3) consists of 728 amino acid residues, the molecule presents two domains, an upper domain rich in a-helices and a lower domain with mixed a-helices and b-sheets. The two domains are separated by a wide cleft which has been proposed to be the substrate binding site (Baral *et al.*, 2008). The two domains are interconnected by a helical loop extending from the lower domain. The catalytic motif (HEXXGH) and the secondary motif (EECRAR／D) are part of the upper domain and are located on a-16 and a-18 helices. A zinc ion, which is critical for the catalytic activity of DPP3, has been

observed in the conserved binding site located in the upper domain. The residues involved in the coordination bond formation with zinc are Glu517, His465 and His460 in yDPP3. A conserved water molecule is observed coordinating the zinc ion. Side chains of Glu461 (in yDPP3) and Glu451 (in hDPP3), essential for hydrolysis of the peptide bond, are in turn hydrogen bonded to the conserved water molecule.

The surface area was calculated to be 27 643.881A˚2.

hDPP3 gene (*DPP3*, NC_000011.9) containing 18 exons and 17 introns is located on chromosome 11q12-13.1 (Fukasawa *et al.*, 2000).

Expression of DPP3 at different stages of fetal development taken together with its ubiquitous distribution suggests a housekeeping role of this peptidase in cellular physiology. However, a marked increase in its activity with sexual maturity (Vanha-Perttula, 1988), histological aggressiveness of the tumour (Simaga *et al.*, 2003), and an increase (11.6-fold) in human retroplacental serum compared with control serum (Shimamori *et al.*, 1986) have been observed. These observations advocate the amenability of hDPP3 expression to regulation.

Localization of DPP3

DPP3 is mainly described as a soluble, cytosolic protein in mammals (Swanson *et al.*, 1978; Ohkubo *et al.*, 1999; Smyth and O'Cuinn, 1994). Proteome analysis of mouse pluripotent stem cells additionally assigned DPP3 as a cytosolic protein with high confidence (Christoforou *et al.*, 2016). However, membrane-associated activity has been described in Pediococcus acidlactici, Drosophila melanogaster (Mazzocco *et al.*, 2001), calf brain (van Amsterdam *et al.*, 1983) and several rat tissues (Lee and Snyder, 1982). Moreover, the translocation of DPP3 into the nucleus was reported under the conditions of oxidative stress (Sobocanec *et al.*, 2016). DPP3 has also been found extracellularly in the cerebrospinal fluid of sheep and human (Sato *et al.*, 2003; Cruz-Diaz *et al.*, 2016), human seminal plasma (Vanha-Perttula, 1988) and retroplacental serum (Shimamori *et al.*, 1986). The presence and activity of DPP3 in human plasma of healthy subjects has been recently reported (Rehfeld *et al.*, 2019). Deniau et al (Deniau *et al.*, 2019) and Takagi *et al.* (Takagi *et al.*, 2019) also described elevated circulating DPP3 (cDPP3) activity in patients with cardiogenic shock. In addition, it was found that elevated cDPP3 levels are associated with a poor outcome in patients suffering from sepsis (Blet *et al.*, 2021; van Lier *et al.*, 2020). Recent reports have further highlighted the role of cDPP3 in cardio-renal disease

progression, which will be discussed in detail below (Depret *et al.*, 2020; Deniau *et al.*, 2020).

Role of DPP3 in terminal stages of protein turnover

Cellular proteins destined for elimination are degraded by the ubiquitin proteosome system, and the resulting peptides (3–24 amino acid residues) are released into the cytosol (York *et al.*, 2003). These peptides contribute to the cytosolic pool of peptides called peptidome, which is subsequently acted upon by a set of downstream cytosolic endoaminopeptidases/exoaminopeptidases and hydrolysed into constituent amino acids. Peptides larger than 16 amino acid residues are degraded by tripeptidyl peptidase II (TPP II) while peptides in the range of six to 17 amino acid residues are hydrolysed by thimet oligopeptidase (TOP). The resulting shorter peptides are then completely hydrolysed to their constituent amino acids by cumulative action of terminal aminopeptidases. Like to these aminopeptidases, DPP3 is very likely to contribute to the cytosolic protein turnover. Peptides four to eight amino acid residues long generated *in vivo* by proteosomal/cytosolic peptidase activity fall within the optimal substrate length of DPP3. This fraction of peptidome can thus be efficiently hydrolysed by DPP3. In addition, its non-specific catalytic behaviour and high affinity for physiologically relevant peptides at cellular pH strongly recommend its inclusion in the afore mentioned repertoire of aminopeptidases. Moreover, the post proline activity of DPP3 (Barsun *et al.*, 2007) can further enable this peptidase to degrade proline containing peptides which are otherwise resistant to most of the aminopeptidases involved in protein turnover. In line with this hypothesis, Zhan *et al.* implicated this peptidase in turnover of lens proteins leading to cataractogenesis in Shumiya cataract rats. A 5–45.5-fold higher DPP3 activity in cataractous lens compared with normal lens further supports its role in this pathology (Swanson *et al.*, 1984) and the peptide turnover. A positive correlation between its activity and age of cataract seems to be the result of increasing oxidative stress [91] which leads to a concomitant increase in DPP3 activity. The increased DPP3 activity, in turn, may contribute to the rapid turnover of oxidized proteins in such tissues (Zhan *et al.* 2001).

Although intracellular peptides have a transient life span of a few seconds within the proteolytic environment of the cytosol, their steady state levels are maintained due to constant turnover of proteins. Interestingly, a small fraction of the peptidome that resists complete degradation is presented on the cell surface complexed with MHC I molecules (Goh and Cooper, 2008). Peptides of this fraction are released as C-terminal trimmed

precursor forms during the protein degradation process. Cytosolic aminopeptidases, most probably including DPP3, trim N-terminal ends of these precursors to their presentable antigenic size (eight to nine amino acids). These peptides then move to endoplasmic reticulum where they complex with MHC I molecules and migrate to the cell surface. Since they are not protected in cytosol (Goh and Cooper, 2008), overexpression of aminopeptidases involved in cytosolic protein turnover including DPP3 can limit the presentation of these antigenic peptides to immune cells, leading to pathologies such as cancer.

Similarly, in view of its presence in human body fluids including serum (Shimamori *et al.*, 1986), this peptidase may participate in assimilation of peptides such as dietary fragments, hormones or their fragments in body fluids. This view has been supported by association of elevated angiotensin hydrolysing DPP3 activity that exists in blood plasma during pregnancy (Simaga *et al.*, 2003).

Hence, considering its likely role in cytosolic peptide turnover, DPP3 like other aminopeptidases (DPP I, DPP IV and metalloaminopeptidases) may thereby potently regulate several physiological／pathological processes associated with these peptides.

DPP3 as a modulator of the renin-angiotensin system

The renin-angiotensin system (RAS) plays a crucial role in the regulation of several physiological parameters, such as blood pressure and electrolyte homeostasis. The coordinated effects on the heart, kidney and blood vessels are mediated by a set of peptides that are generated from angiotensinogen: first, renin releases angiotensin I (ANG I), a decapeptide from the precursor protein (>350 amino acids), and second, angiotensin-converting enzyme removes a dipeptide from the C-terminus to generate angiotensinII (ANG II). ANG II is the primary effector peptide of the RAS, affecting functions of almost all organs, including heart, kidney and vasculature, with both physiological and pathophysiological implications (Mehta and Griendling, 2007). Recently, it has been shown that DPP3 is also present in low levels (median value of 15 ng/mL1) in the circulation of healthy individuals (Rehfeld *et al.*, 2019; Blet *et al.*, 2021). The fact that DPP3, along with its best described substrates angiotensins, is colocated in the circulatory system laid the foundation for the hypothesis that DPP3 may affect the RAS and potentially impact on hemodynamic and the physiology of the cardiovascular system. Wilson *et al.* (Wilson *et al.*, 2015) recently demonstrated DPP3 activity in the culture media of humanHK-2 cells, indicating a possible secretion or release of intracellular DPP3(Cruz-Diaz *et al.*, 2016; Wilson *et al.*, 2015). Furthermore, Wattieux *et al.* (Wattieux *et al.*, 2007) reported increased DPP3

activity in the culture media after anti-Fas receptor (CD95) antibody-mediated cell death. This leakage was attributed to the disruption of plasma membranes, a process known as secondary necrosis (Wattieux *et al.*, 2007). Apoptotic and necrotic cell death occurring in Fas receptor-mediated death pathways have been described previously (Matsamura *et al.*, 2000; Leist *et al.*, 1996; Vercammen *et al.*, 1998). Since cell death plays a major role in critical pathological situations, it could be envisioned that intracellular DPP3might enter the bloodstream due to massive cell death. In a recent article, van Lier *et al.* (2020) similarly suggested a link between progressive cell death and the uncontrolled release of DPP3 into the circulation during shock of various aetiologies (van Lier *et al.*, 2020). ANG II exerts its most important effects through the ANG II receptor type 1 (AT1R) resulting in mostly beneficial outcomes, such as vasoconstriction, Na+/water homeostasis, the activation of the sympathetic nervous system as well as positive inotropic and chronotropic effects on the myocardium. Consequently, ANG II helps maintain blood pressure and the perfusion of vital organs (Mehta *et al.*, 2007; Dinh *et al.*, 2001). The negative effects of ANG II come from its longer exposure effects (which may be pathological). These include proliferation, cardiac hypertrophy and remodelling, hypertrophy and hyperplasia of vascular smooth muscle cells (VSCM) and inflammation (Mehta *et al.*, 2007; Dinh *et al.*, 2001; Forrester *et al.*, 2018). The distinction between these beneficial short-term effects and the pathological longer exposure effects is the basis for the clinical applications targeting the RAS in critical care. The hydrolytic products of ANG II can have similar (ANG III (Fyhrquist and Saijonmaa, 2008) or opposite effects (ANG (1-7)and ANG IV (Forrester *et al.*, 2018). ANG (1-7) is thought to balance the RAS through activation of an alternative RAS pathway, which opposes the activity of ANGII (Forrester *et al.*, 2018). Almost all of the above mentioned angiotensins are substrates of DPP3, which puts DPP3 in a role of altering their distribution and availability. Recently, Jha *et al.* (Jha *et al.*, 2020) used a DPP3-knockout mouse model to investigate the effect of DPP3 on the levels of angiotensin peptides in the serum ("RAS-Fingerprint"). This study revealed that DPP3 deficiency results in elevated levels of ANG II, III, IV and ANG (1-5) causing increased water intake and formation of reactive oxygen species in the kidneys (Jha *et al.*, 2020).

Defence against oxidative stress

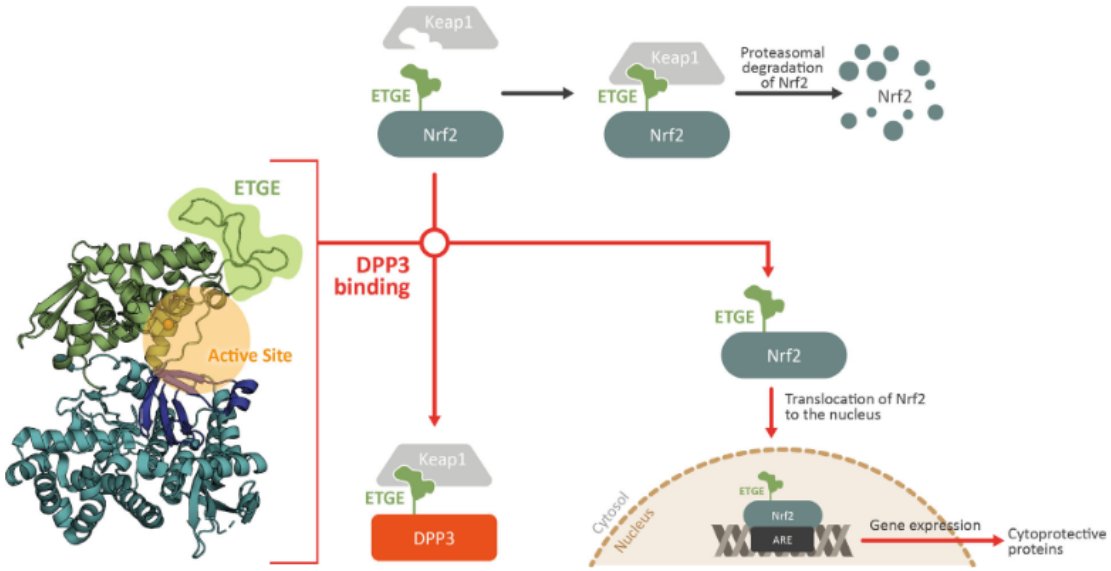


Figure39: *Scheme of DPP3 role in Nrf2-Keap1 patway (from Malovan et., 2022).*

High cellular oxidative stress is associated with nuclear translocation of NF-E2 (nuclear factor erythroid- derived 2) related factor 2 (Nrf2) resulting in transcriptional activation of genes encoding for phase II detoxifying enzymes (Niture *et al.*, 2009). Liu *et al.* in 2007, reported nuclear migration of Nrf2 in response to overexpression of DPP3 in neuroblastoma cells (IMR-32 cells). Although the mechanism by which DPP3 mediates the translocation of Nrf2 to the nucleus has not been elucidated, DPP3 overexpressing cells have been reported to efficiently attenuate the toxic effects of H2O2 and rotenone, thereby demonstrating the cytoprotective effect of this peptidase against oxidative insult. H2O2 is a tumour-derived factor which has been demonstrated in ovarian cancer cells to transcriptionally upregulate the expression of Ets-1 (Wilson *et al.*, 2005), a critical regulator of DPP3 expression (Shukla *et al.*, 2010). Hence, the elevated expression of DPP3 in ovarian cancer may be a result of increased Ets-1 levels, and overexpression of this peptidase may in turn contribute to the expression of phase II detoxifying enzymes such as NAD(P)H: quinine oxidoreductase 1 (NQO 1). Elevated DPP3 may also help in the elimination of proteins that are rendered non-functional due to oxidative damage. These observations taken together with the enhanced expression of DPP3 in cataractous tissues that are characterized by high levels of reactive oxygen species indicate it to be a possible biomarker of cellular oxidative stress.

Liu *et al.* in 2007, further demonstrated that inhibitors of phosphatidyl inositol 3 kinase and protein kinase C (PKC) abolished nuclear migration of Nrf2 in response to DPP3 overexpression. The mechanism involved in migration of Nrf2 to the nucleus by these kinases via DPP3 is still unclear. However, in view of the presence of nine putative PKC phosphorylation sites in the primary structure of hDPP3 (http://myhits.isb-sib. ch/cgi-bin/motif_scan), it appears that the peptidase acts downstream to the kinase. Moreover, because endogenous / synthetic short peptides can mimic protein–protein interactions, they act as potent inhibitors of many peptidases / kinases including PKC (Ferro *et al.*, 2004). Degradation of these inhibitory peptides due to increased DPP3 expression may upregulate their kinase activity, leading to Nrf2 phosphorylation at Ser40 and its subsequent migration to the nucleus. These authors further reported upregulation of the genes encoding chaperone CCT5, micro-RNA-21 (anti- apoptotic), and non-coding RNA MALAT1 in response to elevated expression of DPP3. Micro- RNA-21, a non-coding RNA reported to confer antiapoptotic properties to cancer cells (Chan *et al.*, 2005), is overexpressed in highly malignant human brain tumour (glioblastoma) and cultured glioblastoma cell lines. Similarly, MALAT1 is also a non-coding RNA involved in cancer metastasis (Ji *et al.*, 2001) and alternate splicing regulation (Anko and Neugebauer, 2010).

Angiotensins, the key peptide players of the rennin-angiotensin system (RAS), have vasoactive properties. High DPP3 activity in blood plasma may rapidly scavenge angiotensins (II, III and IV), with a plausible blood pressure lowering effect. Interestingly, dipeptides released by DPP3 mediated hydrolysis of its substrates are inhibitory to angiotensin-converting enzyme (Sentandreu and Toldra, 2005). This inhibition could also reduce the level of functional angiotensins thereby further lowering the blood pressure. In contrast, hydrolysis of hemopressin by DPP3 may potentially contribute to elevation of the same (Dale *et al.*, 2005). In addition to their role in regulating blood pressure, angiotensins also bind to their cognate receptors in brain tissues, potently contributing to the cyclicity of reproductive hormones and sexual behaviour, pituitary secretion, neuritic outgrowth, angiogenesis, kidney function, learning and memory. Thus, DPP3, behaving as signalling scissors in either intracellular or extracellular milieux, may potently modulate an array of diverse biological events mediated by bioactive peptides.

Materials and methods

The Cardiovascular Disease Knowledge Portal (CVDKP)

The Cardiovascular Disease Knowledge Portal (CVDKP http://www.broadcvdi.org/) part of the Human Genetics Amplifier (HuGeAMP), is a software platform and ecosystem of interconnected web portals that integrate, interpret, and present human genetic and genomic data to spark insights into complex diseases.

CVDKP framework is being developed by a team of scientists and software engineers at the Broad Institute of MIT and Harvard. Data in the CVDKP were generated by the Atrial Fibrillation Consortium (AFGen), the Global Lipids Genetics Consortium (GLGC), the Heart Failure Molecular Epidemiology for Therapeutic Targets consortium (HERMES), the Myocardial Infarction Genetics Consortium (MIGen), and the CARDIoGRAMPlusC4D Consortium.

CVDKP was searched for association of high impact single nucleotide polymorphisms (SNPs) in DPP3 gene with traits related to cardiovascular pathologies (last access 20/10/2020).

Pymol

PyMOL is a molecular graphics system with an embedded Python interpreter designed for real-time visualization and rapid generation of high-quality molecular graphics images and animations. It was used to create a model of the altered protein presenting the different single nucleotide polymorphisms identified in the knowledge portals.

Site-Directed Mutagenesis

The pET21a plasmid containing the wildtype DPP3 gene was provided by the Institute for Biochemistry of the Graz University of Technology (Figure 42).

Custom mutagenic primers were designed using NEBaseChanger (version 1.3.2).

The Q5 Site-Directed Mutagenesis Kit, designed for rapid and efficient incorporation of insertions, deletions and substitutions into double-stranded plasmid DNA, was used to introduce the nucleotide substistution of interest in the wild type DPP3 gene, according to the manufacturer's instruction (Figure 40).

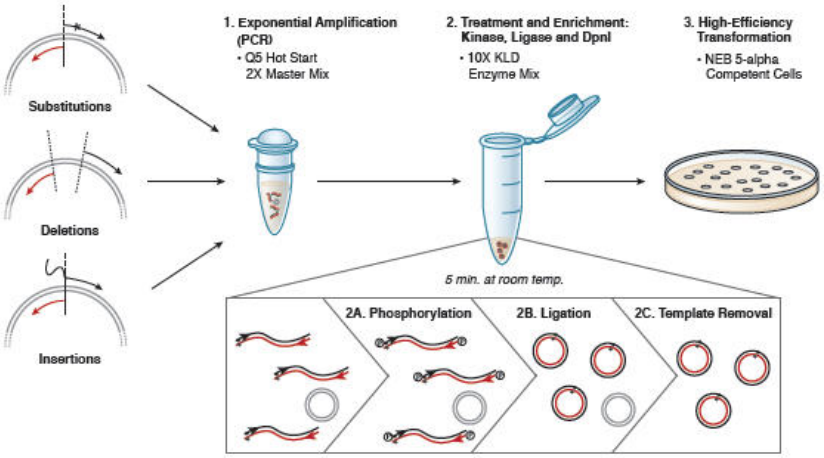


Figure 40 *Q5 Site-Directed Mutagenesis Kit Overview. The first step is an exponential amplification using standard primers and a master mix fomulation of Q5 Hot Start High-Fidelity DNA Polymerase. The second step involves incubation with a unique enzyme mix containing a kinase, a ligase and DpnI. Together, these enzymes allow for rapid circularization of the PCR product and removal of the template DNA. The last step is a high-efficiency transformation into chemically competent cells.*

The next day a single colony was picked up from plate, inoculated in a falcon containing 5 ml of LB media and 5µl of Ampicillin 1M and growth overnight with continuous shaking at 37 °C.

Plasmid Purification

The GeneJET Plasmid Miniprep Kit was used to purify the plasmid from the overnight culture according to the manual. The kit utilizes an exclusive silica-based membrane technology in the form of a spin column to selectively bind DNA molecules at a high salt concentration (Figure 41).

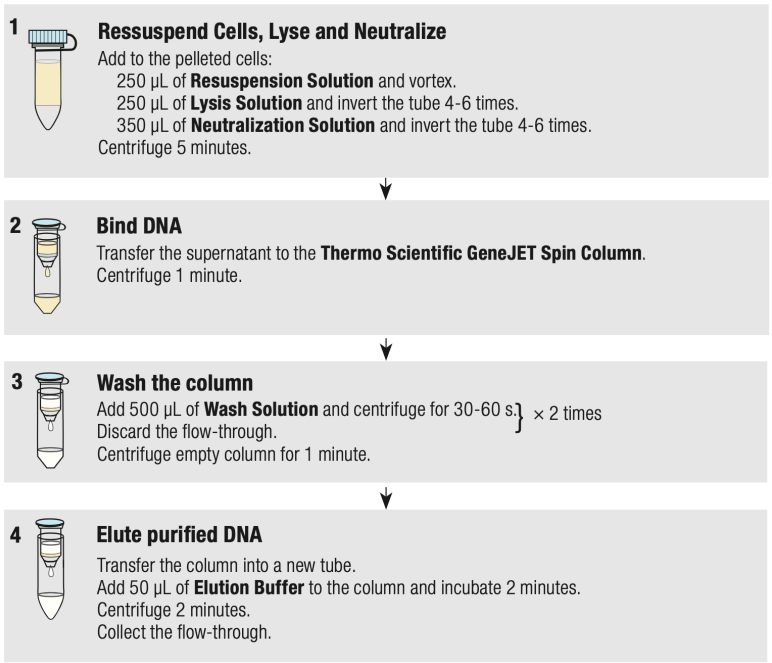


Figure 41 *Pelleted bacterial cells are resuspended and subjected to SDS/alkaline lysis (1) to liberate the plasmid DNA. The resulting lysate is neutralized to create appropriate conditions for binding of plasmid DNA on the silica membrane in the spin column (2). Cell debris and SDS precipitate are pelleted by centrifugation, and the supernatant containing the plasmid DNA is loaded onto the spin column membrane. The adsorbed DNA is washed to remove contaminants, and the pure plasmid DNA is eluted in a small volume of elution buffer (10 mM Tris-HCl, pH 8.5) or water. The purified DNA is ready for immediate use in all molecular biology procedures.*

Sanger sequencing and alignment

In order to confirm the success of nucleotide substitution, the purified plasmids were sent to LGC Genomics (LGC, Biosearch Technologies, Germany) to be sequenced through Sanger sequencing technology. The sequences obtained were then aligned to the wildtype sequence using the CLUSTAL-OMEGA algorithm.

Protein Expression Using the T7 system in BL21(DE3) E. coli cell strand.

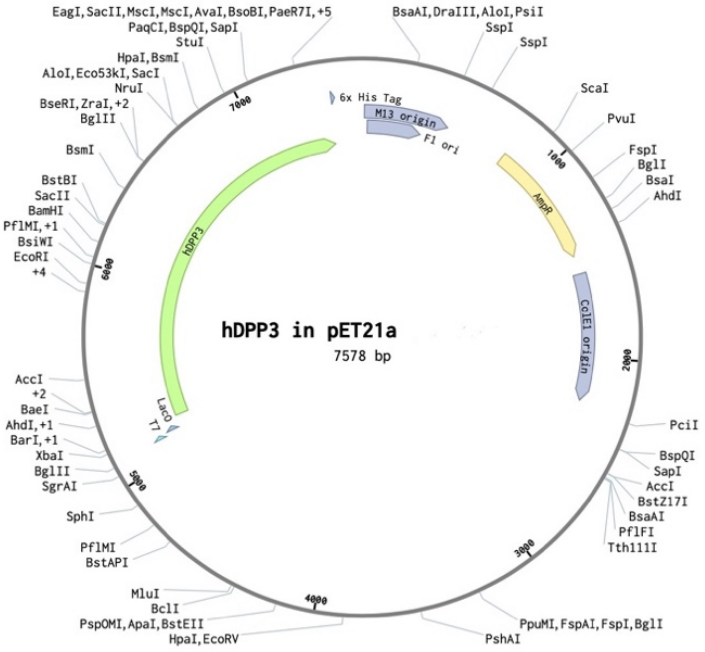


Figure 42: *hDPP3 gene in pET21a plasmid.*

After Sanger confirmation of the successful site directed mutagenesis, the mutated protein was expressed in BL21(DE3) *E. coli* strand produced in house.

Protein expression in the bacterium *E.coli* is the most popular means of producing recombinant protein.

E. coli is a well-established host that offers short culturing time, easy genetic manipulation and low-cost media. *E. coli* has a long history of being capable of producing a wide variety of different types of proteins. Within the realm of *E. coli* expression, the T7 system is the most popular approach for producing proteins. In this system, an expression vector containing a gene of interest cloned downstream of the T7 promoter is introduced into a T7 expression host. T7 expression hosts such as DE3 strains carry a chromosomal copy of the phage T7 RNA polymerase gene. When inducer is added, T7 RNA polymerase is expressed and becomes dedicated to transcription of the gene of interest (Figure 42). The protein presents at the N-terminal a 6x His Tag needed for the purification step.

Protocol:

Day 1 – Trasformation in BL21(DE3) cells

1. Thaw a tube of BL21(DE3) Competent E. coli cells on ice. Mix gently and carefully pipette 100 µl of cells into a transformation tube on ice.

2. Add 5 µl containing 1 pg–100 ng of plasmid DNA to the cell mixture. Carefully flick the tube 4–5 times to mix cells and DNA. Do not vortex.
3. Place the mixture on ice for 30 minutes. Do not mix.
4. Heat shock at exactly 42°C for 2 minutes. Do not mix.
5. Place on ice for 5 minutes. Do not mix.
6. Pipette 950 µl of room temperature LB media into the mixture.
7. Place at 37°C for 60 minutes with continuous shaking (Eppendorf Thermomixer confort).
8. Warm selection plates to 37°C.
9. Spread 50–100 µl of each dilution onto a selection plate and incubate overnight at 37°C.

Day 2 – Media preparation and Overnight Culture

1. Prepare 12 flasks containing 1l LB media each.
2. Sterilize the flasks in autoclave at 120°C for 20 minutes.
3. Wait till the media reaches the room temperature.
4. Resuspend a single colony in 200 ml of LB media containing 200 µl of ampicillin 1M to produce a starter culture.
5. Incubate overnight at 37 °C with continuous shaking (Figure 43).

***Figure 43**: Shaker inside the 37°C room.*

Day 3 – Cells Growth and Protein Expession

1. Inoculate starter culture at a 1:100 dilution into LB media containing ampicillin.
2. Incubate at 37°C with shaking until OD_{600} reaches 0.4–0.8. (Figure 43)
3. .Induce protein expression with 125 µM IPTG for each flask and express protein overnight at 20°C (Figure 44)

Figure 44: *Shaker with temperature control.*

Day 4 - Protein harvesting

1. Transfer the media from the flasks to the centrifuge bottles.
2. Centrifuge for 15 minutes at 5000 rpm 4°C (Figure 45)

Figure 45: *Centrifuge used for harvesting.*

3. Discard the supernatant and transfer the pellet to a Falcon tube.
4. Centrifuge the pellet containing falcons for 30 minutes at 4500 rcf.
5. Proceed to next step or store the pellet at -20°C.

Protein Purification

1. Resuspend the pellet in 3:1 lysis buffer + 10 µl/gr Lysozyme
2. Vortex thoroughly till the solution is homogeneous.
3. Transfer the solution to a sonication flask. Keep on ice.
4. Sonicate for 5 minutes + 2 minutes pause + 5 minutes keeping on ice (Figure 46).

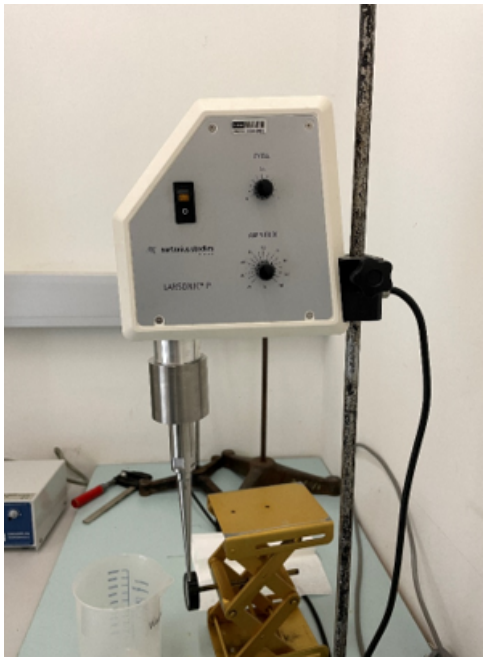
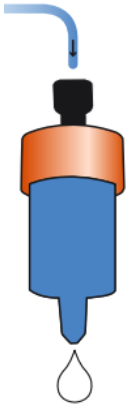


Figure46: *Labsonic P Sonicator.*

5. Centrifuge at 18000 rpm for 45 minutes (Sorvall RC6+ Thermo Scientific, California, USA)
6. Discard the pellet, filter the supernatant with a paper filter and proceed to protein purification with HisTrap^{TM}HP coloumn according to the manual.



HisTrap^{TM} HP (Cytiva Life Science, US) columns are packed with Ni Sepharose $^{®}$ High Performance (HP) affinity resin. This resin consists of highly cross-linked agarose beads to which a chelating group has been coupled. The chelating group is pre-charged with nickel, which selectively retains proteins with exposed histidine groups. Immobilized metal ion affinity chromatography (IMAC) IMAC is based on the interaction of proteins with histidine residues (or Trp and Cys) on their surface with divalent metal ions (e.g., Ni2+, Cu2+, Zn2+, Co2+) immobilized via a chelating ligand. Histidine-tagged proteins have an extra high affinity in IMAC because of the multiple (6 to 10) histidine residues and are usually the strongest binder among all the proteins in a crude sample extract (e.g., a bacterial lysate), while other cellular proteins will not bind or will

bind weakly. IMAC purification begins with equilibration of the column with a binding buffer containing a low concentration of imidazole. The imidazole binds to the immobilized metal ion and becomes the counter ligand. Proteins with histidines bind the column while displacing the imidazole counter ligands. The column is washed using the binding buffer. Bound protein is then eluted using a buffer with high concentration of imidazole. DPP3 is a colourless protein for this reason it is eluted in fraction of 3 ml each. For each fraction Biorad is used to check if these is still protein then SDS page is used to confirm the molecular weight of the protein. (Fig)

Buffer exchange was done using an Amicon MWCO 5000 filter according to the manual.

SDS PAGE

Figure 47*: Biorad™ reaction in presence (wells 1-7) and in absence (well 8) of protein.*

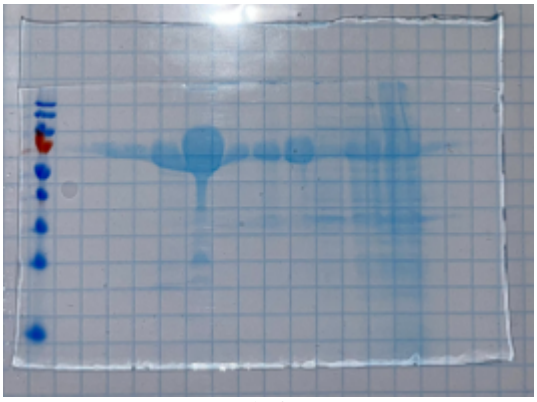


Figure 48*:SDS-PAGE gel of DPP3 protein.*

SDS-PAGE is an analytical technique to separate proteins based on their molecular weight. When proteins are separated by electrophoresis through a gel matrix, smaller proteins migrate faster due to less resistance from the gel matrix. Other influences on the rate of migration through the gel matrix include the structure and charge of the proteins.

In SDS-PAGE, the use of sodium dodecyl sulfate (SDS, also known as sodium lauryl sulfate) and polyacrylamide gel largely eliminates the influence of the structure and charge, and proteins are separated solely based on polypeptide chain length.

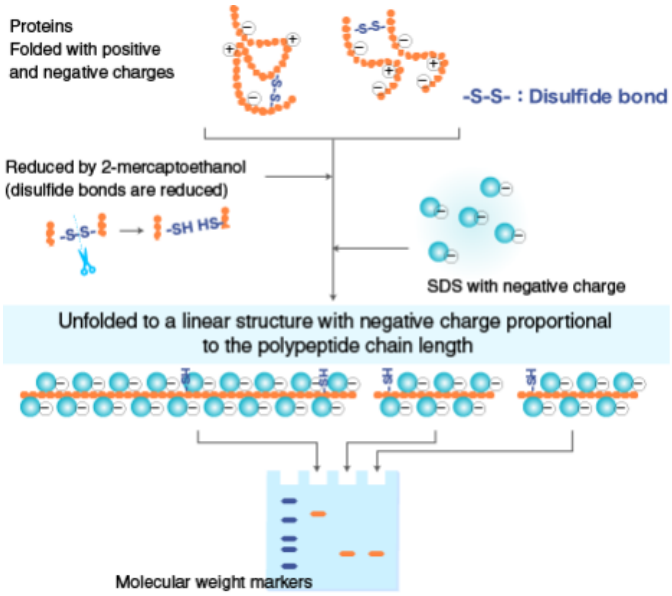


Figure 49: *Principle of SDS-PAGE.*

SDS is a detergent with a strong protein-denaturing effect and binds to the protein backbone at a constant molar ratio. In the presence of SDS and a reducing agent that cleaves disulfide bonds critical for proper folding, proteins unfold into linear chains with negative charge proportional to the polypeptide chain length.

Polymerized acrylamide (polyacrylamide) forms a mesh-like matrix suitable for the separation of proteins of typical size. The strength of the gel allows easy handling. Polyacrylamide gel electrophoresis of SDS-treated proteins allows researchers to separate proteins based on their length in an easy, inexpensive, and relatively accurate manner.

Western Blot

Western blotting was used to verify protein identity and correct molecular weight.

The method is based on building an antibody:protein complex via specific binding of antibodies to proteins immobilized on a membrane and detecting the bound antibody.

Per lane, 20 µg of samples were separated by sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE). The gel consists of 10% seperating gel (pH 8.8) and 4% stacking gel (pH 6.8). The electrophoresis is run at 100 V for 30 minutes and at 160 V for 1 hour. 12 PageRulerTM Prestained Protein Ladder (ThermoFisher Scientific, USA) was used for determination of protein size. The separated proteins were transferred onto a nitrocellulose-membrane (VWR International, USA) under constant 250 mA for one hour and the membrane was then blocked with 5% milk in 1x TBST for 1 hour. Then the membrane was incubated with primary antibody (Table 13) overnight at 4°C and after

washing the membrane was incubated with secondary alkaline phosphatase coupled antibody (Table 13) for 1 hour at room temperature. For detection, 1-StepTM NBT/BCIP substrate (ThermoFisher, USA) was used (Figure 50).

Table 13. *Antibodies used for immunofluorescence.*

Primary Antibody	Species	Dilution	Company (Catalog#)
DPP3 Polyclonal Antibody	Rabbit	1:500	Thermofisher #PA5-21709
Secondary Antibody			
Goat anti-Rabbit IgG, Alexa Fluor 488	Goat	1:500	Thermofisher #A-11008

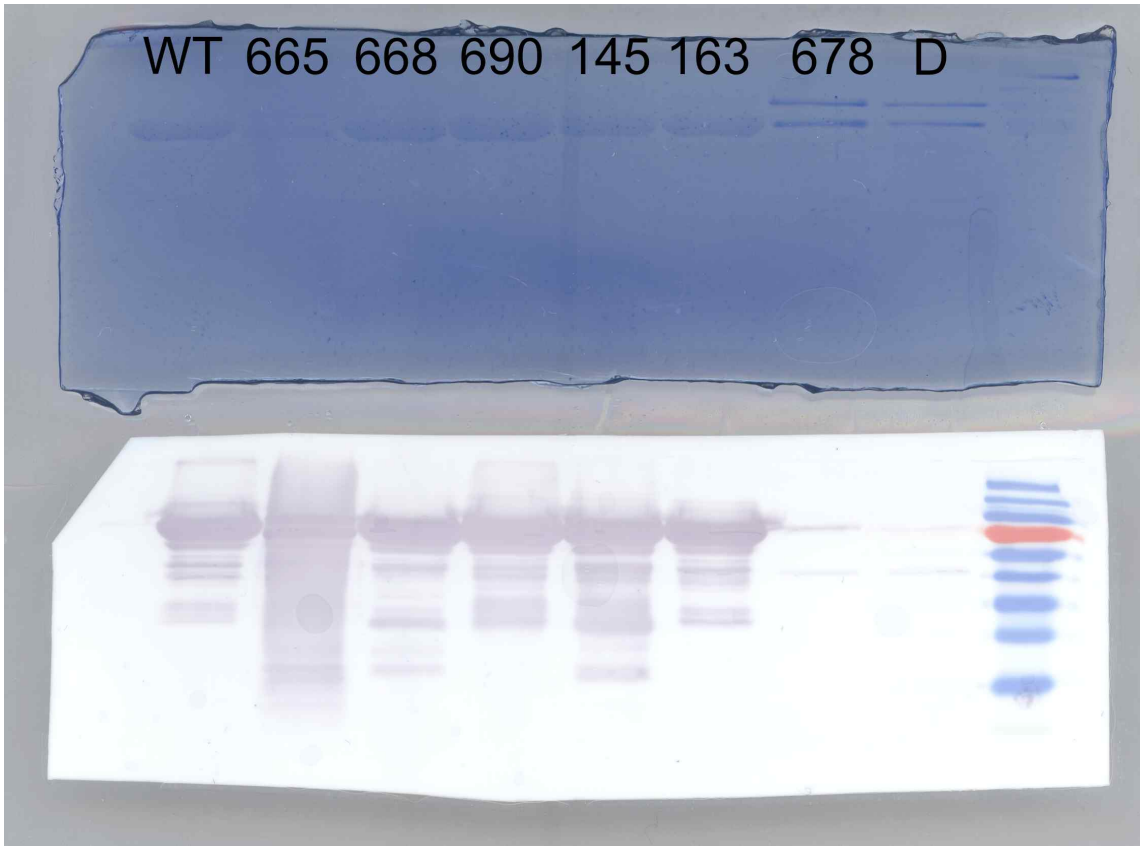


Figure 50*: Western-blot of Wild type and mutated forms of DPP3 protein.*

Activity Assay

To measure DPP3 activity, Arg2-ß-naphtylamide (Arg2-ßNA) was used as a substrate to perform a soluble enzyme activity assay. In each well of an optical plate were added 230 µl of storage buffer containing different concentrations of the substrate (0µM, 5 µM, 10 µM, 20 µM, 40 µM, 80 µM, 100 µM, 200 µM, 300 µM, 500 µM) + 5µl of the protein 50 nm. Immediately after adding the substrate, the fluorescence of the product ß-naphtylamine was measured for 15 minutes (Excitation 332 nm, Emission 420 nm) with Fluorescence Microplate Reader (Molecular Devices, USA).

A custom script written in R was used to compute the K_M and the V_{max}.

Michaelis–Menten kinetics

In biochemistry, Michaelis–Menten kinetics is one of the best-known models of enzyme kinetics. It is named after German biochemist Leonor Michaelis and Canadian physician Maud Menten. The model takes the form of an equation describing the rate of enzymatic reactions, by relating reaction rate v (rate of formation of product, [P]) to [S], the concentration of a substrate S. Its formula is given by:

$$v = \frac{\mathrm{d[P]}}{\mathrm{d}t} = V_{\max} \frac{[\mathrm{S}]}{K_\mathrm{M} + [\mathrm{S}]}$$

This equation is called the Michaelis–Menten equation. Here, Vmax represents the maximum rate achieved by the system, happening at saturating substrate concentration for a given enzyme concentration. When the value of the Michaelis constant K_M is numerically equal to the substrate concentration, then the reaction rate is half of V_{max}. Biochemical reactions involving a single substrate are often assumed to follow Michaelis–Menten kinetics, without regard to the model's underlying assumptions.

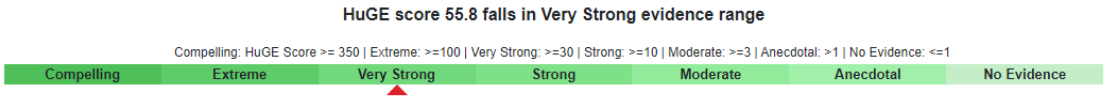
Results

CVDKP (http://www.broadcvdi.org/) was searched for association of high impact single nucleotide polymorphisms (SNPs) in DPP3 gene with traits related to cardiovascular and atherothrombotic diseases.

Genome-wide significant association ($P<5\times10-8$) was observed for *DPP3* with phenotypes linked to lipids (Table 14).

Table 14: Phenotypes significantly associated with DPP3.

Gene	Variants	pValue	Phenotype	Sample Size
DPP3	14	1,09E-20	LDL Cholesterol	1514886
DPP3	14	2,38E-12	Total Cholesterol	1487253
DPP3	13	2,98E-10	nonHDL Cholesterol	925904



In particular, a strong genome-wide significant association with reduced LDL cholesterol levels was observed for two SNPs, rs2305535 and rs11550299, in the UK Biobank Mendelian trait GWAS (Forgetta *et al.*, 2022) and in the GLGC exome chip analysis (Lu X *et al.*, 2017).

In GWAS, it is difficult for rare variants to reach such stringent threshold, even when the sample size is huge. For this reason we extended our analysis to rare (MAF<0.01) missense variant in *DPP3* presenting a p-value within 0.05. A list of these variant is reported in Table 15.

Table 15: *rare missense variant in the region presenting a p-value within 0.05. For each SNP the other CVD associated traits, the global frequency observed in GNOMAD exomes dataset and the Phred-style CADD raw scores are reported.*

rs	Protein change	Trait	gnomAD_AF	CADD
rs201658473	p.R24C	Diastolic blood pressure↓	0,00003183	24,8
rs114567543	p.A61S	Athrial fibrillation↑,HbA1c↑,Fasting Glucose↑	0,003715	20,6
rs145950613	p.E131G	HbA1c↑,Waist circumference↓	0,000199	14,61

rs11550299	p.Q145H	BMI↓,Age-related macular degeneration↑,Athrial fibrillation↓,eGFR-creat (serum creatinine)↓,Height↑,Hip circumference adj BMI↑,Lactate dehydrogenase↑,PC1 dietary pattern↑,PC3 dietary pattern↑,Triglycerides,Waist circumference↑,Waist-hip ratio↓,Red blood cell count↓,Total cholesterol↓	0,2223	16,73
rs112484606	p.L163M	Creatinine↑,eGFR-creat (serum creatinine)↓,Alkaline phosphatase↑	0,001381	24,1
rs149825536	p.M178T	Peripheral vascular disease in type 2 diabetics↓, type 2 diabetes ↑	0,0001551	21
rs151275489	p.A267T	Estimated bone mineral density↓,Red blood cell count↑,HbA1c↓,Systolic blood pressure↑,Total cholesterol↓Neovascular age-related macular degeneration↑↑	0,0009606	15,26
rs1326815078	p.S272R	Estimated bone mineral density,Red blood cell count	0,000008068	21,6
rs201933676	p.I282M	Total cholesterol↓,Height↑	0,0001596	15,27
rs148964668	p.R490Q	BMI↑	0,00004798	15,44
rs34504069	p.H521Y	Albumin↓↓↓↓,BMI↑	0,0006069	18,16
rs150300932	p.R604C	Triglycerides↓	0,00196	23,7
rs139251036	p.R665H	BMI,↑Waist-hip ratio adj BMI↓	0,0001437	23,6
rs146568792	p.S668Y	Age-related macular degeneration↓,Height↓,Red blood cell count↑,Waist-hip ratio↓,type 2 diabetes↑↑,Adiponectin↑	0,001066	23,6
rs12421620	p.E690K	Any cardiovascular disease↑↑,Any stroke↑,Any cancer↑,Type 2 diabetes↑,Hypertension↑,Diastolic blood pressure↓,Systolic blood pressure↑,HDL cholesterol↑,Total cholesterol↓,LDL cholesterol↓,BMI↓,Height↑,Weight↓,Creatinine↑,Microalbuminuria↑,paper breast cancer (pubmed 25803781)	0,07293	23,5
rs747171479	.Asp444His		0.00000891	23,7

For each SNP the other CVD associated traits, the global frequency observed in GNO-MAD exomes dataset and the Phred-style CADD raw scores are reported. The Genome Aggregation Database (gnomAD) is a resource developed by an internation-al coalition of investigators, with the goal of aggregating and harmonizing both exome and genome sequencing data from a wide variety of large-scale sequencing projects, and mak-ing summary data available for the

wider scientific community. (Karczewski *et al.*,2020). The Combined Annotation Dependent Depletion (CADD) tool scores the predicted dele-teriousness of single nucleotide variants and insertion/deletions variants in the human genome by integrating multiple annotations including conservation and functional infor-mation into one metric. Variants with higher score are more likely to be deleterious. (Kircher *et al.*, 2014). Variants were annotated with VEP adapting the script discussed above (see pg. 68).

For rs2305535, rs11550299, and each variant in table 15 a model of the mutation was built using PyMol.

Basing on:

a) Pymol model results

b) in silico predictors of pathogenicity results,

c) available literature,

5 variants were selected: rs2305535, rs11550299, rs139251036, rs12421620 and rs747171479.

The 5 mutated forms of the protein were heterologously expressed in *E.coli*, taking advance from site directed mutagenesis and T7expression system in BL21(DE3) *E. coli* cell strand.

The enzymatic activity of the wild type and 5 mutated proteins was experimentally determined. Below the activity assay for the wild type protein, and PyMol model and activity assay results for the 5 variant are reported.

Wild Type

The activity assay identified a KM= 31 ±3 µM and a Vmax =220 ±7 nmol/s for the wild-type protein (Figure 51).

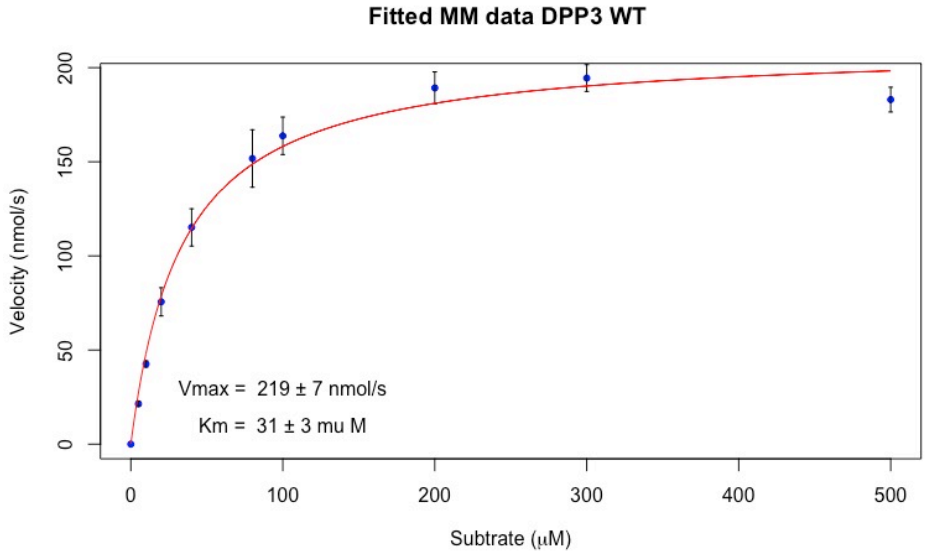


Figure 51: *Fitted Michealis & Menten plot for WT DPP3.*

rs2305535 - chr11_66272237_G/A

It is a missense variant in exon 17 consisting in a nucleotide substitution from guanine to adenine at position 2033 of the coding sequence. It determines the aminoacid substitution from Arginine to Histidine at position 678 of the protein sequence. The Pymol model showed the loss of two interactions with the E702 residue due to the loss of two aminic residues in the side chain (Figure 52). The activity assay identified a KM=32 ±2 µM and a Vmax =243 ± 7 nmol/s (Figure 53).

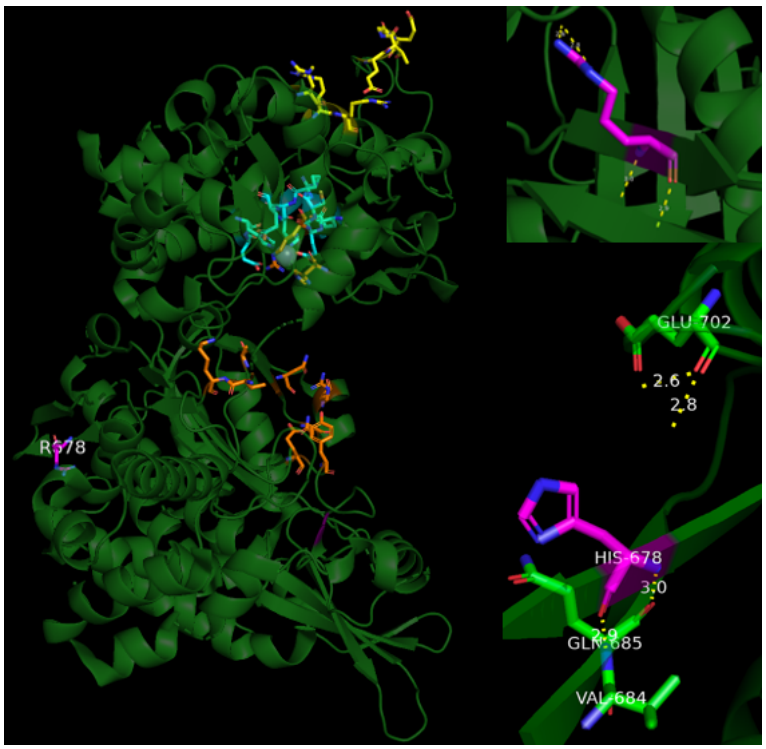


Figure 52: *PyMol model cartoon of the mutated form of the protein. a)localization of the mutated residue b) WT residue c) Mutated residue. (rs2305535)*

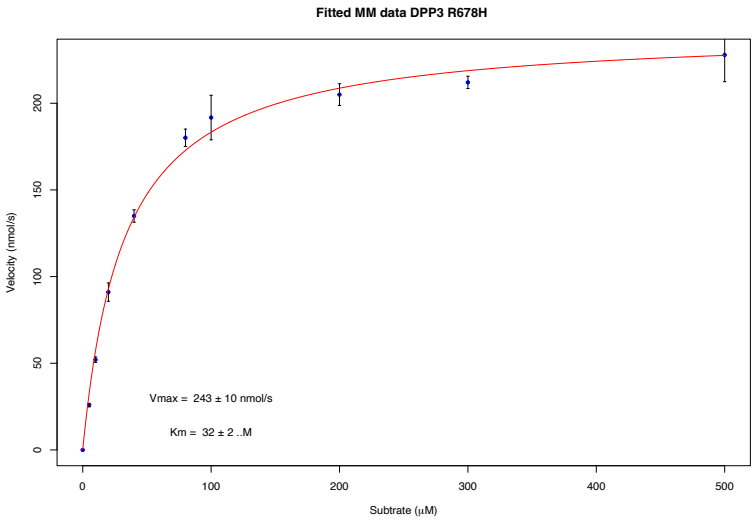


Figure 53: *Fitted Michealis & Menten plot for the mutated form of DPP3 (rs2305535)*

rs11550299 - chr11_66254085_G/T

It is a missense variant in exon 4 consisting in a nucleotide substitution from guanine to thymine at position 435 of the coding sequence. It determines the aminiacid substitution

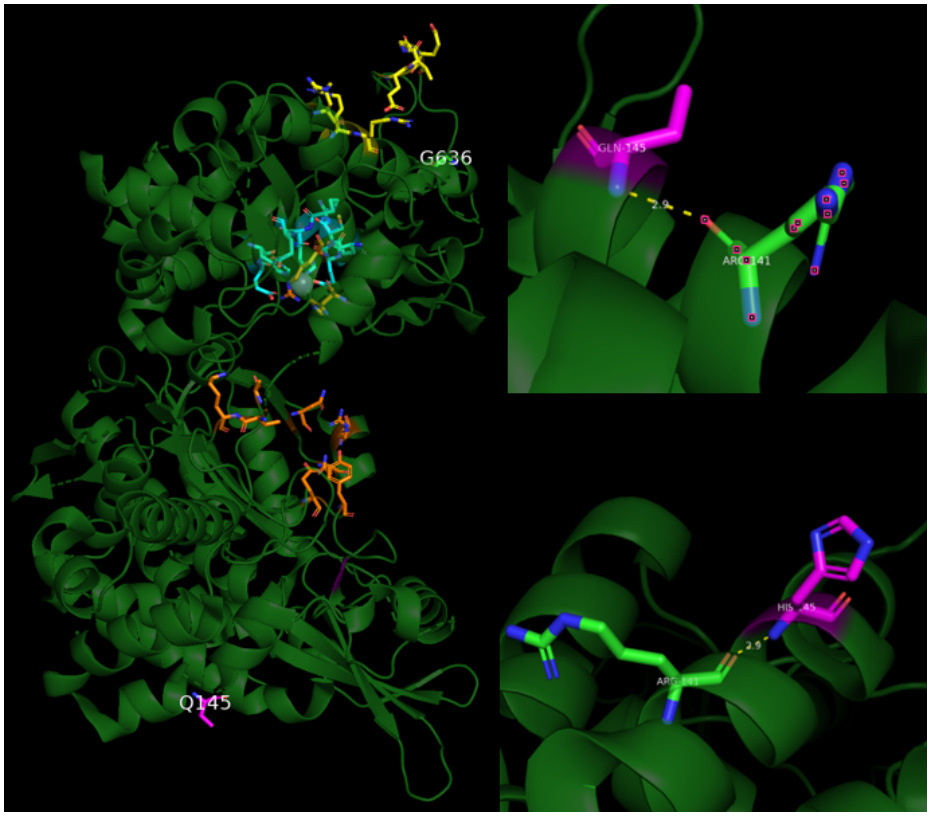


Figure 54: *PyMol model cartoon of the mutated form of the protein. a)localization of the mutated residue b) WT residue c) Mutated residue. (rs11550299)*

from glutamine to histidine at position 145 of the protein sequence. Our Pymol model did not show any differences in interactions in respect to the wild type (Figure 54). The activity assay identified a KM=38 ±2 µM and a Vmax =264 ± 19 nmol/s (Figure 55)

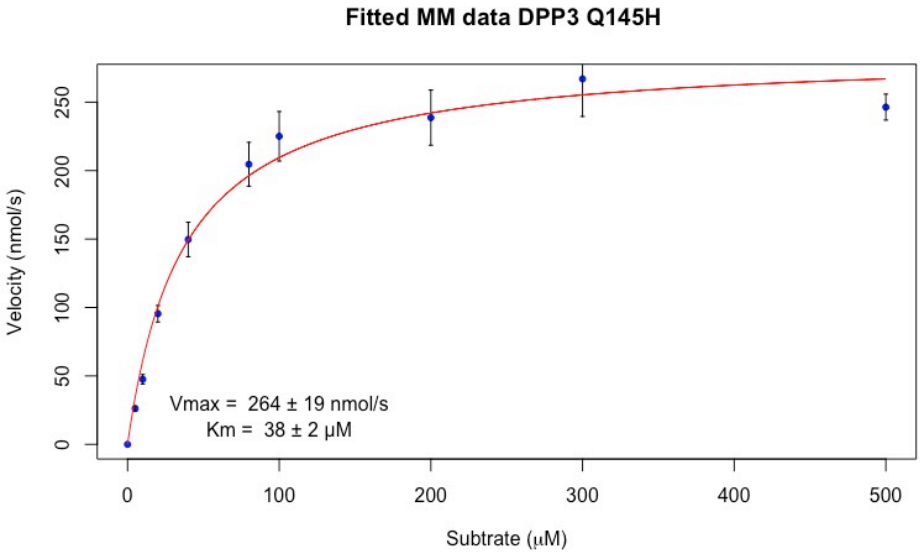


Figure 55: *Fitted Michealis & Menten plot for the mutated form of DPP3 (rs11550299)*

rs139251036 - chr11_66272198_G/A

It is a missense variant in exon 17 consisting in a nucleotide substitution from guanine to adenine at position 1994 of the coding sequence. It determines the aminiacid substitution from arginine to histidine at position 665 of the protein sequence. Our Pymol model showed the loss of several interactions with the R598 and D596 residues due to the loss of two aminic residues in the side chain (Figure 56). The activity assay identified a KM=24 ±8 µM and a Vmax =28 ± 2 nmol/s (Figure 57)

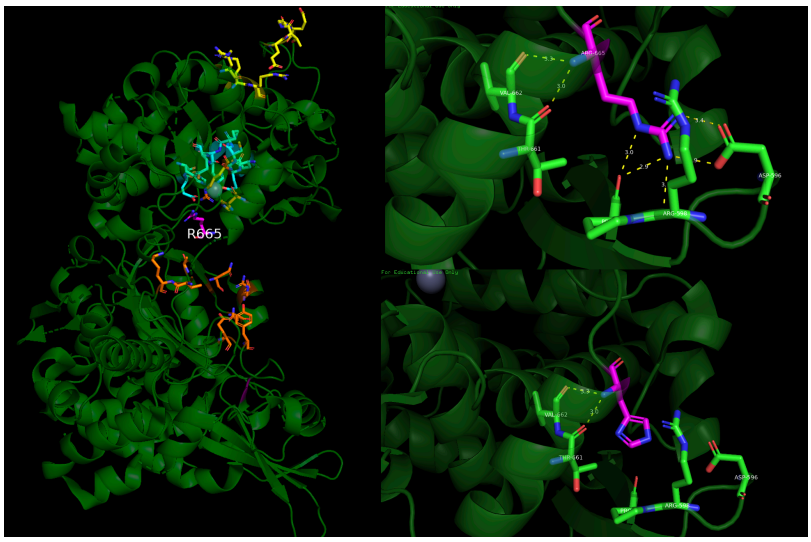


Figure 56: *PyMol model cartoon of the mutated form of the protein. a) localization of the mutated residue b) WT residue c) Mutated residue. (rs139251036)*

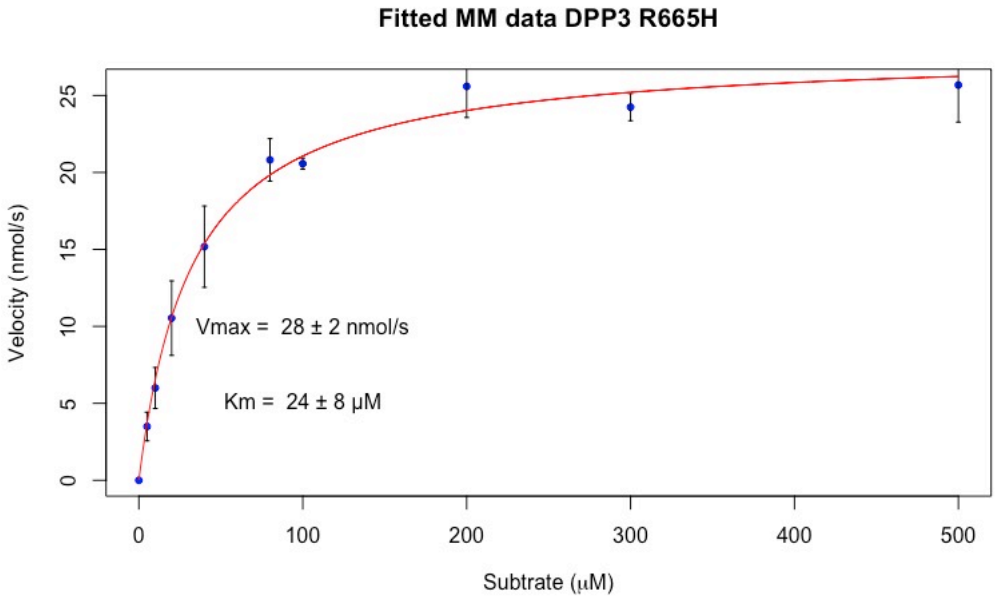


Figure 57: *Fitted Michealis & Menten plot for the mutated form of DPP3 (rs139251036)*

rs12421620 - chr11_66276576_G/A

It is a missense variant in exon 18 consisting in a nucleotide substitution from guanine to adenine at position 2068 of the coding sequence. It determines the aminiacid substitution from glutamic acid to lysine at position 690 of the protein sequence. Our Pymol model showed the loss of the interaction with the S692 residue due to the loss of the Hydroxyl group of the side chain (Figure 58). The activity assay identified a KM=34 ±3 µM and a Vmax =410 ± 28 nmol/s (Figure 59)

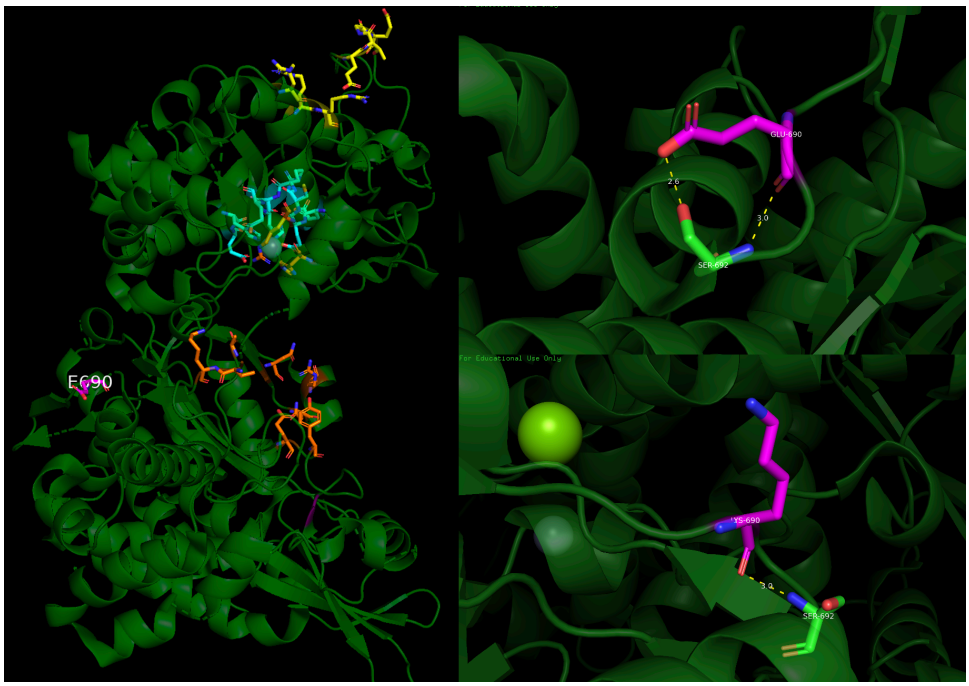


Figure 58: *PyMol model cartoon of the mutated form of the protein. a)localization of the mutated residue b) WT residue c) Mutated residue. (rs12421620)*

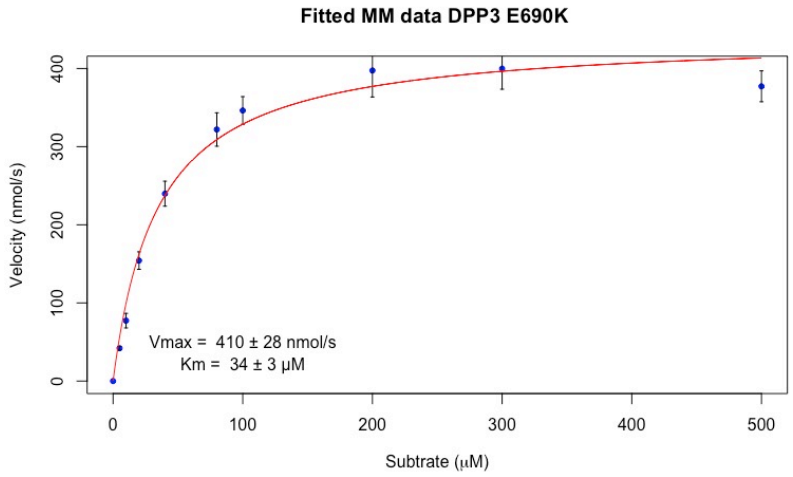


Figure 59: *Fitted Michealis & Menten plot for the mutated form of DPP3 (rs12421620)*

rs747171479 - chr11_66261045_G/C

It is a missense variant in exon 12 consisting in a nucleotide substitution from guanine to cytosine at position 1330 of the coding sequence. It determines the aminiacid substitution from aspartic acid to histidine at position 444 of the protein sequence. Our Pymol model did not show any differences in interactions in respect to the wild type (Figure 60) but showed a big increase of the steric hindrance. The activity assay identified a KM=111 ±14 µM and a Vmax =146 ± 12 nmol/s (Figure 61).

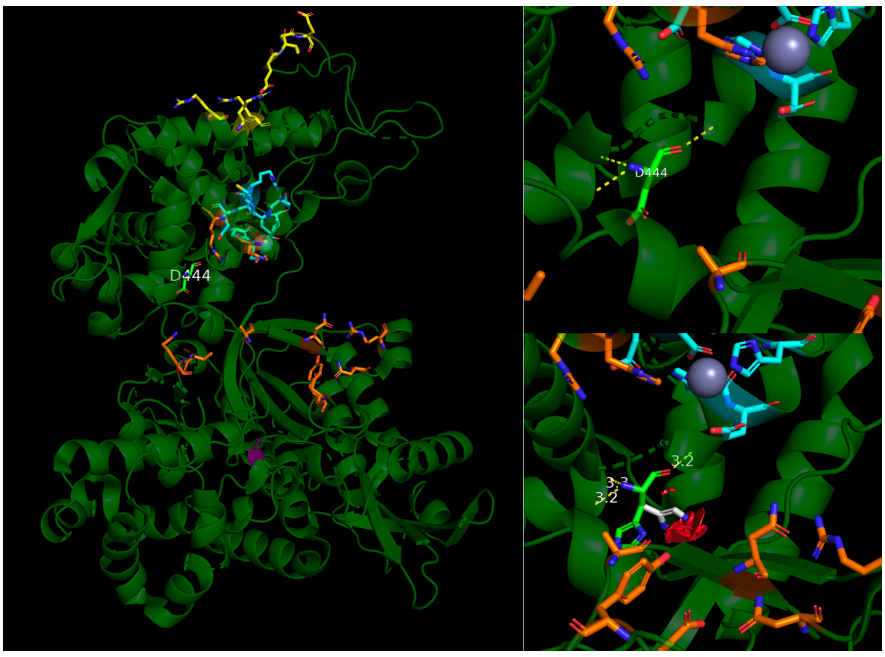


Figure 60: *PyMol model cartoon of the mutated form of the protein. a)localization of the mutated residue b) WT residue c) Mutated residue. (rs747171479)*

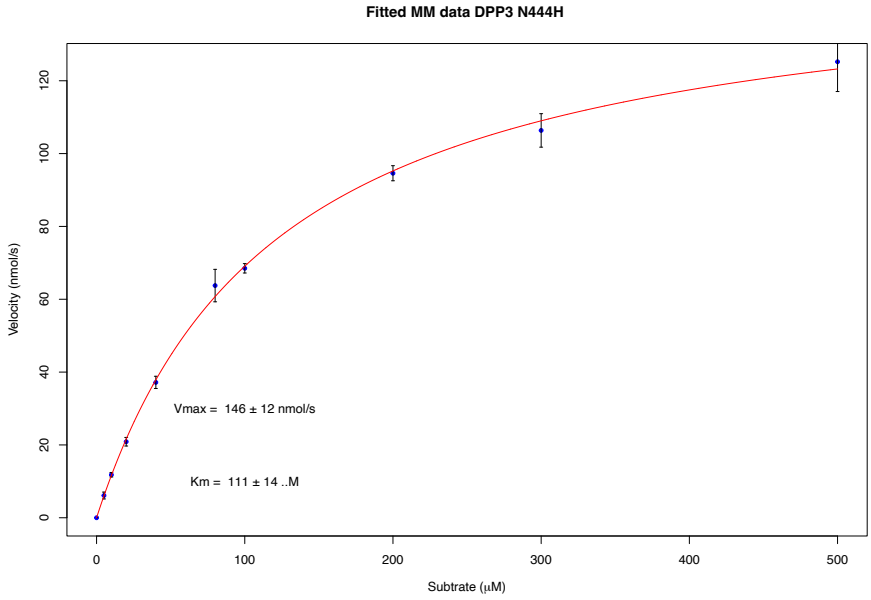


Figure 61: *Fitted Michealis & Menten plot for the mutated form of DPP3 (rs747171479)*

Discussion

Genetic association data from genome-wide association studies (GWAS) are foundational for understanding of complex diseases and traits. Nevertheless, in order to apply these results to diagnosis, drug development, and treatment, the identification of associated loci by GWAS needs to define the actual effector genes that explain those genetic associations. Indeed, most genetic variants/SNPs associated with disease are located outside of coding regions of the genome, so that their impact on genes is not obvious; and even a variant located in a coding region may actually not affect protein structure/function or affect the phenotype because in linkage with different causal variants. In this thesis, an integrated approach of *in silico* and *in vitro* methods was applied to evaluate the effect of *DPP3* variants, resulted associated to lipid metabolism alteration and cardiovascular phenotypes, on enzyme activity.

As concerns, rs11550299 and rs2305535, although the two variants are very strongly associated with the LDL cholesterol phenotype, they are far from every active site of the protein, they are very common, and all *in silico* predictors of pathogenicity consider them benign. In the activity assay they do not alter significantly the Km and Vmax. Therefore, the activity assay support the evidence of a null or moderate direct effect of the two variants on phenotype and support the possibility that the observed association could be attributed to other causal variants in linkage disequilibrium.

Beside in silico prediction - MutationTaster considers the variant to be damaging, as well as GERP score (5.42) and CADD score (23.6) - and the evidence that the rare rs139251036 variants leads to the lost of several hydrogen bounds, the activity assay showed that the mutated protein has a Km comparable to wild type and a Vmax slightly reduced with respect to the wild type. On the other hand, as shown in PyMol picture, the activity result is consistent with the datum that the 665 protein position interested by the aminoacid substitution is located on an unstructured region of the protein, far from every active region.

rs12421620 is a very controversial missense variant; it has a low MAF in the European population (0.001793), but it is extremely common in Latino (MAF=0.4) according to GnomAD. Its association with LDL levels, in GLGC 2021 Lipids GWAS, reaches a p-value of 0.006485. Furthermore, some in silico predictors of pathogenicity consider it damaging (Polyphen, MutationTaster, PROVEAN, GERP, CADD, DANN) while others consider it benign (SIFT, FATHMM). According to activity assay, it seems to have a

possible impact on the Vmax (two fold higher than wild type protein), even if it does not alter the Km. Result of the present thesis are consistent with the hypothesis of an exome analysis from Li *et al.* in breast cancer tissue. Structure analysis of DPP3 by Li and coworkers suggests that the rs12421620 mutant protein has almost similar structure to normal protein, except that the C-terminus has its helix structure changed to a loop structure because of the point mutation. It has been reported that the C-terminal structure of the protein can play an important role in substrate binding with DPP3 (Prajapati, Chauhan,2011). As the rs12421620 variant occurs close to the substrate binding residues, K666 and R669 they hypothesize that the altered structure at C-terminus affects substrate binding and consequently alters protein function.

According to our analyses, rs747171479 showed the strong evidence of its influence on the protein activity. It is extremely rare in all the populations (i.e. gnomAD NFE MAF=0.00000891). The p-value of its association with LDL in AMP T2D-GENES quantitative trait exome sequence analysis is 0.0007995, apparently low for a GWAS, but the beta value of -119.5000 indicates that this variant has an enormous impact on the serum level of LDL cholesterol. The low p-values has to be evaluated in consideration of its rare MAF. Most in silico predictors of pathogenicity consider it to be damaging. Aspartic acid 444 is located in a crucial position of the protein, the big increase in steric hindrance is likely to destabilise the substrate binding site. In activity assay the Km resulted significantly altered.

For a specific and definite evaluation of the obtained results, we should evaluate the protein function on specific substrate of DPP3 [ANGII or ANG(1-7)] through more sophisticated technologies like the isothermal titration calorimetry and cristallography.

Nevertheless, present results suggest the potential utility of the integrated approach of *in silico* and *in vitro* methods to evaluate the effect of variants in genes of interest. Due to the relatively low cost and experimental time of the applied approach, it could represent a possible resource to focus research on functional variants to be studied with more expensive, time consuming and/or ethically critical (animal models) experimental approaches.

Bibliography

1000 Genomes Project Consortium, Abecasis, G.R., Altshuler, D., Auton, A., Brooks, L.D., Durbin, R.M., Gibbs, R.A., Hurles, M.E., McVean, G.A., 2010. A map of human genome variation from population-scale sequencing. Nature 467 (7319), 1061–1073.

100,000 Genomes Project Pilot Investigators, Smedley D, Smith KR, Martin A, *et al.* 100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care - Preliminary Report. N Engl J Med. 2021 Nov 11;385(20):1868-1880.

Aandahl EM, Torgersen KM, Tasken K. CD8_ regulatory T cells—A distinct T-cell lineage or a transient T-cell phenotype? Human Immunology. 2008;69:696-699.

Ab, L. *et al.* National, regional, and global trends in body-mass index since 1980: systematic analysis of health examination surveys and epidemiological studies with 960 country-years and 9Â·1 million participants. The Lancet 377, 557–567 (2011).

Abramic M, Zubanovic M & Vitale L (1988) Dipeptidyl peptidase III from human erythrocytes. Biol Chem Hoppe Seyler 369, 29–38.

Akavia, U.D., Litvin, O., Kim, J., Sanchez-Garcia, F., Kotliar, D., Causton, H.C., Pochanard, P., Mozes, E., Garraway, L.A., Pe'er, D., 2010. An integrated approach to uncover drivers of Cancer. Cell 143 (6), 1005–1017.

Åkerblom M., Sachdeva R., Barde I., Verp S., Gentner B., Trono D., Jakobsson J. MicroRNA-124 is a subventricular zone neuronal fate determinant. J. Neurosci. 2012;32:8879–8889

Al-Ajlan FS, Goyal M, Demchuk AM, Minhas P, Sabiq F, Assis Z, Willinsky R, Montanera WJ, Rempel JL, Shuaib A, Thornton J, Williams D, Roy D, Poppe AY, Jovin TG, Sapkota BL, Baxter BW, Krings T, Silver FL, Frei DF, Fanale C, Tampieri D, Teitelbaum J, Lum C, Dowlatshahi D, Shankar JJ, Barber PA, Hill MD, Menon BK; ESCAPE Trial Investigators. Intra-Arterial Therapy and Post-Treatment Infarct Volumes: Insights From the ESCAPE Randomized Controlled Trial. Stroke. 2016 Mar;47(3):777-81. doi: 10.1161/STROKEAHA.115.012424. PMID: 26892284.

Alawieh A, Elvington A, Zhu H, Yu J, Kindy MS, Atkinson C, Tomlinson S. Modulation of post-stroke degenerative and regenerative processes and subacute protection by site-targeted inhibition of the alternative pathway of complement. J Neuroinflammation. 2015 Dec 30;12:247.

Alba F, Arenas JC, Lopez MA. Properties of rat brain dipeptidyl aminopeptidases in the presence of detergents. Peptides. 1995;16(2):325-9.

Alberico L, Catapano I, Graham G, De Backer O, Wiklund MJ, Chapman H, Drexel AW, Hoes CS, Jennings U, Landmesser T, Pedersen R. 2016 ESC/EAS guidelines for the management of dyslipidaemias. European Heart Journal. 2016;37(39):2999-3058.

Albers GW, Marks MP, Kemp S, *et al.* Thrombectomy for stroke at 6 to 16 hours with selection by perfusion imaging. N Engl J Med 2018;378:708–18.

Alexander, D.H., Novembre, J., Lange, K., 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 19 (9), 1655–1664.

Alexander JH, Lopes RD, Thomas L, *et al.* Apixaban vs. warfarin with concomitant aspirin in patients with atrial fibrillation: insights from the ARISTOTLE trial. Eur Heart J. 2014;35:224–232.

Aliaga E, Silhol M, Bonneau N, Maurice T, Arancibia S, Tapia-Arancibia L (2010) Dual response of BDNF to sublethal concentrations of β- amyloid peptides in cultured cortical neurons. Neurobiol Dis 37(1): 208–217.

Almasi, G.S., Gottlieb, A., 1989. Highly Parallel Computing. Benjamin-Cummings Publishing Co., Inc., Redwood City, CA.

Amemori T, Romanyuk N, Jendelova P, Herynek V, Turnovcova K, Prochazka P, Kapcalova M, Cocks G, Price J, Sykova E. Human conditionally immortalized neural stem cells improve locomotor function after spinal cord injury in the rat. Stem Cell Res Ther. 2013 Jun 7;4(3):68. doi: 10.1186/scrt219. PMID: 23759119; PMCID: PMC3706805.

Anko ML & Neugebauer KM (2010) Long noncoding RNAs add another layer to pre-mRNA splicing regula- tion. Mol Cell 39, 833–834.

Anrather J, Iadecola C. Inflammation and Stroke: An Overview.Neurotherapeutics. 2016 Oct; 13(4):661-670.

Anwar T, Rufail ML, Djomehri SI, Gonzalez ME, Lazo de la Vega L, Tomlins SA, Newman LA, Kleer CG. Next-generation sequencing identifies recurrent copy number variations in invasive breast carcinomas from Ghana. Mod Pathol. 2020 Aug;33(8):1537-1545. doi: 10.1038/s41379-020-0515-2. Epub 2020 Mar 9. PMID: 32152520; PMCID: PMC7390688.

Arabidopsis Genome Initiative, 2000. Analysis of the genome sequence of the flowering plant Arabidopsis Thaliana. Nature 408 (6814), 796–815. https://doi.org/10.1038/35048692.

Auer H, Newsom DL, Kornacker K. Expression Profiling Using Affymetrix GeneChip Microarrays. Methods Mol Biol. 2009;509:35-46.

Aviv RI, d'Esterre CD, Murphy BD, Hopyan JJ, Buck B, Mallia G, Li V, Zhang L, Symons SP, Lee TY. Hemorrhagic transformation of ischemic stroke: prediction with CT perfusion. Radiology. 2009 Mar;250(3):867-77. doi: 10.1148/radiol.2503080257.

Badimon JJ, Lettino M, Toschi V *et al.* Local inhibition of tissue factor reduces the thrombogenicity of disrupted human atheroscle- rotic plaques: effects of tissue factor pathway inhibitor on plaque thrombogenicity under flow conditions. Circulation 1999;99: 1780–7.

Bahar-Shany K, Ravid A, Koren R. Upregulation of MMP-9 production by TNFalpha in keratinocytes and its attenuation by vitamin D. Journal of Cellular Physiology. 2010;222:729-737.

Bak S, Gaist D, Sindrup SH, *et al.* Genetic liability in stroke: a long-term followup study of Danish twins. Stroke 2002;33:769–74.

Baral PK, Jajcanin-Jozić N, Deller S, Macheroux P, Abramić M, Gruber K. The first structure of dipeptidyl-peptidase III provides insight into the catalytic mechanism and mode of substrate binding. J Biol Chem. 2008 Aug 8;283(32):22316-24.

Barr TL, Latour LL, Lee KY, Schaewe TJ, Luby M, Chang GS, El-Zammar Z, Alam S, Hallenbeck JM, Kidwell CS, Warach S. Blood-brain barrier disruption in humans is independently associated with increased matrix metalloproteinase-9. Stroke. 2010 Mar;41(3):e123-8. doi: 10.1161/STROKEAHA.109.570515.

Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., *et al.*, 2013. NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res. 41, D991–D995.

Barsun M, Jajcanin N, Vukelic B, Spoljaric J & Ab- ramic M (2007) Human dipeptidyl peptidase III acts as a post-proline-cleaving enzyme on endomorphins. Biol Chem 388, 343–348.

Baudhuin LM, Lagerstedt SA, Klee EW, Fadra N, Oglesbee D, Ferber MJ. Confirming Variants in Next-Generation Sequencing Panel Testing by Sanger Sequencing. J Mol Diagn. 2015 Jul;17(4):456-61. doi: 10.1016/j.jmoldx.2015.03.004. Epub 2015 May 8. PMID: 25960255.

Beck TF, Mullikin JC; NISC Comparative Sequencing Program, Biesecker LG. Systematic Evaluation of Sanger Validation of Next-Generation Sequencing Variants. Clin Chem. 2016 Apr;62(4):647-54. doi: 10.1373/clinchem.2015.249623. Epub 2016 Feb 4. PMID: 26847218; PMCID: PMC4878677.

Behrendt D, Ganz P. Endothelial function from vascular biology to clinical applications. Am J Cardiol 2002;90:40L–8L.

Bellenguez C, et al; International Stroke Genetics Consortium, Wellcome Trust Case Control Consortium. Genome-wide association study identifies a variant in HDAC9 associated with large vessel ischemic stroke. Nat Genet. 2012;44:328–333.

Benedek A, Cernica D, Mester A, Opincariu D, Hodas R, Rodean I, Keri J, Benedek T. Modern Concepts in Regenerative Therapy for Ischemic Stroke: From Stem Cells for Promoting Angiogenesis to 3D-Bioprinted Scaffolds Customized via Carotid Shear Stress Analysis. Int J Mol Sci. 2019 May 25;20(10). pii: E2574. doi: 10.3390/ijms20102574. Review.

Beręsewicz A, Kurzelewski M. Unstable atherosclerotic plaque. Polish Heart Journal. 2001;54:431-439

Berkhemer OA, Fransen PS, Beumer D, van den Berg LA, Lingsma HF, Yoo AJ, *et al.* A randomized trial of intraarterial treatment for acute ischemic stroke. N Engl J Med. 2015;372(1):11–20. https://doi.org/10.1056/NEJMoa1411587.

Berry JD, Dyer A, Cai X, Garside DB, Ning H, Thomas A, Greenland P, Van Horn L, Tracy RP, Lloyd-Jones DM. Lifetime risks of cardiovascular disease. N Engl J Med. 2012 Jan 26;366(4):321-9.

Biffi A, Sonni A, Anderson CD, *et al.* Variants at APOE influence risk of deep and lobar intracerebral hemorrhage. Ann Neurol 2010; 68: 934–43

Biscetti F, Straface G, Bertoletti G, Vincenzoni C, Snider F, Arena V, Landolfi R, Flex A. Identification of a potential proinflammatory genetic profile influencing carotid plaque vulnerability. Journal of Vascular Surgery. 2015;61:374-381.

Blet A, Deniau B, Santos K, van Lier DPT, Azibani F, Wittebole X, *et al.* Monitoring circulating dipeptidyl peptidase 3 (DPP3) predicts improvement of organ failure and survival in sepsis: a prospective observational multinationa study. Crit Care. 2021;25:1–10.

Boeckh-Behrens T, Kleine JF, Zimmer C, *et al.* Thrombus histology suggests cardioembolic cause in cryptogenic stroke. Stroke 2016; 47: 1864–1871.

Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics. 2003 Jan 22;19(2):185-93.

Boorsma, E. M. *et al.* Dipeptidyl peptidase 3, a marker of the antagonist pathway of the renin–angiotensin–aldosterone system in patients with heart failure. European Journal of Heart Failure 23, 947–953 (2021).

Bowry AD, Brookhart MA, Choudhry NK. Meta-analysis of the effi- cacy and safety of clopidogrel plus aspirin as compared to antiplate- let monotherapy for the prevention of vascular events. Am J Cardiol. 2008;101:960–966.

Calabro P, Willerson JT, Yeh ET. Inflammatory cytokines stimulated C-reactive protein production by human coronary artery smooth muscle cells. Circulation. 2003;108:1930 –1932.

Campbell BC, Christensen S, Butcher KS, Gordon I, Parsons MW, Desmond PM, Barber PA, Levi CR, Bladin CF, De Silva DA, Donnan GA, Davis SM; EPITHET Investigators. Regional very low cerebral blood volume predicts hemorrhagic transformation better than diffusion-weighted imaging volume and thresholded apparent diffusion coefficient in acute ischemic stroke. Stroke. 2010 Jan;41(1):82-8. doi: 10.1161/STROKEAHA.109.562116.

Campbell BCV, Mitchell PJ, Kleinig TJ, *et al.* Endovascular therapy for ischemic stroke with perfusion-imaging selection. N Engl J Med 2015; 372: 1009–1018.

Cancer Genome Atlas Network, "Comprehensive molecular portraits of human breast tumours," Nature, vol. 490, no. 7418, pp. 61–70, 2012.

Cannon CP, Blazing MA, Giugliano RP, McCagg A, *et al.* Ezetimibe added to statin therapy after acute coronary syndromes. New England Journal of Medicine. 2015;372(25):2387- 2397. ---- Sabatine MS, Giugliano RP, Wiviott SD, Raal FJ, *et al.* Efficacy and safety of evolocumab in reducing lipids and cardiovascular events. The New England Journal of Medicine. 2015;372(16):1500-1509.

Centers for Disease Control and Prevention. Underlying Cause of Death, 1999–2018. CDC WONDER Online Database. Atlanta, GA: Centers for Disease Control and Prevention; 2018. Accessed March 12, 2020.

Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin). 2012 Apr-Jun;6(2):80-92.

Chan JA, Krichevsky AM & Kosik KS (2005) MicroR- NA-21 is an antiapoptotic factor in human glioblas- toma cells. Cancer Res 65, 6029–6033.

Chasman DI. New pathway for tissue-type plasminogen activator regu- lation. Arterioscler Thromb Vasc Biol. 2014;34:964–965. doi: 10.1161/ ATVBAHA.114.303499.

Chen Y-C, Huang AL, Kyaw TS, Bobik A, Peter K. Atherosclerotic plaque rupture: Identifying the straw that breaks the Camel's back. Arteriosclerosis, Thrombosis, and Vascular Biology. 2016;36:e63-e72.

Cheng L.C., Pastrana E., Tavazoie M., Doetsch F. miR-124 regulates adult neurogenesis in the subventricular zone stem cell niche. Nat. Neurosci. 2009;12:399–408.

Choy TK, Wang CY, Phan NN, Khoa Ta HD, Anuraga G, Liu YH, Wu YF, Lee KH, Chuang JY, Kao TJ. Identification of Dipeptidyl Peptidase (DPP) Family Genes in Clinical Breast Cancer Patients via an Integrated Bioinformatics Approach. Diagnostics (Basel). 2021 Jul 2;11(7):1204.

Christoforou A, Mulvey CM, Breckels LM, Geladaki A, Hurrell T, Hayward PC, *et al.* A draft map of the mouse pluripotent stem cell spatial proteome. Nat Commun. 2016;7:1–12.

Chu SG, Becker RC, Berger PB, Bhatt DL, Eikelboom JW, Konkle B, Mohler ER, Reilly MP, Berger JS. Mean platelet volume as a predictor of cardiovascular risk: A systematic review and meta-analysis. Journal of Thrombosis and Haemostasis. 2009;8:148-156.

du Clos TW. Function of C-reactive protein. Ann Med. 2000;32:274 –278.

Cochain C, Zernecke A. Macrophages in vascular inflammation and atherosclerosis. European Journal of Physiology. 2017;469:485-499.

Cohen A, Amarenco P. Atherosclerosis of the thoracic aorta: from risk stratification to treatment. Am J Cardiol 2002;90:1333–5.

Constam DB, Tobler AR, Rensing-Ehl A, Kemler I, Hersh LB & Fontana A (1995) Puromycin-sensitive aminopeptidase. Sequence analysis, expression, and functional characterization. J Biol Chem 270, 26931– 26939.

Corti R, Fuster V, Badimon JJ *et al.* New understanding of atherosclerosis (clinically and experimentally) with evolving MRI technology in vivo. Ann NY Acad Sci 2001;947:181–95, discussion 195–8.

Corti R, Fuster V, Badimon JJ. Pathogenetic concepts of acute coronary syndromes. J Am Coll Cardiol 2003;41:7S–14S.

Crisby M, Kallin B, Thyberg J, Zhivotovsky B, Orrenius S, Kostulas V, Nilsson J. Cell death in human atherosclerotic plaques involves both oncosis and apoptosis. Atherosclerosis. 1997;130:17-27.

Cruz-Diaz N, Wilson BA, Pirro NT, Brosnihan KB, Marshall AC, Chappell MC. Identification of dipeptidyl peptidase 3 as the angiotensin-(1–7) degrading peptidase in human HK-2 renal epithel cells. Peptides. 2016;83:29–37.

Dale CS, Pagano Rde L & Rioli V (2005) Hemopres- sin: a novel bioactive peptide derived from the alpha1- chain of hemoglobin. Mem Inst Oswaldo Cruz 100, 105–106.

Dangas G, Badimon JJ, Smith DA *et al.* Pravastatin therapy in hyperlipidemia: effects on thrombus formation and the systemic hemostatic profile. J Am Coll Cardiol 1999;33:1294–304.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R; 1000 Genomes Project Analysis Group. The variant call format and VCFtools. Bioinformatics. 2011 Aug 1;27(15):2156-8.

Dans AL, Connolly SJ, Wallentin L, Yang S, Nakamya J, Brueckmann M, Ezekowitz M, Oldgren J, Eikelboom JW, Reilly PA, Yusuf S. Concomitant use of antiplatelet therapy with dabigatran or warfarin in the randomized evaluation of long-term anticoagulation therapy (RE-LY) trial. Circulation. 2013;127:634–640.

Dargazanli C, Rigau V, Eker O, Riquelme Bareiro C, Machi P, Gascou G, Arquizan C, Ayrignac X, Mourand I, Corlobé A, Lobotesis K, Molinari N, Costes V, Bonafé A, Costalat V. High CD3+ Cells in Intracranial Thrombi Represent a Biomarker of Atherothrombotic Stroke. PLoS One. 2016 May 6;11(5):e0154945. doi: 10.1371/journal.pone.0154945.

Davies MJ, Woolf N. Atherosclerosis: what is it and why does it occur. Br Heart J 1993;69:S3–11.

Day-Williams AG, Zeggini E. The effect of next-generation sequencing technology on complex trait research. Eur J Clin Invest. 2011 May;41(5):561-7.

Deb P, Sharma S, Hassan KM. Pathophysiologic mechanisms of acute ischemic stroke: An overview with emphasis on therapeutic significance beyond thrombolysis. Pathophysiology. 2010 Jun;17(3):197-218.

Dehghan M, Mente A, Teo KK, *et al.* Relationship between healthy diet and risk of cardi-ovascular disease among patients on drug therapies for secondary prevention: A prospec-tive cohort study of 31 546 highrisk individuals from 40 countries. Circulation. 2012;126:2705-2712.

Del Zoppo G.J. 2010. The neurovascular unit, matrix proteases, and innate inflammation. Ann N Y Acad Sci. 1207:46-49.

Deloukas P, Kanoni S, et al; Consortium CAD. Large-scale association analysis identifies new risk loci for coronary artery disease. Nat Genet. 2013;45:25–33.

Deniau B, Rehfeld L, Santos K, Dienelt A, Azibani F, Sadoune M, *et al.* Circulating dipeptidyl peptidase 3 is a myocardial depressant factor: dipeptidyl peptidase 3 inhibition rapidly and sustainably improves haemodynamics Eur J Heart Fail. 2019;22:290–9.

Deniau B, Blet A, Santos K, Ayar PV, Genest M, Kastorf M, *et al.* Inhibition of circulating dipeptidylpeptidase 3 restores cardiac function in a sepsis induced model in rats: a proof of concept study. PLoS One. 2020;15:1–12.

Depret F, Amzallag J, Pollina A, Fayolle-Pivot L, Coutrot M, Chaussard M, *et al.* Circulating dipeptidyl peptidase-3 at admission is associated with circulatory failure, acute kidney injury and death in severely ill burn patients. Crit Care. 2020;24:1–8.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 2011;43:491– 8.

Dharap A., Bowen K., Place R., Li L.C., Vemuganti R. Transient focal ischemia induces extensive temporal changes in rat cerebral microRNAome. J. Cereb. Blood Flow Metab. 2009;29:675–687.

Dichgans M, Pulit SL, Rosand J. Stroke genetics: discovery, biology, and clinical applications. Lancet Neurol. 2019 Jun;18(6):587-599.

Dimmeler S, Haendeler J, Galle J *et al.* Oxidized low-density lipoprotein induces apoptosis of human endothelial cells by activa- tion of CPP32-like proteases. A mechanistic clue to the 'response to injury' hypothesis. Circulation 1997;95:1760–3.

Dinh DT, Frauman AG, Johnston CI, Fabiani ME. Angiotensin receptors: distribution, signalling and function. Clin Sci. 2001;100:481–92.

Dirnagl U. Pathobiology of injury after stroke: the neurovascular unit and beyond. Ann N Y Acad Sci. 2012 Sep;1268:21-5.

Dobrovetsky E, Dong A, Seitova A, Duncan B, Crom- bet L, Sundstrom M, Arrowsmith CH, Edwards AM, Bountra C, Bochkarev A *et al.* (2009) Crystal structure of human dipeptidyl peptidase III. Structural Genomics Consortium (SGC) 2009, to be published.

Doeppner T.R., Doehring M., Bretschneider E., Zechariah A., Kaltwasser B., Müller B., Koch J.C., Bähr M., Hermann D.M., Michel U. MicroRNA-124 protects against focal cerebral ischemia via mechanisms involving Usp14-dependent REST degradation. Acta Neuropathol. 2013;126:251–265.

Do R, Willer CJ, Schmidt EM, *et al.* Common variants associated with plasma triglycerides and risk for coronary artery disease. Nat Genet. 2013;45:1345–1352.

Do R, Stitziel NO, Won HH, et al; NHLBI Exome Sequencing Project. Exome sequencing identifies rare LDLR and APOA5 alleles conferring risk for myocardial infarction. Nature. 2015;518:102–106.

Doran AC, Meller N, McNamara CA. Role of smooth muscle cells in the initiation and early progression of atherosclerosis. Arteriosclerosis, Thrombosis, and Vascular Biology. 2008;28:812-819.

Drouet, L. Atherothrombosis as a Systemic Disease. Cerebrovasc Dis vol. 13 www.karger.com (2002).

Dudoit S, Shaffer J P,and Boldrick J C. Multiple hypothesis testing in microarray experiments. Statistical Science, 18:71-103, 2003.

Dunn P, Albury CL, Maksemous N, Benton MC, Sutherland HG, Smith RA, Haupt LM, Griffiths LR. Next Generation Sequencing Methods for Diagnosis of Epilepsy Syndromes. Front Genet. 2018 Feb 7;9:20. doi: 10.3389/fgene.2018.00020. PMID: 29467791; PMCID: PMC5808353.

Dziedzic, E., Machowski, M., Oleszczak-Kostyra, M. & Dąbrowski, M. J. Athero-thrombosis as a Leading Cause of Acute Coronary Syndromes and Stroke: The Main Killers in Developed Countries. in Atherosclerosis - Yesterday, Today and Tomorrow (InTech, 2018). doi:10.5772/intechopen.71786.

Easton JD. Future perspectives for optimizing oral antiplatelet therapy. Cerebrovasc Dis. 2001;11 Suppl 2:23-8.

Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A. 1998 Dec 8;95(25):14863-8.

Ellis S, Nuenke JM. Dipeptidyl arylamidase III of the pituitary. Purification and characterization. J Biol Chem. 1967 Oct 25;242(20):4623-9.

ENCODE Project Consortium, 2004. The ENCODE (ENCyclopedia of DNA elements) project. Science 306 (5696), 636–640.

Escudero Augusto D, Marqués Alvarez L, Taboada Costa F. [Up-date in spontaneous cerebral hemorrhage]. Med Intensiva. 2008 Aug-Sep;32(6):282-95.

Falk E, Nakano M, Bentzon JF, Finn AV, Virmani R. Update on acute coronary syn- dromes: The pathologists' view. European Heart Journal. 2013;34(10):719-728.

Fayad ZA, Nahar T, Fallon JT *et al.* In vivo magnetic resonance evaluation of atherosclerotic plaques in the human thoracic aorta: a comparison with transesophageal echocardiography. Circulation 2000;101:2503–9.

Ferragina P, Manzini G. (2000, November). Opportunistic data structures with applications. In Proceedings 41st annual symposium on foundations of computer science (pp. 390-398).

Ferrer I, Garcia-Esparcia P, Carmona M, Carro E, Aronica E, Kovacs GG, Grison A, Gustincich S. Olfactory Receptors in Non-Chemosensory Organs: The Nervous System in Health and Disease. Front Aging Neurosci. 2016 Jul 5;8:163. doi: 10.3389/fnagi.2016.00163.

Ferro ES, Hyslop S & Camargo AC (2004) Intracellul- lar peptides as putative natural regulators of protein interactions. J Neurochem 91, 769–777.

Finucane MM, Stevens GA, Cowan MJ, *et al.* National, regional, and global trends in body mass index since 1980: Systematic analysis of health examination surveys and epidemiological studies with 960 country-years and 9.1 million participants. Lancet. 2011;377:557-567.

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., Merrick, J.M., 1995. Whole-genome random sequencing and assembly of haemophilus Influenzae Rd. Science 269 (5223), 496–512.

Flego D, Liuzzo G, Weyand CM, Crea F. Adaptive immunity dysregulation in acute cor- onary syndromes from cellular and molecular basis to clinical implications. Journal of the American College of Cardiology. 2016;68:2107-2117.

Flobmann E, Schulz UGR, Rothwell PM. Systematic review of methods and results of studies of the genetic epidemiology of ischaemic stroke. Stroke 2004;35:212–27.

Forgetta V, Jiang L, Vulpescu NA, Hogan MS, Chen S, Morris JA, Grinek S, Benner C, Jang DK, Hoang Q, Burtt N, Flannick JA, McCarthy MI, Fauman E, Greenwood CMT, Maurano MT, Richards JB. An effector index to predict target genes at GWAS loci. Hum Genet. 2022 Feb 11.

Forrester SJ, Booz GW, Sigmund CD, Coffman TM, Kawai T, Rizzo V, *et al.* Angiotensin II signal transduction: An update on mechanisms of physiology and pathophysiology. Physiol Rev. 2018;98:1627–738.

Francis J, Raghunathan S, Khanna P. The role of genetics in stroke. Postgrad Med J. 2007 Sep;83(983):590-5.

Fraser JF, Collier LA, Gorman AA, Martha SR, Salmeron KE, Trout AL, Edwards DN, Davis SM, Lukins DE, Alhajeri A, Grupke S, Roberts JM, Bix GJ, Pennypacker KR. The Blood And Clot Thrombectomy Registry And Collaboration (BACTRAC) protocol: novel method for evaluating human stroke. J Neurointerv Surg. 2019 Mar;11(3):265-270. doi: 10.1136/neurintsurg-2018-014118.

Fridley, B.L., Lund, S., Jenkins, G.D., Wang, L., 2012. A Bayesian integrative genomic model for pathway analysis of complex traits. Genet. Epidemiol. 36 (4), 352–359.

Friese RS, Rao F, Khandrika S, *et al.* Matrix metalloproteinases: Discrete elevations in essential hypertension and hypertensive end-stage renal disease. Clinical and Experimental Hypertension. 2009;31:521-533.

Fukasawa KM, Fukasawa K & Harda M (2000) Assignment of the dipeptidyl peptidase III gene (DPP3) to human chromosome 11 band q12 fi q13.1 by in situ hybridization. Cytogenet Cell Genet 88, 99–100.

Fuster, V., Badimon, J. J. & Chesebro, J. H. Atherothrombosis: mechanisms and clinical therapeutic approaches. Vascular Medicine vol. 3 (1998).

Fuster V. Epidemic of cardiovascular disease and stroke: the three main challenges. Presented at the 71st scientific sessions of the American Heart Association. Dallas, Texas. Circulation 1999;99: 1132–7.

Fuster V, Moreno PR, Fayad ZA, Corti R, Badimon JJ. Atherothrombosis and high- risk plaque. Part I: Evolving concepts. Journal of the American College of Cardiology. 2005;46:937-954.

Fyhrquist F, Saijonmaa O. Renin-angiotensin system revisited. J Intern Med. 2008;264:224–36.

Gao P, Rong H-H, Lu T, Tang G, Si L-Y, Lederer JA, Xiong W. The CD4/CD8 ratio is associated with coronary artery disease (CAD) in elderly Chinese patients. International Immunopharmacology. 2017;42:39-43.

GenomeAsia100K Consortium. The GenomeAsia 100K Project enables genetic discoveries across Asia. Nature. 2019 Dec;576(7785):106-111.

Gentleman R, Carey V, Huber W, Irizarry R, and Dudoit S. Bioinformatics and Computational Biology Solutions using R and Bioconductor. Springer, New York, 2005.

George, E.I., McCulloch, R.E., 1993. Variable selection via Gibbs sampling. J. Am. Stat. Assoc. 88 (423), 881–889.

George PM, Steinberg GK. Novel Stroke Therapeutics: Unraveling Stroke Pathophysiology and Its Impact on Clinical Treatments. Neuron. 2015 Jul 15;87(2):297-309. doi: 10.1016/j.neuron.2015.05.041

Giesen PL, Rauch U, Bohrmann B *et al.* Blood-borne tissue factor: another view of thrombosis. Proc Natl Acad Sci USA 1999;96:2311–5.

Giusti B, Rossi L, Lapini I, Magi A, Pratesi G, Lavitrano M, Biasi GM, Pulli R, Pratesi C, Abbate R. Gene expression profiling of peripheral blood in patients with abdominal aortic aneurysm. Eur J Vasc Endovasc Surg. 2009 Jul;38(1):104-12.

Glagov S, Weisenberg E, Zarins CK *et al.* Compensatory enlargement of human atherosclerotic coronary arteries. N Engl J Med 1987;316:1371–5.

Glagov S, Zarins C, Giddens DP *et al.* Hemodynamics and athero- sclerosis. Insights and perspectives gained from studies of human arteries. Arch Pathol Lab Med 1988;112:1018–31.

Goldstein JL, Brown MS. The LDL re- ceptor. Arterioscler Thromb Vasc Biol 2009; 29:431-8.

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., *et al.*, 1996. Life with 6000 genes. Science 274 (5287), 546. 563–67.

Goh SY & Cooper ME (2008) Clinical review: the role of advanced glycation end products in progression and complications of diabetes. J Clin Endocrinol Metab 93, 1143–1152.

González RG. Imaging-guided acute ischemic stroke therapy: From "time is brain" to "physiology is brain". AJNR Am J Neuroradiol. 2006 Apr;27(4):728-35. Review. PubMed PMID: 16611754.

Goyal M, Demchuk AM, Menon BK, Eesa M, Rempel JL, Thornton J, *et al.* Randomized assessment of rapid endovascular treatment of ischemic stroke. N Engl J Med. 2015;372(11):1019– 30.

Goyal M, Menon BK, van Zwam WH, *et al.* HERMES collaborators. Endovascular thrombectomy after large-vessel ischaemic stroke: a meta-analysis of individual patient data from five randomised trials. Lancet 2016;387:1723–31.

Gretarsdottir S, Sveinbjornsdottir S, Johnsson HH *et al.* Localization of a suscetibility gene for common forms of stroke to 5q12. Am J Hum Genet. 2002;70:593-603.

Gretarsdottir S, Thorleifsson G, Manolescu A *et al.*: Risk variants for atrial fibrillation on chromosome 4q25 associate with ischemic stroke. Ann Neurol 2008, 64:402-409.

Groom JR, Richmond J, Murooka TT, Sorensen EW, Sung JH, Bankert K, von Andrian UH, Moon JJ, Mempel TR, Luster AD. CXCR3 chemokine receptor-ligand interactions in the lymph node optimize CD4+ T helper 1 cell differentiation. Immunity 2012;37(6):1091- 1103.

Gu, K., Cowie, C. C. & Harris, M. I. M o r ta li ty in Adu lts With and W i t h o u t Diabetes in a National Cohort of the U.S. Population, 1971–1993. DIABETES CA-RE 21, 1138–1145 (1998).

Gullapalli RR, Lyons-Weiler M, Petrosko P, Dhir R, Be cich MJ, LaFramboise WA. Clinical integration of next generation sequencing technology. Clin Lab Med 2012;32:585.

Guo JM, Liu AJ, Su DF. Genetics of stroke. Acta Pharmacol Sin. 2010 Sep;31(9):1055-64.

Gudbjartsson DF, Holm H, Gretarsdottir S *et al.*: A sequence variant in ZFHX3 on 16q22 associates with atrial fibrillation and ischemic stroke. Nat Genet 2009, 41:876-878.

Guttmacher, A. E., Collins, F. S. & Nabel, E. G. genomic medicine Cardiovascular Disease. www.nejm.org (2003).

Hacke W, Albers G, Al-Rawi Y, Bogousslavsky J, Davalos A, Eliasziw M, Fischer M, Furlan A, Kaste M, Lees KR, Soehngen M, Warach S; DIAS Study Group. The Desmoteplase in Acute Ischemic Stroke Trial (DIAS): a phase II MRI-based 9-hour window acute stroke thrombolysis trial with intravenous desmoteplase. Stroke. 2005 Jan;36(1):66-73. doi: 10.1161/01.STR.0000149938.08731.2c. Epub 2004 Nov 29. PMID: 15569863.

Hamzei Taj S., Kho W., Riou A., Wiedermann D., Hoehn M. MiRNA-124 induces neuroprotection and functional improvement after focal cerebral ischemia. Biomaterials. 2016;91:151–165.

Han S, Liu P, Zhang W, Bu L, Shen M, Li H, Fan Y-H, Cheng K, Cheng H-X, Li C-X, G-l J. The opposite-direction modulation of CD4+CD25+ Tregs and T helper 1 cells in acute coronary syndromes. Clinical Immunology. 2007;124:90-97.

Hansson GK, Libby P, Tabas I. Inflammation and plaque vulnerability. Journal of Internal Medicine. 2015;278(5):483-493.

Harston GW, Sutherland BA, Kennedy J, Buchan AM. The contribution of L-arginine to the neurotoxicity of recombinant tissue plasminogen activator following cerebral ischemia: a review of rtPA neurotoxicity. J Cereb Blood Flow Metab. 2010 Nov;30(11):1804-16. doi: 10.1038/jcbfm.2010.149.

Hashimoto J, Yamamoto Y, Kurosawa H, Nishimura K, Hazato T. Identification of dipeptidyl peptidase III in human neutrophils. Biochem Biophys Res Commun. 2000 Jul 5;273(2):393-7.

Hawkins BT, Davis TP. The blood-brain barrier/neurovascular unit in health and disease. Pharmacol Rev. 2005 Jun;57(2):173-85.

Helgadottir A, Thorleifsson G, Manolescu A *et al.*: A common variant on chromosome 9p21 affects the risk of myocardial infarction. Science 2007, 316:1491-3.

Heydari E, Alishahi M, Ghaedrahmati F, Winlow W, Khoshnam SE, Anbiyaiee A. The role of non-coding RNAs in neuroprotection and angiogenesis following ischemic stroke. Metab Brain Dis. 2019 Aug 24.

Hill AA, Brown EL, Whitley MZ, Tucker-Kellogg G, Hunter CP, Slonim DK. Evaluation of normalization procedures for oligonucleotide array data based on spiked cRNA controls. Genome Biol. 2001;2(12):RESEARCH0055.

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A. 2009 Jun 9;106(23):9362-7.

Hiraga A. Gender Differences and Stroke Outcomes. Neuroepidemiology. 2017;48(1-2):61-62. doi: 10.1159/000475451.

Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits. Nat Rev Genet 2005;6:95–108.

Hom J, Dankbaar JW, Soares BP, Schneider T, Cheng SC, Bredno J, Lau BC, Smith W, Dillon WP, Wintermark M. Blood-brain barrier permeability assessed by perfusion CT predicts symptomatic hemorrhagic transformation and malignant edema in acute ischemic stroke. AJNR Am J Neuroradiol. 2011 Jan;32(1):41-8. doi: 10.3174/ajnr. A2244.

Horváth E, Huțanu A, Chiriac L, Dobreanu M, Orădan A, Nagy EE. Ischemic damage and early inflammatory infiltration are different in the core and penumbra lesions of rat brain after transient focal cerebral ischemia. J Neuroimmunol. 2018 Nov 15;324:35-42.

Huang, R.S., Duan, S., Bleibel, W.K., Kistner, E.O., Zhang, W., Clark, T.A., Chen, T.X., *et al.*, 2007. A genome-wide approach to identify genetic variants that contribute to Etoposide- induced cytotoxicity. Proc. Natl. Acad. Sci. U. S. A. 104 (23), 9758–9763.

Huang, R.S., Duan, S., Kistner, E.O., Hartford, C.M., Eileen Dolan, M., 2008. Genetic variants associated with carboplatin-induced cytotoxicity in cell lines derived from Africans. Mol. Cancer Ther. 7 (9), 3038–3046.

Huang J, Huffman JE, Yamakuchi M, et al; Cohorts for Heart and Aging Research in Genome Epidemiology (CHARGE) Consortium Neurology Working Group; CARDIoGRAM Consortium; CHARGE Consortium Hemostatic Factor Working Group. Genome-wide association study for circulating tissue plasminogen activator levels and functional follow-up implicates endothelial STXBP5 and STX2. Arterioscler Thromb Vasc Biol. 2014;34:1093–1101.

Huber W, von Heydebreck A, Sültmann H, Poustka A, Vingron M. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics. 2002;18 Suppl 1:S96-104.

Iborra-Egea O, Montero S, Bayes-Genis A. An outlook on biomarkers in cardiogenic shock. Curr Opin Crit Care. 2020 Aug;26(4):392-397.

International Human Genome Sequencing Consortium, 2004. Finishing the Euchromatic sequence of the human genome. Nature 431 (7011), 931–945.

Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics. 2003 Apr;4(2):249-64.

IST-3 collaborative group, Sandercock P, Wardlaw JM, Lindley RI, Dennis M, Cohen G, Murray G, Innes K, Venables G, Czlonkowska A, Kobayashi A, Ricci S, Murray V, Berge E, Slot KB, Hankey GJ, Correia M, Peeters A, Matz K, Lyrer P, Gubitz G, Phillips SJ, Arauz A. The benefits and harms of intravenous thrombolysis with recombinant tissue plasminogen activator within 6 h of acute ischaemic stroke (the third international stroke trial [IST-3]): a randomized controlled trial. Lancet. 2012 Jun 23;379(9834):2352-63. doi:10.1016/S0140-6736(12)60768-5. Epub 2012 May 23. Erratum in: Lancet. 2012 Aug 25;380(9843):730. PubMed PMID: 22632908; PubMed Central PMCID: PMC3386495.

Jabs WJ, Theissing E, Nitschke M, *et al.* Local generation of C-reactive protein in diseased coronary artery venous bypass grafts and normal vascular tissue. Circulation. 2003;108:1428 –1431.

Jauch, E.C.; Saver, J.L.; Adams, H.P., Jr.; Bruno, A.; Connors, J.J.; Demaerschalk, B.M.; Khatri, P.;McMullan, P.W., Jr.; Qureshi, A.I.; Rosenfield, K.;*et al.*Guidelines for the early management ofpatients with acute ischemic stroke: A guideline for healthcare professionals from the American Heart-Association/American Stroke Association.Stroke2013,44, 870–947

Jawień J. New immunological look at the pathogenesis of atherosclerosis. Polish Archives of Internal Medicine. 2008;118:127-13.

(a) Jeffreys, A.J., Brookfield, J.F., Semeonoff, R., 1985a. Positive identification of an immigration test-case using human DNA fingerprints. Nature 317 (6040), 818–819.

(b) Jeffreys, A.J., Wilson, V., Thein, S.L., 1985b. Hypervariable 'minisatellite' regions in human DNA. Nature 314 (6006), 67–73.

Jeyaseelan K., Lim K.Y., Armugam A. MicroRNA expression in the blood and brain of rats subjected to transient focal ischemia by middle cerebral artery occlusion. Stroke. 2008;39:959–966.

Ji P, Diederichs S, Wang W, Boing S, Metzger R, Schneider PM, Tidow N, Brandt B, Buerger H, Bulk E*et al.* (2003) MALAT-1, a novel noncoding RNA, and thymosin b4 predict metastasis and survival in early-stage non-small cell lung cancer. Oncogene 22, 8031–8041.

Jian Z, Liu R, Zhu X, Smerin D, Zhong Y, Gu L, Fang W, Xiong X. The Involvement and Therapy Target of Immune Cells After Ischemic Stroke. Front Immunol. 2019 Sep 11;10:2167.

Jha R, Battey TW, Pham L, Lorenzano S, Furie KL, Sheth KN, Kimberly WT. Fluid-attenuated inversion recovery hyperintensity correlates with matrix metalloproteinase-9 level and hemorrhagic transformation in acute ischemic stroke. Stroke. 2014 Apr;45(4):1040-5. doi: 10.1161/STROKEAHA.113.004627.

(b) Jha S, Taschler U, Domenig O, Poglitsch M, Bourgeois B, Pollheimer M, *et al.* Dipeptidyl peptidase 3 modulates the renin–angiotensin system in mice. J Biol Chem. 2020;295:13711–23.

Jørgensen AB, Frikke-Schmidt R, Nordestgaard BG, Tybjærg- Hansen A. Loss-of-function mutations in APOC3 and risk of isch- emic vascular disease. N Engl J Med. 2014;371:32–41.

Johnson, W.B., Lindenstrauss, J., 1984. Extensions of Lipschitz mappings into a Hilbert space. Contemp. Math. 26 (189–206), 1.

Jousilahti P, Rastenyte D, Tuomilehto J, *et al.* Parental history of cardiovascular disease and risk of stroke. A prospective follow-up of 14371 middle-aged men and women in Finland. Stroke 1997;28:1361–6.

Jovin TG, Chamorro A, Cobo E, *et al.* Thrombectomy within 8 hours after symptom onset in ischemic stroke. N Engl J Med 2015; 372: 2296–2306.

Kaikita K, Ogawa H, Yasue H *et al.* Tissue factor expression on macrophages in coronary plaques in patients with unstable angina. Arterioscler Thromb Vasc Biol 1997;17:2232–7.

Kaluza D, Kroll J, Gesierich S, *et al.* Histone deacetylase 9 promotes angiogenesis by targeting the anti-angiogenic microRNA-17–92 cluster in endothelial cells. Arterioscler Thromb Vasc Biol 2013;33:533–543.

Kannel WB, Wolf PA, Castelli WP, *et al.* Fibrinogen and risk of cardiovascular disease. The Framingham Study. JAMA. 1987;258: 1183–1186.

Kaplan ZS, Jackson SP. The role of platelets in atherothrombosis. Hematology. American Society of Hematology. Education Program. 2011;2011:51-61.

Karczewski, K.J., Fran-cioli, L.C., Tiao, G. *et al.* The mutational constraint spectrum quantified from varia-tion in 141,456 humans. Nature 581, 434–443 (2020).

Kaufman L and Rousseeuw P J. Finding Groups in Data. An Introduction to Cluster Analysis. John Wiley and Sons, 1990.

Kaur J, Zhao Z, Klein GM, Lo EH, Buchan AM. The neurotox¬icity of tissue plasminogen activator? J Cereb Blood Flow Metab 2004;24:945-963.

Khatri P, Drăghici S. Ontological analysis of gene expression data: current tools, limitations, and open problems. Bioinformatics. 2005 Sep 15;21(18):3587-95.

Khera AV, Kathiresan S. Genetics of coronary artery disease: discovery, biology and clinical translation. Nat Rev Genet 2017; 18: 331–44.

Kim, D., Shin, H., Song, Y.S., Kim, J.H., 2012. Synergistic effect of different levels of genomic data for Cancer clinical outcome prediction. J. Biomed. Inform. 45 (6), 1191–1198.

Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014 Mar;46(3):310-5. doi: 10.1038/ng.2892. Epub 2014 Feb 2.

Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, Weinstock GM, Wilson RK, Ding L. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics. 2009 Sep 1;25(17):2283-5.

Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson RK. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 2012 Mar;22(3):568-76.

Kolodgie FD, Gold HK, Burke AP, Fowler DR, Kruth HS, Weber DK, Farb A, Guerrero LJ, Hayase M, Kutys R, Narula J, Finn AV, Virmani R. Intraplaque hemorrhage and progres- sion of coronary atheroma. The New England Journal of Medicine. 2003;349:2316-2325.

Korninger C, Collen D. Studies on the specific fibrinolytic effect of human extrinsic (tissue-type) plasminogen activator in human blood and in various animal species in vitro. Thromb Haemost. 1981;46.

Kumamoto M, Nakashima Y, Sueishi K. Intimal neovascularization in human coro- nary atherosclerosis: Its origin and pathophysiological significance. Human Pathology. 1995;26:450-456.

Kumar G, Goyal MK, Sahota PK, Jain R (2010) Penumbra, the basis of neuroimaging in acute stroke treatment: current evidence. J Neurol Sci 288(1):13–24.

Kyriakou T, Seedorf U, Goel A, *et al.* A common LPA null allele asso- ciates with lower lipoprotein(a) levels and coronary artery disease risk. Arterioscler Thromb Vasc Biol. 2014;34:2095–2099.

LaDuca H, Farwell KD, Vuong H, Lu HM, Mu W, Shahmirzadi L, Tang S, Chen J, Bhide S, Chao EC. Exome sequencing covers >98% of mutations identified on targeted next generation sequencing panels. PLoS One. 2017 Feb 2;12(2):e0170843. doi: 10.1371/journal.pone.0170843. PMID: 28152038; PMCID: PMC5289469.

Lam CW, Mak CM. Allele dropout caused by a non-primer-site SNV affecting PCR amplification--a call for next-generation primer design algorithm. Clin Chim Acta. 2013 Jun 5;421:208-12. doi: 10.1016/j.cca.2013.03.014. Epub 2013 Mar 21. PMID: 23523590.

Lander ES. Initial impact of the sequencing of the human genome. Nature. 2011 Feb 10;470(7333):187-97.

Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10(3):R25.

Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012 Mar 4;9(4):357-9.

Langsted A, Freiberg JJ, Tybjaerg-Hansen A, Schnohr P, Jensen GB, Nordestgaard BG. Nonfasting cholesterol and triglycerides and associa- tion with risk of myocardial infarction and total mortality: the Copenhagen City Heart Study with 31 years of follow-up. J Intern Med. 2011;270:65– 75.

Lansberg MG, O'Donnell MJ, Khatri P, Lang ES, Nguyen-Huynh MN, Schwartz NE, Sonnenberg FA, Schulman S, Vandvik PO, Spencer FA, Alonso-Coello P, Guyatt GH, Akl EA. Antithrombotic and thrombolytic therapy for ischemic stroke: Antithrombotic Therapy and Prevention of Thrombosis, 9th ed: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines. Chest. 2012 Feb;141(2 Suppl):e601S-e636S. doi: 10.1378/chest.11-2302.

Lee CM & Snyder SH (1982) Dipeptidyl-aminopepti- dase III of rat brain. Selective affinity for enkephalin and angiotensin. J Biol Chem 257, 12043–12050.

Leak RK, Zheng P, Ji X, Zhang JH, Chen J. From apoplexy to stroke: historical perspectives and new research frontiers. Prog Neurobiol. 2014 Apr;115:1-5.

Ledue TB, Rifai N. Preanalytic and analytic sources of variations in C-reactive protein measurement: implications for cardiovascular disease risk assessment. Clin Chem. 2003;49:1258 –1271.

Lee CM, Snyder SH. Dipeptidyl-aminopeptidase III of rat brain. Selective affinity for enkephalin and angiotensin. J Biol Chem. 1982;257:12043–50.

Lee DS, Pencina MJ, Benjamin EJ, *et al.* Association of parental heart failure with risk of heart failure in offspring. N Engl J Med 2006;355:138-47.

Leist M, Gantner F, K€unstle G, Bohlinger I, Tiegs G, Bluethmann H, *et al.* The 55-kD tumor necrosis factor receptor and CD95 independently signal murine hepatocyte apoptosis and subsequent liver failure. Mol Med. 1996;2:109–24.

Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. 2008 Nov;18(11):1851-8.

Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009 Jul 15;25(14):1754-60.

(b)Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078-9.

Li R, Li Y, Kristiansen K, Wang J. SOAP: short oligonucleotide alignment program. Bioinformatics. 2008 Mar 1;24(5):713-4.

Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, Wang J. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics. 2009 Aug 1;25(15):1966-7.

Li J, Ley K. Lymphocyte migration into atherosclerotic plaque. Arteriosclerosis, Thrombosis, and Vascular Biology. 2015;35(1):40-49.

Li Y, Wang X, Vural S, Mishra NK, Cowan KH, Guda C. Exome analysis reveals differentially mutated gene signatures of stage, grade and subtype in breast cancers. PLoS One. 2015 Mar 24;10(3):e0119383.

Lindahl B, Toss H, Siegbahn A, *et al.* Markers of myocardial damage and inflammation in relation to long-term mortality in unstable coronary artery disease. FRISC Study Group. N Engl J Med. 2000;343: 1139 –1147.

Lin K, Zink WE, Tsiouris AJ, John M, Tekchandani L, Sanelli PC. Risk assessment of hemorrhagic transformation of acute middle cerebral artery stroke using multimodal CT. J Neuroimaging. 2012 Apr;22(2):160-6. doi: 10.1111/j.1552-6569.2010.00562.x.

Lindmark E, Diderholm E, Wallentin L, *et al.* Relationship between interleukin 6 and mortality in patients with unstable coronary artery disease: effects of an early invasive or noninvasive strategy. JAMA. 2001;286:2107–2113.

Lipshutz RJ, Fodor SP, Gingeras TR, Lockhart DJ. High density synthetic oligonucleotide arrays. Nat Genet. 1999 Jan;21(1 Suppl):20-4. Liu K, Daviglus ML, Loria CM, *et al.* Healthy lifestyle through young adulthood and the presence of low cardiovascular disease risk profile in middle age: The coronary artery risk development in young adults (CARDIA) study. Circulation. 2012;125:996-1004.Lozano, R. *et al.* Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: A systematic analysis for the Global Burden of Di-sease Study 2010. The Lancet 380, 2095–2128 (2012).

Liu QR, Walther D, Drgon T, Polesskaya O, Lesnick TG, Strain KJ, de Andrade M, Bower JH, Maraganore DM, Uhl GR. Human brain derived neurotrophic factor (BDNF) genes, splicing patterns, and assessments of associations with substance abuse and Parkinson's Disease. Am J Med Genet B Neuropsychiatr Genet. 2005 Apr 5;134B(1):93-103.

Liu Y, Kern JT, Walker JR, Johnson JA, Schultz PG & Luesch H (2007) A genomic screen for activators of the antioxidant response element. Proc Natl Acad Sci 104, 5205–5210.

Liu CM, Wong T, Wu E, Luo R, Yiu SM, Li Y, Wang B, Yu C, Chu X, Zhao K, Li R, Lam TW. SOAP3: ultra-fast GPU-based parallel alignment tool for short reads. Bioinformatics. 2012 Mar 15;28(6):878-9.

Liu XS, Chopp M., Zhang R.L., Zhang Z.G. MicroRNAs in cerebral ischemia-induced neurogenesis. J. Neuropathol. Exp. Neurol. 2013;72:718–722.

Liuzzo G, Biasucci LM, Trotta G, Brugaletta S, Pinnelli M, Digianuario G, Rizzello V, Rebuzzi AG, Rumi C, Maseri A, Crea F. Unusual CD4_CD28null T lymphocytes and recurrence of acute coronary events. Journal of the American College of Cardiology. 2007;50:1450-1458.

Lu K, Alcivar AL, Ma J, Foo TK, Zywea S, Mahdi A, Huo Y, Kensler TW, Gatza ML, Xia B. NRF2 Induction Supporting Breast Cancer Cell Survival Is Enabled by Oxidative Stress-Induced DPP3-KEAP1 Interaction. Cancer Res. 2017 Jun 1;77(11):2881-2892.

Lu X, Peloso GM, Liu DJ, Wu Y, Zhang H, Zhou W, Li J, Tang CS, Dorajoo R, Li H, Long J, Guo X, Xu M, Spracklen CN, Chen Y, Liu X, Zhang Y, Khor CC, Liu J, Sun L, Wang L, Gao YT, Hu Y, Yu K, Wang Y, Cheung CYY, Wang F, Huang J, Fan Q, Cai Q, Chen S, Shi J, Yang X, Zhao W, Sheu WH, Cherny SS, He M, Feranil AB, Adair LS, Gordon-Larsen P, Du S, Varma R, Chen YI, Shu XO, Lam KSL, Wong TY, Ganesh SK, Mo Z, Hveem K, Fritsche LG, Nielsen JB, Tse HF, Huo Y, Cheng CY, Chen YE, Zheng W, Tai ES, Gao W, Lin X, Huang W, Abecasis G; GLGC Consortium, Kathiresan S, Mohlke KL, Wu T, Sham PC, Gu D, Willer CJ. Exome chip meta-analysis identifies novel loci and

East Asian-specific coding variants that contribute to lipid levels and coronary artery disease. Nat Genet. 2017 Dec;49(12):1722-1730.

Lucero MT. Peripheral modulation of smell: fact or fiction? Semin Cell Dev Biol. 2013 Jan;24(1):58-70. doi: 10.1016/j.semcdb.2012.09.001.

Mahmoudi M, Aslani S, Fadaei R, Jamshidi AR. New insights to the mechanisms underlying atherosclerosis in rheumatoid arthritis. International Journal of Rheumatic Diseases. 2017;20:287-297.

Malik R, Chauhan G, Traylor M, *et al.* Multiancestry genome-wide association study of 520 000 subjects identifies 32 loci associated with stroke and stroke subtypes. Nat Genet 2018; 50: 524–37.

Mallat Z, Tedgui A. Current perspective on the role of apoptosis in atherothrombotic disease. Circ Res 2001;88:998–1003.

Malone K, Amu S, Moore AC, Waeber C. Immunomodulatory Therapeutic Strategies in Stroke. Front Pharmacol. 2019; 10:630.

Maki T, Hayakawa K, Pham LD, Xing C, Lo EH, Arai K. Biphasic mechanisms of neurovascular unit injury and protection in CNS diseases. CNS Neurol Disord Drug Targets. 2013 May 1;12(3):302-15.

Mankoo, P.K., Shen, R., Schultz, N., Levine, D.A., Sander, C., 2011. Time to recurrence and survival in serous ovarian tumors predicted from integrated genomic profiles. PLoS One 6(11), e24709.

Marder VJ, Chute DJ, Starkman S, *et al.* Analysis of thrombi retrieved from cerebral arteries of patients with acute ischemic stroke. Stroke 2006; 37: 2086–2093.

Markus HS. Stroke genetics: prospects for personalized medicine. BMC Med. 2012 Sep 27;10:113.

Matsumura H, Shimizu Y, Ohsawa Y, Kawahara A, Uchiyama Y, Nagata S. Necrotic death pathway in Fas receptor signaling. J Cell Biol. 2000;151:1247–55.

Matthijs G, Souche E, Alders M, Corveleyn A, Eck S, Feenstra I, Race V, Sistermans E, Sturm M, Weiss M, Yntema H, Bakker E, Scheffer H, Bauer P; EuroGentest; European Society of Human Genetics. Guidelines for diagnostic next-generation sequencing. Eur J Hum Genet. 2016 Jan;24(1):2-5. doi: 10.1038/ejhg.2015.226. Epub 2015 Oct 28. Erratum in: Eur J Hum Genet. 2016 Oct;24(10):1515. PMID: 26508566; PMCID: PMC4795226.

Mazzocco C, Gillibert-Duplantier J, Neaud V, Fukasawa KM, Claverol S, Bonneu M, Puiroux J. Identification and characterization of two dipeptidyl-peptidase III isoforms in Drosophila melanogaster. FEBS J. 2006 Mar;273(5):1056-64.

Mazzocco C, Fukasawa KM, Raymond A-A, Puiroux J. Purification, partial sequencing and characterization of an insect membrane dipeptidyl aminopeptidase that degrades the insect neuropeptide proctolin. Eur J Biochem. 2001;268:4940–9.

McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 2008;9:356–69.

McCourt CM, McArt DG, Mills K, Catherwood MA, Maxwell P, Waugh DJ, Hamilton P, O'Sullivan JM, Salto-Tellez M. Validation of next generation sequencing technologies in comparison to current diagnostic gold standards for BRAF, EGFR and KRAS mutational analysis. PLoS One. 2013 Jul 26;8(7):e69604. doi: 10.1371/journal.pone.0069604. PMID: 23922754; PMCID: PMC3724913.

McDermott MM, Greenland P, Liu K *et al.* Leg symptoms in peripheral arterial disease: associated clinical characteristics and functional impairment. JAMA 2001;286:1599–606.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010 Sep;20(9):1297-303.

Meade TW, Imeson J, Stirling Y. Effects of changes in smoking and other characteristics on clotting factors and the risk of ischaemic heart disease. Lancet 1987;2:986–8.

Medina P, Navarro S, Bonet E, *et al.* Functional analysis of two haplotypes of the human endothelial protein C receptor gene. Arterioscler Thromb Vasc Biol. 2014;34:684–690.

Meeuwsen JAL, Wesseling M, Imo E, Hoefer IE, de Jager SCA. Prognostic value of cir- culating inflammatory cells in patients with stable and acute coronary artery disease. Frontiers in Cardiovascular Medicine. 2017;4(44):1-10.

Mehta NN, Matthews GJ, Krishnamoorthy P, et al; Chronic Renal Insufficiency Cohort (CRIC) Study Investigators. Higher plasma CXCL12 levels predict incident myocardial infarction and death in chronic kidney disease: findings from the Chronic Renal Insufficiency Cohort study. Eur Heart J. 2014;35:2115–2122.

(b) Mehta N, Qamar A, Qu L, *et al.* Differential association of plasma angio- poietin-like proteins 3 and 4 with lipid and metabolic traits. Arterioscler Thromb Vasc Biol. 2014;34:1057–1063.

(c) Mehta PK, Griendling KK. Angiotensin II cell signaling: Physiological and pathological effects in the cardiovascular system. Am J Physiol - Cell Physiol. 2007;292:82–97.

Michineau S, Franck G, Wagner-Ballon O, Dai J, Allaire E, Gervais M. Chemokine (C-X-C motif) receptor 4 blockade by AMD3100 inhibits experimental abdominal aortic aneurysm expansion through anti-inflam- matory effects. Arterioscler Thromb Vasc Biol. 2014;34:1747–1755.

Miettinen JJ, Kumari R, Traustadottir GA, Huppunen ME, Sergeev P, Majumder MM, Schepsky A, Gudjonsson T, Lievonen J, Bazou D, Dowling P, O Gorman P, Slipicevic A, Anttila P, Silvennoinen R, Nupponen NN, Lehmann F, Heckman CA. Aminopeptidase Expression in Multiple Myeloma Associates with Disease Progression and Sensitivity to Melflufen. Cancers (Basel). 2021 Mar 26;13(7):1527.

Millan, M., *et al.*, 2017. Vessel patency at 24 hours and its relationship with clinical outcomes and infarct volume in REVASCAT trial (randomized trial of revascularization with Solitaire FR device versus best medical therapy in the treatment of acute stroke due to anterior circulation large vessel occlusion presenting within eight hours of symptom onset). Stroke 48 (4), 983e989.

Mishima T, Mizuguchi Y, Kawahigashi Y, Takizawa T, Takizawa T. RT-PCR-based analysis of microRNA (miR-1 and -124) expression in mouse CNS. Brain Res. 2007 Feb 2;1131(1):37-43.

Montaner J. Blood biomarkers to guide stroke thrombolysis. Front Biosci (Elite Ed). 2009 Jun 1;1:200-8.

Montaner J, Ramiro L, Simats A, Hernández-Guillamon M, Delgado P, Bustamante A, Rosell A. Matrix metalloproteinases and ADAMs in stroke. Cell Mol Life Sci. 2019 Aug;76(16):3117-3140. doi: 10.1007/s00018-019-03175-5

Moore HE, Davenport EL, Smith EM, Muralikrishnan S, Dunlop AS, Walker BA, Krige D, Drummond AH, Hooftman L, Morgan GJ *et al.* (2009) Aminopeptidase inhibition as a targeted treatment strategy in myeloma.

Moore KJ, Tabas I. Macrophages in the pathogenesis of atherosclerosis. Cell. 2011;145: 341-355.

Moussaddy A, Demchuk AM, Hill MD. Thrombolytic therapies for ischemic stroke: Triumphs and future challenges. Neuropharmacology. 2018 May 15;134(Pt B):272-279. doi: 10.1016/j.neuropharm.2017.11.010.

Mozaffarian D, Benjamin EJ, Go AS, Arnett DK, Blaha MJ, Cushman M, Das SR, de Ferranti S, Després JP, Fullerton HJ, Howard VJ, Huffman MD, Isasi CR, Jiménez MC, Judd SE, Kissela BM, Lichtman JH, Lisabeth LD, Liu S, Mackey RH, Magid DJ, McGuire DK, Mohler ER, Moy CS, Muntner P, Mussolino ME, Nasir K, Neumar RW, Nichol G, Palaniappan L, Pandey DK, Reeves MJ, Rodriguez CJ, Rosamond W, Sorlie PD, Stein J, Towfighi A, Turan TN, Virani SS, Woo D, Yeh RW, Turner MB; American Heart Association Statistics Committee and Stroke Statistics Subcommittee. Heart disease and stroke statistics–2016 update: a report from the American Heart Association.Circulation. 2016; 133:e38–e360.

Mozaffarian D, Benjamin EJ, Go AS, Arnett DK *et al.*; American Heart Association Statistics Committee; Stroke Statistics Subcommittee. Heart Disease and Stroke Statistics-2016 Update: A Report From the American Heart Association. Circulation. 2016 Jan 26;133(4):e38-360.

Mukhopadhyay, S., George, V., Xu, H., 2010. Variable selection method for quantitative trait analysis based on parallel genetic algorithm. Ann. Hum. Genet. 74 (1), 88–96.

Murray, C. J. L. *et al.* Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990-2010: A systematic analysis for the Global Burden of Disease Study 2010. The Lancet 380, 2197–2223 (2012).

Mutin M, Canavy I, Blann A *et al.* Direct evidence of endothelial injury in acute myocardial infarction and unstable angina by demonstration of circulating endothelial cells. Blood 1999;93:2951–8.

Nabel EG. Cardiovascular disease. N Engl J Med 2003;349:60-72.

Naderi SH, Bestwick JP, Wald DS.Adherence to drugs that prevent cardiovascular disease: Meta-analysis on 376,162 patients. The American Journal of Medicine. 2012;125:882-887.

Narula J, Nakano M, Virmani R, Kolodgie FD, Petersen R, Newcomb R, Malik S, Fuster V, Finn AV. Histopathologic characteristics of atherosclerotic coronary disease and implications of the findings for the invasive and noninvasive detection of vulnerable plaques. JACC. 2013;61(10):1041-1051.

National Institute of Neurological Disorders and Stroke rt-PA Stroke Study Group. Tissue plasminogen activator for acute ischemic stroke. N Engl J Med. 1995 Dec 14;333(24):1581-7. doi: 10.1056/NEJM199512143332401. PMID: 7477192.

Navarro S, Medina P, Bonet E, *et al.* Association of the thrombomodulin gene c.1418C>T polymorphism with thrombomodulin levels and with venous thrombosis risk. Arterioscler Thromb Vasc Biol. 2013;33:1435– 1440.

Nesbit G, Clark W, Oniell O, *et al.* Intra-cranial intra-arterial thrombolysis facilitated by microcatheter navigation through an occluded internal carotid artery. J Neurosurg 1996; 84:387-392.

Nestel PJ, Barnes EH, Tonkin AM, *et al.* Plasma lipoprotein(a) concen- tration predicts future coronary and cardiovascular events in patients with stable coronary heart disease. Arterioscler Thromb Vasc Biol. 2013;33:2902–2908.

Neumaier F, Stoppe C, Veldeman M, Weiss M, Simon T, Hoellig A, Marx G, Clusmann H, Albanna W. Circulatory dipeptidyl peptidase 3 (cDPP3) is a potential biomarker for early detection of secondary brain injury after aneurysmal subarachnoid hemorrhage. J Neurol Sci. 2021 Mar 15;422:117333.

NICE Public Health Guidance 25. Prevention of Cardiovascular Disease. Public health guideline Published; 2010. Available from: http://www.nice.org.uk/guidance/PH25

Nielsen R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet 2011;12:443–51.

Niesten JM, van der Schaaf IC, van Dam L, *et al.* Histopathologic composition of cerebral thrombi of acute stroke patients is correlated with stroke subtype and thrombus attenuation. PLoS One 2014; 9: e88882.

Nilsson J. Atherosclerotic plaque vulnerability in the statin era. European Heart Journal. 2017;38(21):1638-1644.

Niture SK, Kaspar JW, Shen J & Jaiswal AK (2009) Nrf2 signaling and cell survival. Toxicol Appl Pharmacol 244, 37–42.

Nogueira RG, Jadhav AP, Haussen DC, *et al.* Thrombectomy 6 to 24 hours after stroke with a mismatch between deficit and infarct. N Engl J Med 2018;378:11–21.

O'Donnell CJ, Nabel EG. Genomics of cardiovascular disease. N Engl J Med. 2011 Dec 1;365(22):2098-109.

Ohkubo I, Li Y-H, Maeda T, Yamamoto Y, Yamane T, Du P-G, *et al.* Dipeptidyl peptidase III from rat liver cytosol: purification, molecular cloning and immunohistochemical localization. Biol Chem.1999;380:1421–30.

Oliver, G. R., Hart, S. N. & Klee, E. W. Bioinformatics for clinical next generation sequencing. Clinical Chemistry vol. 61 124–135 (2015).

Osende JI, Badimon JJ, Fuster V *et al.* Blood thrombogenicity in type 2 diabetes mellitus patients is associated with glycemic control. J Am Coll Cardiol 2001;38:1307–12.

Ketelhuth DFJ, Hansson GK. Adaptive response of T and B cells in atherosclerosis. Circulation Research. 2016;118:668-678.

Koenig W. Fibrin (ogen) in cardiovascular disease: an update. Thromb Haemost 2003;89:601–9.

Ouriel K. Peripheral arterial disease. Lancet 2001;358:1257–64

Parikh, N. I. *et al.* Parental occurrence of premature cardiovascular disease pre-dicts increased coronary artery and abdominal aortic calcification in the Fra-mingham Offspring and third generation cohorts. Circulation 116, 1473–1481 (2007).

Parsons ME, Pennington RJT. Separation of rat muscle aminopeptidases. Biochem J.1976;155:375–81.

Perez-de-Puig I, Miró-Mur F, Ferrer-Ferrer M, Gelpi E, Pedragosa J, Justicia C, Urra X, Chamorro A, Planas AM. Neutrophil recruitment to the brain in mouse and human ischemic stroke. Acta Neuropathol. 2015 Feb; 129(2):239-57.

Pezzini A, Grassi M, Del Zotto E, Archetti S, Spezi R, vergani V, Assanelli D, Caimi L, Padovani A. Cumulative effects of predisposing genotypes and their interaction with modifiable factors on the risk of ischemic stroke in young adults. Stroke 2005;36:533-539.

Pham VL, Adel MS, Gouzy-Darmon C, Hanquez C, Beinfeld MC, Nicolas P, Etchebest C & Foulon T (2007) Aminopeptidase B, a glucagon-processing enzyme: site directed mutagenesis of the Zn2+-binding motif and molecular modeling. BMC Biochem 8, 21.

Piyamongkol W, Bermúdez MG, Harper JC, Wells D. Detailed investigation of factors influencing amplification efficiency and allele drop-out in single cell PCR: implications for preimplantation genetic diagnosis. Mol Hum Reprod. 2003 Jul;9(7):411-20. doi: 10.1093/molehr/gag051. PMID: 12802048.

Popova T, Mennerich D, Weith A, Quast K. Effect of RNA quality on transcript intensity levels in microarray analysis of human post-mortem brain tissues. BMC Genomics. 2008 Feb 25;9:91. doi: 10.1186/1471-2164-9-91

Powers WJ, Rabinstein AA, Ackerson T, Adeoye OM, Bambakidis NC, Becker K, Biller J, Brown M, Demaerschalk BM, Hoh B, Jauch EC, Kidwell CS, Leslie-Mazwi TM, Ovbiagele B, Scott PA, Sheth KN, Southerland AM, Summers DV, Tirschwell DL; American Heart Association Stroke Council. Guidelines for the Early Management of Patients With Acute Ischemic Stroke: 2019 Update to the 2018 Guidelines for the Early Management of Acute Ischemic Stroke: A Guideline for Healthcare Professionals From the American Heart Association/American Stroke Association. Stroke. 2019 Oct 30:STR0000000000000211. doi: 10.1161/STR.0000000000000211.

Prajapati SC, Chauhan SS. Dipeptidyl peptidase III: a multifaceted oligopeptide N-end cutter. FEBS J. 2011 Sep;278(18):3256-76.

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., Reich, D., 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. 38 (8), 904–909.

Pritchard, J.K., Stephens, M., Donnelly, P., 2000. Inference of population structure using multilo- cus genotype data. Genetics 155 (2), 945–959.

Quackenbush J. Microarray data normalization and transformation. Nat Genet. 2002 Dec;32 Suppl:496-501.

Rader DJ. Human genetics of atherothrombotic disease and its risk factors. Arterioscler Thromb Vasc Biol. 2015 Apr;35(4):741-7.

Raj, A., Stephens, M., Pritchard, J.K., 2014. FastSTRUCTURE: variational inference of popula- tion structure in large SNP data sets. Genetics 197 (2), 573–589.

Ramos E, Patiño P, Reiter RJ, Gil-Martín E, Marco-Contelles J, Parada E, de Los Rios C, Romero A, Egea J. Ischemic brain injury: New insights on the protective role of melatonin. Free Radic Biol Med. 2017 Mar;104:32-53.

(a)Rauch U, Osende JI, Chesebro JH *et al.* Statins and cardiovascular diseases: the multiple effects of lipid-lowering therapy by statins. Atherosclerosis 2000;153:181–9.

(b)Rauch U, Crandall J, Osende JI *et al.* Increased thrombus formation relates to ambient blood glucose and leukocyte count in diabetes mellitus type 2. Am J Cardiol 2000;86:246–9.

Rauch U, Osende JI, Fuster V *et al.* Thrombus formation on atherosclerotic plaques: pathogenesis and clinical consequences. Ann Intern Med 2001;134:224–38.

Rayasam A, Hsu M, Kijak JA, Kissel L, Hernandez G, Sandor M, Fabry Z. Immune responses in stroke: how the immune system contributes to damage and healing after stroke and how this knowledge could be translated to better cures? Immunology. 2018 Jul; 154(3):363-376.

Razvi SSM, Bone I. Single gene disorders causing ischaemic stroke. J Neurol 2006;253:685– 700.

Rehfeld L, Funk E, Jha S, Macheroux P, Melander O, Bergmann A. Novel methods for the quantification of dipeptidyl peptidase 3 (DPP3) concentration and activity in human blood samples. J Appl Lab Med. 2019;3:943–53.

Ren X, Yu J, Guo L, Ma H. Dipeptidyl-peptidase 3 protects oxygen-glucose deprivation/reoxygenation-injured hippocampal neurons by suppressing apoptosis, oxidative stress and inflammation via modulation of Keap1/Nrf2 signaling. Int Immunopharmacol. 2021 Jul;96:107595.

Reiner AP, Lange EM, Jenny NS, *et al.* Soluble CD14: genomewide asso- ciation analysis and relationship to cardiovascular risk and mortality in older adults. Arterioscler Thromb Vasc Biol. 2013;33:158–164.

Ridker PM. Clinical application of C-reactive protein for cardiovascular disease detection and prevention. Circulation. 2003;107:363–369.

Ridker PM, Brown NJ, Vaughan DE, Harrison DG, Mehta JL. Established and emerging plasma biomarkers in the prediction of first atherothrombotic events. Circulation. 2004 Jun 29;109(25 Suppl 1):IV6-19.

Ritchie, M.D., Holzinger, E.R., Li, R., Pendergrass, S.A., Kim, D., 2015. Methods of integrating data to uncover genotype-phenotype interactions. Nat. Rev. Genet. 16 (2), 85–97.

Ritchie ME, Silver J, Oshlack A, Holmes M, Diyagama D, Holloway A, Smyth GK. A comparison of background correction methods for two-colour microarrays. Bioinformatics. 2007 Oct 15;23(20):2700-7.

Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., *et al.*, 2015. Integrative analysis of 111 reference human epigenomes. Nature 518 (7539), 317–330.

Robciuc MR,Maranghi M,LahikainenA,etal.Angptl3deficiencyisassociated with increased insulin sensitivity, lipoprotein lipase activity, and decreased serum free fatty acids. Arterioscler Thromb Vasc Biol. 2013;33:1706–1713.

Roffi M, Patrono C, Collet J-P, Mueller C, Valgimigli M, Andreotti F, Bax JJ, Borger MA, Brotons C, Chew DP. 2015 ESC guidelines for the management of acute coronary syn- dromes in patients presenting without persistent ST-segment elevation: Task force for the Management of Acute Coronary Syndromes in patients presenting without persis- tent ST-segment elevation of the European Society of Cardiology (ESC). European Heart Journal. 2016;37(3):267-315.

Romeo S, Pennacchio LA, Fu Y, *et al.* Population-based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase HDL. Nat Genet. 2007;39:513–516.

Ross R. Atherosclerosis--an inflammatory disease. N Engl J Med. 1999 Jan 14;340(2):115-26.

Rothberg MB, Celestin C, Fiore LD, Lawler E, Cook JR. Warfarin plus aspirin after myocardial infarction or the acute coronary syndrome: meta-analysis with estimates of risk and benefit. Ann Intern Med. 2005;143:241–250.

Sabra A, Bessoule JJ, Atanasova-Penichon V, Noël T, Dementhon K. Host-pathogen interaction and signaling molecule secretion are modified in the dpp3 knockout mutant of Candida lusitaniae. Infect Immun. 2014 Jan;82(1):413-22.

Sacco RL, Ellenberg JH, Mohr JP, *et al.* Infarcts of undetermined cause: the NINCDS Stroke Data Bank. Ann Neurology 1989;25:382–90.

Sambola A, Osende J, Hathcock J *et al.* Role of risk factors in the modulation of tissue factor activity and blood thrombogenicity. Circulation 2003;107:973–7.

Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 1977;74(12):5463-5467. doi:10.1073/pnas.74.12.5463

Sardellaa G, Accapezzatob D, Di Romaa A, Francavillab V, Di Russoa C, Iannuccib G, Sirinianc MI, Giacomellid L, Fedelea F, Parolib M. Altered trafficking of CD8+ memory T cells after implantation of rapamycin-eluting stents in patients with coronary artery disease. Immunology Letters. 2005;96:85-91.

Sato H,Kimura K,Yamamoto Y& Hazato T(2003) Activity of DPP III in human cerebrospinal fluid derived from patients with pain. Masui 52, 257–263.

Saver JL, Goyal M, Bonafe A, Diener HC, Levy EI, Pereira VM, *et al.* Stent-retriever thrombectomy after intravenous t-PA vs. t-PA alone in stroke. N Engl J Med. 2015;372(24):2285–95. https://doi.org/10.1056/NEJMoa1415061.

Selvakumar P, Lakshmikuttyamma A, Das U, Pati HN, Dimmock JR & Sharma RK (2009) NC2213: a novel methionine aminopeptidase 2 inhibitor in human colon cancer HT29 cells. Mol Cancer 8, 65.

Sentandreu MA & Toldra F (2005) Generation of ACE inhibitory peptides by pork muscle dipeptidyl peptidase I and III. Food Chem (http://hdl.handle.net/10261/3036).

Serbina NV, Jia T, Hohl TM, Pamer EG. Monocyte-mediated defense against microbial pathogens. Annu Rev Immunol. 2008; 26():421-52.

Shah PK, Falk E, Badimon JJ, Fernandez-Ortiz A, Mailhac A, Villareal-Levy G, Fallon JT, Regnstrom J, Fuster V. Human monocyte-derived macrophages induce collagen breakdown in fibrous caps of atherosclerotic plaques. Potential role of matrix-degrading metalloproteinases and implications for plaque rupture. Circulation. 1995;92:1565-1569.

Shah R, Hellkamp A, Lokhnygina Y, Becker RC, Berkowitz SD, Breithardt G, Hacke W, Halperin JL, Hankey GJ, Fox KA, Nessel CC, Mahaffey KW, Piccini JP, Singer DE, Patel MR; ROCKET AF Steering Committee Investigators. Use of concomitant aspirin in patients with atrial fibrillation: findings from the ROCKET AF trial. Am Heart J. 2016;179:77–86. doi: 10.1016/j.ahj.2016.05.019

Shaked I, Hanna DB, Gleißner C, *et al.* Macrophage inflamma- tory markers are associated with subclinical carotid artery disease in women with human immunodeficiency virus or hepatitis C virus infection. Arterioscler Thromb Vasc Biol. 2014;34:1085–1092.

Shaw LJ, Hausleiter J, Achenbach S, et al; CONFIRM Registry Investigators. Coronary computed tomographic angiography as a gate- keeper to invasive diagnostic and surgical procedures: results from the

multicenter CONFIRM (Coronary CT Angiography Evaluation for Clinical Outcomes: an International Multicenter) registry. J Am Coll Cardiol. 2012;60:2103–2114.

Shedden K. Confidence levels for the comparison of microarray exper- iments. Stat Appl Genet Mol Biol, 3:Article32, 2004.

Shen R, Seshan VE. FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. Nucleic Acids Res. 2016 Sep 19;44(16):e131. doi: 10.1093/nar/gkw520. Epub 2016 Jun 7. PMID: 27270079; PMCID: PMC5027494.

Shimamori Y, Watanabe Y & Fujimoto Y (1986) Purification and characterization of dipeptidyl aminopepti- dase III from human placenta. Chem Pharm Bull 34, 3333–3340.

Shukla AA, Jain M & Chauhan SS (2010) Ets-1 ⁄ Elk-1 is a critical mediator of dipeptidyl-peptidase III tran- scription in human glioblastoma cells. FEBS J 277, 1861–1875.

Sikkema-Raddatz B, Johansson LF, de Boer EN, Almomani R, Boven LG, van den Berg MP, van Spaendonck-Zwarts KY, van Tintelen JP, Sijmons RH, Jongbloed JD, Sinke RJ. Targeted next-generation sequencing can replace Sanger sequencing in clinical diagnostics. Hum Mutat. 2013 Jul;34(7):1035-42. doi: 10.1002/humu.22332. Epub 2013 Apr 29. PMID: 23568810.

Simaga S, Babic D, Osmak M, Sprem M & Abramic M (2003) Tumor cytosol dipeptidyl peptidase III activity is increased with histological aggressiveness of ovarian primary carcinomas. Gynecol Oncol 91, 194–200.

Simons N, Mitchell P, Dowling R, Gonzales M, Yan B. Thrombus composition in acute ischemic stroke: a histopathological study of thrombus extracted by endovascular retrieval. J Neuroradiol. 2015 Apr;42(2):86-92. doi: 10.1016/j.neurad.2014.01.124.

Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genet. 2014 Feb;15(2):121-32. doi: 10.1038/nrg3642. PMID: 24434847.

Singer OC, Humpich MC, Fiehler J, Albers GW, Lansberg MG, Kastrup A, Rovira A, Liebeskind DS, Gass A, Rosso C, Derex L, Kim JS, Neumann-Haefelin T; MR Stroke Study Group Investigators. Risk for symptomatic intracerebral hemorrhage after thrombolysis assessed by diffusion-weighted magnetic resonance imaging. Ann Neurol. 2008 Jan;63(1):52-60. PubMed PMID: 17880020.

Smith WS, Sung G, Starkman S, *et al.* Safety and efficacy of mechanical embolectomy in acute ischemic stroke: results of the MERCI trial. Stroke 2005;36: 1432–38.

Smyth M, O'Cuinn G. Dipeptidyl aminopeptidase III of guinea-pig brain: specificity for short oligopeptide sequences. J Neurochem. 1994 Oct;63(4):1439-45.

Smyth GK, Michaud J, Scott HS. Use of within-array replicate spots for assessing differential expression in microarray experiments. Bioinformatics. 2005 May 1;21(9):2067-75.

Smyth SS, Mueller P, Yang F, Brandon JA, Morris AJ. Arguing the case for the autotaxin-lysophosphatidic acid-lipid phosphate phosphatase 3-signaling nexus in the development and complications of atheroscle- rosis. Arterioscler Thromb Vasc Biol. 2014;34:479–486.

Sobocanec S, Filic V, Matovina M, Majhen D, Safranko ZM, Hadzija MP, *et al.* Prominent role of exopeptidase DPP III in estrogen-mediated protection against hyperoxia in vivo. Redox Biol. 2016;8:149–59.

Soejima H, Ogawa H, Yasue H *et al.* Heightened tissue factor associated with tissue factor pathway inhibitor and prognosis in patients with unstable angina. Circulation 1999;99:2908–13.

Soderberg C, Giugni TD, Zaia JA, Larsson S, Wahlberg JM & Moller E (1993) CD13 (human aminopeptidase N) mediates human cytomegalovirus infection. J Virol 67, 6576–6585.

Souza LC, Payabvash S, Wang Y, Kamalian S, Schaefer P, Gonzalez RG, Furie KL, Lev MH. Admission CT perfusion is an independent predictor of hemorrhagic transformation in acute stroke with similar accuracy to DWI. Cerebrovasc Dis. 2012;33(1):8-15. doi: 10.1159/000331914.

Stary HC, Chandler AB, Glagov S *et al.* A definition of initial, fatty streak, and intermediate lesions of atherosclerosis. A report from the Committee on Vascular Lesions of the Council on Arteriosclerosis, American Heart Association. Circulation 1994;89: 2462–78.

Steinhubl SR, Berger PB, Mann 3rd JT *et al.* Early and sustained dual oral antiplatelet therapy following percutaneous coronary interven- tion: a randomized controlled trial. JAMA 2002;288:2411–20.

Stevens AJ, Taylor MG, Pearce FG, Kennedy MA. Allelic Dropout During Polymerase Chain Reaction due to G-Quadruplex Structures and DNA Methylation Is Widespread at Imprinted Human Loci. G3 (Bethesda). 2017 Mar 10;7(3):1019-1025. doi: 10.1534/g3.116.038687. PMID: 28143949; PMCID: PMC5345703.

Stoll G, Bendszus M. Inflammation and atherosclerosis. Novel insights into plaque formation and destabilization. Stroke. 2006;37:1923-1932.

Strom SP, Lee H, Das K, Vilain E, Nelson SF, Grody WW, Deignan JL. Assessing the necessity of confirmatory testing for exome-sequencing results in a clinical molecular diagnostic laboratory. Genet Med. 2014 Jul;16(7):510-5. doi: 10.1038/gim.2013.183. Epub 2014 Jan 9. PMID: 24406459; PMCID: PMC4079763.

Sumii T, Lo EH. Involvement of matrix metalloproteinase in thrombolysis-associated hemorrhagic transformation after embolic focal ischemia in rats. Stroke. 2002 Mar;33(3):831-6.

Sun Y., Gui H., Li Q., Luo Z.M., Zheng M.J., Duan J.L., Liu X. MicroRNA-124 protects neurons against apoptosis in cerebral ischemic stroke. CNS Neurosci. Ther. 2013; 19:813–819.

Sun Y, Luo ZM, Guo XM, Su DF, Liu X. An updated role of microRNA-124 in central nervous system disorders: a review. Front Cell Neurosci. 2015 May 20; 9:193.

Suzuki T, Kohro T, Hayashi D, Yamazaki T, Nagai R. Frequency and impact of lifestyle modification in patients with coronary artery disease: The Japanese coronary artery dis-ease (JCAD) study. American Heart Journal. 2012;163:268-273.

Suzuki Y, Nagai N, Umemura K. A Review of the Mechanisms of Blood-Brain Barrier Permeability by Tissue-Type Plasminogen Activator Treatment for Cerebral Ischemia. Front Cell Neurosci. 2016 Jan 25; 10:2. doi: 10.3389/fncel.2016.00002.

Swanson AA, Albers-Jackson B, McDonald JK. Mammalian lens dipeptidyl aminopeptidase III. Biochem Biophys Res Commun. 1978;84:1151–9.

Swanson AA, Davis RM & McDonald JK (1984) Dipeptidyl peptidase III of human cataractous lenses. Partial Purification. Curr Eye Res 3, 287–291.

Szmitko PE, Wang CH, Weisel RD, *et al.* New markers of inflammation and endothelial cell activation: Part I. Circulation. 2003;108:1917–1923.

Takagi K, Blet A, Levy B, Deniau B, Azibani F, Feliot E, *et al.* Circulating dipeptidyl peptidase 3 and alteration in haemodynamics in cardiogenic shock: results from the OptimaCC trial. Eur J Heart Fail. 2019;22:279–86.

Tan CE, Glantz SA. Association between smoke-free legislation and hospitalizations for cardiac, cerebrovascular, and respiratory diseases: A meta-analysis. Circulation. 2012;126:2177-2183.

Tang, H., Peng, J., Wang, P., Risch, N.J., 2005. Estimation of individual admixture: analytical and study design considerations. Genet. Epidemiol. 28 (4), 289–301.

Teslovich TM, Musunuru K, Smith AV, *et al.* Biological, clinical and pop- ulation relevance of 95 loci for blood lipids. Nature. 2010;466:707–713.

Thim, T., Hagensen, M. K., Bentzon, J. F. & Falk, E. From vulnerable plaque to atherothrombosis. in Journal of Internal Medicine vol. 263 506–516 (2008).

Tice JA, Browner W, Tracy RP, *et al.* The relation of C-reactive protein levels to total and cardiovascular mortality in older U.S. women. Am J Med. 2003;114:199 –205.

Tong Y, Huang Y, Zhang Y, Zeng X, Yan M, Xia Z, Lai D. DPP3/CDK1 contributes to the progression of colorectal cancer through regulating cell proliferation, cell apoptosis, and cell migration. Cell Death Dis. 2021 May 22;12(6):529.

Torres JL, Ridker PM. Clinical use of high-sensitivity C-reactive protein for the prediction of adverse cardiovascular events. Curr Opin Cardiol. 2003;18:471– 478.

Toschi V, Gallo R, Lettino M *et al.* Tissue factor modulates the thrombogenicity of human atherosclerotic plaques. Circulation 1997;95:594–9.

Tournier-Lasserve E. New players in the genetics of stroke. N Engl J Med. 2002 Nov 21;347(21):1711-2.

Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and

solutions. Nat Rev Genet 2012;13:36 – 46.

Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A. 2001 Apr 24;98(9):5116-21.

Ulahannan D, Kovac MB, Mulholland PJ, Cazier JB, Tomlinson I. Technical and implementation issues in using next-generation sequencing of cancers in clinical practice. Br J Cancer 2013;109:827–35.

van Amsterdam JGC, van Buuren KJH, Soudijn W. Purification and characterization of enkephalin degradating enzymes from calf-brain striatum. Biochem Biophys Res Commun. 1983;115:632–41.

Vandenberg JI, King GF & Kuchel PW (1985) Enkephalin degradation by human erythrocytes and hemolysates studied using 1H NMR spectroscopy. Arch Biochem Biophys 242, 515–522.

Vanha-Perttula T (1988) Dipeptidyl peptidase III and alanyl aminopeptidase in the human seminal plasma: origin and biochemical properties. Clin Chim Acta 177, 179–195.

van Lier D, Kox M, Pickkers P. Promotion of vascular integrity in sepsis through modulation of bioactive adrenomedullin and dipeptidyl peptidase 3. J Intern Med. 2020;289:1–15.

Vazeux G, Wang J, Corvol P & Llorens-Cortes C (1996) Identification of glutamate residues essential for catalytic activity and zinc coordination in aminopepti- dase A. J Biol Chem 271, 9069–9074.

Varbo A, Benn M, Tybjærg-Hansen A, Jørgensen AB, Frikke-Schmidt R, Nordestgaard BG. Response: lipoprotein subclass profiling reveals pleiotropy in the genetic variants of lipid risk factors for

coronary heart disease: a note on Mendelian randomization studies. J Am Coll Cardiol. 2013;62:1908–1909.

Veerbeek JM, van Wegen E, van Peppen R, van der Wees PJ, Hendriks E, Rietberg M, Kwakkel G. What is the evidence for physical therapy poststroke? A systematic review and meta-analysis. PLoS ONE. 2014; 9:e87987. [PubMed: 24505342]

Vercammen D, Brouckaert G, Denecker G, Van de Craen M, Declercq W, Fiers W, *et al.* Dual signaling of the Fas receptor: initiation of both apoptotic and necrotic cell death pathways. J Exp Med. 1998;188:919–30.

Viles-Gonzalez, J. F., Fuster, V. & Badimon, J. J. Atherothrombosis: A wide-spread disease with unpredictable and life-threatening consequences. European Heart Journal vol. 25 1197–1207 (2004).

Virmani R, Kolodgie FD, Burke AP, Farb A, Schwartz SM. Lessons from sudden coro- nary death: A comprehensive morphological classification scheme for atherosclerotic lesions. Arteriosclerosis, Thrombosis, and Vascular Biology. 2000;20(5):1262-1275.

Virmani R, Burke AP, Farb A, Kolodgie FD. Pathology of the vulnerable plaque. Journal of the American College of Cardiology. 2006;47:C13-C18.

van der Wal AC, Becker AE, van der Loos CM *et al.* Site of intimal rupture or erosion of thrombosed coronary atherosclerotic plaques is characterized by an inflammatory process irrespective of the dominant plaque morphology. Circulation 1994;89:36–44.

Voelkerding KV, Dames SA, Durtschi JD. Next-generation sequencing: from basic research to diagnostics. Clin Chem 2009;55:641–58.

Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010 Sep;38(16):e164.

Wang X, Peter K. Molecular Imaging of Atherothrombotic Diseases: Seeing Is Believing. Arterioscler Thromb Vasc Biol. 2017 Jun;37(6):1029-1040.

Wang A, Abramowicz AE. Endovascular thrombectomy in acute ischemic stroke: new treatment guide. Curr Opin Anaesthesiol. 2018 Aug;31(4):473-480. doi: 10.1097/ACO.0000000000000621.

Wang J, Huang Q, Ding J, Wang X. Elevated serum levels of brain-derived neurotrophic factor and miR-124 in acute ischemic stroke patients and the molecular mechanism. 3 Biotech. 2019 Nov;9(11):386.

Wattiaux R, Wattiaux-De Coninck S, Thirion J, Gasingirwa MC, Jadot M. Lysosomes and Fasmediated liver cell death. Biochem J. 2007;403:89–95.

Weiss N, Keller C, Hoffmann U *et al.* Endothelial dysfunction and atherothrombosis in mild hyperhomocysteinemia. Vasc Med 2002;7: 227–39.

Wildgruber M, Swirski FK, Zernecke A. Molecular imaging of inflam- mation in atherosclerosis. Theranostics. 2013;3:865–884.

Wilson LA, Gemin A, Espiritu R & Singh G (2005) ets-1 is transcriptionally up-regulated by H2O2 via an antioxidant response element. FASEB J 19, 2085–2087.

(b) Wilson BA, Cruz-Diaz N, Marshall AC, Pirro NT, Su Y, Gwathmey TYM, *et al.* An angiotensin-(1–7) peptidase in the kidney cortex, proximal tubules, and human HK-2 epithelial cells that is distinct from insulin-degrading enzyme. Am J Physiol - Ren Physiol. 2015;308:F594–601.

Winship IR, Murphy TH. Remapping the somatosensory cortex after stroke: insight from imaging the synapse to network. Neuroscientist. 2009; 15:507–524.

de Winter RJ, Koch KT, van Straalen JP, *et al.* C-reactive protein and coronary events following percutaneous coronary angioplasty. Am J Med. 2003;115:85–90.

Woo D, Falcone GJ, Devan WJ, *et al.* Meta-analysis of genome-wide association studies identifies 1q22 as a susceptibility locus for intracerebral hemorrhage. Am J Hum Genet 2014; 94: 511–21.

World Health Organization. Mortality estimates by cause, age, and sex for the year 2008 [Internet]. Geneva: WHO; 2011. Available from: http://www.whoint/healthinfo/global_ burden_disease/en/ [Accessed: 2017-08-20].

Wu H, Kerr M K, Cui X, Churchill G A. MAANOVA: a software package for the analysis of spotted cDNA microarray experiments. The Analysis of Gene Expression Data: Methods and Software, pages 313-341, 2003.

Wu C, Dwivedi DJ, Pepler L, *et al.* Targeted gene sequencing identi- fies variants in the protein C and endothelial protein C receptor genes in patients with unprovoked venous thromboembolism. Arterioscler Thromb Vasc Biol. 2013;33:2674–2681.

Xu H, Ruff CT, Giugliano RP, Murphy SA, Nordio F, Patel I, Shi M, Mercuri M, Antman EM, Braunwald E. Concomitant use of single anti- platelet therapy with edoxaban or warfarin in patients with atrial fibril- lation: analysis from the ENGAGE AF-TIMI 48 trial. J Am Heart Assoc. 2016;5:e002587.

Yamamoto Y, Hashimoto J, Shimamura M, Yama- guchic T & Hazato T (2000) Characterization of tynorphin, a potent endogenous inhibitor of dipeptidyl peptidase III. Peptides 21, 503–508.

Yang Y, Rosenberg GA. Blood-brain barrier breakdown in acute and chronic cerebrovascular disease. Stroke. 2011 Nov;42(11):3323-8.

Yang J, Zhang X, Chen X, Wang L, Yang G. Exosome Mediated Delivery of miR-124 Promotes Neurogenesis after Ischemia. Mol Ther Nucleic Acids. 2017 Jun 16;7:278-287.

Yilmaz F, Köklü E, Yilmaz FK, Gencer ES, Alparslan AS, Yildirimtürk Ö. Evaluation of mean platelet volume and platelet distribution width in patients with asymptomatic intermediate carotid artery plaque. Polish Heart Journal. 2017;75(1):35-41.

Young JY, Schaefer PW. Acute ischemic stroke imaging: a practical approach for diagnosis and triage. Int J Cardiovasc Imaging. 2016 Jan;32(1):19-33.

York IA, Mo AX, Lemerise K, Zeng W, Shen Y, Abra- ham CR, Saric T, Goldberg AL & Rock KL (2003) The cytosolic endopeptidase, thimet oligopeptidase, destroys antigenic peptides and limits the extent of MHC class I antigen presentation. Immunity 18, 429–440.

Young JY, Schaefer PW. Acute ischemic stroke imaging: a practical approach for diagnosis and triage. Int J Cardiovasc Imaging. 2016 Jan;32(1):19-33. doi: 10.1007/s10554-015-0757-0. Epub 2015 Sep 11. Review. PubMed PMID: 26362874.

Yu Y, Wu BL, Wu J, Shen Y. Exome and whole-genome sequencing as clinical tests: a transformative practice in molecular diagnostics. Clin Chem 2012;58:1507–9.

Yuan C, Zhang SX, Polissar NL *et al.* Identification of fibrous cap rupture with magnetic resonance imaging is highly associated with recent transient ischemic attack or stroke. Circulation 2002;105: 181–5.

Yusuf S, Zhao F, Mehta SR *et al.* Effects of clopidogrel in addition to aspirin in patients with acute coronary syndromes without ST segment elevation. N Engl J Med 2001;345:494–502.

Zhan H, Yamamoto Y, Shumiya S, Kunimatsu M, Nishi K, Ohkubo I & Kani K (2001) Peptidases play an important role in cataractogenesis: an immunohisto- chemical study on lenses derived from Shumiya cata- ract rats. Histochem J 33, 511–521.

Zidar DA, Mudd JC, Juchnowski S, Lopes JP, Sparks S, Park SS, Ishikawa M, Osborne R, Washam JB, Chan C, Funderburg NT, Owoyele A, Alaiti MA, Myttle Mayuga M, Orringer C, Costa MA, Simon DI, Tatsuoka C, Califf RC, Newby LK, Lederman MM, Kent J, Weinhold KJ. Altered maturation status and possible immune exhaustion of CD8 T lymphocytes in the peripheral blood of patients presenting with acute coronary syndromes. Arteriosclerosis, Thrombosis, and Vascular Biology. 2016;36:389-397.

Zheng J, Zhang H, Banerjee S, Li Y, Zhou J, Yang Q, Tan X, Han P, Fu Q, Cui X, Yuan Y, Zhang M, Shen R, Song H, Zhang X, Zhao L, Peng Z, Wang W, Yin Y. A comprehensive assessment of Next-Generation Sequencing variants validation using a secondary technology. Mol Genet Genomic Med. 2019 Jul;7(7):e00748. doi: 10.1002/mgg3.748. Epub 2019 Jun 4. PMID: 31165590; PMCID: PMC6625156.

Zhou, D., Bousquet, O., Lal, T.N., Weston, J., Sch€olkopf, B., 2004. Learning with local and global consistency. In: Thrun, S., Saul, L.K., Sch€olkopf, B. (Eds.), Advances in Neural Information Processing Systems 16. MIT Press, pp. 321–328.

Zivin JA, Fisher M, DeGirolami U, Hemenway CC, Stashak JA. Tissue plasminogen activator reduces neurological damage after cerebral embolism. Science. 1985.

Zuccato C, Ciammola A, Rigamonti D, Leavitt BR, Goffredo D, Conti L, MacDonald ME, Friedlander RM, Silani V, Hayden MR, Timmusk T, Sipione S, Cattaneo E. Loss of huntingtin-mediated BDNF gene transcription in Huntington's disease. Science. 2001 Jul 20;293(5529):493-8.