Universidad Autónoma de Madrid

Faculty of Medicine

Department of Biochemistry

Interrogating human disease by transcriptomic and proteomic profiling in a sheep model for atrial fibrillation

Alba Álvarez-Franco

Doctoral thesis directed by Dr. Miguel Manzanares Fourcade at CNIC

Madrid May 2021

A mi familia.

Acknowledgements

Desde donde hay que empezar a contar para llegar hasta aquí? Es complicado. Supongo que elegiría Salamanca y la Facultad de Biología y la oportunidad que me dieron mis padres para hacer lo que quiera y donde quisiera, que no es poco. Me vine de Salamanca a Madrid, para hacer un máster en biología molecular, al laboratorio de María, para hacer algo que me motivaba a más no poder, "cosas de replicación transcripción y cromatina", en un bicho de lo más curioso. Mucho ordenador y algún que otro experimento. Me lo pasé muy bien, aprendí mucho. María, aunque no haya sido más largo, ha sido un genial haber podido ser parte de tu grupo y disfrutar del buen ambiente que había, gracias por confiar en mí, por el pensamiento crítico y por empujarme a hacer mejor las cosas.

Y seguidamente empezó la tesis, Miguel me dió la oportunidad de entrar en su grupo con una beca de doctorado, para estudiar cosas de ~~cromatina~~corazón. Miguel, esto ya es un estandar pero es verdad, no eres solo un jefe, eres un gran mentor y creas escuela. Muchas cosas son a valorar positivamente. Más allá de las evidentes me quedo con que no te ciñes a "el tema", eso me ha ayudado a tener la mente más abierta; nos motivas para que participemos en todo, charla o fiesta, de todo se aprende; los errores son necesarios; y los logros, por pequeños que sean se celebran. Gracias por tener puerta del despacho siempre abierta o la cámara del zoom siempre activa y ponerme aquellas reuniones todos los jueves. Gracias por la confianza y la libertad de estos 5 años largos. Me voy sintiendo que he crecido mucho, ciéntificamente y como persona dentro y fuera de la ciencia.

A mis compañeros de laboratorio, miembros pasados y presentes, ha sido genial ser parte del funKGen, sois todos científicos y personas incríbles, cada uno en su especialidad. Sergio, de mayor quiero ser como tú. Eres una maravillosa persona y un científico incansable y tenaz, da igual que a veces sea aburrido. Espero que me lleguen noticas sobre ti del Qmul. Melisa, me ha encantado aprender cosas sobre CTCF de ti, gracias por tu alegría y por motivarme, nunca olvidaré que convertiste mi primer congreso en una experiencia sobre ruedas. Claudio, gracias por haber estado siempre dispuesto a enseñarme todo lo que sabes, que me parece increíble! Recuerdo esos valiosísimos libros de histología. Me lo he pasado muy bien discutiendo de todo contigo y tengo la suerte de siguir disfrutado de tus consejos de experto montañero. Julio, admiro lo inquieta y despierta que es tu cabeza, básicamente, ha sido un placer y aún más instructivo haberte visto en tu salsa discutiendo de ciencia. Isa, he de decir, para mi vergüenza, que nunca he entendido muy bien lo que haces... pero si he entendido que eres incansable, y que cada vez que haces algo, lo haces hasta un poquito mejor, aunque a priori no parezca posible. Estoy segura de que triunfarás allá donde vayas. Raquel, he disfrutado compartiendo proyecto contigo, no con todo el mundo se

consigue gestionar tan bien las cosas complicadas. Admiro el esmero que le has puesto siempre a todo y tu resistencia. Aurora, me dió la vida en su día que te apiadases de mi y me llevaras a cultivos. Aunque breve, me ha encantado coincidir con tu sonrisa en el laboratorio. Mariajo, eres una luchadora que no se rinde, que consigue sus objetivos. El mérito que tiene hacer lo que haces y conciliar con tu cada año más numerosa familia, me parece alucinante. Me alegro de haber trabajado estrechamente contigo, aunque no todos los momentos han sido fáciles, el marcador es muy favorable. Disfruta mucho esta nueva etapa y que lo que decidas que venga te llene. Jesús, eres un tío brillante y tienes una energía increíble, y por supuesto tienes un piquito de oro. He disfrutado mucho contigo en el labo y te he visto crecer durante estos años (me alegro de que hayas superado lo del reggaeton). No me cabe duda de que encontrarás lo que te motiva. A todos los estudiantes de grado les agradezco el tiempo que hemos compartido, Antonio L, proxímamente en las mejores salas, Sara, fue genial tener a otra bioinformática en el grupo, espero que tus pasos te lleven lejos, Marcos, (cuidado con el escalón!), el futuro vecino de la resistencia e incipiente bioinformático. Gonzalo, aportando la perspectiva clínica al grupo, me alegro de haber poder contado contigo, sobretodo con tus conocimientos sobre el sistema circulatorio. A los que quedan, que son la resistencia y la semilla de lo que venga, les agradezco los animos, los pequeños detalles y las discusiones científicas y no tan científicas de la etapa final. María, eres estupenda, tan pronto te resucita un marrón de projecto, te hace un western, te "baila un swing" o te hace gabinete psicológico. Creo que eres una científica increíble y una mejor persona. Antonio, eres otro tío brillante, con la fuerza de sacarte tú solo un proyecto de debajo del brazo a base de perseverancia y resilencia y encima con exito. La meiga que hay en mi te augura muchos exitos. Marta, eres una sonrisa andante, con ganas de aprender de probar y de superarte, da gusto compartir espacio contigo. No me cabe duda de que tienes un potencial asombroso para conseguir todo lo que te propongas, incluso hasta hacerte un 8a.

Gracias a todas las unidades del CNIC, que nos facilitan diariamente el trabajo. Especialmente a la Unidad de Bioinformática en la que he tenido la suerte de poder tener un sitio para trabajar. El ambiente no podría ser mejor, y las ganas de ayudar y compartir que tenéis todos os hace muy especiales. Gracias Fátima, por permitirme estar ahí y participar de vuestras reuniones y seminarios, desde luego ha sido muy fructífero para mí. Gracias Manuel, Felipe, Carlos T., Juan Carlos, Jorge y a todos los juniors pasados y presentes: Víctor, Álvaro, Mateo y JP. Especialmente gracias a los Fernandos, he disfrutado muchísimo compartiendo las comidas con vosotros mientras arreglábamos la política madrileña. Fernando, señor del cluster, gracias por haberme salvado tantas veces, no solo arreglando problemas sino explicando y ayudandome a entender "por qué esta

torre no funciona". Gracias a la Unidad de Proteómica, esta tesis no habría salido a flote sin su esfuerzo y sus habilidades técnicas. Especialmente a Ricardo y a Enrique.

Quisiera también agradecer la aportación de todos los investigadores del CNIC y a todos los que he conocido a lo largo de mi vida, sea en etapas pasadas, o en congresos, por sus preguntas y comentarios que me han ayudado a encaminar este proyecto.

Gracias a todos los que han estado ahí durante todos estos años, acompañandome en esta etapa. Me temo que mi estupenda gestión del tiempo, como siempre, me ha jugado una mala pasada y me ha impedido agradeceros como se merece, pero prometo que solo fallo en esta versión depositada.

Abstract

Atrial fibrillation (AF) is a progressive cardiac arrhythmia that increases the risk of hospitalization and adverse cardiovascular events. AF is a complex disease with multiple risk factors and associated comorbidities [74, 126] and remarkably, among them, AF is associated with a 5-fold risk of stroke [100]. The global incidence is expected to double the current rates by 2050, and as the global population increases its longevity, understanding the consequences of AF on an aged population becomes more critical. Although AF has been under extensive research for more than 50 years, there is a clear demand for more inclusive and large-scale approaches to understand the molecular drivers responsible for AF, as well as the fundamental mechanisms governing the transition from paroxysmal to persistent and permanent forms. In this doctoral thesis project, we have taken advantage of a well-established model of tachypacing-induced long-standing AF in the sheep to analyse in vivo the molecular changes that occur in the atria and systemically during the progression of AF from paroxysmal to its long-lasting forms, always in comparison with paired surgically operated sinus rhythm controls. We have performed transcriptomic and proteomic profiling of atrial tissue and cardiomyocytes as well as profiling the blood proteome. For the analysis of data, we have developed our own pipelines, benchmarked different tools and carried out correlation-based integration analysis and hierarchical modelling of longitudinal data in a Bayesian framework. We demonstrate that the hallmarks of AF-induced atrial remodelling change only at early transitional stages at the molecular level, but remain unaltered at later stages of the disease and that the left atrium undergoes significantly more profound changes in its expression programme than the right atrium. By dissecting the short time window between the paroxysmal and persistent forms, we proved that electrical remodelling occurring in the left atria is sufficient to trigger molecular changes in less than a few hours and we have confirmed that the pro-thrombotic state, inflammation, and lipid metabolism are activated systemically as the result of AF per se, beyond being these processes associated with pre-existing comorbidities. This pattern of dynamic changes in gene and protein expression replicate the electrical and structural remodelling previously shown in animal models and in humans, and uncover novel mechanisms potentially relevant for disease treatment.

Resumen

La Fibrilación Auricular (FA) es una arritmia cardiaca progresiva, que aumenta el riesgo de hospitalización y de aparición de eventos cardiovasculares adversos. FA es una enfermedad compleja, con multiples factores de riesgo y comorbilidades asociadas [74, 126] y destacando entre todos ellos, la FA está asociada con un incremento en 5 veces de riesgo de accidente cerebrovascular [100]. Está previsto que en 2050, la incidencia poblacional se duplique, y teniendo en cuenta la tendencia al incremento de la experanza de vida, entender las consecuencias de la FA en una población cada vez más envejecida es imperativo. A pesar de ser sujeto de exhaustiva investigación durante más de 50 años, existe una clara demanda de abordajes más inclusivos y a gran escala que nos permitan comprender el funcionamiento de los actores moleculares responsables de la FA, así como los mecanismos fundamentales que gobiernan la transición de las formas paroxísticas a las persistentes y permanentes. En este proyecto de tesis doctoral, hemos aprovechado un modelo bien establecido de FA de larga duración inducida por taquicardia en ovejas, para así analizar **in vivo** los cambios moleculares que se producen en las aurículas y de manera sistémica durante la progresión de la FA desde paroxística a su formas duraderas, todo ello siempre en comparación con los controles emparejados del ritmo sinusal operados quirúrgicamente. Hemos perfilado a nivel transcriptómico y proteómico el tejido auricular y los cardiomiocitos pertenecientes a este, así como perfilado el proteoma sanguíneo. Para el análisis de datos, hemos desarrollado nuestras propias *pipelines*, hemos comparado diferentes herramientas y realizado un análisis de integración basados en correlación así como modelado jerárquicamente datos longitudinales en un contexto bayesiano. Con todo ello, hemos demostrado que las características distintivas de la remodelación auricular inducida por la FA cambian solo en las primeras etapas de transición a nivel molecular, pero permanecen inalteradas en las etapas posteriores de la enfermedad y que la aurícula izquierda sufre cambios significativamente más profundos en su programa de expresión que la aurícula derecha. Al diseccionar la corta ventana de tiempo entre las formas paroxística y persistente, hemos demostramos que el remodelado eléctrico que ocurre en las aurículas izquierdas es suficiente para desencadenar cambios moleculares en menos de unas pocas horas y hemos confirmado que el estado protrombótico, la inflamación y el transporte lipoproteico o el metabolismo, se activan sistémicamente como consecuencia de la FA per se, más allá de ser estos procesos asociados a comorbilidades preexistentes. Este patrón de cambios dinámicos en la expresión de genes y proteínas reproduce la remodelación eléctrica y estructural demostrada previamente en ovejas y en humanos, y en último término, descubre nuevos mecanismos potencialmente relevantes para el tratamiento de enfermedades.

Index

Figures

Acronyms

AF Atrial Fibrillation.

CI Credible Interval.

HM Hierarchical Models.

LA Left Atrium.

LAA Left Atrial Appendage.

MLM Multilevel Models.

pAF persistent AF.

PG Protein Group and Gene.

PLA Posterior Left Atrium.

PSIS-LOO Pareto-Smoothed Importance Sampling leave-one-out.

RA Right Atrium.

RAA Right Atrial Appendage.

SR Sinus rhythm.

Chapter 1

Introduction

1 What is Atrial Fibrillation?

Among cardiovascular diseases, Atrial Fibrillation (AF) is the most common heart rhythm disorder that influences quality of life significantly, causes considerable morbidity, and contributes to overall mortality. Its socio-economic impact is very high, as European countries allocate around 0.28% to 2.6% of their national healthcare-expenses to AF, mainly to in-patient care and in-hospital procedures [22, 51, 102]. Considered as one of the most important public health issues, atrial fibrillation is a complex disease with multiple risk factors and associated comorbidities [74, 126]. AF is associated with a 5-fold risk of stroke, and AF-related strokes are 2.5-fold more likely to be fatal [100]. Nearly 2000 million new cases of AF have been registered worldwide between 1997 and 2017, increasing the global incidence of AF by 31% in the past two decades [14], and the future perspectives are to double the current rates by 2050 [74]. As the global population increases its longevity, understanding the consequences of AF on an aged population becomes more critical. Also, gender differences have been reported in AF, such as distinct electrophysilogy properties [183], possibly caused by the different incidence of several risk factors [19, 138]. As such, the overall incidence of AF in women is lower, although its prevalence in women older than 75 years of age is higher, due to their increased life-expectancy [200]. Lifestyle choices also have a strong influence on AF risk. Both a sedentary lifestyle and extreme physical activity are detrimental [193]. Sedentarism enhances other risk factors of AF such as obesity, elevated BMI, diabetes mellitus, hypertension or obstructive sleep apnea, all of which are independent risk factors for AF [28, 87, 157]. Smoking and second hand exposure to cigarette smoke are associated with incident AF, multiplying by 2 the incidence in the case of current smoking [35]. These external factors or extrinsic drivers ultimately contribute to disease development by affecting the exposure of the heart to mechanical, chemical or electrical cues, and finally increasing arrhythmia susceptibility [126].

1.1 Initiation, maintenance and progression of AF

Normal heart beating is directed by the cardiac conduction system. The electrical impulse starts at the sinoatrial node SA, an specialized myocardial structure found in the atrial wall at the junction of the superior cava vein and the right atrium [154]. The wave of depolarization spreads across the atria, passing through the intra-atrial conduction network, causing them to contract and pump blood into ventricles. Upon reaching the atrioventricular node (AV), this wave penetrates through the membranous part of the intraventricular septum to bifurcate into the right and left bundles, populated with Purkinje fibres at their terminal branches. AV is the unique conduction path between atria and ventricles, and it introduces a delay in the conduction between chambers, allowing

proper ventricular filling. Finally, ventricles contract and pump blood to the body, as a result of the coordinated rhythmic contraction and relaxation of the heart.

AF is a supraventricular tachyarrhythmia [103], meaning that abnormal electrical activity within the atria, becomes chaotic and uncoordinated inducing fibrillation, which prevents proper ventricular excitation and impairs normal cardiac function. AF begins with the appearance of aberrant electrical impulses in the atria, not governed by the sinoatrial node (SA), which triggers the propagation of reentrant waves in a refractory region, known as the vulnerable substrate. These two events, initial ectopic firing and reentry, are the major arrhythmogenic mechanisms sustaining AF [141, 202]. However, with time, the vulnerable substrate undergoes incremental structural remodelling, stabilizing the reentry pattern and diminishing the importance of the ectopic firing [197]. Thus, progressive and everlasting atrial remodelling leads to long-term perpetuation of AF, captured by the concept of "AF begets AF" [251], where the fibrillating myocardium causes remodelling that further sustains fibrillation. Considering that AF is a progressive disorder, subtypes are consequently classified according to the duration of AF episodes, and in agreement with the temporal pattern of the arrhythmia. Episodes can be paroxysmal if they terminate spontaneously, persistent if lasting for more than 7 days, or permanent in case of long-term episodes lasting more than one year [69, 110].

1.2 Molecular basis of AF

Despite being well establish that ectopic activity and reentry are the fundamental mechanisms governing AF [141, 202], the underlying downstream and upstream pathophysiological pathways of the local atrial and systemic AF-promoting mechanism, remain poorly understood and more work is needed to provide insights with translational potential [94]. Accumulating evidence points towards electrophysiological abnormalities, atrial dilation and fibrosis, neurohumoral dysregulation, hemodynamics or inflammation, as some of the molecular mechanisms behind initiation and maintenance of the disease [83].

1.2.1 Electrophysiological remodelling

Ectopic activity is likely caused by delayed afterdepolarizations (DADs), associated with Ca^{2+}-handling abnormalities and early afterdepolarizations (EADs), promoted by delayed repolarization [84]. The electrical remodelling of ion currents leads to the decreasing L-type calcium current ($I_{Ca,L}$) and the increasing potassium currents (I_{K1}, I_{KS} and I_{K2P}, resulting in the shortening

of the effective refractory period [84, 201]. This allows cells to be ready for reactivation earlier, which favours the reentry, or in other words, the continuous self-excitation of cardiac tissue around a functional or structural obstacle, the vulnerable substrate [93]. What leads to differential behaviour of the ion channels and sensors, requires further study.

1.2.2 Structural remodelling: atrial dilation and fibrosis

Structural remodelling of the atria due to atrial dilation, fibrosis or loss of cell-to-cell coupling via gap junctions, exacerbates substrate vulnerability, promoting the stabilization of reentrant circuits and reducing conduction velocity [83]. The precise mechanisms by which atria become enlarged and stretched [48], fibroblast proliferates and differentiate towards procollagen-secreting myofibroblasts [159], the excessive extracellular matrix deposition or the impairment of connexin proteins [95], remain to elucidate at the molecular level.

1.2.3 Neurohumoral regulators

Autonomic nervous system influences cardiac electrophysiology and atrial fibroblast function [124]. In AF patients, sympathetic stimulation via beta-adrenoreceptors [128], parasympathetic stimulation via cholinergic muscarinergir receptors [125], or combined sympathetic and vagal activation induce ectopic activity contributing to AF initiation [208]. The Renin-Angiotensin-Aldosterone System, via angiotensin II receptors, leads to gene expression modulation of profibrotic pathways, upregulating among others, transcription factors such as Elk-1, c-fos, STAT1 or STAT3 [159].

1.2.4 The thrombotic state

AF disrupts the normal hemostasis of the circulation system, which might lead to thrombus formation, peripheral embolization and stroke [100]. The major causes of this pro-thromboembolic state are related to 1) hypercoagulability, 2) vessel wall and 3) blood flow abonormalities, described in literature as the Virchow's triad [240]. First, hypercoagulability refers to the activation of the coagulation cascade, platelet activation and impaired fibrinolysis, which are the most significant causes of clot formation. Thrombogenesis mostly takes places in the left atria (LA), and in particular in the region of the appendage (LAA) [253]. There, the presence of pro-thrombotic biomarkers is enhanced in comparison with the blood peripheral system, which may require a longer duration of AF to become manifested. Despite this state might be considered a pre-existing comorbidity, platelet are activated by the disease itself, only after 15 minutes of AF. The second component is related

to structural changes of the endothelium. In response to cardiac injury, the endocardium suffers oedematous and fibrous thickening, and the inner endothelial layer communicates damage releasing several factors, such as the von Willebrand factor vWF, which promotes clotting [147]. During AF episodes, the LA velocity becomes slower and turbulent, together with increase of blood stasis and thus contributing to clot formation [143]. For instance, vWF is degraded in normal conditions by a metallopeptidase (ADAMTS13) which is flow dependent [6]. Given its importance an clinical implications, an improved understanding of the factors implicated in the pro-thrombotic state will allow better diagnosis and prognosis of AF.

1.2.5 Inflamation

Inflammation is a risk factor for AF [12], however, increasing evidences suggest that inflammation as well is implicated in the pathophysiology of atrial remodelling, leading to a vicious circle [113, 261]. The impact of inflammation at different phases of AF is unknown and demands further studies. Moreover, there is a clear link between thrombogeneis and inflammation, with several players shared and interacting between both systems [101]. These mechanistic links are complex and need to be elucidated.

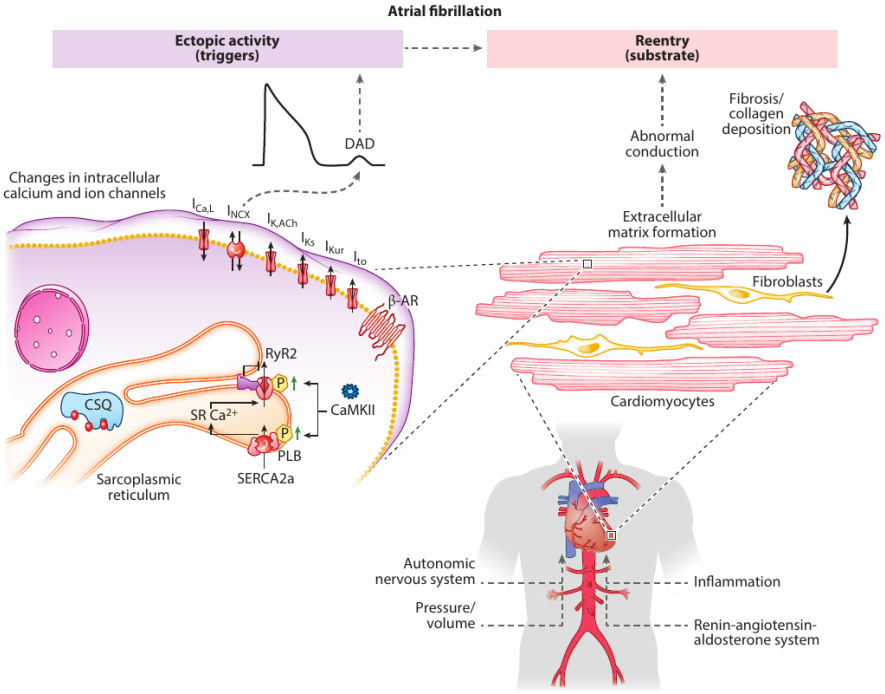


Figure 1.1: **Major pathophysiological mechanisms underlying atrial fibrillation (AF)-promoting ectopic activity and reentry.** Adapted from [84].

2 Approaches for study Atrial Fibrillation

2.1 Genome Wide Association Studies of Atrial fibrillation

GWAS rely on the fact that genetic variation leads to differences in phenotypes. To locate variation, the entire genome of the population subjected to analysis is interrogated using microarrays that genotype single nucleotide polymorphisms (SNPs). A robust test of association with numerous factors influencing the statistical design and the final outcome is conducted. Below a standard threshold of significance, the allele is accepted to be likely associated with the trait (tag SNP). Once a tag SNP is identified, the linkage disequilibrium (LD) block at the locus of that polymorphism is identified as a disease-risk locus. However, since LD blocks span from a few to hundreds of kilobases [7, 47, 184, 243], a clear limitation of GWAS is the difficulty in narrowing down of the associated region, thus not being able to specify which are the causal variants. Furthermore, because by design GWAS search for associations of SNPs that are common in the population, each individual variant would only account for little phenotypic differences in the trait or disease under study.

Since the release of the first genome-wide association study (GWAS) for AF in 2007, more than 20 studies have tested the association of genomic locations (loci) with this trait by means of comparing individuals with and without AF [20, 41, 60, 61, 77, 79, 134, 135, 167, 168, 191, 213, 230, 231]. The first GWAS performed by Gudbjartsson *et al.* discovered two strong associations on chromosome 4q25 at the PITX2 locus. Variants on this region increase the risk of AF, and this finding has been replicated in all the posterior studies. As the number of samples and the number of polymorphisms tested increased from hundreds to millions, so did the number of associations found (Figure 1.2). Recently, the simultaneous publication of the two largest studies to date, more than triplicated the number of previously reported associations, identifying more than 100 hundred loci and 150 genes putatively associated to AF, many of them replicated in distinct ethnic populations [168, 191].

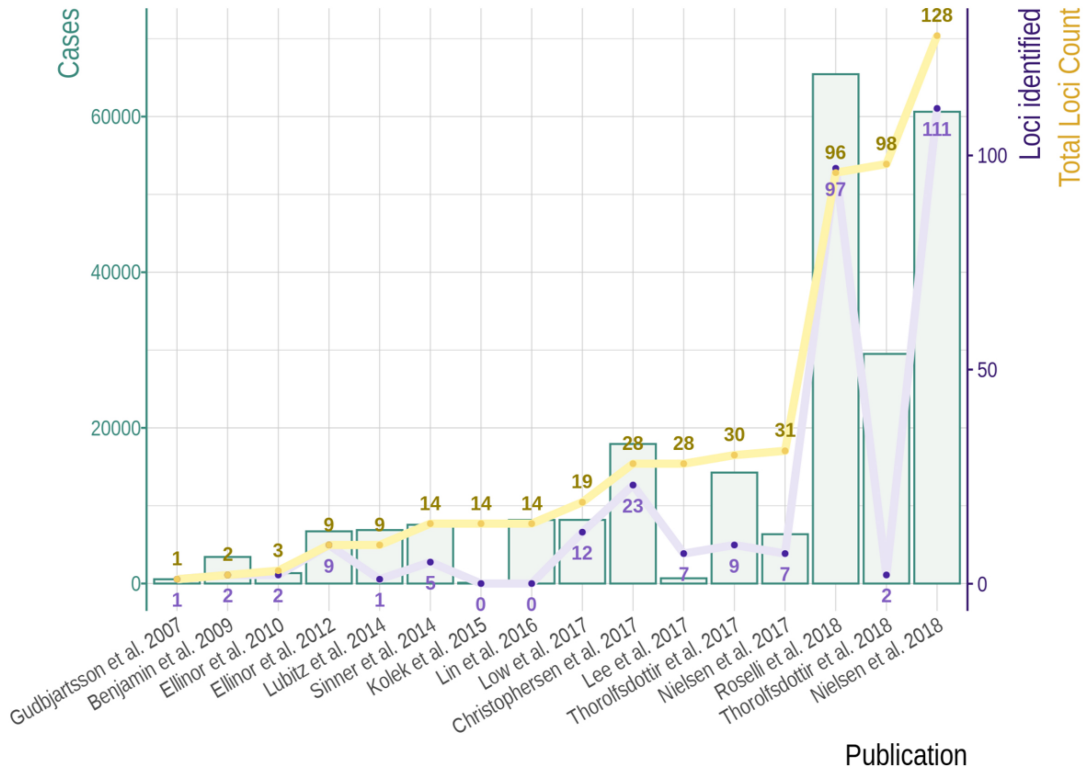


Figure 1.2: **The steady increase in numbers of genome-wide association studies of atrial fibrillation.** Bars represent the number of AF patients (Cases) recruited in each GWAS (original publication indicated below). Number of loci identified in each study are coloured in purple, while the cumulative number of newly discovered AF-associated loci appear in yellow (Total Loci Count).

Similar to any other complex trait, the heritability of AF is not explained by single genetic factors. Instead, thousands of them contribute to disease susceptibility additively and to different extents [247]. Those contributing modestly do not usually reach the standard threshold of significance of an association test until larger cohorts are used [99]. Along this line, work based on the UK biobank population estimated that the contribution of genomic variation to AF susceptibility is over 22%. However, the 25 loci associated with AF at that time only explained 5.3% [68].

One of the fundamental benefits of these genome-wide approach is that they are unbiased and not driven by predetermined candidate genes. Therefore, GWAS have uncovered unsuspected new pathways involved in AF. In addition to ion channels coding genes, genes for connexins (GJA5, GJA1), cardiac transcription factors (GATA4, GATA6, TBX5, NKX2-5, PITX2), and members of the renin-angiotensin pathways (ENPEP, C9orf3/AOPEP) are also in the spotlight since a number of associations in the vicinities of these genes has been described [168, 191]. Because of adding new players to the understanding of the disease, GWAS are a powerful tool in precision medicine

that will help to improve diagnoses or design new therapies based on the mechanisms responsible for these associations.

As a general principle for all GWAS, increasing sample size as well as the incorporation of populations of non-European descendants leads to a steady increase in the number of genotype-phenotype associations found in the studies, as is exemplified by AF (Figure 1.2). However, these strategies are probably becoming exhausted, until more efficient and affordable methodologies are developed for genotyping. Then, associations might reveal rare SNPs or structural variants such as small insertions or deletions (indels) that are not identified by standard GWAS. This discordance might result in the tag SNP being a proxy of a rarer event, hindering the identification of the causal genetic effector. The increasingly cheaper whole genome sequencing will contribute to the better fine-mapping of associations [78]. Also, other genomic structural variations such as copy number variants, have not been widely explored and might give insight into AF risk loci characterization [235].

2.2 Transcriptomic studies of Atrial Fibrillation

In recent years, numerous transcriptional studies have been carried out using both microarrays and RNA-seq to identify the genes responsible for the initiation and remodeling processes behind AF [217]. Those performed in human atrial tissue, and due to obvious practical and ethical concerns, have been restricted to individuals requiring surgical intervention [117]. Of those patients, most suffered from cardiovascular artery disease, valvular disease or heart failure, and in some cases, were under treatment for the previously mentioned or related comorbidities. In addition, although left atrial tissue has been demonstrated to undergo more profound remodeling, the difficulty in obtaining left atria samples has led to an overrepresentation of right atrial tissue in published transcriptomic studies. Moreover, the collection of samples at different time points of disease progression is impossible for obvious reasons. Because of these inherent limitations to the experimental design, the main questions concerning the transcriptional differences between sinus rhythm individuals and patients with paroxysmal or persistent AF remain largely unknown in the most relevant target tissue of the disease.

Nevertheless, these studies using human atrial tissue have provided important insight into the molecular mechanisms underlying AF. Most of the published work points towards similar pathways through which atrial remodeling occurs: ion channel alteration [40, 59, 70, 96, 236], contractile dysfunction [4], increased oxidative stress [104], development of cardiac fibrosis [1, 115, 211,

262], increased risk of thromboembolic events [115, 224, 236]) and inflammation [261]. However, when examining the degree of overlap of differentially expressed genes identified in the various studies, it is unexpectedly small (Figure 1.2). We examined publications using left [52, 96, 180, 211, 228, 236, 254, 260] and right atria [34, 96, 104, 106, 229], finding little overlap between genes identified as differentially expressed (Figure 1.3A). Most probably the disparities between experimental design, sample number, microarray platforms and use of different technologies (RNA-seq) explains in part the small coincidence observed. However, and while most of the genes are identified in only one of the publications examined, those found in at least three independent studies (10 genes for left atria, and 2 for right atria; Figure 1.3B) could be of interest. These include NPPB, coding for natriuretic peptide B and located next to NPPA on the genome, with whom it shares regulatory mechanisms [140]; GPR22 and RGS6, that code for G-protein signaling components; NTM, that codes for Neurotrimin, a promising novel marker of heart failure in patients [31]; or COLQ, that is found in at least three independent studies of both left or right atria, and codes for a variant collagen that anchors acetylcholinesterase to the basal lamina, and is strongly expressed in human atria as observed in GTEx data.

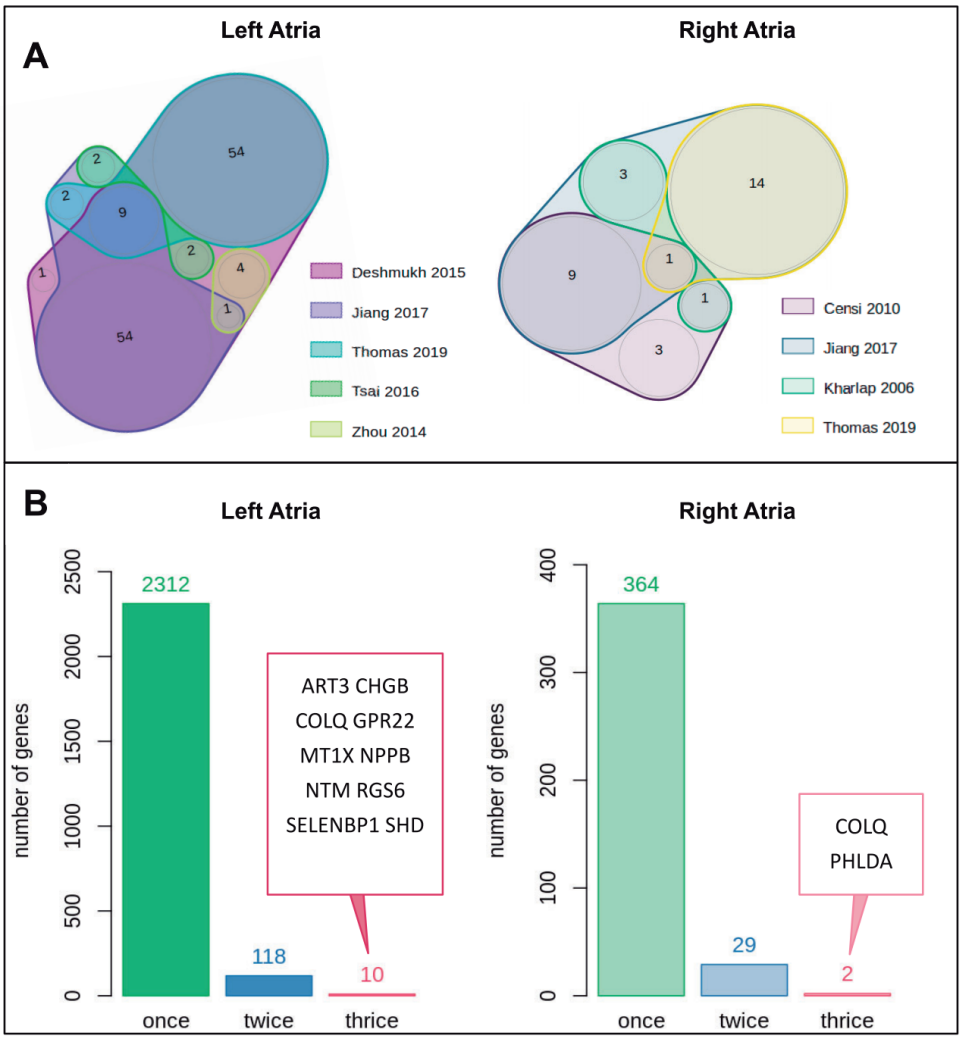


Figure 1.3: **Small overlap of transcriptomic studies of AF in humans.** (A) Multi-Venn diagram showing the overlap of genes reported as differentially expressed between sinus rhythm and AF cases (paroxysmal, susceptibility or persistent) in left or right atrial tissue. For clarity, we only used those genes identified in at least two studies to construct the overlaps. Colour legend indicates the study from which the genes were obtained. (B) Number of differentially expressed genes identified in one, two or three transcriptional studies of human samples from the left or right atria. Those gens identified in three independent studies are shown.

2.2.1 Computational approaches for transcriptomic studies

Since its development around 2007 [62, 127], RNA sequencing has become essential for transcriptome-wide analysis. Even though the arrival of new applications or variants (spatial transcriptomics, sinlge-cell transcriptomics, variants in read length, etc.), several gold-standards have been established for computational analysis. Above the 95% of publications use Illumina short-read sequencing technologies of cDNA, which implies extraction of RNA, cDNA synthesis, adaptor ligation, PCR amplification, sequencing and analysis [42, 216]. Analysis can be summarized in fourth steps,

Introduction

starting from the raw sequence reads, of one or two ends, generated by a sequencing platform. First, reads are pre-processed to clean adaptor contaminants or possible artifactual reads consequence of library preparation or sequencing itself, with tools such as *Cutadapt* [146] or *Trimmomatic* [25]. Those reads are then mapped to the reference genome, that can be either the whole genome sequence or simply the known transcriptome, depending on the experimental demands. The performance of the most popular aligners are *BWA* [122], *TopHat* [234], *Bowtie* [116] or *STAR* [54], have been subjected continuously to benchmarking, being *STAR* one of the aligners obtaining better evaluation scores [18, 226]. Next, to estimate gene expression, the simplest approach will be the counting of mapped reads onto genes. However, since read counts levels are affected by transcript and read length, total read number among other possible biases, normalization must be done to infer properly transcription abundances. Calculation of RPKMs (reads per kilobase of exon model per million reads), FPKMs (fragments per kilobase of exon model per million mapped reads) or TPMs (transcripts per million), is broadly use for that purpose. Other tools such as *RSEM* [120], compute ML abundance estimations using the Expectation-Maximization algorithm for its statistical model. The final step is usually differential expression analysis, which implies previous normalization considering total effective counts and the heterogeneity of the count distribution. Algorithms such as TMM [187] or *DESeq* [133] take that considerations into account. Many methods exist to account for differential expressed genes, and those have been extensively used in the literature, being *DESeq2* [133] and *voom-limma* [186], some of the best scored [15, 44].

2.3 Proteomics of Atrial Fibrillation

Proteomics includes several methodologies for the identification and quantification of proteins, being complementary to other massive techniques such as transcriptomics. Although transcriptomic profiling provides valuable information, mRNA and protein abundance do not necessarily show a high correlation [139]. Moreover, information regarding post-translational modifications (PTMs), which constitute and extra layer of regulation, can not be extracted from there. Deciphering the proteome status under normal and pathological becomes important for early diagnosis, prognosis and to monitor progression. However, proteomic profiling is particularly challenging at the methodological level. Protocols are themselves complex [36], protein dynamic range largely exceeds that of mRNA expression [245] and protein abundance is more variable through development, cell type or in response to stimuli than mRNA [225]. Furthermore, the mathematical and statistical tools for the analysis of large proteomic datasets are far less developed and standardized as compared to those used in transcriptomic studies.

To date, few proteomic studies have been conducted in the field of AF, and these result in patchy information. As mentioned in the previous section, the difficulties of sample collection, leads to inconsistency in the experimental design, reduced number of subjects and to the uneven representation of RA and LA tissue. Indeed, two-dimensional gel electrophoresis (2-DE) has been preferentially used as the proteomic separation technology, which, although providing relevant results, is fairly insensitive to low abundances, expensive and allows the study of a reduced number of proteins at a time. Initial publications [50, 114] identified a total of 3 proteins involved in contractility (MLC-2V, MLC-1 and MLC-2). Posterior studies [150, 258], suggested metabolic remodelling and mitochondrial dysfunction to have a role in the pathophysiology of AF, providing respectively, 17 and 223 candidates found diferentially expressed in LAA and RAA compared to SR patients. Another publication by [130], studied the differences between both atrial appendages in patients with rheumatic mitral valve disease. The authors found 17 proteins diferentially expressed both in the RA and LA compared with expression in SR samples, whereas 15 and 14, were different only in the RA or LA respectively. Obviously, these initial studies fell short of identifying the full range of changes of protein expression occurring during AF progression, resulting in a limited number of candidates, and a small overlap between differential expressed proteins, as already was reviewed [49, 218].

Since the development of gel-free mass-spectrometry-based proteomics (MS), some high-throughput experiments have been carried out in the field of AF. For instance, Doll and colleagues [55] using as a reference their map of the healthy human heart, identified 4147 proteins in total present in the LA samples of patients suffering AF. Of those, 104 were downregulated in AF samples, including collagens (COL1A2, COL3A1), mitochondrial, or contractile proteins (TNNT2, HRC, MYH6, SCN5A, and SRL). On the other hand, 307 were upregulated, involving processes such as ribonucleoprotein complexes and transcription. Despite the importance of AF proteomics, the small number of patients included, only 3, makes it difficult to figure out a general signature of the condition, specially if we take into account the large human genetics variability.More recently in a proteomic profiling of blood samples among 349 participants from the Framingham Heart Study Offspring, was carried out the identification of non-invaisve biomarkers of AF prognosis [111]. This study identified a total of 1373 proteins in the plasma proteome, associating ADAMTS13 and NT-proBNP to the incident AF.

2.3.1 Computational approaches for proteomic studies

Mass spectrometry-based proteomics (MS) approaches have the ability of interrogate the entire proteome, not only protein abundance, but also the peptidome or degradome, and can do it at the level of primary sequence, protein-protein interactions or post-translational modifications. MS has become the preferred method for studying complex protein samples [2, 8]. The most common MS-based method is bottom-up or shotgun proteomics [8, 149]. Here, the complex mixture of proteins is isolated and cleaved, typically by trypsin, into small peptides, to be subsequently fractionated using chromatography (LC), or other separation techniques. Eluted peptides from the column are ionized and transfer to the MS, where these are separated by their mass to charge ratio (m/z). The MS has the ability to be performed in tandem (MS/MS), meaning two consecutive acquisitions of MS. The first MS (MS1) is the mass spectrum of the intact peptide or precursor. Next, that isolated and ionized precursor is fragmented into smaller pieces, whose fragmentation pattern is analyzed (MS2). These measurements are recorded by a detector and transformed to intensities as a function of their m/z values. Therefore, the output or raw data, of a LC-MS/MS experiment, is a collection of mass spectrum (MS1 and MS2), reporting ion intensity at different m/z values. From the MS2, mass differences between peaks are correlated with residue masses and thus peptide sequence is deduced [45].

Peptide identification using MS2 data, utilizes genome annotation databases to query experimental mass estimations (experimental sequences) against theoretical mass data (theoretical sequences from the database), by a scoring function [67]. The goal here, is bring the most likely peptide to the top of the candidate list and output a set of scored psms (peptide-spectrum matches), along with the respective E-values estimates. Numerous scoring algorithms and different bioinformatic tools exists [75], as for instance Mascot [112], *Sequest* [219], X!tandem [23], *MaxQuant* [46], *QuiXoT* [182], *MSGF+* [108], etc. To validate and optimize peptide identification results, several post-processing algorithms [237] where developed, such as *PeptideProphet* or *Percolator* [227]. Next, FDR values are computed using the target-decoy approach, which means that target and a decoy (reverse or scrambled sequences) were both included in the database to later extract the false positive information [81]. Usually, psms below 1% FDR are considered for the further protein identification. The last critical step in the identification branch of the analysis, is protein inference, where psms are assembling to a list of confident proteins [89]. Two main problems arise at this step, peptides shared by multiple proteins (degerated peptides) and the existence of proteins only associated to one single peptide ("one-hit wonders"). Regarding generated peptides, it becomes complex to known to which protein or protein group they belongs and whether all the related pro-

teins are truly present. In the case of "one-hit wonders", those are simply not reliable, and proving their existence is challenging. Multiple algorithms provide different solutions to resolve this ambiguity, among them we have *ProteinProphet* [163], FIDO [204] or *EPIFANY* [178]. In light of this, it become apparent that the identification capacity is largely associated to the improvement genome annotation and the selection of a proper search database is crucial for the analysis [33].

There are several methods for quantification in the field of proteomics, that can be grouped primarily in labelled and label-free. The main difference between them is that in label-free the samples are run separated in different MS experiments, whereas labelled methods are combined prior to the MS run and consequently, more attention has to be put in the experimental design. A clear advantage is that it reduces technical variation in the experimental workflow [175]. Among the labelled techniques, a popular one is the isobaric chemical labeling using tandem mass tags (TMT) [256], which enables multiplexing several samples. For example, TMT-10 quantifies ten samples simultaneously. Basically, these isobaric tags, meaning all tags having the same mass, are added discriminatory to each sample or peptide digestion and then, samples are pooled together in the MS run. During the second fragmentation in the second MS step, sample-specific and pre-known reporter ions of different m/z values are generated for each tag, making possible the relative quantification from all the experimental conditions included. This strategy minimizes the number of missing peptide quantification values in each TMT experiment and takes advantage of the whole set of samples to obtain information for the peptide identification. Some care has to be taken with missing values at these level, specially if we take into account the subsequent steps of the protein identification branch explained before. Peptide abundances are relative to each psm and can not be compared in absolute terms, even for the same peptide sequence. Thereby, allowing missing values in some of the samples at this step, will propagate the error up to the estimation of protein abundances.

2.4 Data integration

Myriad high-throughput technologies and platforms are available to asses global transcripts, proteins, metabolites or players of the epigenomic regulation, resulting in several levels of quantitative data, that most of the times were generated separately. If the magnitude of this data in itself is complex to frame at the biological level, more complex become due to technical variation, which can even distort measurements of the same biological molecules. A major challenge is the integration of such multi-dimensional data, to properly comprehend the relationship among these layers. Usually, integration leads to better results from a statistical and a biological point of view [21].

When we confront high-dimensional data, the typical problem that arises is that points in the high-dimensional space are highly sparse distributed. The more number of features or dimensions, the more amount of data is required to generalize accurately, following an exponential relation. Dimensionality reduction techniques are employed to transform data from a high-dimensional space into a low-dimensional space, without losing the meaningful properties of the original data. Several approaches have been used in omics data, mainly for exploratory data analysis [153]. Among them, PCA is probably the most popular linear transformation, and tSNE or UMAP the most popular non-linear ones.

Methods proposed in literature for joint analysis, can be categorized as union of datasets [151], comparisons made at the functional level [176], topological network approaches regarding upstream regulators [179], merging of datasets in individual domains [76], missing value estimation by non-linear optimization [166], multiple regression analysis to predict contribution of sequence features in mRNA-protein correlation [165], clustering approaches such as coupled clustering [189], dynamic models based on Boolean networks, linear models or Bayesian networks [246, 249], etc.

Among the available procedures, we focus in multiple co-inertia analysis, which identifies co-relationships between multiple high dimensional datasets [152]. It starts from a set of tables where either features (for example genes) or measurements (for example experimental conditions) are matched and data is scaled to non-negative values. Then, a non-symmetric correspondence analysis (NSC) is performed to each dataset independently, transforming data onto the same lower dimensional space. The last step is a generalization of the Co-Inertia analysis, repeated until the desired number of principal components is obtained. The aim is maximizing the sum of the squared covariance between scores of each table with synthetic axes. Although this methods provide robust and consistent results [185, 212], further efforts in the improvement of protocols or experimental design have to be done to obtain a better understanding of the interactions between datasets and end up with standardize procedures.

2.5 Multilevel modelling for longitudinal data analysis

Longitudinal experiments involving repeated measurements of the same subjects over time are increasingly frequent in the cardiovascular field. In this type of data, observations belong to different clusters, groups or hierarchies and each cluster have its own properties, meaning different mean response, distinct sensitivity to predictors. Multilevel Models (MLM) or Hierarchical Models (HM)

are more suitable in general terms than classical linear regression, because are designed to handle this dependence among nested data structures.

Correlation among time is a common characteristic of repeated measurements on a given subject. We expect the values of consecutive time-points to be more similar than distant ones. Another important feature in longitudinal data is the heteroscedascity, or heterogeneous variability over time. These features violate the assumptions of independence and homoscedascity of more the standard statistical techniques such as linear regression. HM are more flexible and superior in this context, pooling information across clusters and thus providing better estimates for repeated sampling, better estimates when design of the study is unbalanced or being able to accommodate incomplete data. Additionally, HM provides estimates of the variation across subsamples and avoid the averaging by retaining variation [66].

HM with longitudinal data, allow us to model the behaviour of a dependent variable Y in which we are interested, said the protein abundance, whose changes in abundance may occur for clustered data, said through time and among different sheep replicates. There must be variables that change among sheep representing a certain level. However, these variables remain unchanged for the population, representing a higher level. HM recognise those hierarchies allowing for residual components at each level. In this example, part of the residual variance is accommodated into a between-sheep and within-sheep component. These between-sheep variability, leads to correlation between protein abundance measurements taken through time from the same sheep.

2.6 Experimental Models

Sample availability is a significant limitation when studying human cardiac diseases, which puts in value the use of other disease models instead. AF has been studied using a wide variety of cellular and animal models, all of which present advantages and disadvantages. While cultured cardiomyocytes and small animal models are suitable for genetic manipulation and require less time and costs, they do not reflect the complexity of a whole heart, do not develop AF, or have substantial differences in size and pace, respectively. Large animal models are much more similar to humans electrophysiologically despite presenting obvious disadvantages such as costs, space, nursing, lack of easy genetic modification tools, and ethical issues regarding their use. AF has been spontaneously detected in some animals such as horses, where it is found in regular animals (>2%) and more frequently in athletic horses [119, 170], and also dogs, a model in which atrial remodelling and inflammation have been extensively studied [92, 121, 255].

2.6.1 An *in vivo* model for studying AF

In this work, we made use of a well-established experimental model of intermittent atrial tachypacing, generated by [146]. A pacemaker was implanted subcutaneously on the RA with an unique atrial lead inserted in the RA appendage, and an implantable loop recorder was placed subcutaneously on the LA. Upon activation, the pacemaker induces AF by burst tachypacing, meaning that pacing is enabled 30 seconds at 20Hz, following that 10 seconds sensing. Pacemakers resumed pacing only if AF stopped and SR detection, while AF is not yet self-sustaining. We group sheep as transition, chronic or long-standing persistent AF and control based on the clinical classification of AF described before.

By measurement and monitoring of the dominant frequency (DF), we were able to identify the time at which AF stabilizes and sheep reaches the persistent AF [64, 146]. DF is the highest magnitude sinusoidal component extracted from the electrogram. It recognizes the fibrillatory signal with a higher frequency regarding its surroundings [71, 118], meaning the signals emerging from the reentrant sources, or rotors [141]. The rate of DF increase, predicts this timing of stabilization, which coincides with the time of persistent AF, and our transition group.

In this doctoral thesis project, we made use a wide range of different experimental and computational approaches, taking advantage of a model of tachypacing-induced long-standing AF in the sheep that will enable us to better understand the underlying molecular mechanisms behind AF and will offer an opportunity to improve prognosis, diagnosis and to translate our findings into the clinics.

Chapter 2

Objectives

In this doctoral thesis project, we aimed to understand the molecular networks behind atrial fibrillation (AF) and unravel the precise timings when transitions of the different AF forms take place in a model of tachypacing-induced long-standing AF in the sheep. With this general aim in mind, we specified the following objectives:

- To study the transcriptomic and proteomic changes during the progression of AF towards permanente states in atrial tissue and cardiomyocytes, understanding how these are connected to the electrophysiological and structural remodelling that takes place distinctly in both atrial appendages.
- To identify the transcriptomic changes taking place in the posterior left atrial wall at the time when persistent atrial fibrillation is reached, and compare them with those characterized in the left atrium appendage.
- To develop custom pipelines for data processing, integration and visualization of transcriptomic and proteomic data.
- To investigate the dynamics of the serum proteome in the time window of paroxysmal AF through the onset of persistent AF, locally in the heart and systemically in the peripheral circulation.
- To develop custom pipelines and statistical models for the proteomic and longitudinal data analysis.

Chapter 3

Materials and Methods

1 Experimental animals

We induced atrial fibrillation using a tachypacing device implanted in the right atrium of the sheep heart. Peacemaker implantation and pacing protocol were performed as described [64, 146]. Briefly, 6-8 month-old sheep (∼40 kg) instrumented with a subcutaneous pacemaker, with an atrial lead inserted transvenously into the right atria appendage. In addition, an implantable loop recorder (ILR) was placed subcutaneously on the left side of the sternum. Induction of anaesthesia was done with intravenous propofol (4-6 mg/kg), and maintained with isoflurane gas (5-10 ml/kg). After several days recovery, sheep were assigned to either control sham-operated or atrial tachypacing groups. The pacemaker was programmed to induce AF by burst tachypacing (30-sec pacing, 20 Hz, twice diastolic threshold) followed by 10-sec sensing. Pacemakers resumed pacing only if AF stopped and sinus rhythm was detected. The pacing device also recorded intracardiac electrograms (EGMs) to accurately confirm the occurrence of AF, generate histograms, and follow the evolution of AF from the first episode of paroxysmal AF to eventual confirmed establishment of persistent AF, at which time the pacemaker was switched off. We use the clinical definition of persistent AF [146], as episodes lasting longer than 7 days without reversal to SR. Three groups of animals were used: transition (13.75±4.50 days of self-sustained AF without reversal to sinus rhythm), chronic or long-standing persistent AF (289.25±50.29 days of self-sustained AF without reversal to sinus rhythm) and control group (sham operated or not operated animals, always in sinus rhythm). Animal procedures were approved by the University of Michigan Committee on Use and Care of Animals and conformed to National Institutes of Health guidelines.

2 Atrial tissue and cardiomyocyte isolation

For euthanasia, hearts were removed under anesthesia in the operation room by thoracotomy and placed in cold cardioplegic solution. Left and right atrial appendages (LAA and RAA, respectively) or posterior left atrial wall (PLA) were dissected. Samples for whole tissue analysis were snap-frozen. Cardiomyocyte isolation from both LAA and RAA was performed as previously described 2. Briefly, tissue samples were placed into a stock solution (120 mM NaCl, 5.4 mM KCl, 5mM MgSO4, 5 mM Pyruvate, 20 mM Glucose, 20 mM Taurine, 20 mM HEPES, 5 mM itrilotriacetic acid) and were cut with scissors into 1mm, 3 pieces. Pieces were shaken at 37º C for 12 min bubbling with 100% O2, changing the solution every 3 minutes, and transferring to a calcium free protease digestion solution for 45 minutes. After, protease digestion, pieces were transferred to collagenase type 1 (Worthington) and samples containing isolated cells were taken at 15, 30 and 45

minutes. Cardiomyocyte suspension was decanted and centrifuged 2 min at 500 rpm. Supernatant was discarded and pellets were resuspended in KB solution (50 mM L-Glutamic acid, 70 mM KOH, 30 mM KCl, 10 mM L-Aspartic acid-K, 10 mM KH2PO4, 2 mM MgSO4-7H2O, 20 mM Glucose, 20 mM Taurine, 5 mM Creatine, 0.5 mM EGTA, 10 mM HEPES) and centrifuged one more time. Supernatant was aspirated and pellets were snap-frozen in liquid nitrogen. Sample collection was conducted by members of the Jalife group in the Center for Arrhythmia Research, Michigan.

3 Plasma collection

Central blood was collected from the right atrium at the time of device implantation (baseline sinus rhythm) under general anaesthesia with isoflurane. A steerable sheath was inserted into the right jugular vein over a wire and positioned in the right atrium at the coronary sinus orifice under fluoroscopic guidance. Central blood was also collected at the time of euthanasia (persistent atrial fibrillation) during median sternotomy under deep sedation with propofol immediately prior to removal of the heart. Peripheral blood was collected weekly from conscious sheep via venipuncture of the cephalic vein. Both central and peripheral blood samples were centrifuged shortly after collection for separation of serum and plasma which were then aliquoted and frozen. Plasma collection was conducted by members of the Jalife group in the Center for Arrhythmia Research, Michigan.

4 RNA isolation and sequencing

We isolated RNA from 30-35 mg of LAA, RAA and PLA tissue using Quiagen RNeasy Mini Kit (#74106) following manufacturer's recommendations. Disruption and homogenization of the tissue was performed using a T10 ULTRA-TURRAX homogenizer (IKA Works Inc.) with 600 μl of RLT lysis buffer supplemented with β-mercaptoethanol. We isolated RNA from 2-3 frozen cardiomyocyte pellets using Quiagen RNeasy Mini Kit (#74106) following manufacturer's recommendation for samples with low cell number. In both cases, DNase digestion was performed during RNA extraction using Qiagen RNase-Free DNase Set (#79254). We checked tissue RNA concentration on a NanoDrop ND-1000 and cardiomyocyte RNA concentration was determined on a NanoDrop 2000. RNA quality was checked with an Agilent 2100 Bioanalyzer (Agilent Technologies, Palo Alto, CA). Using an Illumina Hiseq 2500 sequencer the CNIC Genomic Unit performed RNA-seq. Read quality was determined with *FASTQC* and adapter sequences were trimmed with *cutadapt* v1.14 4. *RSEM* v1.3.1 5 [120] was used to map reads against the reference sheep genome (Ovis aries v75 Ensembl) and to calculate estimated counts were used as RNA abundance metric. 3 sam-

ples from control, transition and chronic groups were used for LAA and RAA RNA-seq, while 6 samples from controls and transition groups were used for PLA RNA-seq. RNA isolation was conducted by Raquel Rouco, member of Manzanares group, and the sequencing protocol by the Genomic Unit at CNIC.

5 LC-MS/MS proteomics

Pellets of frozen cardiomyocytes were resuspended and homogenized in lysis buffer (10mM Tris-HCL pH7.4, 1 mM EDTA, 0.32 M Sucrose, 2% SDS, 50 mM DTT) supplemented with protease and phosphatase inhibitors (Roche), boiled for 5 min and cleared by centrifugation. Protein extracts were treated with 50 mM iodoacetamide (IAM) and digested with trypsin (FASP digestion kit, Expedeon). Peptides were labeled with 10 plex-TMT reagents (Thermo Fisher Scientific), desalted on OASIS HLB extraction cartridges (Waters Corp.), fractioned (High pH reversed-phase peptide fractionation kit, Thermo Fisher Scientific) and dried. Each fraction of the labeled peptide samples was analyzed using an Easy nano-flow HPLC system (Thermo Fisher Scientific) coupled via a nanoelectrospray ion source (Thermo Fisher Scientific, Bremen, Germany) to a Q Exactive HF mass spectrometer (Thermo Fisher Scientific, Bremen, Germany). C18-based reverse phase separation was used with a 2-cm trap column and a 50-cm analytical column (EASY column, Thermo). Peptides were loaded in buffer A (0.1% formic acid) and eluted with a 240 min linear gradient of buffer B (80% acetonitrile, 0.1% formic acid) at 200 nL/min. Mass spectra were acquired in a data-dependent manner, with an automatic switch between MS and MS/MS using a top 15 method. MS spectra were acquired in the Orbitrap analyzer with a mass range of 400–1500 m/z and 60,000 resolution. HCD fragmentation was performed at 27 of normalized collision energy and MS/MS spectra were analyzed at 60,000 resolution in the Orbitrap. 3 samples from control, transition and chronic groups were used. Experiments were conducted by the Proteomic Unit at CNIC.

5.1 QuiXoT

Peptide identification was performed using the probability ratio method [145], and FDR was calculated using inverted databases. The relative abundance of each protein was estimated from ion intensities of peptides with an FDR below 1% and expressed in units of standard deviation according to their estimated variances (Zq values), as previously described [161]. Obtention of Zq values was conducted by Proteomic Unit at CNIC.

5.2 Maxquant

The data from the 36 10-plex TMT experiments were analysed per batch using *Maxquant* [46]. The FDR threshold was set to 0.01 for the psm and protein level. Peptides and proteins were identified using the *Ovis aries* Uniprot (SwissProt & TrEMBL) reference besides our custom composite database. All settings were default with precursor-ion tolerance set to 20 ppm, maximun of 2-missed cleavages, full trypsin specificity. Cysteine carbamidomethylation was used as a fixed modification and acetylation of protein N-termini (TMT) and oxidation of methionine were selected as variable modifications. The run parameters are available. *Maxquant* was run using mono on a cluster under CentOS 7.

5.3 Custom pipeline

Our custom pipeline was developed under bash scripting, chaining *OpenMS tools* and primary software. All the scripts are available here.

6 Transcriptomics and proteomics data integration

Ensembl gene IDs and Ensembl protein IDs were used to map transcripts and proteins respectively against the sheep genome (Ovis aries v75 ensembl). We detected 13187 atrial tissue transcripts, 13262 cardiomyocytes transcripts and 7284 cardiomyocytes proteins. We considered only those features (genes or proteins) with homologous orthology to human with confidence equal to 1, and we only kept for further analysis the human Ensembl gene with the highest similarity for every unique ID, thus avoiding redundancy in the annotation and gaining accuracy. Therefore, we kept 11962 unique gene symbols from the human genome (11483 atrial tissue mRNA, 11318 atrial cardiomyocyte mRNA, and 3830 atrial cardiomyocyte proteins), corresponding to homologues of sheep genes. LC/MS-MS data was processed with *QuiXoT* [161] to obtain quantitative Zq. Statistical analysis to compare RNA-seq and LC/MS-MS data was conducted in R. We applied Multiple Co-Inertia Analysis [152] as a concatenation-based integration method for unsupervised dimensionality reduction. Previously, we normalized mRNA expression from atrial tissue and isolated cardiomyocytes, as well as protein abundance from cardiomyocytes. We arranged them in three matrices where the rows corresponding to the IDs and the columns correspond to the 18 samples: tissue mRNA, cardiomyocyte mRNA and cardiomyocyte proteins samples (each in triplicate) from each of the three experimental conditions (control, transition and chronic), and from LAA and RAA. After analysis of the resulting 18 principal components (PCs), we retained the

projected values for PC1-PC3, recapitulating 42% of the total variability of the original datasets.

7 Feature selection and Gaussian Mixture Models

Differential expression analysis was performed for each of the experiments (atrial tissue RNA-seq, atrial cardiomyocyte RNA-seq, atrial cardiomyocyte proteomics) using the *voom-limma* procedure in R [186]. Only those genes differentially expressed with an adjusted P-value of 0.05 (Benjamini-Hochberg correction) in at least one comparison were selected. Density-based spatial clustering of applications with noise (implemented in the *fpc* library in R (Flexible Procedures for Clustering) as *dbscan* function, with reachability distance of 0.4 and reachability minimum number of points equal to 12) was computed to finally select 10% of features that more strongly associated with the 3 projections selected from the co-inertia analysis. The union of these two criteria determined the features of interest (genes and proteins) to be used for all subsequent analysis. Gaussian Mixture Model (GMM) implemented in the R library *clusterR* was used for clustering the selected features by their projected PC values. The projected values were transformed to spherical coordinates for the clustering step, and then transformed back to cartesian coordinates for interpretation, using the *pracma* package in R (Practical Numerical Math Functions). The optimal cluster number was determined using the Bayesian Information Criterion.

8 GO term enrichment analysis

For Gene Ontology [10] and KEGG [9] functional enrichment analysis of the different gene clusters we used *clusterProfiler* library in R, retaining as significant those enriched terms with an adjusted P value of 0.05 (Bonferroni correction). Semantic similarity among GO terms, based on the Relevance method with a cutoff of 0.3, was estimated to reduce redundancy of the enriched terms using the *GoSemSim* package in R. Enrichments for plasma samples were conducted without including all the identified PGs as universe, otherwise no term was found enriched.

9 Analysis of Epigenetic modifiers

For analysis of epigenetic modifiers, we recovered 142 genes from the cardiomyocyte RNA-seq experiment out of the 166 genes annotated in the curated database of epigenetic modifiers, *dbEM* [158]. The log2 cpm (counts per million) expression values were plotted using the *heatmaply* package in R.

10 Transposable elements analysis

Transposable elements (TEs) were annotated in gtf format and manually curated with the help of the repeat masker track from the UCSC Genome Browser, based on the Aug 2012 Oar3.1 version of the sheep genome. To obtain the relative abundances of TEs, reads from the cardiomyocytes RNA-seq experiment were mapped and quantified using *STAR v.2.6.1a_08-27* [54] and *TEtranscripts v.2.0.3* [97] respectively. Differential expression between TEs was estimated with *DESeq2* [133]. Measurements of coverage were calculated based on this annotation file. Data related to the de novo annotations of sheep transposable elements is available upon request.

11 GWAS enrichment analysis

For gene set enrichment analysis of genome-wide association studies, we performed a hypergeometric test to assess the association between our set of selected features (4409) and the genes annotated in three different GWAS collections. Genes detected in any of our datasets were used as universe. The resulting p-values were adjusted for multiple testing (Benjamini-Hochberg procedure). The electrophysiological (668) and myocardial (212) GWAS sets of genes were obtained from the public GWAS Catalog hosted at the NHGRI-EBI. For that purpose, we collected annotations in the database of cardiovascular diseases with electrophysiological and myocardial phenotypes, respectively. A third GWAS collection (240) was obtained by merging the results from two recent meta-analysis of AF-associated genes [168, 191]. Gene lists as well as GWAS Catalog annotations used are available in Supplemental Table S4.

12 Analysis of PLA RNA-seq

Differential expression analysis, annotation and GO functional enrichment analysis were conducted as described above for the atrial appendage experiments, comparing in this case, control and transition sheep. We applied hierarchical clustering, Euclidean distance and method complete, on the 2185 differentially expressed genes and labeled the samples as control, fast and slow. Analysis of the chromatin signature was performed as described above.

13 Western Blot

Cells were harvested in RIPA buffer (50mM Tris-HCl pH 8, 150 mM NaCl, 0,1% SDS, 1% NP-40, 0,5 Sodium Deoxycholate) containing protease inhibitors (Complete ULTRA tablet; Roche; 06538304001). Protein concentration was quantified using BioRad DC protein assay (BioRad, 5000112) and 15 μg of each sample was resolved in 15% SDS-polyacrilamide gels. Proteins were then transferred to Polyvinylidene difluoride (PVDF) membrane (Immobilon©-P, Millipore, IPVH100010). Membranes were blocked with 5% non-fat dry milk in TBS-T (50 mM Tris-HCl pH 7.6, 150 mM NaCl, 0.1% Tween-20) and incubated overnight with the corresponding antibody (Histone H3: abcam; ab1791, Histone H4: Santa Cruz Biotechnology; sc-377520, Tnnt2: DSHB, CT3 supernatant). Bands were quantified using ImageJ and normalized to TNNT2. Experiments were conducted by MT. Experiments were conducted by Maria Tiana, member of Manzanares group.

14 Bayesian Hierarchical Modelling

14.1 Bayesian Inference

Probabilistic models were implemented in *brms* package [27], an interface in R to fit Bayesian generalized (non-)linear multivariate multilevel models using the probabilistic programming language *Stan* [32] on the back-end. As a considerable advantage, brms uses *lme4*-like formula syntax. The joint posterior distributions of the parameters were approximated using Hamiltonian Monte Carlo [57, 162] and its extension, the No-U-Turn Sampler (NUTS), which tends to scale particularly well for complex multilevel models. Whenever possible, we run 4 chains, along 4000 iterations, using 3000 of them for warming up. MCMC algorithm was evaluated at each occasion by inspection of R-hat convergence diagnostic and Bulk- and Tail-Effective Sample Sizes. All the models used the Gaussian likelihood function. As a good practice with brms, cause it presumes predictors are mean centered, we replaced the default intercept with a b Intercept parameter. All the scripts of this section are available here.

14.2 Mathematical Notation

We denoted PGs as k, time-point of the AF progression as j and individuals as i. We used γ terms for population-level effects, and ζ terms for group-level effects. The first subscript refers to L1 coefficient, and the latter to L2 coefficient, in numerical order. Those are prefixed respectively as

*b_* and a *r_* in the joint posterior distribution. Variance components were prefixed as *sd_* or *sigma* in the joint posterior distribution and noted as σ^2 in our models, with the corresponding subscripts. Note that for simplicity, we refer our variable of interest as *PGab*.

14.3 Model comparison

Model comparison was performed via Pareto smoothed importance-sampling leave-one-out cross-validation as an efficient way to approximate true LOO-CV [239]. We used the respective unconditional means model as null model for comparison. When the difference in ELPD is larger than twice the estimated standard error, the top model is expected to have better predictive performance than the bottom model. We use as threshold an 2 or 1.5 ELPD difference.

14.4 Proteins that change through progression

We accommodated the next L2 models and L3 models to address the question: How does protein abundance change over time.

14.4.1 L2 models

L2 models end up to estimate $PGab_{ij}$ which is the protein abundance at $j'th$ time-point for the sheep $i'th$. L2 specifies simultaneously a pair of subsidiary models: level-1 sub-model of the individual growth model, describing how each sheep -all the i's- changes over time with its individual growth parameters, and level-2 sub-model, describing how these changes differ across sheep. We specified both, L1 and L2, together with the composite model. In addition, the brms formula syntax is detailed below. Group-Level variance was not specified for any other parameter than the intercept Π_{0i}. The models version including a random slope was implemented, however in all the cases dropping the ζ_{10} we obtained a better performance. Although some of these models are not referenced in the result section, we included their specification for a better understanding of the L3 section.

Dataset

PG abundances, already standardized, were arranged as a tibble of an entry per PG, containing the corresponding protein/peptide intensity and metadata information. In the case of proteins we got 292 entries. Intensities from right atria samples are excluded of this analysis.

m.0 Constant or Unconditional Means model

The m.0 represents the "no change" trajectory, known as a polynomial function of zero order. Each trajectory is flat and different individuals can have different intercepts. Therefore, a collection of "no change" trajectories is enclosed in this model.

$L1: \quad PGab_{ij} = \Pi_{0i} + \epsilon_{ij}$

$L2: \quad \Pi_{0i} = \gamma_{00} + \zeta_{0i}$

$$PGab_{ij} = \gamma_{00} + \zeta_{0i} + \epsilon_{ij}$$

$$brms : \text{PGab} \sim 0 + \text{Intercept} + (1|\text{sheep})$$

We assume that the residual variance ϵ_{ij} follows:

$$\epsilon_{ij} \sim Normal(0, \sigma^2_\epsilon)$$

and that the variance component for the varying-intercept ζ_{0i} follows:

$$\zeta_{0i} \sim Normal(0, \sigma^2_0)$$

At level-1, Π_{0i} is the mean PG abundance of the sheep $i'th$ or intercept. At level-2, we capture the interindividual differences in change trajectories and the remaining unexplained variance. Composite specification of m.0, without predictors at any level, contains three model parameters: 1) an intercept γ_{00} representing the mean PG abundance across all time-points and sheep, 2) the residuals σ^2_ϵ which denotes the amount the abundance on occasion $j'th$ deviates from sheep i's mean, or in other words, the between-sheep variability and 3) σ^2_0, the amount sheep i's mean deviates from the population mean, or within-sheep variability. Note that m.0 stochastic components ζ_{0i} and ϵ_{ij} are preserved in the next models.

m.1 Linear or Unconditional Growth model

The "linear change" trajectory is known as a first order polynomial in time.

$$L1: \quad PGab_{ij} = \Pi_{0i} + \Pi_{1i} time_{ij} + \epsilon_{ij}$$

$$L2: \quad \Pi_{0i} = \gamma_{00} + \zeta_{0i}$$

$$\Pi_{1i} = \gamma_{10}$$

$$PGab_{ij} = \gamma_{00} + \gamma_{10} time_{ij} + \zeta_{0i} + \epsilon_{ij}$$

$$brms : \text{PGab} \sim 0 + \text{Intercept} + \text{time} + (1|\text{sheep})$$

The m.1 model introduces time as predictor $time_{ij}$ at level-1, assessing how much of the within-sheep variability can be attributed to systematic changes over AF progression. In other words, Π_{0i} is the Sinus Rhythm status of PG abundance of sheep i and Π_{1i} is sheep's the mean change per time-point in *PGab* of sheep $i'th$ during the AF progression. Now, the composite specification shows 4 model parameters to estimate, where γ_{00} means the mean PG abundance for the population in SR and γ_{10} is the population average of rate of change through progression, together with the stochastic components ζ_{0i} and ϵ_{ij}, already explained.

m.2 Second degree polynomial or Unconditional Quadratic Growth model

To capture nonlinear patterns of time, we include in the next models (m.2,m.3 and m.4), several level-1 predictors that collectively represent a polynomial function of time.

$$L1: \quad PGab_{ij} = \Pi_{0i} + \Pi_{1i} time_{ij} + \Pi_{2i} timeQ_{ij} + \epsilon_{ij}$$

$$L2: \quad \Pi_{0i} = \gamma_{00} + \zeta_{0i}$$

$$\Pi_{1i} = \gamma_{10}$$

$$\Pi_{2i} = \gamma_{20}$$

$$PGab_{ij} = \gamma_{00} + \gamma_{10} time_{ij} + \gamma_{20} timeQ_{ij} + \zeta_{0i} + \epsilon_{ij}$$

$$brms : \text{PGab} \sim 0 + \text{Intercept} + \text{time} + \text{time}^2 + (1|\text{sheep})$$

The m.2 model expands m.1 with $timeQ_{ij}$ as level-1 predictor. This second order polynomial change trajectory includes terms for both predictors, time and the square of time, in conjunction

with three growth parameters: 1) Π_{0i} the intercept, 2) Π_{1i} now the instantaneous rate of change when SR, and the new Π_{2i} which is called the curvature parameter and describes the change in the rate of change. The stochastic components, ζ_{0i} and ϵ_{ij}, remain the same.

m.3 Third degree polynomial or Cubic Growth model

$$
\begin{aligned}
L1: \quad & PGab_{ij} = \Pi_{0i} + \Pi_{1i} time_{ij} + \Pi_{2i} timeQ_{ij} + \Pi_{3i} timeC_{ij} + \epsilon_{ij} \\
L2: \quad & \Pi_{0i} = \gamma_{00} + \zeta_{0i} \\
& \Pi_{1i} = \gamma_{10} \\
& \Pi_{2i} = \gamma_{20} \\
& \Pi_{3i} = \gamma_{30}
\end{aligned}
$$

$$PGab_{ij} = \gamma_{00} + \gamma_{10} time_{ij} + \gamma_{20} timeQ_{ij} + \gamma_{30} timeC_{ij} + \zeta_{0i} + \epsilon_{ij}$$

$$brms: \text{PGab} \sim 0 + \text{Intercept} + \text{time} + \text{time}^2 + \text{time}^3 + (1|\text{sheep})$$

The m.3 model expands m.2 with $timeC_{ij}$ as level-1 predictor. Includes a 3 order power of time increasing the complexity of the polynomial trajectory and the new growth parameter Π_{3i}. The stochastic components ζ_{0i} and ϵ_{ij} remain the same.

m.4 fourth degree polynomial

$$
\begin{aligned}
L1: \quad & PGab_{ij} = \Pi_{0i} + \Pi_{1i} time_{ij} + \Pi_{2i} timeQ_{ij} + \Pi_{3i} timeC_{ij} + \Pi_{4i} timeF_{ij} + \epsilon_{ij} \\
L2: \quad & \Pi_{0i} = \gamma_{00} + \zeta_{0i} \\
& \Pi_{1i} = \gamma_{10} \\
& \Pi_{2i} = \gamma_{20} \\
& \Pi_{3i} = \gamma_{30} \\
& \Pi_{4i} = \gamma_{40}
\end{aligned}
$$

$$PGab_{ij} = \gamma_{00} + \gamma_{10} time_{ij} + \gamma_{20} timeQ_{ij} + \gamma_{30} timeC_{ij} + \gamma_{40} timeF_{ij} + \zeta_{0i} + \epsilon_{ij}$$

$$brms: \text{PGab} \sim 0 + \text{Intercept} + \text{time} + \text{time}^2 + \text{time}^3 + \text{time}^4 + (1|\text{sheep})$$

The m.4 model expands m.3 with $timeF_{ij}$ as level-1 predictor. Includes a 4 order power of time increasing even more the complexity of the polynomial trajectory and the new growth parameter Π_{4i}. The stochastic components ζ_{0i} and ϵ_{ij} remain the same.

m.5 General Time model

$$L1: \quad PGab_{ij} = \Pi_{0i} + \Pi_{1i}t1_{ij} + \Pi_{2i}t2_{ij} + \Pi_{3i}t3_{ij} + \Pi_{4i}t4_{ij} + \Pi_{5i}t5_{ij} + \Pi_{6i}t6_{ij} + \epsilon_{ij}$$

$$L2: \quad \Pi_{0i} = \gamma_{00} + \zeta_{0i}$$

$$\Pi_{1i} = \gamma_{10}$$

$$\Pi_{2i} = \gamma_{20}$$

$$\Pi_{3i} = \gamma_{30}$$

$$\Pi_{4i} = \gamma_{40}$$

$$\Pi_{5i} = \gamma_{50}$$

$$\Pi_{6i} = \gamma_{60}$$

$$PGab_{ij} = \gamma_{00} + \gamma_{10}t1_{ij} + \gamma_{20}t2_{ij} + \gamma_{30}t3_{ij} + \gamma_{40}t4_{ij} + \gamma_{40}t5_{ij} + \gamma_{40}t6_{ij} + \zeta_{0i} + \epsilon_{ij}$$

$$brms: \quad \text{PGab} \sim 0 + \text{Intercept} + \text{t.fact} + (1|\text{sheep})$$

$$\text{PGab} \sim 0 + \text{Intercept} + \text{t1 + t2 + t3 + t4 + t5 + t6} + (1|\text{sheep})$$

In the m.5 model, time was treated as a discrete variable, being each time point its own category. Note that the t.fact variable has its time point values saved as a factor, which simplify the code. This syntax is equivalent to a set of J-1 time-period dummy variables.

Our prior belief

We specified relatively informative prior distributions for the parameters described above, see Figure3.1. As long as our data was previously Z-scored, we set a normal prior with mean 0 and standard deviation 1 for the γ_{00}. We further set a Student's T prior distribution on the remaining population-level parameters, both for the polynomial terms and for the time-period parameters, with 3 degrees of freedom, mean 0 and standard deviation 1.5, assuming that changes in abundance over time are smooth and bounded to [-2.5,2.5]. We continued to use the Student's T for all the stochastic components, with the lower-limit set to 0 in these two cases.

$$p(\gamma_{00}) = Normal(\ \gamma_{00} \mid 0, 1\ )$$

$$p(\gamma_{i0}) = t(\ \gamma_{i0} \mid 3, 0, 1.5\ ) \quad where\, \gamma_{i0} = [\gamma_{10}, \gamma_{20}, ..., \gamma_{60}]$$

$$p(\sigma_0) = t(\ \sigma_0 \mid 3, 0, 1.25\ )$$

$$p(\sigma_\epsilon) = t(\ \sigma_\epsilon \mid 3, 0, 1.5\ )$$

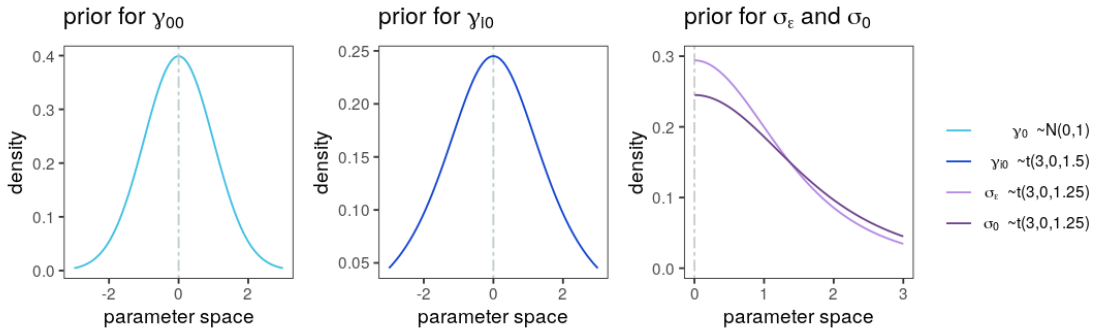


Figure 3.1: L2 priors

14.4.2 L3 models

The previous L2 models are expanded to L3 models by pooling protein-level information across time and replicates in a joint model. Thus, information across proteins is borrow to facilitate inference, with proteins at the highest level of the hierarchical model. Basically, previous L2 models are relocated in a L3 multilevel context, preserving the specification of the growth parameters of previous section. Here, we model $PGab_{ijk}$, which represents the PG abundance of the protein $k'th$, measured in the sheep $i'th$ at the time $j'th$. PG trajectories are now estimated as group-level effects instead of as population-level effects. We included two group-level terms for both, PG and PG:sheep (meaning each PG by each sheep), and growth parameters are included in the former, whereas the later accounts for the variability among sheep by PG in all our models as a single intercept. For simplification, we only include the brms syntax.

Dataset

PG abundances, already standardized, were arranged in a unique tibble structure containing the corresponding protein/peptide intensity and metadata information. In the case of proteins we got 292 entries. Intensities from right atria samples are excluded of this analysis.

M.0 Constant or Unconditional Means model

$$brms : \text{PGab} \sim 0 + \text{Intercept} + (1|\text{PG}) + (1|\text{PG:sheep})$$

A grand global mean for PG abundance over time, sheep and proteome, is represented by the Intercept, as the unique population-level parameter. This is common to all the subsequent L3 models

but M.6. The last term (1 | PG:sheep), accounts specifically for the variation across replicates for a given PG, and this one is common to all the subsequent L3 models. The variability among sheep per every PG is estimated with the term (1 | PG). We can extract the corresponding group-level effects of this level, to obtain the mean of PG abundance over time and sheep per every PG. These are the constant trajectories estimated by a constant L3 model. Annotated in brms syntax, the stochastic components of the model are: 1) the variation across the proteome (sd_PG_Intercept); 2) the variability among sheep per every PG (sd_PG:sheep_Intercept), and 3) the residual variance (sigma). The subsequent models, expanded M.0 progressively to accommodate the longitudinal variation inside parameters within the PG grouping factor.

M.1 Linear Growth model

$$brms : \text{PGab} \sim 0 + \text{Intercept} + (1 + \text{time}|\text{PG}) + (1|\text{PG:sheep})$$

M.1 expands M.0 by introducing time as predictor within the PG grouping factor as (1 + time | PG). Now at this level, the variance over the PGs abundances in SR and the variance over the rates of change through progression are estimated, additionally to its corresponding covariance. If we use ζ_{0i} and ζ_{1i} to notate these group-level parameters, the typical way to express the multivariate distribution would be:

$$\begin{pmatrix} \zeta_{0i} \\ \zeta_{1i} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{10} & \sigma_1^2 \end{pmatrix} \right]$$

We can extract the corresponding group-level effects of this PG level, to obtain the PG abundance at SR and the rate of change of each PG. These are the linear growth trajectories estimated by the M.1 model.

M.2 Second degree polynomial or Quadratic Growth model

$$brms : \text{PGab} \sim 0 + \text{Intercept} + (1 + \text{time} + \text{timeQ}|\text{PG}) + (1|\text{PG:sheep})$$

M.2 expands M.1 by introducing timeQ as predictor within the PG grouping factor as (1 + time + timeQ | PG). Now at this level, we additionally estimate the variance over the curvature parameters, together with to the corresponding covariances. If we use ζ_{0i}, ζ_{1i} and ζ_{2i} to notate these group-level parameters, the multivariate distribution would be:

$$\begin{pmatrix} \zeta_{0i} \\ \zeta_{1i} \\ \zeta_{2i} \end{pmatrix} \sim N \left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & \sigma_{01} & \sigma_{02} \\ \sigma_{10} & \sigma_1^2 & \sigma_{12} \\ \sigma_{20} & \sigma_{21} & \sigma_2^2 \end{pmatrix} \right]$$

We can extract the corresponding group-level effects of this PG level, to obtain the PG abundance at SR, the rate of change and the curvature parameter of each PG. These are the quadratic growth trajectories estimated by the M.2 model.

M.3 Third degree polynomial or Cubic Growth model

$$brms : \text{PGab} \sim 0 + \text{Intercept} + (1 + \text{time} + \text{timeQ} + \text{timeC}|\text{PG}) + (1|\text{PG:sheep})$$

M.3 expands M.2 by introducing timeC as predictor within the PG grouping factor as (1 + time + timeQ + timeC | PG). Now at this level, we additionally estimate the variance over the new growth parameters, together with to the corresponding covariances. If we use ζ_{0i}, ζ_{1i}, ζ_{2i} and ζ_{3i} to notate these group-level parameters, the multivariate distribution would be:

$$\begin{pmatrix} \zeta_{0i} \\ \zeta_{1i} \\ \zeta_{2i} \\ \zeta_{3i} \end{pmatrix} \sim N \left[\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & \sigma_{01} & \sigma_{02} & \sigma_{03} \\ \sigma_{10} & \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{20} & \sigma_{21} & \sigma_2^2 & \sigma_{23} \\ \sigma_{30} & \sigma_{31} & \sigma_{32} & \sigma_3^2 \end{pmatrix} \right]$$

We can extract the corresponding group-level effects of this PG level, to obtain the growth parameters of each PG. These are the cubic growth trajectories estimated by the M.3 model.

M.4 Fourth degree polynomial

$$brms : \text{PGab} \sim 0 + \text{Intercept} + (1 + \text{time} + \text{timeQ} + \text{timeC} + \text{timeF}|\text{PG}) + (1|\text{PG:sheep})$$

M.4 expands M.3 by introducing timeF as predictor within the PG grouping factor as (1 + time + timeQ + timeC + timeF | PG). Now at this level, we additionally estimate the variance over the new growth parameters, together with to the corresponding covariances. If we use ζ_{0i}, ζ_{1i}, ζ_{2i}, ζ_{3i} and ζ_{4i} to notate these group-level parameters, the multivariate distribution would be:

$$\begin{pmatrix} \zeta_{0i} \\ \zeta_{1i} \\ \zeta_{2i} \\ \zeta_{3i} \\ \zeta_{4i} \end{pmatrix} \sim N \left[\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & \sigma_{01} & \sigma_{02} & \sigma_{03} & \sigma_{04} \\ \sigma_{10} & \sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{20} & \sigma_{21} & \sigma_2^2 & \sigma_{23} & \sigma_{24} \\ \sigma_{30} & \sigma_{31} & \sigma_{32} & \sigma_3^2 & \sigma_{34} \\ \sigma_{40} & \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_4^2 \end{pmatrix} \right]$$

We can extract the corresponding group-level effects of this PG level, to obtain the growth parameters of each PG. These are the fourth degree growth trajectories estimated by the M.4 model.

M.5 General Time model

$$brms: \quad \text{PGab} \sim 0 + \text{Intercept} + (1 + \text{t.fact}|\text{PG}) + (1|\text{PG:sheep})$$

M.5 treats time as a discrete variable, being each time point its own category. We introduced t.fact within the PG grouping factor as (1 + t.fact | PG). On this occasion, we estimate the variance over the PGs abundances in SR and every variance over the 6 discrete-time parameters, additionally to the corresponding covariances.

If we use ζ_{0i}, ζ_{1i},... to ζ_{6i} to notate these group-level parameters, the multivariate distribution would be:

$$\begin{pmatrix} \zeta_{0i} \\ \zeta_{1i} \\ \zeta_{2i} \\ \zeta_{3i} \\ \zeta_{4i} \\ \zeta_{5i} \\ \zeta_{6i} \end{pmatrix} \sim N \left[\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & \sigma_{01} & \sigma_{02} & \sigma_{03} & \sigma_{04} & \sigma_{05} & \sigma_{06} \\ \sigma_{10} & \sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} & \sigma_{15} & \sigma_{16} \\ \sigma_{20} & \sigma_{21} & \sigma_2^2 & \sigma_{23} & \sigma_{24} & \sigma_{25} & \sigma_{26} \\ \sigma_{30} & \sigma_{31} & \sigma_{32} & \sigma_3^2 & \sigma_{34} & \sigma_{25} & \sigma_{26} \\ \sigma_{40} & \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_4^2 & \sigma_{45} & \sigma_{46} \\ \sigma_{50} & \sigma_{51} & \sigma_{52} & \sigma_{53} & \sigma_{54} & \sigma_5^2 & \sigma_{56} \\ \sigma_{60} & \sigma_{61} & \sigma_{62} & \sigma_{63} & \sigma_{64} & \sigma_{65} & \sigma_6^2 \end{pmatrix} \right]$$

We can extract the corresponding group-level effects of this PG level, to obtain the rate change of each time-point regarding SR, for each PG. These are the discrete-time trajectories estimated by the M.5 model.

M.6 General Time with global time predictors

$$brms: \quad \text{PGab} \sim 0 + \text{Intercept} + \text{t.fact} + (1 + \text{t.fact}|\text{PG}) + (1|\text{PG:sheep})$$

This M.6 model expands the M.5, by adding population-level time predictors. In addition to M.5 parameters, the time-point parameters are estimated globally for the whole proteome.

Extracting PG trajectories from M.6

To obtain the averaged trajectories of PGs from M.6 model, the corresponding parameter estimates were added: Intercept (t0) was added to every single time coefficient (t1,t2,t3,t4,t5 and t6). We calculated these 6 abundances at the population-level (global parameters), and per every group-level term (meaning every particular PG). Finally, each population parameter was added to its equivalent PG parameter. Note that parameter estimates were added for each sampling of the posterior probability distribution, to calculate in the later step, the proper mean, standard deviation, and Credible Interval (CI) within the parameter falls with a probability of 95%.

Our prior belief

We specified relatively informative prior distributions for the parameters described above, see Figure3.2. As long as our data was previously Z-scored, we set a normal prior with mean 0 and standard deviation 1 for the global intercept (b_Intercept). We further set a Student's T prior distribution on the population-level time-period parameters of M.5 and M.6 (b_), with 3 degrees of freedom, mean 0 and standard deviation 1.5, assuming that changes in abundance over time are smooth and bounded to [-2.5,2.5]. We continued to use the Student's T for all the stochastic components, with the lower-limit set to 0 in these two cases (sd_ and sigma_). For the correlation among the group-level variance parameters (cor_), we used the Lewandowski-Kurowicka-Joe (LKJ) distribution setting η as 0.7, expecting values through time to be fairly correlated.

$$p(b\_Intercept) = Normal(\,0, 1\,)$$

$$p(b\_) = t(\,3, 0, 1.5\,)$$

$$p(sd\_) = t(\,3, 0, 1.25\,)$$

$$p(sigma\_) = t(\,3, 0, 1.5\,)$$

$$p(cor\_) = LkjCorr(\ 0.7\ )$$

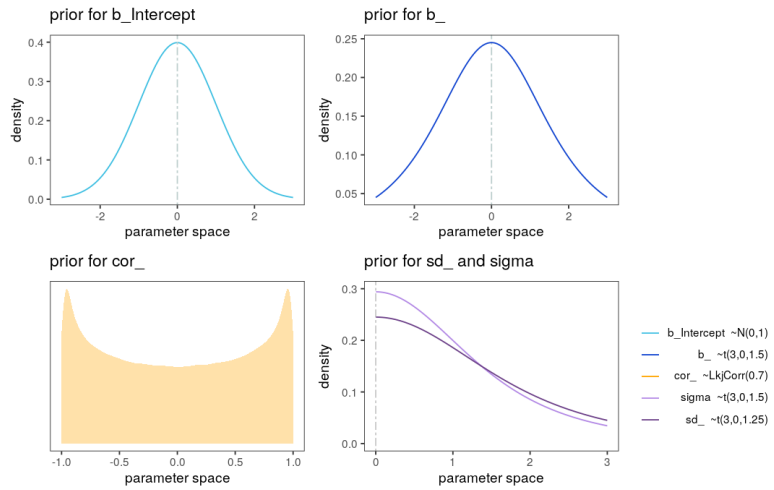


Figure 3.2: ...

14.4.3 Building a final model for proper comparison

m.6 General Time model with M.6 covariance as prior

The model m.6 its a variant of the m.5 model, which takes advantage of the information about the whole proteome through time and sheep, learned from the M.6 model by pooling together all the identified PGs. The variance and covariance components of the PG grouping factor were extracted from M.6. This information, together with the population-level time-point estimates, was set as a multivariate normal prior to fit the m.6 model. The multivariate normal prior follows a $X \sim Normal_J(\mu, \Sigma)$, where μ is a the column vector containing the global means for each time-point and Σ the corresponding covariance matrix, both extracted from M.6.

$$\begin{pmatrix} t0 \\ t1 \\ t2 \\ t3 \\ t4 \\ t5 \\ t6 \end{pmatrix} \sim N \left[\begin{pmatrix} \mu_{t0} \\ \mu_{t1} \\ \mu_{t2} \\ \mu_{t3} \\ \mu_{t4} \\ \mu_{t5} \\ \mu_{t6} \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & \sigma_{01} & \sigma_{02} & \sigma_{03} & \sigma_{04} & \sigma_{05} & \sigma_{06} \\ \sigma_{10} & \sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} & \sigma_{15} & \sigma_{16} \\ \sigma_{20} & \sigma_{21} & \sigma_2^2 & \sigma_{23} & \sigma_{24} & \sigma_{25} & \sigma_{26} \\ \sigma_{30} & \sigma_{31} & \sigma_{32} & \sigma_3^2 & \sigma_{34} & \sigma_{25} & \sigma_{26} \\ \sigma_{40} & \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_4^2 & \sigma_{45} & \sigma_{46} \\ \sigma_{50} & \sigma_{51} & \sigma_{52} & \sigma_{53} & \sigma_{54} & \sigma_5^2 & \sigma_{56} \\ \sigma_{60} & \sigma_{61} & \sigma_{62} & \sigma_{63} & \sigma_{64} & \sigma_{65} & \sigma_6^2 \end{pmatrix} \right]$$

14.5 Differences between Right atrium and Peripheral blood

14.5.1 L2 models

L2 models end up to estimate $PGab_{ij}$ which is the protein abundance at $j'th$ time-point for the sheep $i'th$. Again, we specified both, L1 and L2, together with the composite model. In addition, the brms formula syntax is detailed below. Our prior belief for this section was set as before with L2 models.

For this purpose, we include a new level-2 predictor in the subsequent models, which is loc_i, a categorical predictor that takes the value 0 when blood is peripheral and 1 when it was sampled from the right atrium.

$$loc_i = \begin{cases} 0 & \text{if blood i is collected peripherally} \\ 1 & \text{if blood i is collected at the right atrium} \end{cases}$$

Dataset

PG abundances, already standardized, were arranged as a tibble of an entry per PG, containing the corresponding protein intensity and metadata information. In the case of proteins we got 292 entries. Only initial and final intensities from preipheral and RA samples are included in this analysis.

l.0 Constant or Unconditional means model

$$L1: \quad PGab_{ij} = \Pi_{0i} + \epsilon_{ij}$$

$$L2: \quad \Pi_{0i} = \gamma_{00} + \zeta_{0i}$$

$$PGab_{ij} = \gamma_{00} + \zeta_{0i} + \epsilon_{ij}$$

$$brms: \text{PGab} \sim 0 + \text{Intercept} + (1|\text{sheep})$$

Model l.0 reassembles the m.0 model including only the data specified above. This constant models assumes no changes through AF progression, without differences between RA and peripheral blood.

We assume that the residual variance ϵ_{ij} follows:

$$\epsilon_{ij} \sim Normal(0, \sigma^2_\epsilon)$$

and that the variance component for the varying-intercept ζ_{0i} follows an univariate normal distribution:

$$\zeta_{0i} \sim Normal(0, \sigma^2_0)$$

l.1 Constant or Conditional means model

$$L1: \quad PGab_{ij} = \Pi_{0i} + \epsilon_{ij}$$

$$L2: \quad \Pi_{0i} = \gamma_{00} + \gamma_{01} loc_i + \zeta_{0i}$$

$$PGab_{ij} = \gamma_{00} + \gamma_{01} loc_i + \zeta_{0i} + \epsilon_{ij}$$

$$brms: \text{PGab} \sim 0 + \text{Intercept} + \text{loc} + (1|\text{sheep})$$

Model l.1 expands l.0 by adding the level-2 predictor loc_i. When γ_{01} is equal to 1, meaning that blood was collected from the right atrium, the estimated loc_i value is added to the RA abundance. l.1 accounts for differences between RA and peripheral blood that are constant over progression.

l.2 Linear or Unconditional growth model

$$L1: \quad PGab_{ij} = \Pi_{0i} + \Pi_{1i} time_{ij} + \epsilon_{ij}$$

$$L2: \quad \Pi_{0i} = \gamma_{00} + \zeta_{0i}$$

$$\Pi_{1i} = \gamma_{10}$$

$$PGab_{ij} = \gamma_{00} + \gamma_{10} time_{ij} + \zeta_{0i} + \epsilon_{ij}$$

$$brms: \text{PGab} \sim 0 + \text{Intercept} + \text{time} + (1|\text{sheep})$$

Model l.2 reassembles the m.1 model, including only the data specified above. This linear model assumes a linear growth over AF progression, without differences between RA and peripheral blood.

l.3 Linear or Conditional Growth model -global effect of loc-

$$L1: \quad PGab_{ij} = \Pi_{0i} + \Pi_{1i} time_{ij} + \epsilon_{ij}$$

$$L2: \quad \Pi_{0i} = \gamma_{00} + \gamma_{01} loc_i + \zeta_{0i}$$

$$\Pi_{1i} = \gamma_{10}$$

$$PGab_{ij} = \gamma_{00} + \gamma_{01} loc_i + \gamma_{10} time_{ij} + \zeta_{0i} + \epsilon_{ij}$$

$$brms: \text{PGab} \sim 0 + \text{Intercept} + \text{loc} + \text{time} + (1|\text{sheep})$$

The l.3 expands the l.2 model, adding the possibility of find differences between RA and peripheral at SR time, with the term $\gamma_{01} loc_i$.

l.4 Linear or Conditional Growth model -loc interacts with time-

$$L1: \quad PGab_{ij} = \Pi_{0i} + \Pi_{1i} time_{ij} + \epsilon_{ij}$$

$$L2: \quad \Pi_{0i} = \gamma_{00} + \zeta_{0i}$$

$$\Pi_{1i} = \gamma_{10} + \gamma_{11} loc_i$$

$$PGab_{ij} = \gamma_{00} + \gamma_{10} time_{ij} + \gamma_{11} loc_i time_{ij} + \zeta_{0i} + \epsilon_{ij}$$

$$brms: \text{PGab} \sim 0 + \text{Intercept} + \text{time} + \text{time:loc} + (1|\text{sheep})$$

The l.4 expands the l.3 model, adding the possibility of find differences between RA and peripheral at pAF time, with the term $\gamma_{11} loc_i time_{ij}$.

l.5 Linear or Conditional Growth model - global effect of loc and interaction with time-

$$L1: \quad PGab_{ij} = \Pi_{0i} + \Pi_{1i} time_{ij} + \epsilon_{ij}$$

$$L2: \quad \Pi_{0i} = \gamma_{00} + \gamma_{01} loc_i + \zeta_{0i}$$

$$\Pi_{1i} = \gamma_{10} + \gamma_{11} loc_i$$

$$PGab_{ij} = \gamma_{00} + \gamma_{01} loc_i + \gamma_{10} time_{ij} + \gamma_{11} loc_i time_{ij} + \zeta_{0i} + \epsilon_{ij}$$

$$brms: \text{PGab} \sim 0 + \text{Intercept} + \text{loc*time} + (1|\text{sheep})$$

The l.5 joints l.3 and l.4 models, adding the possibility of find differences between RA and periph-

eral before AF has started and at the time of pAF.

14.6 Measuring the magnitude of change over progression and location

g.1 magnitude of changes L2

The g.1 expands the m.5 model, where time was treated as a discrete variable, adding the level-2 predictor loc_i. We used as syntax, a set of J-1 time-period dummy variables instead of the categorical variable t.fact, to model the proper interaction of the predictor location with the time-point 6.

$$
\begin{aligned}
L1: \quad & PGab_{ij} = \Pi_{0i} + \Pi_{1i}t1_{ij} + \Pi_{2i}t2_{ij} + \Pi_{3i}t3_{ij} + \Pi_{4i}t4_{ij} + \Pi_{5i}t5_{ij} + \Pi_{6i}t6_{ij} + \epsilon_{ij} \\
L2: \quad & \Pi_{0i} = \gamma_{00} + \gamma_{01}loc_i + \zeta_{0i} \\
& \Pi_{1i} = \gamma_{10} \\
& \Pi_{2i} = \gamma_{20} \\
& \Pi_{3i} = \gamma_{30} \\
& \Pi_{4i} = \gamma_{40} \\
& \Pi_{5i} = \gamma_{50} \\
& \Pi_{6i} = \gamma_{60} + \gamma_{61}loc_i
\end{aligned}
$$

$$
PGab_{ij} = \gamma_{00} + loc_i + \gamma_{10}t1_{ij} + \gamma_{20}t2_{ij} + \gamma_{30}t3_{ij} + \gamma_{40}t4_{ij} + \gamma_{50}t5_{ij} + \gamma_{60}t6_{ij} + loc_i t6_{ij} + \zeta_{0i} + \epsilon_{ij}
$$

$$
brms: \text{PGab} \sim 0 + \text{Intercept + loc + t1 + t2 + t3 + t4 + t5 + t6 + t6:loc} + (1|\text{sheep})
$$

G.1 dummies included L3

We expand this g.1 model to a three level model pooling all the proteins together in the G.1. Again for simplicity we only include the brms syntax.

$$
\begin{aligned}
brms: \quad & \text{PGab} \sim 0 + \text{Intercept + loc + t1 + t2 + t3 + t4 + t5 + t6 + loc:t6 +} \\
& (1 + \text{loc + t1 + t2 + t4 + t5 + t6 + t6:loc}|\text{PG}) + (1|\text{PG:sheep})
\end{aligned}
$$

14.6.1 Extracting PG trajectories from G.1

To obtain the averaged trajectories of PGs from G.1 model, the corresponding parameter estimates were added: Intercept (t0) was added to every single time coefficient (t1,t2,t3,t4,t5 and t6). Next, Intercept+loc to obtain t0 abundances in the cavity (t0c), and loc+t6+t6:loc for t6 cavity values (t6c). We calculated these 8 abundances at the population-level (global parameters), and per every group-level term (meaning every particular PG). Finally, each population parameters was added to its equivalent PG parameter. Note that parameter estimates were added by each sampling of the posterior probability distribution, to calculate in the later step, the proper mean, standard deviation, and Credible Interval (CI) within the parameter falls with a probability of 95%.

14.6.2 Probability calculation of differentially expressed proteins

To count the time-points that change in contrast to SR in peripheral blood, we calculated the corresponding probabilities of that event being different than 0 within a credibility of the 95%. First, the population-level parameters of each sampling were added to their equivalent group-level parameters (meaning by PG). If the mean of the resulting distribution is positive, the number of samples below zero was divided by the total number of samples of the posterior distribution. Thus, when the resulting value is smaller than 0.05, the PG was considered upregulated in contrast to SR at peripheral blood. Instead, if the mean of the distribution is negative, the number of samples above zero was divided by the total number of samples of the posterior distribution. Thus, when the resulting value is smaller than 0.05, the PG was considered downregulated in contrast to SR at peripheral blood.

15 Clustering of longitudinal data

Clustering of PGs was performed using the longitudinal data partitioning algorithm *klmShape* [73]. This method is a variation of the k-means algorithms in which a shape-respecting distance and a shape-respecting mean are used. Number of clusters was restricted to 2.

16 Annotation and identification of Uncharacterized proteins

Data from the Human Protein Atlas was downloaded here and crossed against the corresponding PG candidates. Gene symbol entry was used to intersect the data. The Human Protein Atlas database includes information regarding the distribution of the proteins across all major tissues and organs,

expression of protein-coding genes in single cell types, categorizes proteins detected in the blood cell types and proteins secreted by human tissues and subcellular localization of proteins in single cells among others. Uncharacterized protein were searched against the NCBI Protein Reference Sequences database by using the *blastp (protein-protein BLAST)* algorithm.

Chapter 4

Results

1 Transcriptome and proteome mapping of the sheep atria

1.1 Experimental Design

To understand the dynamics of expression that take place during the progression of AF, we took advantage of a well-established model of tachypacing-induced long-standing AF in the sheep [64, 146]. We sampled tissue from both the LAA and RAA from three male sheep each as follows: control, transition (1–2 weeks of persistent self-sustained AF in the absence of tachypacing), and chronic (more than 10 months of self-sustained persistent AF) groups (Figure 4.1A and Table S1). We used LAA and RAA whole-tissue samples for transcriptomic profiling by RNA-seq. We also isolated CMs from LAA and RAA samples and performed both RNA-seq and LC-MS/MS proteomic profiling. Sample collection and cardiomyocyte isolation was conducted by members of the Jalife group in the Center for Arrhythmia Research, Michigan, whereas RNA extractions were performed by Raquel Rouco and Western Blots by Maria Tiana, both members of the Manzanares.

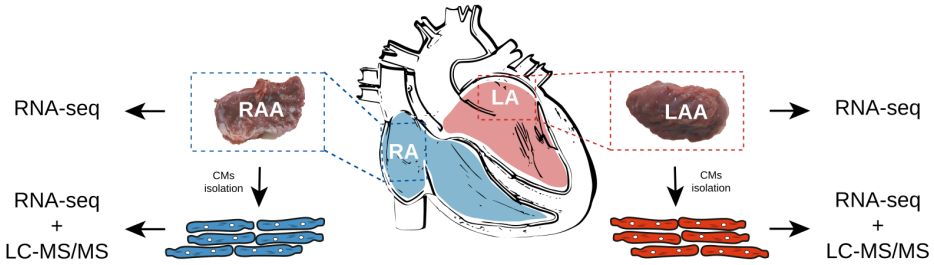


Figure 4.1: **Schematic diagram of the experimental strategy.** Three samples were collected in each case and for each analysis. Collected samples from left (LA) and right (RA) atria and the analysis that were carried out.

1.2 Distinct molecular changes occur rapidly at the transition to early persistent AF

Overall, we detected 13187 and 13262 transcripts from whole tissue (Figure 4.2) and CMs samples (Figure 4.3), respectively, of which a total of 1367 and 2479 genes were differentially expressed (5% false discovery rate, FDR) in at least one of the comparisons carried out (Table S2). On the other hand, we identified 7283 proteins in CMs (Figure 4.4), from which 581 had significant differential abundances (5% FDR). We initially performed comparisons of both transcriptomic and proteomic profiles between control, transition and chronic groups from LAA and RAA. Pairwise correlations of LAA tissue RNA-seq showed that changes to the control condition increase with disease progression (from the transition to the chronic group); however, when we compared transition and chronic groups, we recovered no differentially expressed genes (Figure 4.2). We observed

the same trend for the comparisons of both RNA-seq (Figure 4.3) and LC-MS/MS (Figure 4.4) of LAA CMs. As for the RAA, changes for both tissue and CMs were much lower but followed the same trend and again we detected no changes between transition and chronic states. Therefore, this initial analysis suggests that the LAA undergoes more profound changes during its progression to persistent AF than the RAA. It is also noteworthy that no transcriptomic or proteomic changes are observed from early persistent AF (transition) to long-standing persistent AF (chronic) in LAA or RAA for any of the comparisons.

To explore further the above observation, we compared the degree of change for all expressed genes and proteins (measured as log-fold changes) between control and transition groups to that of control versus chronic or transition versus chronic. In whole atrial appendage tissue RNA-seq, there is a positive linear relationship between the changes occurring from control to transition and control to chronic, indicating that the same trend in expression variation occurs along disease progression (Figure 4.2), bottom left panels). However, these changes level out and lead to a flat or even slightly descending relation when we compare transition to chronic states (Figure 4.2), bottom right panels). We observe the same behaviour for both RNA-seq and proteomic data from purified CMs (Figure 4.3 and 4.4, bottom panels).

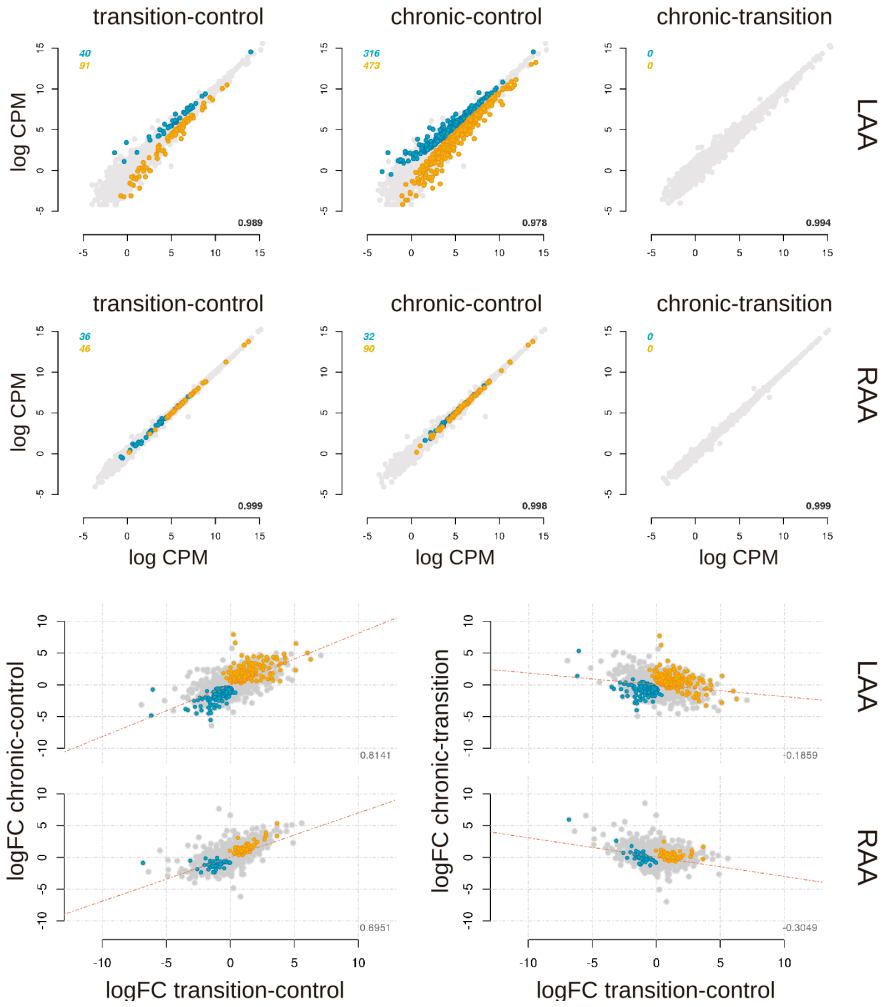


Figure 4.2: **Atria RNA-seq.** Upper panels show the correlation of mean expression values as log of counts per million (CPM) in atrial tissue RNA-seq between transition and control (left panel), chronic and control (middle panel) and chronic and transition (right panel) from left (LAA) and right (RAA) atrial appendages. Blue and yellow indicate up- and downregulated genes for each comparisons. The Pearson coefficient of correlation is indicated on the lower right corner of each plot. Lower panels depict the progression of changes in gene expression in atrial tissue along persistent AF. Shown is the linear regression adjustment of control-to-transition logFC (fold-change) to those of control-to-chronic (left panels) and transition-to-chronic (right panels), in left (upper panels) and right (lower panels) atrial tissue. Blue and yellow indicate up- and downregulated genes. The R2 value is indicated on the lower right corner of each plot.

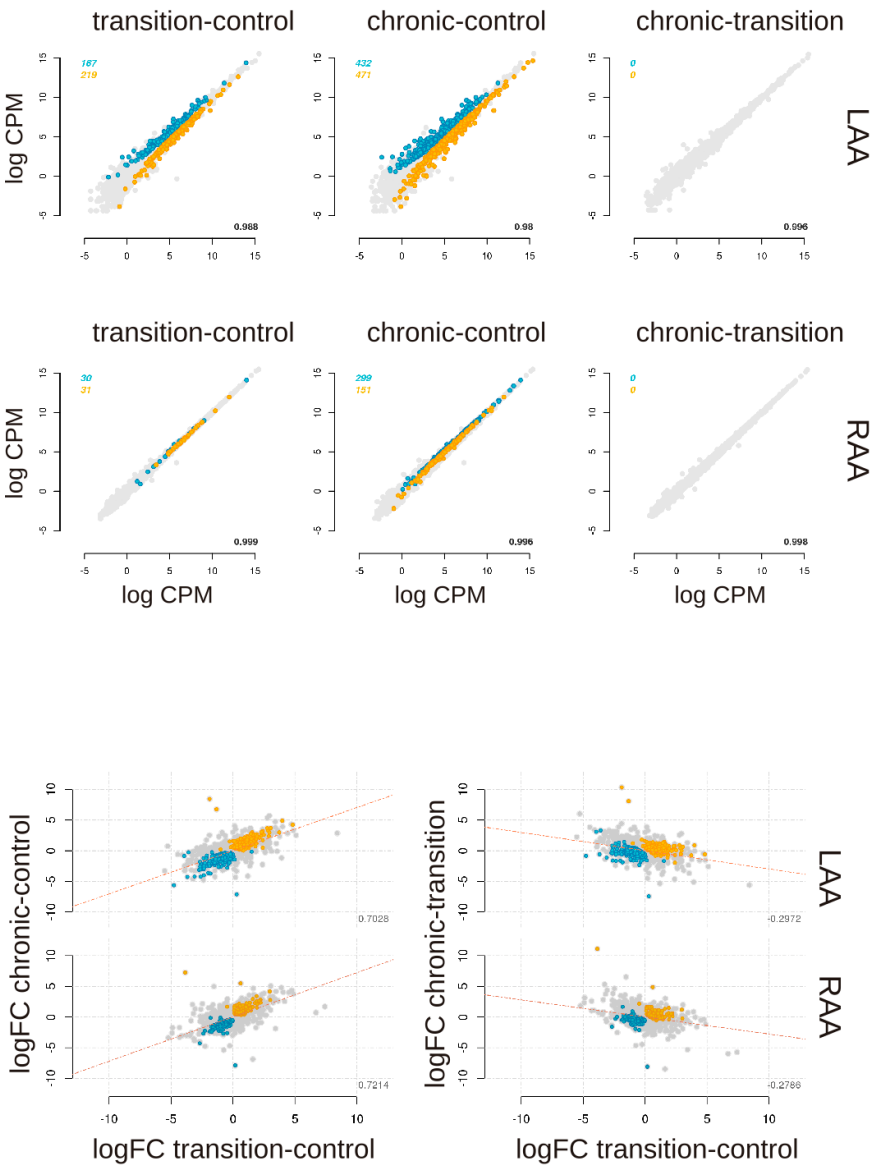


Figure 4.3: **Cardiomyocyte RNA-seq**. Upper panels show the correlation of mean expression values as log of counts per million (CPM) in cardiomyocyte RNA-seq between transition and control (left panel), chronic and control (middle panel) and chronic and transition (right panel) from left (LAA) and right (RAA) atrial appendages. Blue and yellow indicate up- and downregulated genes for each comparisons. The Pearson coefficient of correlation is indicated on the lower right corner of each plot. Lower panels depict the progression of changes in gene expression in cardiomyocytes along persistent AF. Shown is the linear regression adjustment of control-to-transition logFC (fold-change) to those of control-to-chronic (left panels) and transition-to-chronic (right panels), in left (upper panels) and right (lower panels) cardiomyocytes. Blue and yellow indicate up- and down-regulated genes. The R2 value is indicated on the lower right corner of each plot.

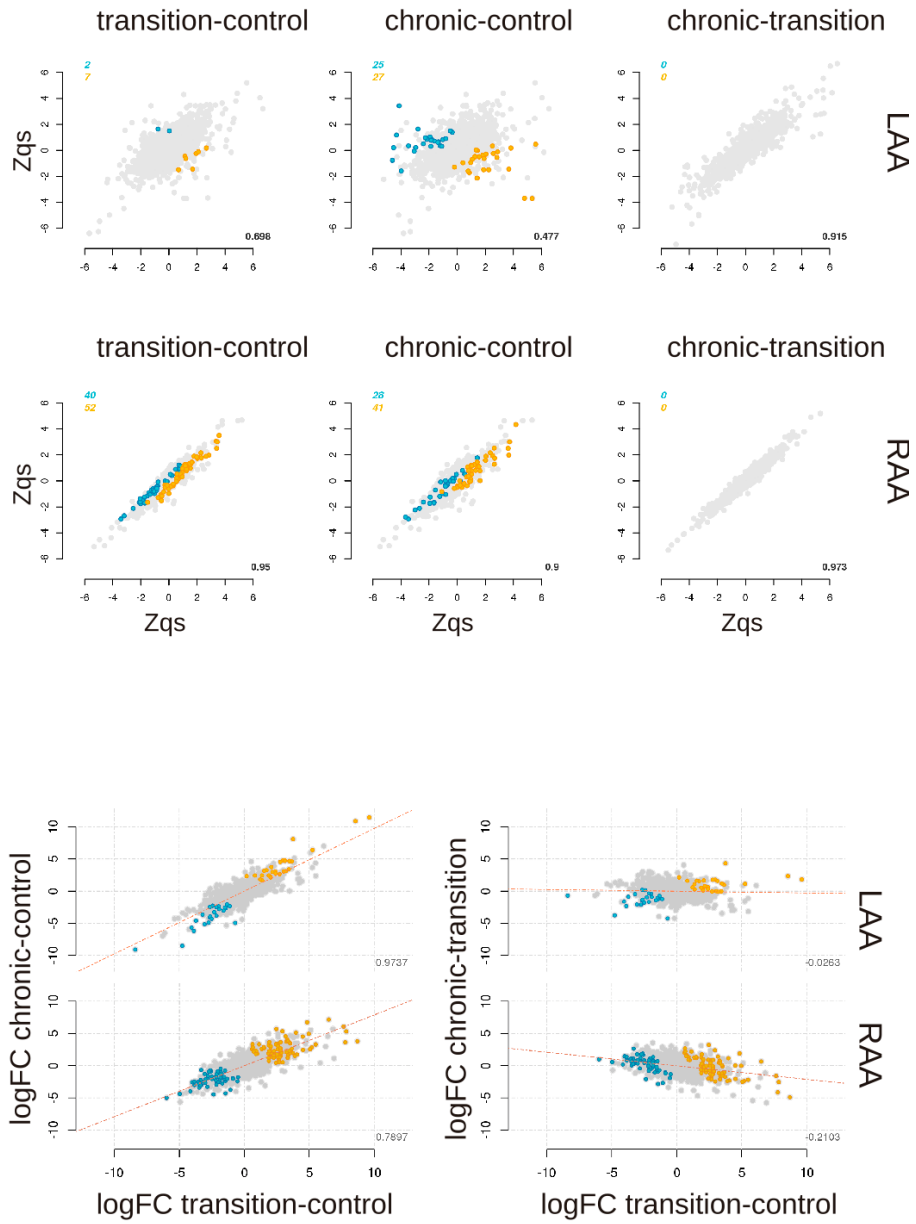


Figure 4.4: **Cardiomyocyte LC-MS/MS**. Upper panels show the correlation of mean expression values as relative protein abundance (Zqs) in cardiomyocyte LC-MS/MS between transition and control (left panel), chronic and control (middle panel) and chronic and transition (right panel) from left (LAA) and right (RAA) atrial appendages. Blue and yellow indicate up- and down-regulated genes for each comparisons. The Pearson coefficient of correlation is indicated on the lower right corner of each plot. Lower panels depict the progression of changes in gene expression in cardiomyocytes along persistent AF. Shown is the linear regression adjustment of control-to-transition logFC (fold-change) to those of control-to-chronic (left panels) and transition-to-chronic (right panels), in left (upper panels) and right (lower panels) cardiomyocytes. Blue and yellow indicate up- and downregulated genes. The R2 value is indicated on the lower right corner of each plot.

This analysis thus confirms that the mayor events related to AF progression occur during early phases of the disease and later stabilize as the animal moves from the transition towards the chronic state. We believe this pattern could not be secondary to mayor morphological changes taking place during AF progression, such as the increase in atrial dimensions that occur by dilation of the myocardium.[146] Measurement of the differences in atrial area shows that it increases steadily from control to chronic sheep, and does not level out from transition to chronic as does occur with transcriptional and proteomic changes (Figure 4.5, top panel). Remarkably, the temporal dynamics of the above changes in gene and protein expression corresponded closely with the electrical and structural remodelling we demonstrated previously in the sheep [146], where dominant frequency measurements in chronic (long-standing persistent AF) sheep did not significantly change compared with the values recorded at transition (early persistent) stage (Figure 4.5, lower panel).

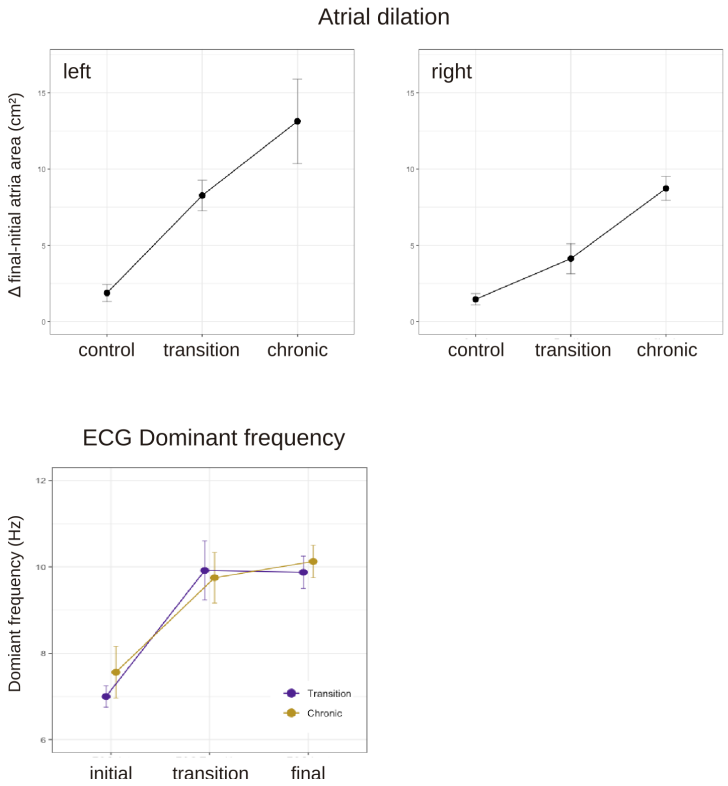


Figure 4.5: **Increase in atrial area and dominant frequency in a sheep model of AF progression**. In Upper panels, atria dilatation (cm2) calculated as the difference between final and initial atrium area in control (n=4), transition (n=7) and chronic (n=4) sheep for both left and right atria. Lower panel shows the Dominant Frequency (Hz) measured through surface ECG during AF progression. n=3 for transition, n=4 for chronic.

To assess how well changes in mRNA and protein expression correlated in our system, we compared log fold-changes between control and transition, as well as control and chronic stages of disease progression from transcriptomic and proteomic data we generated from LAA and RAA CMs. Correlation was relatively low but evident (Figure 4.6), and in line to what has been previously reported

[203, 242].

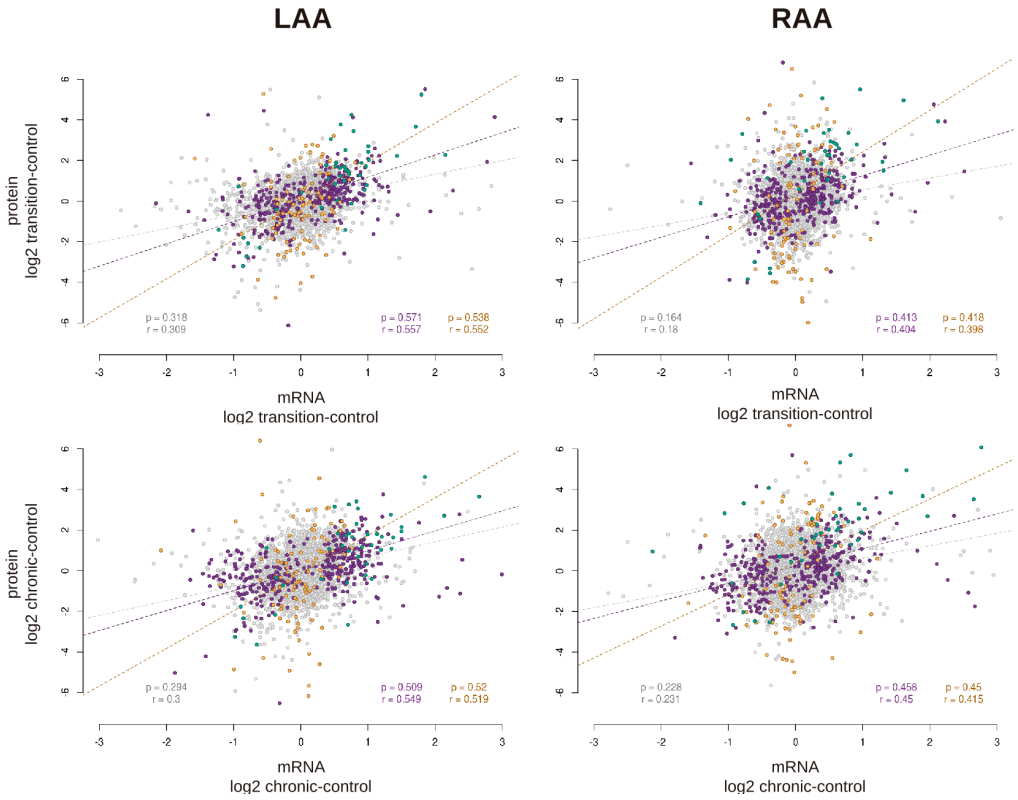


Figure 4.6: **Data variability and correlation of transcriptomic and proteomic data from a sheep model of AF progression.** Relative gene expression in transition (upper panels) or chronic (lower panels) compared to control cardiomyocytes from LAA (left panels), modestly correlates with their paired relative protein abundance, in line with previous reports. Correlation coefficients from the RAA cardiomyocytes (right panels) are slightly lower. Gray dots indicate common features between RNA-seq and LC-MS/MS data, purple are differentially expressed genes (DEGs), orange differentially expressed proteins (DEPs), and green are common DEGs and DEPs. Pearson (r) and Spearman (p) coefficients of correlation are indicated on the lower part of each plot, using the same color code as above.

1.3 A three-component model explains molecular variation during AF progression

To understand these early changes in the expression dynamics, we performed an integrative data analysis. Given the complexity of the data and the multiple layers of information (control, transition, and chronic disease states; LAA and RAA samples; whole atrial appendage tissue and isolated CMs; transcriptomics and proteomics; Figure 4.2), we also applied dimensionality reduction methods after data integration [153]. We first analysed each technical dataset (atrial appendage tissue RNA-seq, CM RNA-seq, and CM LC-MS/MS) separately by non-symmetric correspondence analysis, which transforms each dataset into a series of unsupervised lower dimensional units. In this way, we obtained 18 principal components that behaved in a similar fashion in the three datasets (Figure 4.7A). To integrate the three experiments, we performed multiple co-inertia analysis [152]

among datasets obtaining a ranked order of pseudo-eigenvalues that explained the variability of the data.

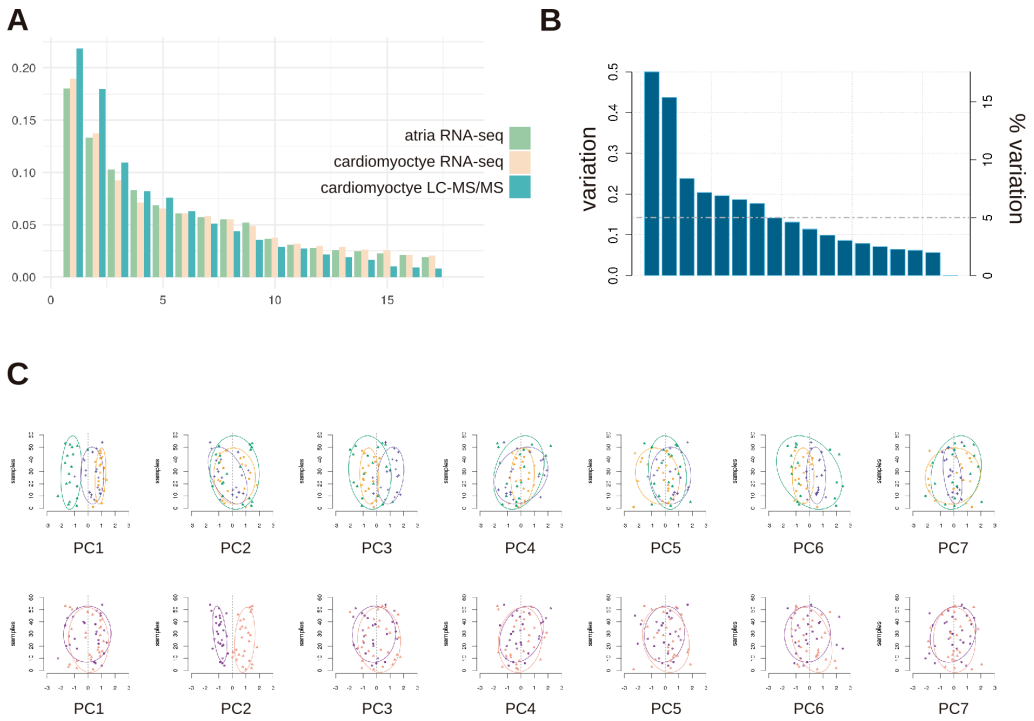


Figure 4.7: **Dimensionality reduction of transcriptomic and proteomic data.** (A) Eigenvalues (y-axis) of each separate experiment (atrial tissue RNA-seq, cardiomyocyte RNA-seq and cardiomyocyte LC-MS/MS) identified with non-symmetrical correspondence analysis. (B) Pseudo-eigenvalues after maximization of the squared covariance between experiments by co-inertia analysis. Amount of variability contained in that PC (left) and percentage of variance explained (right) are indicated on the y-axis. (C) Characterization of the topmost principal components and grouping of individual samples based on experimental design. From left to right, contribution of the seven principal components (PC1-PC7) that explain more than 5% of variance each. In the top row, samples are shape and color coded for disease state (control, green triangles; transition, purple crosses; chronic, yellow circles); in the bottom row, by tissue of origin (left, purple circles; right, yellow triangles). Ordering of the samples along the y-axis is arbitrary

We next explored the biological sources of variation that could underlie each of these components, and found that the progression of the disease from control to chronic explains the first principal component (PC1), accounting for 17.6% of the total variation in the data (disease progression; Figure 4.8A,B, 4.7C). The second component (PC2, 15.4% of variation) represents the regional differences between LAA and RAA (left/right identity; Figure 4.8A, 4.7C). Interestingly, we observed that the third component (PC3, 8.4% of variation) groups together samples from control and chronic individuals, separating them from transition individuals (transition state; Figure 4.8B). This was an unexpected finding, but it reinforced our previous observations (see above) that the mayor

changes occurring during disease progression occur during the transition from paroxysmal to persistent AF. We could not identify other possible effects that would account for further components of the variation (Figure 4.7)C), so we decided to use the first three components, which together explain more than 40% of the total variability, to model and analyse the molecular mechanisms underlying AF progression.

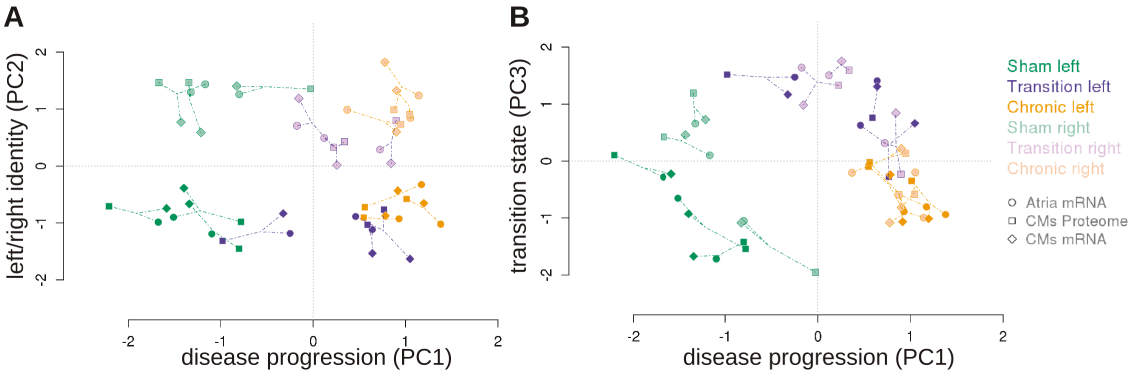


Figure 4.8: **Co-inertia analysis of multidimensional data identifies of the main components that drive variability in the sheep AF model.** (A) Distribution of transcriptomic and proteomic samples (three in each case) in relation to principal components PC1 (disease progression) and PC2 (left/right identity). Lines connect paired samples, obtained from the same individual. Control, green; transition, purple; chronic, yellow. LAA samples, dark colors; RAA samples, light colors. Atrial tissue RNA-seq, circles; cardiomyocyte RNA-seq, diamonds; cardiomyocyte LC-MS/MS, squares. (B) Distribution of transcriptomic and proteomic samples in relation to components PC1 (disease progression) and PC3 (transition state). Legend as in A.

1.4 Defining the molecular features responsible for atrial divergence and disease progression

We subsequently characterized the molecular changes occurring during AF progression using two different criteria to select features (genes and/or proteins) from our RNA-seq and proteomic analysis. On one hand, we selected features showing differential expression in any of the pairwise comparisons (Table S2 and Figure 4.9A, B) significantly detected at 5% FDR (Benjamini-Hochberg correction). This resulted in a list of 3278 differentially expressed genes and proteins (DEG/P; Table S2 and Figure 4.9D). On the other hand, we selected features showing extreme values (outer 10%) in our three-component space (Figure 4.9C), which totalled 1790 extreme-value genes and proteins that were highly variant on PC1, PC2 and/or PC3 (exG/P; Table S2 and Figure 4.9D). Both groups overlapped in 658 features, including key factors with atrial-restricted expression (such as PITX2 or BMP10) or related to atrial physiology (IL6R, KCNN2, NPPA, or RCAN1). In total, we retained 4410 features for further analysis.

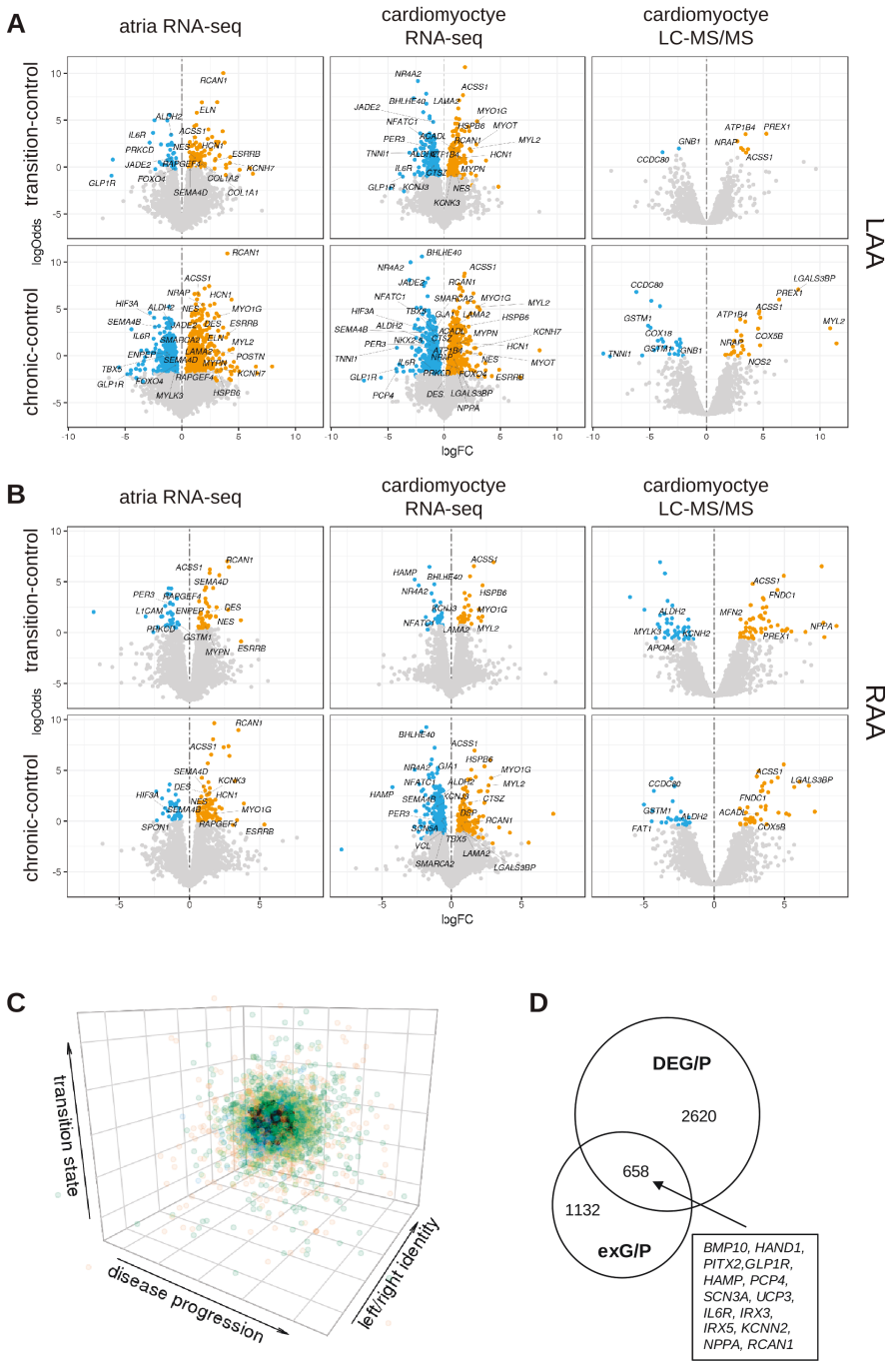


Figure 4.9: **Selection of differentially expressed and extreme-value genes and proteins for further analysis.** (A, B) Volcano plots of transition versus control (upper row) and chronic versus control (lower row) pairwise comparisons for atria RNA-seq /left panels), cardiomyocyte RNA-seq (middle panels) and cardiomyocyte LC-MS/MS (right panels), from left (A, LAA) and right (B, RAA) atrial appendage. Differentially expressed genes or proteins at 5% FDR are shown in orange (upregulated in transition or chronic as compared to controls) or blue (downregulated in transition or chronic as compared to controls). (C) Distribution in the 3-component pseudo-eigenvalue space of all features, where those with 10% extreme values are shown in color (green, atria RNA-seq; blue, cardiomyocyte RNA-seq; orange, cardiomyocyte LC-MS/MS). (D) Venn diagram showing number of features and overlap between differentially expressed genes or proteins (DEG/P) and features with extreme values (exG/P). Representative features included in the overlapping set are listed.

To better understand the biological relevance of these features and their behaviour, selected features were analysed by unsupervised clustering using GMM and separated into 31 independent clusters (g_0-g_30; Table S2) distributed along the three-component space as defined above (Figure 4.10). The majority of clusters showed contribution to more than one component, but those contributing to the third component (PC3, transition state) axis were less abundant.

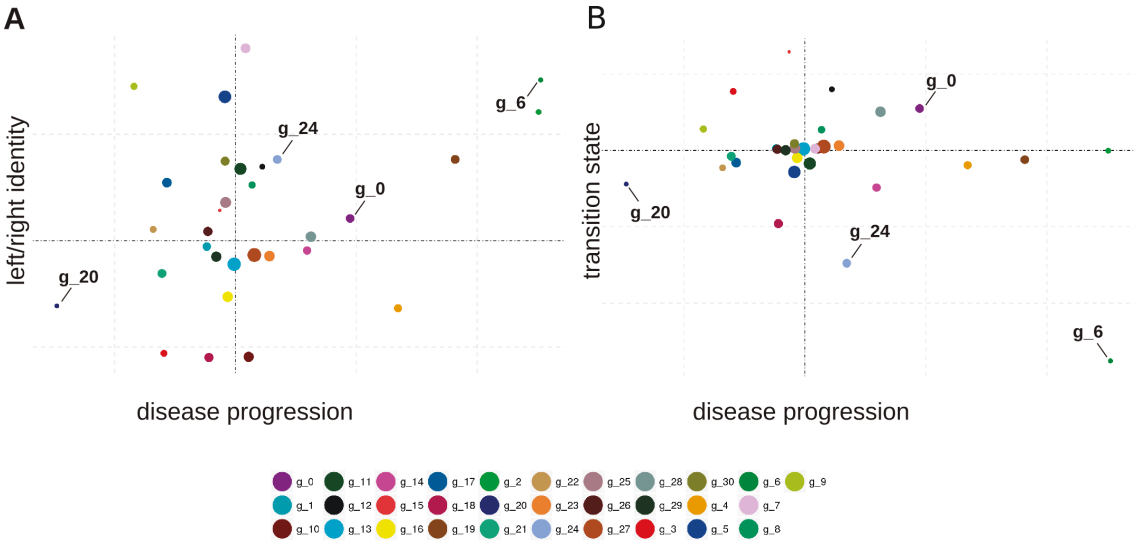


Figure 4.10: **GMM clustering.** Position of each of the thirty one clusters identified by GMM unsupervised clustering along the axis that define disease progression and left/right identity (A) or transition state (B). The size of each cluster represented on the plot correlates to the number of features (genes and proteins) that it includes. Colour legend is shown below. Arrows indicate the position of representative clusters (see next Figure.11)

Examination of mean expression of genes or proteins from selected clusters showed direct relationship to the location of the cluster in the three-component space (Figure 4.11). For example, cluster 0 (g_0) that includes among others the genes coding for the calcineurin regulator RCAN1 or Galectin-3 binding protein (LGALS3BP), contributes to disease progression and to a lesser extent to left/right identity and to transition state. Expression levels increase in all conditions from control to transition, and remain mainly stable from transition to chronic (Figure 4.11A). Cluster g_6 is interesting in that it shows an extreme position along all three components (Figure 4.11B) and includes PITX2, a left atrial marker and the gene most strongly associated to AF by GWAS [77], or PCP4, a calmodulin regulator with specific expression in cardiac Purkinje cells [107]. Cluster g_20 is mostly related to disease progression, showing a trend for down-regulation of expression (Figure 4.11C). This cluster includes receptors for glucagon-like peptide 1 (GLP1R) or relaxin (RXFP1), both related to the control of insulin secretion. Cluster g_24, where we find key features related with atrial physiology such as the potassium channel gene KCNH2, or the nuclear receptor gene NR4A3, is strongly associated with the transition state. As such, in CM RNA-seq,

we observe diminished expression in transition samples as compared to controls, which does not occur in chronic sheep (Figure 4.11D). Thus, this three-component model, allows us to explore the contribution of significative features to better understand the progression of AF.

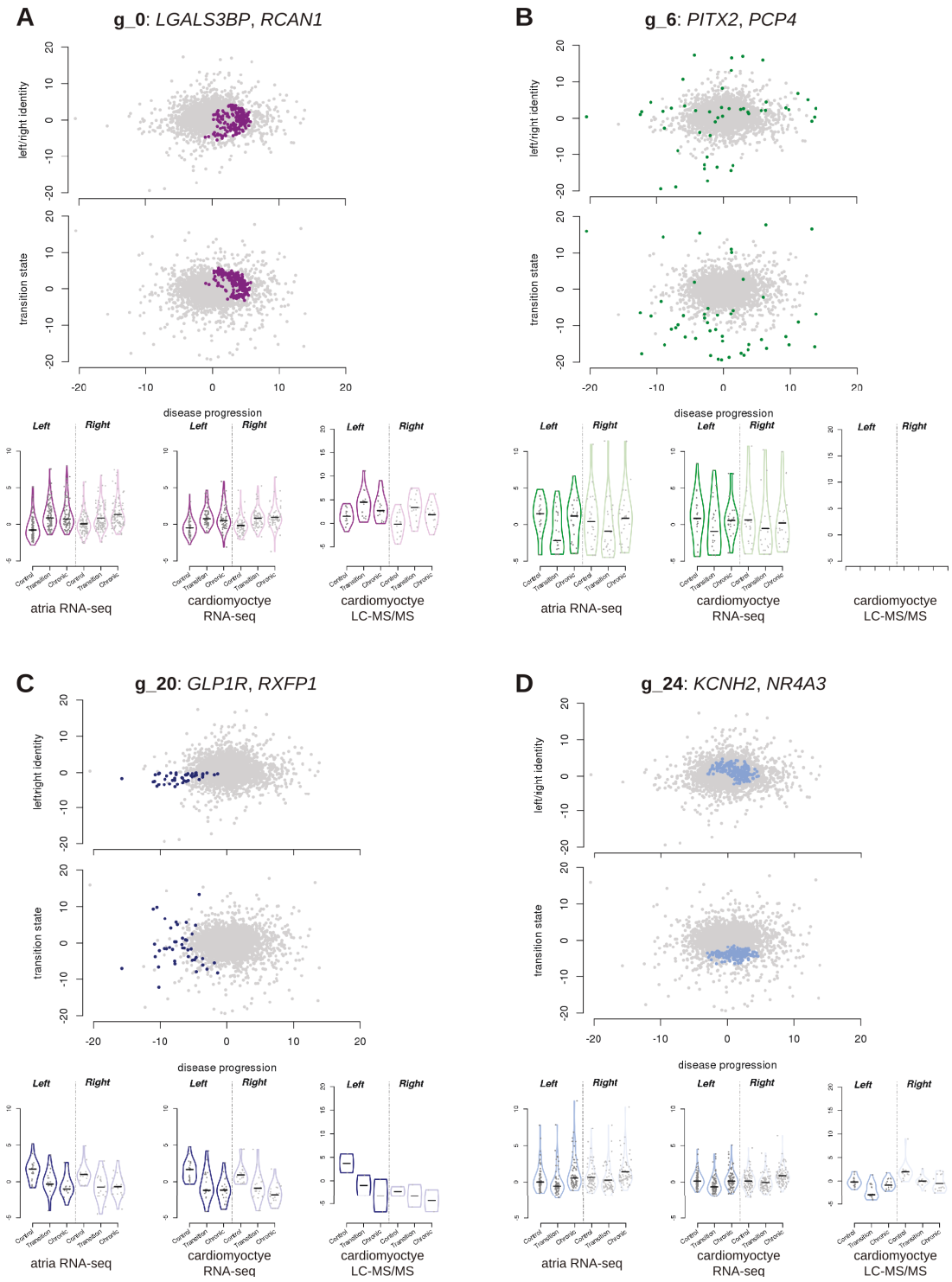


Figure 4.11: **Distribution and expression of representative GMM clusters in the 3-component space of AF progression.** The position of all individual features of the specified GMM clusters (A, g_0; B, g_6; C, g_20; D, g_24) along the disease progression axis and left/right identity (top) or transition state (middle). Below, violin plots depicting the expression of the features from the specified cluster in each individual experiment (atria RNA-seq, cardiomyocyte RNA-seq, and cardiomyocyte LC-MS/MS), condition (control, transition and chronic) for both LAA and RAA; mean expression is indicated by a horizontal black line. No proteomic data was available for cluster g_6. Colour legend of the GMM clusters is as in Figure 4.10

1.5 Distinct genetic programmes underlie cell type-specific variation in AF

To gain further insight into the molecular changes occurring during AF progression, we carried out functional enrichment analysis of each of the GMM clusters described above. We searched for statistically enriched Gene Ontology [10] terms among the features of each cluster, and discerned how much each of the three different experiments was contributing to this enriched annotation (Figure 4.12 and Table S3). We count how many times each experiment was represented by a feature, into every enriched term and calculated the respective ratios. Thus, we were able to identify changes in gene or protein expression coming mainly from what we reckon are non-CM cells (those terms enriched in atria RNA-seq data but not in CM data), or post-transcriptional regulatory events (enriched in LC-MS/MS but not RNA-seq data).

As such, terms related to extracellular matrix were enriched in several clusters, showing a trend of increased gene expression during AF progression but mainly in whole-atrial appendage tissue and not in CMs (Figure 4.12A). This suggests that these changes occur in non-CM cells, most surely atrial fibroblasts, and relate to the increased fibrosis that has been described during AF progression [4, 146].Genes belonging to other broad categories, such as inflammation or ion channels, change in both atrial tissue and CMs (Figure 4.12B and C), suggesting a complex interplay between different cell types during disease progression. Ion channels genes show an interesting pattern of expression changes. While various potassium (KCNJ3, KCNJ5), calcium (CACNA1C) or sodium (SCN5A) channel genes show a decrease in expression in CMs in AF, most prominent when comparing control and transition samples (Figure 4.12), other components such as HCN2 or KCNH7 are increased. These data are compatible with previous results in sheep atrial tissue showing increased or decreased protein levels of ion channels upon transition to persistent AF [146].

We also found a large group of annotations related to heart muscle and myofibril structure (Figure 4.12D). As expected, changes in these genes were detected almost exclusively in CM samples. Again, here we find genes with different behaviours. Those genes belonging to cluster g_29 and annotated as contractile fibre contain genes such as cardiac muscle alpha actin (ACTC1) or myosin heavy chain 2 (MYH2), and show a general trend of down-regulation. On the other hand, genes included in cluster g_13 (annotated as myofibril) show specific up-regulation in LAA CMs as compared to RAA in chronic sheep. Genes in this cluster include those coding for Titin (TTN), Myomesin (MYOM1) or Myosin heavy chain beta (MYH7). Titin is the largest protein in humans and is essential for normal myocardial function. GWAS studies have found loss of function variants in TTN to be statistically associated with a diagnosis of early-onset AF [39, 167]. MYH7 encodes

the slow molecular motor β-MyHC29 that expresses only in the atria during cardiac development and not in the adult, but its expression is elevated in atrial myocytes of patients with chronic AF as well as in the ovine model of chronic AF [30]. Thus, our analysis identifies many of the previously reported changes in human and other animal models.

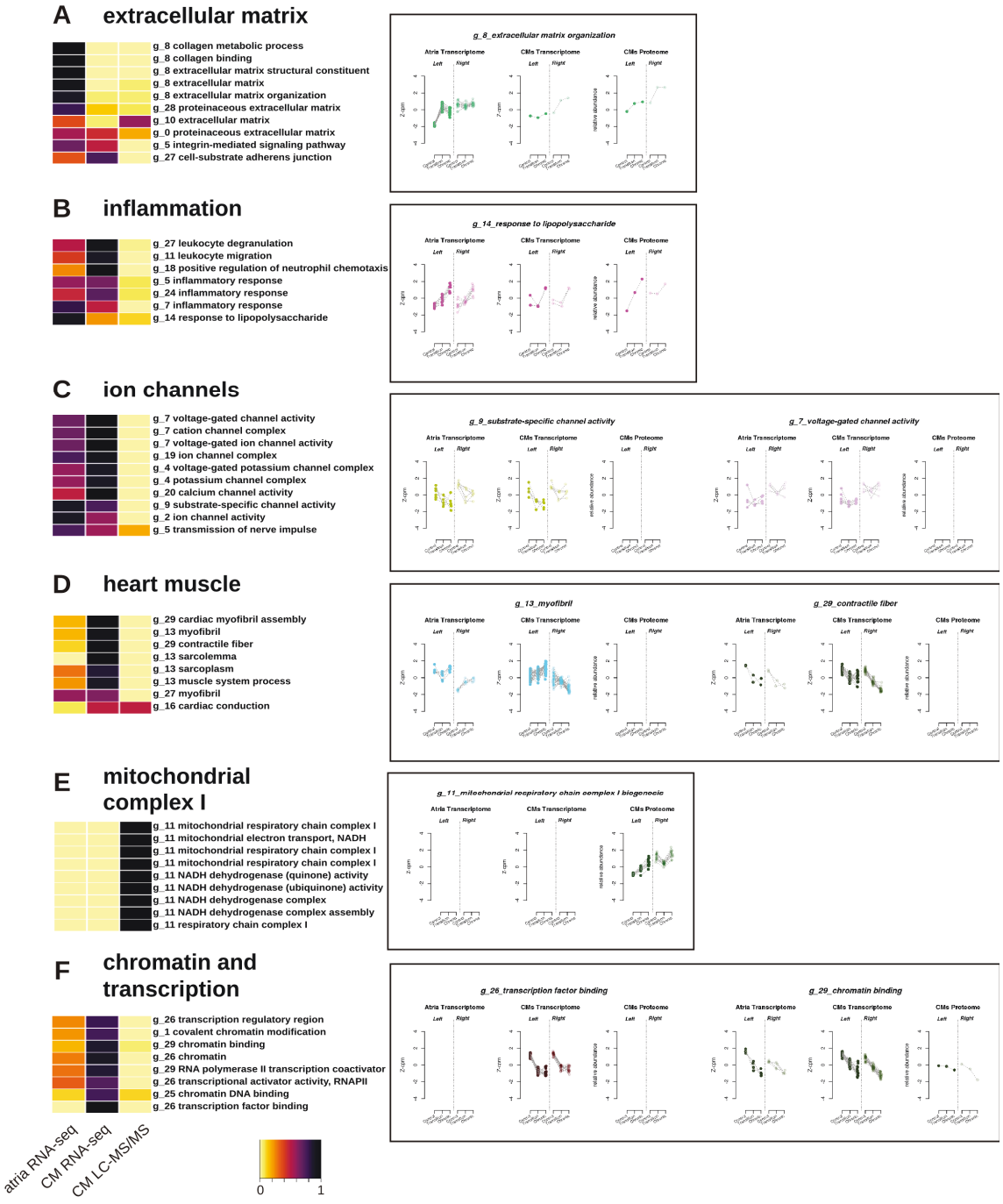


Figure 4.12: **Functional annotation of the molecular changes taking place during AF progression.** (A-F) Broad functional categories of enriched terms are shown, together with heatmaps (left) indicating each individual Gene Ontology Term and the GMM cluster enriched for that term (full listings and details are provided in Table S3). The heatmap represent the contribution of the different datasets to each specific GO term (pale yellow, no contribution; black, all features contribute). On the right, representative examples of enriched GO terms in each dataset. Gene expression levels are expressed as z-scores of counts per million (cpm) and protein levels as relative abundance.

1.6 Chromatin dysregulation in cardiomyocytes is a hallmark of AF

An unexpected observation was the enrichment in terms related to chromatin, that were present in four independent clusters (g_1, g_25, g_26, and g_29; Table S3). The general trend was for a decrease in expression in both LAA and RAA, mainly in CMs (Figure 4.12F). We analysed the expression in CM RNA-seq data of 142 genes encoding chromatin related factors [158]. We identified an overall decrease in expression in transition and chronic sheep as compared to controls (Figure 4.13A). Down-regulated genes included those coding for different histone modifying enzymes (methyl-transferases, de-methylases, acetyl-transferases), related to both active and inactive chromatin and transcription, as well as nucleosome remodellers such as the NuRD complex (Figure 4.13A).

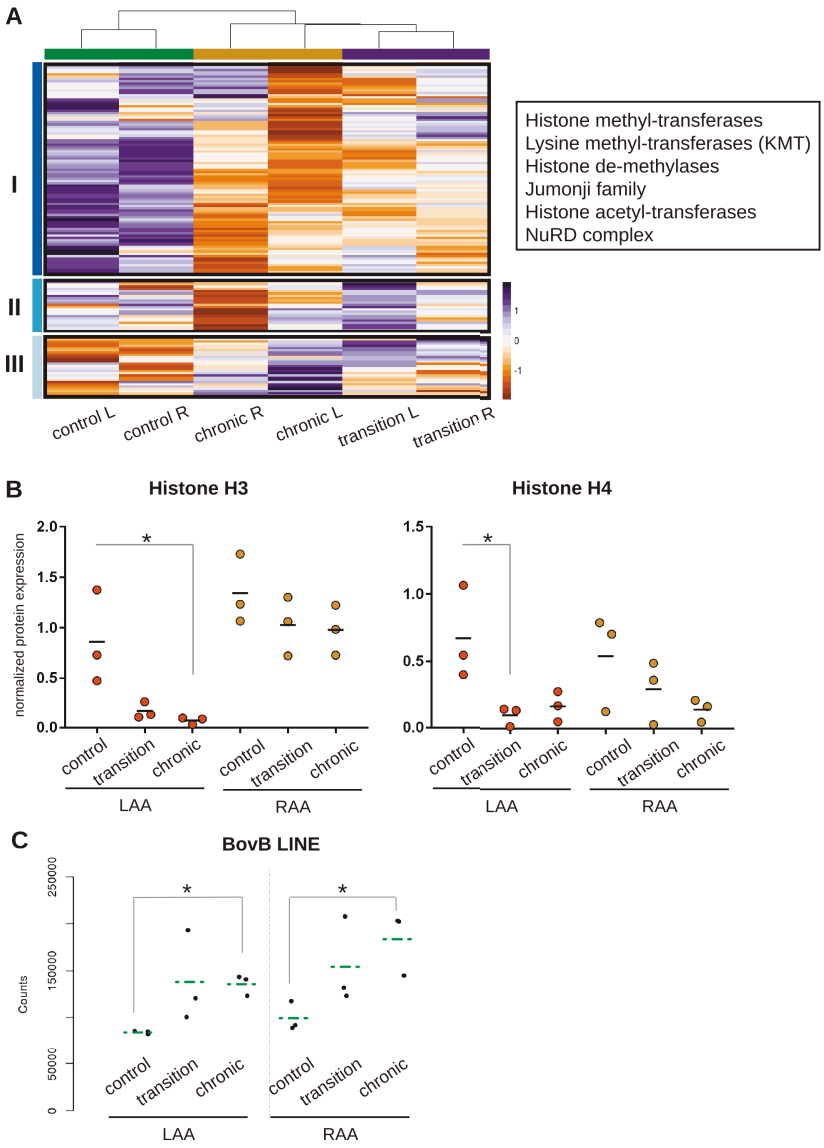


Figure 4.13: **Cardiomyocyte chromatin is disorganized in AF.** (A) Heatmap showing the expression (as z-scores) of 142 genes encoding for chromatin remodellers in cardiomyocytes from LAA and RAA of control, transition and chronic AF sheep. Three main clusters are observed (left), with cluster I showing decreased expression in transition and chronic conditions. This cluster include mayor histone modifiers and nucleosome remodellers (shown on the right). (B) Quantification of the expression of Histone 3 (left) and Histone 4 (right) during AF progression in cardiomyocytes form right and left atria, as measured by Western Blot. Values were normalized to those of TNNT2 as cardiomyocyte marker. n=3; * p-value < 0,05, Student's unpaired t-test. (C) BovB transposable element transcript abundance (counts) in the RNA-seq data from LAA and RAA cardiomyocytes during AF progression. n=3; * p-value < 0,05, DEseq default method.

This observation led us to ask if there was a global dysregulation of chromatin in CMs from sheep with induced AF. We first measured the total amount of histones present in CMs by western blot (Figure 4.13B and Figure 4.14A), and found that there was an important decrease of both Histone 3 and Histone 4 in transition and chronic individuals. Interestingly, the decrease in histones was

much more pronounced in LAA than RAA (Figure 4.13B).

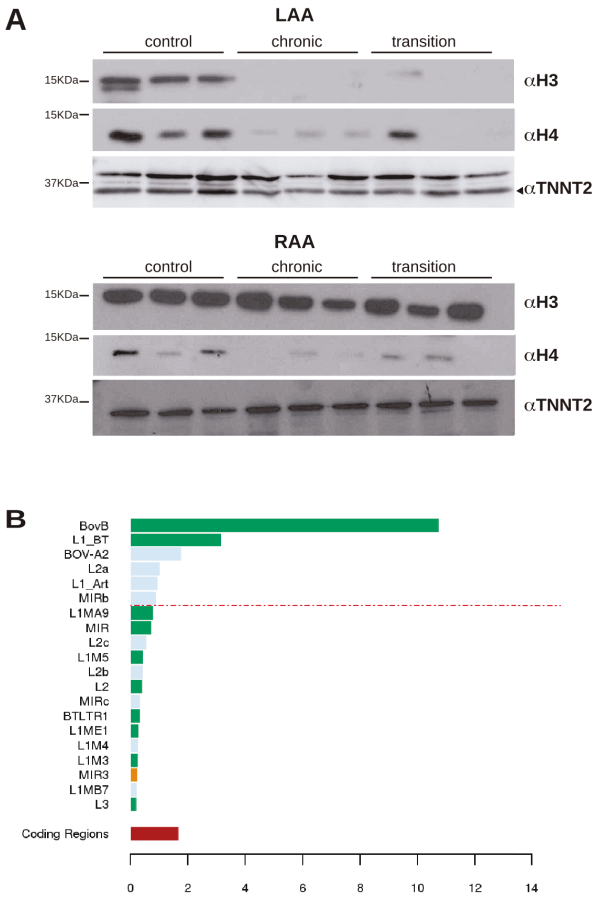


Figure 4.14: **Histone expression during AF progression and TE distribution in the sheep genome.** (A) Western Blot for Histone 3, Histone 4 and TNNT2 in three separate cardiomyocyte samples from left and right atria appendages (LAA and RAA, respectively) from control, chronic and transition AF sheep. Quantification of these blots is shown in Figure 6B. Molecular weights (kilodaltons, KDa) are shown on the left. (B) Annotation of transposable elements (TEs) in the sheep genome. The percentage of the Ovis aries genome represented by each category of TEs is shown on the x-axis; the percentage of the genome corresponding to protein coding regions is shown at the bottom for comparison. Low complexity, simple repeats, rRNA, scRNA, snRNA, srpRNA and tRNA were filtered out, as well as elements located in not-assigned chromosomes. The dashed line represents 50% coverage of total TEs. Different colours depict TE classes: green, LINEs; pale blue, SINEs; orange, LTRs.

The general decrease in chromatin remodellers together with lower amounts of histones suggested that an overall de-compaction of chromatin could be occurring in CMs during AF. Therefore, we decided to examine the expression of TEs from the sheep genome as a proxy for chromatin deregulation. Under normal circumstances, TEs are silenced except in very early stages of development [188]. However, it has been recently proposed that TE expression can occur in some pathological states and also during organismal ageing [252]. We reanalysed our transcriptomic data to assess TE expression, as this information is filtered out during standard pre-processing of RNA-seq data.

In first place, we annotated the sheep genome's complement of TEs (4.14B), finding that the most abundant was the homologue of bovine BovB long interspersed element (LINE) [244]. Expression of BovB LINE was increased in CMs during progression of AF, both in transition and chronic individuals and in left and right atria (Figure 4.13C). Altogether, these results suggest that progression of AF results in a global disorganization of chromatin, with a reduction of histones and remodelling factors, leading to a de-repression of TE expression.

1.7 Changes in the posterior left atrium mirror those in the atrial appendage

So far, our analysis was based on tissue and cells from the atrial appendage, but we wished to know if the molecular changes occurring in this tissue were representative of those happening in other anatomical locations of the atria. We took advantage of samples available from control and transition sheep from the PLA, which together with the pulmonary veins are the prime substrate for AF initiation and maintenance. Furthermore, we obtained six samples to analyse by RNA-seq from each condition.

We first compared the changes in gene expression between control and transition states from the PLA and the LAA. We observed a very high correlation for all genes, both for LAA tissue and isolated CMs (Figure 4.15 A). Furthermore, many of the critical genes we had identified as differentially expressed in the LAA, such as RCAN1, LGALS3, or PCP4, were also changing in the PLA (Figure 4.15B and Table S4). Unsupervised hierarchical clustering resulted in 12 different clusters showing a clear difference in expression between control and transition PLA (Figure 4.15C). Functional enrichment of GO terms (Table S5) showed how terms related to heart contraction and the CM were up-regulated in transition samples, while those related to blood vessel development down-regulated (Figure 4.15C). Furthermore, changes in expression in the PLA of genes coding for chromatin related factors (Figure 4.16A) showed a very similar trend to that we had previously observed in genes differentially expressed in the atrial appendage along disease progression (Figure 4.13A). Overall, this analysis shows that the molecular changes that take place during AF progression are very similar between different anatomical locations of the atria.

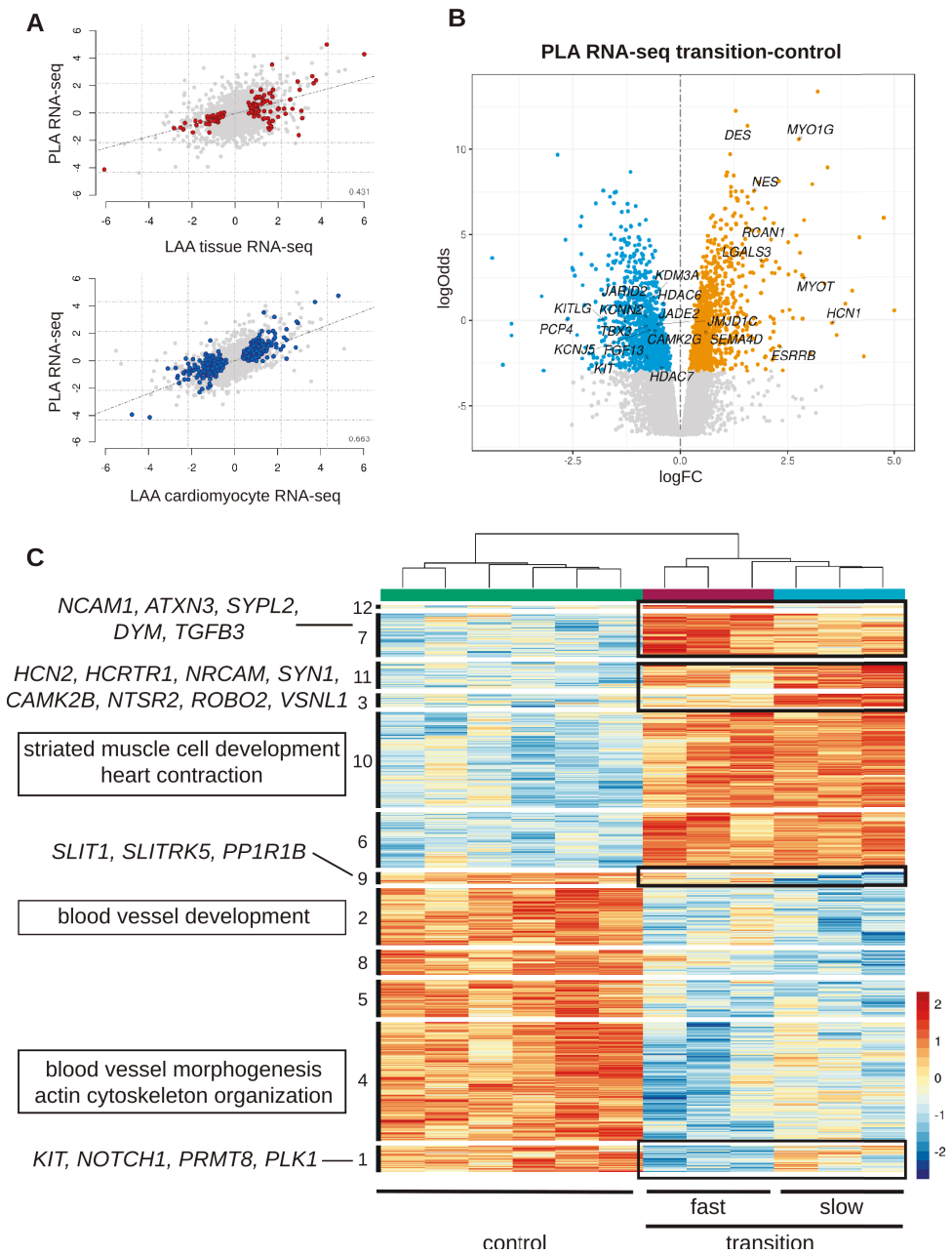


Figure 4.15: **Transcriptomic profiling of posterior left atria tissue.** (A) Correlation of the logFC of expression in transition versus control of PLA tissue with LAA tissue (upper panel) and with LAA cardiomyocytes (lower panel). Differentially expressed genes in LAA are shown in red and blue, respectively. Pearson correlation values are indicated in the right bottom corner of each graph. (B) Volcano plot of transition versus control for PLA tissue. Differentially expressed genes at 5% FDR are shown in orange (upregulated in transition as compared to controls) or blue (downregulated in transition as compared to controls). (C) Heatmap showing the expression (as z-scores) of the 2185 genes found differentially expressed in the PLA when comparing transition versus control sheep (n=6). Two main branches of the clustering segment the differentially expressed genes into downregulated and upregulated for this comparison (transition-control). Various clusters suggest the existence of two different gene expression patterns, for fast and slow sheep to reach persistent AF (indicated as burgundy and blue bars on top of the heatmap, respectively). Genes and GO terms related to individual clusters are indicated on the left

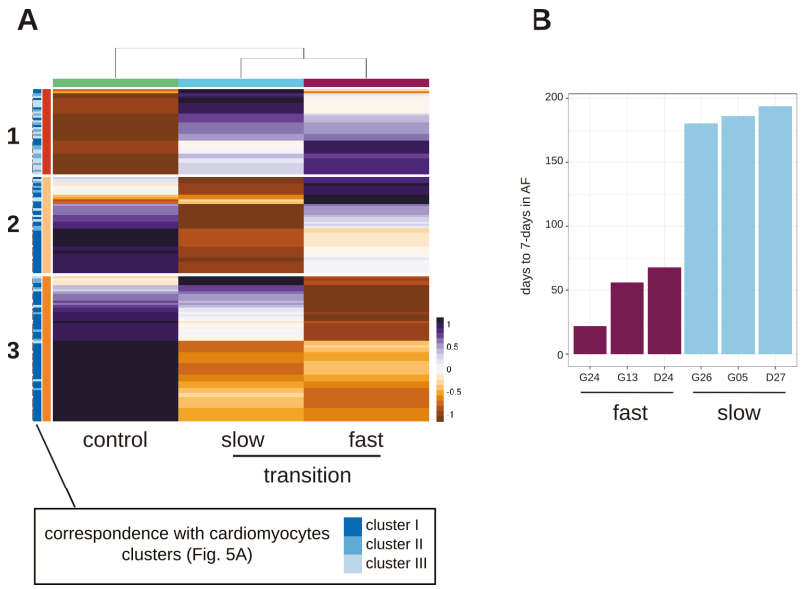


Figure 4.16: **Transcriptomic profiling of posterior left atria tissue.** (A) Heatmap showing the expression (as z-scores) of 142 genes encoding for chromatin remodelers in PLA tissue of control and transition AF sheep. Three main clusters are observed (left), with cluster 1 and 2 showing decreased expression in transition condition. Different expression patterns between slow and fast sheep arise in the three clusters. Clustering for the same gene set observed in cardiomyocytes (Figure 5A) is depicted in different shades of blue, showing a large overlap and concordance. (B) Days for each transition sheep used for PLA RNA-seq to reach persistent AF (defined as 7 days in AF without tachypacing).

1.8 Gene expression identifies differences in the rate of AF progression of individual sheep

An unexpected observation from the clustering analysis was that transition samples separated into two subgroups (top burgundy and blue bars, Figure 6C), and that a number of the clusters generated showed differences between these subgroups (clusters 1, 7, 9, 11, and 12; boxed in Figure 4.15C). This subgrouping was also evident when we analysed the expression of chromatin factors (Figure 4.16A). When searching for a possible explanation of the differences in the transition samples, we observed that these were perfectly matched by the rate of AF progression in these sheep. This was measured as the days taken for an individual sheep from the start of tachypacing to a period of 7 days in AF with no tachypacing (the criterion used to define transition samples). Three sheep took more than 180 days, and were classified as slow progressing, while other three took less than 70 days and were considered fast (Figure 4.16B and Table S1).

Those clusters showing differences in expression between slow and fast transition sheep were not enriched for any particular GO term. However, a careful inspection of the genes included in them (Table S4) showed an unsuspected enrichment in genes related to neural cells, such as those coding

for the adhesion molecules NCAM1 or NRCAM, the ion channel HCN2, components of signalling pathways such as TGFB3 or CAMK2B, or members of the Slit/Robo axon guidance signalling pathway such as SLIT1 and ROBO2 (Figure 4.15C). Finally, it was interesting to find that genes belonging to cluster 1, that show a stronger down-regulation in fast progressing sheep compared to controls than slow sheep, include prominent regulators of cell proliferation such as KIT, NOTCH1, PRMT8, or PLK1 (Figure 4.15C). Future studies will be required to explore if reduced proliferation of cells in the PLA leads to a faster progression of AF to the permanent condition.

1.9 Molecular changes that occur during disease progression are enriched for AF risk-associated genes

GWAS have highlighted the genetic basis for a predisposition to AF, and inform in an unbiased manner of possible molecular mechanisms underlying the disease [4, 41, 61]. On the other hand, our study interrogates the molecular changes that occur as a consequence of the disease, and provides information of the molecular mechanisms responsible for its progression. We asked to what extent did these two processes overlap, and if genes that increase susceptibility to AF were also changing during disease progression.

To do so, we first obtained from the public NHGRI-EBI GWAS Catalogue non-redundant lists of genes associated to electrophysiological cardiovascular diseases (CVD), including AF, and electrophysiological traits such as PR or QT interval, and as a control genes associated to myocardial CVD (such as myocardial infarction or heart failure), obtaining 668 and 212 genes respectively (Table S6). We also used a third GWAS group of the non-redundant genes listed in two recently published independent GWAS meta-analyses which have extended previous AF associations to hundreds of loci [168, 191] (240 genes, Table S6). We then compared these different lists to the set of selected features (4409 differentially expressed genes and proteins) identified in our study of the LAA and RAA, and as a control, the set of all expressed features (11960). We found that genes associated to electrophysiological CVD and traits, as well as those identified as associated to AF by meta-GWAS, are enriched in the set of features that change during AF progression in our sheep model, while genes associated with myocardial CVD are not (Figure 4.17A). Among these are genes coding for ion channels (such as KCNJ5 or SCN5A), developmental transcription factors (PITX2, TBX5), chromatin regulators (such as KDM3A or MBD5), structural proteins (MYOCD, MYOT), and cell-to-cell communication (GJA1 or IL6R) proteins (Figure 4.17B). These results show that gene regulatory networks and molecular pathways that are involved in the genetic predisposition to AF are also altered because of disease progression towards an atrial cardiomyopathy

due to external factors, including stress and inflammation.

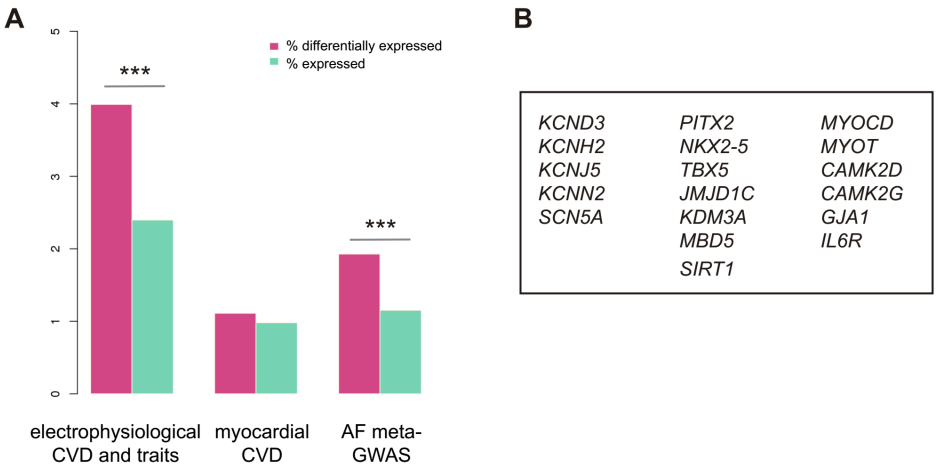


Figure 4.17: **Overlap of intrinsic genetic determinants and extrinsic genetic changes in AF.** (A) Graph showing the percentage of differentially expressed (pink) and all expressed (green) features (genes and proteins) in the sheep AF model that are present in the selected list of genes associated by GWAS to electrophysiological CVD and traits, myocardial CVD, and genes associated to AF in two recent meta-analysis loci 15, 16. Differences between both sets was assessed by a hypergeometric test with Benjamini-Hochberg correction for multiple testing. *** p-value < 1e-04. (B) Representative genes included in the overlap between differentially expressed features in the sheep AF model and AF-associated genes, coding for ion channels (left row), developmental transcription factors and chromatin regulators (middle row), and other cellular components (right row)

1.10 Supplementary interactive file

We provide the user-friendly and interactive web application, AfibOmics Browser, to enhance and facilitate data sharing. Through the app, inspection of the processed data and plenty of tools for visualization and comparison with external data are available. Hence, our molecular map for AF is available as a curated database, open to the user by simply mouse navigation in any web browser.

2 Modelling the progression of Atrial Fibrillation: from the non-tachypacing to permanent-AF

In the first chapter of this thesis, we observed that the transcriptome and proteome undergo significant changes at early stages of AF progression, during the transition from paroxysmal to persistent, remaining stable during follow-up. Those changes correspond closely with the electrical remodelling that occurs in the sheep, whereas structural remodelling in the form of interstitial fibrosis appears more gradually and is belatedly manifest once self-sustained persistent AF (pAF) has stabilized. We reasoned that a more exhaustive analysis of the early stages of disease progression becomes necessary to examine in detail the time prior to the transition between an initial phase of electrical remodelling and a second more gradual phase of structural remodelling. In light of that, we decided to perform a longitudinal analysis of the serum proteome dynamics in our previously described sheep model, covering this temporal window of disease between paroxysmal and persistent AF. This non-invasive procedure, allow us to follow the progression of every sheep, from a sinus rhythm (SR) state before the tachypacing device was implanted and the AF induced, towards the onset of persistent AF, when the dominant frequency reaches its maximum value.

For that purpose, we design our specific proteomic pipeline to analyse LC-MS/MS profiling and accommodate these longitudinal proteomic measurements to multilevel models within a Bayesian framework.

2.1 Experimental Design

2.1.1 Sample collection

In our design, we collected blood samples from 6 sheep longitudinally, tracking every subject with 7 time-points during AF progression, as shown in Figure 4.18A. These time-points were taken in the short window between the sham and the transition sheep, at the onset of AF when disease evolves from paroxysmal to persistent. Peripheral blood was collected from the cephalic vein. Together with that, central blood samples from the right atrium (RA) were collected, at the initial and final time-points of the experiment. At the starting point the sheep remains at baseline sinus rhythm (SR) and device was not already implanted (t_0). Next, there was a recovery period that took around 50 days, as can be appreciated in Figure 4.18C. The pacemaker was programmed to induce AF by burst tachypacing and device was turn on that day before interrogation protocol and collection of the first sample (t_1). The remaining samples were taken progressively until the sheep reaches

persistent AF (t_{pAF}), and becomes a transition sheep.

2.1.2 AF progression as time

AF has been largely proved to be a progressive cardiac arrhythmia [146, 206, 221]. Dominant frequency increase progressively from SR towards pAF, being a reflection of atrial electrical remodelling in the form of action potential duration abbreviation [64, 146], (Figure 4.18B). However, this time window takes shorter or longer regarding the subject, as it has been shown in patients [123] and the sheep, undergoing analogous electrophysiological and morphological changes as result. In our experimental design, the 6 sheep were categorized in fast, medium or slow, as a function of the time required to reach pAF (4-50 weeks). We equally distributed our time-points during this window of time, and thus, accommodate time as a variable of AF progression, as shown in Figure 4.18C. Since samples were collected weekly, we were able to select *a posteriori* the convenient ones to complete our experimental design.

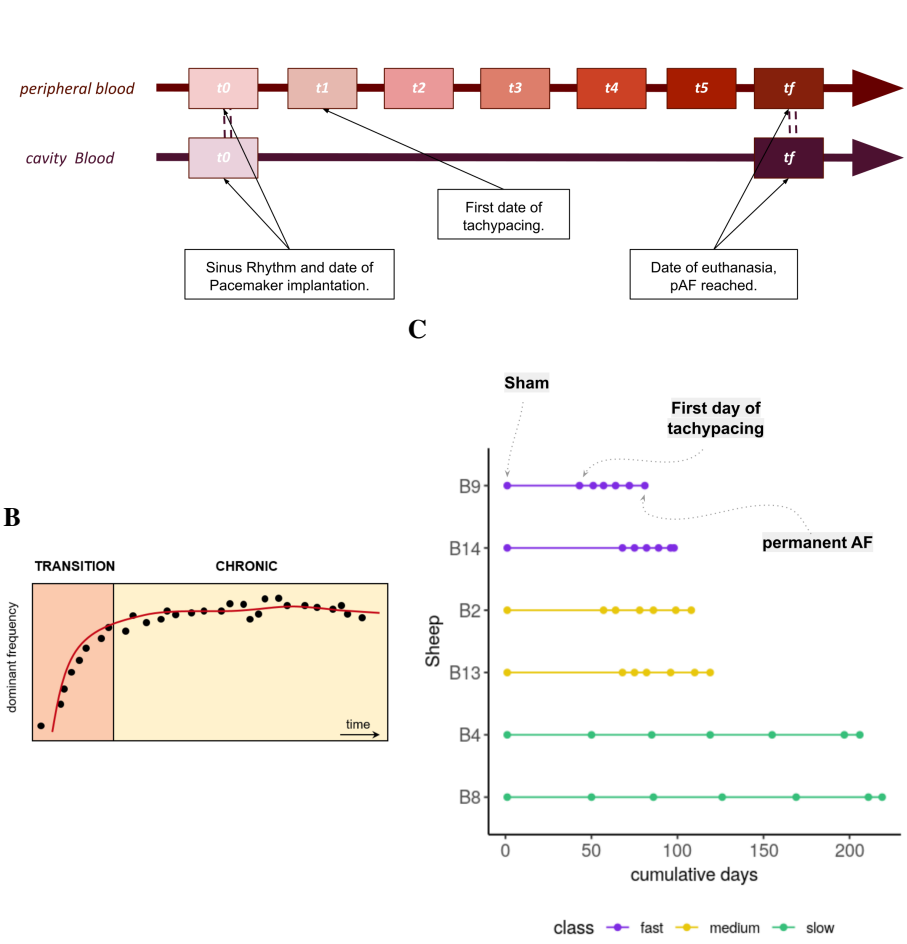


Figure 4.18: **Schematic diagram of the experimental design.** (A) Six time-points were employed for the analysis. Peripheral and right atria (RA) blood samples are illustrated as arrows and t_0 and t_{pf} correspond to SR and pAF samples respectively. Pacing protocol starts at t_1. (B) Representation of the evolution of DF measured in the sheep over progression. (C) Sample collection timings measured as cumulative days since device activation.

2.1.3 Tandem Mass Tag experiments

We performed tandem mass tag (TMT-10plex) based quantitative proteomics, an isobaric labelling technique [256]. Samples were multiplexed in batches of 10 each, which diminish instrument runtime and the variability consequence of the mass spectrometer itself [129]. A total of 6 fractions per sample were processed, to increase resolution and better cross-examine the dynamic range of plasma samples, thereby, 36 TMT-10plex experiments were carried out. Each TMT-10plex batch contains the whole collection of samples belonging to a single sheep, as depicted the Figure 4.19 Additionally, we included an internal standard (IS) in each cassette to capture the variability caused by the batch and correct the data. Since the experimental design was highly complex, samples were not randomly distributed in the different channels, although we recognise that would have been a more worthwhile design to control the variability of labelling reagents.

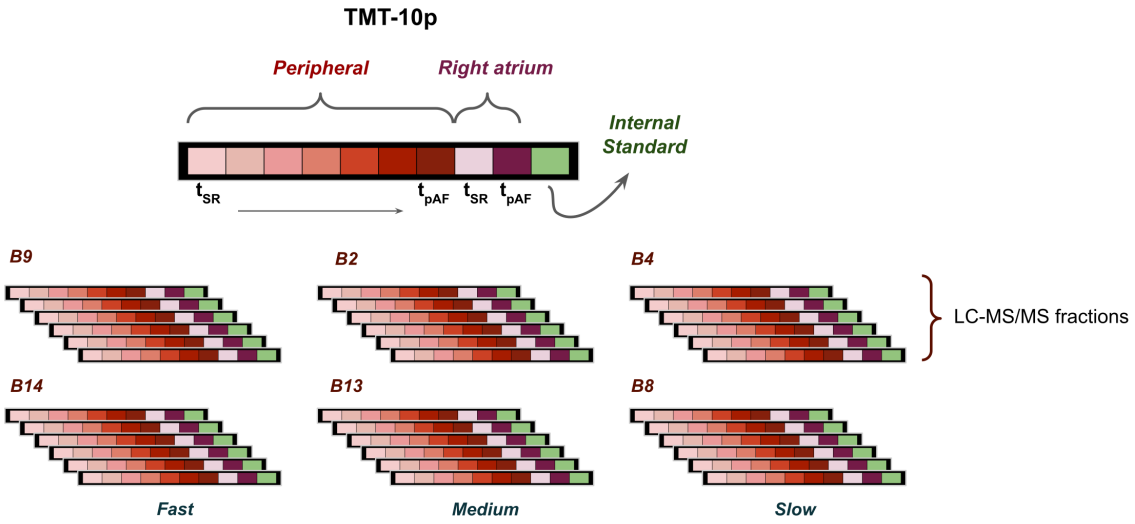


Figure 4.19: **Schematic diagram of the proteomic TMT design.** The 36 TMT-10plex cassettes used in the experiment are depicted as squares. The ten samples are represented as smaller subdivisions in different colors, being red the experimental conditions and green the Internal Standard. (t_{SR}) refers to SR samples and (t_{pAF}) to pAF samples. The order in the illustration corresponds to the TMT 10-channel labeling reagents (TMT-126, 127N, 127C, 128N 128C, 129N, 129C, 130N, 130C, 131)

2.2 A pipeline for proteomic analysis

To analyse this LC-MS/MS shotgun proteomics data, we develop our own pipeline, based on several open-access tools. Most of the subsequent steps were performed under OpenMS [192], an open-source software library, importantly, useful for data management in addition to the analysis itself. OpenMS integrates the TOPP (The OpenMS Proteomics Pipeline), and the UTILIS tools, which are small applications that can be chained to create pipelines tailored for a specific problem. Given that this software is in continuous development, some of the required utilities were not fully available or updated at the date this analysis was performed, and in some cases, input and output files collide between tools. For those cases, we made use of the original programs and switched between suitable formats whenever possible. All the corresponding scripts and settings are available here.

Our pipeline comprises four primary parts, 1) a pre-processing initial step, 2) an identification branch, led by MSGF+ [109] and percolator [227], in charge of the main search and the re-scoring of psms respectively, 3) a quantitative branch that extracts the isobaric labeling information and maps them to the corresponding identified psms and 4) a room for data integration. A general overview of the pipeline is depicted in Figure 4.20. The main pipeline uses data collected for ev-

ery sheep as input independently, because we are interested in obtaining separate abundances per subject. Nevertheless, our 6 sheep act as replicates in our design, and we aim to end up with a common set of protein groups that allows us to integrate the entire dataset. For that purpose, we also modified the original pipeline to perform the protein inference step with all the samples at once. Note that the main pipeline is labelled with black arrows in the diagram, whereas the pipeline for inference does it with the grey ones. Following is a step-by-step description of the pipeline we have created and used for the analysis. Because of its relevance, we thought it would be of interest to include it here.

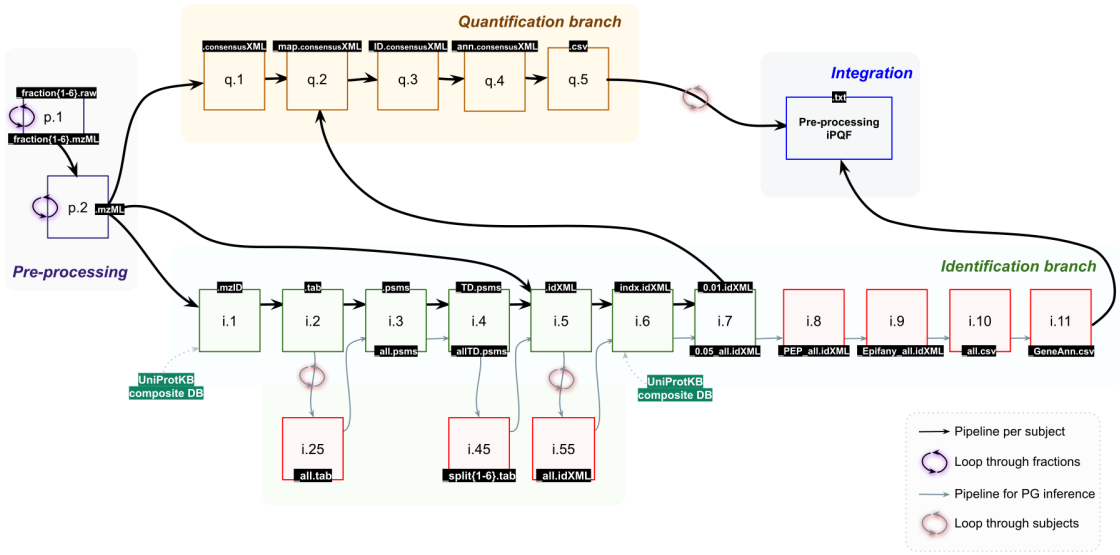


Figure 4.20: **General overview of the proteomic pipeline.** The four main parts of the pipeline are depicted in different colors and sections and the additional steps for the PG inference in red. Boxes represent sub-processes and file extension of the subsequent output are indicated in the upper and lower part for the main and PG inference pipeline respectively. Arrows represent the connection of output/input files between sub-processes. Custom database for identification is depicted in green.

2.2.1 Pre-processing

We started the analysis from the raw Thermo files, generated directly by a mass spectrometer, in our case, from the Thermo Scientific Orbitrap Fusion Tribrid, which was configured as MS2.

p.1: **File conversion of mass spectrometer data**

Firstly, our 36 raw files must be converted to a mzML format, a variant of the XML format for encoding raw spectrometer output and a systematic description of the conditions under this data was acquired and transformed. Few tools are freely available in Unix OS for this conversion, one is the Proteowizard's msconvert utility. However, its ability is limited by the need for vendor-specific licenses. Thus, people under Unix had to use official Docker images to perform this step.

p.1: **Dealing with fractions**

We concatenated our 6 fractions in a common mzML file per sheep, by using FileMerger. Since we are exclusively interested in the joint results per subject, we create a temporal gap between files, inserting 30 seconds between the retention time (RT) ranges of the distinct fractions to avoid overlapping values. We are losing meta information and violating the specification of the mzML file, however, it is an intermediate step that improves the searching and the rescoring of psms. Presumably, not an elegant solution, however, most software does not work properly with fractions, concatenated files or even with multiple mzML files at once, and any other strategy will increase substantially and unreasonably the complexity of the analysis. This situation was already discussed by the OpenMS developer team here. In any case, the original files are preserved and a key is produced in case we need to trace back the original RT values.

2.2.2 Identification branch

i.1: **Main search with MS-GF+**

We selected MS-GF+ as our MS/MS database search engine [109], one of the most sensitive search tools, increasing significantly the number of psms compared with other commonly used methods, and importantly, universal, regarding the type of proteomic protocol, free and open-source. The search was done against a target-decoy database, for computation of q-Values. Precursor mass tolerance was set as 20 ppm. Full trypsin specificity was employed with a maximum of 2-missed cleavages and 2 tolerable termini, at length between 6 and 40 amino acids. Cysteine carbamidomethylation was used as a fixed modification. Acetylation of protein N-termini (TMT) and oxidation of methionine were selected as variable modifications. Note that we added the addFeatures option, to report the specifics scores demanded by the posterior percolator step.

Database generation: Search was done against a custom-built database, composed of the genome reference sequences of the sheep and two related species, an approach that provides better results than using the sheep proteome alone, diminishing the impact of imprecise genome annotations [37]. UniProtKB reference proteomes of *Ovis aries* (238210 sequences), together with the cow *Bos taurus* (434384) and the goat *Capra hircus* (377919), besides a reduced list of typical contaminants, were downloaded. The typical list of contaminants usually contains bovine serum proteins, appropriated for cell lysate experiments using fetal calf serum. Some of these proteins that have high homology to plasma proteins might hide the identification of true proteins in our samples[REF]. The reduced list was obtained

from the common Repository of Adventitious Proteins (cRAP). It contains 65 contaminants, mainly keratins and trypsin. gt-sequniq from GenomeTools was applied to filter out redundant proteins in our composite database, containing at the end a total of 90795 sequences.

i.2: **Dealing with formats: PIN**

The mzid MS-GF+ output, containing identified psms, and including target and decoy positive identifications, were converted to percolator input (PIN) with the msgf2pin utility.

i.3: **Percolator as psms post-processor**

We chose to use percolator [227], a post-processor of search results, compatible with most of the popular engines, free and open-source, that improves search results[REF]. The semi-supervised learning algorithm of percolator, trains a SVM classifier on a random subset of the psms and use the resulting score vectors to re-assign reliable statistical confidence q-values and calculates posterior error probabilities, to peptides and psms. We run percolator without the inference step, since our purpose is simply to recalibrate all valid psms.

i.4: **Editing the POUT**

Decoy, besides target identifications, must be incorporated in the percolator output file (POUT). Then, POUT was manually edited to have those psms added without corrupt the file format.

i.5: **Dealing with formats: idXML**

POUT was converted to a suitable identification file format (idXML) with IDFileConverter. The concatenated mzML file was required to recover the metainformation, lost as a consequence of Percolator specific formats.

i.6: **Restoring the internal database**

PeptideIndexer was used to refresh the protein entries of all peptide hits, utilising the concatenated version of our target decoy database.

i.7: **Getting the valid psms**

To conclude the identification branch, IDFilter filters out the identified psms with a FDR vale below 0.01.

2.2.3 Quantitative branch

q.1: **Extracting the isobaric labelling information**

IsobaricAnalyzer was used to obtain isobaric labelling information from the concatenated

mzML file. We employed the correction matrix supplied by the manufacturer of our labelling kits to perform isotope correction by non-negative least squares. Important to notice, we did not perform any normalization at this step. The output contains the consensus features, each one representing one relevant MS2 scan.

q.2: **Mapping valid psms**

Next, the 0.01-FDR peptide identifications, obtained in the i.7 step, were assigned via IDMapper to the consensus features obtained previously in the isobaric quantification. This mapping step is based on retention times and mass-to-charge values per MS2 scan, to map consensus features to valid psms.

q.3: **Filter**

Psm records matching to more than one feature were amended with IDConflictResolver. Only the hit with the best score remains annotated .

q.4: **Clean**

Some peptides still stay unassigned, so those were filtered out with FileFilter.

q.5: **Export**

Finally, annotated and quantified peptide results were obtained with the help of TextExporter and "in house" code to create readable and comprehensible plain text file. These csv files, one per sheep, include quantification per channel of every psms that pass the filters, the corresponding scores and metainformation.

2.2.4 Performance of the main pipeline

We compare the performance of our pipeline against MaxQuant [46], QuiXoT [161] and the results obtained by the use of MS-GF+ [109] without post-processor step in the Figure 4.21. MS-GF+ identified more psms than any other search engine, and the use of percolator substantially improved the amount of psms and peptides identified with a q-value below 0.01. Notice that we also tested the behaviour of this methods with our composite database and using simply the *Ovis aries* (Oar) genome sequences. The composite database outperformed the Oar version regardless of the method.

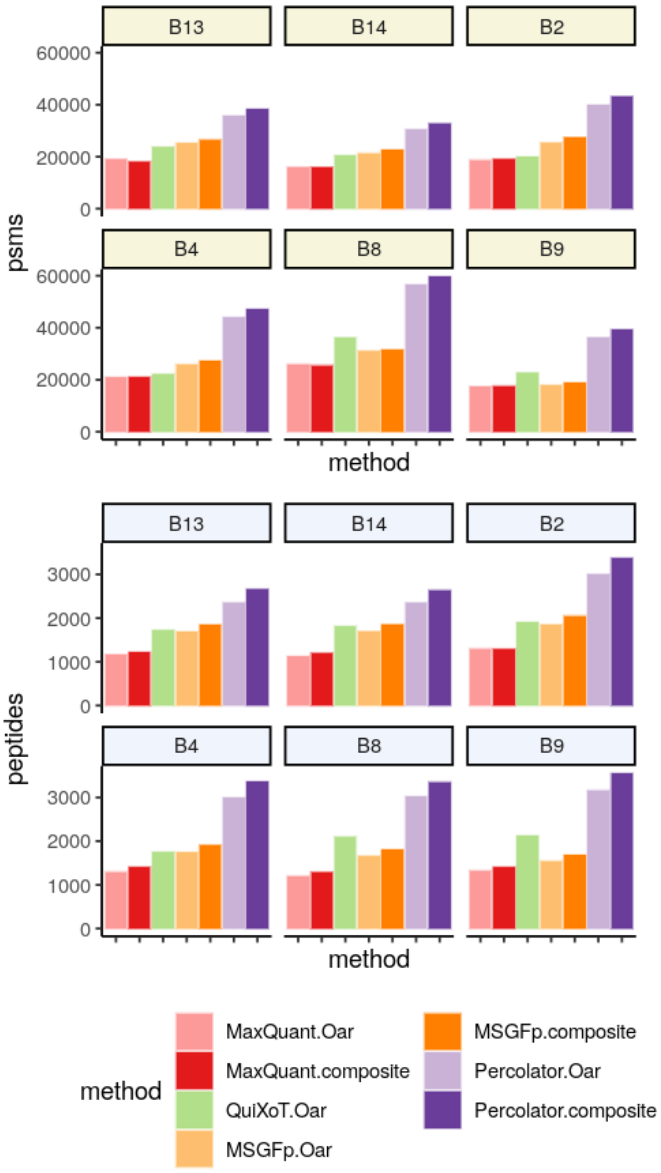


Figure 4.21: **Performance of the main proteomic pipeline.** Results of the 7 different strategies for peptide identification carried out by sheep were compared and represented as barplots in different colors. Upper an lower panels represent the number of valid identifications of psms and peptides, respectively.

2.2.5 Pipeline for PG inference

Since we are working with 6 sheep replicates, our final aim is put together all this information. We expect having common protein changes through progression in all the sheep of the study. Given the nature of protein inference, it is quite probable that the protein groups pinpointed per sheep, if we do the inference one by one, will not match perfectly in a great majority of the cases, and we would lose that data. We reasoned that a more accurate and a better performance in the inference step would be achieved, if all data was analysed together. Therefore, protein inference was done on

a concatenated version of our whole dataset. To do so, we modified the main pipeline described above, adding some intermediate and final steps, as follows:

i.25: **Concatenating the PINs**

The previously obtained PIN files, from the main pipeline, were manually concatenated together, keeping the format specifications. Notice that Percolator was run again with this input, to obtain reliable scores for the whole set of data.

i.45: **Splitting the POUTs**

Prior to convert the POUT to an idXML, this must be split by sheep, because the per-individual mzML is required at this step, see i.5. The alternative of creating a common mzML file would result in a much more complex solution.

i.55: **Merging in a common idXML**

We use IDMerger to build a common idXML file, which contains the psms from whole experiment, worthy for the next steps of the pipeline.

i.8: **Setting proper scores**

The idXML files have a single score associated at every time, however multiple scores are stored as metadata. Once filtered by FDR, and previous to inference, we switched this flag to the posterior error probability calculated by Percolator wiht IDScoreSwitcher.

i.9: **Protein inference**

Protein Inference was performed with Epifany [178]. This algorithm made use of a Bayesian probabilistic model to group and score proteins. Epifany has been published recently, showing very impressive results in comparison with other methods for inference, including the popular FIDO [205].

i.10: **Exporting the inferred protein groups**

Finally, inference results were obtained with the help of TextExporter and "in house" scripts to create readable and comprehensible plain text file. This csv consists of a table that matches peptide sequences and protein groups, the corresponding scores and metainformation.

i.11: **Annotation to Genes: PG**

The inference algorithm constructs a graph based on the peptides sequences and the proteins that share peptides to finally assemble the protein groups. Gene information is missing there, fact that becomes tricky under our particular circumstances. We chose to use 3 different species, and predominantly, the final protein groups include the homologous proteins

codified from the same genes in the 3 animal models. However, sometimes those relations are not captured in the graph because a peptide that links the proteins is missing and thereby, the same protein might be found in different protein groups. As a final step, we decided to connect the Epifany protein groups by genes. We named them PG for the remainder of the text.

2.2.6 Integration to PGs

Once we obtained the valid psms per sheep in addition to a common relational table to match those peptide sequences together into PGs, the entire information of the experiment was integrated in a common dataset. For this fourth and last part of the proteomic pipeline, we made use of the R library MSnbase [72], a popular library to handle proteomic data in R.

Proteomic data per sheep, containing the peptide quantification, was treated as a MSnSet2 objects, one per batch. Each was cleaned of possible target-contaminant entries and metadata was edited according to specifications. Given the nature of isobaric labelling, missing values per TMT might result in an inconvenience, causing abundances to be distorted. Abundances are relative and summarization methods do not deal properly with missing data. Therefore, we applied the strictest criterion and decided to remove all the entries with at least an NA. Data was normalized using variance stabilization (VSN) [90], which eliminates the dependency between mean intensity and variance and brings samples onto the same scale. This normalization method performs consistently well with this sort of values and has been proved to reduce variation between technical replicates in real data [238]. Figure 4.22A and Figure 4.22B show the effect of normalization for B9 sheep.

We chose the iPQF method for peptide-to-PG summarization [65], precisely developed to deal with isobaric labelling protocols. Firstly, psms are weighted according to spectra features like charge, search score, length redundancy, etc along with the quantitative values for protein ratio estimation. Lastly, psms are combined into PGs by median polish, when at least 3 psms support the PG. The IS channel was specified to calculate PG abundance ratios. The distribution of PG abundances and corresponding PCA can be checked in Figure 4.22C and Figure 4.22D. Data was standardized as Z-scores per PG for every sheep individually, Figure 4.22E and Figure 4.22F. Through these successive steps we can observe how data relations change, and some of the variability regarding time is now retained within the first principal component in Figure 4.22F.

After data integration through our relational table, we ended with a total of 1194 PG identified, of

which 292 are below the 0.01 threshold of FDR. Those PGs correspond to 4039 and 693 proteins, respectively, and to a total of 3565 unique peptides. For the next modelling step, we only kept those PG which appear in at least 3 of the sheep, finally capturing 292 PGs.

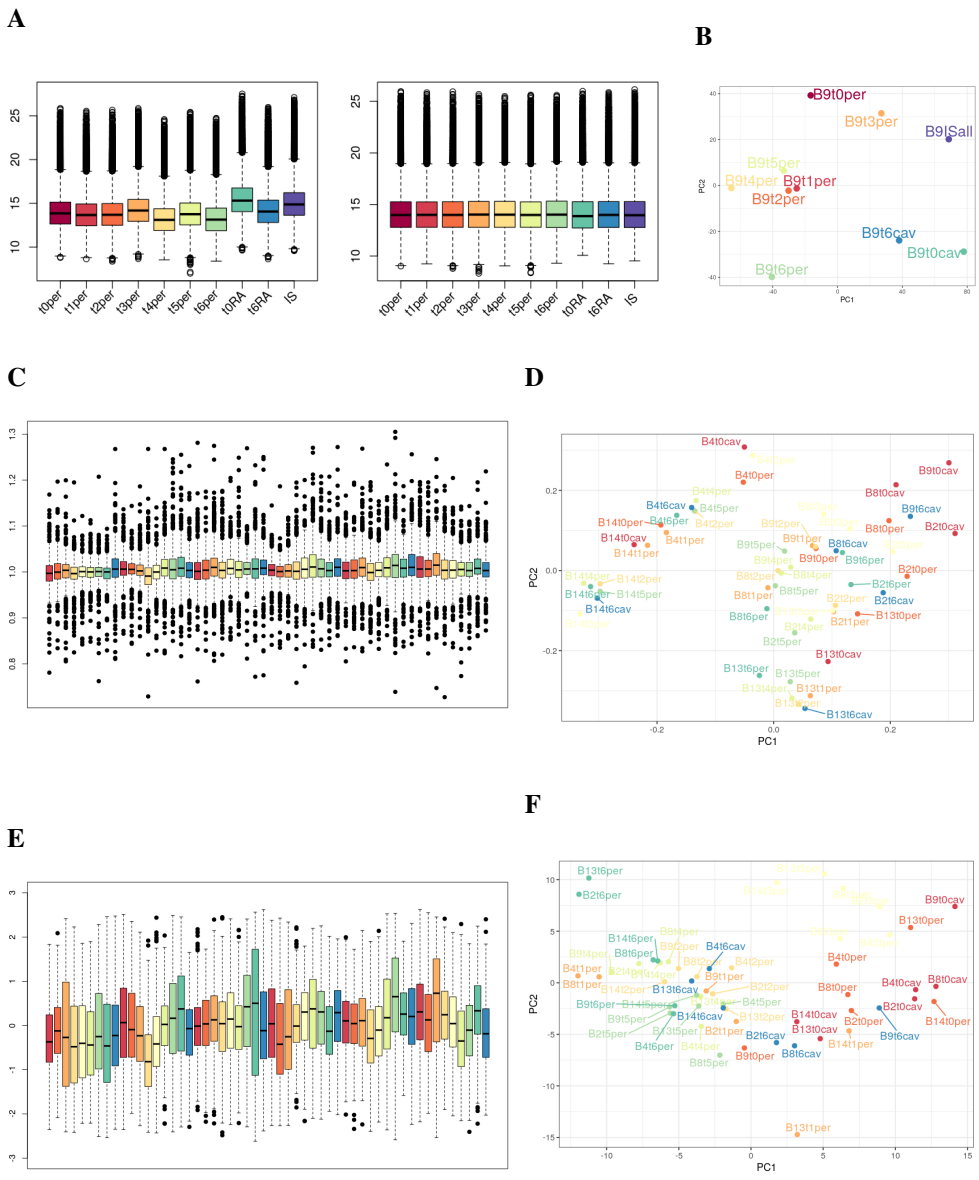


Figure 4.22: **Filtering, normalization and integration post-processing.** Conditions are represent in different colors. (A) Raw psm abundances of the sheep B9 are depicted in the left panel and filtered and vsn-normalized psms in the right panel. (B) PCA of the corresponding normalized data. (C) Boxplot representing iPQF-integrated data and (D) the corresponding PCA. IS were excluded in the plots after integration. (E) Boxplot representing Z-scored abundances and (F) the corresponding PCA.

2.3 Probabilistic modelling

We aim to model and examine the behaviour of the proteome through disease progression, from baseline SR towards persistent AF. For that purpose, we chose to fit Multilevel Models (MLM) within a framework of Bayesian inference. The basic idea behind this approach, is that our data is clustered by sheep replicates, time-points and proteins (PG), in different observational units, and these clusters share certain attributes and similarities that should be incorporated in the inference process. Depending on the question, we implemented several models, being time our level-one in the hierarchy and sheep the level-two. In some cases, these L2 models were expanded with a third level (L3 models), in which all the PG are jointly modeled. Going further, we took into account the different locations (peripheral and right atrial blood), to distinguish what happens systemically and in the atrial environment. We started from the already 292 identified protein, and our main strategies were model comparison and probability calculation to find the differential PG abundances among time. All the scripts of this section are available here.

2.3.1 Proteins that change over progression of AF

Firstly, we aimed to find the PGs that change over progression in peripheral blood. Our strategy begins by obtaining the model that better fits our data using three-level hierarchy, which includes all the information of the proteome at once. Afterwards, we compare the former against a null model in which no changes in PG over time are presumed, but in this case, by fitting L2 models, protein by protein.

2.3.1.1 Pooling information from the whole proteome

Here, we provide a brief overview of the L3 models employed, but full specification is available in the Method section 2.4.3.2. We aim to estimate $PGab_{ijk}$, which represents the PG abundance of the protein k, measured in the sheep i at the time j. For that, all the dataset information, regarding the 292 identified PGs, is pooled together in each of these L3 models. Our strategy systematically expands the previous model by adding more complexity, building up incrementally, up to fit our full theoretical multilevel model of change.

We started from the constant or unconditional means model, designed as M.0, which assumes no changes in the trajectory of proteins and acts as our null model. A grand mean for PG abundance over progression is estimated globally, for the whole population, together with a mean for each PG, computed as group-level terms. 3 sources of variation are encapsulated in M.0 terms: 1) the

variation across the proteome, still there despite normalization (sd_PG_Intercept); 2) the variability among sheep per every PG (sd_PG:sheep_Intercept), and 3) the remaining variation or residual variance (sigma), comprising the longitudinal variation, the variation attributed to other possible covariates and surely, the noise of the proteomic data. Their posterior density estimates can be seen in Figure 4.23. Observe that the variability due to sheep and proteome is close to zero, and the main source of variation is the within sheep and protein.

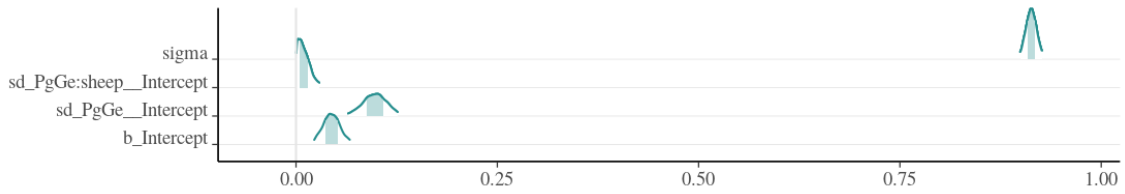


Figure 4.23: **Main sources of variation in L3 models.** CI are represented as shaded areas under the estimated posterior density curves. The global mean of the proteome (b_Intercept) and the variances in L3 models (sd_PgGe:sheep_Intercept, sd_PgGe_Intercept and sigma) were depicted. Inner shadow areas represent 50% CI and density curves 95% CI.

Now, we expanded M.0 progressively to accommodate the longitudinal variation inside parameters. The M.1 is a unconditional growth model, assuming a linear change trajectory over progression, in which an initial PG abundance at SR is estimated, globally and for each PG, as always will be for these L3 models. Additionally, a unique rate of change is estimated per PG. Going further in complexity, we expect changes of PG abundance to be nonlinear for some PGs. Thus, with the M.2 (quadratic), M.3 (cubic) and M.4 (fourth degree) models, we subsequently add increasing orders of polynomial terms at level-one, building systematically more and more intricate patterns of abundance through time. Lastly, we fit a discrete-time model, treating time as a categorical variable. M.5 provides, additionally to the SR abundance, a different rate of change per time-period category. We observed that M.5 time-point predictors per PG were not centered, therefore we fit M.6, which expands the previous one by adding the global time-point predictors as population-level effects.

Model comparison via calculation of the PSIS-LOO Information Criteria in Figure 4.24, shows that M.6 is marginally superior to the M.5, and both outperform the rest by far. PSIS-LOO comparison indicates that fit improves with the increasing complexity of the temporal level-one specification, as it grows the number of parameters in the model. Since any alternative model to the discrete-time one is barely close, and despite having a more complex parametrization, we chose M.6 as our full theoretical multilevel model of change. Selecting the M.6 suppose many advantages for this kind of analysis. Importantly, it yields comprehensible coefficients, contrary to the polynomial ones,

which are not easy to interpret. Straight questions can be done, such as accounting for the number of significant changes between time-points, the magnitude of those changes, etc. Moreover, we notice that polynomial terms force temporal patterns that do not correspond to the truth of the data.

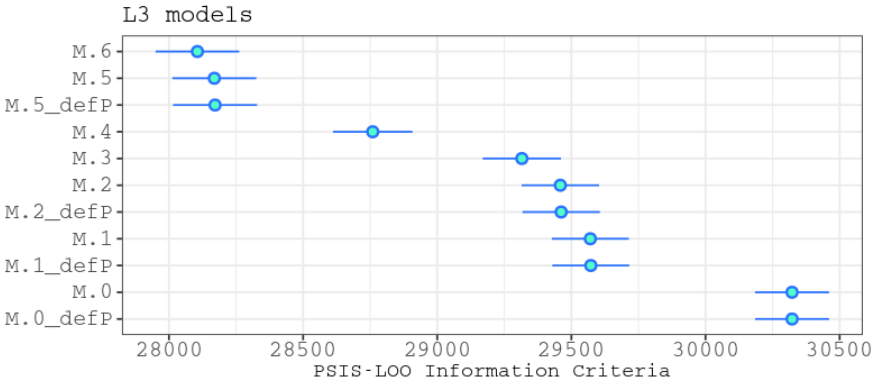


Figure 4.24: **Comparison of L3 models.** L3 models were compared using the PSIS-LOO information criterion, dots represent the PSIS-LOO estimates and horizontal lines the associated standard errors.

By inspection of the M.6, we soon realised on the artifactual behaviour at t3. Although the global mean for the t3 rate of change (b_t_fact3 in Figure 4.25A) and its variance (t_fact3 diagonal of Figure 4.25B) seems pretty normal, the odd behaviour of t3 becomes clear when observing its covariance (t_fact3 in Figure 4.25B) or correlation (t_fact 3 in Figure 4.25C). At this point of the analysis, we had already observed a distinctive behavior of this 128N channel, however, we were not able to discriminate before that it has no biological foundation. Carefully inspection of MS2 chromatograms evidenced its aberrant behaviour and at last, we decided to exclude the channel for the analysis.

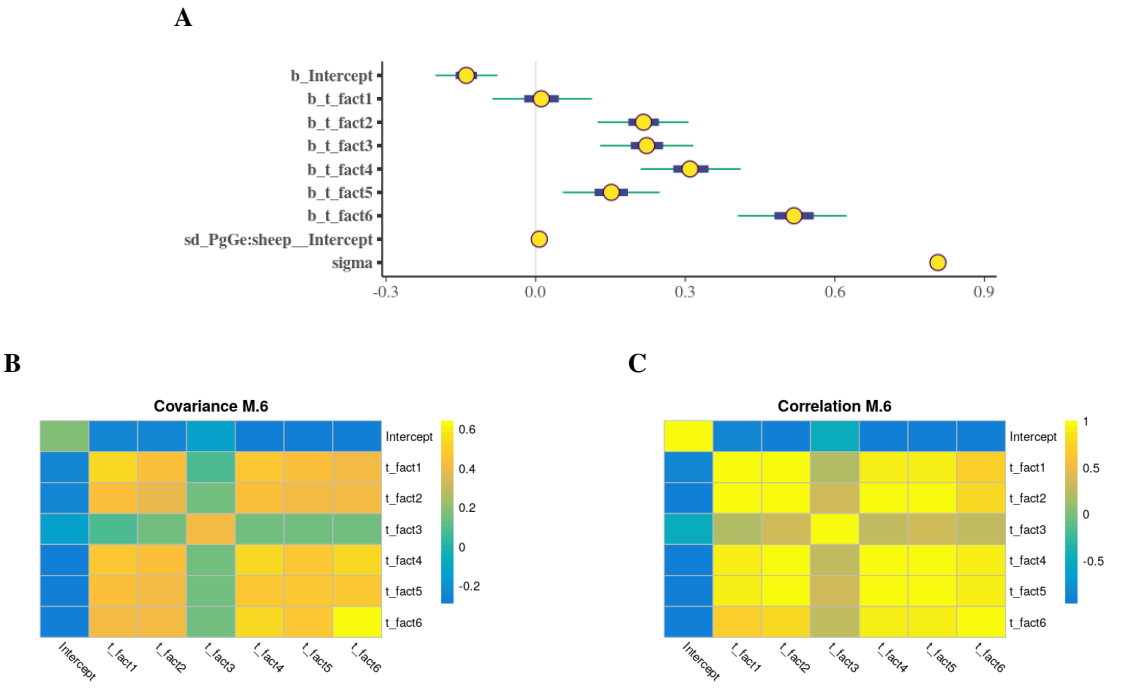


Figure 4.25: **Overview of the M.6 model.** (A) Coefficient plot of the population-level posterior probabilities, residual variance (sigma) and the variability among sheep per every PG (sd_PgGe:sheep_Intercept). The estimated mean is represented as yellow dots and 50% and 95% CIs as thicker and inner segments, respectively. (B) Heatmap representing the covariance matrix of M.6 of the group-level time-point terms, where variances are in the diagonal together with the corresponding covariances. (C). Heatmap representing the correlation matrix of the group-level time-point terms of M.6.

L3 model comparison without t3, showed again that M.6 gets the best performance. Nevertheless, the differences in deviance are less abrupt, probably due to t3 inclusion misshaped the natural PG trendline, becoming also atypical for polynomial functions. Global coefficients in Figure 4.27A remained the same, as expected. The covariance matrix in Figure 4.27B, also shown as correlations in Figure 4.27C, exhibit the expected temporal pattern, with highly correlated time-points. The lower variance is found at SR, whereas the largest is at the time of pAF, indicating distinct and increasing magnitudes of change in the peripheral proteome over progression.

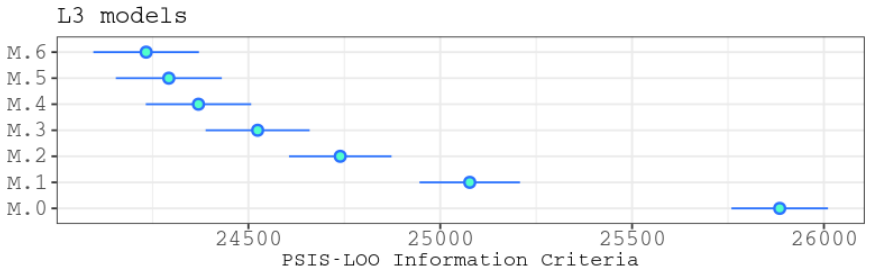


Figure 4.26: **Comparison of L3 models without t3.** L3 models were compared using the PSIS-LOO information criterion, dots represent the PSIS-LOO estimates and horizontal lines the associated standard errors.

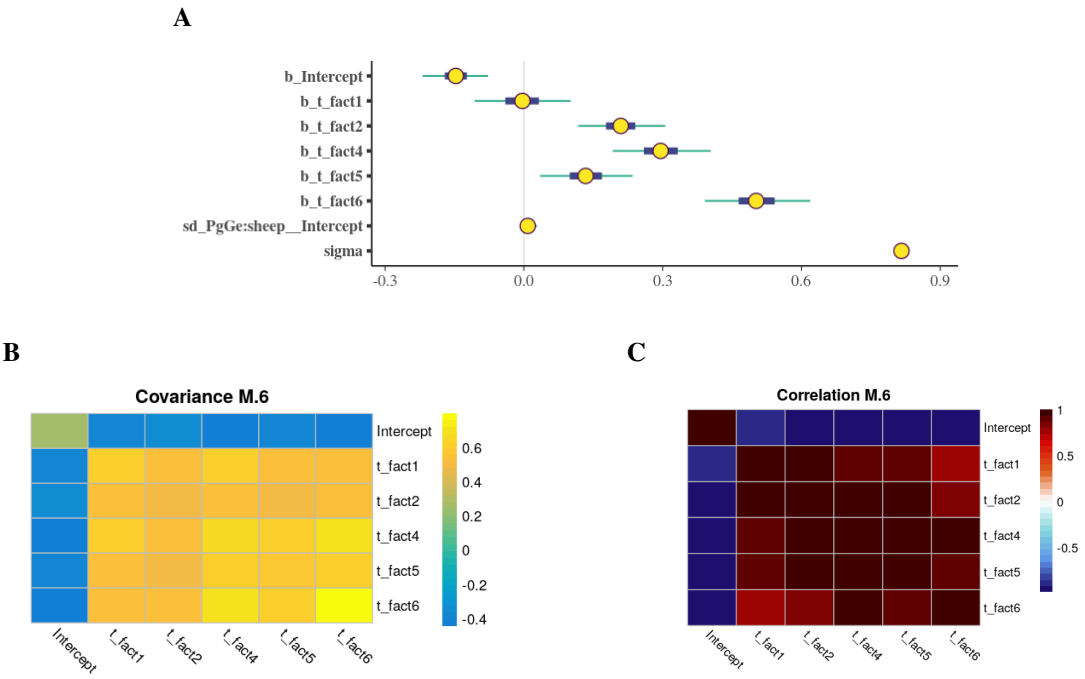


Figure 4.27: **Overview of the M.6 model without t3.** (A) Coefficient plot of the population-level posterior probabilities, residual variance (sigma) and the variability among sheep per every PG (sd_PgGe:sheep_Intercept). The estimated mean is represented as yellow dots and 50% and 95% CIs as thicker and inner segments, respectively. (B) Heatmap representing the covariance matrix of M.6 of the group-level time-point terms, variances are in the diagonal together with the corresponding covariances. (C). Heatmap representing the correlation matrix of the group-level time-point terms of M.6.

2.3.1.2 Comparison of L2 models

To obtain the proteins that change through progression, we implemented an L2 version of the L3 models described previously, parametrized equally in terms of the time pattern that represent, and

omitting the third proteome-level. Here, we end to estimate $PGab_{ij}$, which represents the PG abundance of a given protein, measured in the sheep i at the time j. L2 models are named in lower case with matching enumeration according to level-one specification. Thus, to identified the PGs that change, we examine one PG at a time.

The m.0 is the null model, and assumes no changes for a particular PG over progression. The mean of the PG abundance is calculated, together with 2 sources of variation: 1) the variation among sheep (sd_sheep_Intercept) and 2) the remaining variation or residual variance (sigma), comprising the longitudinal variation, the variation attributed to other possible covariates and surely, the noise of the proteomic data. As an example, the posterior density estimates of this coefficients for the PG_23 are depicted in Figure 4.28. Variability among sheep is close to zero, and the main source of variation is within sheep. The m.0 is expanded to accommodate the longitudinal variation inside parameters, exactly as we did before for the level-one parametrization. The method section includes the full specification of the models: m.1) unconditional linear growth, m.2) unconditional quadratic growth, m.3) unconditional cubic growth, m.4) fourth degree polynomial growth, m.5) discrete-time and m.6) the discrete-time version which makes use of the M.6 covariance structure among time-points.

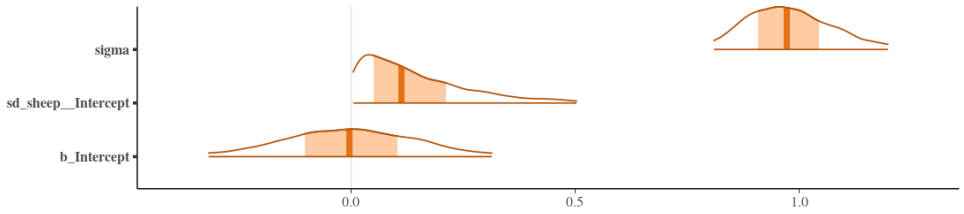


Figure 4.28: **Main sources of variation in L2 models.** CI are represented as shaded areas under the estimated posterior density curves. The global mean of the exemplified PG (PG_23) (b_Intercept) and the variances in L2 models (sd_sheep_Intercept and sigma) were depicted. Inner shadow areas represent 50% CI and density curves 95% CI.

The m.6 takes advantage of the M.6 model, and uses the global information that we learned about the proteome over time. To do so, the covariance values among time-points, showed in Figure 4.27B, were extracted together with the global time-point coefficients, to set them as a multivariate normal prior of the time-point coefficients into the m.6 model. For simplicity, m.1, m.2 and m.4 were excluded of the subsequent comparison. Of the polynomial models, the one exemplified was m.3, due to the cubic pattern was proved to perform well and resemble the nature of this sort of longitudinal proteomic data in several publications [131].

Model comparison was performed PG by PG using the next evaluation criteria: when the difference

in ELPD is larger than twice the estimated standard error, the top model is expected to have better predictive performance than the bottom model. We have reasoned previously, the discrete-time was the top-rated parametrization, and in light of m.6 model outperforming the m.5 fit at every single instance, this m.6 became our full theoretical multilevel model of change for L2 models. In other words, the change for that PG is considered significant when the m.6 obtains at least the previous difference in deviance compared to the m.0. Under this assumption, we found a total of 58 PG that truly change through AF progression in the plasma proteome. Being a bit more permissive, we use as threshold an 1.5 ELPD difference, and the list of candidates grows into 109. Table S7 includes the candidates and the corresponding estimations.

Figure 4.29, 4.30 and 4.31 show the top 3 PGs that change abundance through disease progression. Each figure depicts the L2 models containing t3 (panels B to D) and the ones just below excluding that channel (panels F to G). Comparing the third degree polynomial trendline and the coefficients of the discrete-time model, makes evident how the former misshaped the pattern and masks the erratic behaviour of t3. Some PG perfectly fit to the cubic nature, as PON1, obtaining a better performance than the m.6.

The top ranked protein is the W5NV14 (Figure 4.29), is an uncharacterized protein in our referenced database. BlastP search revealed an identity above 75% with the human Immunoglobulin lambda variable 1-40. Immunoglobulins are critical part of the immune response, however its PG abundance tend to decrease. PG_18 (Figure 4.30) represents the transcription factor SREBF1, which has a role in cholesterol biosynthesis and lipid homeostasis. This TF binds the SRE1 motif, found in the promoters of genes involved in sterol biosynthesis. The protein is initially attached to the nuclear membrane and endoplasmic reticulum, while its cleveage is being inhibited by sterols [207]. SREBP-1 has been proved to regulate parasympathetic stimulation of the heart, providing protection from arrhythmia and sudden death [172]. SREBF1 is upregulated from the beginning and stays stable until pFA. Another interesting PG is PON1 (Figure 4.31), a protein secreted from the liver to the bloodstream, stabilized by Ca+ and with has a function as hydrolase of toxic metabolites. PON1 has a protective role, avoiding oxidation of low density lipoproteins (LDL), and athero-protective effects. PON1 is upregulated early upon electrophysiological remodelling, probably as a healthy response to induced metabolic changes/cardiac injury.

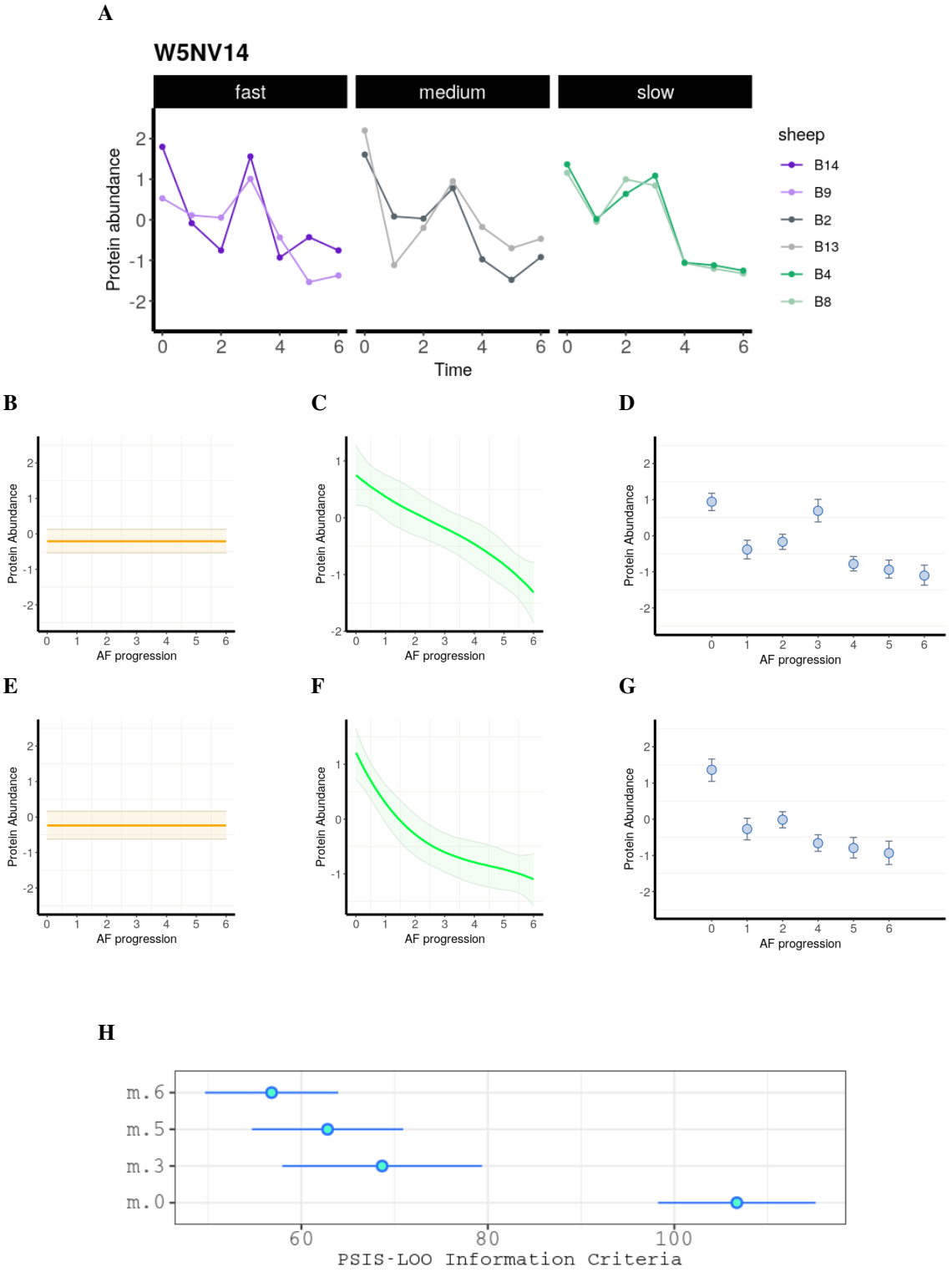


Figure 4.29: **Modelling the trajectory of W5NV14 over peripheral progression.** (A) The z-scored abundances of W5NV14 are represented in different facets regarding how long it takes to reach pAF. (B) The orange line represents the mean response of the PG through progression and the shaded area its 95% CI, estimated with the null model m.0, (C) Cubic trajectory estimated with m.3 is depicted in green together with its corresponding 95% CI and (D) time-point coefficients estimated with m.6 where blue dot represents the mean and segments the 95% CIs. (E), (F) and (G) which do not include t3, correspond respectively to the three previous panels.

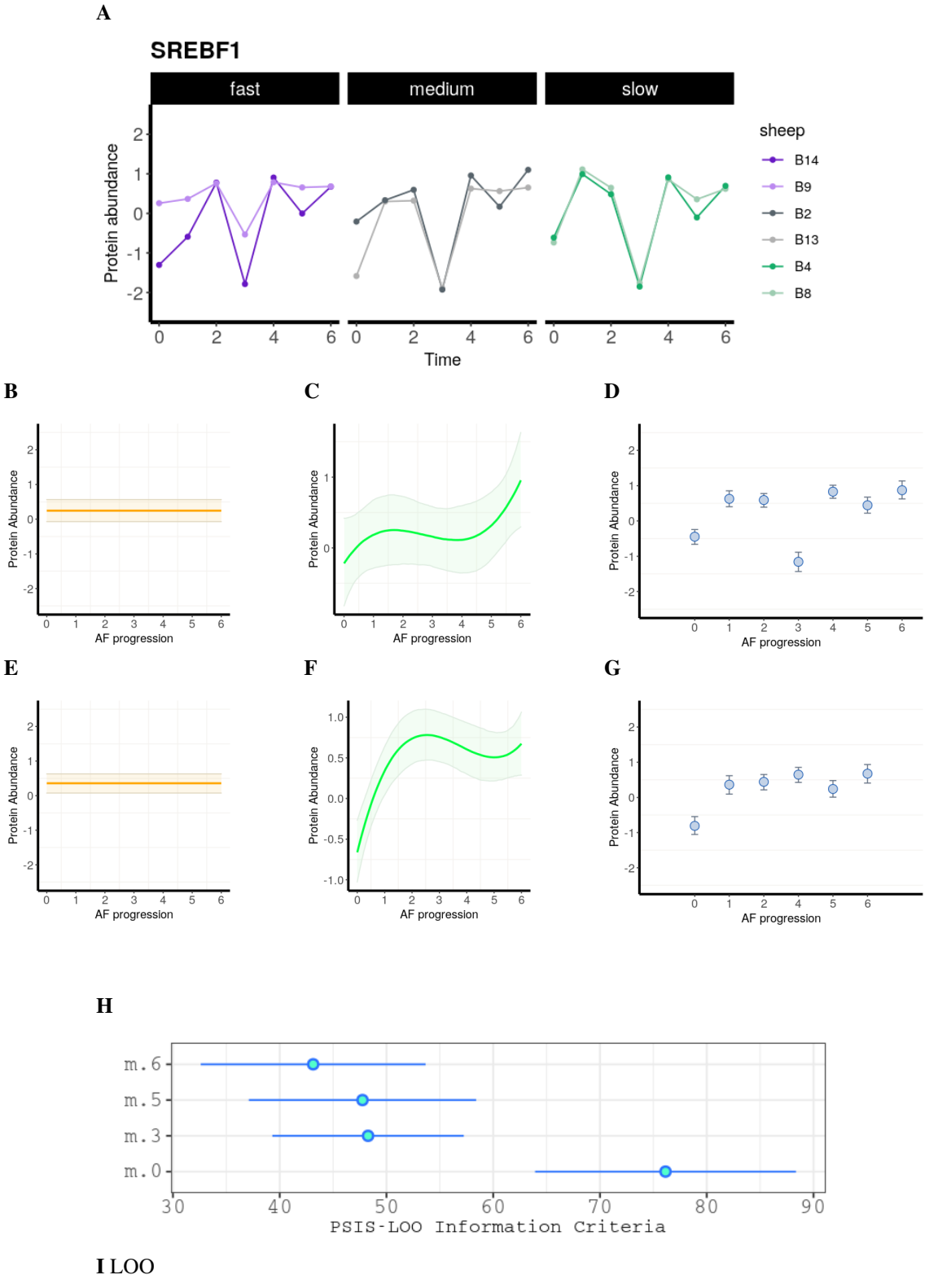


Figure 4.30: **Modelling the trajectory of SREBF1 over peripheral progression.** (A) The z-scored abundances of SREBF1 are represented in different facets regarding how long it takes to reach pAF. (B) The orange line represents the mean response of the PG through progression and the shaded area its 95% CI, estimated with the null model m.0, (C) Cubic trajectory estimated with m.3 is depicted in green together with its corresponding 95% CI and (D) time-point coefficients estimated with m.6 where blue dot represents the mean and segments the 95% CIs. (E), (F) and (G) which do not include t3, correspond respectively to the three previous panels.

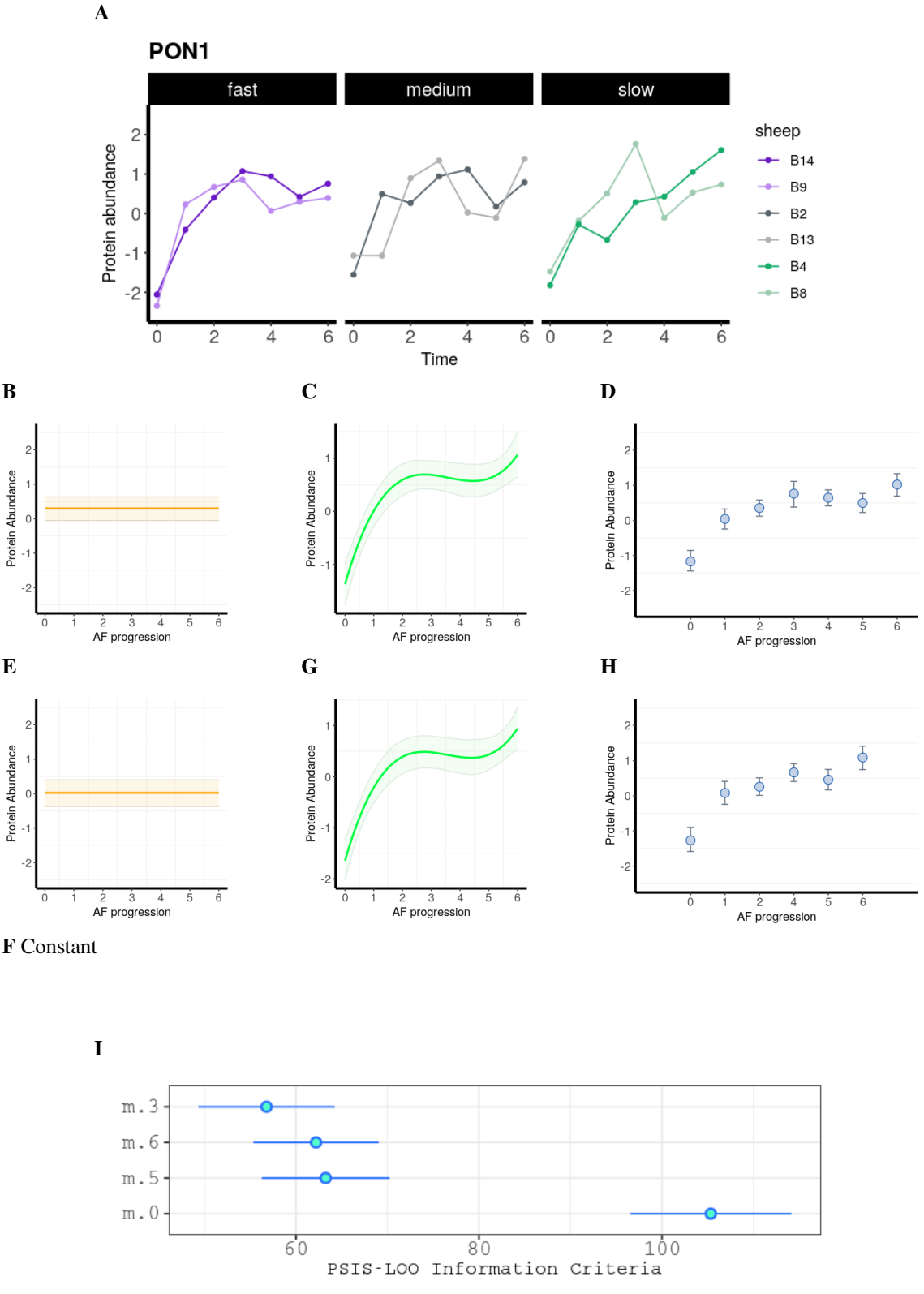


Figure 4.31: **Modelling the trajectory of PON1 over peripheral progression.** (A) The z-scored abundances of PON1 are represented in different facets regarding how long it takes to reach pAF. (B) The orange line represents the mean response of the PG through progression and the shaded area its 95% CI, estimated with the null model m.0, (C) Cubic trajectory estimated with m.3 is depicted in green together with its corresponding 95% CI and (D) time-point coefficients estimated with m.6 where blue dot represents the mean and segments the 95% CIs. (E), (F) and (G) which do not include t3, correspond respectively to the three previous panels.

2.3.1.3 The main paths of AF progression

In order to find the predominant trends in the proteome, our 109 candidates were subjected to cluster analysis. We applied a longitudinal data partitioning algorithm that takes into account the shapes of the trajectories rather than on classical distances [73], basically, a k-Means adapted for longitudinal data using shape-respecting distance. For this purpose, we recovered the PG trajectories estimated with the M.6 model. The number of clusters was selected based on the clinical relevance and observation, because to date, there is no proper quality metric in the context of respecting-shape partitioning [73]. We obtained two main clusters, comprising 70 and 39 PG respectively, shown in Figure 4.32A.

The majority of PGs, included in cluster 1, undergo an immediate and pronounced increase of their relative abundances between SR and the time in which the pacemaker starts the pacing protocol. That trend is sustained or even augmented towards pAF. Contrary, the PGs included in cluster 2 drop sharply at the first time-point, and after a sligth increase, remain stable until pAF or increase progressively in some cases. We observed that time-point 1 is the inflection point for most PGs, and that the major changes, consequence of electrical stimulation, take place in a extremely narrow time window, to progressively continue with that tendency or to recover initial levels of expression. None of the detected PGs remains steady during the first window of time to change later its pattern. This quick response to stimuli of the plasma proteome has been seen before other works [174].

GO [10] term and KEGG [9] pathway enrichment analysis was performed on clusters to interpret these set of PGs collectively, as shown in Figure 4.32. Overall, four main process arise quickly induced by the electrophysiological remodelling: 1) the coagulation cascade, 2) the complement system (CS), 3) inflammation and 4) the lipid/lipoprotein transport and metabolism. Additionally, the presence of cardiac proteins in the bloodstream was evidenced.

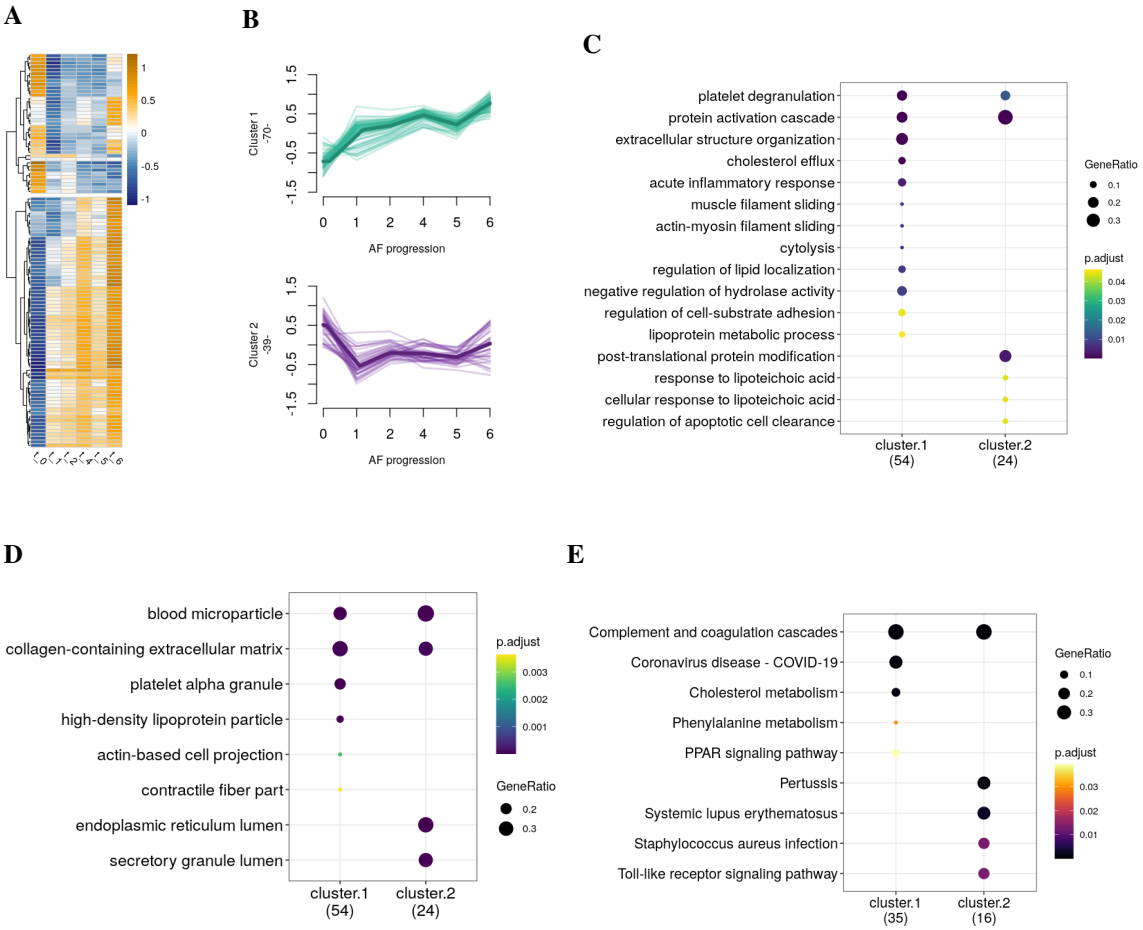


Figure 4.32: **Clustering and enrichment analysis of the mains paths of AF over peripheral progression.** (A) and (B) represent the data segmentation by the klm-shape algorithm, view as a heatmap in the former and as trajectories in the later. Estimates were obtained from the M.6 model. Subsequent figures represent as dotplots the enrichment by clusters of (C) Biological Processes, (D) Cellular Compartments and (E) KEGG pathways. The size of the dots correlates with the number of genes from the experiment included in the term and color scale with the adjusted P-value.

Since the total number of candidates is low, obtaining much information from a GO term analysis becomes complex, and partitioning this data with a larger number of clusters did not work properly by not capturing subtle differences. Although enrichment analysis provides and overall picture of the proteome through progression, we reasoned that our data should be approached at a single protein level, so we crossed our candidate list with several databases, including the Human Protein Atlas. To annotate the uncharacterized candidates, we made use of blastp. Table S7 includes the annotated information of the proteome clustering.

Most of the proteins, 79 out of the 109, belong to the human secretome, and only just the remaining 30 are predicted to be found in the cell membrane, inside the cell or as part of the extracellular matrix. Indeed, this classification might be inaccurate sometimes, seeing that some proteins exist

in both, soluble and insoluble states. Overall, liver metabolism is clearly enhanced and synthesizes the majority of the plasma secreted proteins (component system, most of the components of coagulation cascade, some acute phase components, etc), together with immune cells but these to less extent.

Nearly all the proteins identified of the coagulation cascade (Figure 4.33) were upregulated, like for instance the von Willebrand factor (vWF), released by the endothelial cells to communicate endothelial damage, as well as to strengthen the union of collagen and platelets during plug formation [164]. At the time of platelet degranulation, platelet-activation factor 4 (PF4) and vWF, among others, are released to bloodstream. Coagulation factors X (F10) and XIII (FIIIA1 and FIIIB) have their abundance increased, whereas only factor IX (F9) abundance decreased, and factor II (F2), mantained constant relative abundance. Clotting cascade leads to conversion of soluble fribinogen (FGB) into insoluble fibrin strands, which starts to increase in the between t1 and t2 of progression. On the other hand, Kininogen (KNG1), a member of the kallikrein/kinin system of the coagulation cascade and precursor of active peptide bradykinin, has a fluctuating pattern through time.

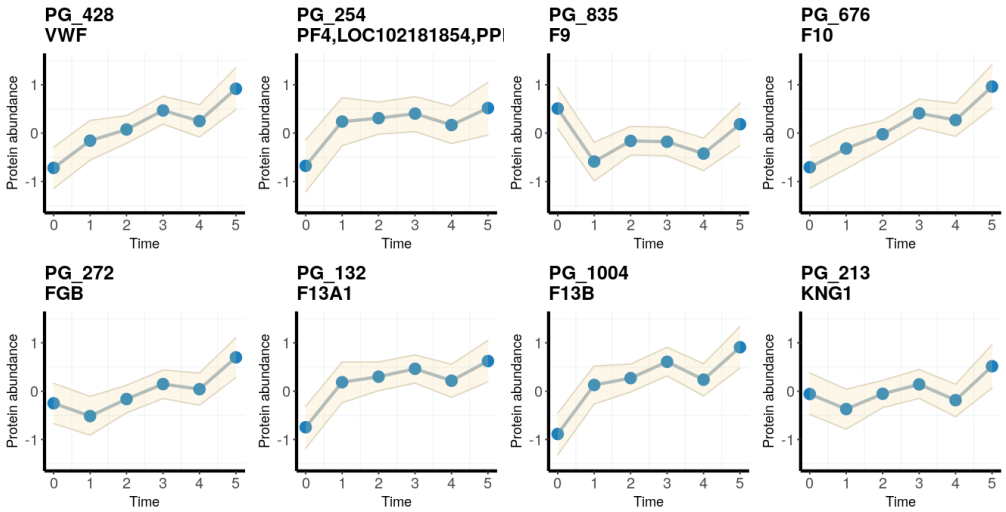


Figure 4.33: **Upregulation of the Coagulation cascade in the peripheral proteome.** Proteins of the coagulation cascade identified in the experiment are illustrated as trajectories over time. Estimates were obtained from the M.6 model. Blue dots represent the mean and the shaded area the corresponding 95% CIs.

Regarding the complement system (CS) (Figure 4.34), we observed a progressively upregulation of C1QA, which recognizes antigen-bound immunoglobulins to initiate the classical pathway of the CS. Notably, of the immunoglobulins found in our data, 7 decreased their abundance progressively since SR, while just one increased (data not shown). Following the same pattern of upregulation we found C8B, member of the membrane attack complex; complement factor D (CFD), member of

the alternative pathway; vitronectin (VTN) and clusterin (LOC101113728). The last two are negative regulators of membrane insertion of C5b-7 and C9 polymerization. The opposite behaviour was seen in tree PGs: C3, a central reaction in both classical and alternative pathways; C5, which initiates the late complement components reactions; and properidin (CFP), which positively regulates the alternative pathway and stabilizes C3/C5 convertases. These proteins are downregulated at the initial stage of disease, to further increase slightly their abundances without recovering initial levels.

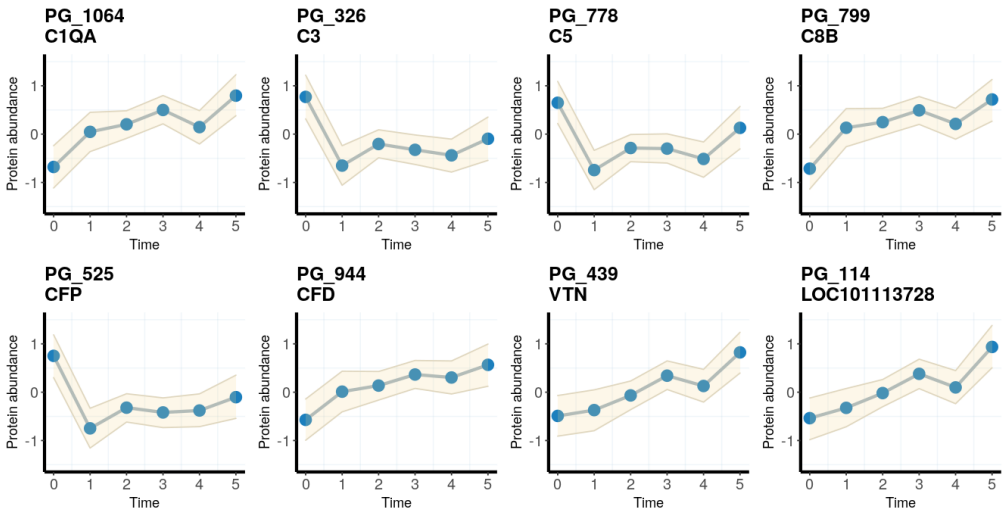


Figure 4.34: **Deregulation of the Complement system in the peripheral proteome.** Proteins of the coagulation cascade identified in the experiment are illustrated as trajectories over time. Estimates were obtained from the M.6 model. Blue dots represent the mean and the shaded area the corresponding 95% CIs.

We identified several proteins directly related to the inflammatory process, among them, proteins of the acute phase response (APPs) drew our attention (Figure 4.35). Positive APPs, which are proteins expressed in the APP, display a similar pattern of downregulation during the first time period, to later gradually increase. Positive APPs include the C-reactive protein (CRP), haptoglobin, (HP), ceruloplasmin (CP), serum amyloid A (SAA or LOC101120613) C3, and fibrinogen (see C3 and FGB in Figure 4.33). The unique negative APP identified was albumin (ALB), which decreases consistently since the first stage of AF. Another APP related protein, Fibronectin (FN1), follows the same pattern than the positive ones. Various receptors of the inmune response were deregulated similarly to the positive APPs. That is the case of CD14, co-receptor with LBP of LPS, that triggers the acute-phase response to gram-negative bacterial infection by internalization of TLR4 inducing cytokine release. The same behaviour is shown by FGL1, a ligand that binds the LAG-3 receptor of T-cells to inhibit T-cell immunity or the lubricin (PRG4), a proteoglycan secreted to the extracelular matrix, that can bind to and regulate the activity of TLRs and CD5-like molecule (LOC443475),

which mediates in the relationship between lipid homeostasis and immune response. On the other hand, the IL1RAP co-receptor of the interleukin-1 family and the Oncostatin M receptor (OSMR) which binds members of the il-6 family, become more abundant starting from the first stage of disease. Lime1 (W5NRL9), a transmembrane adaptor of the BCR and TCR mediated signaling, displayed the same tendency. Additionally, we identified S100A4 which is found intracellular in a wide range of cells, but also has a extracellular role promoting pro-inflammatory pathways. This protein is categorized as DAMP (alarmins or damage-associated molecular patterns), being released to the extracellular space from cells in response to stress. Finally, PGLYRP2 hydrolyzes bacterial cell wall peptidoglycan after bacteria or cytokine induction.

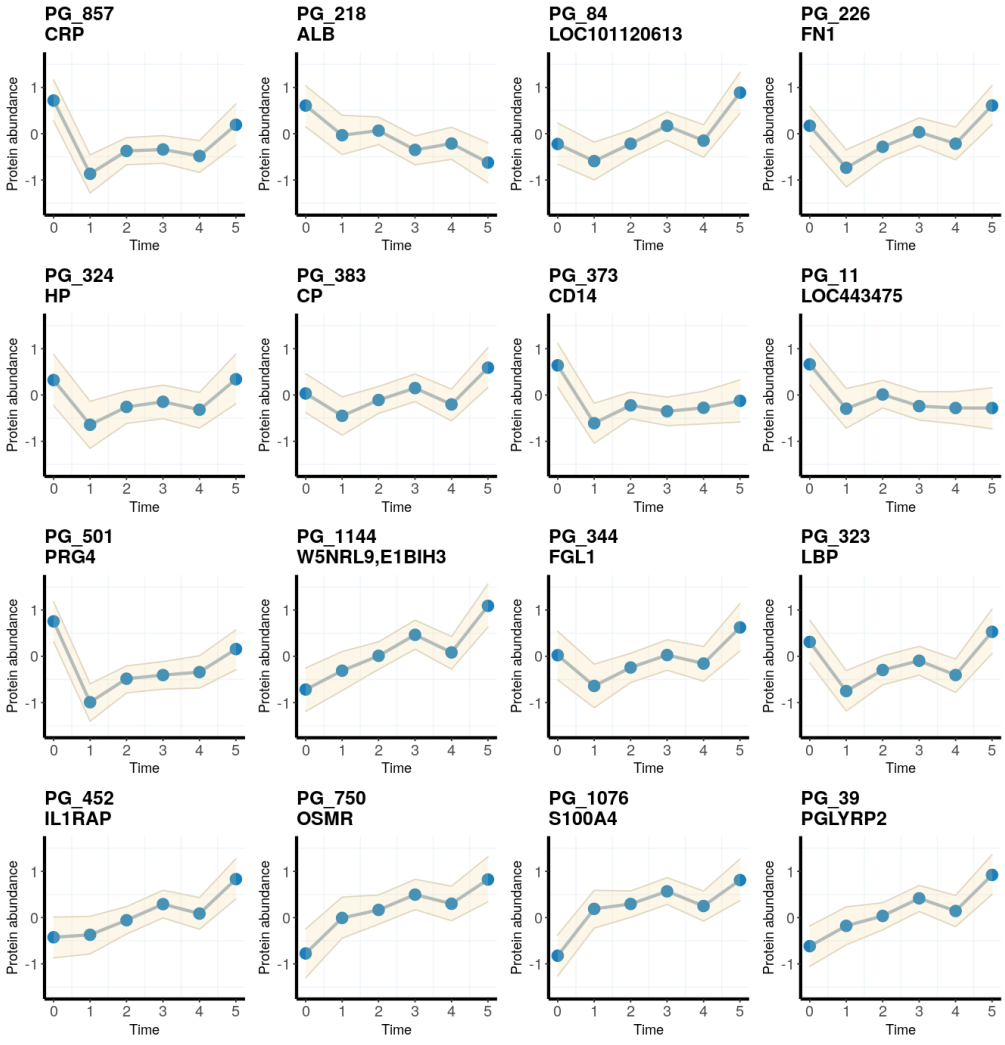


Figure 4.35: **Players of inflammation response detected in the peripheral proteome.** Estimates were obtained from the M.6 model. Blue dots represent the mean and the shaded area the corresponding 95% CIs.

The relative abundance of apolipoproteins varies with the progression of AF, Figure 4.36. Most of these proteins, APOA (A0A452G804), APOA2, APOC3, APOC4, APOD and APOM, boost

abruptly their abundances in the first time-point and then increase slowly. APOE, although increases too, does it more progressively. APOH is the only one that shows the opposite behaviour, and its abundance diminished over disease. Aditionally, we included the upregulated AZPG1 protein, which encodes the ZAG glycoprotein that promotes extensive lipogenesis and increases hepatic lipid levels. This adipokine promotes the browning of white adipose tissue, and increasing lipolysis. Furthermore, it influences glucose metabolism and insulin resistance.

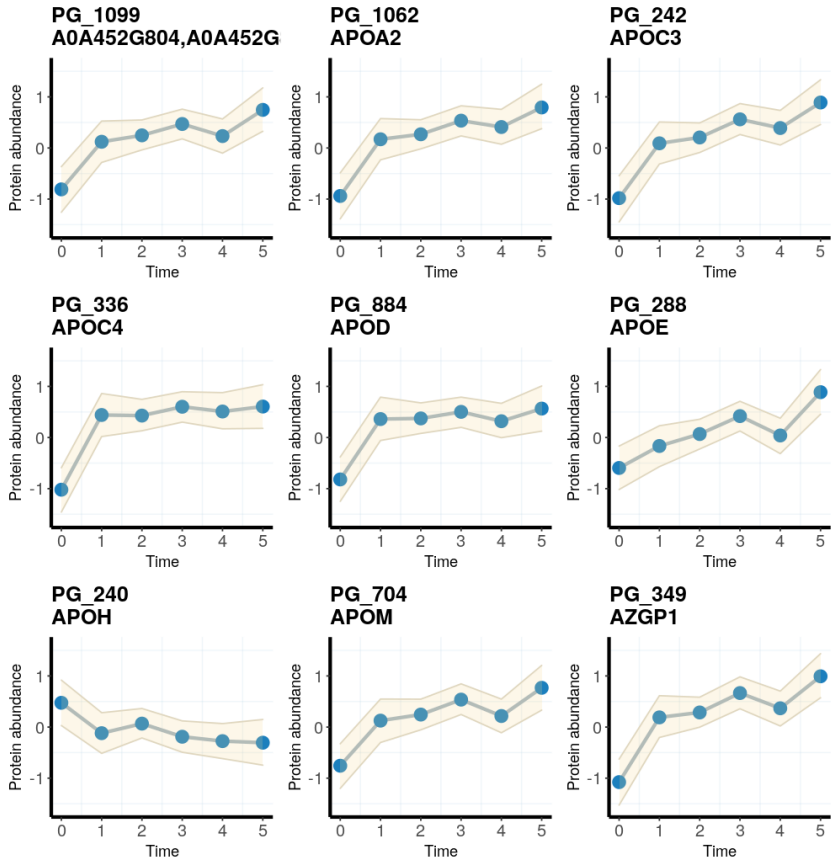


Figure 4.36: **Upregulation of lipid/lipoprotein transport in the peripheral proteome.** Estimates were obtained from the M.6 model. Blue dots represent the mean and the shaded area the corresponding 95% CIs.

We observed gradual increase of cardiac proteins of the troponin complex (TPM1, TPM2), myosins (MYL1, MYLPF) and of the alpha cardiac muscle 1 actin (ACTC1) in the bloodstream, most probably as a result of damage to heart muscle cells (Figure 4.37). Additionally, the muscle creatin kinase of muscle (CKM), also released by damaged myocardial cells into the blood, a well-known serum marker for myocardial infarction.

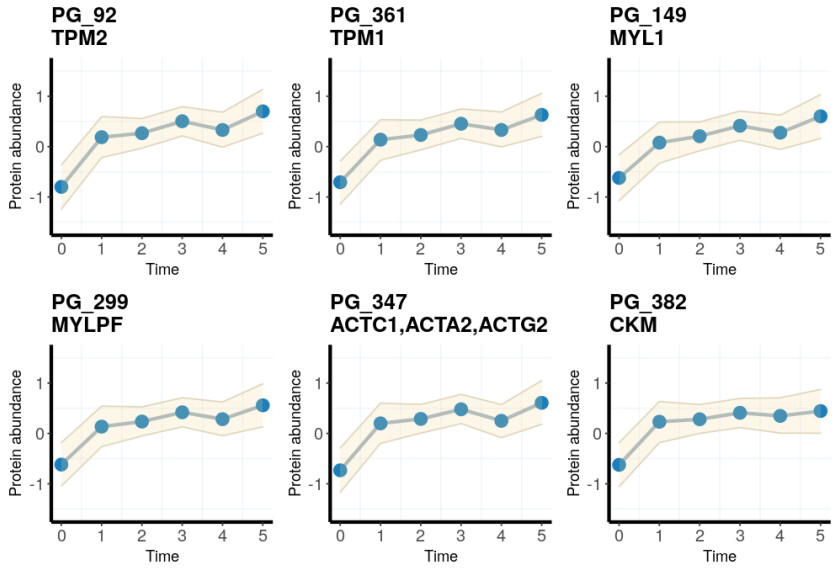


Figure 4.37: **Presence of heart debris in the peripheral proteome.** Estimates were obtained from the M.6 model. Blue dots represent the mean and the shaded area the corresponding 95% CIs.

Finally, we consider other proteins that, although related in many ways, do not exactly belong to the main processes described above. Again, this group of proteins can be separated into the two main groups, those that initially decreases and those that constantly increase. See Figure 4.38. Three candidates displayed the first pattern: Afamin (AFM), a member of the albumin family that mainly carries vitamin E or Wnt family members among others; Lactate Dehydrogenase B (LDHB) enzyme for interconversion between lactate and pyruvate, highly expressed in the atrial tissue; and Glutathione Peroxidase 3 (GPX3), which protect cells against oxidative damage. Otherwise increasing, we found five distinct amino oxidases, such as AOC3, also known as vascular adhesion protein (VAP-1). This protein can be found soluble or membrane-bound. Its soluble form, which acts as semicarbazide-sensitive amine oxidase (SSAO), has a pathogenic role, producing ammonia, aldehyde and hydrogen peroxide which might initiate oxidative stress [196]. PLTP is involucrated in the transfer of lipids between lipoproteins and in the cholesterol uptake from tissues. Fibulin 1 (FBLN1), might be secreted as part of the the extracellular matrix or even incorporated into clots. PEPD is a protease of collagen metabolism and biotinidase (BTD) is involved in the recycling of protein-bound biotin. SNAPC2 is essential for the transcription of snRNA genes, found in the nucleus of all cells. IGFBP2, in its plasma form, binds the insluline-like growth factors IGF-I and IGF-II acting as a pericellular modulator. The Parvalbumin (PVALB) although gene transfer have prove that involved directly in the relaxation process in fast muscle, it is mostly expressed in GABAergic neurons [137] and its presence in cardiomyocytes remains controversial. Finally, HRG which tethers plasminogen to the cell surface, regulates the plasmin/plasminogen system.

Plasmin disolves the fibrin of blood clots by cleaving proteins like fibrin, thrombospondin or the von Willebrand factor.

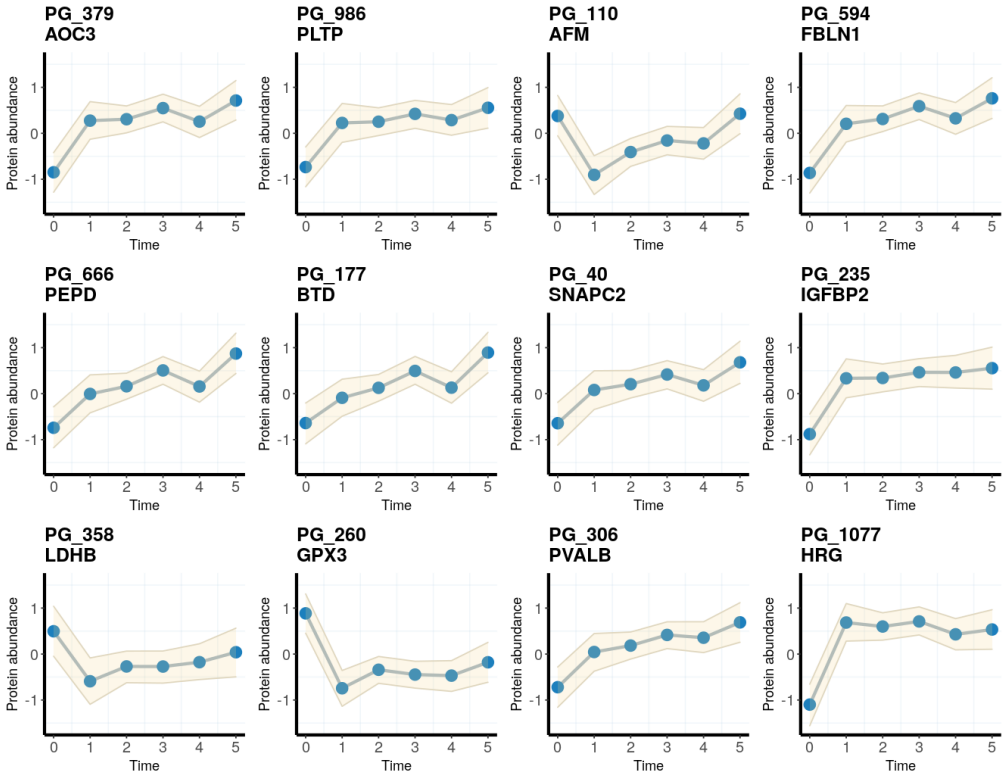


Figure 4.38: **Biomarkers affected by AF in the peripheral proteome.** Estimates were obtained from the M.6 model. Blue dots represent the mean and the shaded area the corresponding 95% CIs.

2.3.2 Proteins that change between cavity and peripheral

We next wondered if PGs behave differently in the blood taken from the right atrium (RA) compared to peripheral blood samples. To address that question, we applied conditional linear growth models including only in our data the initial time-point of SR and the final one of pAF, from both locations. We fit one protein at a time. Finally, later we included a third level for the whole proteome, and the whole set of data, to frame the magnitude of these changes in the AF progression.

2.3.2.1 Comparison of L2 models

Models were built up incrementally adding level-2 predictors and, once again, our strategy to identify the PGs was based in model comparison. We aimed to estimate $PGab_{ij}$, which represents the PG abundance of a given protein, measured in the sheep i at the time j. For the full description of these models, see the Method section **2.4.4**.

The l.0 constant model, was used as the null. As always, it assumes no changes, but in this case between SR and pAF and between peripheral and RA. Next, l.1 expand the previous with a level-2 predictor that account for the difference in PG abundance between RA and peripheral blood. l.2 is a unconditional linear growth model, that allows only for changes through time without considering differences between location. l.3 expands the former allowing differences in LA and peripheral just at time of SR, whereas l.4 allows that differences just at pAF time. Finally, l.5 considers differences at both, SR and pAF through AF progression. We fit the described models for every PG, and evaluated the performance using the PSIS-LOO IC, under the same principles as before. A Table S8 with the candidates and the preferred model is available, including 51 and 75 with a ratio of 2 and 1.5 respectively. 16 of the proteins that change between right atria and peripheral blood did not change in previous peripheral progression analysis. Of the total 75 candidates, 27 fit preferably as the more complex model (l.5), 23 fit the l.4, 16 better fit the l.3, 7 the l.2 and only 2 l.1. In Figure 4.39 examples of each are shown.

The l.1 model obtained the best score for maltase-glucoamylase (MGAM), a protein not expected in the secretome. It has constant abundance over AF, however, more protein is found in the peripheral bloodstream. Therefore, this pattern seems to be independent of disease. As we mentioned before, several circulating plasma apolipoproteins, where identified in the analysis. APOH, which was the only one decreasing, showed the same tendency and abundance levels in the RA cavity. The remaninig apolipoproteins, although increase through progression, tend to decrease (APOA, APOC3 and APOD) or stay constant (APOE, APOA2) in the RA. APOM is the unique apolipoprotein increasign in both locations. Lumican (LUM) the Ig-like protein A0A452EKN2 and the Insulin receptor substrate 4 (IRS4) fit preferentially the l.3 model, and having different abundance starting levels display similar trends in both location. Lumican is known to be secreted by cardiac fibroblasts under condition of heart failure, induced by inflammation or mecanical stimuli, important for cardiac remodelling and fibrosis [156]. IRS4 is a regulatory factor in the BMP pathway, enhancing myogenic differentiation of muscle precursor cells [56]. Two of the proteins that obtained a better performance with the l.4 model were Suprabasin (SBSN) and the previously described alarmin S100A4. Despite initial equal abundances in RA and peripheral, these proteins behave differently over progression of AF. Suprabasin, whose role remains controversial, might be involved in keratinocyte differenciation, however, is a positive marker of several diseases, and seems to act as a signalling protein. Finally, proteins like the muscle creatin kinase (CKM) and AZGP1, were better adjusted to l.5, displaying different abundances at SR and different rate of change. CKM seems to be initially more abundant in peripheral blood, to finally reach the same levels in both locations.

Curiously, the rest of proteins related to heart damage, (Figure 4.37), exhibit the same pattern. The AZPG1, precursor of the adipokine ZAG, increases its abundance mainly in peripheral blood.

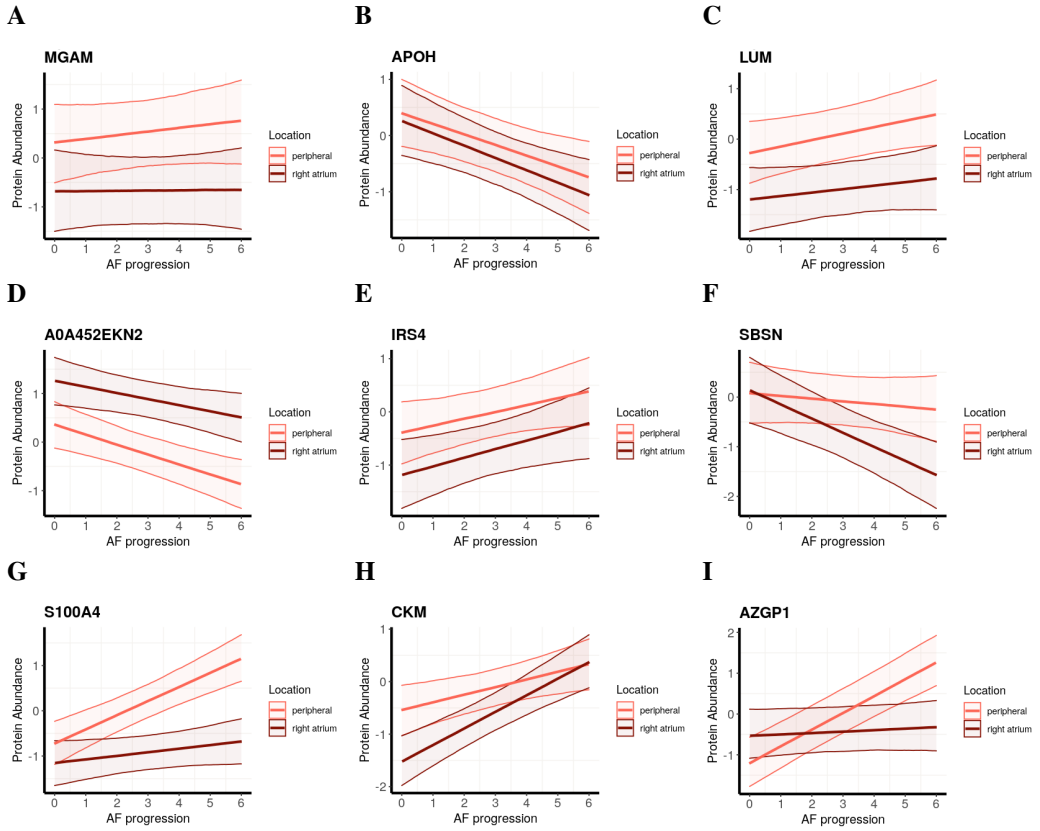


Figure 4.39: **Different protein abundances between the peripheral and RA locations.** Conditional linear trajectories are depicted in dark red for the RA and light for the peripheral samples. Shade areas represent the 95% CIs. Estimates were obtained from the l.5 model in all the cases to illustrate differences, whenever was the winner model. (A) exemplifies l.1 behaviour, (B) l.2, (C), (D) and (E) l.3, (F) and (G) l.4 and (H) and (I) l.5.

2.3.3 Exploring the magnitude of changes

To explore the magnitude of variation between all the data-points at once and thus be able to compare variation in the RA versus any other time-point, we fit the G.1 model, which expands g.1 to a third level (see the full specification in the methods section). Here, discrete time-points were modelled as dummy variables to allow for the proper interactions between predictors. After inspection of global variances (diagonal of the covariance matrix in Figure 4.40A), we observed that initial variances are the smallest of the entire progression, of both the peripheral and RA samples. However, these two proteome states anti-correlate (Figure 4.40B), meaning that variability is low at the beginning but the proteome state has a tendency to be slightly opposite for several PGs. The larger variances in the proteome are found at the pAF time in peripheral blood, whereas at this time blood proteome of the RA remains more homogeneous than most of the peripheral time-points of

the progression. As we observed with our M.6 model, variances through peripheral samples are quite homogeneous, and time-points have highly correlated proteomes over time, as expected. Notably, the proteome state in the RA is considerably different to all of the peripheral blood samples and those differences are enhanced at pAF when compared with SR. The proteome at SR in RA correlates a bit with peripheral progression, whereas at pAF in RA, the proteome anti-correlates and that behaviour is enhanced through progression, being each time more and more anti-correlated.

Overall, the global idea is that the peripheral proteome is remodeled early in t_1, and those changes are sustained or even enhanced through progression until the permanent AF is reached. The proteome of the RA evolves from a distinct state in SR that remarkably is more similar to the peripheral response to disease than to the peripheral SR proteome. Finally, the pAF-proteome in the RA arrives in an opposite state compared to the peripheral proteome.

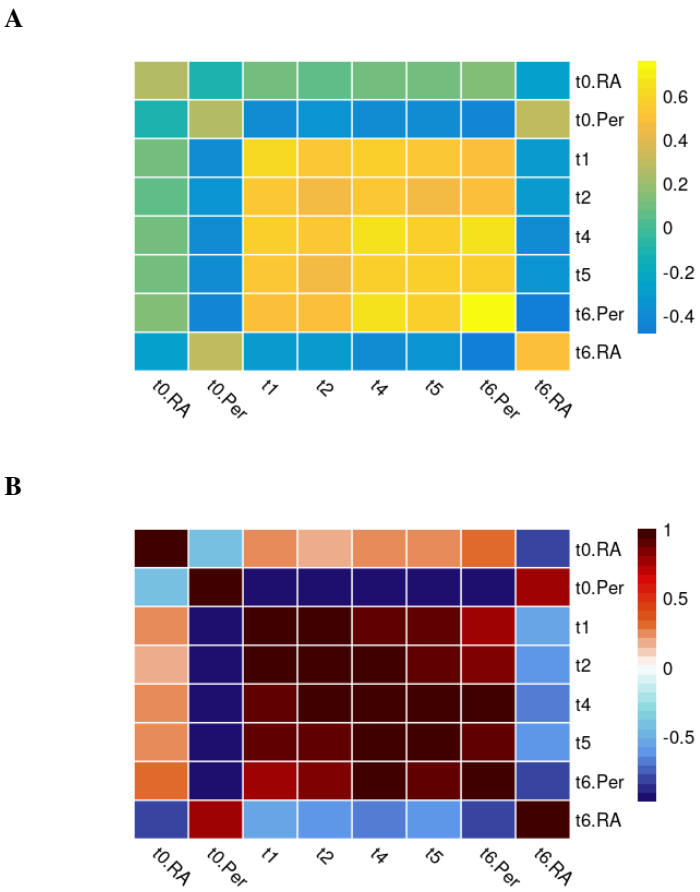
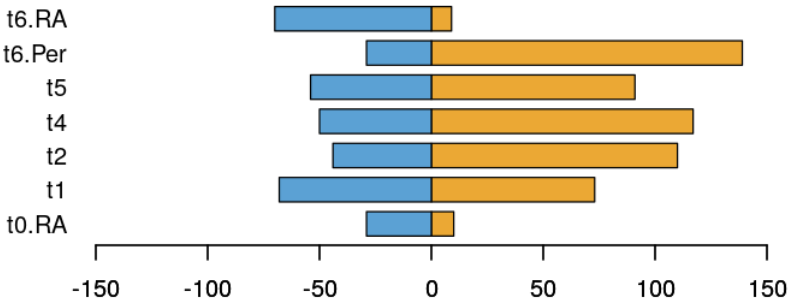


Figure 4.40: **Global changes of the serum proteome.** (A) Heatmap representing the covariance matrix of G.1 of the group-level time-point terms, variances are in the diagonal together with the corresponding covariances. (B). Heatmap representing the correlation matrix of the group-level time-point terms of G.1.

2.3.3.1 Global dynamics of the serum proteome

Using the G.1 model, we calculated the probability of a PG having changed its abundance significantly. To do so, we contrast the abundance of SR peripheral blood against the remaining time-points in peripheral and RA locations. The barplot counting the number of changes in Figure 4.41A, shows how protein abundance upregulation dominates through peripheral progression, and that upregulation increases progressively, except a bit more slightly at t5. Proteins whose abundance was reduced are the minority, and the largest number of downregulated PGs in peripheral blood occurred at the beginning of the pacing protocol. Only at this time, the number of down- and up-regulated proteins is similar. Comparing RA against peripheral SR samples, reveals that both proteomes differ at the beginning, although not as much as they differ during AF progression. Moreover, those differences are enhanced at the final time-point of pAF.

A



B

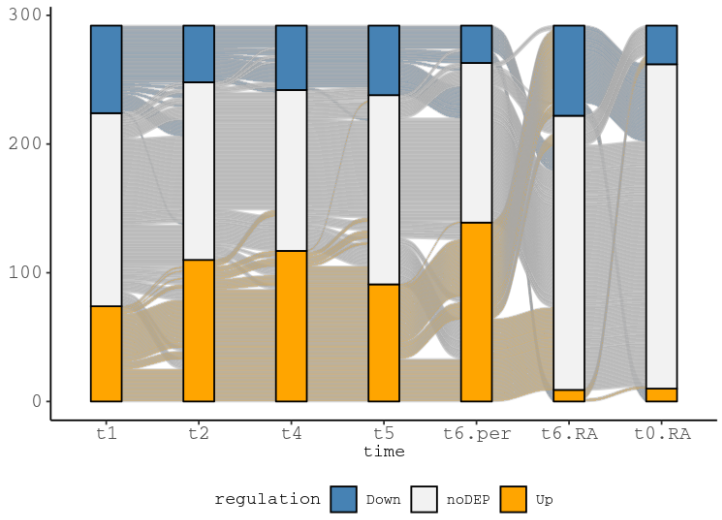


Figure 4.41: **Global dynamics of the serum proteome.** (A) Barplot representing the number of proteins differentially expressed in comparison to SR peripheral proteome. Blue means downregulation and orange upregulation. (B) Frequency distribution of the differentially expressed proteins over progression of AF. Proteins are connected by couloured lines, meanign blue downregulation, orange upregulation and gray no differential expression detected.

By inspection of the alluvial plot (Figure 4.41B), we observed that there is not a switch between up- and down- regulation proteins, or in other words, most of the proteins deregulated through time preserve their trend over AF progression. A switch in trends is only found between peripheral and RA at $t6$. In light of what we anticipated globally by examining the correlation matrix, it means that the proteome reaches the pAF with very distinct outcomes between locations. Going further, it is noticeable that the status of RA proteome changes completely with AF.

To understand this distinct behavior between peripheral and RA proteome, we must have in mind that peripheral blood was collected at the cephalic vein, before reaching the liver where many signaling molecules find their receptors, and to some extend blood is filtered. Moreover, as we and others [199] have proved previously, the LA undergoes earlier and more profound changes than RA, and LA appendage is by far the preferable location for thrombus formation during AF [210]. Therefore, summarising what we observe here, the peripheral proteome is a readout of the left atrium whereas the RA proteome is a readout of hepatic clearance. We suggest that LA is the injured tissue that triggers the systemic response that we observed and that RA has not been already as damaged to trigger that response at the onset of persistent AF.

Chapter 5

Discussion

I Prelude:

open questions in Atrial Fibrillation

Atrial fibrillation has been under extensive research for more than 50 years. However, to date, important questions remain to be elucidated. The precise role of the myriad molecular players already linked to AF has not been dealt with in-depth. Furthermore, although its progressive nature is clearly established, how the transition from paroxysmal to persistent and long-lasting permanent forms takes place is not well understood.

The advent of genome-wide assays in the last decades, and its increasing availability, evidence that the classical research approaches focusing on one experimental condition or one molecular player, have been exhausted. Sequencing-based assays provide genome-wide data, performing the scoring of hundreds to thousands of molecules simultaneously, most of the time in a comparable fashion. In the last decade, with the publication of transcriptional, proteomic and genome-wide association studies, our understanding of the biological functions affected in AF has increased considerably, as we have previously reviewed here and here.

Nevertheless, such data-driven approaches, where data leads to the hypothesis itself, are bioinformatically complex to analyze. Indeed, there are many challenges to overcome before formulating a biological question, not only related to computational problems themselves or to the translation of statistical knowledge into code but to management in general science and in particular in bioinformatics. A few that we usually have to confront are the absence of multiplatform software, inadequate support for data exchange, lack of common interfaces to handle data, an excessive number of vendor's licenses in computational science, missing or deficient documentation, inconsistency in identifiers for biological entities among databases, the insufficient visibility of some tools together with the overwhelming number of tools that perform a similar analysis, etc [68]. Taken together, besides the option of sharing code with publications, it contributes to the actual crisis in reproducibility. Although academic science is uninterruptedly evolving and this is the basis of constant improvement, some care should be taken to coordinate and standardize procedures.

Additionally to technological advances, another typical drawback to overcome when studying cardiac disease, as also occurs with other human diseases, is sample availability. Subjects from which cardiac samples can be obtained are included in the studies because they require surgical intervention, in most cases unrelated to the condition under study as is the case of AF. While the left atrium

is more revealing [199], the right atrial tissue is overrepresented due to ease of access [218]. Paired control samples, crucial given human genetic variability, are impossible to obtain due to obvious reasons. Altogether, this causes experimental designs to be unappropriated and explains the small overlap that we have shown beforehand within proteomic and transcriptomic experiments. This is where experimental models prove their value. Different cell or small animal experimental models have been used to understand the pathophysiology of AF [4, 169]. But so far, large animal models [64, 148] have proven to be the most adequate in terms of ease in pacing and similarity to reproduce experimentally human AF-molecular complexity [64, 146]. Importantly, sheep models have shown their utility to study molecular determinants of disease progression [220, 221].

In this work, we have taken advantage of a well-established, clinically relevant large animal model to analyse systemically in vivo the molecular changes that occur in the atria during the progression of AF from paroxysmal to its long-lasting forms, always in comparison with paired surgically operated sinus rhythm controls. We have performed transcriptomic and proteomic profiling of atrial tissue and cardiomyocytes as well as profiling the blood proteome. For the analysis of data, we have developed our own pipelines, benchmarked different tools and carried out correlation-based integration analysis and hierarchical modelling in a Bayesian framework. We demonstrate that the hallmarks of AF-induced atrial remodelling change only at early transitional stages at the molecular level, but remain unaltered at later stages of the disease and that the left atrium undergoes significantly more profound changes in its expression program than the right atrium. By dissecting the short time window between the paroxysmal and persistent forms, we proved that the remodelling occurring in the left atrium is sufficient to promote a systemic response in less than a few hours and we confirmed that the pro-thrombotic state, inflammation, and lipid metabolism are activated systemically as the result of AF per se, beyond being these processes associated with pre-existing comorbidities.

II First movement: *atrial molecular mapping of the tachypacing-induced long-standing AF*

In a first approach, we performed transcriptomic and proteomic profiling of both atrial tissue and isolated CMs from LAA and RAA of sheep that had been in self-sustained persistent AF for a short period (7 days) and from animals for more than a year in self-sustained persistent AF. We have also analysed the transcriptomic changes driven by AF in the PLA, finding them very similar to those in

the LAA. By comparing these two time-points of disease progression to control animals, we were able to explore the networks and pathways underlying the fundamental mechanism of AF.

The initial global analysis of the differentially expressed genes and proteins provided an unexpected finding. The changes in gene or protein expression during AF occurred early, during the transition from paroxysmal to persistent, but thereafter were unchanged for up to one year despite AF persistence. Although more than double differentially expressed features are found in chronic-control regarding transition-control, the trend of change through progression is absolutely sustained for every single feature. These differences in number are the consequence of genes and proteins bordering on statistical significance. Consequently, no changes were found between transition and chronic. This occurs both in whole atrial appendage tissue and isolated CMs, and for either gene or protein expression.

Another interesting finding from our global analysis is that changes in the LAA are more pronounced than in the RAA, in line with the view that AF is a left atria disease [199], with subsequent changes occurring in the right atrium. The left atrium undergoes significantly more profound changes in its gene expression program than the right atrium, and correlation decreases slightly in the left atrium when comparing transition to control or even slightly more if chronic are compared to control.

Globally, this temporal dynamics of the changes in gene and protein expression corresponded closely with the electrical and structural remodelling previously demonstrated in the sheep by Martins and colleagues [146]. These changes were remarkably similar to those obtained retrospectively in patients undergoing remote transmissions of AF frequency via implantable cardioverter-defibrillator/cardiac resynchronization therapy with defibrillator or pacemaker devices [123]. In both animal and human studies, the duration of the AF episodes and the electrical activation frequency increased progressively during the time window between paroxysmal and persistent. After this transition, the DF did not increase after 1 year of self-sustained persistent AF in the sheep or 3.4 years in patients. DF gradual increase correlates with the shortness of the action potential duration reflecting electrical remodelling. Atrial fibrosis and hypertrophy of cardiomyocytes being longer and wider, increase also progressively but not after transition (or the onset of persistent) becomes evident. Importantly, cardiomyocyte hypertrophy in the RA evolves slower than in the LA. These observations correlate with what we observed at the molecular level, pointing to electrical remodelling as the first manifestation and structural remodelling accumulating more gradually, although

underlying molecular changes have already occurred at the initial phase. At the end of the transition period, this results in atrial dilatation and fibrosis [5, 146]. The rate of increase in DF was different for each animal or patient, reaching a maximum at the onset of persistent AF, meaning that electrical remodelling takes shorter or longer regarding individual variability. However, these differences result in the same outcome, at the electrophysiological and morphological levels, and also in the molecular changes we observe in our data.

For the integration of transcriptome and proteome data of atrial tissue and CMs, we used multiple co-inertia (MCIA) [152], which is a correlation-based method for the joint analysis of multiple data. Correlation-based approaches seek correlative links or anchors among datasets, to then concatenate all these features together onto a common space, in which the new axes are meaningful and thereafter making comparisons with those new values possible. MCIA, in particular, maximizes the covariance between eigenvectors. Therefore, these new axes resume the common main sources of variability or principal components, and features are proximal in the new space regarding their similarity in expression patterns.

Changes in the transcriptome and proteome are not expected to be simultaneous [203, 242], and different sources of samples, although biologically related and matched as in our experimental setting, might vary to a great extent. Add to all of this that many associations are hidden by noise and different magnitudes of variability among datasets. Thereby, we found features belonging to the same pathways which did not show a correlative link and the other way around. The potential of correlation-based methods is undisputed, but some care has to be taken. That is why we subsequently performed clustering, enrichment and contribution analysis, to consider the players of each enriched pathway also regarding the experiment they belong to, but using altogether the information from a correlated group of features.

Of the drivers of the variability underlying our data, we were able to identify the first two components as disease progression and left/right identity, and related the third one to a ‘transition state’, retaining almost half of the variability. We hypothesize that the transition state reflects the early changes taking place in the initial period of AF, which in the sheep model are reflected by a rapid increase in dominant frequency. Most of the features linked to the third component were found unannotated or as pseudogenes in our reference database and, certainly, these features need further exploration and characterization. Only 17 out of the 50 top features were referenced Among them, we found promising candidates, which again requires further study.

The functional annotation of the data, and the comparison of whole tissue and CM-specific RNA-seq, allowed us to identify processes and pathways that have been previously associated with AF. For example, fibrosis, inflammation and changes in ion channels have been previously described as part of the mechanisms responsible for the perpetuation of AF [94], and contractile dysfunction in AF (referred to as atrial cardiomyopathy) is emerging as an important contributor to the disease [80, 209].

Equally, we observed changes in genes related to calcium-handling, which have a critical role in AF [84, 160]. The calcium ion-channel subunit encoding gene CACNA1C shows reduced expression in CMs during AF progression, which correlates well with functional studies using the same model [146, 221]. On the other hand, RCAN1, involved in calcineurin signalling, is one of the most up-regulated genes in our analysis, while PCP4, which exerts an opposite effect in this pathway [107], is strongly down-regulated. It has been shown that rapid atrial activity results in Ca2+ loading, which in turn triggers the Ca2+-dependent calmodulin–calcineurin–NFAT pathway to cause the down-regulation of ICaL and action potential duration reduction in the atrial cardiomyocytes [181]. RCAN1 has also been implicated in TRPC1/C4 channel-mediated activation of the calcineurin-NFAT pathway [29], and identified as top hub gene in human AF samples [222].

Our analysis has revealed changes to chromatin as a major consequence of AF in the sheep. Nuclear proteins as CENP-E and CBX1 arise from the analysis as features associated to the transition state pointing towards chromatin remodelling. The former is a centromeric associate protein, that unlike the other members of its family, is not expressed in interphase but at the time of prometaphase when the nuclear membrane breaks into vesicles. The latter is found bound to the methylated lysine residues of the histone H3 in the heterochromatin and having a role in epigenetic repression is usually found CBX1, also known as HP1-β. Global down-regulation of chromatin factors, together with a drop in core histone levels and an increase in the expression of TEs, suggestive of chromatin decompaction, all occur during AF progression. Several of these epigenomic-related changes have been associated with the decline that occurs during ageing [26]. Reduction of histones can lead to loss of heterochromatin with derepression of TEs, which would normally be silenced [215]. Hypomethylation of DNA also accompanies chromatin decompaction and has been observed in other CVD such as atherosclerosis [13, 257], and overall epigenomic changes have been argued to be a cause for the progression of common human diseases [24, 63]. It is interesting that the analysis of subcellular structures in a goat model of AF revealed dispersed heterochromatin in the nucleus of AF CMs, compared to the normal clustered aggregates found in CMs from sinus rhythm animals,

as one of the earliest changes [11]. Therefore, our results suggest that a general decrease in the nuclear organization is a hallmark of AF, and could be explored as an early marker of the disease in humans [190]. Furthermore, chromatin remodelling would lead to altered epigenomic states that reinforce the disease-related gene expression programme.

Additionally, we found relevant the presence of DIO1 coming from the atrial transcriptome, which remarks the importance of Thyroid hormone metabolism, although its role in AF is controversial in the literature [91, 259]. The enzyme product of this gene converts the thyroid hormone T4 into its activated and circulation version (T3) and is required for heart growth. DIO1 was found upregulated after ischemia-reperfusion injury at the transcriptomic and proteomic level. However, under the presence of circulating T3, the protein levels decrease in the damaged tissue, while transcript even increases its expression [195].

When analysing the transcriptomic data from the PLA, we were able to find differences in gene expression that correlate with sheep atria that transition to persistent AF at different rates. Interestingly, among these, we identified many genes related to neural cells, which together with other observations on the role of the neural system in AF [38, 82], suggests a neural input on how quickly AF progresses. We also observed that fast progressing sheep had lower expression of proliferation-related genes. These observations open up novel avenues that could help to identify and treat those patients that will progress more rapidly to permanent, and therefore more adverse, forms of AF.

There is substantial evidence in the literature indicating that AF is partially heritable. Classical genetics have documented several chromosomal loci and genetic mutations in myocardial sodium, potassium, and potassium-adenosine triphosphate channels linked to atrial arrhythmia [98, 136]. In our study, a number of those genes (such as NPPA, MYL4, PRKAG2, LMNA, SCN5A, and KCNH2) are differentially expressed in the sheep atria during AF progression. In fact, the increase in AF frequency during the progression to persistent AF is a reflection of electrical remodelling in the form of action potential duration abbreviation brought about by differential changes to ion channels, such as decreases in sodium and L-type calcium currents or increase in inward rectifier potassium current [146]. On the other hand, more than 100 new genetic loci have been associated with an increased risk of AF by GWAS, pointing to yet unexplored mechanisms underlying the disease [17]. The analysis of the variants underlying these associations will reveal how they underpin specific atrial substrates or conditions that can modify molecular functions leading to the progression of AF [142]. This source of variation might help explain the large variability in the rate of

AF remodelling and progression that has been observed in animals and in patients. Nevertheless, research efforts taking together personal genetic profiles, clinical risk factors and monitoring of AF progression [248] along with atrial cell remodelling [142] should help to better understand the risk and progression of AF. It may also help to predict the rate of AF progression and the time to completion of atrial electrical remodelling, thus improving stratification and the personalized care of AF patients.

III Second movement:
from the non-tachypacing to permanent-AF

In our second approach, we performed proteomic profiling of plasma samples collected from both the peripheral cephalic vein and the right atrium cavity (RA). In the case of peripheral samples, those cover the whole progression from a control SR stage, through the window of paroxysmal AF until the onset of persistent AF (pAF). Since cavity blood collection is a more invasive process, RA samples complemented the experiment at the time of device implantation (SR) and euthanasia pAF. By monitoring protein abundance through paroxysmal progression, we have witnessed the immediate systemic response consequence of AF per se, pinpointing various biomarkers that might be translated into the clinic for a better prognosis of the disease.

Although constantly evolving and its contribution to biosciences is not in dispute, proteomics is still challenging, either at the technical or analytical level. High-throughput proteomic data is largely noisy, and complex samples as indeed plasma proteome, are highly heterogenic and variable in abundance, becoming almost impossible the identification of trace amounts of proteins of biological relevance such as transcription factors. Error propagation, starting from the differences in the efficiency of protein extraction to the semi-stochastic nature of the MS/MS analysis or the elevated presence of missing values, forces abundances to not be comparable among proteins even of the same sample [241]. Isobaric quantification, in particular, is affected by ratio compression, which causes ratios between conditions to converge towards the mean value, affecting the accuracy of the quantification and thus the detection of changes [85]. Additionally, labelling efficiency may distort the relative abundances among conditions, as we have observed at t3, finally having to remove that sample from our experimental data. Altogether, it causes the computational analysis to be multistep and conservative, and to yield a small proportion of true candidates, compared to other high-throughput approaches such as transcriptomics. To maximize the outcomes, we developed our own pipeline, chaining high-performance software as MSGF+ [109], percolator [227] or EPI-

FANY [178], and also creating a custom database for the main search and adapted protein group inference to our particular experimental design. We proved that our results outperform other more conventional approaches such as MaxQuant [46] and QuixoT [161], already at the peptide identification level.

Given the hierarchical structure of the experimental design, clustered by sheep replicates, time and proteins (PG), we found the implementation of multilevel models (MLM) a convenient choice for the statistical analysis. MLM take into account information of the different observation units for the inference process and, for this reason, are routinely used in the literature for longitudinal data analysis [66, 131]. Furthermore, we jointly modelled the protein trajectories by pooling the information of the entire proteome, which is expected to provide more accurate estimations than treating proteins independently [43]. Overall, we employed MLM to interrogate the data, resuming specific questions in distinct models, such as: which are the proteins that change through the progression of AF? Which proteins differ in abundance between peripheral and right atrium locations? How does the proteome correlate over time? or when and where do the largest and the smallest remodelling of the proteome take place?

In our analysis, we used AF progression as a relative metric of time instead of the absolute measure of regular days since the pacing protocol starts, confirming that the proteome reacts to disease progression regardless of how long it takes to reach pAF. Importantly, this is in line with the pace of electrical and structural remodelling [146, 221] , and with our previous results regarding molecular changes in the atrial tissue and cardiomyocytes. Given that we accommodated replicate variability within our models, we truly believe that the candidates arising from the analysis change their abundances as a function of this progression. Nevertheless, we recognise that this needs further inspection of the proteins that change over regular time and are not a side effect of the disease but of the experimental procedure or randomness.

Clustering analysis found that two main groups summarise the behaviour of the peripheral plasma proteome, one increasing along progression, and another initially decreasing during the period immediately after the device activation. In light of this, and by observing the global variances of the proteome , that suddenly increase in t1 and the high correlation of the peripheral proteome afterwards, together with regulation status been preserved over time, we can state that t1 is the inflection point in our experimental conditions revealing that major remodelling takes places extremely early

in the AF progression.

Functional annotation of the data pointed to four main biological processes altered in our model: coagulation, the complement system, inflammation and lipid/lipoprotein transport and metabolism. However, since the number of candidates identified by the analysis was low and subsequently lower for those found deregulated, we had to complement the analysis by crossing information with several annotation databases. This provided important clues about underrepresented trends and biomarkers, that although do not belong to any of the main pathways affected, provide insights into the pathophysiology of the disease.

One important finding of this work is that the electrical stimulation that mimics the triggers of AF and therefore begets AF, generates a prothrombotic or hypercoagulable state per se in an extremely narrow time window in healthy and not aged individuals absent of comorbidities. Upon cardiac injury in the atria, endothelium results damaged, and the subendothelial collagen of the endocardium becomes exposed [105]. Circulating platelets bind that collagen and the von Willebrand factor (vWF), which is released by the own endothelial cells to also communicate damage, strengthening the link [177]. At the time of platelet activation, granules containing the platelet-activating factor (PF4) are discharged to the bloodstream [164]. The last step in the coagulation cascade is the conversion of soluble fibrinogen (FGB) into fibrin, forming the insoluble clot [171]. All the above mentioned biomarkers, among other coagulation factors, were found upregulated over AF progression in peripheral blood. To face the prothrombic state, fibrinolysis plays an important role in activating plasmin, which cleaves fibrin to prevent intravascular thrombosis. The alpha-2-antiplasmin (SERPINF2) is the major plasmin inhibitor in the circulation and was found upregulated through progression in our data, as well as FXIIIa, which cross-links this antiplasmin to plasmin, altering drastically its susceptibility to lysis [86] and impairing fibrinolysis. The detection of these biomarkers evidence that AF alone is promoting the prothrombotic state. Akar and colleagues [3] found that platelets and markers of thrombin generation were activated in the human heart after 15 minutes of AF, which goes in line with our observation of this remodelling happening at the initial time-point of our experimental setting.

Another important pathway arising from the analysis was inflammation, which already has been associated with the initiation, progression and maintenance of AF [53, 88]. Again, we observed in the sheep AF itself to induce inflammation, as occurs in patients with lone AF [233]. The presence of pro-inflammatory cytokines was notorious, presumably released locally by infiltrated immune

cells as an early response to tissue injury. Although those small (glyco)proteins are not detected in our data, the abundance of receptors such as OSMR and ILRAP, which binds IL-6 and IL-1 family members respectively, was increased over progression. These signalling molecules, among many other functions, promote the synthesis in the liver of positive acute-phase proteins, such as the C-reactive protein (CRP), haptoglobin, (HP), ceruloplasmin (CP), serum amyloid A (SAA or LOC101120613) C3, and fibrinogen (FGB) [198, 223]. Curiously, in our model, these biomarkers were deceased in the initial time-point compared to SR, to then later increase. We found discrepancies regarding the role of CRP, whose levels are predictive of the developmental AF. Akar et al [3] identified thrombin markers 15 min after AF induction, and pro-inflammatory markers such as CPR and IL6 were found in equal abundance under induced AF and under SR patients. This suggests different temporal dynamics for both responses, but further studies are required to reach more solid conclusions. On the other hand, negative acute-phase proteins, such as albumin (ALB) and Fibronectin (FN1), followed the expected decreasing trajectory[198]. Released increasingly through progression also by injured cells we found S100A4, an alarmin that induces the expression of proinflammatory genes via cell surface Toll-like receptors. Additionally, in close relation with TLR function, we detected the LBP, CD14 and CD5-like proteins following the same pattern as positive acute-phase proteins. Although here we evidenced important candidates, the dynamics of the inflammatory response as a consequence of stresses such as AF is complex to understand, and more efforts are needed to uncover specific targets of the inflammatory pathway that could be useful to treat AF. Moreover, there is an interplay between thrombogenesis and inflammation, starting with the fact that most of the proteins of these three major processes are synthesised by the liver, which drastically altered its metabolism as a consequence of AF, to provide a systemic response and orchestrated response [101, 113]. Additionally, it is known that IL-6 promotes megakaryocyte maturation, leading to platelet release and CRP activates platelet function [223]. In addition, the complement system (CS), deregulated in our model, is known to induce platelet activation and the other way around, platelets can activate the CS. Going further, the same system activates the inflammatory response via components such as C3a or C5a. The functional and mechanistic connections between these systems remain to be elucidated [101].

AF is known to affect cardiomyocyte energy demand due to irregular and high-frequency excitation and contraction, causing them to be under metabolic stress. AF is suggested to cause relative ischemia in the myocardium [214, 250] and in consonance, causes accumulation of lactate in the atrium. We observed reduced expression of Lactate dehydrogenase B (LDHB), which is the specific heart subunit forming the isozyme LDH1 that preferable catalyzes pyruvate from lactate [232]. The

mutation of this protein in mouse models is linked to the downregulation of mitochondrial functions and increase of hypoxia inducing factor-1α. Lower oxygen supply pushes the cardiomyocytes to be more efficient in their cardiac work, thus fatty acid oxidation desists to be the predominant source of energy in favour of glucose metabolism, which produces less ATP but consumes less oxygen in balance [132]. In agreement with this metabolic economy, slow-contracting myofilaments that demand less energy and are expressed in the embryo, such as β-chain tropomyosin, β-myosin heavy chain and myosin light-chain, increase their expression in AF patients [155]. Nevertheless, extra sources of energy are required to deal with this high demand and in consequence, the storage molecular phosphocreatine, from which to generate ATP, is mobilized by the muscle creatine kinase (CKM) [16], upregulated in our data. Its presence in the blood is a marker of muscle damage, as is the presence of several myofilaments that follow the same pattern in out model (TPM1, TPM2, MYL1, MYLPF and ACTC1). Going further, inflammation drives profound changes in lipid and lipoprotein metabolisms, as is well-known in atherosclerosis, where LDL plasma concentrations increase and HDL are known to protect against lesion development [14]. In this regard, apolipoproteins in our sheep model of AF boost abruptly their abundances in the first time-point, together with the upregulation of the transcription factor SREBPF1 which modulates cellular lipogenesis [58, 173]. Under oxysterol deficiency, this factor is activated by proteolysis and translocated to the nucleus promoting LDL receptor gene activation. This is probably linked to an uptake of LDL particles by the atrial tissue which delivers cholesterol metabolites. Additionally, we observed an increase in the expression of AZGP1, the precursor of the lipid-mobilizing adipokine ZAG. ZAG abundance in serum correlates with serum triglycerides and adipocyte fatty acid-binding protein levels whereas anticorrelates with high-density lipoprotein-cholesterol levels [194]. Furthermore, ZAG interacts with beta-adrenergic receptors predominantly found in adipocytes and is thought to act locally inducing lipid utilization [144], in our case, in the atrial tissue.

By comparing the peripheral and RA proteome, we observed that, rather counterintuitively and as previously seen with LAA and RAA, peripheral plasma undergoes significantly more profound changes than the right atrium. Most intra-atrial thrombus formation during AF takes places in the LA, in particular in the appendage (LAA) [210], which is long, with a narrow intel that favours blood stasis. Therefore, we suggest this is the main location from where injured tissue signalling is taking place. Blood flows from LA to the left ventricle towards the body, and before reaching the liver, is collected at the cephalic vein. Once at the liver, blood composition changes due to hepatic clearance. Proteins synthesized in the liver are released and many signalling proteins find their receptors, thus altering their relative abundances. The serum collected at RA reflects this effect, and

given that tissue is probably not as damaged as the LA, at least at the time between paroxysmal and persistent AF, we observed a general reduction in levels of biomarkers. In agreement, biomarkers of cardiac damage, myofilaments or CKM, although increased in both locations shown lower levels in RA, as well as the adipokine AZPG1 or the alarmin S1004A, whose abundance levels are modulated over time in the peripheral sample, whereas are sustained in RA at lower levels. We were able to observe such differences in abundances within the cardiovascular system even at the same collection time, due to the high sensitivity and reliability of our proteomic experimental design and analysis.

IV Coda:

summary, limitations and perspectives

Continued and subsequent episodes of paroxysmal AF, cause a progressive accumulation of molecular changes underlying the electrical and metabolic remodeling of the atrial myocardium reaching an irreversible state when AF becomes permanent. Through this early progression, the injured left atrial injured tissue which undergoes a more profound remodelling, drives a systemic response and thus causes the remodelling of the blood proteome. Then, after this transition from paroxysmal to persistent AF, no more molecular change occurs in the atria, except changes in chromatin and DNA indicating their senescence. Figure 5.1, summarized our vision of the AF progression.

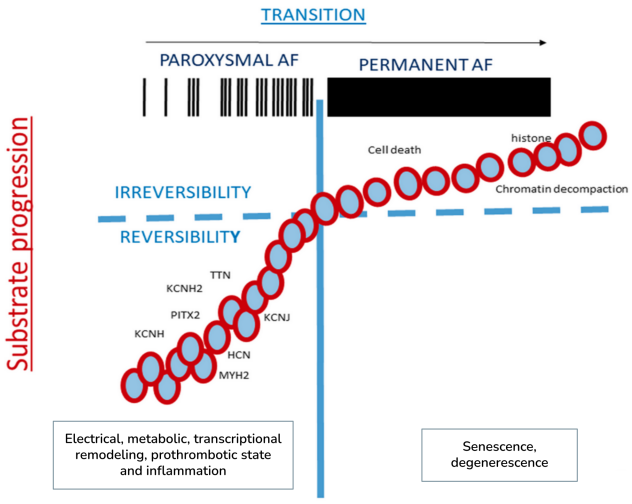


Figure 5.1: **Diagram depicting the progression of atrial fibrillation** Diagram depicting the progression of atrial fibrillation. Figure adapted from here. Progression of (AF) from paroxysmal to persistent and permanent AF together with the associated molecular remodeling (circle). TTN, titin; KCNH2, potassium voltage-gated channel subfamily H; PITX2 Paired like homodomain 2; KCNJ, potassium voltage-gated channel subfamily J; KCNH, potassium voltage-gated channel subfamily H; HCN, hyperpolarization activated cyclic-nucleotide-gated channels; MYH2, Myosine heavy chain

Our experimental large-animal model has proven its value through this work reproducing the progression of AF observed in humans. It points to systemic changes taking place only a few hours after AF induction, probably as it is happening in the atria tissue, regardless our first project did not characterize those shorter time windows of paroxysmal progression. For the AF induction, the pacemaker paces 30 seconds and hold for 10 seconds, only stopping this burst if the stimuli are generated by the tissue, autonomously. When we observed these dynamics after sheep interrogation, we realised that the timings at which the tissue becomes the greatest source of stimulation come later than the early times at which the pacing protocol starts and coincide with the blood proteome remodelling. An important conclusion that can be extracted from this evidence, is that molecular changes are the consequence of the ceaseless stimuli itself, independently of being those generated by an external source or autonomously by the own tissue. Therefore, our model simulates the existence of a very aggressive ectopic firing or trigger in the atria that uninterruptedly generates paroxysmal episodes, being this more similar to the AF charge suffered by the subject after the persistent stage of the disease. In light of this, one might think that an experimental setting where the pacemaker was programmed to induce more sporadic episodes will be more realistic. However, it becomes complex, because patients have each a distinctive progression rate and even our experimental sheep take longer or shorter to reach the time-points in the serum experiment. For sure, all these considerations and open questions need further study.

We have generated a detailed molecular map of atrial fibrillation (AF) progression in a clinically relevant large-animal model, by analysing atrial tissue and serum samples through a robust and well-designed experimental setting. Such data would be very difficult if not impossible to obtain from patients. Our results provide a framework for a comprehensive molecular analysis of the disease, pointing to novel avenues of research towards identifying early events that can lead to therapeutically targets to prevent AF-induced atrial remodelling.

We make public two comprehensive resources: 1) An user-friendly and interactive web application that assembles all the analysis presented in the first part of this thesis project, allowing for the inspection of the processed data and plenty of tools for visualization and comparison with external data. AfibOmics Browser enhances and facilitates data sharing by simply mouse navigation in any web browse. 2) The supplementary annotated Tables S7 and S8 that recap the outcome of the blood proteome analysis, useful for the future testing of local and systemic biomarkers released to the bloodstream during the early onset of Atrial Fibrillation. We plan to integrate these results in the above mention interactive web application shortly.

Chapter 6

Conclusions

1. The major molecular events related to AF progression in atrial tissue and cardiomyocytes, occur during early phases of the disease, during the transition from paroxysmal to persistent, and later stabilize as the animal moves from the transition towards the chronic state.

2. The temporal dynamics of gene and proteomic changes and measurements of atrial area points to electrical remodelling as the first manifestation and structural remodelling accumulating more gradually, although underlying molecular changes have already occurred at the initial phase.

3. Atrial fibrillation is a left atrium disease, where the left atrium undergoes significantly more profound molecular changes in its expression program besides an earlier manifestation of them than the right atrium.

4. The molecular variation concerning disease progression, atria divergence, sample source and profiling technique, can be retained in a common and meaningful space, enabling us to interrogate altogether genes and proteins grouped by similar expression patterns, without losing this experimental setting information.

5. Fibrosis, ion channel deregulation, inflammation, mitochondrial metabolism impairment and contractile dysfunction are hallmarks of AF-induced atrial remodelling occurring already at early transition stages.

6. Global downregulation of chromatin factors, together with a drop in core histones leading to de-repression of TE expression, is another hallmark of AF and could be explored as an early marker of the disease in human.

7. Changes in the posterior left atrium mirror those in the atrial appendage and suggest differential remodelling of the neuronal innervation of the autonomous nervous system during AF development at the PLA, regarding the rate of AF progression.

8. Peptide identification strategy chaining MSGF+ and percolator outperforms popular search engines such as MaxQuant or QuiXoT.

9. Multilevel models in a Bayesian framework are a powerful and flexible tool for the interrogation of complex longitudinal data.

10. Left atrial remodelling, as consequence of AF induction, causes a systemic response and remodelling of the blood proteome, which is driven by the injured tissue local signalling and the subsequent liver response.

11. The remodelling of the proteome occurs in a few hours after AF induction and can be summarized in two main groups of proteins that increase or decrease their abundance to remain sustained or gradually enhanced towards the onset of persistent AF.

12. AF induce per se a prothrombotic state involving players of the coagulation cascade, complement system and immune response together with the mobilization of lipids and lipoproteins into the bloodstream.

Chapter 7

Conclusiones

1. Los eventos moleculares más relevantes tienen lugar en las fases tempranas de la enfermedad, en el tejido auricular y en los cardiomiocitos derivados de este, sucediendo concretamente durante la transición de FA paroxística a persistente, y posteriormente estabilizándose mientras el animal evoluciona de la fase de transición hacia la crónica.

2. La dinámica temporal de los cambios en genes proteínas así como las variaciones en el área de las aurículas, apuntan a que el remodelado eléctrico es la primera manifestación de la enfermedad y que el remodelado estructural ocurre de manera más gradual, aunque los cambios moleculares relativos a ambos procesos haya ocurrido previamente en la fase inicial.

3. La Fibrilación Auricular es una enfermedad de la aurícula izquierda, en la cual esta sufre cambios en sus patrones de expresión mucho más acentuados además de una más temprana manifestación de los mismos si comparamos con la aurícula derecha.

4. La variación consecuencia de la progresión de la enfermedad, la divergencia entre aurículas, el tipo de muestra y la técnica empleada, puede ser retenida en un espacio de magnitudes comunes, que nos permite interrogar a la vez genes y proteínas que se agrupan por similitud de su patrón de expresión, todo ello sin desestimar la información experimental inicial.

5. La fibrosis, la desregulación de los canales iónicos o del metabolismo mitocondrial, la inflamación o la disfunción del aparato contractil son marcas del remodelado causado por la FA inducida que tienen lugar de manera temprana en la fase de transición.

6. La bajada de expresión global de los factores de la cromatina, así como de las histonas, llevando a la des-represión de los elementos transponibles, es otra marca de la FA y podría ser explorado como marcador temprano de la enfermedad en humanos.

7. Los cambios en la pared posterior de la aurícula izquierda son reflejo de aquellos que ocurren en la orejuela izquierda y sugieren a la vez remodelado en la inervación neural del sistema nervioso autónomo durante la FA en función de cúal sea el ratio de progresión de la enfermedad.

8. La identificación de péptidos encadenando herramientas como MSGF+ y percolator tiene un rendimiento superior respecto a otros populares motores de búsqueda tales como MaxQuant o QuiXoT.

9. Los modelos jerárquicos bayesianos son herramientas robustas y flexibles para la interrogación de datos longitudinales complejos.

10. El remodelado en la aurícula izquierda, consecuencia de la inducción de la FA, causa una respuesta sistémica y un remodelado del proteoma sanguíneo, el cual es orquestado por la señalización local desde el tejido lesionado y la subsecuente respuesta hepática.

11. El remodelado del proteoma ocurre en pocas horas tras la indución de la FA y puede sintetizarse en dos grandes grupos de proteínas que aumentan o disminuyen su abundancia inicialmente para posteriormente mantenerse o seguir cambiando muy gradualmente hasta que se alcanza la FA persistente.

12. La FA induce per se un estado protrombótico incluyendo elementos de la cascada de coagulación, el sistema del complemento y la respuesta inmune así como la movilización de lípidos y lipoproteínas en el torrente sanguíneo.

Chapter 8

References

References

[1] O. Adam, D. Lavall, K. Theobald, M. Hohl, M. Grube, S. Ameling, M. A. Sussman, S. Rosenkranz, H. K. Kroemer, H. J. Schäfers, M. Böhm, U. Laufs, *Journal of the American College of Cardiology* **2010**, *55*, 469–480.

[2] R. Aebersold, M. Mann, Mass spectrometry-based proteomics, **2003**.

[3] J. G. Akar, W. Jeske, D. J. Wilber, *Journal of the American College of Cardiology* **2008**, *51*, 1790–1793.

[4] M. Allessie, J. Ausma, U. Schotten, Electrical, contractile and structural remodeling during atrial fibrillation, **2002**.

[5] M. A. Allessie, N. M. S. De Groot, R. P. M. Houben, U. Schotten, E. Boersma, J. L. Smeets, H. J. Crijns, *Circulation: Arrhythmia and Electrophysiology* **2010**, *3*, 606–615.

[6] N. Ammash, E. A. Konik, R. D. McBane, D. Chen, J. I. Tange, D. E. Grill, R. M. Herges, T. G. McLeod, P. A. Friedman, W. E. Wysokinski, *Arteriosclerosis Thrombosis and Vascular Biology* **2011**, *31*, 2760–2766.

[7] E. C. Anderson, J. Novembre, *American Journal of Human Genetics* **2003**, *73*, 336–354.

[8] T. E. Angel, U. K. Aryal, S. M. Hengel, E. S. Baker, R. T. Kelly, E. W. Robinson, R. D. Smith, *Chemical Society Reviews* **2012**, *41*, 3912–3928.

[9] K. F. Aoki-Kinoshita, M. Kanehisa, *Methods in Molecular Biology* **2007**, *396*, 71–91.

[10] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, G. Sherlock, Gene ontology: Tool for the unification of biology, **2000**.

[11] J. Ausma, M. Wijffels, F. Thoné, L. Wouters, M. Allessie, M. Borgers, *Circulation* **1997**, *96*, 3157–3163.

[12] R. J. Aviles, D. O. Martin, C. Apperson-Hansen, P. L. Houghtaling, P. Rautaharju, R. A. Kronmal, R. P. Tracy, D. R. Van Wagoner, B. M. Psaty, M. S. Lauer, M. K. Chung, *Circulation* **2003**, *108*, 3006–3010.

[13] A. Baccarelli, M. Rienstra, E. J. Benjamin, *Circulation: Cardiovascular Genetics* **2010**, *3*, 567–573.

[14] L. Badimon, G. Vilahur, *Annals of the New York Academy of Sciences* **2012**, *1254*, 18–32.

[15] B. Baik, S. Yoon, D. Nam, *PLoS ONE* **2020**, *15*, DOI 10.1371/journal.pone.0232271.

[16] M. F. Baird, S. M. Graham, J. S. Baker, G. F. Bickerstaff, Creatine-kinase- and exercise-related muscle damage implications for muscle performance and recovery, **2012**.

[17] A. Bapat, C. D. Anderson, P. T. Ellinor, S. A. Lubitz, Genomic basis of atrial fibrillation, **2018**.

[18] G. Baruzzo, K. E. Hayer, E. J. Kim, B. DI Camillo, G. A. Fitzgerald, G. R. Grant, *Nature Methods* **2017**, *14*, 135–139.

[19] E. J. Benjamin, D. Levy, S. M. Vaziri, R. B. D'agostino, A. J. Belanger, P. A. Wolf, *JAMA: The Journal of the American Medical Association* **1994**, *271*, 840–844.

[20] E. J. Benjamin, K. M. Rice, D. E. Arking, A. Pfeufer, C. Van Noord, A. V. Smith, R. B. Schnabel, J. C. Bis, E. Boerwinkle, M. F. Sinner, A. Dehghan, S. A. Lubitz, R. B. D'Agostino, T. Lumley, G. B. Ehret, J. Heeringa, T. Aspelund, C. Newton-Cheh, M. G. Larson, K. D. Marciante, E. Z. Soliman, F. Rivadeneira, T. J. Wang, G. Eiríksdottir, D. Levy, B. M. Psaty, M. Li, A. M. Chamberlain, A. Hofman, R. S. Vasan, T. B. Harris, J. I. Rotter, W. H. Kao, S. K. Agarwal, B. H. Stricker, K. Wang, L. J. Launer, N. L. Smith, A. Chakravarti, A. G. Uitterlinden, P. A. Wolf, N. Sotoodehnia, A. Köttgen, C. M. Van Duijn, T. Meitinger, M. Mueller, S. Perz, G. Steinbeck, H. E. Wichmann, K. L. Lunetta, S. R. Heckbert, V. Gudnason, A. Alonso, S. Kääb, P. T. Ellinor, J. C. Witteman, *Nature Genetics* **2009**, *41*, 879–881.

[21] M. Bersanelli, E. Mosca, D. Remondini, E. Giampieri, C. Sala, G. Castellani, L. Milanesi, *BMC Bioinformatics* **2016**, *17*, 15.

[22] A. Bhat, S. Khanna, H. H. Chen, G. C. Gan, C. R. MacIntyre, T. C. Tan, *Heart Rhythm* **2020**, *0*, DOI 10.1016/j.hrthm.2020.06.015.

[23] R. D. Bjornson, N. J. Carriero, C. Colangelo, M. Shifman, K. H. Cheung, P. L. Miller, K. Williams, *Journal of Proteome Research* **2008**, *7*, 293–299.

[24] H. T. Bjornsson, M. Daniele Fallin, A. P. Feinberg, *Trends in Genetics* **2004**, *20*, 350–358.

[25] A. M. Bolger, M. Lohse, B. Usadel, *Bioinformatics* **2014**, *30*, 2114–2120.

[26] L. N. Booth, A. Brunet, The Aging Epigenome, **2016**.

[27] P. C. Bürkner, *Journal of Statistical Software* **2017**, *80*, 1–28.

[28] N. Calvo, P. Ramos, S. Montserrat, E. Guasch, B. Coll-Vinent, M. Domenech, F. Bisbal, S. Hevia, S. Vidorreta, R. Borras, C. Falces, C. Embid, J. M. Montserrat, A. Berruezo, A. Coca, M. Sitges, J. Brugada, L. Mont, *Europace* **2015**, *18*, 57–63.

[29] J. E. Camacho Londoño, Q. Tian, K. Hammer, L. Schröder, J. Camacho Londoño, J. C. Reil, T. He, M. Oberhofer, S. Mannebach, I. Mathar, S. E. Philipp, W. Tabellion, F. Schweda, A. Dietrich, L. Kaestner, U. Laufs, L. Birnbaumer, V. Flockerzi, M. Freichel, P. Lipp, *European Heart Journal* **2015**, *36*, 2257–2266.

[30] S. Cañón, R. Caballero, A. Herraiz-Martínez, M. Pérez-Hernández, B. López, F. Atienza, J. Jalife, L. Hove-Madsen, E. Delpón, A. Bernad, *Journal of Molecular and Cellular Cardiology* **2016**, *99*, 162–173.

[31] T. H. Cao, P. A. Quinn, J. K. Sandhu, A. A. Voors, C. C. Lang, H. M. Parry, M. Mohan, D. J. L. Jones, L. L. Ng, *The Lancet* **2015**, *385*, S26.

[32] B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. A. Brubaker, J. Guo, P. Li, A. Riddell, *Journal of Statistical Software* **2017**, *76*, 1–32.

[33] J. D. Cavalcoli, Genomic and proteomic databases: Large-scale analysis and integration of data, **2001**.

[34] F. Censi, G. Calcagnini, P. Bartolini, A. Giuliani, *PLoS ONE* **2010**, *5*, DOI 10.1371/journal.pone.0013668.

[35] A. M. Chamberlain, S. K. Agarwal, A. R. Folsom, S. Duval, E. Z. Soliman, M. Ambrose, L. E. Eberly, A. Alonso, *Heart Rhythm* **2011**, *8*, 1160–1166.

[36] K. Chandramouli, P.-Y. Qian, *Human Genomics and Proteomics* **2009**, *1*, DOI 10.4061/2009/239204.

[37] S. Chemonges, R. Gupta, P. C. Mills, S. R. Kopp, P. Sadowski, *Proteome Science* **2017**, *15*, 1–16.

[38] P. S. Chen, L. S. Chen, M. C. Fishbein, S. F. Lin, S. Nattel, *Circulation Research* **2014**, *114*, 1500–1515.

[39] S. H. Choi, L. C. Weng, C. Roselli, H. Lin, C. M. Haggerty, M. B. Shoemaker, J. Barnard, D. E. Arking, D. I. Chasman, C. M. Albert, M. Chaffin, N. R. Tucker, J. D. Smith, N. Gupta, S. Gabriel, L. Margolin, M. A. Shea, C. M. Shaffer, Z. T. Yoneda, E. Boerwinkle, N. L. Smith, E. K. Silverman, S. Redline, R. S. Vasan, E. G. Burchard, S. M. Gogarten, C. Laurie, T. W. Blackwell, G. Abecasis, D. J. Carey, B. K. Fornwalt, D. T. Smelser, A. Baras, F. E. Dewey, C. E. Jaquish, G. J. Papanicolaou, N. Sotoodehnia, D. R. Van Wagoner, B. M. Psaty, S. Kathiresan, D. Darbar, A. Alonso, S. R. Heckbert, M. K. Chung, D. M. Roden, E. J. Benjamin, M. F. Murray, K. L. Lunetta, S. A. Lubitz, P. T. Ellinor, *JAMA - Journal of the American Medical Association* **2018**, *320*, 2354–2364.

[40] A. H. Christensen, F. C. Chatelain, I. G. Huttner, M. S. Olesen, M. Soka, S. Feliciangeli, C. Horvat, C. F. Santiago, J. I. Vandenberg, N. Schmitt, S. P. Olesen, F. Lesage, D. Fatkin, *Journal of Molecular and Cellular Cardiology* **2016**, *97*, 24–35.

[41] I. E. Christophersen, M. Rienstra, C. Roselli, X. Yin, B. Geelhoed, J. Barnard, H. Lin, D. E. Arking, A. V. Smith, C. M. Albert, M. Chaffin, N. R. Tucker, M. Li, D. Klarin, N. A. Bihlmeyer, S. K. Low, P. E. Weeke, M. Müller-Nurasyid, J. G. Smith, J. A. Brody, M. N. Niemeijer, M. Dörr, S. Trompet, J. Huffman, S. Gustafsson, C. Schurmann, M. E. Kleber, L. P. Lyytikäinen, I. Seppälä, R. Malik, A. R. Horimoto, M. Perez, J. Sinisalo, S. Aeschbacher, S. Thériault, J. Yao, F. Radmanesh, S. Weiss, A. Teumer, S. H. Choi, L. C. Weng, S. Clauss, R. Deo, D. J. Rader, S. H. Shah, A. Sun, J. C. Hopewell, S. Debette, G. Chauhan, Q. Yang, B. B. Worrall, G. Paré, Y. Kamatani, Y. P. Hagemeijer, N. Verweij, J. E. Siland, M. Kubo, J. D. Smith, D. R. Van Wagoner, J. C. Bis, S. Perz, B. M. Psaty, P. M. Ridker, J. W. Magnani, T. B. Harris, L. J. Launer, M. B. Shoemaker, S. Padmanabhan, J. Haessler, T. M. Bartz, M. Waldenberger, P. Lichtner, M. Arendt, J. E. Krieger, M. Kähönen, L. Risch, A. J. Mansur, A. Peters, B. H. Smith, L. Lind, S. A. Scott, Y. Lu, E. B. Bottinger, J. Hernesniemi, C. M. Lindgren, J. A. Wong, J. Huang, M. Eskola, A. P. Morris, I. Ford, A. P. Reiner, G. Delgado, L. Y. Chen, Y. D. I. Chen, R. K. Sandhu, M. Li, E. Boerwinkle, L. Eisele, L. Lannfelt, N. Rost, C. D. Anderson, K. D. Taylor, A. Campbell, P. K. Magnusson, D. Porteous, L. J. Hocking, E. Vlachopoulou, N. L. Pedersen, K. Nikus, M. Orho-Melander, A. Hamsten, J. Heeringa, J. C. Denny, J. Kriebel, D. Darbar, C. Newton-Cheh, C. Shaffer, P. W. Macfarlane, S. Heilmann-Heimbach, P. Almgren, P. L. Huang, N. Sotoodehnia, E. Z. Soliman, A. G. Uitterlinden, A. Hofman, O. H. Franco, U. Völker, K. H. Jöckel, M. F. Sinner, H. J. Lin, X. Guo, M. Dichgans, E. Ingelsson, C. Kooperberg, O. Melander, R. J. Loos, J. Laurikka, D. Conen, J. Rosand, P. Van Der Harst, M. L. Lokki, S. Kathiresan, A. Pereira, J. W. Jukema, C. Hayward, J. I. Rotter, W. März, T. Lehtimäki, B. H. Stricker,

M. K. Chung, S. B. Felix, V. Gudnason, A. Alonso, D. M. Roden, S. Kääb, D. I. Chasman, S. R. Heckbert, E. J. Benjamin, T. Tanaka, K. L. Lunetta, S. A. Lubitz, P. T. Ellinor, *Nature Genetics* **2017**, *49*, 946–952.

[42] A. Conesa, P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, A. McPherson, M. W. Szcześniak, D. J. Gaffney, L. L. Elo, X. Zhang, A. Mortazavi, A survey of best practices for RNA-seq data analysis, **2016**.

[43] E. M. Conlon, B. L. Postier, B. A. Methé, K. P. Nevin, D. R. Lovley, *PLoS ONE* **2012**, *7*, 52137.

[44] J. Costa-Silva, D. Domingues, F. M. Lopes, RNA-Seq differential expression analysis: An extended review and a software tool, **2017**.

[45] J. S. Cottrell, Protein identification using MS/MS data, **2011**.

[46] J. Cox, M. Mann, *Nature Biotechnology* **2008**, *26*, 1367–1372.

[47] M. J. Daly, J. D. Rioux, S. F. Schaffner, T. J. Hudson, E. S. Lander, *Nature Genetics* **2001**, *29*, 229–232.

[48] A. M. De Jong, A. H. Maass, S. U. Oberdorf-Maass, D. J. Van Veldhuisen, W. H. Van Gilst, I. C. Van Gelder, Mechanisms of atrial structural changes caused by stretch occurring before and during early atrial fibrillation, **2011**.

[49] A. I. De Souza, A. J. Camm, *Circulation: Arrhythmia and Electrophysiology* **2012**, *5*, 1036–1043.

[50] A. I. De Souza, S. Cardin, R. Wait, Y. L. Chung, M. Vijayakumar, A. Maguy, A. J. Camm, S. Nattel, *Journal of Molecular and Cellular Cardiology* **2010**, *49*, 851–863.

[51] J. A. Delaney, X. Yin, J. D. Fontes, E. R. Wallace, A. Skinner, N. Wang, B. G. Hammill, E. J. Benjamin, L. H. Curtis, S. R. Heckbert, *SAGE Open Medicine* **2018**, *6*, 205031211875944.

[52] A. Deshmukh, J. Barnard, H. Sun, D. Newton, L. Castel, G. Pettersson, D. Johnston, E. Roselli, A. M. Gillinov, K. McCurry, C. Moravec, J. D. Smith, D. R. Van Wagoner, M. K. Chung, *Circulation: Arrhythmia and Electrophysiology* **2015**, *8*, 32–41.

[53] W. Y. Ding, D. Gupta, G. Y. Lip, Atrial fibrillation and the prothrombotic state: Revisiting Virchow's triad in 2020, **2020**.

[54] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, T. R. Gingeras, *Bioinformatics* **2013**, *29*, 15–21.

[55] S. Doll, M. Dreßen, P. E. Geyer, D. N. Itzhak, C. Braun, S. A. Doppler, F. Meier, M. A. Deutsch, H. Lahm, R. Lange, M. Krane, M. Mann, *Nature Communications* **2017**, *8*, 1–13.

[56] G. Dörpholz, A. Murgai, J. Jatzlau, D. Horbelt, M. P. Belverdi, C. Heroven, I. Schreiber, G. Wendel, K. Ruschke, S. Stricker, P. Knaus, *Scientific Reports* **2017**, *7*, 1–17.

[57] S. Duane, A. D. Kennedy, B. J. Pendleton, D. Roweth, *Physics Letters B* **1987**, *195*, 216–222.

[58] D. Eberlé, B. Hegarty, P. Bossard, P. Ferré, F. Foufelle, SREBP transcription factors: Master regulators of lipid homeostasis, **2004**.

[59] P. Ellinghaus, R. J. Scheubel, D. Dobrev, U. Ravens, J. Holtz, J. Huetter, U. Nielsch, H. Morawietz, *Journal of Thoracic and Cardiovascular Surgery* **2005**, *129*, 1383–1390.

[60] P. T. Ellinor, K. L. Lunetta, N. L. Glazer, A. Pfeufer, A. Alonso, M. K. Chung, M. F. Sinner, P. I. W de Bakker, M. Mueller, S. A. Lubitz, E. Fox, D. Darbar, N. L. Smith, J. D. Smith, R. B. Schnabel, E. Z. Soliman, K. M. Rice, D. R. Van Wagoner, B.-M. M. Beckmann, C. Van Noord, K. Wang, G. B. Ehret, J. I. Rotter, S. L. Hazen, G. Steinbeck, A. V. Smith, L. J. Launer, T. B. Harris, S. Makino, M. Nelis, D. J. Milan, S. Perz, T. T. Esko, A. Köttgen, S. Moebus, C. Newton-Cheh, M. Li, S. Möhlenkamp, T. J. Wang, W. H. Linda Kao, R. S. Vasan, M. M. Nöthen, C. A. MacRae, B. H. Ch Stricker, A. Hofman, A. G. Uitterlinden, D. Levy, E. Boerwinkle, A. Metspalu, E. J. Topol, A. Chakravarti, V. Gudnason, B. M. Psaty, D. M. Roden, T. Meitinger, H.-E. E. Wichmann, J. C. M Witteman, J. Barnard, D. E. Arking, E. J. Benjamin, S. R. Heckbert, S. Kääb, P. I. De Bakker, M. Mueller, S. A. Lubitz, E. Fox, D. Darbar, N. L. Smith, J. D. Smith, R. B. Schnabel, E. Z. Soliman, K. M. Rice, D. R. Van Wagoner, B.-M. M. Beckmann, C. Van Noord, K. Wang, G. B. Ehret, J. I. Rotter, S. L. Hazen, G. Steinbeck, A. V. Smith, L. J. Launer, T. B. Harris, S. Makino, M. Nelis, D. J. Milan, S. Perz, T. T. Esko, A. Köttgen, S. Moebus, C. Newton-Cheh, M. Li, S. Möhlenkamp, T. J. Wang, W. H. Linda Kao, R. S. Vasan, M. M. Nöthen, C. A. MacRae, B. H. Ch Stricker, A. Hofman, A. G. Uitterlinden, D. Levy, E. Boerwinkle, A. Metspalu, E. J. Topol, A. Chakravarti, V. Gudnason, B. M. Psaty, D. M. Roden, T. Meitinger, H.-E. E. Wichmann, J. C. Witteman, J. Barnard, D. E. Arking, E. J. Benjamin, S. R. Heckbert, S. Kääb, **2010**, *42*, 240–244.

[61] P. T. P. T. Ellinor, K. L. K. L. Lunetta, C. M. C. M. C. M. C. M. Albert, N. L. Glazer, M. D. Ritchie, A. V. Smith, D. E. Arking, M. Müller-Nurasyid, B. P. Krijthe, S. A. Lubitz, J. C. Bis, M. K. Chung, M. Dörr, K. Ozaki, J. D. Roberts, J. D. G. D. G. D. G. Smith, A. Pfeufer, M. F. Sinner, K. Lohman, J. Ding, N. L. Smith, J. D. G. D. G. D. G. Smith, M. Rienstra, K. M. Rice, D. R. Van Wagoner, J. W. Magnani, R. Wakili, S. Clauss, J. I. Rotter, G. Steinbeck, L. J. Launer, R. W. Davies, M. Borkovich, T. B. Harris, H. Lin, U. Völker, H. Völzke, D. J. Milan, A. Hofman, E. Boerwinkle, L. Y. Chen, E. Z. Soliman,

B. F. Voight, G. Li, A. Chakravarti, M. Kubo, U. B. Tedrow, L. M. Rose, P. M. Ridker, D. Conen, T. Tsunoda, T. Furukawa, N. Sotoodehnia, S. Xu, N. Kamatani, D. Levy, Y. Nakamura, B. Parvez, S. Mahida, K. L. Furie, J. Rosand, R. Muhammad, B. M. Psaty, T. Meitinger, S. Perz, H.-E. E. Wichmann, J. C. M. Witteman, W. H. L. Kao, S. Kathiresan, D. M. Roden, A. G. Uitterlinden, F. Rivadeneira, B. McKnight, M. Sjögren, A. B. Newman, Y. Liu, M. H. Gollob, O. Melander, T. Tanaka, B. H. C. Stricker, S. B. Felix, A. Alonso, D. Darbar, J. Barnard, D. I. Chasman, S. R. Heckbert, E. J. Benjamin, V. Gudnason, S. Kääb, M. Muller-Nurasyid, B. P. Krijthe, S. A. Lubitz, J. C. Bis, M. K. Chung, M. Dorr, K. Ozaki, J. D. Roberts, J. D. G. D. G. D. G. Smith, A. Pfeufer, M. F. Sinner, K. Lohman, J. Ding, N. L. Smith, J. D. G. D. G. D. G. Smith, M. Rienstra, K. M. Rice, D. R. Van Wagoner, J. W. Magnani, R. Wakili, S. Clauss, J. I. Rotter, G. Steinbeck, L. J. Launer, R. W. Davies, M. Borkovich, T. B. Harris, H. Lin, U. Volker, H. Volzke, D. J. Milan, A. Hofman, E. Boerwinkle, L. Y. Chen, E. Z. Soliman, B. F. Voight, G. Li, A. Chakravarti, M. Kubo, U. B. Tedrow, L. M. Rose, P. M. Ridker, D. Conen, T. Tsunoda, T. Furukawa, N. Sotoodehnia, S. Xu, N. Kamatani, D. Levy, Y. Nakamura, B. Parvez, S. Mahida, K. L. Furie, J. Rosand, R. Muhammad, B. M. Psaty, T. Meitinger, S. Perz, H.-E. E. Wichmann, J. C. M. Witteman, W. H. L. Kao, S. Kathiresan, D. M. Roden, A. G. Uitterlinden, F. Rivadeneira, B. McKnight, M. Sjogren, A. B. Newman, Y. Liu, M. H. Gollob, O. Melander, T. Tanaka, B. H. C. Stricker, S. B. Felix, A. Alonso, D. Darbar, J. Barnard, D. I. Chasman, S. R. Heckbert, E. J. Benjamin, V. Gudnason, S. Kaab, M. Müller-Nurasyid, B. P. Krijthe, S. A. Lubitz, J. C. Bis, M. K. Chung, M. Dörr, K. Ozaki, J. D. Roberts, J. D. G. D. G. D. G. Smith, A. Pfeufer, M. F. Sinner, K. Lohman, J. Ding, N. L. Smith, J. D. G. D. G. D. G. Smith, M. Rienstra, K. M. Rice, D. R. Van Wagoner, J. W. Magnani, R. Wakili, S. Clauss, J. I. Rotter, G. Steinbeck, L. J. Launer, R. W. Davies, M. Borkovich, T. B. Harris, H. Lin, U. Völker, H. Völzke, D. J. Milan, A. Hofman, E. Boerwinkle, L. Y. Chen, E. Z. Soliman, B. F. Voight, G. Li, A. Chakravarti, M. Kubo, U. B. Tedrow, L. M. Rose, P. M. Ridker, D. Conen, T. Tsunoda, T. Furukawa, N. Sotoodehnia, S. Xu, N. Kamatani, D. Levy, Y. Nakamura, B. Parvez, S. Mahida, K. L. Furie, J. Rosand, R. Muhammad, B. M. Psaty, T. Meitinger, S. Perz, H.-E. E. Wichmann, J. C. M. Witteman, W. H. L. Kao, S. Kathiresan, D. M. Roden, A. G. Uitterlinden, F. Rivadeneira, B. McKnight, M. Sjögren, A. B. Newman, Y. Liu, M. H. Gollob, O. Melander, T. Tanaka, B. H. C. Stricker, S. B. Felix, A. Alonso, D. Darbar, J. Barnard, D. I. Chasman, S. R. Heckbert, E. J. Benjamin, V. Gudnason, S. Kääb, *Nature Genetics* **2012**, *44*, 670–675.

[62] S. J. Emrich, W. B. Barbazuk, L. Li, P. S. Schnable, *Genome Research* **2007**, *17*, 69–73.

[63] A. P. Feinberg, Phenotypic plasticity and the epigenetics of human disease, **2007**.

[64] D. Filgueiras-Rama, N. F. Price, R. P. Martins, M. Yamazaki, U. M. R. Avula, K. Kaur, J. Kalifa, S. R. Ennis, E. Hwang, V. Devabhaktuni, J. Jalife, O. Berenfeld, *Circulation: Arrhythmia and Electrophysiology* **2012**, *5*, 1160–1167.

[65] M. Fischer, B. Y. Renard, *Bioinformatics* **2016**, *32*, 1040–1047.

[66] G. M. Fitzmaurice, C. Ravichandran, A primer in longitudinal data analysis, **2008**.

[67] A. M. Frank, *Journal of Proteome Research* **2009**, *8*, 2241–2252.

[68] J. C. Fuller, P. Khoueiry, H. Dinkel, K. Forslund, A. Stamatakis, J. Barry, A. Budd, T. G. Soldatos, K. Linssen, A. M. Rajput in EMBO Reports, *Vol. 14*, European Molecular Biology Organization, **2013**, pp. 302–304.

[69] V. Fuster, L. E. Rydén, R. W. Asinger, D. S. Cannom, H. J. Crijns, R. L. Frye, J. L. Halperin, G. N. Kay, W. W. Klein, S. Lévy, R. L. McNamara, E. N. Prystowsky, L. S. Wann, D. G. Wyse, R. J. Gibbons, E. M. Antman, J. S. Alpert, D. P. Faxon, V. Fuster, G. Gregoratos, L. F. Hiratzka, A. K. Jacobs, R. O. Russell, S. C. Smith, W. W. Klein, A. Alonso-Garcia, C. Blomström-Lundqvist, G. De Backer, M. Flather, J. Hradec, A. Oto, A. Parkhomenko, S. Silber, A. Torbicki, ACC/AHA/ESC guidelines for the management of patients with atrial fibrillation: Executive summary a report of the american college of cardiology/american heart association task force on practice guidelines and the european society of cardiology committee, **2001**.

[70] N. Gaborit, M. Steenman, G. Lamirault, N. Le Meur, S. Le Bouter, G. Lande, J. Léger, F. Charpentier, T. Christ, D. Dobrev, D. Escande, S. Nattel, S. Demolombe, *Circulation* **2005**, *112*, 471–481.

[71] L. Gadenz, J. Hashemi, H. Shariat, L. Gula, D. P. Redfearn, Clinical role of dominant frequency measurements in atrial fibrillation ablation - A systematic review, **2017**.

[72] L. Gatto, S. Gibb, J. Rainer, MSnbase, e cient and elegant R-based processing and visualisation of raw mass spectrometry data, **2020**.

[73] C. Genolini, R. Ecochard, M. Benghezal, T. Driss, S. Andrieu, F. Subtil, *PLoS ONE* **2016**, *11*, e0150738.

[74] A. S. Go, E. M. Hylek, K. A. Phillips, Y. C. Chang, L. E. Henault, J. V. Selby, D. E. Singer, *Journal of the American Medical Association* **2001**, *285*, 2370–2375.

[75] K. C. Gracia, H. Husi in Computational Biology, Codon Publications, **2019**, pp. 119–142.

[76] D. Greenbaum, C. Colangelo, K. Williams, M. Gerstein, Comparing protein abundance and mRNA expression levels on a genomic scale, **2003**.

[77] D. F. Gudbjartsson, D. O. Arnar, A. Helgadottir, S. Gretarsdottir, H. Holm, A. Sigurdsson, A. Jonasdottir, A. Baker, G. Thorleifsson, K. Kristjansson, A. Palsson, T. Blondal, P. Sulem, V. M. Backman, G. A. Hardarson, E. Palsdottir, A. Helgason, R. Sigurjonsdottir, J. T. Sverrisson, K. Kostulas, M. C. Ng, L. Baum, W. Y. So, K. S. Wong, J. C. Chan, K. L. Furie, S. M. Greenberg, M. Sale, P. Kelly, C. A. MacRae, E. E. Smith, J. Rosand, J. Hillert, R. C. Ma, P. T. Ellinor, G. Thorgeirsson, J. R. Gulcher, A. Kong, U. Thorsteinsdottir, K. Stefansson, *Nature* **2007**, *448*, 353–357.

[78] D. F. Gudbjartsson, H. Helgason, S. A. Gudjonsson, F. Zink, A. Oddson, A. Gylfason, S. Besenbacher, G. Magnusson, B. V. Halldorsson, E. Hjartarson, G. T. Sigurdsson, S. N. Stacey, M. L. Frigge, H. Holm, J. Saemundsdottir, H. T. Helgadottir, H. Johannsdottir, G. Sigfusson, G. Thorgeirsson, J. T. Sverrisson, S. Gretarsdottir, G. B. Walters, T. Rafnar, B. Thjodleifsson, E. S. Bjornsson, S. Olafsson, H. Thorarinsdottir, T. Steingrimsdottir, T. S. Gudmundsdottir, A. Theodors, J. G. Jonasson, A. Sigurdsson, G. Bjornsdottir, J. J. Jonsson, O. Thorarensen, P. Ludvigsson, H. Gudbjartsson, G. I. Eyjolfsson, O. Sigurdardottir, I. Olafsson, D. O. Arnar, O. T. Magnusson, A. Kong, G. Masson, U. Thorsteinsdottir, A. Helgason, P. Sulem, K. Stefansson, *Nature Genetics* **2015**, *47*, 435–444.

[79] D. F. Gudbjartsson, H. Holm, S. Gretarsdottir, G. Thorleifsson, G. B. Walters, G. Thorgeirsson, J. Gulcher, E. B. Mathiesen, I. Njølstad, A. Nyrnes, T. Wilsgaard, E. M. Hald, K. Hveem, C. Stoltenberg, G. Kucera, T. Stubblefield, S. Carter, D. Roden, M. C. Ng, L. Baum, W. Y. So, K. S. Wong, J. C. Chan, C. Gieger, H. E. Wichmann, A. Gschwendtner, M. Dichgans, G. Kuhlenbäumer, K. Berger, E. B. Ringelstein, S. Bevan, H. S. Markus, K. Kostulas, J. Hillert, S. Sveinbjörnsdóttir, E. M. Valdimarsson, M. L. Løchen, R. C. Ma, D. Darbar, A. Kong, D. O. Arnar, U. Thorsteinsdottir, K. Stefansson, *Nature Genetics* **2009**, *41*, 876–878.

[80] J. B. Guichard, S. Nattel, Atrial Cardiomyopathy: A Useful Notion in Cardiac Disease Management or a Passing Fad?, **2017**.

[81] N. Gupta, N. Bandeira, U. Keich, P. A. Pevzner, *Journal of the American Society for Mass Spectrometry* **2011**, *22*, 1111–1120.

[82] G. Gussak, A. Pfenniger, L. Wren, M. Gilani, W. Zhang, S. Yoo, D. A. Johnson, A. Burrell, B. Benefield, G. Knight, B. P. Knight, R. Passman, J. J. Goldberger, G. Aistrup, J. Andrew

Wasserstrom, Y. Shiferaw, R. Arora, *JCI Insight* **2019**, *4*, DOI 10.1172/jci.insight.130532.

[83] J. Heijman, D. Linz, U. Schotten, Dynamics of Atrial Fibrillation Mechanisms and Comorbidities, **2021**.

[84] J. Heijman, N. Voigt, S. Nattel, D. Dobrev, *Circulation Research* **2014**, *114*, 1483–1499.

[85] A. Hogrebe, L. Von Stechow, D. B. Bekker-Jensen, B. T. Weinert, C. D. Kelstrup, J. V. Olsen, *Nature Communications* **2018**, *9*, 1–13.

[86] R. Al-Horani, *Cardiovascular & Hematological Agents in Medicinal Chemistry* **2015**, *12*, 91–125.

[87] F. B. Hu, M. F. Leitzmann, M. J. Stampfer, G. A. Colditz, W. C. Willett, E. B. Rimm, *Archives of Internal Medicine* **2001**, *161*, 1542–1548.

[88] Y. F. Hu, Y. J. Chen, Y. J. Lin, S. A. Chen, Inflammation and the pathogenesis of atrial fibrillation, **2015**.

[89] T. Huang, J. Wang, W. Yu, Z. He, Protein inference: A review, **2012**.

[90] W. Huber, A. Von Heydebreck, H. Sültmann, A. Poustka, M. Vingron in Bioinformatics, *Vol. 18*, Oxford University Press, **2002**.

[91] G. Iervasi, A. Pingitore, P. Landi, M. Raciti, A. Ripoli, M. Scarlattini, A. L'Abbate, L. Donato, *Circulation* **2003**, *107*, 708–713.

[92] Y. Ishii, R. B. Schuessler, S. L. Gaynor, K. Hames, R. J. Damiano, *Journal of Thoracic and Cardiovascular Surgery* **2017**, *153*, 1357–1365.

[93] J. Jalife, O. Berenfeld, M. Mansour, Mother rotors and fibrillatory conduction: A mechanism of atrial fibrillation, **2002**.

[94] J. Jalife, K. Kaur, Atrial remodeling, fibrosis, and atrial fibrillation, **2015**.

[95] M. M. Jennings, J. K. Donahue, Connexin remodeling contributes to atrial fibrillation, **2013**.

[96] Y. Y. Jiang, H. T. Hou, Q. Yang, X. C. Liu, G. W. He, *Scientific Reports* **2017**, *7*, DOI 10.1038/s41598-017-10590-w.

[97] Y. Jin, O. H. Tam, E. Paniagua, M. Hammell, *Bioinformatics* **2015**, *31*, 3593–3599.

[98] D. P. Judge, The complex genetics of atrial fibrillation, **2012**.

[99] A. S. Kaler, L. C. Purcell, *BMC Genomics* **2019**, *20*, DOI 10.1186/s12864-019-5992-7.

[100] W. B. Kannel, E. J. Benjamin, Status of the Epidemiology of Atrial Fibrillation, **2008**.

[101] J. C. Kaski, A. L. Arrebola-Moreno, *Revista Española de Cardiología (English Edition)* **2011**, *64*, 551–553.

[102] S. V. Katikireddi, C. L. Niedzwiedz, F. Popham, D. Srinivasa, V. Katikireddi, *Open* **2012**, *2*, 1790.

[103] D. G. Katritsis, G. Boriani, F. G. Cosio, G. Hindricks, P. Jäis, M. E. Josephson, R. Keegan, Y. H. Kim, B. P. Knight, K. H. Kuck, D. A. Lane, G. Y. Lip, H. Malmborg, H. Oral, C. Pappone, S. Themistoclakis, K. A. Wood, C. Blomström-Lundqvist, B. Gorenek, N. Dagres, G. A. Dan, M. A. Vos, G. Kudaiberdieva, H. Crijns, K. Roberts-Thomson, Y. J. Lin, D. Vanegas, W. R. Caorsi, E. Cronin, J. Rickard, European heart rhythm association (EHRA) consensus document on the management of supraventricular arrhythmias, endorsed by Heart Rhythm Society (HRS), Asia-Pacific Heart Rhythm Society (APHRS), and Sociedad Latinoamericana de Estimulación Cardiaca y Elect, **2017**.

[104] M. D. Kertai, W. Qi, Y. J. Li, F. W. Lombard, Y. Liu, M. P. Smith, M. Stafford-Smith, M. F. Newman, C. A. Milano, J. P. Mathew, M. V. Podgoreanu, *Journal of Molecular and Cellular Cardiology* **2016**, *92*, 109–115.

[105] A. A. Khan, G. N. Thomas, G. Y. Lip, A. Shantsila, Endothelial function in patients with atrial fibrillation, **2020**.

[106] M. S. Kharlap, A. V. Timofeeva, L. E. Goryunova, G. L. Khaspekov, S. L. Dzemeshkevich, V. V. Ruskin, R. S. Akchurin, S. P. Golitsyn, R. S. Beabealashvilli in Annals of the New York Academy of Sciences, *Vol. 1091*, Blackwell Publishing Inc., **2006**, pp. 205–217.

[107] E. E. Kim, A. Shekhar, J. Lu, X. Lin, F. Y. Liu, J. Zhang, M. Delmar, G. I. Fishman, *Journal of Clinical Investigation* **2014**, *124*, 5027–5036.

[108] K. H. Kim, Y. Nakaoka, H. G. Augustin, G. Y. Koh, *Cell Reports* **2018**, *23*, 2455–2466.

[109] S. Kim, P. A. Pevzner, *Nature Communications* **2014**, *5*, 1–10.

[110] P. Kirchhof, S. Benussi, D. Kotecha, A. Ahlsson, D. Atar, B. Casadei, M. Castella, H. C. Diener, H. Heidbuchel, J. Hendriks, G. Hindricks, A. S. Manolis, J. Oldgren, B. A. Popescu, U. Schotten, B. Van Putte, P. Vardas, S. Agewall, J. Camm, G. Baron Esquivias, W. Budts, S. Carerj, F. Casselman, A. Coca, R. De Caterina, S. Deftereos, D. Dobrev, J. M. Ferro, G. Filippatos, D. Fitzsimons, B. Gorenek, M. Guenoun, S. H. Hohnloser, P. Kolh, G. Y. Lip, A. Manolis, J. McMurray, P. Ponikowski, R. Rosenhek, F. Ruschitzka, I. Savelieva, S. Sharma, P. Suwalski, J. L. Tamargo, C. J. Taylor, I. C. Van Gelder, A. A. Voors, S. Windecker, J. L. Zamorano, K. Zeppenfeld, 2016 ESC Guidelines for the management of atrial fibrillation developed in collaboration with EACTS, **2016**.

[111] D. Ko, M. D. Benson, D. Ngo, Q. Yang, M. G. Larson, T. J. Wang, L. Trinquart, D. D. McManus, S. A. Lubitz, P. T. Ellinor, R. S. Vasan, R. E. Gerszten, E. J. Benjamin, H. Lin, *Journal of the American Heart Association* **2019**, *8*, DOI 10.1161/JAHA.118.010976.

[112] T. Koenig, B. H. Menze, M. Kirchner, F. Monigatti, K. C. Parker, T. Patterson, J. J. Steen, F. A. Hamprecht, H. Steen, *Journal of Proteome Research* **2008**, *7*, 3708–3717.

[113] P. Korantzopoulos, K. P. Letsas, G. Tse, N. Fragakis, C. A. Goudis, T. Liu, Inflammation and atrial fibrillation: A comprehensive review, **2018**.

[114] L. P. Lai, J. L. Lin, C. S. Lin, H. M. Yeh, Y. G. Tsay, C. F. Lee, H. H. Lee, Z. F. Chang, J. J. Hwang, S. U. Ming-Jai, Y. Z. Tseng, S. K. Huang, *Journal of Cardiovascular Electrophysiology* **2004**, *15*, 214–223.

[115] G. Lamirault, N. Gaborit, N. Le Meur, C. Chevalier, G. Lande, S. Demolombe, D. Escande, S. Nattel, J. J. Léger, M. Steenman, *Journal of Molecular and Cellular Cardiology* **2006**, *40*, 173–184.

[116] B. Langmead, S. L. Salzberg, *Nature Methods* **2012**, *9*, 357–359.

[117] J. D. Lanzer, F. Leuschner, R. Kramann, R. T. Levinson, J. Saez-Rodriguez, Big Data Approaches in Heart Failure Research, **2020**.

[118] R. Latchamsetty, A. G. Kocheril, *Journal of atrial fibrillation* **2009**, *2*, 204.

[119] A. A. Leroux, J. Detilleux, C. F. Sandersen, L. Borde, R. M. Houben, A. Al Haidar, T. Art, H. Amory, *Journal of Veterinary Internal Medicine* **2013**, *27*, 1563–1570.

[120] B. Li, C. N. Dewey, *BMC Bioinformatics* **2011**, *12*, 1–16.

[121] D. Li, S. Fareh, T. K. Leung, S. Nattel, *Circulation* **1999**, *100*, 87–95.

[122] H. Li, R. Durbin, *Bioinformatics* **2009**, *25*, 1754–1760.

[123] J. M. Lillo-Castellano, J. J. González-Ferrer, M. Marina-Breysse, J. B. Martínez-Ferrer, L. Pérez-Álvarez, J. Alzueta, J. G. Martínez, A. Rodríguez, J. C. Rodríguez-Pérez, I. Anguera, X. Viñolas, A. García-Alberola, J. G. Quintanilla, J. M. Alfonso-Almazán, J. García, L. Borrego, V. Cañadas-Godoy, N. Pérez-Castellano, J. Pérez-Villacastín, J. Jiménez-Díaz, J. Jalife, D. Filgueiras-Rama, *Europace* **2020**, *22*, 704–715.

[124] D. Linz, A. D. Elliott, M. Hohl, V. Malik, U. Schotten, D. Dobrev, S. Nattel, M. Böhm, J. Floras, D. H. Lau, P. Sanders, Role of autonomic nervous system in atrial fibrillation, **2019**.

[125] D. Linz, C. Ukena, F. Mahfoud, H. R. Neuberger, M. Böhm, Atrial autonomic innervation: A target for interventional antiarrhythmic therapy?, **2014**.

[126] G. Lippi, F. Sanchis-Gomar, G. Cervellin, *International Journal of Stroke* **2021**, *16*, 217–221.

[127] R. Lister, R. C. O'Malley, J. Tonti-Filippini, B. D. Gregory, C. C. Berry, A. H. Millar, J. R. Ecker, *Cell* **2008**, *133*, 523–536.

[128] G. Liu, A. Papa, A. N. Katchman, S. I. Zakharov, D. Roybal, J. A. Hennessey, J. Kushner, L. Yang, B. X. Chen, A. Kushnir, K. Dangas, S. P. Gygi, G. S. Pitt, H. M. Colecraft, M. Ben-Johny, M. Kalocsay, S. O. Marx, *Nature* **2020**, *577*, 695–700.

[129] T. Y. Liu, H. H. Huang, D. Wheeler, Y. Xu, J. A. Wells, Y. S. Song, A. P. Wiita, *Cell Systems* **2017**, *4*, 636–644.e9.

[130] Y. Liu, Q. Shi, Y. Ma, Q. Liu, The role of immune cells in atrial fibrillation, **2018**.

[131] J. J. Locascio, A. Atri, *Dementia and Geriatric Cognitive Disorders Extra* **2011**, *1*, 330–357.

[132] G. D. Lopaschuk, J. S. Jaswal in Journal of Cardiovascular Pharmacology, *Vol. 56*, **2010**, pp. 130–140.

[133] M. I. Love, W. Huber, S. Anders, *Genome Biology* **2014**, *15*, 550.

[134] S.-K. K. Low, A. Takahashi, Y. Ebana, K. Ozaki, I. E. Christophersen, P. T. Ellinor, A. Consortium, S. Ogishima, M. Yamamoto, M. Satoh, M. Sasaki, T. Yamaji, M. Iwasaki, S. Tsugane, K. Tanaka, M. Naito, K. Wakai, H. Tanaka, T. Furukawa, M. Kubo, K. Ito, Y. Kamatani, T. Tanaka, **2017**, *49*, 953–958.

[135] S. A. Lubitz, K. L. Lunetta, H. Lin, D. E. Arking, S. Trompet, G. Li, B. P. Krijthe, D. I. Chasman, J. Barnard, M. E. Kleber, M. Dörr, K. Ozaki, A. V. Smith, M. Müller-Nurasyid, S. Walter, S. K. Agarwal, J. C. Bis, J. A. Brody, L. Y. Chen, B. M. Everett, I. Ford, O. H. Franco, T. B. Harris, A. Hofman, S. Kääb, S. Mahida, S. Kathiresan, M. Kubo, L. J. Launer, P. W. Macfarlane, J. W. Magnani, B. McKnight, D. D. McManus, A. Peters, B. M. Psaty, L. M. Rose, J. I. Rotter, G. Silbernagel, J. D. Smith, N. Sotoodehnia, D. J. Stott, K. D. Taylor, A. Tomaschitz, T. Tsunoda, A. G. Uitterlinden, D. R. Van Wagoner, U. Völker, H. Völzke, J. M. Murabito, M. F. Sinner, V. Gudnason, S. B. Felix, W. März, M. Chung, C. M. Albert, B. H. Stricker, T. Tanaka, S. R. Heckbert, J. W. Jukema, A. Alonso, E. J. Benjamin, P. T. Ellinor, *Journal of the American College of Cardiology* **2014**, *63*, 1200–1210.

[136] S. A. Lubitz, X. Yin, J. D. Fontes, J. W. Magnani, M. Rienstra, M. Pai, M. L. Villalon, R. S. Vasan, M. J. Pencina, D. Levy, M. G. Larson, P. T. Ellinor, E. J. Benjamin, *JAMA - Journal of the American Medical Association* **2010**, *304*, 2263–2269.

[137] E. K. Lucas, S. J. Markwardt, S. Gupta, J. H. Meador-Woodruff, J. D. Lin, L. Overstreet-Wadiche, R. M. Cowell, *Journal of Neuroscience* **2010**, *30*, 7227–7235.

[138] C. Magnussen, T. J. Niiranen, F. M. Ojeda, F. Gianfagna, S. Blankenberg, I. Njølstad, E. Vartiainen, S. Sans, G. Pasterkamp, M. Hughes, S. Costanzo, M. B. Donati, P. Jousilahti, A. Linneberg, T. Palosaari, G. De Gaetano, M. Bobak, H. M. Den Ruijter, E. Mathiesen, T. Jørgensen, S. Söderberg, K. Kuulasmaa, T. Zeller, L. Iacoviello, V. Salomaa, R. B. Schnabel, *Circulation* **2017**, *136*, 1588–1597.

[139] T. Maier, M. Güell, L. Serrano, Correlation of mRNA and protein in complex biological samples, **2009**.

[140] J. C. Man, K. Van Duijvenboden, P. H. Krijger, I. B. Hooijkaas, I. Van Der Made, C. De Gier-De Vries, V. Wakker, E. E. Creemers, W. De Laat, B. J. Boukens, V. M. Christoffels, *Circulation Research* **2021**, 115–129.

[141] R. Mandapati, A. Skanes, J. Chen, O. Berenfeld, J. Jalife, *Circulation* **2000**, *101*, 194–199.

[142] S. A. Mann, R. Otway, G. Guo, M. Soka, L. Karlsdotter, G. Trivedi, M. Ohanian, P. Zodgekar, R. A. Smith, M. A. Wouters, R. Subbiah, B. Walker, D. Kuchar, P. Sanders, L. Griffiths, J. I. Vandenberg, D. Fatkin, *Journal of the American College of Cardiology* **2012**, *59*, 1017–1025.

[143] M. Markl, D. C. Lee, N. Furiasse, M. Carr, C. Foucar, J. Ng, J. Carr, J. J. Goldberger, *Circulation: Cardiovascular Imaging* **2016**, *9*, DOI 10.1161/CIRCIMAGING.116.004984.

[144] M. P. Marrades, J. A. Martínez, M. J. Moreno-Aliaga, *Journal of Physiology and Biochemistry* **2008**, *64*, 61–66.

[145] S. Martínez-Bartolomé, P. Navarro, F. Martín-Maroto, D. López-Ferrer, A. Ramos-Fernández, M. Villar, J. P. García-Ruiz, J. Vázquez, *Molecular and Cellular Proteomics* **2008**, *7*, 1135–1145.

[146] R. P. Martins, K. Kaur, E. Hwang, R. J. Ramirez, B. C. Willis, D. Filgueiras-Rama, S. R. Ennis, Y. Takemoto, D. Ponce-Balbuena, M. Zarzoso, R. P. O'Connell, H. Musa, G. Guerrero-Serna, U. M. R. Avula, M. F. Swartz, S. Bhushal, M. Deo, S. V. Pandit, O. Berenfeld, J. Jalife, *Circulation* **2014**, *129*, 1472–1482.

[147] N. Masawa, Y. Yoshida, T. Yamada, T. Joshita, G. Ooneda, *Virchows Archiv A Pathological Anatomy and Histopathology* **1993**, *422*, 67–71.

[148] R. H. Al-Mashhadi, C. B. Sørensen, P. M. Kragh, C. Christoffersen, M. B. Mortensen, L. P. Tolbod, T. Thim, Y. Du, J. Li, Y. Liu, B. Moldt, M. Schmidt, G. Vajta, T. Larsen, S. Purup, L. Bolund, L. B. Nielsen, H. Callesen, E. Falk, J. G. Mikkelsen, J. F. Bentzon, *Science Translational Medicine* **2013**, *5*, DOI 10.1126/scitranslmed.3004853.

[149] R. Matthiesen, J. Bunkenborg, Introduction to mass spectrometry-based proteomics, **2013**.

[150] M. Mayr, S. Yusuf, G. Weir, Y. L. Chung, U. Mayr, X. Yin, C. Ladroue, B. Madhu, N. Roberts, A. De Souza, S. Fredericks, M. Stubbs, J. R. Griffiths, M. Jahangiri, Q. Xu, A. J. Camm, *Journal of the American College of Cardiology* **2008**, *51*, 585–594.

[151] J. P. McRedmond, S. D. Park, D. F. Reilly, J. A. Coppinger, P. B. Maguire, D. C. Shields, D. J. Fitzgerald, *Molecular and Cellular Proteomics* **2004**, *3*, 133–144.

[152] C. Meng, B. Kuster, A. C. Culhane, A. M. Gholami, *BMC Bioinformatics* **2014**, *15*, DOI 10.1186/1471-2105-15-162.

[153] C. Meng, O. A. Zeleznik, G. G. Thallinger, B. Kuster, A. M. Gholami, A. C. Culhane, *Briefings in Bioinformatics* **2016**, *17*, 628–641.

[154] T. Mikawa, R. Hurtado, Development of the cardiac conduction system, **2007**.

[155] J. Modrego, L. Maroto, J. Tamargo, L. Azcona, P. Mateos-Cáceres, A. Segura, R. Moreno-Herrero, N. Pécerescerесrez-Castellanos, E. Delpón, J. Pécerescerесrez-Villacastín, E. Rodríguez, C. MacAya, A. J. López-Farrécerescerес, *Journal of Cardiovascular Electrophysiology* **2010**, *21*, 859–868.

[156] N. Mohammadzadeh, I. G. Lunde, K. Andenæs, M. E. Strand, J. M. Aronsen, B. Skrbic, H. S. Marstein, C. Bandlien, S. Nygård, J. Gorham, I. Sjaastad, S. Chakravarti, G. Christensen, K. V. Engebretsen, T. Tønnessen, *Scientific Reports* **2019**, *9*, 1–13.

[157] C. R. Moreno, F. A. Carvalho, C. Lorenzi, L. S. Matuzaki, S. Prezotti, P. Bighetti, F. M. Louzada, G. Lorenzi-Filho in Chronobiology International, *Vol. 21*, Taylor and Francis Inc., **2004**, pp. 871–879.

[158] J. S. Nanda, R. Kumar, G. P. S. Raghava, J. Singh Nanda, R. Kumar, G. P. S. Raghava, *Scientific Reports* **2016**, *6*, 19340.

[159] S. Nattel, Molecular and Cellular Mechanisms of Atrial Fibrosis in Atrial Fibrillation, **2017**.

[160] S. Nattel, D. Dobrev, The multidimensional role of calcium in atrial fibrillation pathophysiology: Mechanistic insights and therapeutic opportunities, **2012**.

[161] P. Navarro, M. Trevisan-Herraz, E. Bonzon-Kulichenko, E. Núñez, P. Martínez-Acedo, D. Pérez-Hernández, I. Jorge, R. Mesa, E. Calvo, M. Carrascal, M. L. Hernáez, F. García, J. A. Bárcena, K. Ashman, J. Abian, C. Gil, J. M. Redondo, J. Vázquez, *Journal of Proteome Research* **2014**, *13*, 1234–1247.

[162] R. M. Neal in Handbook of Markov Chain Monte Carlo, CRC Press, **2011**, pp. 113–162.

[163] A. I. Nesvizhskii, A. Keller, E. Kolker, R. Aebersold, *Analytical Chemistry* **2003**, *75*, 4646–4658.

[164] T. A. Nevzorova, E. R. Mordakhanova, A. G. Daminova, A. A. Ponomareva, I. A. Andrianova, G. Le Minh, L. Rauova, R. I. Litvinov, J. W. Weisel, *Cell Death Discovery* **2019**, *5*, 106.

[165] L. Nie, G. Wu, F. J. Brockman, W. Zhang, *Bioinformatics* **2006**, *22*, 1641–1647.

[166] L. Nie, G. Wu, W. Zhang, *Biochemical and Biophysical Research Communications* **2006**, *339*, 603–610.

[167] J. B. Nielsen, L. G. Fritsche, W. Zhou, T. M. Teslovich, O. L. Holmen, S. Gustafsson, M. E. Gabrielsen, E. M. Schmidt, R. Beaumont, B. N. Wolford, M. Lin, C. M. Brummett, M. H. Preuss, L. Refsgaard, E. P. Bottinger, S. E. Graham, I. Surakka, Y. Chu, A. H. Skogholt, H. Dalen, A. P. Boyle, H. Oral, T. J. Herron, J. Kitzman, J. Jalife, J. H. Svendsen, M. S. Olesen, I. Njølstad, M.-L. L. Løchen, A. Baras, O. Gottesman, A. Marcketta, C. O'Dushlaine, M. D. Ritchie, T. Wilsgaard, R. J. Loos, T. M. Frayling, M. Boehnke, E. Ingelsson, D. J. Carey, F. E. Dewey, H. M. Kang, G. R. Abecasis, K. Hveem, C. J. Willer, C. O'Dushlaine, M. D. Ritchie, T. Wilsgaard, R. J. Loos, T. M. Frayling, M. Boehnke, E. Ingelsson, D. J. Carey, F. E. Dewey, H. M. Kang, G. R. Abecasis, K. Hveem, C. J. Willer, **2018**, *102*, 103–115.

[168] J. B. Nielsen, R. B. Thorolfsdottir, L. G. Fritsche, W. Zhou, M. W. Skov, S. E. Graham, T. J. Herron, S. McCarthy, E. M. Schmidt, G. Sveinbjornsson, I. Surakka, M. R. Mathis, M. Yamazaki, R. D. Crawford, M. E. Gabrielsen, A. H. Skogholt, O. L. Holmen, M. Lin, B. N. Wolford, R. Dey, H. Dalen, P. Sulem, J. H. Chung, J. D. Backman, D. O. Arnar, U. Thorsteinsdottir, A. Baras, C. O'Dushlaine, A. G. Holst, X. Wen, W. Hornsby, F. E. Dewey, M. Boehnke, S. Kheterpal, B. Mukherjee, S. Lee, H. M. Kang, H. Holm, J. Kitzman, J. A. Shavit, J. Jalife, C. M. Brummett, T. M. Teslovich, D. J. Carey, D. F. Gudbjartsson, K. Stefansson, G. R. Abecasis, K. Hveem, C. J. Willer, Biobank-driven genomic discovery yields new insight into atrial fibrillation biology, **2018**.

[169] K. Nishida, G. Michael, D. Dobrev, S. Nattel, Animal models for atrial fibrillation: Clinical insights and scientific opportunities, **2010**.

[170] H. Ohmura, A. Hiraga, T. Takahashi, M. Kai, J. H. Jones, *Journal of the American Veterinary Medical Association* **2003**, *223*, 84–88.

[171] S. Palta, R. Saroa, A. Palta, Overview of the coagulation system, **2014**.

[172] H. J. Park, S. P. Georgescu, C. Du, C. Madias, M. J. Aronovitz, C. M. Welzig, B. Wang, U. Begley, Y. Zhang, R. O. Blaustein, R. D. Patten, R. H. Karas, H. H. Van Tol, T. F. Osborne, H. Shimano, R. Liao, M. S. Link, J. B. Galper, *Journal of Clinical Investigation* **2008**, *118*, 259–271.

[173] H. J. Park, Y. Zhang, C. Du, C. M. Welzig, C. Madias, M. J. Aronovitz, S. P. Georgescu, I. Naggar, B. Wang, Y. B. Kim, R. O. Blaustein, R. H. Karas, R. Liao, C. E. Mathews, J. B. Galper, *Circulation Research* **2009**, *105*, 287–294.

[174] B. L. Parker, J. G. Burchfield, D. Clayton, T. A. Geddes, R. J. Payne, B. Kiens, J. F. Wojtaszewski, E. A. Richter, D. E. James, *Molecular and Cellular Proteomics* **2017**, *16*, 2055–2068.

[175] V. J. Patel, K. Thalassinos, S. E. Slade, J. B. Connolly, A. Crombie, J. C. Murrell, J. H. Scrivens, *Journal of Proteome Research* **2009**, *8*, 3752–3759.

[176] P. Perco, I. Mühlberger, G. Mayer, R. Oberbauer, A. Lukas, B. Mayer, *Electrophoresis* **2010**, *31*, 1780–1789.

[177] F. Peyvandi, I. Garagiola, L. Baronciani, Role of von Willebrand factor in the haemostasis, **2011**.

[178] J. Pfeuffer, T. Sachsenberg, T. M. Dijkstra, O. Serang, K. Reinert, O. Kohlbacher, *Journal of Proteome Research* **2020**, *19*, 1060–1072.

[179] E. Piruzian, S. Bruskin, A. Ishkin, R. Abdeev, S. Moshkovskii, S. Melnik, Y. Nikolsky, T. Nikolskaya, *BMC Systems Biology* **2010**, *4*, 41.

[180] M. Prondzynski, M. D. Lemoine, A. T. Zech, A. Horváth, V. Di Mauro, J. T. Koivumäki, N. Kresin, J. Busch, T. Krause, E. Krämer, S. Schlossarek, M. Spohn, F. W. Friedrich, J. Münch, S. D. Laufer, C. Redwood, A. E. Volk, A. Hansen, G. Mearini, D. Catalucci, C. Meyer, T. Christ, M. Patten, T. Eschenhagen, L. Carrier, *EMBO Molecular Medicine* **2019**, *11*, e11115.

[181] X. Y. Qi, Y. H. Yeh, L. Xiao, B. Burstein, A. Maguy, D. Chartier, L. R. Villeneuve, B. J. Brundel, D. Dobrev, S. Nattel, *Circulation Research* **2008**, *103*, 845–854.

[182] A. Ramos-Fernández, D. López-Ferrer, J. Vázquez, *Molecular and Cellular Proteomics* **2007**, *6*, 1274–1286.

[183] P. Rautaharju, S. H. Zhou, S. Wong, H. P. Calhoun, G. S. Berenson, R. Prineas, A. Davignon, *Canadian Journal of Cardiology* **1992**, *8*, 690–695.

[184] D. E. Reich, M. Cargili, S. Boik, J. Ireland, P. C. Sabeti, D. J. Richter, T. Lavery, R. Kouyoumjian, S. F. Farhadian, R. Ward, E. S. Lander, *Nature* **2001**, *411*, 199–204.

[185] L. Revilla, A. Mayorgas, A. M. Corraliza, M. C. Masamunt, A. Metwaly, D. Haller, E. Tristan, A. Carrasco, M. Esteve, J. Panes, E. Ricart, J. J. Lozano, A. Salas, *PLoS ONE* **2021**, *16*, DOI 10.1371/journal.pone.0246367.

[186] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, G. K. Smyth, *Nucleic Acids Research* **2015**, *43*, e47.

[187] M. D. Robinson, A. Oshlack, *Genome Biology* **2010**, *11*, 1–9.

[188] D. Rodriguez-Terrones, M. E. Torres-Padilla, Nimble and Ready to Mingle: Transposon Outbursts of Early Development, **2018**.

[189] S. Rogers, M. Girolami, W. Kolch, K. M. Waters, T. Liu, B. Thrall, H. S. Wiley, S. H. Wiley, *Bioinformatics* **2008**, *24*, 2894–900.

[190] M. Rosa-Garrido, D. J. Chapski, T. M. Vondriska, Epigenomes in Cardiovascular Disease, **2018**.

[191] C. Roselli, M. D. Chaffin, L. C. Weng, S. Aeschbacher, G. Ahlberg, C. M. Albert, P. Almgren, A. Alonso, C. D. Anderson, K. G. Aragam, D. E. Arking, J. Barnard, T. M. Bartz, E. J. Benjamin, N. A. Bihlmeyer, J. C. Bis, H. L. Bloom, E. Boerwinkle, E. B. Bottinger, J. A. Brody, H. Calkins, A. Campbell, T. P. Cappola, J. Carlquist, D. I. Chasman, L. Y. Chen, Y. D. I. Chen, E. K. Choi, S. H. Choi, I. E. Christophersen, M. K. Chung, J. W. Cole, D. Conen, J. Cook, H. J. Crijns, M. J. Cutler, S. M. Damrauer, B. R. Daniels, D. Darbar, G. Delgado, J. C. Denny, M. Dichgans, M. Dörr, E. A. Dudink, S. C. Dudley, N. Esa, T. Esko, M. Eskola, D. Fatkin, S. B. Felix, I. Ford, O. H. Franco, B. Geelhoed, R. P. Grewal, V. Gudnason, X. Guo, N. Gupta, S. Gustafsson, R. Gutmann, A. Hamsten, T. B. Harris, C. Hayward, S. R. Heckbert, J. Hernesniemi, L. J. Hocking, A. Hofman, A. R. Horimoto, J. Huang, P. L. Huang, J. Huffman, E. Ingelsson, E. G. Ipek, K. Ito, J. Jimenez-Conde, R. Johnson, J. W. Jukema, S. Kääb, M. Kähönen, Y. Kamatani, J. P. Kane, A. Kastrati, S. Kathiresan, P. Katschnig-Winter, M. Kavousi, T. Kessler, B. L. Kietselaer, P. Kirchhof, M. E. Kleber, S. Knight, J. E. Krieger, M. Kubo, L. J. Launer, J. Laurikka, T. Lehtimäki, K. Leineweber, R. N. Lemaitre, M. Li, H. E. Lim, H. J. Lin, H. Lin, L. Lind, C. M. Lindgren, M. L. Lokki, B. London, R. J. Loos, S. K. Low, Y. Lu, L. P. Lyytikäinen, P. W. Macfarlane, P. K. Magnusson, A. Mahajan, R. Malik, A. J. Mansur, G. M. Marcus, L. Margolin,

K. B. Margulies, W. März, D. D. McManus, O. Melander, S. Mohanty, J. A. Montgomery, M. P. Morley, A. P. Morris, M. Müller-Nurasyid, A. Natale, S. Nazarian, B. Neumann, C. Newton-Cheh, M. N. Niemeijer, K. Nikus, P. Nilsson, R. Noordam, H. Oellers, M. S. Olesen, M. Orho-Melander, S. Padmanabhan, H. N. Pak, G. Paré, N. L. Pedersen, J. Pera, A. Pereira, D. Porteous, B. M. Psaty, S. L. Pulit, C. R. Pullinger, D. J. Rader, L. Refsgaard, M. Ribasés, P. M. Ridker, M. Rienstra, L. Risch, D. M. Roden, J. Rosand, M. A. Rosenberg, N. Rost, J. I. Rotter, S. Saba, R. K. Sandhu, R. B. Schnabel, K. Schramm, H. Schunkert, C. Schurman, S. A. Scott, I. Seppälä, C. Shaffer, S. Shah, A. A. Shalaby, J. Shim, M. B. Shoemaker, J. E. Siland, J. Sinisalo, M. F. Sinner, A. Slowik, A. V. Smith, B. H. Smith, J. G. Smith, J. D. Smith, N. L. Smith, E. Z. Soliman, N. Sotoodehnia, B. H. Stricker, A. Sun, H. Sun, J. H. Svendsen, T. Tanaka, K. Tanriverdi, K. D. Taylor, M. Teder-Laving, A. Teumer, S. Thériault, S. Trompet, N. R. Tucker, A. Tveit, A. G. Uitterlinden, P. Van Der Harst, I. C. Van Gelder, D. R. Van Wagoner, N. Verweij, E. Vlachopoulou, U. Völker, B. Wang, P. E. Weeke, B. Weijs, R. Weiss, S. Weiss, Q. S. Wells, K. L. Wiggins, J. A. Wong, D. Woo, B. B. Worrall, P. S. Yang, J. Yao, Z. T. Yoneda, T. Zeller, L. Zeng, S. A. Lubitz, K. L. Lunetta, P. T. Ellinor, *Nature Genetics* **2018**, *50*, 1225–1233.

[192] H. L. Röst, T. Sachsenberg, S. Aiche, C. Bielow, H. Weisser, F. Aicheler, S. Andreotti, H. C. Ehrlich, P. Gutenbrunner, E. Kenar, X. Liang, S. Nahnsen, L. Nilse, J. Pfeuffer, G. Rosenberger, M. Rurik, U. Schmitt, J. Veit, M. Walzer, D. Wojnar, W. E. Wolski, O. Schilling, J. S. Choudhary, L. Malmström, R. Aebersold, K. Reinert, O. Kohlbacher, OpenMS: A flexible open-source software platform for mass spectrometry data analysis, **2016**.

[193] G. Rozen, S. M. Hosseini, M. I. Kaadan, Y. Biton, E. K. Heist, M. Vangel, M. C. Mansour, J. N. Ruskin, *Journal of the American Heart Association* **2018**, *7*, DOI 10.1161/JAHA.118.009024.

[194] S. T. Russell, M. J. Tisdale, *Endocrinology* **2012**, *153*, 4696–4704.

[195] L. Sabatino, C. Kusmic, G. Iervasi, *Molecular and Cellular Biochemistry* **2020**, *475*, 205–214.

[196] M. Salmi, S. Jalkanen, Vascular adhesion protein-1: A cell surface amine oxidase in translation, **2019**.

[197] D. Sanchez-Quintana, J. Ramon Lopez-Mínguez, G. Pizarro, M. Murillo, J. Angel Cabrera, *Current Cardiology Reviews* **2012**, *8*, 310–326.

[198] L. E. Sander, S. D. Sackett, U. Dierssen, N. Beraza, R. P. Linke, M. Müller, J. M. Blander, F. Tacke, C. Trautwein, *Journal of Experimental Medicine* **2010**, *207*, 1453–1464.

[199] M. Sardana, D. Lessard, C. W. Tsao, N. I. Parikh, B. A. Barton, G. Nah, R. C. Thomas, S. Cheng, N. B. Schiller, J. R. Aragam, G. F. Mitchell, A. Vaze, E. J. Benjamin, R. S. Vasan, D. D. McManus, *Journal of the American Heart Association* **2018**, *7*, DOI 10.1161/JAHA.117.008435.

[200] R. B. Schnabel, X. Yin, P. Gona, M. G. Larson, A. S. Beiser, D. D. McManus, C. Newton-Cheh, S. A. Lubitz, J. W. Magnani, P. T. Ellinor, S. Seshadri, P. A. Wolf, R. S. Vasan, E. J. Benjamin, D. Levy, *The Lancet* **2015**, *386*, 154–162.

[201] U. Schotten, S. Verheule, P. Kirchhof, A. Goette, Pathophysiological mechanisms of atrial fibrillation: A translational appraisal, **2011**.

[202] R. B. Schuessler, T. M. Grayson, B. I. Bromberg, J. L. Cox, J. P. Boineau, *Circulation Research* **1992**, *71*, 1254–1267.

[203] B. Schwanhüusser, D. Busse, N. Li, G. Dittmar, J. Schuchhardt, J. Wolf, W. Chen, M. Selbach, *Nature* **2011**, *473*, 337–342.

[204] O. Serang, M. J. MacCoss, W. S. Noble, *Journal of Proteome Research* **2010**, *9*, 5346–5357.

[205] O. Serang, W. S. Noble, *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **2012**, *9*, 809–817.

[206] A. J. Shah, M. Hocini, Y. Komatsu, M. Daly, S. Zellerhoff, L. Jesel, S. Amaroui, K. Ramoul, A. Denis, N. Derval, F. Sacher, P. Jais, M. Haissaguerre, The progressive nature of atrial fibrillation: A rationale for early restoration and maintenance of sinus rhythm, **2013**.

[207] W. Shao, P. J. Espenshade, Expanding roles for SREBP in metabolism, **2012**.

[208] O. F. Sharifov, V. V. Fedorov, G. G. Beloshapko, A. V. Glukhov, A. V. Yushmanova, L. V. Rosenshtraukh, *Journal of the American College of Cardiology* **2004**, *43*, 483–490.

[209] M. J. Shen, R. Arora, J. Jalife, Atrial Myopathy, **2019**.

[210] J. Shirani, J. Alaeddini, *Cardiovascular Pathology* **2000**, *9*, 95–101.

[211] M. I. Sigurdsson, L. Saddic, M. Heydarpour, T. W. Chang, P. Shekar, S. Aranki, G. S. Couper, S. K. Shernan, J. D. Muehlschlegel, S. C. Body, *BMC Medical Genomics* **2017**, *10*, DOI 10.1186/s12920-017-0270-5.

[212] A. Simats, L. Ramiro, T. García-Berrocoso, F. Briansó, R. Gonzalo, L. Martín, A. Sabé, N. Gill, A. Penalba, N. Colomé, A. Sánchez, F. Canals, A. Bustamante, A. Rosell, J. Montaner, *Molecular and Cellular Proteomics* **2020**, *19*, 1921–1935.

[213] M. F. Sinner, N. R. Tucker, K. L. Lunetta, K. Ozaki, J. G. D. Smith, S. Trompet, J. C. Bis, H. Lin, M. K. Chung, J. B. Nielsen, S. A. Lubitz, B. P. Krijthe, J. W. Magnani, J. Ye, M. H. Gollob, T. Tsunoda, M. Müller-Nurasyid, P. Lichtner, A. Peters, E. Dolmatova, M. Kubo, J. G. D. Smith, B. M. Psaty, N. L. Smith, J. W. Jukema, D. I. Chasman, C. M. Albert, Y. Ebana, T. Furukawa, P. W. Macfarlane, T. B. Harris, D. Darbar, M. Dörr, A. G. Holst, J. H. Svendsen, A. Hofman, A. G. Uitterlinden, V. Gudnason, M. Isobe, R. Malik, M. Dichgans, J. Rosand, D. R. Van Wagoner, E. J. Benjamin, D. J. Milan, O. Melander, S. R. Heckbert, I. Ford, Y. Liu, J. Barnard, M. S. Olesen, B. H. Stricker, T. Tanaka, S. Kääb, P. T. Ellinor, *Circulation* **2014**, *130*, 1225–1235.

[214] E. I. Skalidis, M. I. Hamilos, I. K. Karalis, G. Chlouverakis, G. E. Kochiadakis, P. E. Vardas, *Journal of the American College of Cardiology* **2008**, *51*, 2053–2057.

[215] S. Song, F. B. Johnson, Epigenetic mechanisms impacting aging: A focus on histone levels and telomeres, **2018**.

[216] R. Stark, M. Grzelak, J. Hadfield, RNA sequencing: the teenage years, **2019**.

[217] M. Steenman, Insight into atrial fibrillation through analysis of the coding transcriptome in humans, **2020**.

[218] M. Sühling, C. Wolke, C. Scharf, U. Lendeckel, Proteomik und Transkriptomik bei Vorhofflimmern, **2018**.

[219] D. L. Tabb, The SEQUEST Family Tree, **2015**.

[220] Y. Takemoto, R. J. Ramirez, K. Kaur, O. Salvador-Montañés, D. Ponce-Balbuena, R. Ramos-Mondragón, S. R. Ennis, G. Guerrero-Serna, O. Berenfeld, J. Jalife, *Journal of the American College of Cardiology* **2017**, *70*, 2893–2905.

[221] Y. Takemoto, R. J. Ramirez, M. Yokokawa, K. Kaur, D. Ponce-Balbuena, M. C. Sinno, B. C. Willis, H. Ghanbari, S. R. Ennis, G. Guerrero-Serna, B. C. Henzi, R. Latchamsetty, R. Ramos-Mondragon, H. Musa, R. P. Martins, S. V. Pandit, S. F. Noujaim, T. Crawford, K. Jongnarangsin, F. Pelosi, F. Bogun, A. Chugh, O. Berenfeld, F. Morady, H. Oral, J. Jalife, Galectin-3 Regulates Atrial Fibrillation Remodeling and Predicts Catheter Ablation Outcomes, **2016**.

[222] N. Tan, M. K. Chung, J. D. Smith, J. Hsu, D. Serre, D. W. Newton, L. Castel, E. Soltesz, G. Pettersson, A. M. Gillinov, D. R. Van Wagoner, J. Barnard, *Circulation: Cardiovascular Genetics* **2013**, *6*, 362–371.

[223] T. Tanaka, M. Narazaki, T. Kishimoto, *Cold Spring Harbor Perspectives in Biology* **2014**, *6*, 16295–16296.

[224] H. Tao, J. J. Yang, Z. W. Chen, S. S. Xu, X. Zhou, H. Y. Zhan, K. H. Shi, *Toxicology* **2014**, *323*, 42–50.

[225] T. Tebaldi, A. Re, G. Viero, I. Pegoretti, A. Passerini, E. Blanzieri, A. Quattrone, *BMC Genomics* **2012**, *13*, 220.

[226] M. Teng, M. I. Love, C. A. Davis, S. Djebali, A. Dobin, B. R. Graveley, S. Li, C. E. Mason, S. Olson, D. Pervouchine, C. A. Sloan, X. Wei, L. Zhan, R. A. Irizarry, *Genome Biology* **2016**, *17*, 74.

[227] M. The, M. J. MacCoss, W. S. Noble, L. Käll, *Journal of the American Society for Mass Spectrometry* **2016**, *27*, 1719–1727.

[228] A. M. Thomas, C. P. Cabrera, M. Finlay, K. Lall, M. Nobles, R. J. Schilling, K. Wood, C. A. Mein, M. R. Barnes, P. B. Munroe, A. Tinker, *Physiological Genomics* **2019**, *51*, 323–332.

[229] A. Thomas, W. Schänzer, M. Thevis, Immunoaffinity techniques coupled to mass spectrometry for the analysis of human peptide hormones: advances and applications, **2017**.

[230] R. B. Thorolfsdottir, G. Sveinbjornsson, P. Sulem, A. Helgadottir, S. Gretarsdottir, S. Benonisdottir, A. Magnusdottir, O. B. Davidsson, S. Rajamani, D. M. Roden, D. Darbar, T. R. Pedersen, M. S. Sabatine, I. Jonsdottir, D. O. Arnar, U. Thorsteinsdottir, D. F. Gudbjartsson, H. Holm, K. Stefansson, *Journal of the American College of Cardiology* **2017**, *70*, 2157–2168.

[231] R. B. Thorolfsdottir, G. Sveinbjornsson, P. Sulem, J. B. Nielsen, S. Jonsson, G. H. Halldorsson, P. Melsted, E. V. Ivarsdottir, O. B. Davidsson, R. P. Kristjansson, G. Thorleifsson, A. Helgadottir, S. Gretarsdottir, G. Norddahl, S. Rajamani, B. Torfason, A. S. Valgardsson, J. T. Sverrisson, V. Tragante, O. L. Holmen, F. W. Asselbergs, D. M. Roden, D. Darbar, T. R. Pedersen, M. S. Sabatine, C. J. Willer, M. L. Løchen, B. V. Halldorsson, I. Jonsdottir, K. Hveem, D. O. Arnar, U. Thorsteinsdottir, D. F. Gudbjartsson, H. Holm, K. Stefansson, *Communications Biology* **2018**, *1*, 1–9.

[232] C. Tian, Y. J. Kim, S. Hali, O. S. Choo, J. S. Lee, S. K. Jung, Y. U. Choi, C. B. Park, Y. H. Choung, *Cell Death and Disease* **2020**, *11*, 1–13.

[233] K. Toutouzas, M. Drakopoulou, P. Dilaveris, S. Vaina, K. Gatzoulis, J. Karabelas, M. Riga, E. Stefanadi, A. Synetos, K. Vlasis, C. Stefanadis, *International Journal of Cardiology* **2009**, *134*, 345–350.

[234] C. Trapnell, L. Pachter, S. L. Salzberg, *Bioinformatics* **2009**, *25*, 1105–1111.

[235] C. T. Tsai, C. S. Hsieh, S. N. Chang, E. Y. Chuang, K. C. Ueng, C. F. Tsai, T. H. Lin, C. K. Wu, J. K. Lee, L. Y. Lin, Y. C. Wang, C. C. Yu, L. P. Lai, C. D. Tseng, J. J. Hwang, F. T. Chiang, J. L. Lin, *Nature Communications* **2016**, *7*, DOI 10.1038/ncomms10190.

[236] F. C. Tsai, Y. C. Y. M. Lin, S. H. Chang, G. J. Chang, Y. J. Hsu, Y. C. Y. M. Lin, Y. S. Lee, C. L. Wang, Y. H. Yeh, *International Journal of Cardiology* **2016**, *222*, 104–112.

[237] C. Tu, Q. Sheng, J. Li, D. Ma, X. Shen, X. Wang, Y. Shyr, Z. Yi, J. Qu, *Journal of Proteome Research* **2015**, *14*, 4662–4673.

[238] T. Välikangas, T. Suomi, L. L. Elo, *Briefings in Bioinformatics* **2018**, *19*, 1–11.

[239] A. Vehtari, J. Lampinen, *Neural Computation* **2002**, *14*, 2439–2468.

[240] R. L. K. Virchow, *Canton Mass: Science History Publications/USA* **1846**.

[241] C. Vogel, E. M. Marcotte, Absolute abundance for the masses, **2009**.

[242] C. Vogel, E. M. Marcotte, *Nature Reviews Genetics* **2012**, *13*, 227–232.

[243] J. D. Wall, J. K. Pritchard, Haplotype blocks and linkage disequilibrium in the human genome, **2003**.

[244] A. M. Walsh, R. D. Kortschak, M. G. Gardner, T. Bertozzi, D. L. Adelson, *Proceedings of the National Academy of Sciences of the United States of America* **2013**, *110*, 1012–1016.

[245] D. Wang, B. Eraslan, T. Wieland, B. Hallström, T. Hopf, D. P. Zolg, J. Zecha, A. Asplund, L.-h. Li, C. Meng, M. Frejno, T. Schmidt, K. Schnatbaum, M. Wilhelm, F. Ponten, M. Uhlen, J. Gagneur, H. Hahne, B. Kuster, *Molecular Systems Biology* **2019**, *15*, e8503.

[246] Y. Wang, T. Joshi, X. S. Zhang, D. Xu, L. Chen, *Bioinformatics* **2006**, *22*, 2413–2420.

[247] L. C. Weng, S. H. Choi, D. Klarin, J. G. Smith, P. R. Loh, M. Chaffin, C. Roselli, O. L. Hulme, K. L. Lunetta, J. Dupuis, E. J. Benjamin, C. Newton-Cheh, S. Kathiresan, P. T. Ellinor, S. A. Lubitz, *Circulation: Cardiovascular Genetics* **2017**, *10*, DOI 10.1161/CIRCGENETICS.117.001838.

[248] L. C. Weng, S. R. Preis, O. L. Hulme, M. G. Larson, S. H. Choi, B. Wang, L. Trinquart, D. D. McManus, L. Staerk, H. Lin, K. L. Lunetta, P. T. Ellinor, E. J. Benjamin, S. A. Lubitz, *Circulation* **2018**, *137*, 1027–1038.

[249] A. V. Werhli, M. Grzegorczyk, D. Husmeier, *Bioinformatics* **2006**, *22*, 2523–2531.

[250] C. W. White, R. E. Kerber, H. R. Weiss, M. L. Marcus, *Circulation Research* **1982**, *51*, 205–215.

[251] M. C. Wijffels, C. J. Kirchhof, R. Dorland, M. A. Allessie, *Circulation* **1995**, *92*, 1954–1968.

[252] J. G. Wood, S. L. Helfand, Chromatin structure and transposable elements in organismal aging, **2013**.

[253] M. Yamamoto, Y. Seo, N. Kawamatsu, K. Sato, A. Sugano, T. Machino-Ohtsuka, R. Kawamura, H. Nakajima, M. Igarashi, Y. Sekiguchi, T. Ishizu, K. Aonuma, *Circulation: Cardiovascular Imaging* **2014**, *7*, 337–343.

[254] Y. H. Yeh, C. T. Kuo, Y. S. Lee, Y. M. Lin, S. Nattel, F. C. Tsai, W. J. Chen, *Heart Rhythm* **2013**, *10*, 383–391.

[255] L. Yue, P. Melnyk, R. Gaspo, Z. Wang, S. Nattel, *Circulation Research* **1999**, *84*, 776–784.

[256] J. Zecha, S. Satpathy, T. Kanashova, S. C. Avanessian, M. H. Kane, K. R. Clauser, P. Mertins, S. A. Carr, B. Kuster, *Molecular and Cellular Proteomics* **2019**, *18*, 1468–1478.

[257] D. Zhang, X. Hu, J. Li, F. Hoogstra-Berends, Q. Zhuang, M. A. Esteban, N. de Groot, R. H. Henning, B. J. Brundel, *Journal of Molecular and Cellular Cardiology* **2018**, *125*, 39–49.

[258] P. Zhang, W. Wang, X. Wang, X. Wang, Y. Song, Y. Han, J. Zhang, H. Zhao, *PLoS ONE* **2013**, *8*, e60210.

[259] Y. Zhang, E. I. Dedkov, D. Teplitsky, N. Y. Weltman, C. J. Pol, V. Rajagopalan, B. Lee, A. Martin Gerdes, *Circulation: Arrhythmia and Electrophysiology* **2013**, *6*, 952–959.

[260] J. Zhou, J. Gao, Y. Liu, S. Gu, X. Zhang, X. An, J. Yan, Y. Xin, P. Su, *International Heart Journal* **2014**, *55*, 71–77.

[261] X. Zhou, S. C. Dudley, *Frontiers in Cardiovascular Medicine* **2020**, *7*, DOI 10.3389/fcvm.2020.00062.

[262] R. Zou, M. Yang, W. Shi, C. Zheng, H. Zeng, X. Lin, D. Zhang, S. Yang, P. Hua, *Cellular Physiology and Biochemistry* **2018**, *47*, 1299–1309.

Chapter 9

Publications

The work described in this thesis is included in the following publications:

- **Alvarez-Franco A**, Martí-Gómez C, Ramirez RJ, Guerrero-Serna G, Rouco R, Magni R, Jalife J, Manzanares M. Modelling the serum proteome: from the non-tachypacing to permanent-AF. To submit.

- Victorino J*, **Alvarez-Franco A***, Jalife J, Manzanares M. Functional genomics and epigenomics of atrial fibrillation. J Mol Cell Cardiol. 2021 Apr 19;157:45-55. doi: 10.1016/j.yjmcc.2021.04.003. Epub ahead of print. PMID: 33887329. (*co-first authors).

- **Alvarez-Franco A***, Rouco R*, Ramirez RJ, Guerrero-Serna G, Tiana M, Cogliati S, Kaur K, Saeed M, Magni R, Enriquez JA, Sanchez-Cabo F, Jalife J, Manzanares M. Transcriptome and proteome mapping in the sheep atria reveal molecular features of atrial fibrillation progression. Cardiovasc Res. 2020 Oct 29:cvaa307. doi: 10.1093/cvr/cvaa307. Epub ahead of print. PMID: 33119050. (*co-first authors)

The collaboration in other research projects during the development of the thesis has resulted in the following publications:

- Andreu M, **Alvarez-Franco A***, Portela M*, Gimenez-Llorente D, Cuadrado A, Badia-Careaga C, Tiana M, Losada A, Manzanares M. Correct establishment of 3D chromatin structure following fertilization and the metabolic switch at the morula-to-blastocyst transition require CTCF, (*these authors contributed equally to this work). To submit.

- Sainz de Aja J, Menchero S, Rollan I, Barral A, Tiana M, Jawaid W, Cossio I, **Alvarez A**, Carreño-Tarragona G, Badia-Careaga C, Nichols J, Göttgens B, Isern J, Manzanares M. The pluripotency factor NANOG controls primitive hematopoiesis and directly regulates Tal1. EMBO J. 2019 Apr 1;38(7):e99122. doi: 10.15252/embj.201899122. Epub 2019 Feb 27. PMID: 30814124; PMCID: PMC6443201.

- Gomez-Velazquez M, Badia-Careaga C, Lechuga-Vieco AV, Nieto-Arellano R, Tena JJ, Rollan I, **Alvarez A**, Torroja C, Caceres EF, Roy AR, Galjart N, Delgado-Olguin P, Sanchez-Cabo F, Enriquez JA, Gomez-Skarmeta JL, Manzanares M. CTCF counter-regulates cardiomyocyte development and maturation programs in the embryonic heart. PLoS Genet. 2017 Aug 28;13(8):e1006985. doi: 10.1371/journal.pgen.1006985. PMID: 28846746; PMCID: PMC5591014.