Université de Montréal

Multiomic strategies for the discovery of molecular determinants of atrial fibrillation

Par

Francis J.A. Leblanc

Département de biochimie et médecine moléculaire

Faculté de médecine

Thèse présentée en vue de l'obtention du grade de

doctorat en bioinformatique

13 septembre 2023

Université de Montréal

Institut de cardiologie de Montréal, Faculté de médecine

Cette thèse intitulée

Multiomic strategies for the discovery of molecular determinants of atrial fibrillation

Présenté par

Francis J.A. Leblanc

A été évaluée par un jury composé des personnes suivantes

Sylvie Hamel
Président-rapporteur

Guillaume Lettre
Directeur de recherche

Sébastien Lemieux
Codirecteur

Yoshiaki Tanaka
Membre du jury

Sebastien Theriault
Examinateur externe

Résumé

La fibrillation auriculaire (FA) est l'arythmie cardiaque la plus répandue dans le monde et est associée à une hausse de morbidité et une mortalité importante. Des progrès substantiels dans notre compréhension de l'étiologie de la maladie ont été réalisés au cours des deux dernières décennies, conduisant à une amélioration du traitement et de la gestion de la maladie. Cependant, le fardeau de la FA continue d'augmenter. De plus, les mécanismes moléculaires et cellulaires sous-jacents à l'initiation et à la progression de la FA restent incomplètement compris.

Dans cette thèse, mon objectif était de caractériser de nouveaux déterminants moléculaires et cellulaires de la FA en utilisant une approche multiomique. J'ai d'abord utilisé le séquençage de l'ARN (RNAseq) pour l'ARN total et les micro-ARN (miRNA) afin de dévaluer l'effet de la FA sur l'expression génique dans deux modèles canins de FA. Ces résultats ont impliqué le locus orthologue humain 14q32 et son lien potentiel avec la signalisation du glutamate. Dans le chapitre trois, j'ai démontré les lacunes actuelles des modèles statistiques utilisés pour prédire l'effet régulateur des régions de chromatine ouvertes sur l'expression des gènes dans des essais multiomiques à noyau unique et suggéré des alternatives montrant un meilleur pouvoir prédictif. Dans le chapitre quatre, j'ai utilisé des analyses par locus quantitatifs d'expression (eQTL) pour caractériser les variants génétiques communs associés à la FA. Grâce à des analyses de colocalisation, une cartographie fine et un multiome à noyau unique, j'ai justifié mécaniquement l'effet de variants non-codants et fait la priorisation de gènes candidats, notamment *GNB4*, *MAPT* et *LINC01629*. Enfin, dans le chapitre cinq, j'ai fourni une caractérisation approfondie des gènes persistants de FA différentiellement exprimés au niveau cellulaire et identifié les facteurs de transcription potentiels impliqués dans leur régulation.

En résumé, l'utilisation d'une approche multiomique a permis de découvrir de nombreuses nouvelles voies cellulaires et génétiques modifiées au cours de la FA ainsi que des gènes candidats impliqués dans le risque génétique de la FA. Ces résultats fournissent des informations et ressources importantes pour concevoir de nouvelles stratégies thérapeutiques, impliquant à la fois des cibles génétiques et nouvelles voies cellulaires pour lutter contre cette maladie cardiaque commune.

Mots-clés: Fibrillation auriculaire, bioinformatique, multiomique, génétique humaine, génomique, RNAseq, ATACseq, technologie *single cell*.

Abstract

Atrial fibrillation (AF) is the most prevalent cardiac arrhythmia worldwide and is associated with important morbidity and mortality. Substantial advancement in our understanding of the disease etiology have been made in the past two decades leading to improved treatment and management of the disease, however, AF burden continues to increase. Moreover, the molecular and cellular mechanisms underlying AF initiation and progression remain incompletely understood.

In this thesis I aimed to characterise novel molecular and cellular determinants of AF using a multiomic approache. In chapter two, I used RNA sequencing (RNAseq) for total RNA and micro-RNAs (miRNA) to decipher the effect of AF in two canine models, which implicated the orthologue human locus 14q32 and its potential role in glutamate signaling regulation. In chapter three, I demonstrated current shortcomings of statistical models used to predict the regulatory effect of open chromatin regions on gene expression in single nuclei multiomic assays and suggested alternatives showing better predictive power. In chapter four, I used expression quantitative loci (eQTL) to characterize AF associated common genetic variants. Through co-localization analyses, fine-mapping and single nuclei multiome, I mechanistically substantiated non-coding variants and prioritized strong candidate genes including *GNB4*, *MAPT* and *LINC01629*. Finally, in chapter five, I provided a deep characterization of persistent AF differentially expressed genes (DEGs) at the cellular level and identified potential transcription factors involved in their regulation.

In summary, using a multiomic approach unraveled numerous new cellular and gene pathways altered during AF and candidate genes implicated in AF genetic risk. These findings provide important insights and data resources to design novel therapeutic strategies, targeting both genetically derived candidate genes and cellular pathways to address this pervasive cardiac disease.

Keywords: Atrial fibrillation, bioinformatics, multiomics, genomics, human genetics, RNAseq, ATACseq, single cell technology.

Table of content

List of tables

List of figures

Chapter 2

Chapter 3

Chapter 4

Chapter 5

List of abbreviations and acronyms

AAD: Anti-arrhythmic drugs

ACh: Acetylcholine

AF : Atrial Fibrillation

AHA: American Heart Association

AMPA: α-amino-3-hydroxy-5 methylisoxazole-4-propionate

APD: Action potential duration

ATACseq: Assay for transposase-accessible chromatin with sequencing

AVN: Atrioventricular node

bp: Base pairs

caQTL: Chromatin accessibility quantitative trait loci

CBP: p300–CREB-binding protein

CCB: Calcium channel blockers

cDNA: Coding DNA

ChiP-seq: Chromatin immunoprecipitation

CM: Cardiomyocytes

CRIPSRa: CRISPR activator

CRIPSRi: CRISPR inhibotor

CRISPR: Clustered regularly interspaced short palindromic repeats

CVD: Cardiovascular diseases

DAD: Delayed afterdepolarizations

dCas9: Dead Cas9

DEG: Differentially expressed genes

DGE: Differential gene expression

dTTPs: Deoxythymidine triphosphate

dUTPs: Deoxyuridine triphosphate

EAD: Early afterdepolarizations

ECG: Electrocardiogram

ECM: Extracellular matrix

eGenes: eQTL gene

ENCODE: Encyclopedia of DNA elements project

eQTL: Expression quantitative loci

ESC: European society of cardiology

FPKM: Fragments per kilobase of transcript per million mapped reads

gDNA: Genomic DNA

GO: Gene Onthology

GSEA: Gene set enrichment analysis

GTEx: Genotype-tissue expression

GWAS: Genome wide association studies

HDAC: Histone deacetylase

Hi-C: High-throughput chromosome conformation capture

iGluR: Ionotropic glutamate receptors

iPSC-CM: Induced pluripotent stem cells derived cardiomyocytes

KEGG: Kyoto Encyclopedia of Genes and Genomes

KO: Knock out

LA: Left atrium

LAA: Left atrium appendage

LD: Linkage disequilibrium

lncRNA: long non-coding RNA

MAF: Minor allele frequency

mGluR: Metabotropic glutamate receptors

miRNA: Micro-RNAs

mQTL: Methylation quantitative trait loci

mRNA: Messenger RNA

NE: Norepinephrine

NGS: Next generation sequencing

NMDA: N-methyl-D-aspartate

oligo: Oligonucleotides

ORA: Over-representation analysis

PBMCs: Peripheral blood mononuclear cells

PCA: Principal component analysis

PCHi-C: Promoter capture Hi-C

PIP: Posterior inclusion probability

polyA: Polyadenylated

pQTL: Protein quantitative trait loci

PRS: Polygenic risk score

PV: Pulmonary veins

QC: Quality control

qPCR: Quantitative polymerase chain reaction

RA: Right atrium

RAA: Right atrial appendage

Ribo-seq: Ribosome-sequencing

RNAseq: RNA sequencing

rRNA: Ribosomal RNA

scRNAseq: Single cell RNAseq

sgRNA: Single guide RNAs

SN: Sinus node

snATACseq: Single nuclei ATACseq

SNP: Single nucleotide polymorphisms

snRNAseq: Single nuclei RNAseq

SR: Sinus rhythm

TF: Transcription factor

t-SNE: T-distributed stochastic neighbor embedding

TSS: Transcription starting site

UMAP: Uniform manifold approximation and projection

UMI: Unique molecular identifiers

To my familiy

Acknowledgements

First and foremost, I express my deep gratitude to my supervisor, Dr Guillaume Lettre, for being so generous of his time, for sharing his contagious love of science and for his mentorship. I will continue to be inspired by his skillful balance of rigor and levity.

I thank my co-supervisor, Dr Sebastien Lemieux, for being a wonderful and insightful critic of bioinformatics and for his support.

I am also grateful to Melissa Beaudoin for her passionate mentorship in the wet-lab and to Ken Sin Lo for his generous bioinformatic troubleshooting insights.

I wish to thank Dr Stanley Nattel, Dr Svetlana Relly and Dr Patrice Naud for their numerous collaborations and contributions to my work.

I also wish to express my appreciation to my thesis committee members, Dr Eric Thorin and Dr Julie Hussin for their astute comments and guidance during our meetings.

I express my gratitude to the members of my jury Dr Sylvie Hamel, Dr Yoshiaki Tanaka and Dr Sebastien Theriault for their thoughtful review and comments.

I am thankful for the scholarships offered from the Fonds de Recherche en Santé du Québec (FRQS), Montreal Heart Institute Foundation (MHIF), Canadian Institutes of Health Research (CIHR) and Université de Montréal.

I am also thankful for the continuous support from my family and friends. I wish to address special gratitude to Simon Del Testa for his clinical perspective.

To my mother, Manon Gauthier, thank you for all your support and empathy. Your courage, energy, and sincerity has fueled my resolve and given me more to aspire to.

Above all, I thank my wife, ma complice, editor and best friend Lauren Montpetit for her unwavering patience, support, comfort, and encouragement. This journey would not have been one I wish to undertake without you. I will forever be grateful for the sacrifices you made during these years.

Chapter 1: Introduction

Atrial fibrillation (AF) is the most common cardiac arrhythmia, affecting one in three to one in five individuals in their lifetime[1]. Its prevalence is expected to double by 2050[1]. It increases the risk of death by about 2-fold, but its toll is perhaps more significantly felt by its impairment on quality of life, affecting more than 60% of AF patients[1]. Tremendous progress in treatment and management have been achieved in the last two decades, but halting the progression of AF remains a challenge. Furthermore, current options often are invasive, poorly tolerated or not indicated, which compresses the number of healthy years in late life.

While the disease's emergence in late life underlines the importance of its environmental component, AF also has an important genetic component. To date, hundreds of genetic loci have been associated with AF. For some, a gene can confidently be prioritized as the causal association, but for the majority, causality remains to be established. A comprehensive understanding of the genomic etiology of AF may provide the tools to find novel, more targeted, therapeutics. The expanding omic toolkit can provide orthogonal lines of evidence to narrow disease causing genes. Alternatively, these methods can also help understand non-genetically driven mechanisms. The combination of these methods applied to AF models and humans is the basis of the efforts outlined in this thesis.

1.1 Atrial Fibrillation: Advances and Ongoing Challenges

1.1.1 AF comorbidities and epidemiology

The European society of cardiology defines AF as "*a supraventricular tachyarrhythmia with uncoordinated atrial electrical activation and consequently ineffective atrial contraction*". Ineffective atrial contractions sometimes reaching up to 600/minute can occur without symptoms or can produce debilitating consequences such as chest pain, light-headedness, fatigue, or shortness of breath. Stagnating blood in the atria facilitates blood clothing and increases the risk of stroke. AF patients have an increased risk of stroke of about five-fold[2]. Together, this leads to AF related healthcare costs estimated at $28 billion/year in the US in 2020[3]. Despite important progress made in AF management, treatment and basic knowledge, AF burden is quickly increasing. Between 1990 and 2010, the number of disability adjusted life-years from AF increased by ~19% and mortality by 2-fold[4]. A better understanding of the mechanisms of AF is imperative to develop new, more specific pharmacological therapies.

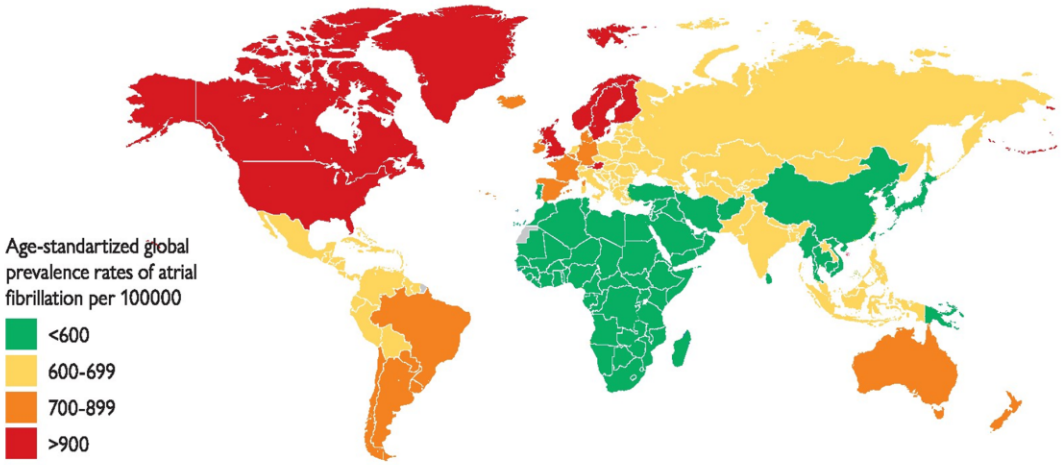
1.1.1.1 Prevalence and risk

In the United States (US), 1 in 3 Caucasians and 1 in 5 African Americans will suffer from AF during their lifetime[5]. Notably, African Americans, Hispanics and Asians appear to have similar AF risk, while Europeans' risk is markedly higher[6]. AF is most correlated with age (**Fig. 1**), with less than 0.5% of individuals affected before age 50, about 1% at age 60, and over 10% at age 80[7]. Among obese individuals, the risk of developing AF is 50% higher[8]. In an era where populations are both aging and getting more obese (obesity rate increasing from 30.5% in 1999-2000 to 37.7% in 2013-2014 in the US[9]), new strategies to reduce the burden of AF are necessary. Other important risk factors for AF are hypertension, smoking, having other cardiovascular diseases (CVD), being diabetic, male sex, and having a genetic predisposition. Many of these risk factors are modifiable. The risk of complications from CVD in general can be significantly altered based on the 7 behaviors and health factors (blood lipids, smoking, blood pressure, blood glucose levels, body mass index, exercise, and diet) of the American Heart Association (AHA)[5].

A

GLOBAL PREVALENCE OF AF

(globally, 43.6 million individuals had prevalent AF/AFL in 2016)



B

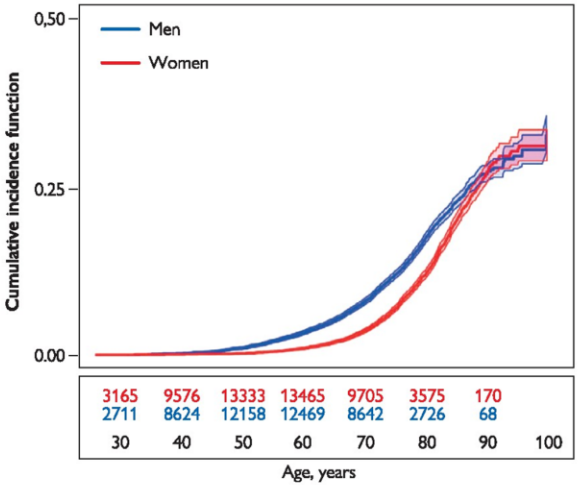
LIFETIME RISK for AF
1 in 3 individuals

of European ancestry
at index age of 55 years
37.0% (34.3% to 39.6%)

C

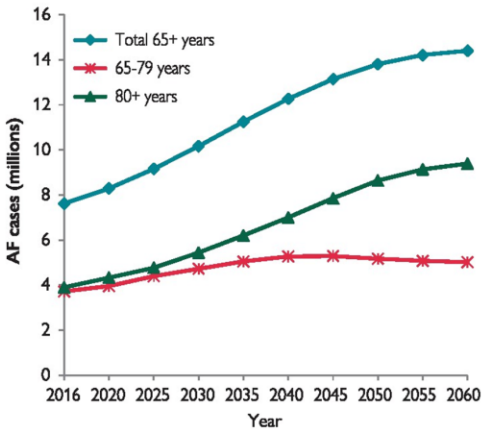
AF is more common in males

Cumulative incidence curves and 95% CIs
for AF in women and men with death as a competing risk



D

Projected increase in AF prevalence among elderly in EU 2016-2060



E

Lifetime risk of AF increases with increasing risk factor burden[a]

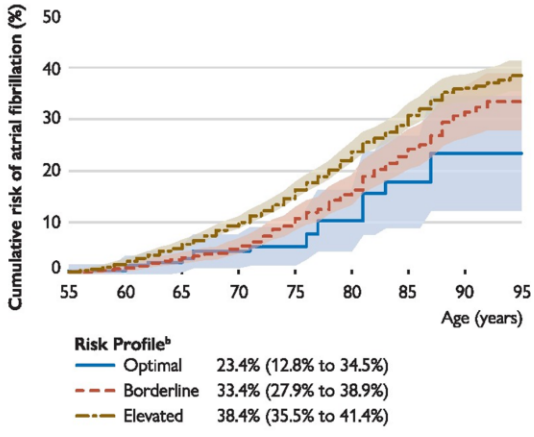


Figure 1. Global AF prevalence.

(A) World map of AF prevalence. **(B)** Symbolic representation of AF lifetime risk in Europeans. **(C)** Male/Female cumulative incidence curves by age group. **(D)** Projected increase in AF to year 2060 stratified by age groups. **(E)** Lifetime cumulative risk of AF by age, stratified by risk factor. AF = atrial fibrillation; AFL = atrial flutter; BP = blood pressure; CI = confidence interval; EU = European Union. [a]Smoking, alcohol consumption, body mass index, BP, diabetes mellitus (type 1 or 2), and history of myocardial infarction or heart failure. [b]Risk profile: optimal − all risk factors are negative or within the normal range; borderline − no elevated risk factors but >1 borderline risk factor; elevated − >1 elevated risk factor. Reproduced from the 2020 ESC guidelines[1].

1.1.2 The cardiac conduction system

The conduction system of the heart is composed of specialized structures with distinct properties. A normal heartbeat begins with the depolarization of the pacemaker cells at the *sinus node* (SN) (**Fig. 2**), setting the sinus rhythm. The depolarization of adjacent cells is induced through the flow of ions in shared gap junctions. The conduction of the signal initiated at the SN to the cardiomyocytes (CM) of the atria happens from top to bottom through the *Bachmann bundle* in the left atrium (LA) and through the *internodal pathway* in the right atrium (RA), allowing the synchronous contraction of the atria and the expulsion of blood into the ventricles. The presence of a fibrous membrane between the atria and the ventricles normally limits the transmission of the impulse between the upper and lower chambers of the heart. To reach the ventricles, the impulse passes through the *atrioventricular node* (AVN) and then spreads through the *His bundle*, the *bundle branches* and finally the *Purkinje fibers* in the ventricles, producing a contraction that ejects the blood out of the ventricles.

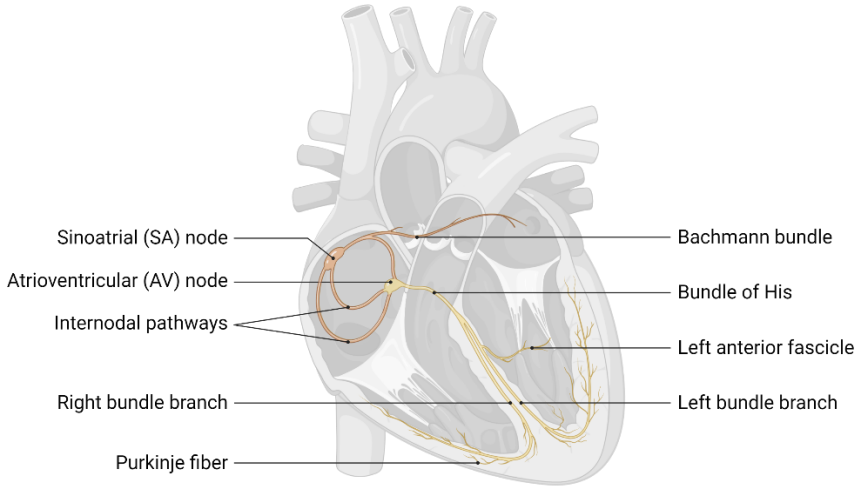


Figure 2. The conduction system of the heart.

Created with BioRender.

1.1.2.1 Cellular properties

Both the SN and AVN have spontaneous depolarization properties (automaticity). This is caused by the presence of leaky channels (hyperpolarization-activated cyclic nucleotide–gated; HCN channels), which facilitates the passive inflow of the positive ions Na^+, called the "funny" current (I_f). The SN has a faster intrinsic rate, which normally dictates the rate of depolarization at the AVN, but in the event of a disconnect between the two, the AVN can initiate depolarization. To allow the atria to completely empty before the contraction of the ventricles, the AVN also slows down the conduction of action potentials[10]. This is achieved through the presence of fewer gap junctions between cells. The other structures act as action potential highways, with different conduction speeds generally proportional to the size of their cells and number of gap junctions connecting them. For instance, the Purkinji cells, the largest cells of this system, have the fastest conduction rate at ~4 m/sec compared to ~1m/sec in the Bachmann bundle and internodal pathway and 0.3 m/sec in the atrial muscle[10]. Lastly, while the cells of the conduction system are considered CM, they do not possess contractile properties, unlike myocardial cells.

1.1.2.2 Cardiac action potentials

Cardiac contractions are modulated by the ion flow of CM with distinct patterns in nodal cells, atrial and ventricular myocardial cells (**Fig. 3A**). The main ion actors are K^+, Na^+ and Ca^{2+}. In the initial polarized state (resting phase or phase 4), CM are positively charged on their outer surface ($[Ca^{2+}]$ and $[Na^+]$ high) and negatively on their cytoplasmic face ($[K^+]$ high). This ion gradient is maintained by active transport of ion channels. To the exception of the pacemaker, change in voltage potential initiates the action potential. In the SN, this process is triggered by I_f. Elsewhere, the depolarization from a surrounding cell brings the membrane potential to a threshold allowing the opening of the fast gated Na^+ channels, followed by slow L-type Ca^{2+} channels. The inward currents of Na^+ (I_{Na}) and Ca^{2+} (I_{CaL}) are responsible for the sharp depolarization (phase 0). Around peak depolarization, Na^+ channels close and K^+ channels open, releasing K^+ outside the cell (transient outward currents; I_{to}). This results in a mild repolarization (early repolarization or phase 1). The plateau phase follows (phase 2), where the membrane potential remains relatively stable due to I_{CaL} neutralizing the potassium ionic currents (I_K). More Ca^{2+} is then released from

the sarcoplasmic reticulum into the cytoplasm by ryanodine receptors (RyR). A process known as calcium-induced calcium release. The high cytoplasmic [Ca^{2+}] allows the contraction of myofilaments, which produce the heart's contractions (**Fig. 3B**). Repolarization follows (phase 3), with the activation of the delayed rectifier K^+ channels increasing further the outflow of K^+ ions (I_{Kr}, I_{Ks} and I_{Kur}). Finally, the original ionic concentrations are reestablished by the Na^+/K^+ ATPase pump and Na^+/Ca^{2+} exchanger[11]. Importantly, during repolarization, CM are insensitive to a new signal, corresponding to the refractory period. Differences in isoforms and concentrations of these channels in nodal cells, atrial CM and ventricular CM lead to their differences in electrophysiologic properties (**Fig. 3A**).

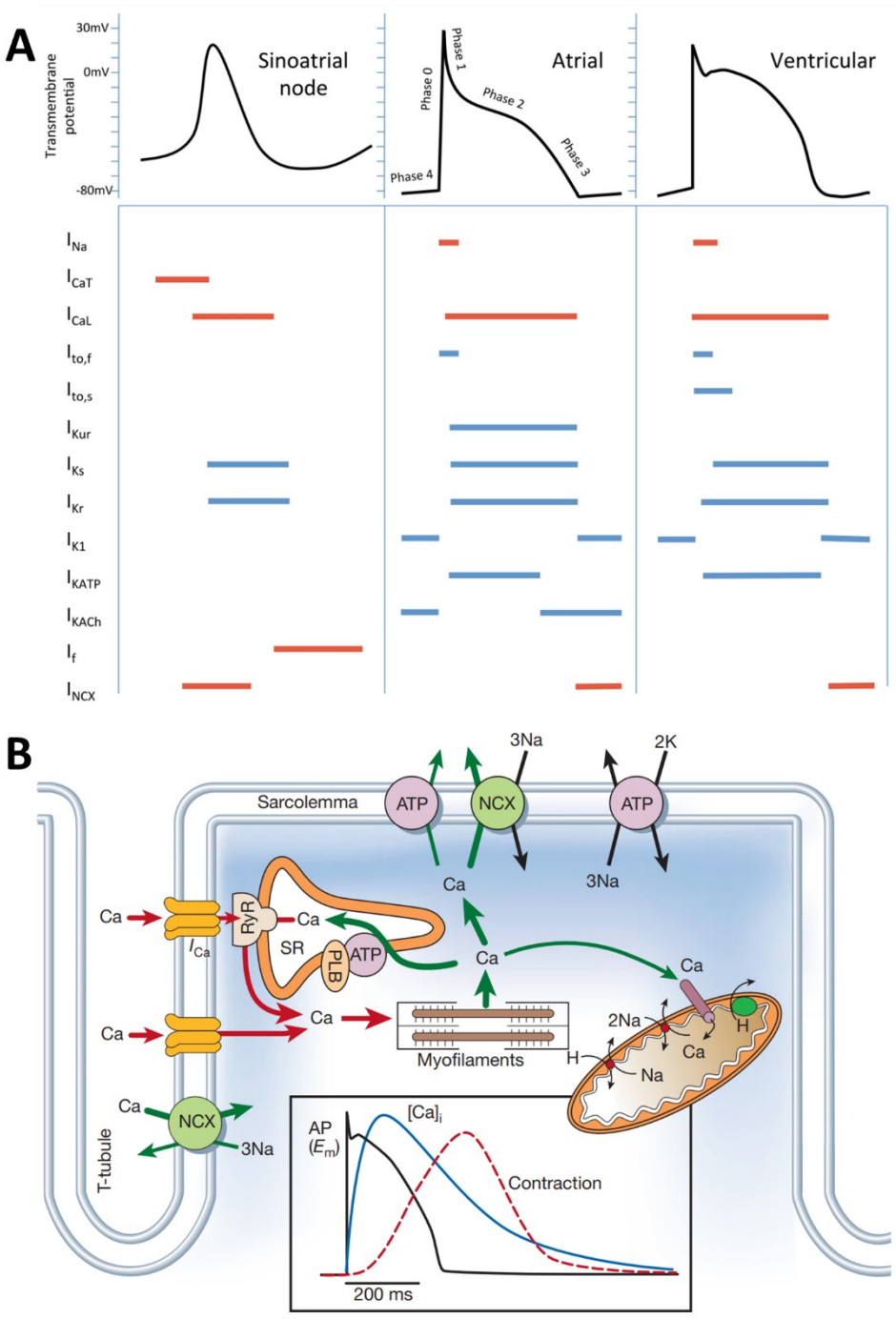


Figure 3. Cardiac action potentials.

A) From JD Lane et al. 2017[12]. Sinus node, atrial and ventricular cardiomyocytes action potentials. Colored lines indicate the phase of the action potential that the current participates in. Inward currents are in red, outward currents in blue. Currents - I_{Na}, inward Na^+; I_{CaT}, T-type Ca^{2+}; I_{CaL}, L- type Ca^{2+}; $I_{to,f}$, fast transient outward; $I_{to,s}$, slow transient outward; I_{Kur}, ultra-rapid K^+ delayed rectifier; I_{Ks}, slow K^+ delayed rectifier; I_{Kr}, rapid K^+ delayed rectifier; I_{K1}, inward rectifier; I_{KATP}, ADP-activated K^+ channel; I_{KACh}, muscarinic-gated K^+ channel; I_f, "funny" current; I_{NCX}, Na^+/Ca^{2+} exchange current. **B)** From DM Bers et al. 2002[13]. Ion exchanges

leading to depolarization and contraction of cardiomyocytes. SR; sarcoplamic reticulum, AP; potential action, NCX; Na^+/Ca^{2+} exchanger, PLB; phospholamban, ATP; ATPase.

1.1.2.3 Autonomic system regulation

Cardiac demand varies broadly with the activity of daily living, which requires the precise modulation of its contraction rate and force. Without extrinsic inputs, the adult human heart beats at approximately 100 time per minute (bpm)[14], in contrasts to the normal heart rate of approximately 70 bpm. This regulation occurs in part through a constant battle between the antagonistic sympathetic ("fight or flight") and parasympathetic ("rest and digest") systems. Other determinants such as thyroid hormones and lifestyle factors will be discussed in the *altered automaticity* section.

The heart is innervated by several sympathetic autonomic extrinsic nerves from the cervical and upper thoracic spine and a single parasympathetic nerve (vagus nerve) originating from the brainstem. Historically, parasympathetic innervations were thought to be mostly limited to the SN and AVN, as opposed to the sympathetic innervation which also reaches the atria and ventricles. A more recent consensus suggests that parasympathetic innervation also occurs in the atria and ventricles[15,16]. Extrinsic stimuli detected by the central nervous system can then trigger the release of acetylcholine (ACh) and norepinephrine (NE), acting directly on pacemaker cells and CM polarization and contractility. For instance, physical activity increases sympathetic activity and NE release through afferent signaling by the mechanoreceptors, chemoreceptors, baroreceptors, and thermoreceptors. The downstream molecular cascade of NE is mediated by the beta-adrenergic receptors, which results in the phosphorylation of L-type calcium channels, stimulating the inflow of calcium, increasing the depolarization rate and effectively decreasing the action potential duration (APD). Higher $[Ca^{2+}]$ also enables more actin-myosin cross-bridges in myofilaments, subsequently enhancing the force of contraction. Conversely, at rest, when cardiac demand is low, ACh released from parasympathetic nerves results in the inhibition of the pacemaker I_f through the activation of the muscarinic receptor 2 (M2), increasing the APD. In turn, this decreases the heart rate[17-19].

Dysfunctions of the systems overviewed above can lead to irregular cardiac contractions, called arrhythmias. The consequences of these heart rate and rhythm irregularities can range from

benign to life-threatening. In the next section, I'll briefly discuss the different mechanisms leading to arrhythmias, their causes and their impact on heart function.

1.1.3 Mechanisms of arrhythmias

Arrhythmias can be classified by rate, location, duration, and mechanism of action. Arrhythmias causing abnormally fast rates are called tachyarrhythmias (resting heart rate > 100 bpm) and those with abnormally slow rates are called bradyarrhythmia (resting heart rate < 60 bpm). Classification by localization generally partitions the heart in two, with the supraventricular arrythmias designating all arrhythmias above the ventricles (including the AVN), and ventricular arrhythmias. Classifications by duration are more useful for AF and will be detailed in its own section (see section 1.1.4). As for their mechanisms, three broad categories are generally described (**Fig. 4-5**); altered automaticity, triggered activity, and re-entry (altered automaticity and triggered activity are sometimes jointly referred as "ectopic activities")[20-23]. The following sub-sections will detail these mechanisms and put them in context with the most important arrhythmias.

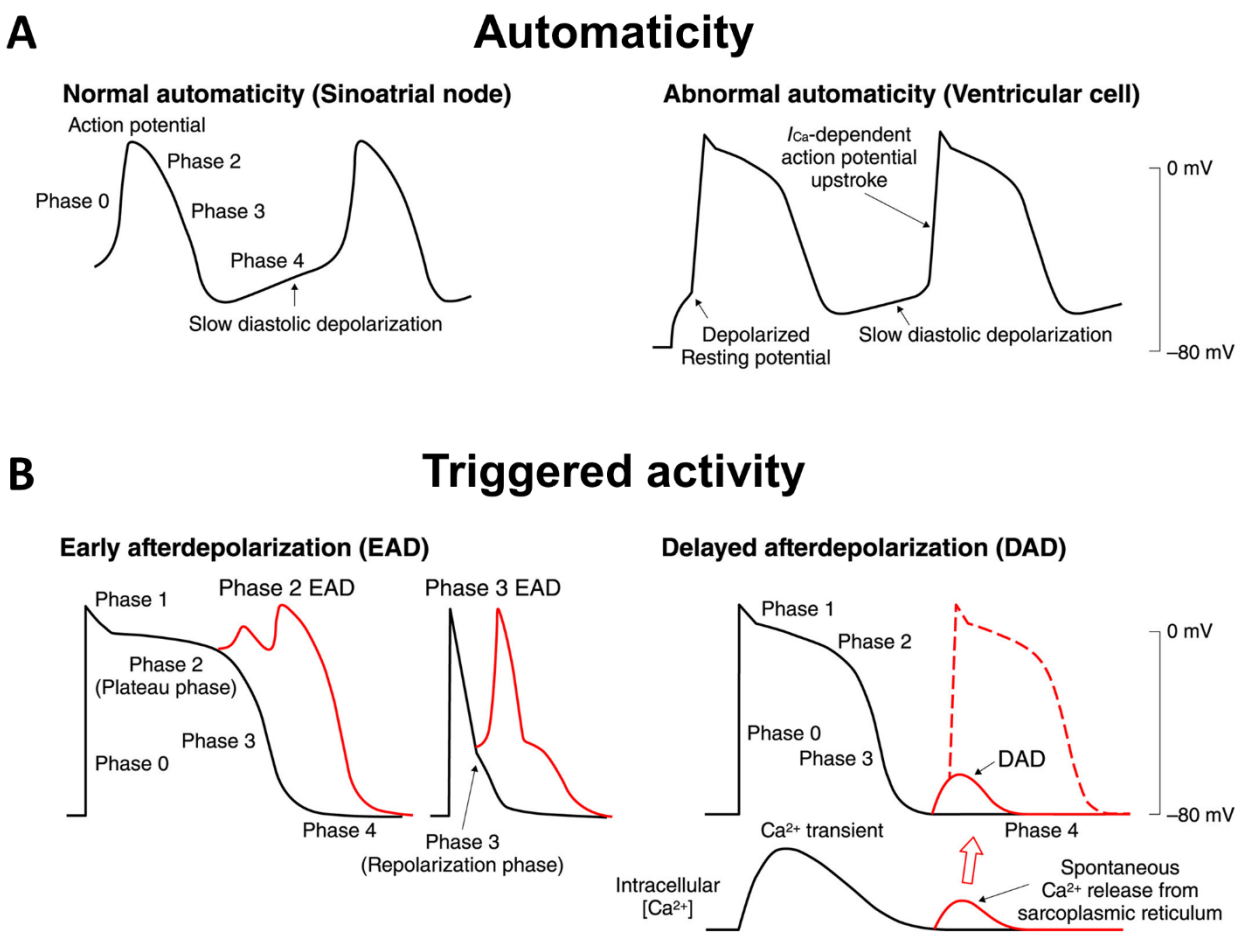


Figure 4. Mechanisms of arrhythmias.

A) Action potential of a ventricular cell abnormal automaticity shows a comparable phase 4 slow diastolic depolarization to sinoatrial cells. **B**) Normal action potential patterns (black) and abnormal afterdepolarizations patterns (red). On the right panel, the dashed line shows a triggered activity. Adapted from Ono, K et al. 2022[22]. I_{Ca}; C^{a2+} current.

1.1.3.1 Altered automaticity

As mentioned above, automaticity can be physiologically altered. For instance, in endurance athletes, resting heart rates below 60 bpm are frequent. In these individuals, the activity of the vagus nerve is increased (referred to as vagal tone). The prevailing hypothesis with regards to this phenomenon has been an afferent signal from baroreceptors, but studies in animal models also suggest that SN cells adapt to this stimulus by repressing I_f through the expression of the SN dominant form of HCNs, HCN4[24]. Various degrees of automaticity-driven rhythm patterns synchronized with normal breathing are also observed, called respiratory sinus arrhythmia. These are generally symptomless.

Non-physiologic factors can also induce abnormal impulses that may emerge from a latent pacemaker or from cells without automaticity properties under normal circumstances (**Fig. 4A**). Latent pacemakers at the AVN or His bundle, are called junctional rhythms. Some foci in the atria outside of the cardiac conduction system also possess automaticity, such as cells around the pulmonary veins, the mitral and tricuspid valves and the inferior vena cava[25]. Abnormal blood volume, electrolytes levels and metabolic activity can all trigger tachyarrhythmia or bradyarrhythmia at both ends of their spectrum. Perhaps most critically, potassium levels above 6 mmol/L (hyperkalemia) or below 3 mmol/L (hypokalemia) directly affect the cellular resting membrane potential, which can potentially be lethal. Dyskalemia can also alter the function of potassium channels, causing a delay or accelerating repolarization[26]. This, in turn, may allow the emergence of triggered activities (see below). Similarly, hyperthyroidism and hypothyroidism can also cause tachycardia and bradycardia respectively. The thyroid effector hormone triiodothyronine (T_3) can cause direct changes in ion homeostasis through the beta-adrenergic receptors and the Na^+/K^+ ATPase[27] but also changes in gene expression encoding proteins involved in myofilaments, adrenergic receptors ion pumps and ion channels[28].

1.1.3.2 Triggered activity

Positive fluctuations of the membrane potential occurring between beats are called afterdepolarizations (**Fig. 4B**). A triggered activity occurs when an afterdepolarization is strong enough to initiate a new action potential. Early afterdepolarizations (EAD) occur in phases 2 or 3 of the action potential cycle, while delayed afterdepolarizations (DAD) occur in phase 4. Together, DADs and EADs triggered activities depend on the activation of I_{Na}.

Because the membrane potential is relatively neutral during the plateau phase, it is more sensitive to small current alterations, making phase 2 EADs more frequent than phase 3 EADs. While less frequent, phase 3 EADs occur later and are more likely to propagate to non-refractory CM. Long QT syndrome is characterized by a longer APD in ventricular CM. EADs in these patients are thought to be the most common cause of torsades de pointes[29]. This can lead to an often-fatal arrhythmia, ventricular fibrillation, characterized by irregular heart rate preventing systemic circulation. Other factors such as hypokalemia, heart failure, bradycardia and medication can also increase APD and the likelihood of EADs[20,22,26]. The molecular basis for EADs is

generally associated with abnormally elevated inward currents, such as I_{Ca} and I_{NCX}, or decreased outward currents (I_K) during the plateau phase or the repolarization phase[20,29,30].

DADs are generally associated with abnormally high [Ca^{2+}] in the sarcoplasmic reticulum, referred to as calcium overload. Calcium overload has been associated with hypokalemia, hypertrophy and heart failure[31-33]. Sudden outflow of Ca^{2+} in the cytoplasm can activate the I_{NCX}, exporting one Ca^{2+} ion for three Na^+ imported ions, thus increasing membrane potential. If threshold is reached, an action potential is triggered. Multiple lines of evidence implicate the cardiac RyR isoform RyR2 as likely cause for this release, either through genetic mutations or upstream beta-adrenergic signalling[34]. Subthreshold DADs may also predispose to re-entry[35] (discussed in the next section). DADs are thought to be a cause for ventricular arrhythmias and AF.

1.1.3.3 Re-entry

A re-entry occurs when non-refractory cells are depolarized by one or multiple self-sustaining circulating currents that are independent of automaticity. These circuits can be anatomical (also called macro-re-entry) or functional (also called micro-re-entry). Because re-entry cannot propagate to refractory cells, it is dependent on the presence of slower, suitably long conduction pathways or the presence of a shorter refractory period. Importantly, re-entry usually occurs in conjunction with other mechanisms, such as triggered activities.

Anatomical circuits

Anatomical circuits (**Fig. 5A**) can be congenital, such as the Wolff-Parkinson-White syndrome. This condition is characterized by the presence of an accessory pathway linking an atrium to the ventricles. Under normal circumstances, this condition is asymptomatic because the action potential is synchronously descending (anterograde) in the septum and in the ventricular wall through the accessory pathway, dying off in the ventricles. Concurrently, if a block or path of resistance prevents the action potential to reach the accessory pathway before the impulse through the AVN reaches it, a retrograde conduction into the atrium can occur and cause a self-sustaining re-entrant circuit (**Fig. 5A**), resulting in atrioventricular re-entrant tachycardia. The AVN is also amenable to re-entry when a path of resistance is present (slow vs fast pathways), resulting in rapid impulses from the AVN and i.e., atrioventricular re-entrant nodal tachycardia[36]. Anatomical re-entry can also occur within the atria, causing atrial flutter[37].

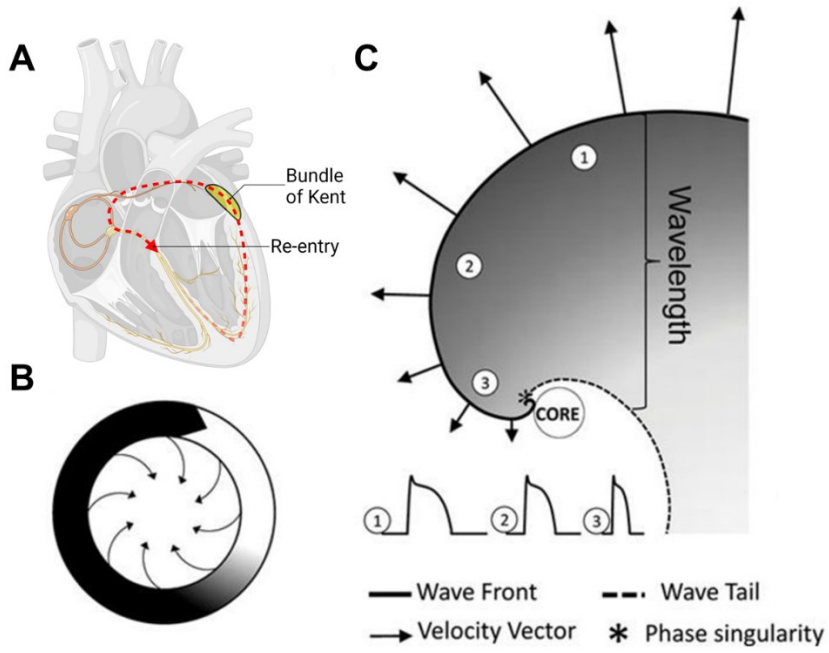


Figure 5. Anatomical and functional re-entry mechanisms.

A) Wolff-Parkinson-White syndrome example of anatomical retrograde re-entry around an anatomical obstacle. Created with BioRender. **B-C**) Schematic representation of functional re-entry, from SV Pandit et al.[38]. **B**) Leading circle theory with centripetal forces pointing inwards toward a refractory center. **C**) Rotor model showing a refractory core from which a rotating depolarizing wave with increasing velocity emerges.

Functional circuits

Functional re-entry does not depend on the presence of specific anatomical structures, but instead, relies on the heterogeneous electrophysiologic properties of the substrate. Furthermore, micro-circuits are not necessarily fixed and can trigger the emergence of multiple competing micro-circuits, which are more likely to cause fibrillation. Many models of micro-re-entry have been proposed over the years such as the leading circle, anisotropic re-entry, the rotor and others[21,39].

The first attempt to explain functional re-entry was posited by Allessie et al. in 1976[40]. Studying re-entry in a rabbit model without anatomical block, they formulated the leading circle model (**Fig. 5B**). Their model allows the theoretical formation of smaller circuits by removing the necessity of an excitable gap between the wavefront and the tail of the circuit. Instead, they proposed that the tail would be partially refractory. A permanently refractory core was also

proposed, resulting from the constant centripetal impulses from the wavefront. Today, rotors are a preferred model (discussed below).

The consideration of cardiac fiber orientation and its implications in conduction velocity is a more recent development. Cardiomyocytes are generally shaped as long cylinders with higher concentrations of gap junctions in the intercalated disks connecting the shorter faces of neighboring CMs. The cardiac muscle is anisotropic, given that an action potential is conducted with greater velocity longitudinally than transversally to cardiac fibers. If a longitudinal path is slowed or blocked long enough for transversal fibers to end their refractory period, a slower transversal current can cause re-entry. Alterations in cell shape and sizes can contribute to increase anisotropy[41]. Fibrosis, a frequent cardiac remodeling mechanism, may also further increase the transvers vs longitudinal conduction velocity difference and gap junction lateralization (re-localization to the longitudinal face) or dysregulation[42]. Therefore, conditions altering those properties like ischemia and heart failure are though to contribute to anisotropic re-entry and may promote arrhythmias[41,43].

The different layers of the heart have different action potential patterns due to variations in ion channels concentrations (most notably affecting I_K). Decreasing APD is observed in cardiac layers in the following order myocardium, endocardium, and epicardium. Moreover, the epicardium is prone to heterogeneous action potentials due to the presence of a strong I_{to}, which in some cases inhibits I_{CaL}, drastically reducing the APD[44]. In Brugada syndrome and myocardial ischemia, the endo to epicardium APD difference is further exacerbated, increasing the likelihood of re-entry due to endo-epicardial asynchrony, potentially leading to AF and ventricular fibrillation[45].

While the leading circle model was useful to explain re-entry in the absence of anatomical obstacles, today the rotor model is favored due its closer agreement with cardiac mapping studies and mathematical models[46]. Rotors (or spiral waves) have hurricane like features (**Fig. 5C**). At their center, a core of hyperpolarized cells remains unexcited despite frequent subthreshold action potential flickering. From the wavefront inwards, a decreasing velocity gradient is formed because of a progressive “dilution” of the depolarizing source necessary to trigger an increasing number of excitable connected cells. At the core, it reaches a critical point called phase singularity, where the stimulus becomes insufficient to trigger an action potential. Rotors can form when a current wave

encounters a pocket of refractory cells (e.g., from a triggered activity) bending the wave and creating a curved wavefront. With enough curvature, the wave begins to rotate around a central core, forming a rotor. Rotors can be fixed or drifting. Encountering obstacles, they can break into multiple rotors and lead to fibrillation. Rotors are thought to be involved in many arrhythmias such as ventricular tachycardia, ventricular fibrillation, and AF[20].

1.1.4 Molecular basis of AF pathophysiology

AF is a progressive disease where both frequency and duration of event tend to increase over time, a result of cardiac tissue remodeling. The European society of cardiology (ESC) 2020 AF classification based on progression suggests 5 types of AF; 1-first diagnosis, 2-paroxysmal (where the normal rhythm is regained without medical intervention within 7 days), 3-persistent (sustained AF during >7 days), 4-long-standing persistent (>12 months with a control strategy) and 5-permanent (an accepted state when the risks of interventions are considered greater than the benefits of cardioversion)[1]. Previous nomenclature used chronic AF encompassing persistent and long-standing persistent AF. For the purpose of describing the disease progression, I will mostly use paroxysmal and persistent AF to distinguish remodeling mechanisms occurring under transient vs sustained AF respectively.

The first theories of AF mechanisms were established in the beginning of the 20th century[47]. Remarkably, these theories still hold their ground today[39]. The mother wave, multiple wavelets and trigger focus theories all have shown to have some basis for AF (**Fig. 6**). Refinements of these concepts are now the subject of debates. The etiology of AF remains incomplete and may differ across individuals. Evidence for both ectopic activities and re-entry have been reported[36]. Located at the back of the LA, the pulmonary veins (PV) have been a recognized as important foci of AF since 1997[48]. They are now prioritized in most ablation procedures, but have lower success rates as AF progresses[49]. The nature of re-entry can be caused by a single, fixed circuit or by an array of drifting circuits. Rotors have taken the center stage as the potential source of AF propagation in the past few decades but results of clinical trial investigating the effect of their ablation remains mixed[50-52]. Others argue for the epicardial to endocardial conduction asynchrony to be the major source of re-entry[53]. Regardless of these mechanistic uncertainties, an amenable substrate is required for their emergence, which is generated through different types of remodeling and accelerated by risk factors. Identifying the molecular drivers of remodeling may empower the

development of more targeted pharmacological therapies and tailored management. On this front, significant advancements have been made in the past two decades. Electrical remodeling has long been a suspected driver of AF, but more recent body of work also suggest a role for fibrosis, proteostasis, mitochondrial function and inflammation. The following subsections discuss the main molecular bases of these remodeling events.

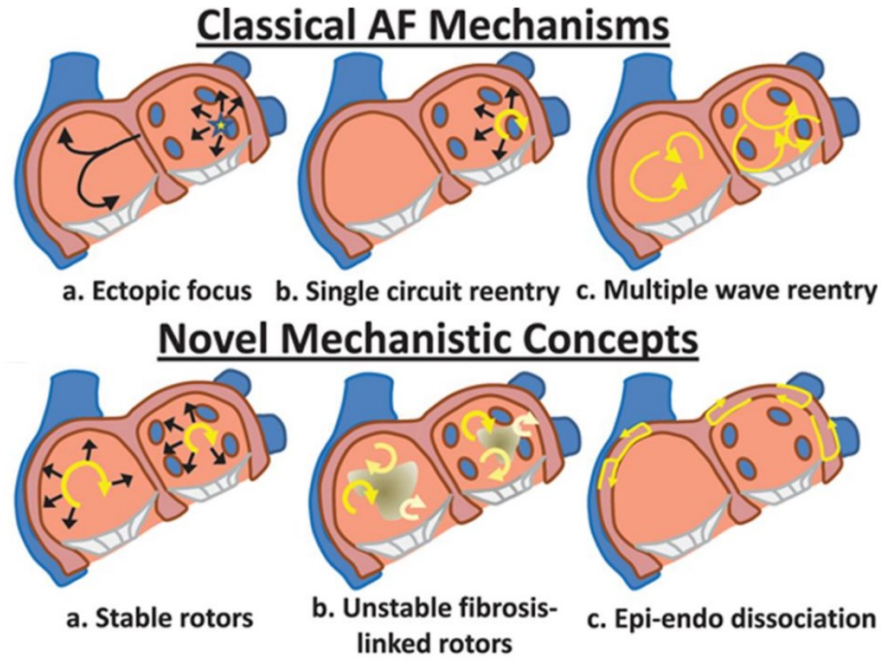


Figure 6. Classical and modern AF mechanisms.

Cartoon representations of the atria and mechanisms of AF, with the action potential propagation depicted with black arrows and re-entrant circuits in yellow arrows. Adapted from S Nattel et al. 2017[39].

1.1.4.1 Electrical remodeling

The role of electrical remodeling in AF has been extensively studied, but most knowledge is based on observations made in persistent AF. The high success rates of PV ablations in paroxysmal AF clearly point to their important contribution at this stage, but recurrences suggest other factors are involved. For persistent AF, poorer outcomes suggest a greater importance of other contributing factors such as increased structural and electrical remodeling.

Persistent AF leads to important CM adaptations caused by the substantial stress of sustained rapid atrial contractions. There is an extensive body of research centrally implicating

calcium in this process. First, under physiologic conditions, atrial CM have higher sarcoplasmic Ca^{2+} contents and a faster restoration of its [Ca^{2+}] at diastole than ventricular CMs[54]. By itself, this may increase the likelihood of DADs in the atria. During sustained AF, DADs triggered activities can be facilitated by the calcium–calmodulin-dependent protein kinase II (CaMKII) increased phosphorylation of RyR2, leading to more frequent sarcoplasmic reticulum Ca^{2+} leakage[55]. Furthermore, constant triggering of CMs at supraphysiological rates increases calcium loading. This triggers the transcriptional repression of the voltage-dependent L type calcium channel alpha 1C subunit, leading to a reduced I_{CaL}, a shorter APD and facilitated re-entry[56].

Beyond calcium-related remodeling, the atrial specific G-protein-gated K+ channel current (I_{KACh})[57] also appears to be altered during sustained AF. These channels increase the repolarization rate when activated by the ACh/M2 axis, effectively reducing the APD. In dogs, I_{KACh} is higher in the LA compared to the RA and even higher in appendages, likely because of increased expression of the M2 receptor[58]. In isolated human CMs, I_{KACh} appeared constitutively activated only in persistent AF patients[59]. Further, a rotor modeling study showed that ACh promotes rotor stability and frequency[60].

While less is known of the electrical remodeling at the source of paroxysmal AF, there is strong evidence for the contribution of PV. Some studies have shown that CMs located at the PV sleeves have reduced APD and a higher resting membrane potential, properties that may promote re-entry[61,62]. PV fiber orientations may also contribute to re-entry through anisotropy. Another study in dogs suggest the presence of cells with pacemaker properties[63], providing a basis for altered automaticity impacting PV. Afterdepolarization triggered activities have been reported in PVs under a broad range of AF related conditions such as autonomic nerve stimulation or ACh exposure, tachypacing, thyroid hormone exposure and hyperthermia[61]. While PV isolation has a good success rate, many paroxysmal AF patients have recurrences, suggesting that a different mechanism is involved. To date, there is little evidence of electrical or other types of remodeling that could explain these recurrences, underlying the need for more studies in this domain.

It is to be noted that many studies are conducted in the RA, but there are considerable differences between the left and RA. Clinically, greater remodeling is generally observed in the LA. The LA is also more innervated by the autonomic system and shows higher sensitivity to ACh.

1.1.4.2 Structural remodeling

Alteration of the extracellular matrix (ECM) plays a major role in facilitating AF[64]. Collagen deposition can occur between CMs (reactive fibrosis or interstitial) or after CM death (reparative fibrosis). Interstitial fibrosis can increase anisotropy as collagen is usually deposited between CMs longitudinally, while patches of fibrosis replacing dead CMs may disrupt longitudinal impulse conduction and promote rotor formation, both facilitating re-entry. The mediator of these processes is fibroblasts. Beyond ECM homeostasis, fibroblasts can also produce cytokines and growth factors with paracrine effects. Fibroblast proliferation or differentiation into the pro-fibrotic myofibroblast state can cause inflammation, cardiac injury, and pressure overload.

In the context of AF, the role of the vasoconstrictor hormone Angiotensin II (AngII) is perhaps the most well documented for its effect on fibroblast due to the strong AF risk increase of hypertension (present in 60-80% of persistent AF patients)[65]. AngII is converted from Angiotensin by the enzyme angiotensin-converting enzyme (ACE). In fibroblasts, AngII triggers the transcription of transforming growth factor beta 1 (TGF-β1), a central pro-fibrotic modulator. In a meta-analysis, ACE inhibitors were successful as primary and secondary prevention of AF, reducing incidence by 24% and 27% respectively[66]. Both inhibited AngII signaling in fibroblasts and a reduction of pressure overload leading to mechanical stretch of the atria may play a role in this outcome. Mechanical stretch can increase atrial size and remodeling, which facilitates re-entry. Increased atrial size is also associated with fibrosis. While many have hypothesized that mechanical stretch itself may trigger fibroblast extra cellular matrix remodeling through some mechanosensory pathway intrinsic to fibroblasts, a recent review questions the consistency of the results on the matter[67].

Perhaps less intuitively, fibroblasts can alter electrophysiological properties of CMs directly through gap-junction-mediated heterogeneous cell coupling[68]. Fibroblasts express a variety ion channel allowing for Na^+, K^+ and Ca^{2+} currents[69]. Moreover, the resting membrane potential of fibroblasts is much higher than CMs (−31 to −16 mV[69]), affecting ion flow through gap junctions directionally dependent on the action potential phase. Of particular significance, fibro-CM coupling might play an acute role in the SN, where high concentrations of fibroblasts have been reported[68]. While co-culture models showed increased abnormal automaticity[70], how this translate in vivo requires further investigation.

1.1.4.3 Mitochondrial dysfunction

AF is also associated with a strong metabolic shift with evidence implicating the mitochondria[71,72]. The adenosine triphosphate (ATP) content in strongly depleted in the atria of AF patients and more strongly so in the LA[73]. Concordantly, mitochondrial function and structure are also compromised. The poly(ADP-ribose) polymerase-1 (*PARP1*) is a DNA repair enzyme using NAD^+ as substrate. Its (hyper)activity is associated with the depletion of nicotinamide adenine dinucleotide (NAD^+) pools, an essential coenzyme in mitochondria for an array of processes[74]. A group investigating the effect of tachypacing on HL-1 CMs and Drosophila found that it lead to PARP1 activation and NAD+ depletion[75], which has expected consequences on mitochondrial function. Moreover, oxidative stress also appears to be a consequence of AF. Researchers found that RyR2 oxidation was increased in chronic AF compared to control patients and that a mouse model with induced RyR2 leakage lead to mitochondrial dysfunction and reactive oxygen species (ROS) production[76].

1.1.4.4 Inflammation

Inflammation is correlated with AF progression and is a predictor of ablation outcomes[77,78]. Inflammatory cytokines such as tumor necrosis factor alpha (TNF-α), interleukin 1 beta (IL- 1β) and interleukin 6 (IL-6) promote fibrosis and electrical remodeling, but also may promote protein misfolding[78]. TNF-α is a key inflammatory cytokine implicated in immune regulation cell proliferation, inhibition and apoptosis[79]. It can also induce differentiation of cardiac fibroblast into collagen producing myofibroblasts, promoting fibrosis and AF substrate[79]. One major transcription factor responsible for the transcription of many of these cytokines is nuclear factor kappa-light-chain-enhancer of activated B cells (NFκB). In left atrial appendages (LAA) of AF patients, it was shown that NFκB was over-expressed and more phosphorylated (activated) compared to the sinus rhythm group[80]. More recent evidence also implicates the NLR family pyrin domain containing 3 (NLRP3) inflammasome in AF. Inflammasomes are oligomers, known to mediate the innate immune response by catalyzing the maturation of IL-1 family of cytokines[81]. A recent study showed that in CMs of both paroxysmal and persistent AF patients, NLRP3 had increased activity[82]. Further, the investigators showed that NLRP3 knock-in in mice promoted ectopic activity, reduced atrial refractory period, and increased sarcoplasmic reticulum Ca2+ leaks, all features of AF. Thus, there is strong evidence implicating inflammation in AF, but research on its role in AF is still in its infancy.

Protein homeostasis (proteostasis), through the regulation of misfolded or unfolded proteins can also bolster inflammatory processes[78]. Together, lysosomes and proteasomes are responsible for protein degradation. Overload of the endoplasmic reticulum (ER stress) can lead to increased misfolding and activation of the unfolded protein response. Both proteasomal degradation through ubiquitination and lysosomal degradation through autophagy can ensue. Recent research provided evidence of increased ER stress and activation autophagy in persistent AF patients[83]. These investigators further showed that blocking ER stress could rescue I_{Ca} in tachypaced HL-1 CMs. Moreover, there is increasing evidence for the involvement of protein aggregates in CVD. For instance, while most studied for its role in Alzheimer's disease, tau aggregates have also been found to promote diastolic dysfunction[84].

While the current pace of discoveries is encouraging, many of AF's molecular determinants remain to be discovered. In addition, there is an urgent need to translate this knowledge to actionable treatments for patients. In the next section, I will cover the current treatment options AF patients.

1.1.5 Treatment and management of AF

The recent "AF Better Care (ABC) pathway" approach[85], supported in the ESC 2020 guidelines[1], emphasize three axes for an integrated AF management; (A) avoid stroke; (B) better symptom management; (C) cardiovascular and comorbidity optimization. Multiple randomized control trials showed marked effectiveness of this approach compared to the conventional approach[1]. Special considerations should be taken for hemodynamically unstable patients but fall outside the scope of this work. Below I discuss current best practices for stable patients with an emphasis on pharmacologic options.

1.1.5.1 Avoid stroke

Because the most significant risk of AF is stroke, anticoagulants are generally the first line of treatment. Anticoagulation therapy must balance bleeding and stroke/embolism risks, assessed with the HAS-BLED and CHA_2DS_2-VASc scores respectively. Warfarin, the most widely used vitamin K antagonist, reduces cloth formation by impairing the production of coagulation factors[86]. However, dosage must be closely monitored because of its narrow therapeutic dose and because of the high inter-individual metabolism variability which depends on both environmental factors

and gene variants[87]. Several new anticoagulants have recently been indicated as superior to warfarin, such as apixaban and rivaroxaban, further reducing mortality rates, and strokes, which may be explained by the difficult-to-achieve therapeutic dosage of warfarin[88]. Generally, oral anticoagulants are well tolerated and safe, reinforcing their place as first line treatment.

1.1.5.2 Pharmacologic rhythm and rate control

The main options to achieve better symptom management are restoring the sinus rhythm (rhythm control) or settling for regulating the ventricular rate (rate control) at the AVN[1]. Beta-blockers, calcium channel blockers (CCB) and digoxin are the prevailing rate control medication. Beta-blockers reduce sympathetic nervous system activity by inhibiting the activation of beta-adrenergic receptors, thereby reducing heart rate and blood pressure. CCB reduce I_{Ca} which leads to slower AVN conduction rates. Non-dihydropyridine CCB are generally used for rate control as they have lower vascular effects and greater cardiac specificity. Digoxin acts through the inhibition of the Na^+/K^+ ATPase, increasing cardiac contractility and reducing heart rate[89]. Beta-blockers and CCB are suggested as first line drugs. While these drugs are generally better tolerated than anti-arrhythmic drugs (AAD), similar side effects can occur such as dizziness, weakness, fatigue, nausea, or insomnia. Many AAD are available which are classified by their action (Vaughan-Williams classification) on the various ion channels[90]. One of the most widely used among them is Amiodarone, which appears to act on multiple currents[91]. However, serious adverse effects can occur with prolonged use in up to 50% of patients such as pulmonary toxicity, liver damage and thyroid dysfunction[92].

The decision to rhythm control or rate control has historically been made based on symptoms and concomitant diseases. Given that prolonged atrial tachyarrhythmia is known to increase remodeling and promote AF progression, rhythm control appears as an appealing rational to stop AF progression. However, the 2002 AFFIRM trial showed no statistical difference between rate and rhythm control for mortality, while rhythm control drugs showed increased hospitalization rates and serious adverse effects such as torsade de pointes and bradycardia[93]. Until very recently, the results from this study have shaped the decision to limit the use rhythm control drugs around symptom management. In the 2020 EAST-AFNET 4 trial, early rhythm control was evaluated against rate control in patient diagnosed with AF within 1 year, which were predominantly first diagnoses or paroxysmal[94]. The primary endpoint was composite death from cardiovascular

outcomes including stroke, heart failure and acute coronary syndrome. Both ablation and AAD were used for rhythm control. The study showed a significant reduction in composite outcome for rhythm control, with more patients maintaining sinus rhythm after two years. Adverse events remained higher in the rhythm control group. A subsequent analysis showed no difference in outcomes between symptomatic and asymptomatic patients[95]. These new results support the use of rhythm control in early diagnosis.

Overall, a significant number of issues associated with these interventions remain, mainly due to their non-specific action. The discovery of more specific treatments are required to reduce risks and side effects.

1.1.5.3 Cardiac ablation

The risk and efficacy of ablations to restore sinus rhythm have significantly improved and have been shown to be superior to AAD for quality of life[96]. The PV ablation followed the discoveries of Dr. Michel Haïssaguerre[97]. In this procedure, a series of radiofrequency induced lesions are performed around the PV. Since then, the procedure has been refined, increasing the ablated area around the PV to include nerve bundles that may cause recurrences[98]. Today, ablations further benefit from 3D mapping, guiding ablation when needed and more recent catheters including force sensing[99] and cryoballoon catheters[100]. Ablation is indicated for an increasing proportion of patients and is especially effective early in the progression of AF. Despite these advancements, success rates are 50-60% and 80% in paroxysmal AF and 40% and 60% in persistent AF after a single or multiple procedures respectively[101]. Moreover, complications rates are 5–7%, including serious complications such as cardiac tamponade and stroke[101].

For patients with debilitating symptoms in whom alternatives have failed, ablation of the AVN with pacemaker rate control may be indicated. This relatively safe procedure has shown improvements in quality of life and left ventricular ejection fraction[1]. There is currently no randomized control trial for this procedure and its effect on atrial remodeling remains elusive.

1.1.5.4 Early detection and modifiable risk optimization

Early detection is becoming a priority to reduce AF progression through cardiac remodeling. The emergence of affordable and accessible technologies monitoring cardiac electrical activity (via electrocardiogram; ECG) promises to have a significant impact on disease

control. In 2019, The WATCH AF Trial showed that AF detection using over-the-counter smartwatches allowed for a sensitivity of 93.7% and specificity of 98.2%[102]. This is of particular importance considering that some individuals remain asymptomatic for years. Lifestyle interventions such as increasing cardiovascular fitness or weight loss are extremely potent in reducing AF recurrence, often showing greater freedom from AF than rhythm or rate control[103]. Paired with advances in catheter-based ablation, early detection and aggressive comorbidity management promises to have a major impact on the burden of AF.

Besides modifiable risk management, the current paradigm still falls short of a treatment that can significantly reduce or reverse atrial remodeling. In the following sections I will discuss how the combination of omic technologies can expand our current molecular understanding of AF, improve risk stratification and help personalize management.

1.2 Multiomic strategies for molecular target discovery

Biology has been transitioning from a qualitative to quantitative science at an increasing rate. Sequencing technologies have outpaced Moore's law since the development of next generation sequencing (NGS), bringing down the cost of sequencing a human genome from ~7M$ in 2008 to ~500$ in 2022 (**Fig. 7**)[104]. This enabled the emergence of unbiased whole transcriptome and epigenome sequencing. In parallel, microarray technology also blossomed, allowing for the genotyping of hundreds of thousands of variants at even lower costs. Using this technology, millions of individuals have now been genotyped. More recently, advances in microfluidic technology further expanded the applications of NGS to single cells or nuclei. The convergence of these technologies is turning biology into "Big Data", on which evermore powerful models can be trained.

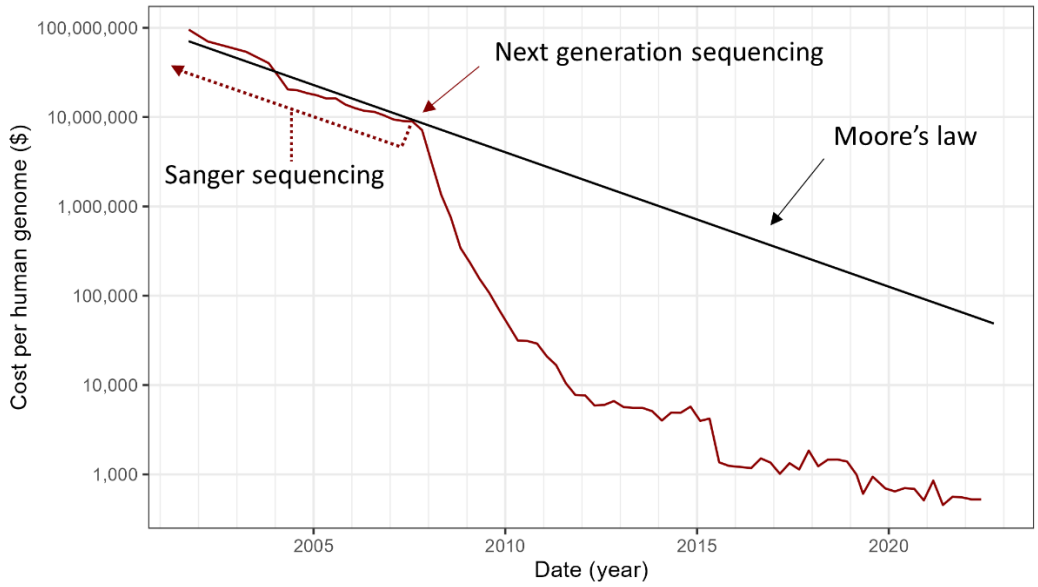


Figure 7. Sequencing costs over time.

Adapted from[104].

1.2.1 Genetics

Inheritance has fascinated humans for millennia. Theories attempting to explain the apparent transmission of one's traits to one's offspring date back at least to ancient Greece. Pythagoras' "spermism" theory posited that male sperm contributed all information necessary for conception through a mystical imbuement of his traits over his lifetime, while the female provided

only nurture[105]. Two centuries later, Aristotle refuted major flaws of this theory, such as the obvious unaccounted sex-specific traits of female offspring. Instead, he proposed that females' equivalent of males' sperm was menstrual blood, providing the "messages" for conception. The essence of these theories echoed in history, finding its way millennia later in Darwin's Pangenesis gemmule theory of inheritance in 1868[106], three years after Gregor Mendel had published the laws of inheritance[107]. Deemed the father of modern genetics, Mendel's work would revolutionize biological science, but remained undiscovered for more than 30 years.

The power of genetics in understanding traits and diseases resides in the lifetime exposure to one's inherited genetic variations (called "variants" hereafter). Contrary to environmental variables, which change over time and can be difficult to measure accurately, variants are constant throughout life (aside from somatic mutations). Causal associations for given diseases can thereby be inferred directly in humans, in some cases providing molecular targets to treat the associated disease. This strategy has been shown to increase the success rate of drugs in clinical trials by more than two-fold[108].

1.2.1.1 Mendelian genetics

High impact mutations, such as those leading to the gain or loss of a gene's function can be sufficient to be fully deterministic on a phenotype (monogenic or Mendelian disease). Since Thomas Morgan's work on fruit flies and that of Botstein et al. on human polymorphisms[109], causal genes of Mendelian diseases can be traced in families using linkage analyses (using chromosomal markers and recombination likelihood to narrow down the causal genetic locus). Today exome sequencing or whole genome sequencing have largely replaced this tedious process. Nevertheless, the identification of rare variants causing "familial" AF has provided numerous candidate genes.

Rare mutations have been shown to cause AF, generally with early onset, in a highly deterministic way (high penetrance), referred as familial AF (recently exhaustively reviewed[53]). The largest category of genes affected appear to be ion channel genes, most notably potassium channels. Rare mutations have been reported in almost every I_K, such as the ultrarapid delayed rectifier potassium current I_{Kur} encoded by *KCNA5*, an interesting target that has gathered a lot of attention given its atrial-specific expression[110]. Phase 0 I_{Na} sodium channels are also affected, including the cardiac specific *SNCA5*, also associated with ventricular arrhythmias[111]. However, most frequently, familial AF rare mutations appear to affect the sarcomeric protein Titin (*TTN*),

perhaps unsurprisingly given its central role in myocyte contraction and that it is the longest human protein[112]. Other categories of genes affected include cardiac transcription factor genes including *NKX2-5*, *PITX2*, and *TBX5*, gap junction genes such as connexin 40 and 43 (*GJA5* and *GJA1* respectively) as well as proteins that form the cellular cytoskeleton and others[53]. While illuminating, familial AF is rare and may not reflect the complex etiology of its more prevalent form occurring later in life. This more colloquial form of AF also bears strong genetic determinants, albeit with less penetrance.

1.2.1.2 Complex traits

Contrary to the colors of Mendel's peas, the expression of complex traits is multifactorial, involving the interaction of both genetic and environmental factors (phenotype = genotype + environment). These traits are considered polygenic, where multiple genes contribute a small fraction to the phenotype. The fraction of the trait's variance in a population that is due to genetic factors is deemed its heritable component (broad-sense heritability), which may differ across populations. For instance, a population with heterogeneous access to food may have a greater environmental component for height than one where it is more homogeneous. Francis Galton first exposed the bell-shaped distribution of polygenic traits, such as height. His ancestral law of heredity postulated that a trait was the sum of diminishing contributions from one's closest to most distant ancestors, such that the parents genetically contributed ½ of the phenotype, the grandparents ¼ and so on, summing to 1[113]. Height may be the quintessential example of a polygenic trait, with more than 12,000 variants now shown to contribute to the phenotype[114]. Deciphering such small effects on a phenotype can only be accomplished through large sample sizes, making the field of genetics of complex traits one of populations study. To identify the genomic loci contributing to these traits, a technology allowing to probe the entire genome in thousands of individuals would be required.

1.2.1.3 Genome-wide association studies

In the late 1990s, a method to obtain human genotype information at more than a thousand sites in parallel was published and then commercialized by Affymetrix[115]. Probes of around 25 nucleotides were designed to match specific sites in the genome that were known to be heterogeneous. Fragmented DNA would then bind to its matching probe and emit fluorescence. Imaging of these arrays, where each probe's locations on the chip is known, could then infer

thousands of genotypes in a single assay. Refinements of this technology improved accuracy and throughput, which now usually contains more than 600,000 curated markers. Today the cost to genotype an individual can be below 50$, enabling cohorts of millions of individuals.

Because of this, genome-wide association studies (GWAS) have been growing in scope since the mid-2000s. The success of GWAS was built on our ability to leverage correlations between genomic regions (haplotype blocks) to find traits and diseases associations. The linkage disequilibrium (LD) of 2 variants is a measure of their correlation determined by the frequency of recombination events occurring between them and is therefore anti-correlated with the distance between single nucleotide polymorphisms (SNPs). Without these haplotypes, each known SNP (now more than 84 million[116]) of the human genome would have to be probed independently in each individual to then be tested against a trait. Instead, using a selection of SNPs curated to represent each haplotype (tag SNPs), most genetic variations can be imputed using microarray (genotyping). In individuals of European descent, the HapMap project estimated that 94% of common variants (minor allele frequency (MAF) > 5%) could be captured with 250,000 tag SNPs[117] while the same number of tag SNPs captured only 70% common variants in individuals of African ancestry. This is due to LD blocks being of different sizes across populations. Bottlenecks, such as the out of Africa event, reduce the genetic diversity of populations that emerged from them, which has occurred multiple times in human history. Because of this, individuals of African ancestry have smaller LD blocs, which can help deconvolute loci with very high LD. Unfortunately, most GWAS studies to date have focused on individuals of European descent. This limitation of early human genetics is now the front and center issue to address in the field[118].

1.2.1.4 Fine-mapping

Fine-mapping strategies are designed to dissect patterns of association in a locus down to the causal SNP(s). Today, Bayesian fine-mapping methods are preferred to early educated guess (arbitrary SNP correlation thresholds and visual locus inspection) and penalized regression (lasso and elastic net) as they allow for the integration of a growing number of genomic annotations as priors e.g., conservation across species, DNA methylation, chromatin states, gene expression, splicing ratios and transcription factor binding sites[119]. An additional advantage of Bayesian fine-mapping is that it provides intuitive quantities on the probability of a set of variants to be causal at

a locus, the posterior inclusion probability (PIP) which sums to 1. A credible set of variants at a locus generally includes the top SNPs for which the PIPs sum to 95% at each locus. In 2020, Weissbrod et al. published PolyFun[120], a method that built on previous popular annotation informed fine-mapping methods (CAVIAR and PAINTOR)[121,122], drastically scaling the number of annotations included as priors (reaching 187). Other strategies include multi-trait fine-mapping (flashfm[123]) and leveraging multi-ancestry cohorts (MsCAVIAR[124]), which have shown a reduction of the credible set size.

Despite these advances, fine-mapping remains challenging. To deconvolute the signal, greater sample size as well as higher SNPs density are usually required to reach similar power to GWAS. Moreover, each method brings its own set of assumptions and limitations. Namely, that a set of variants is causal for multiple traits or in multiple populations, which can be false. More general assumptions include that we have information on the causal variant, which can sometimes be filtered out due to low coverage or poor imputation quality. Furthermore, not all methods allow testing for multiple causal variants at a locus.

1.2.1.5 AF GWAS

For most individuals, AF is a complex disease, with a relatively strong genetic component. Based on common SNPs, the heritability of AF is estimated at 22% in the European population[125]. As of June 2023, there were 28 GWAS on AF in GWAS catalog (https://www.ebi.ac.uk/gwas/). Importantly, there is considerable sample overlap across these studies, since the UK biobank generally accounts for most samples of the larger datasets. Nevertheless, two large AF GWAS published in 2018 by Roselli et al.[126] and Nielsen et al.[127] drastically increased the number of AF associated loci, suggesting the association of 138 independent loci[128]. Since then, another publication by Miyazawa et al. in 2023 made an important contribution with an additional 160,098 individuals of East Asian ancestry[129]. Consistently, the strongest association is found in an intergenic region about 150kb from the *PITX2* gene, conferring between 1.5 and 2-fold increased risk of AF. PITX2 (paired like homeodomain 2) is a transcription factor essential for development and establishment of the left-right axis[130]. Downregulation of *PITX2* has been associated with both pro and anti-arrhythmogenic phenotypes[131]. The causal mechanism for the association remains to be determined. As for all GWAS, most sentinel SNPs (SNP with the lowest p-value at each locus) are in non-coding regions. Only one and two were found in coding regions in Roselli et al. and

Nielsen et al. respectively, with the most predominantly affected regions being introns (at 52% and 65% respectively) and intergenic regions (at 20% and 23% respectively).

While the identification of causal genes from GWAS remains an important challenge, these datasets can be leveraged for other means, such as better risk stratification or for precision medicine. For risk stratification, combining the effect of a number of SNPs for a given phenotype allows the creation of polygenic risk score (PRS)[132]. This method can explain a larger proportion of the phenotype's heritability, as it sometimes allows for the inclusion of potentially suggestive SNPs that do not pass the GWAS threshold of $P<5\times10^{-8}$. The value of PRS combined with clinical risk factors has recently been underscored for AF risk stratification. In 4606 individuals from the Framingham Heart Study, tertile stratification of patients using both the PRS and clinical risk factors outperformed clinical risk factors alone (the incidence of AF was 22% vs 33% respectively in the low-risk group and 48% vs 43% respectively in the high-risk group)[133]. In the UK Biobank, a PRS showed that the top 6.1% of the population with the highest PRS is 3 times more at risk of developing AF[134]. PRSs may also be useful for evaluating the effectiveness of medical interventions. According to a Korean study, individuals with a high PRS had a 2.66 times higher risk of AF recurrence after catheter ablation than the low PRS group[135]. Overall, the personalization of treatments and the management of patients at risk of AF may be improved with genetic information[134].

Although labeling causal genes with the closest gene approach may yield relatively good predictions[136,137], non-coding region do not necessarily regulate the closest gene or may regulate multiple genes. Moreover, a given locus may contain multiple causal SNPs, further complexifying this task. I will cover different strategies leveraging other modalities that aim to overcome this challenge in the sections below.

1.2.2 Transcriptomics

Our ability to read the code of life was enabled by Frederick Sanger in 1977[138]. His electrophoresis method allowed to “read” the sequence of short DNA fragments (~1000bp). This method, later called first generation sequencing, was used to sequence the first human genome during the human genome project. Such method limited the scope of studies to a small number of genes selected based on strong hypotheses (candidate gene approach). Later came gene expression microarrays, when most coding gene sequences were known, probes could be designed to

hybridize coding DNA (cDNA) reverse-transcribed from RNA and be imaged similarly to the genotyping technology. Transcriptomic microarrays allowed for high-throughput, unbiased and hypothesis free investigation of gene expression. This is still used today for its cost effectiveness but has some limitations compared to sequencing based transcriptomics (see next section).

NGS, or second generation sequencing, was developed in the early 2000s. The 454 pyrosequencing method refined and commercialized by Roche would eventually be displaced by Illumina's sequencing-by-synthesis, now allowing parallel sequencing of billions of short sequences (reads) of around 150bp in a single run[139,140]. In this process, fragmented cDNA or genomic DNA (gDNA) is first sparsely ligated onto plate-bound adapters and then amplified locally, creating "spots" of dense identical sequences. Complementary strands of de-hybridized single strand templates are then re-synthesized using fluorescent nucleotides with images captured for every base addition. Each spot provides enough redundance for an accurate fluorescence readout. For whole transcriptomes RNA sequencing (RNAseq), the resulting list of reads can then be aligned to a reference genome for quantification of gene expression. As for whole genome sequencing, reads are assembled into novel genomes. This greatly accelerated discoveries such as new species genomes but also of novel transcripts, issue of a never-before seen combination of exons, or from non-coding regions previously thought to have no function.

Third generation sequencing or long reads sequencing was Nature's 2022 method of the year[141]. While the initial technology produced sequences of low quality, impairing the adoption rate, innovations such a Pacific Biosystem's HiFi now rivals Illumina's high fidelity[142]. Long reads recently allowed telomere to telomere sequencing of the human genome[143], truly completing the journey of the human genome project. On the RNA front, long reads are appealing to improve allelic imbalance analyses, gene isoform detection and quantification.

1.2.2.1 RNAseq

For most genomics laboratories, RNAseq has now become a standard protocol to assess the effectiveness of a treatment. Counting samples from human origin only, there are now more than 1.5M RNAseq samples in the Sequence Read Archive database[144]. Variations of RNAseq can be tailored to answer specific questions such as differential gene expression (DGE) analysis, the identification of novel isoforms, the identification of transcription start site (CAGE and RAMPAGE), the identification of translated regions of mRNAs (Ribosome-sequencing; Ribo-seq)

and others[145]. Among these applications, DGE analysis is the most broadly used and meaningful in this research. I will discuss important considerations to be made from the experimental design through bioinformatic analyses.

Experimental design

Important considerations should be made upfront for downstream statistical analyses. The probability to reject the null hypothesis in the presence of a true signal (power) is a function of sample size, effect size, variance, and false positive tolerance. For microarray-based assays, consideration should be made on the number of probes to test and correct for. This differs for RNAseq based on the number of genes or transcripts included in the annotation (transcript level analyses can sometimes increase the number of tests by three-fold). When positive controls are available, effect sizes and variance may be estimated with quantitative polymerase chain reaction (qPCR). Under low variance experimental settings such as cell cultures, a lower sample size may be necessary, while high variance settings such as in the human population should require greater sample size. Estimates under controlled environments and simulations suggest that between four and six samples are generally advised[146-148].

Ribosomal RNA (rRNA) can account for up to 90% of cell's RNA content[149]. To avoid over-sequencing rRNA, enrichment of mRNA is done either using oligo(dT) primers to selectively amplify polyadenylated (polyA) RNAs or using rRNA depletion can be done through enzymatic digestion or magnetic beads. Importantly, the polyA amplification is known to increase 3' bias[150,151]. This may limit downstream applications. For instance, different transcript isoforms may have the same 3' end, splicing events will not be represented evenly in the transcripts or differential exon usage away from the 3' end may be missed. Moreover, RNA species without polyA tails such as some non-coding RNAs will not be sequenced. Under these considerations, polyA amplification may be interesting to mostly limit sequencing to coding genes and reduce sequencing costs or reduce multiple test burden. Furthermore, choosing a protocol that retains strand information has been shown to yield superior results[152]. Strandedness is preserved by changing deoxythymidine triphosphate (dTTPs) for deoxyuridine triphosphate (dUTPs) during the reverse transcription step. This later allows the digestion of the complementary strand and the identification of genomic strand that produced the RNA. For antisense overlapping genes, this provides additional information increasing quantification accuracy.

For sequencing parameters, sequencing depth, read length and single vs paired end reads must be considered. Under the simplest DGE design, when only highly expressed genes with high inter-group differences are of interest, 50bp single end reads at low sequencing depth of around 5M reads could be considered[153]. Conversely, if a more thorough characterization of all transcripts with lower expression levels and fold change differences between groups is required, 150bp paired end reads at high sequencing depth of around 50-100M reads should be considered. The latter also opens possibilities for isoform level analyses and splicing events detections.

RNAseq pipeline

Sequencing data is generally provided as FASTQ files from the sequencing center. RNAseq pipelines generally consist of; i-read alignment and quantification, ii-differential expression, and iii-functional enrichment analysis. When a good reference genome is available, splice-aware alignment tools such as STAR[154] and HISAT2[155] or pseudoalignment tools such as Kallisto[156] and Salmon[157] can be used. Splice-aware alignment methods account for splicing events by allowing reads mapping at splicing junctions to be split across exons, improving mapping accuracy and sensitivity compared to predecessors such as BWA. HISAT2 was shown to be more memory efficient and more accurate than STAR, to the cost of a marked reduction in the number of aligned reads due to a higher mismatch stringency, favoring STAR for mapping sensitivity[158,159].

Next, gene or transcript quantification of the aligned reads can be done using a variety of tools such as StringTie2[160], CuffLinks[161], RSEM[162] and FeatureCounts[163]. Some of these tools will provide counts and others a normalized value such as FPKM (fragments per kilobase of transcript per million mapped reads). Stringtie2 and Cufflinks are common choices to assemble and quantify novel isoforms, when necessary, while the others will solely rely on established annotations. Superseding the alignment step, pseudoalignment is a very fast (~100 fold faster than alignment + quantification[156]) and efficient graph-based method providing direct count estimates for genes or transcripts of a provided transcriptome. These algorithms use short sequences of a length *k* (k-mers) to build a de Bruijn graph on which each read will attempt to traverse. Counts are attributed to the best matching branches, corresponding to specific transcripts. These methods compare favorably to alignment-based methods with possible advantages on specificity and sensitivity[156,164,165].

After quantification, a feature (gene or transcript) by sample matrix is used for downstream analysis. DESeq2[166], edgeR[167] and limma-voom[168] are the most widely adopted statistical models to compare expression levels between conditions or groups. Modeling gene expression using counts has been shown to yield better results than normalized values such as FPKMs[153]. Both DESeq2 and edgeR use the negative binomial distribution and generalized linear model to model counts, while limma-voom use an empirical Bayes model. Careful consideration of the covariates to include in the model should be made. Principal component analysis (PCA) is a mandatory first step that may provide insight for such a decision. Further considerations on the appropriate filters to apply for lowly expressed features may help reduce false positives or reduce the multiple test burden. In general, these methods provide similar results but dataset specificities may warrant the use of a tool over another[153,169].

Functional or pathway enrichment analyses are a common way to obtain an overview of the main cellular functions, pathways or components that may be altered between groups when there is many differentially expressed genes (DEG). The two most used strategies are over-representation analysis (ORA) and gene-set enrichment analysis (GSEA). ORA is most often used for its speed and simplicity. ORA tools usually perform a Fisher's exact test to determine if an input list of DEG is enriched in an annotated gene set representing specific cellular pathways compared to a background set of genes. As for GSEA, the input list of DEG is first ranked (based on log fold change, -log10 p-value or other metrics) and then tested against the annotation using Wilcoxon rank-sum statistic or other methods[170]. Facilitating DEG interpretations, a vast number of gene set libraries have been curated over the years, with platforms such as EnrichR hosting more than 200. Some of the most widely used include Kyoto Encyclopedia of Genes and Genomes (KEGG), Gene Onthology (GO) or Reactome. These libraries are regularly updated, and more recent versions should be prioritized. Lastly, choosing the appropriate tool and set of background gene can also significantly impact results and should be carefully considered[171].

1.2.2.2 AF differentially expressed genes in humans

As of August 2020, there were at least 24 studies conducted on human samples comparing gene expression between AF and sinus rhythm patients[172], the majority being microarrays with only three RNAseq studies. The microarray study by Deshmukh et al. stands out for its dataset size, with a total of 239 samples analysed[173]. Individuals were grouped according to the presence

of a previous AF diagnosis and their rhythm at the time of surgery i.e., AF patient in sinus rhythm (AF/SR), AF patient in AF rhythm (AF/AF) and no AF diagnosis in sinus rhythm (NoAF). Their DGE analysis on 11 806 transcript-specific probes, found ~5-fold more differentially expressed transcripts in AF/AF group than in AF/SR group when compared to NoAF (1011 vs 190 respectively) which largely confirmed the absence of differential expression of ion channel genes in the AF/SR group, and their dysregulation in the AF/AF group. Functional enrichment analyses suggested increased cellular stress in the AF/SR group compared to NoAF, with enrichment for oxidoreductase activity and transcription factor target genes involved in cardiac remodeling such as CREB/ATF and SRF. The relatively few details provided in the methods limit the interpretation of these results. Very recently, Zeemering et al. published an RNAseq study of LAA or right atrial appendage of 195 patients using a similar design[174]. Because their initial quality control (QC) indicated a dominant effect of heart failure, they further stratified the DGE analyses using this variable. No difference between paroxysmal AF patients and controls were found. For persistent AF, 35 genes were DEG with and without heart failure. There was little concordance in pathways enriched for patients with and without heart failure. Most notably, the researchers found a coalescence of robust persistent AF DEG at the *IFNG* locus including *IFNG, IL26*, *IL22* and *MDM1*. Another study looked at the difference between LAA and PV junction in persistent AF and sinus rhythm controls[175]. Interestingly, *PITX2* had higher expression levels in the PV compared to the LAA (both at protein and RNA level) but did not change with AF. Higher remodeling in the LAA was suggested since oxidative stress and fibrosis pathways were higher.

Taken together, AF transcriptomic studies commonly involved functional enrichment in ion channels and contractility dysfunction, inflammation, oxidative stress and fibrosis[176]. On the other hand, there is a markedly low congruence of DEGs among these studies. An assessment of DEGs across studies by Victorino et al. showed that only 10 DEG in LA and only 2 DEG in the RA replicated in 3 studies, none of which were ion channels. Among these genes we find *NPPA* and *NPPB*, known to be induced upon cardiac stress[177], and *RGS6* and *COLQ*, both involved in regulating parasympathetic response[178,179]. The overall lack of reproducibility across these studies underlines the challenges of transcriptomic studies in humans, which have highly heterogeneous environmental exposures and genetics, and reinforces the need for greater sample sizes to ensure reproducibility.

1.2.2.3 Trash or treasure? The mystery of the non-coding genome

The rather arbitrary definition of a coding gene was limited to open reading frames coding for at least 100 amino acids (aa), bringing the number of estimated coding genes to around 25,000 in the human genome. The rationale for this threshold was that shorter sequences would be unlikely to produce functional protein structures due to the low complexity of the sequence, while conveniently reducing the genomic search space. The inexplicably high proportion of non-coding DNA in the human genome (98.5%), and other eukaryotes, was famously touted as "junk" DNA by Susumo Ohno in 1972[180]. Along with the discovery of novel gene isoforms, NGS provided irrefutable evidence of transcription in non-coding DNA regions. Djebali et al. showed that at least ¾ of the genome was transcribed at some point in time[181]. Novel RNA species were discovered such as small nuclear RNA, circular RNA, PIWI-interacting RNA, microRNA (miRNA), long non-coding RNA (lncRNA) and several others[182], the most numerous and the most studied for their effect on cardiovascular diseases being miRNAs and lncRNAs.

The role of miRNA as a regulator of gene expression in AF has been a particular focus in recent years. These small RNAs of about 22 nucleotides sequester or induce the degradation of messenger RNA (mRNA) depending on their binding affinity, frequently at the 3' end of the mRNA. More than 17,000 miRNAs (around 50,000 today) from 140 species in 2010 have been identified and grouped in the miRBase database[183,184]. Furthermore, it is estimated that around half of mammalian mRNAs have conserved regions allowing their regulation by miRNA[185]. Prediction of miRNA mRNA targets can be obtained from multiple databases using tools such as miRNAtap[186].

In both humans and mice, studies have shown that the dysregulation miRNAs such as miR-1, miR-328, and miR-21 could be sufficient to cause AF or to normalize rhythm by restoring their endogenous levels[187,188]. Their dysregulation has been associated with factors exacerbating AF, including the level of expression of ion channels impacting I_{Ca} and I_K, the expression of gap junction proteins, and signaling via the ERK/MAPK pathway promoting the differentiation of myofibroblasts and atrial fibrosis[188].

lncRNAs form a rather heterogeneous category of RNAs with more than 200 nucleotides in length and no open reading frame longer than 100aa. LncRNAs can be intergenic, intronic or antisense to a coding gene (other genomic categorizations exist albeit with some inconsistencies).

Some of the main obstacles to the annotation of these RNAs are their generally low expression levels, their irregular splicing rate, their low stability, their low conservation rate, and irregular polyadenylation. The main functions of lncRNAs currently reported are as i-transcriptional modulators, ii-splicing modulators, iii-translational regulators, either by sequestering miRNAs or mRNAs, or by modulating mRNA stability, iv-precursors of microproteins (open reading frames <100aa) or as v-extracellular messengers[189]. Adding to the many challenges in annotating lncRNAs, most of these transcripts seem to have a functional effect in a specific cell-type. SJ Liu et al. 2017 reported that among 499 lncRNAs influencing cell growth, 89% had a specific effect in only one of 7 cell lines[190].

Studies assessing the impact of lncRNAs on AF have so far been limited in scope. Using a mouse model of AF, a group found that the lncRNA KCNQ1OT1 induced the translation of CACNA1C (calcium voltage-gated channel subunit alpha1 C, an ion channel subunit) by sequestering miR-384[191]. Other investigators found that overexpression of the lncRNA NRON, initially identified in the blood patients with heart failure, reduced fibrosis compared to controls by reducing fibroblast proliferation[192]. Interestingly, using 80 human hearts including some dilated cardiomyopathy patients, investigators found that 22% of the 783 expressed lncRNAs encoded a microprotein using Ribo-seq[193], with most localizing to the mitochondria.

While much has been learned with transcriptomic studies, they have several limitations. First, unlike genetically derived gene candidates, there is generally no evidence of causality for DEGs. Second, the dominant factor influencing RNAseq done in bulk tissues is generally its cell-type composition[194]. Batch effects are therefore very difficult to avoid and may warrant further inclusion of covariates in the model which may increase requirements in the number of samples needed to detect an effect. Result interpretation is generally limited to the whole tissue, without information on the cell-type at the source of DEG differences. To overcome this, recent methods leveraging single cell RNAseq offer ways to re-analyze bulk RNAseq and infer the cell-type influence (see section 1.2.5). Lastly, transcriptional changes may not reflect protein level changes[195,196]. Therefore, combining transcriptomic evidence with other lines of evidence is often necessary.

1.2.3 Quantitative trait loci analyses

Among the set of annotations used to infer causality, linking SNPs to gene expression is generally seen as one of the most attractive because it facilitates the interpretation of the SNP mechanism and provides an actionable target in the context of disease (even long non-coding RNAs can now be targeted using anti-sense oligonucleotides in humans[197]). While protein abundance is considered closer to the phenotype, proteomics still does not provide the same ease of use and throughput. Moreover, large differences in gene expression may only result in small phenotypic differences at the population level, increasing the power to link SNP and gene expression.

Expression quantitative trait loci (eQTL) analyses aim to find associations between gene expression and polymorphic alleles (**Fig. 8**). This is generally achieved through linear regression with tools such as Matrix eQTL[198] or FastQTL[199], but other models have been proposed such as linear mixed models and non-linear models[200]. QTLs are generally classified as *cis* (proximal to the gene and expected to have a direct mechanism of action) or *trans* (distal to the gene and expected to have an indirect mechanism of action, but rare direct actions have been suggested[201]). *Trans*-eQTLs may occur within and across chromosomes. In addition to considerations discussed in the RNAseq section, restriction of the genomic window to create SNP-gene pairs is an important parameter to determine, testing all genes and SNPs otherwise resulting in billions of tests. While there is evidence of eQTLs occurring at distances greater than 1Mb, studies consistently show an exponential decay of SNPs regulatory potential against distance with >90% of lead eQTLs found within 100kb of the gene body[201,202]. Conveniently, the QTL framework is conceptually applicable to other modalities such as protein abundance (pQTL), chromatin accessibility (caQTL), DNA methylation (mQTL) and others[200] but these modalities are more rarely used.

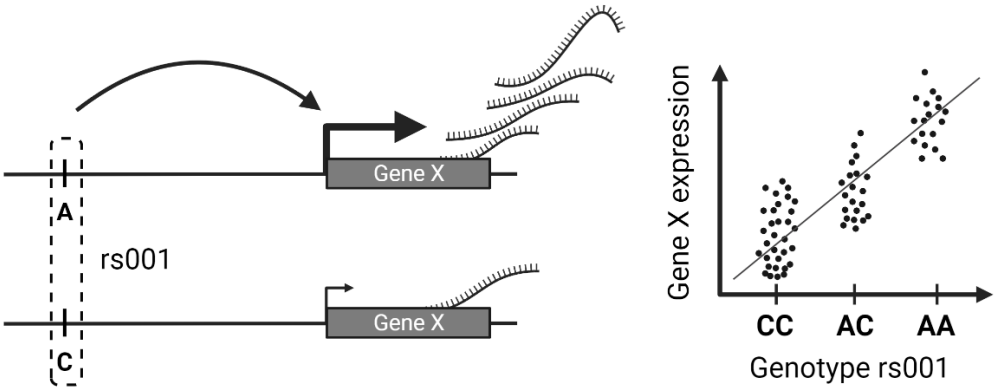


Figure 8. Schematic representation of a *cis*-eQTL.

The arched arrow represents an activation of gene X transcription through the recruitment of a transcription factor at rs001. Created with BioRender.

The genotype-tissue expression (GTEx), lunched in 2010, remains an invaluable resource, cumulating gene expression on 44 human tissues and paired genotype data in 449 individuals[203]. Other recent efforts have superseded this database in number of participants, namely eQTLgen[201] and eQTL Catalogue[204]. Strikingly, almost all (80% of protein coding genes and 67% of lncRNAs[205]) genes have eQTLs (eGenes). These figures could even be an underestimation as others have shown that bulk eQTLs have lower power to detect cell-type specific eQTLs[206,207] and that single cell data finds a greater proportion of cell-type specific eQTLs[208]. Moreover, eQTLs may be condition specific. The ubiquity of eQTLs complexifies the attribution of causality for a given GWAS locus to an associated eGene. On the other hand, Gamazon et al. have shown sticking enrichments (median of 1.7 and a maximum of 6-fold enrichment for height) in GWAS of complex traits (n=18) for the strongest eQTL of each gene in GTEx data v6[209]. Moreover, the same study showed that up to 35% of the heritability of these traits were captured by eQTLs in a multi-tissue analysis. While attractive for its simplicity and interpretability, attributing a causal gene to a GWAS locus often requires further evidence due to LD and the sheer abundance of eQTLs. Colocalization is a statistical method (generally Bayesian) that evaluates the likelihood of two signals coming from different measurements to have the same causal SNP(s). Many tools are available to conduct colocalization analyses using summary statistics such as COLOC[210], MCOLOC[211] and others[212,213], some allowing for the colocalization of more than two measurements or for more than 1 causal SNP at a locus. This can provide evidence that the GWAS and eQTL signals at a locus are shared and that the associated eGene is causal. Transcriptome wide association study (TWAS) provides an alternative to colocalization. Here, a reference panel of paired gene expression and genotype data is leveraged to impute gene expression from larger GWAS cohorts. TWAS has the advantage of integrating all genetic contributions, including sub-genome-wide significant signals, to a gene's expression instead of assessing each GWAS and eQTL signals independently. While TWAS or colocalization offer interpretable results (list of genes associated with the GWAS trait), the prioritized genes are not necessarily causal, in part due to single variants frequently being eQTLs for multiple genes[214]. Mendelian randomization can further complement these methods by providing inference of causality between the variants

(instrumental variable) and the trait (outcome) through an exposure (such as gene expression) with the condition that the instrumental variable's effect on the outcome is exclusively mediated through the exposure. The other key assumptions are the independence of association of the instrumental variable and the absence of association with confounders.

1.2.3.1 AF eQTLs

Initial eQTL AF studies used smaller sample sizes or conducted replication of GWAS results using a subset of probes or candidate genes. To date, two studies from the Ellinor group provided the bulk of insights gained from AF eQTLs. In 2014, they performed a meta-analysis of 16 AF GWAS with a replication in two cohorts of European and Japanese ancestry and reported eQTLs in LA for AF SNPs with *GJA1* (connexin 43, the dominant gap junction connexin in the heart) and *TBX5*, while *CAND2* was only an eGene in skeletal muscle[215]. In 2018, the same group conducted a transcriptome wide eQTL analysis (profiling SNPs located 250kb from the gene's transcription starting site (TSS)) of LAA from 235 and 30 individuals of European and African ancestry respectively[216]. They found 15,906 eGenes (~66% of tested genes) and 12 eQTL from AF GWAS SNPs (with the eGenes *PRRX1, SNRNP27, CEP68, FKBP7, KCNN2, FAM13B, CAV1, ASAH1, MYOZ1, C11ORF45, TBX5,* and *SYNE2*). Interestingly, while most eQTLs were conserved across other tissues, their comparison of AF eQTLs in LAA and right atrial appendages (RAA) suggests that the effects of these eQTLs have greater impact in the LAA. Others have confirmed the association of *MYOZ1* and *CAV1* in RAA and post-operative AF[217,218]. A more recent study compared the effects of AF eQTLs and pQTLs[195]. Interestingly, only 32% of lead eQTLs were also pQTLs, while for all SNP-gene pairs, only 8% were common to both. The investigators further found that loci with both eQTL and pQTL enriched for SNPs disrupting TSS, while those that were eQTLs only tended to disrupt splicing sites or TF biding sites or enhancers and pQTL only were enriched for exonic regions. Their *trans*-QTL analysis further implicated *NKX2-5* target genes, which has been shown to bind multiple AF loci by others[219].

Together, these studies suggest multiple AF candidate genes, but further validations are required such as colocalization of GWAS and eQTL signals, integration of other modalities and functional validation in the causal cell-type.

1.2.4 Epigenomics

Defined as *above* the genome, (*epi*)genomics encompasses techniques aimed at studying the regulatory mechanisms of the genome. DNA is regulated, compacted, and protected by histones, together forming the major constituents of the chromatin. Looping around histones assembled in octamers (nucleosomes) takes around 150bp of DNA. The density of nucleosomes along a given string of DNA largely dictates its propensity for gene expression. Regions sparsely populated by nucleosomes (euchromatin), generally located in the center of the nucleus, favor gene expression through increased accessibility to transcription factors (TF) DNA binding motifs that recruit the necessary machinery to initiate transcription. Conversely, densely packed nucleosomes are observed in the heterochromatin, which tends to be associated with the nuclear lamina in periphery of nuclei and inhibit gene expression. Histone N-terminal tail modifications most frequently occur on lysine and arginine. Acetylation is generally considered an activation mark while methylation's effect can go both ways depending on the site and number of methyl group added. Methylation can also directly occur on DNA, generally on cytosine in regions of high GC repeats (CpG islands), which is often associated with repression of gene expression.

The advent of NGS has enabled a surge of high-throughput methods to probe the epigenome and decipher its mechanisms (**Fig. 9**). Most notably, the encyclopedia of DNA elements project (ENCODE) lunched in 2003[220] and Roadmap Epigenomics Mapping Consortium lunched in 2008[221] pioneered some of the efforts in epigenomics. As of 2019, there was 9,239 such assays from more than 500 tissues and cell-types made publicly available by ENCODE[222]. Some of the most popular methods used to probe the epigenome include chromatin immunoprecipitation (ChiP-seq), high-throughput chromosome conformation capture (Hi-C) and assay for transposase-accessible chromatin with sequencing (ATACseq), which I will briefly detail.

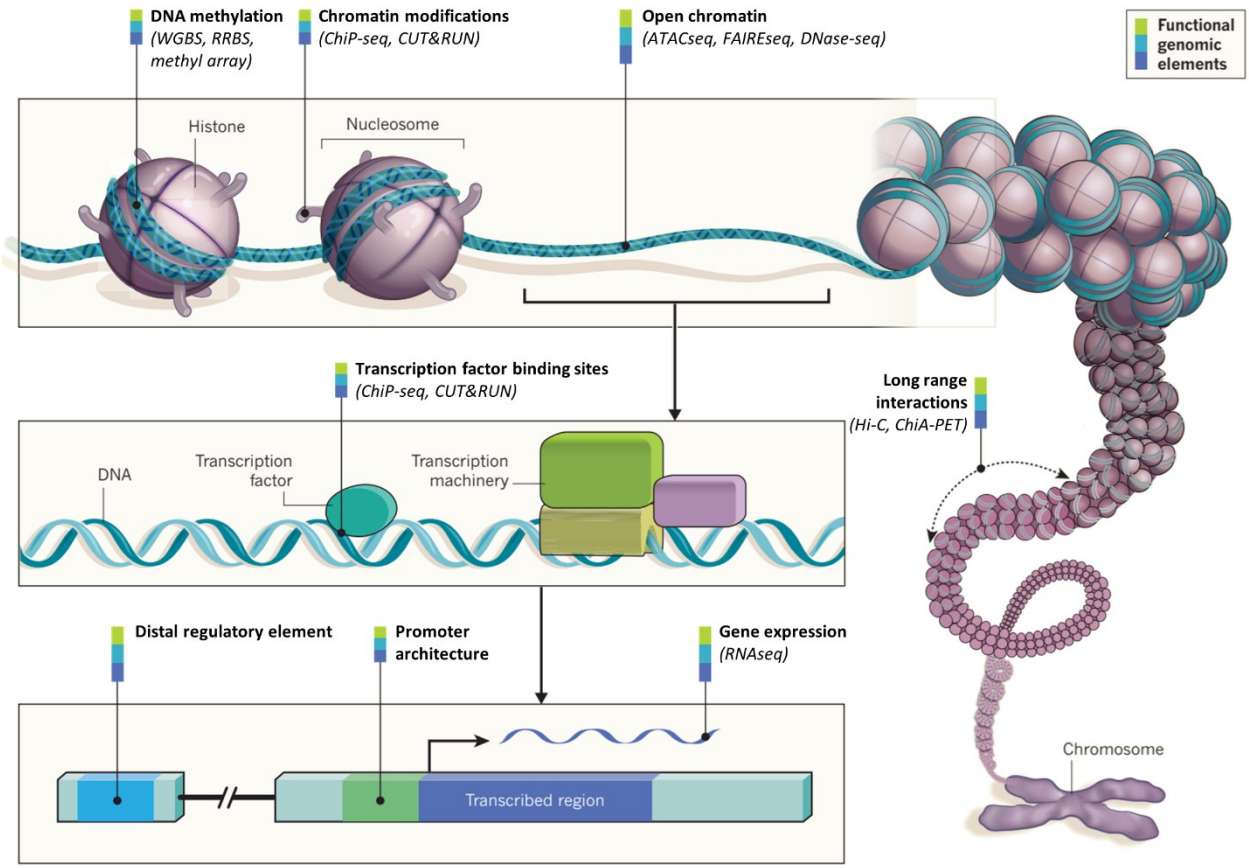


Figure 9. The landscape of epigenomic research tools.

WGBS; whole genome bisulfite sequencing, RRBS; reduced-representation bisulfite sequencing, ChiP-seq; chromatin immunoprecipitation sequencing, CUT&RUN; cleavage under targets & release using nuclease, ATACseq; assay for transposase-accessible chromatin with sequencing, FAIREseq; formaldehyde-assisted isolation of regulatory elements sequencing, DNase-seq; DNase I hypersensitive sites sequencing, Hi-C; high-throughput chromosome conformation capture, ChiA-PET; chromatin interaction analysis with paired-end tag sequencing. Adapted from Ecker et al.[223].

ChiP-seq is generally used to probe DNA elements bound by histones with specific marks such as H3K27ac (active enhancer), H3K27me3 (repression), H3K4me3 (active promoter), or transcriptional regulators such as the repressor CTCF. In this process, proteins are first crosslinked to the DNA, followed by chromatin fragmentation, immunoprecipitation using antibody for the protein of interest, protein digestion and reverse-crosslinking. DNA fragments are then sent for NGS and mapped to the genome. More recent methods can facilitate and improve data quality such as CUT&RUN[224]. Used in conjunction with RNAseq, ChiP-seq marks can be correlated with gene expression to locate promoters and enhancers. DNA motifs of TF can be inferred from their bound DNA and, with RNAseq, used to study their effect on gene regulation.

Hi-C is the culmination of chromosome conformation methods (preceded by 3C, 4C and 5C) aimed at capturing 3D chromosome-chromosome interactions. Genes can be regulated through elements hundreds of kilobases away[225]. This method breaks the linear conception of DNA. Linearly distant elements maintained in close 3D proximity by proteins can be sequenced together. This is achieved by crosslinking proteins to DNA, followed by restriction enzyme digestion of DNA, ligation to form circular chimeric DNA, fragmentation, and sequencing. Because distant elements are sequenced together, their split alignment onto the linear genome is depicted as arched links. These interactions can often occur within non-coding regions, which is less informative than promoter-enhancer interactions. To increase the resolution of this method, promoter capture Hi-C (PCHi-C) uses bait probes designed to target promoters which enriches promoter containing chimeric DNA fragments[225].

ATACseq captures regions of open chromatin deprived of nucleosomes, generally considered to activate gene expression. This method has gained tremendous popularity for its lower number of cells (nuclei are used) required and the simplicity of its protocol compared to its predecessor DNase-seq[226-228]. Open chromatin DNA fragments are retrieved using "tagmentation", where adapters are inserted in those regions with a Tn5 transposase followed by amplification and sequencing (this method is discussed in greater depth in the single-cell section 1.2.5.3 snATACseq).

The combination of these modalities has proven to be a powerful tool, stemming more than 2,000 publications in 2019 from researchers using ENCODE data[222]. Visualization of genomic loci of interest can easily be browsed online using this combined information with online tools such as the UCSC genome browser (http://genome.ucsc.edu). This is especially helpful to make predictions on the effect of non-coding SNPs or make selections of candidates for validation.

1.2.4.1 Insights into AF molecular etiology from the epigenome

In an epigenomic analysis profiling 7 histone marks in humans, researchers identified 15,545 enhancers unique to the LA[229]. These enhancers enriched for motifs related to profibrotic SMAD/TGF-β signaling and cardiac TF homeobox genes such as the PITX and NKX2 families. Using ChromHMM (a tool modeling histone marks into distinct chromatin functional states), they found that most AF SNPs were in quiescent regions, but when situated within actively regulated zones, they predominantly intersect the enhancer state (enriched for the H3K27ac mark). Another

group formulated a scoring scheme to prioritize AF genes using PCHi-C, topologically associated domains, atrial RNAseq and eQTLs[230]. With this strategy they selected *PITX2*, *DHX38*, *CAV1*, *SLK*, *TBX5*, *PRRX1*, *FAM13B* and *GJA1* as the strongest candidates at their respective loci. Others suggest the ERRg (estrogen related receptor gamma) motif as most enriched at AF loci, which was subsequently shown to be involved in calcium transient and contraction rate[129]. In a tour de force study, combining ATACseq, RNAseq and ChiP-seq of NKX2-5 from induced pluripotent stem cells derived CM (iPSC-CM) lines, Benaglio et al. provided robust evidence for electrophysiological traits modulation by NKX2-5[219]. Most notably, among the 14 SNPs associated to ECG traits by GWAS that showed allele specific NKX2-5 ChiP-seq signals, two AF eQTL loci *GNB4*[129] and *CAV1* were further validated by electrophoretic mobility shift assay, which confirmed allele specific affinity of NKX2-5 at rs7612445 and rs3807989 respectively. Lastly, although not in the context of high throughput epigenomics, altered DNA methylation, histone acetylation and filaments acetylation have been reported in AF models with some histone deacetylase (HDAC) inhibitors currently in clinical trial showing promising results as therapeutics for AF and other CVD[176,231-233].

High-throughput bulk omic analyses remain a relatively young group of technologies which continues to evolve at a rapid pace. The next frontier of developments was to transition these technologies to single cell resolution.

1.2.5 Single-cell/nucleus omics

Building on the success of NGS, developments in microfluidics and combinatorial indexing now enable most technologies used in bulk to be used at single cell resolution. There are now single cell adaptations of RNAseq, ATACseq, CUT&TAG, Methyl-seq, Hi-C and other NGS methods[234-238]. Moreover, it is now possible to probe multiple modalities at once in the same cell or nucleus[239-241]. Initial publications reported sequencing on a few hundred cells, but since 2015, the growth rate of the number of cells analyzed per study has grown astronomically, now reaching 10s of millions of cells jointly modeled[242] (**Fig. 10**). This is the result of reciprocal advances in bioinformatics and combinatorial indexing. Single cell technologies are revealing the previously underappreciated cellular heterogeneity of complex tissues. To name a few, this has facilitated discoveries of novel cell-types, cell-type specific gene expression signatures and regulatory

elements accessibility and TF activity, and condition specific cellular interactions. Together, these developments continue to push the boundaries of genomics and biological knowledge.

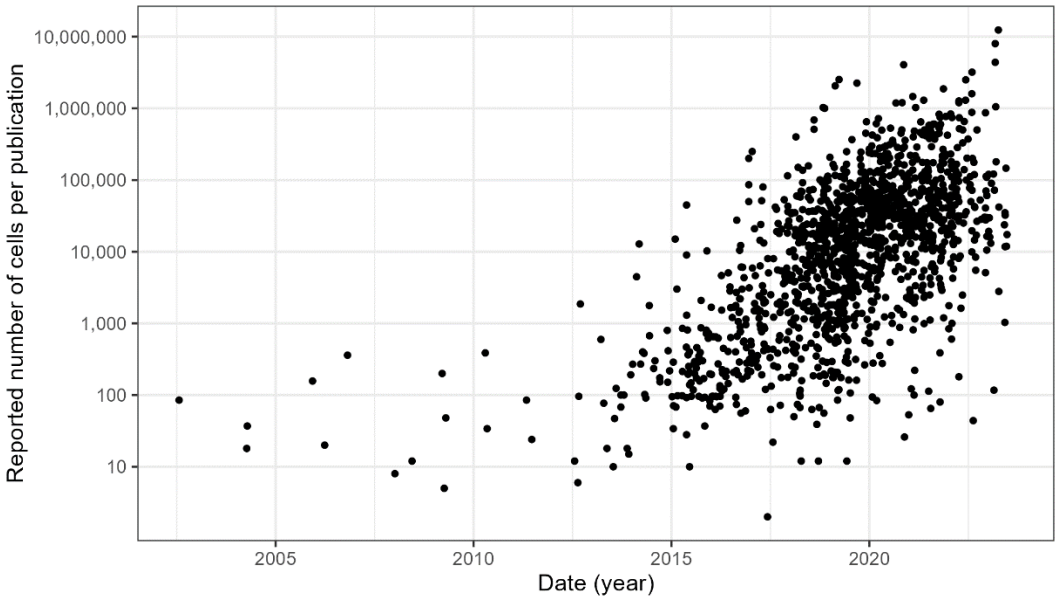


Figure 10. The growth of single cell datasets.

Data from Svensson et al.[242].

1.2.5.1 Multiome sample preparation and study design

Single cell RNAseq (scRNAseq, or single nuclei RNAseq; snRNAseq) is the most widely used single cell omic method and is currently most often done using the 10X microfluidic solution[242]. Alternatively, sequential addition of barcodes through a series of cell suspension splitting and pooling can be used to lower costs and increase throughput at the expense of increased methodological complexity and reduced gene detection sensitivity[243]. For the scope of this work, I will be reviewing aspects of the microfluidic technology focusing on the newly developed multiome assay, which entails the paired measurement of RNAseq and ATACseq modalities in the same nuclei. From study design to biological insights, the process involves multiple steps; i-dissociation of tissue and preparation of a nuclei suspension, ii-single nucleus barcoding, iii-library preparation and sequencing, and iv-bioinformatic analyses (**Fig. 11**). I will detail this process focusing on some of the key steps and study design choices having the most important consequences on the resulting dataset.

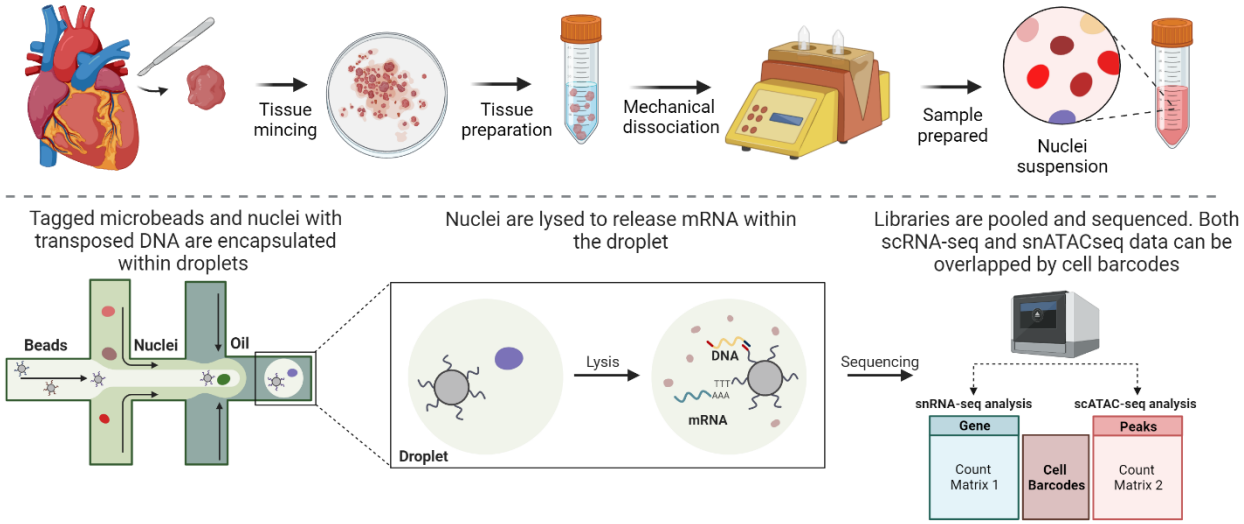


Figure 11. Single nuclei multiome.

(**top**) Schematic representation of sample preparation for the obtention of a nuclei suspension. (**bottom**) Schematic representation of lipid droplet formation in a microfluidic chromium controller and data structure obtained post sequencing. Created with BioRender.

For RNAseq, the choice of using cells or nuclei often depends on the available study material. Cells have more mRNA, less intronic RNA and will produce libraries with slightly higher complexity than nuclei, but generally yields similar results compared to nuclei[244]. On the other hand, the use of nuclei has gained in adoption for a variety of methodological advantages. While cell cultures are generally easy to dissociate into cell suspensions, preservation of cell integrity and representation during enzymatic tissue digestion requires a lot of optimization. Nuclei are generally easier to isolate, only requiring mechanical dissociation executed at low temperature (~4°C), which may also help preserve more fragile cell-types in the data[245,246]. Single-cell isolation also requires fresh tissue, while nuclei can be isolated from snap frozen tissues, broadening its range of applications. Moreover, some cells can be too large to fit into lipid droplets such as CM. Lastly, some protocols, such as single nuclei ATACseq (snATACseq), require nuclei as starting material. Because the multiome application include ATACseq and therefore requires nuclei, I will only refer to nuclei in the following steps, but generally they also apply to cells when allowed by the technology. Important steps such as reduction of extracellular debris, minimizing extra-nuclear RNA, preserving nuclei integrity and reducing nuclei clumps must be optimized during sample processing to obtain data of quality[245,247].

Once a satisfactory nuclear preparation is obtained, DNA transposition is done in bulk and nuclei are loaded onto microfluidic channels that will encapsulate into a lipid droplet, a single nucleus, a gel bead and the necessary reagents for reverse transcription. Approximately 90% of droplets are expected to be empty to reduce the chances of encapsulating two nuclei or more within a droplet (doublet or multiplet). New protocols can make use of tagged lipids or antibodies to multiplex samples and load considerably higher concentrations of nuclei in a channel while allowing downstream deconvolution of doublets[248]. Multiplexing can help reduce costs or increase the number of samples per reaction while maintaining the same number of targeted nuclei. For comparison between conditions, more samples may be an attractive choice to increase statistical power, while for rare cell-type identification, loading more nuclei per sample could be prioritized. Once the droplets are formed, nuclei are lysed to release their RNA and DNA content which is captured by the millions of oligonucleotides (oligos) contained in gel beads. For RNAseq applications, oligos are composed of a bead-specific barcode to identify specific nuclei, unique molecular identifiers (UMI) to identify individual transcripts and a ploy(dT) sequence to capture mRNAs. Given the polyA capture strategy, the same limitations apply to those discussed in the RNAseq section 1.2.2.1. For the ATACseq application the UMI and poly(dT) sequences are exchanged for a sequence complementary to tagmentation adapters. Once the mRNA is converted to cDNA, a series of PCR amplifications, size selection and adaptor ligation are done in bulk separately for both modalities and sent for sequencing.

The multiome assay is 3' sequencing based, mostly suited for gene expression quantification, but other technologies allow for full-length or 5' sequencing[247]. Detection of different isoforms is further enabled by the advent of long read sequencing in single cells[249,250]. Sequencing depth indications may vary depending on the application but a minimum of 20,000 read pairs per nucleus is generally recommended for RNAseq libraries and 25,000 read pairs per nucleus for ATACseq. Up to 90% RNAs may not be captured for sequencing through the processes mentioned above (so called dropout events), making single cell datasets very sparse[251]. Depending on the research focus, quantification of rare transcripts may require greater sequencing depth.

To process and analyze the resulting data, many tools developed for bulk RNAseq and ATACseq have been repurposed for single nucleus usage. Bioinformatic pipelines generally involve raw data processing (steps leading to the count matrices generation; read alignment, peak

calling, empty droplet calling, UMI deduplication and gene/peak quantification), preprocessing (filtering, normalization, batch correction and dimensionality reduction), clustering and cell-type annotation, and downstream analyses. The downstream analyses are diverse. I will briefly discuss the most broadly adopted ones but will emphasize the ones most relevant to this thesis i.e., DGE, deconvolution, TF activity and linking open chromatin to gene expression. Of note, up to the clustering stage, the multiome RNAseq and ATACseq modalities are generally processed independently. Thus, I will review these steps independently.

1.2.5.2 Processing and analysis of snRNAseq data

Raw data processing

Like bulk sequencing data, single nuclei raw data is generally obtained from the sequencing centers in the form of FASTQ files. RNAseq read alignment can be done using 10X Genomics software Cell Ranger[252] or open source tools such as STARsolo[253]. Pseudoalignment using kallisto | bustools[254] is another option, but an unspliced annotation including introns is necessary to adequately quantify nuclei data. These tools generally streamline all raw data processing steps. After alignment, unique barcodes must be labeled as having captured nuclear RNA or background RNA. An attempt to correct barcodes' PCR and sequencing errors is made, but when this fails to make a confident attribution to whitelisted barcodes, it is rejected. To determine if a barcode is associated with a nucleus or an empty droplet, a threshold can be set based on the total UMI per barcode. Ambiguously classified barcodes can further be filtered by comparing their gene profiles to low rank barcodes (few UMI per barcode) profiles which confidently correspond to background[252]. Ranked distribution plots (knee plots) can further inform appropriate threshold selection[255]. Subsequently, further error correction and collapsing of UMIs is done to account for PCR duplicates and replication errors. Lastly, the deduplicated UMI reads can be quantified against the reference.

Preprocessing

At this stage, operations are mostly performed on the gene by nuclei matrix. Further QC is necessary to remove doublets, low quality nuclei and/or to correct for background RNA. Nuclei containing a low number of detected genes may be empty droplets or be from dying cells, while high mitochondrial count percentage may indicate an incomplete cell lysis or clump. Inspection of

the distribution of these variables can often reveal appropriate outlier filtering thresholds. Alternatively, metrics such as the median absolute deviations[256] or dedicated tools[257] can also be used to select appropriate filtering approaches. Multiple tools can identify doublets[258]. A common strategy is to create archetypal doublets combing the gene expression profiles from two nuclei clusters and score nuclei based on their similarity to these archetypes. Currently scDblFinder appears to offer the best performance[259]. Correction for background RNA is a more difficult task which should be carefully evaluated. Every sample preparation contains extra-nuclear RNA (which can range from 3% to 35%[260]) enriched for highly expressed genes that will be contributing to each nucleus transcriptome. Various tools such as SoupX[261], DecontX[262] and CellBender[263] offer solutions to correct for the background RNA contribution in the count matrix, which can be estimated from barcodes flagged as background. In some cases, background correction may help identify rarer cell-types[264], but it can also introduce systematic biases which can influence downstream analyses[260]. Therefore, it should be used with caution. Usually, it is advisable to use initial filters and corrections of low stringency to avoid losing information as outliers can often be later identified through clustering. Lastly, to account for variations in RNA content and sequencing depths in each nucleus and enable their comparison, a variety of normalization schemes have been proposed. Conveniently, a simple scaled log transformation has been shown to provide equal or better results than more sophisticated alternatives[265].

After count matrix QC and normalization, nuclei clustering and annotation are performed. These processes are streamlined using toolkits such as Seurat[266] (R) or Scanpy[267] (Python). While multiplexing samples and splitting them across microfluidic chip channels is an attractive new option to reduce batch effect, to date, samples and channels remain confounded variables in most studies. A balance must therefore be struck between the removal of batch effects and the preservation biological variance. This spurred the development of numerous batch effect correction methods such as scVI[268] and Harmony[269] with application specific performance[270]. Clustering of nuclei in distinct groups assumes the presence of distinct cell populations. If instead a continuum is expected, a trajectory analysis is likely more appropriate. Dimensionality reduction using PCA on most informative genes usually precludes these steps. The number of informative genes and PCs can impact downstream results and should be optimized[234]. Commonly, a k-nearest neighbor graph is built, where edges are drawn between nearby nuclei in the PC space. Clusters (or communities) of user-defined resolution are then built using modularity optimization methods

such as the Louvain algorithm where clusters of increasing size are iteratively built by aggregating nearby groups of nuclei. With labeled clusters, cell-type annotation can be accomplished through an increasingly accelerated process. This is due to the growing number of expert curated datasets, now allowing for accurate cell-type predictions using tools such as Azimuth[266] or scArches[271]. Finally, visualization of the resulting clusters is facilitated by methods such as uniform manifold approximation and projection (UMAP)[272] and t-distributed stochastic neighbor embedding (t-SNE)[273] which attempt to further reduce the PC space to a human readable 2 or 3 dimensions. While informative, these methods distort the multidimensional space to accommodate lower dimensions and should not be relied upon for results interpretation.

Downstream analyses

Options for downstream analyses are expanding rapidly with the growing number of modalities and datasets that can be co-leveraged. Of special interest to the work outlined here are single nuclei DGE and deconvolution methods, but cell-cell communication, differential cell-type composition, gene regulatory network as well as other types of analyses[234] are other options that can be contemplated by investigators.

The appropriate model to use for DGE analyses in snRNAseq remains a hot topic[274-277]. While it is generally less problematic to compare gene expression between cell-types (because these differences are usually large), the smaller differences observed between conditions can be more difficult to capture. The widely used Wilcoxon rank-sum test is not appropriate as nuclei from the same sample are not independent. To account for this bias, nuclei-level models such as MAST attempt to model dropouts while allowing the inclusion of other covariates such as sex[278]. Alternatively, models used for bulk RNAseq can be used by aggregating counts by sample and cell-type (so called pseudobulk). Overall, pseudobulk appears to perform well in most cases with the added benefit of being fast and easy to implement[279]. Pathway analyses are a common subsequent analysis done on the DEG list, mostly using similar tools as for bulk RNAseq (see the RNAseq section 1.2.2.1).

The growth of snRNAseq illuminated cell-type specific gene signatures. These signatures can be used to estimate cell-type proportions of millions of publicly available bulk RNAseq samples[144]. So called deconvolution tools such as CIBERSORTx[280] and MuSiC[281] show consistently good results across a broad range of conditions[282]. While these methods work well

when a few cell-types with distinct signatures are included, they have poor performance to predict the abundance of cell states that are more correlated. A few tools[283,284] have attempted to resolve this issue, but this remains an area requiring more research.

1.2.5.3 snATACseq

While snRNAseq provides meaningful cell-type specific gene expression information, it fails to capture the intricacies of non-coding regulatory elements, where most GWAS SNPs are located. snATACseq has now enabled the discovery of millions of cell-type specific open chromatin regions (hereafter simply referred as peaks), across the human body (and other species[285,286]) and developmental stages[287]. Many steps within snATACseq bioinformatic pipelines are conceptually akin to snRNAseq such as the read alignment, batch correction, clustering and differential expression. However, there are key differences to understand such as the usage of fixed-width genomic bins vs peak calling, differences in QC metrics, different usage of dimensionality reduction methods and different downstream applications such as TF activity inferences.

Cell Ranger, ArchR[288] and Signac[289] are commonly used tools streamlining the analysis of snATACseq. Contrarily to gene expression, what defines a feature is less apparent. Two main strategies are employed to define features, either establishing fixed-width genomic bins (~500bp) or selecting specific genomic regions of variable widths that show signal to noise enrichment (widths can range from a few thousand bp to ~200bp). Each strategy has its own set of advantages. Using fixed-width bins facilitates their comparison, integration across samples and can speed up processing[288], while using variable-width peaks creates less features, less sparsity and can often be used similarly as the gene by nuclei matrix. Once the features are defined, fragments (DNA sequences corresponding to transpositions) are quantified against the newly created reference to create the peak by nuclei matrix. For nuclei QC, different metrics are used such as the ratio of mitochondrial fragment, fragment falling within peaks vs the rest of the genome, the enrichment of fragments around TSS (TSS enrichment score) and the presence of nucleosome periodicity.

Once a first filter is performed, dimensionality reduction is executed. Given that there are only two copies of DNA that can produce transposed fragments as opposed to potentially thousands of RNA copies, snATACseq is even more sparse than snRNAseq. For this reason, alternative methods that came from text search research such as latent semantic indexing

(implemented in ArchR and Signac) can better summarize peak information by importance and outperforms PCA[290]. Then, clustering can be done similarly to snRNAseq. Post-clustering, a second peak calling step performed on a per-cluster or cell-type basis can help identify peaks specific to less abundant cell-types[289]. To annotate nuclei, a common strategy is to infer gene expression by summing all counts overlapping each gene[291], but in the case of the multiome assay, the paired gene expression modality can conveniently be used.

Lastly, snATACseq offers different downstream analysis possibilities such as the inference of TF activities through enrichment of their motifs. This can provide insight as to which TF may be a key regulator of cell-type identity or modulator of pathological remodeling. From databases such as JASPAR[292] and CisBP[293], curated motifs derived from ChiP-seq experiments can be queried and quantified across peaks. Many tools can then compute motif enrichments[226], either through contrasting groups of cells by comparing motifs occurrences in peaks with different accessibilities across cell-types or conditions, or at the nuclei level by computing a motif activity score for each nucleus[294].

1.2.5.4 Linking open chromatin to gene expression

Paired measurements of open chromatin and gene expression can serve as a bridge[295] to match nuclei of unimodal snATACseq and snRNAseq datasets or it can serve to improve gene regulatory network construction[296,297]. Of particular significance here, a direct correlation of peak accessibility and gene expression can be obtained. This provides an unmatched resolution of potential peak effects, which can complement other lines of evidence such as eQTLs to strengthen the selection of GWAS candidate genes associated to non-coding SNPs.

Given the recent development of the multiome assay, only a few methods to model such interactions have been proposed[288,298,299]. ArchR relies purely on peak-gene correlations to establish links. Signac attempts to correct for Tn5 insertion biases by creating a null distribution of trans-peaks matching with GC content and coverage. This method was proposed by Ma et al., which they justified by showing that correcting for these variables resulted in a tighter peak-gene distance distribution[299]. Another method called scREG derives a *cis*-regulatory scores within each cell-type by weighting counts of gene-peak pairs with distance[298]. These scores were better predictors of CD14 positive monocyte eQTLs than a simple correlation metric.

Inherent limitations in evaluating the performance of these methods include the paucity of ground truth and multiome datasets. For instance, eQTL signals do not guarantee their regulatory potential because of LD. Another limitation is the inherent sparsity of the matrices, which creates very weak correlation coefficients (very few significant links exceed 0.1)[299]. To circumvent this, some investigators have proposed to impute missing values[300] or to create "metacells"[301,302] (conceptually analogous to pseudobulk but with smaller communities of ~50-200 cells). Furthermore, it remains an open question whether such links should be evaluated across or within cell-types. More work is needed to establish sound statistical frameworks for such analyses.

1.2.5.5 AF clues from single-cell omics

The field of single cell omics remains in its infancy with relatively few studies focused on AF[303-305]. Most of what has been discovered by single cell omic studies with significance for AF comes from research focused on describing the cardiac cellular landscape[287,306-308], heart development[309], heart failure[310], myocardial infarction[311,312] and cardiomyopathies[313-315].

The cardiac cellular landscape

Most studies identify the following cardiac cells; CM, fibroblasts, endothelial cells, endocardial cells, pericytes, smooth muscle cells, myeloid cells, lymphoid cells, adipocytes, neuronal cells and mesothelial cells (epicardium). In a landmark 2020 study, Litviňuková et al. provided the most comprehensive assessment of cardiac cells to date, profiling six regions of the human heart in 14 individuals[306]. Notably, the atria appear to have less CM, more fibroblasts and more neuronal cells than ventricles (30% vs 49% for CM, 24% vs 16% for fibroblasts, 17% vs 21% for mural cells, 12% vs 7.8% for endothelial cells, 10% vs 5% for immune cells and 2.3% vs 0.6% for neuronal cells in the atria vs ventricular regions respectively). Novel atrial marker genes were identified, including *ALDH1A2*, *ROR2* and *SYNPR*. While only a few studies have profiled atrial cells, there appears to be less CM heterogeneity in the atria compared to ventricles as judged by the number of DEG and their fold changes across the reported CM states[306]. Conversely, a study in Pitx2 mutant mice showed important CM heterogeneity in the PV, within which a CM state showed pacemaker gene expression (Tbx3, Tbx18, and Shox2)[304]. The heterogeneity of fibroblasts was shown to be more complex than previously expected, while remaining relatively constant across the different chambers[306]. Commonly reported fibroblasts states include resident (quiescent) fibroblasts marked by *DCN*, pro-fibrotic/activated fibroblasts marked by the expression of *POSTN*

and *TNC*, pro-inflammatory fibroblasts marked by the expression of *CCL2* and *THBS1*, and other states (*PCOLCE2*$^{+}$, *SERPINE1*$^{+}$)[306,310,314]. Annotated neural cell may also contain some previously unappreciated cardiac glial cells[316]. These cells were shown to interact with pacemaker cells in the SN[308] and may be implicated in post-ablation AF recurrences[317].

Disease cell states

Under multiple conditions (myocardial infarction and cardiomyopathies), a similar profile of stressed ventricular CM have been reportedly more prevalent (also found in normal hearts[306]), which is marked by *ANKRD1* (increased ankyrin repeat domain 1), *XIRP2* (xin actin binding repeat containing 2), *NPPA* and *NPPB* (DEG in multiple bulk AF RNAseq[176]) in mice and humans[310,312,314,318]. AnkRD1 myocardial overexpression in mice was shown to induce sino-atrial developmental defects with progressive atrial dilation and loss of contractility[319]. *XIRP2* rare variants were found in patients with Brugada Syndrome and showed impaired conduction in knock out (KO) mice[320]. Interestingly, in dilated cardiomyopathy patients, while the CM proportions were reduced, there was no changes for fibroblast proportions despite increased fibrosis, suggesting that fibrosis may be mediated by a transition to a pro-fibrotic state rather than proliferation of fibroblasts[314].

Cell-type specific gene prioritization for AF GWAS loci

Using snATACseq, it is now possible to infer each cell-type's genetic contribution to a phenotype using cell-type specific peaks and overlapping GWAS SNPs. In 2021, Hocker et al. performed unimodal snRNAseq and snATACseq in the 4 cardiac chambers[307]. In comparing all atrial and ventricular nuclei, they noted that most peak accessibility differences occurred in CM, with enrichments for the TF motifs of *TBX5* and *GATA4* in atrial CMs. While there was a considerable overlap between atrial and ventricular peaks, more than 16,000 peaks were shown to differ in accessibility[307]. Furthermore, it was shown that AF GWAS loci are predominantly found in CM open chromatin regions[305,307] and when compared against ventricular or fetal CM the strongest enrichment was found in adult atrial CM[287]. In line with rare mutations found in familial AF patients, this confirms that CM are likely to be the main driver of AF genetic risk.

To prioritize AF SNPs, Hocker et al. fine mapped 111 AF GWAS loci and selected SNPs overlapping CM peaks. Among 38 fine-mapped SNPs found to overlap CM peaks, the investigators showed that a *KCNH2* enhancer modulated the action potential duration in human

pluripotent stem cell derived CM[307]. Another group suggested a novel candidate gene selection method leveraging single nuclei annotations[305]. In this process, a gene posterior probability score is derived from the sum of its associated weighted fine mapped SNPs (weights attributed by SNP position within the gene structures, distance to TSS and evidence of peak regulation on the target gene). Using ventricular CM peaks to prioritize AF genes, they identified 46 genes with strong causal probability including known ones such as *TBX5*, *PITX2,* but also novel ones implicated in processes such as ephrin signaling and MAPK signaling.

Together, these studies exemplify how, in a narrow time frame, single cell omics can fill important gaps in knowledge. Much can still be gained from increasing the relatively few numbers of atrial samples currently available and a comprehensive single cell omics study in AF patients has yet to be done. Furthermore, profiling atrial tissue using the multiome assay may improve candidate gene selection compared to unimodal ATACseq and RNAseq matching predictions.

1.3 Research questions and thesis outline

The etiology of AF remains incompletely understood. Genetics provides causal associations directly in humans but suffers from the important difficulty to identify the effector genes as most GWAS SNPs are in non-coding regions. In addition, LD complexifies the identification of the causal SNP(s). Using a diversified omic approach can provide additional information that may allow the prioritization of SNPs through regulatory elements. Moreover, multiomic information resolved at the single cell level can deconvolute bulk signals and enhance our understanding of genomic regulatory processes involved in AF.

Hypothesis: Integrative multiomic approaches with single cell resolution can expand knowledge on AF etiology and identify novel therapeutic targets.

Principal goal: Identify novel molecular and cellular determinants of AF.

Specific objectives:

- Chapter 2: Identify DEG in early-stage AF canine models (Published[321]). My contribution to this work was to data generation:0%, data analysis:80%, redaction:50%.
- Chapter 3: Optimize statistical models used to infer regulatory potential of open chromatin regions using the recently developed single nuclei multiome assay (Published[322]). My contribution to this work was to data generation:0% (only public data was used), data analysis:100%, redaction:60%.
- Chapter 4: Identify AF GWAS candidate genes and the genomic regulatory mechanism linking them to AF SNP using eQTL and single nuclei multiome (In review at: iScience). My contribution to this work was to data generation:75%, data analysis:90%, redaction:60%.
- Chapter 5: Define robust AF DEGs and identify their cellular origins and TF regulators using a comprehensive multiomic characterization the LAA cellular landscape (in preparation). My contribution to this work was to data generation:100%, data analysis:100%, redaction:95%.

Expected impact:

- Identify novel genes, pathways and TFs implicated in AF.
- Improve the methodology involved in multiome data analysis which may benefit other areas of research.
- Delineate cell-type specific gene circuitry to inform future validation studies.

Chapter 2: Transcriptomic profiling of canine atrial fibrillation models after one week of sustained arrhythmia

Francis J.A. Leblanc, BS[1,2], Faezeh Vahdati Hassani, MS[1,2], Laura Liesinger, MS[3,4], Xiaoyan Qi, PhD[2], Patrice Naud, PhD[2], Ruth Birner-Gruenberger, PhD[3,4,5], Guillaume Lettre, PhD[1,2]*, Stanley Nattel, MD[1,2,6]*

*Share senior authorship.

Reference: Leblanc, Francis JA, et al. "Transcriptomic profiling of canine atrial fibrillation models after one week of sustained arrhythmia." *Circulation: Arrhythmia and Electrophysiology* 14.8 (2021): e009887.

Affiliations

[1]Faculty of Medicine, Université de Montréal, Montreal, Quebec, Canada

[2]Montreal Heart Institute, Montreal, Quebec, Canada

[3]Medical University of Graz, Diagnostic and Research Institute of Pathology, Graz, Austria

[4]BioTechMed-Graz, Omics Center Graz, Austria

[5]Technische Universität Wien, Institute of Chemical Technologies and Analytical Chemistry, Vienna, Austria

[6]Institute of Pharmacology, West German Heart and Vascular Center, Faculty of Medicine, University Duisburg-Essen, Essen, Germany

Correspondence to: Guillaume Lettre, PhD or Stanley Nattel, MD; 5000 Belanger St. E., Montreal, Quebec, Canada H1 T1C8. **guillaume.lettre@umontreal.ca or Stanley.Nattel@icm-mhi.org**

Journal Subject Terms: Arrhythmias, Atrial Fibrillation, Electrophysiology, Catheter Ablation

2.1 ABSTRACT

Background: Atrial fibrillation (AF), the most common sustained arrhythmia, is associated with increased morbidity, mortality, and health-care costs. AF develops over many years and is often related to substantial atrial structural and electrophysiological remodeling. AF may lack symptoms at onset and atrial biopsy samples are generally obtained in subjects with advanced disease, so it is difficult to study earlier-stage pathophysiology in humans.

Methods: Here, we characterized comprehensively the transcriptomic (miRNAseq and mRNAseq) changes in the left atria of two robust canine AF-models after one week of electrically-maintained AF, without or with ventricular rate-control via atrioventricular node-ablation/ventricular pacing.

Results: Our RNA-sequencing experiments identified thousands of genes that are differentially expressed, including a majority that have never before been implicated in AF. Gene-set enrichment analyses highlighted known (e.g. extracellular matrix structure organization) but also many novel pathways (e.g. muscle structure development, striated muscle cell differentiation) that may play a role in tissue remodeling and/or cellular trans-differentiation. Of interest, we found dysregulation of a cluster of non-coding RNAs, including many microRNAs but also the *MEG3* long non-coding RNA orthologue, located in the syntenic region of the imprinted human *DLK1-DIO3* locus. Interestingly (in the light of other recent observations), our analysis identified gene-targets of differentially expressed microRNAs at the *DLK1-DIO3* locus implicating glutamate signaling in AF pathophysiology.

Conclusions: Our results capture molecular events that occur at an early stage of disease development using well-characterized animal models, and may therefore inform future studies that aim to further dissect the causes of AF in humans.

Keywords: Atrial fibrillation, Atrial remodeling, Canine models, Transcriptomics, miRNA targets, Glutamate signaling, *DLK1-DIO3* locus, *MEG3*.

2.2 INTRODUCTION

Atrial fibrillation (AF) is the most common sustained arrhythmia, with an estimated lifetime risk of 22%-26% and association with increased morbidity and mortality[323]. Despite advances in antiarrhythmic therapies, their suboptimal efficacy and adverse effects have limited their use[324]. Therefore, there is a need to further characterize fundamental arrhythmia mechanisms in order to discover new therapeutic targets[324]. Although AF is known to be a final common endpoint of atrial remodeling resulting from a variety of heart diseases, it can also be, in turn, a cause of remodeling. This vicious cycle is called "AF begets AF"[325] and explains the progressive nature of this arrhythmia and the complexity of its management.

Atrial remodeling is characterized by ion channel dysfunction, Ca^{2+} handling abnormalities, and structural changes, which result in AF induction and maintenance[326,327]. Heart disease, and even rapid atrial activity itself, cause the development of atrial fibrosis, which is a hallmark of structural remodeling. The degree of fibrosis is positively correlated with the persistence of AF[328]. Atrial cardiomyocytes subjected to rapid activation release factors that induce fibroblast-to-myofibroblast differentiation that leads to increased collagen synthesis[329].

Any arrhythmia causing a rapid ventricular rate, including AF, is a well-recognized inducer of ventricular dysfunction, so-called "arrhythmia-induced cardiomyopathy"[330]. Heart failure enhances atrial stretch and sympathetic tone, making AF more resistant to rate- or rhythm-control treatments[331]. AF promotion results from the rapid atrial rate, but rapid ventricular rates due to inadequate rate-control also promote AF-related atrial remodeling with a different profile from the remodeling produced by rapid atrial rate alone[332]. Radiofrequency atrioventricular node ablation (AVB) with right ventricular pacing is a nonpharmacological strategy for rate control that can improve symptoms and outcomes[333].

Our previous work in canine AF models showed that maintaining AF for one week by rapid atrial pacing activates fibroblasts, collagen gene expression and cardiomyocyte ion channel changes, without yet causing fibrosis[334]. Continued electrical maintenance of AF for 3 weeks produces fibrosis, but electrically-maintained AF with ventricular rate control through AVB produces less profibrillatory remodeling than 3 weeks of AF alone[332]. However, how AF with and without AVB impact atrial remodeling at the molecular level has not yet been assessed comprehensively. To answer this question, we took advantage of our well-characterized AF dog

models and performed RNA-sequencing (RNA-seq) of cardiomyocyte-enriched atrial samples after one week to capture the molecular actors of atrial remodeling. In comparison with control (CTL) dogs, we found thousands of mRNAs, long non-coding RNA (lncRNA) and microRNA (miRNA) that are differentially expressed (DE) in the atria of the canine AF models. Pathway analyses of the transcriptomic data highlighted known biological processes, but also potential novel modulators of arrhythmia initiation which may shed new light on our understanding of AF in humans.

2.3 METHODS

All results and R code are available at https://github.com/lebf3/DogAF.

2.3.1 Canine atrial fibrillation model

A total of 18 adult mongrel dogs of either sex, weighing 18- 32 kg, were obtained from LAKA Inc and randomly assigned to control (CTL) group (n=6) and two canine AF-models (n=6/group) (**Supplemental Table I**). We selected 6 animals per group based on previously published RNAseq studies and our previous experiments with these models. Further, the number of DE genes identified in our analyses (post hoc) indicates that this sample size is sufficient to capture the main transcriptional changes that occur in the atrium of these dog models. Animals were handled in accordance with the "Guide for the Care and Use of Laboratory Animals" established by the National Institutes of Health as approved by the Montreal Heart Institute Ethics Committee (2016-47-01, 2019-47-03 for control dogs, 2015,47-01, 2018.47.12 for AF dogs).

To induce AF, animals were subjected to atrial tachypacing without (AF)[335] and with (AF+AVB)[336] atrioventricular-node ablation under 0.07 mg/kg acepromazine (IM), 5.3 mg/kg ketamine (IV), and 0.25 mg/kg diazepam (IV), and 1.5% isoflurane anesthesia. In the AF group, a bipolar pacing lead with fluoroscopic guidance was placed in the right atrial appendage (RAA). In the AF+AVB group, pacing leads were inserted into the RAA and right ventricular apex. Pacing leads were connected to a subcutaneous pacemaker implanted in the neck (right side). In the AF+AVB group, radiofrequency catheter ablation was used to create AF+AVB. For this purpose, a quadripolar catheter with fluoroscopic guidance was placed across the tricuspid valve via the right femoral vein. Radiofrequency energy was then used to perform ablation when action potential at the His bundle was detected. Twenty-four to seventy-two hours after surgery, dogs in the AF

group were subjected to AF-maintaining atrial tachypacing at 600 bpm for seven days. In the AF+AVB group, RA and right ventricle were paced at 600 and 80 bpm, respectively. In animals of the CTL group, no pacemaker was inserted. No adverse event was recorded and no dog was excluded.

2.3.2 Enrichment of dog atrial cardiomyocytes

Cardiomyocytes were enriched from the left atrium (LA) with enzymatic digestion through the coronary artery-perfused Langendorff system, as previously described[337]. Briefly, dogs were anesthetized with 2 mg/kg morphine (IV) and 120 mg/kg alpha-chloralose and mechanically ventilated. Hearts were aseptically and quickly removed after intra-atrial injection of 10,000 U heparin and placed in Tyrode's solution containing 136 mM NaCl, 5.4 mM KCl, 2 mM CaCl2, 1 mM MgCl2, 10 mM dextrose, 5 mM HEPES, 0.33 NaH2PO4 (pH was adjusted to 7.3 with NaOH). The left coronary artery of the isolated heart was cannulated, and the LA was dissected free and perfused with 100% oxygenated Tyrode's solution (37°C, 1.8 mM Ca^{2+}). The arterial branches were ligated to have a leak-free system, and LA tissues were perfused with Ca^{2+}-free Tyrode's solution for ~10 minutes, followed by ~1-hour perfusion with ~0.45 mg/mL collagenase (CLSII, Worthington, Lakewood, NJ) and 0.1% bovine serum albumin (Sigma–Aldrich, Oakville, ON) in Ca^{2+}-free Tyrode's solution for enzyme digestion. Digested tissue was removed from the cannula and cut into small pieces, and atrial cardiomyocytes were harvested.

2.3.3 RNA-seq/miRNA-seq

2.3.3.1 Library preparation and sequencing

mRNA and miRNA libraries were prepared at Genome Québec. mRNA libraries were made with the NEBNext_dual kit (rRNA-depleted stranded) and sequenced on NovaSeq 6000 S2 PE100 Illumina platform generating 32-123M Paired-end reads per sample. miRNA libraries were prepared with TruSeq smRNA and sequenced on the HiSeq 4000 SR50 Illumina platform generating 10-12M reads per sample.

Bioinformatic processing and DE analysis

The complete analysis can be found at https://github.com/lebf3/Dog_AF_transcriptomic. Briefly, mRNA reads were pseudomapped on reference transcriptome CanFam3.1.98 with

Kallisto[156] with the options quant -t 5 -b 100 and the rest as default. We aggregated transcripts by genes with tximport[338] and quantified with DESeq2[339]. Genes with 0 reads in more than 12 samples were removed. Shrunken $\log_2$ transformed expression corrected for library size, with and without fibroblast fraction as a covariate (within DEseq2's model) were then analyzed for DE with Wald test for all pairwise comparisons of CTL, AF and AF+AVB and likelihood ratio test for total assay DE. We did not adjust our differential gene expression analyses for biological sex because no genes where DE between female and male dogs in our experiment. We plotted the PCA with fibroblast fraction as a covariate from $\log_2$ transformed expression values corrected with Limma's removeBatchEffect() function[340] for visualization of fibroblast effect on the top 1000 most variable genes. We then compared sets of GENEIDs found to be up (L2FC > 0 & $p < 0.01$) or down (L2FC < 0 & $p < 0.01$) in all possible contrasts.

For miRNAs, we trimmed reads using fastp[341] with default settings and aligned them to CanFam3.1.98 genome with STAR v2.7.1a[154] according to ENCODE protocol[342]. DEseq2 DE analysis was then conducted with the same parameters as described above for mRNAs.

2.3.3.2 Deconvolution of RNA-seq data

To account for potential tissue heterogeneity, we used a murine atrial gene signature matrix described in Donovan et al.[343] and our matrix of gene expression in Fragments Per Kilobase of exon model per Million reads mapped (FPKM) in CIBERSORTx online tool[344]. We then performed nonparametric Wilcoxon test on all possible comparisons for fibroblast fraction with a statistical significance threshold of $p < 0.05$.

2.3.3.3 Gene set enrichment analyses

For each gene sets described above, we performed hypergeometric testing against the human Gene Ontology (GO) Biological Processes (BP) from Molecular Signatures Database v7.1 with the HypeR package.

2.3.3.4 miRNA target prediction

For DE miRNA present in the 5 most cited miRNA databases (DIANA, Miranda, PicTar, TargetScan, and miRDB), we defined genes as targets if they were: i-annotated with a human homolog in the ensemble database, ii-predicted targets by at least 3 out of 5 databases queried with the MiRNAtap package, iii-DE (mRNA FDR < 0.01), iv-inversely correlated (Pearson's r<-0.5)

log_2 expression, corrected for the fibroblast fraction (expression values corrected with Limma's removeBatchEffect() function). We then performed a GSEA with the remaining 82 predicted targets of the miRNA located on the syntenic region of the Dlk1-Dio3 locus (CanFam3.1 Chr8:68961744-69696779) as described above.

2.3.3.5 RNA-seq and miRNA-seq DE genes comparison between human AF patients and canine AF models

DE genes in our canine AF models with annotated human orthologues in the ENSEMBL database were compared to a meta-analysis of miRNA DE in human AF and a large RNAseq study on left atrial appendages obtained from 261 patients undergoing valve surgery[216,345]. Precursors of the human miRNAs listed in Shen et al. Table S8 *(n= 53, 21 upregulated and 32 downregulated*)[25] were retrieved with the R package miRBaseConverter and then compared across species. Because only 5 miRNA were found to overlap with human DE miRNA, only mRNA genes found to be DE in human and our canine AF models are represented as an Upset plot. The counts for human mRNA data were downloaded from GEO database (GSE69890). DE testing was conducted as described above for the 3 groups; no AF (CTL, n=50), AF in AF rhythm (AF, n=130), and AF in sinus rhythm (AF.SR, n=81), with inclusion of sex as a covariate.

2.3.3.6 Mitochondrial genes DE in canine AF models

The human MitoCarta3.0[346] database was queried for genes with mitochondrial localization in the heart (n=539). We represented DE genes with likelihood ratio test FDR<0.01 from that list as volcano plots.

2.3.4 Proteomics

Dog cardiomyocytes were lysed by sonication, reduced and alkylated. Protein was precipitated, resuspended, quantified and subjected to tryptic digest. Peptides (500 ng) were analyzed by reverse phase nano-HPLC coupled to a Bruker maXis II mass spectrometer (positive mode, mass range 150 - 2200 m/z, collision induced dissociation of top 20 precursors). LC-MS/MS data were analyzed for protein identification and label-free quantification using MaxQuant[347] (1.6.1.0) against the public database UniProt with taxonomy Canis lupus familiaris and common contaminants (downloaded on 01.08.2019, 29809 sequences) with carbamidomethylation on Cys as fixed and oxidation on Met as variable modification with decoy database search included (mass

tolerance 0.006 Da for precursor, 80 ppm for product ions; 1 % PSM and protein FDR, match between runs enabled, minimum of 2 ratio counts of quantified razor and unique peptides).

2.3.4.1 DE analysis and correlation

Proteins with > 3 missing values per treatment were removed. The remaining missing intensities were replaced with random values taken from the Gaussian distribution centered around a minimal value from the 10th quantile with the *DEP* package's Minprob function, to simulate a relative label-free quantification (LFQ) value for those low abundant proteins. Two-sample t-tests with subsequent multiple testing correction by FDR were used to identify DE proteins (p<0.01) with the fibroblast fraction as covariate using the Limma package.

Because proteomic processing does not always converge to a single protein, only 755 genes out of the 1029 in the proteomic matrix were correlated to their corresponding RNA-seq data. We compared overlapping genes' mean $\log_2$ transformed expression in proteomic and RNA-seq. The distribution of mean RNA-seq expression of the 755 overlapping genes was then compared to the full mean RNA-seq gene expression values.

2.4 RESULTS

2.4.1 RNA-sequencing of cardiomyocyte-enriched atrial samples from canine AF models

We analyzed data from three groups of six dogs. The first group (CTL) was the control group without atrioventricular ablation (AVB) nor pacemaker, in the second group (AF), right atrial-tachypacing at 600 beats per minute (b.p.m.) was used to maintain AF electrically for one week, and the third group (AF+AVB) included dogs with electrically-maintained AF for one week in the presence of AVB and ventricular pacing at 80 b.p.m. to control the ventricular rate. We reasoned that transcriptomic profiling of atria from these animals should allow us to discover the molecular changes that occur over the first week after the onset of AF, and play a role in the development of the tissue remodeling accompanying the transition from paroxysmal to persistent AF.

Initial analysis of bulk RNA-seq data hinted at some heterogeneity of cellular composition across samples. Therefore, we estimated the fraction of the major cell-types in each sample using

an *in-silico* deconvolution technique implemented in CIBERSORTx (Fig. 1A)[344]. Because of the induced tissue remodeling due to the AF treatments, we found that both AF and AF+AVB dogs had more fibroblasts in their atria than CTL animals (Fig. 1B). To emphasize the transcriptional differences between conditions that are not a result of variable cellular composition, we included the fibroblast fraction as a covariate in all subsequent DE analyses. Correction for this confounding variable reduced inter-group variability (Fig. 1C-D).

2.4.2 Proteomic analysis largely confirms the transcriptomic results

To validate our RNA-seq results, we took advantage of mass spectrometry (MS)-based protein quantification results from the same 18 dog atrial cardiomyocyte-enriched cell extractions that were generated in a parallel study (detailed proteomic results will be presented elsewhere). After stringent quality control, we obtained relative quantification for 755 proteins. For these genes, the relative RNA and protein levels were strongly correlated (Pearson's r=0.49, P=1.57x10^{-46}) (Fig. 2A). Many of the genes that are well-correlated encode abundant cardiomyocyte proteins, such as titin (*TTN*), myosin light chain-4 (*MYL4*), desmin (*DES*), and tropomyosin-1 (*TPM1*). We found that RNA-seq could profile transcripts with a wider range of expression profiles, whereas MS-based proteomics preferentially captured proteins whose genes are expressed at high levels. (Fig. 2B).

2.4.3 Transcriptomic changes in cardiomyocyte-enriched atrial samples

Pairwise comparisons of gene expression levels between the three groups of dogs identified 434, 5971, and 7867 genes that are DE (false discovery rate (FDR) <0.01) in atrial cardiomyocyte-rich fractions in AFvsCTL, AF+AVBvsCTL, and AFvsAF+AVB, respectively (Fig. 3A-B). All differential gene expression level results are available in Supplemental Table II and https://github.com/lebf3/DogAF). Many genes previously implicated in AF are dysregulated in both AF and AF+AVB dogs when compared to controls, thus validating the experimental design. This includes *FHL1* involved in myofilament regulation[348], *SORBS2* involved in intercalated disc gap junction regulation[349], and *KCNA5*, which regulates atrial action potential repolarization[350]. Previous studies have established an important role for mitochondrial dysfunction in the etiology of AF[346]. Accordingly, we identified 54 genes that encode mitochondrial proteins that are DE in our AF canine models (Supplemental Figure I). In particular in the AF+AVB group, we noted the up-regulation of two key beta-oxidation genes (*CPT1A* and

ACADL) and the down-regulation of the electron transport chain genes *COX17* and *NDUFA8*. However, our data also implicates genes not previously recognized to be involved in AF, such as leukocyte receptor cluster member-8 (*LENG8*), transcription elongation regulator-1 (*TCERG1*), ligand dependent nuclear receptor corepressor (*LCOR*), formin-binding protein-4 (*FNBP4*), and *ENSCAFG00000049959* (orthologue of the lncRNA *MEG3*)(Fig. 3A, Supplemental Table III, and https://github.com/lebf3/DogAF).

To understand what pathways are modulated in the atria of these canine AF models, we performed gene set enrichment analyses (GSEAs) on the DE genes (Fig. 3C and Supplemental Table IV). In AFvsCTL, we noted an up-regulation of genes associated with profibrotic pathways (*e.g.* extracellular structure organization, biological adhesion, response to wounding) and a down-regulation of genes implicated in angiogenesis, such as blood vessel morphogenesis. Genes implicated in muscle biology were up-regulated in the AF+AVBvsCTL analysis (*e.g.* muscle structure development, striated muscle cell differentiation) whereas the same comparison implicated down-regulated genes involved in ion transport and signaling pathways (*e.g.* sensory perception). We confirmed that this enrichment was not due to a smaller fraction of cardiac neurons found in the atria of AF+AVB dogs (Kruskal-Wallis' P=0.32). Because of the large overlap in genes that are down-regulated in AF+AVBvsCTL and up-regulated in AFvsAF+AVB (Fig. 3B), we identified similar pathways in the GSEA for these two comparisons (in Fig. 3C, compare AF+AVBvsCTL and AFvsAF+AVB). Finally, genes that were down-regulated in the AFvsAF+AVB analysis implicated genes with more generic functions in gene expression and chromatin modifications, such as the histone-lysine N-methyltransferase *SETD5* and the DNA methyltransferase *TET2*. Dysregulation of the expression of these chromatin-related genes and pathways is consistent with the extensive transcriptomic changes observed in the atria of AF+AVB dogs, in sheep models of AF (Supplemental Figure II) as well as in AF patients[229,351].

2.4.4 Dysregulation of miRNA expression

Because miRNA play important roles in AF biology[352] but are not detected in standard RNA-seq protocols, we performed in parallel miRNA-seq on the same dog samples. We found 31, 19 and 21 miRNA that are DE (FDR <0.01) in AFvsCTL, AF+AVBvsCTL and AFvsAF+AVB, respectively (Fig. 4A, Supplemental Table II). When comparing miRNA expression in the two AF models, *MIR185* on the dog chromosome 26 was the most DE miRNA with strong up-regulation

in the atria of AF animals. We also noted that 11 of the most strongly DE miRNA in the AF+AVBvsCTL and AFvsCTL analyses (*MIR136*, *MIR411*, *MIR370*, *MIR127*, *MIR493*, *MIR494*, *MIR485*, *ENSCAFG00000025655* (96.20% identity to hsa-mir-379), *MIR758*, *MIR543*, *MIR889*) mapped to the chr8:68,900,000-69,700,000 region in the dog reference genome CanFam3.1 (Fig. 4B). This region, highly conserved in mammals, is syntenic to the imprinted 14q32 region in humans (also known as the *DLK1-DIO3* locus)[353]. The lncRNA *MEG3*, which we described above as being over-expressed in the AF canine models is also located in the same *DLK1-DIO3* syntenic dog locus.

The dysregulation of the expression of lncRNA and miRNA at the same locus suggested that they might co-regulate the expression of genes implicated in the same biological pathway(s). To address this possibility, we used in silico predictions to infer the DE mRNA that are possible direct targets of these DE miRNA located at the syntenic *DLK1-DIO3* locus. For this analysis, we focused on DE miRNAs and DE mRNAs that are predicted to physically interact by at least three out of five databases and that have expression levels that are negatively correlated in the RNA-seq/miRNA-seq experiments (Pearson's $r < -0.5$). Using these filters, we identified 82 potential target genes for the DE miRNAs at this locus, with most genes targeted by a single miRNA (Fig. 4C). GSEA with these 82 genes indicated a common role in synaptic signaling involving glutamate signaling (Fig. 4D and Supplemental Table V). Some of the key genes within these pathways are metabotropic glutamate receptor-1 and -8 (*GRM1*, *GRM8*), glutamate ionotropic receptor delta type subunit-1 (*GRID1*), glutamate ionotropic receptor AMPA type subunit-1 (*GRIA1*), and corticotropin-releasing factor-binding protein (*CRHBP*).

2.4.5 Partial differential transcriptomic overlap between human and dog AF atrial samples

To assess the ability of our canine AF models to capture early transcriptomic changes which might be missed by profiling the atria of human AF patients who have developed the disease over years, we compared the DE genes identified in AF dogs with results from human left atrial transcriptomic profiling experiments[216,345]. Hsu et al. performed bulk RNA-seq experiments on left atrial appendages from 261 patients who underwent cardiac surgery to treat AF, valve disease, or other cardiac disorders. For differential gene expression analyses, these patients were divided between no AF (n=50), AF in AF rhythm (n=130), and AF in sinus rhythm (n=81).

When we intersected this list of human DE genes with the list of DE genes in our AF dog models (with clear human orthologs), we identified 668 genes (Fig. 5 and Supplemental Table VI). We found the strongest overlap between dog AF+AVB and human AF in AF rhythm. Of note, most of the strongest signals in our dog study are also present in this human study (*LENG8*, *SORBS2*, *BMP10*, *FNBP4* and glutamate receptor-related genes (*GRM1*, *GRM8*, *GRIA1*, *GRIK2*, *GRID1*)). For miRNA, we compared our results with data from a large meta-analysis involving 40 articles and 283 DE miRNA in AF (in different tissues and species)[345]. Of the 53 AF-associated miRNAs that were previously identified in human cardiac tissues and showed consistent results in the meta-analysis, we found five miRNAs in our analyses of the dog transcriptomic datasets (*MIR144*, *MIR142*, *MIR146B*, *MIR223* and *MIR451*). These miRNAs have not been characterized functionally yet for a role in AF. Generally, the canine AF group matched better the directionality of change reported in the human AF DE miRNA meta-analysis.

2.5 DISCUSSION

In this study, we used a transcriptomic approach to comprehensively assess the molecular architecture of AF-induced remodeling with and without AVB from atria cardiomyocyte-enriched samples. We validated the robustness of our RNAseq data by correlating it with a proteomic analysis, which showed a strong correlation in gene expression, with well-defined cardiomyocyte genes being most highly expressed in both datasets (e.g. TTN, MYL4). We confirmed the involvement of known AF factors like the reactivation of developmental pathways, but also found a strong and novel association with microRNAs and lncRNA from the *DLK1-DIO3* locus, including the *MEG3* canine orthologue. This finding is concordant with the many chromatin remodeling genes dysregulated in our models, which is an emerging phenotype of AF both in human and sheep models[354].

2.5.1 Molecular remodeling in AF with versus without AVB

We did not expect to find a smaller number of DE genes in the AFvsCTL analysis, given our previous observation that AF treatment alone without AVB results in more important tissue remodeling[332]. One possible explanation is that our prior histological studies were done in AF animals treated for three weeks,[10] whereas the results presented here reflect RNA changes after

one week of AF. The transcriptomic changes in the AF+AVB group show that cells are under active chromatin modification, indicating ongoing adaptation to the stimulus. This is not observed in the AF group (lacking AVB), which may indicate that this adaptation has already occurred. This idea would be consistent with the down-regulation of chromatin-related genes recently noted in the atria of sheep AF models[354], and may be a result of earlier establishment of profibrotic transdifferentiation in AF compared to AF+AVB canine models.

Our analyses highlighted many genes not previously implicated in AF. While the functions of some of these genes remain uncharacterized (*e.g. LENG8* in AF+AVB), we can speculate on the activities of others. For instance, *TCERG1* and *FNBP4*, which are up-regulated in the atria of AF+AVB dogs, encode co-regulated proteins that are involved, respectively, in RNA splicing and translation[355]. The up-regulation of these genes, with general actions on gene transcripts, may (partly) explain why more genes are dysregulated in AF+AVB animals when compared to the CTL or AF groups (**Fig. 3B**). Another interesting candidate is *LCOR*, which is up-regulated in the AF+AVB model and encodes a transcriptional cofactor that interacts with PPARγ and RXRα to control gene expression[356]. While further experiments are needed to determine the extent by which *LCOR* modifies gene expression in AF+AVB and contributes to the pathology, our RNA-seq experiments detected the dysregulation of two of its likely targets based on the literature: *CPT1A* (see above) and the cell cycle regulator *CDKN1A*[357] .

2.5.2 Potential role of non-coding genes at the *DLK1-DIO3* locus in early AF

The highly-conserved *DLK1-DIO3* locus hosts two differentially DNA-methylated regions modulating the expression of its non-coding RNA clusters, where in humans the maternal allele is hypomethylated with concomitant expression on the hypermethylated paternal allele of non-coding RNA and other protein-coding genes (*DLK1*, *RLT1*, and *DIO3*)[353]. In both AF+AVBvsCTL and AFvsCTL, we found a large proportion (58% and 23%, respectively) of DE miRNA at this locus, underlying its importance in AF-related adaptation. We also found dysregulation of the *MEG3* lncRNA canine orthologue at this locus. *MEG3* is a highly expressed lncRNA that has been studied in various pathologies, including cancer[358] and more recently cardiovascular diseases[359]. Non-coding RNAs at this locus have been shown to mediate various cardiac developmental programs[353]. More specifically, *MEG3* can contribute to the recruitment of the Polycomb repressive complex-2 (*PRC2*)[360], a key chromatin modulating factor. Of particular

interest, Mondal et *al*. showed that through interaction with the H3-Lys-27 methyltransferase *EZH2*, *MEG3* can repress TGF-beta target genes, which are known to promote a profibrotic response[361]. Data have been presented that suggest an important role of *EZH2* and/or *EZH2*-regulated genes in AF[362].

2.5.3 Glutamate receptor regulation by miRNAs from the *DLK1-DIO3* locus

Our GSEA analysis-predicted gene targets of DE miRNA at the *DLK1-DIO3* locus suggest a role for glutamate signaling in AF. Immunostaining has confirmed the presence of glutamate receptors on cardiomyocytes[363]. Glutamate was also found to be significantly increased in AF patient left atrial appendages[364]. Glutamate signaling is important in vagal afferent neurons[365], and remodeling of the glutamate system in AF may relate to the extensive previous evidence of autonomic dysfunction in AF patients[366]. Moreover, a recent study has shown fundamental roles for glutamatergic receptors in rat atrial cardiomyocytes and induced pluripotent stem cell-derived atrial cardiomyocytes, including a reduction in cardiomyocyte excitability after *GRIA3* knockdown[367]. Therefore, the *DLK1-DIO3* miRNA cluster may be an adaptative regulator of cardiomyocyte excitability or of neural cells in the presence of AF.

2.5.4 Limitations

We used cardiomyocyte-enriched samples in an attempt to obtain clearer results from the transcriptomic analysis by excluding extrinsic variability due to changes in cell composition. However, while our samples are enriched in cardiomyocytes, they do not constitute a pure cardiomyocyte preparation. A disadvantage is that variability due to changes in cell composition is not eliminated. On the other hand, our cardiomyocyte-enriched (but not pure) samples allow us to detect potential features of AF related to non-cardiomyocyte cells, such as autonomic dysregulation mediated by neural cells; however, we cannot unambiguously attribute DE genes to transcriptomic changes in a specific cell-type. In part, we were able to control for fibroblast composition by adjustment through analysis for expression of fibroblast-related RNA-expression patterns. Nevertheless, features underlined here should be confirmed in pure cell cultures or single cell transcriptomic assays. A second limitation is the difficulty in extrapolating our findings to gene expression changes in humans. We found only modest overlap of DE genes in our model compared to reported DE gene patterns in human; several factors could explain this (*e.g.* differences in biospecimen preparation, tissue heterogeneity, fundamental differences between

dog and human AF pathology). It is also possible that different transcriptomic programs may be involved at the initiation of arrhythmia and tissue remodeling (AF and AF+AVB dog models) when compared with those dysregulated in the atria once the pathology has been present for years.

We compared our transcriptomic results with proteomic data on the same samples and found a very high level of correlation (**Fig. 2**). These results are consistent with a large measure of transcriptomic control over protein expression and validate the relevance of transcriptomic analysis of these data. A further in-depth look at the proteomic signature in these models would be of interest but is beyond the scope of the present manuscript.

2.6 CONCLUSIONS

Understanding the pathophysiology of chronic human diseases such as AF is challenging because they develop over many years and initially present with only unremarkable pre-clinical symptoms. In this study, we took advantage of two well-characterized canine AF models to chart the transcriptomic changes that occur at the earlier phases of arrhythmia. Despite the inherent limitations in relating dog models to human AF, our results offer interesting new hypotheses for future testing, including in man. In particular, the up-regulation of miRNAs at the *DLK1-DIO3* locus after 1 week of AF suggests that they may be early biomarkers of tissue remodeling and/or adaptation in the atria.

2.7 DECLARATIONS

2.7.1 Sources of Funding

This work was funded by the Fonds de Recherche en Santé du Québec (FRQS), the Canada Research Chair Program, the Montreal Heart Institute Foundation (MHIF), Heart and Stroke Foundation of Canada (grant #18-0022032), Canadian Institutes of Health Research (CIHR) (grant # 148401), the Austrian Science Fund (FWF) (projects KLI645, W1226 and F73 to RBG). F.L. received scholarships from the CIHR, FRQS, MHIF and Université de Montréal.

2.7.2 Disclosures

None

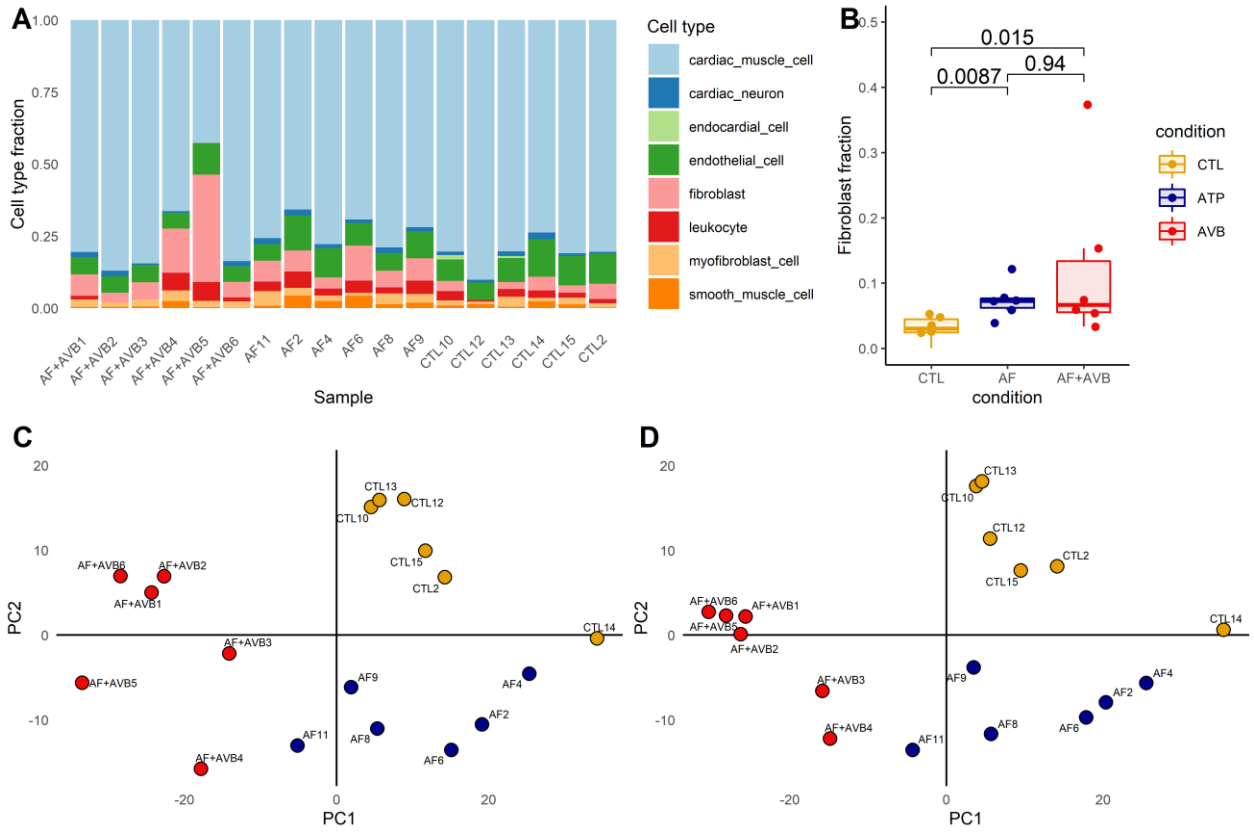


Figure 1. Deconvolution of canine atria cell composition using bulk RNA-sequencing.

(**A**) We inferred cell fractions with CIBERSORTx and an atrial-specific gene signature matrix obtained using orthologous murine genes[343]. We present cell fractions for each dog sample that we analyzed in this study. CTL, control; AF, Atrial-tachypacing; AF+AVB, AF with Atrio-Ventricular Block. (**B**) When we group animals per treatment arm, we observed a significantly higher fraction of fibroblasts in the atrial fibrillation dog models (AF and AF+AVB) than in the control animals (AFvsCTL Wilcoxon's test P=0.0087 and AF+AVBvsCTL P=0.015). Principal component analysis of the top 1000 most variable genes expressed in canine atria before **(C)** and after **(D)** correction for fibroblast fraction show treatment-dependent clustering after correction for cell composition.

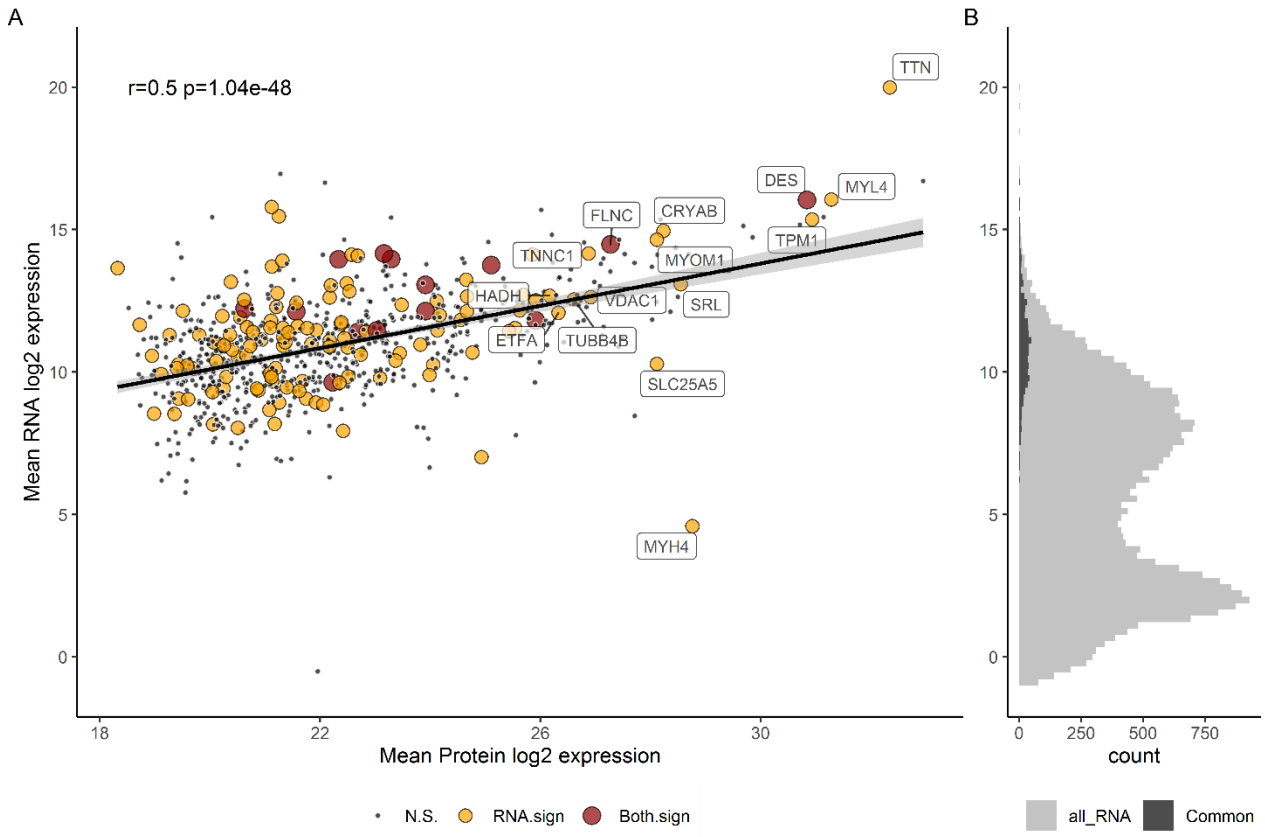


Figure 2. Validation of highly expressed RNA by proteomics.

(**A**) In 18 atrial samples, 755 genes (N.S.= 619, RNA.sign=122, Both.sign=14) found in both datasets are highly correlated at the protein (x-axis) and RNA (y-axis) levels (Pearson's r=0.49, P=1.57x10^{-46}). For reference, we annotated 15 genes that are differentially expressed in the RNA-seq experiment and have high protein expression levels. N.S., not differentially expressed in the RNA-seq or proteomic experiment; RNA.sign, genes that are differentially expressed in the RNA-seq assay only; Both.sign, differentially expressed genes in both the RNA-seq and proteomic experiments. DE genes in the RNAseq dataset have an FDR < 0.01 (likelihood ratio test) and proteomics dataset an FDR < 0.05 (F-test). The grey area around the line corresponds to the 95% confidence interval. (**B**) Relative expression level of all transcripts measured in the RNA-seq experiment. The histogram shows that genes that are present in both the RNA-seq and proteomic experiment are highly expressed (Common, dark grey) in comparison to the expression levels of all transcripts measured (all_RNA, light grey).

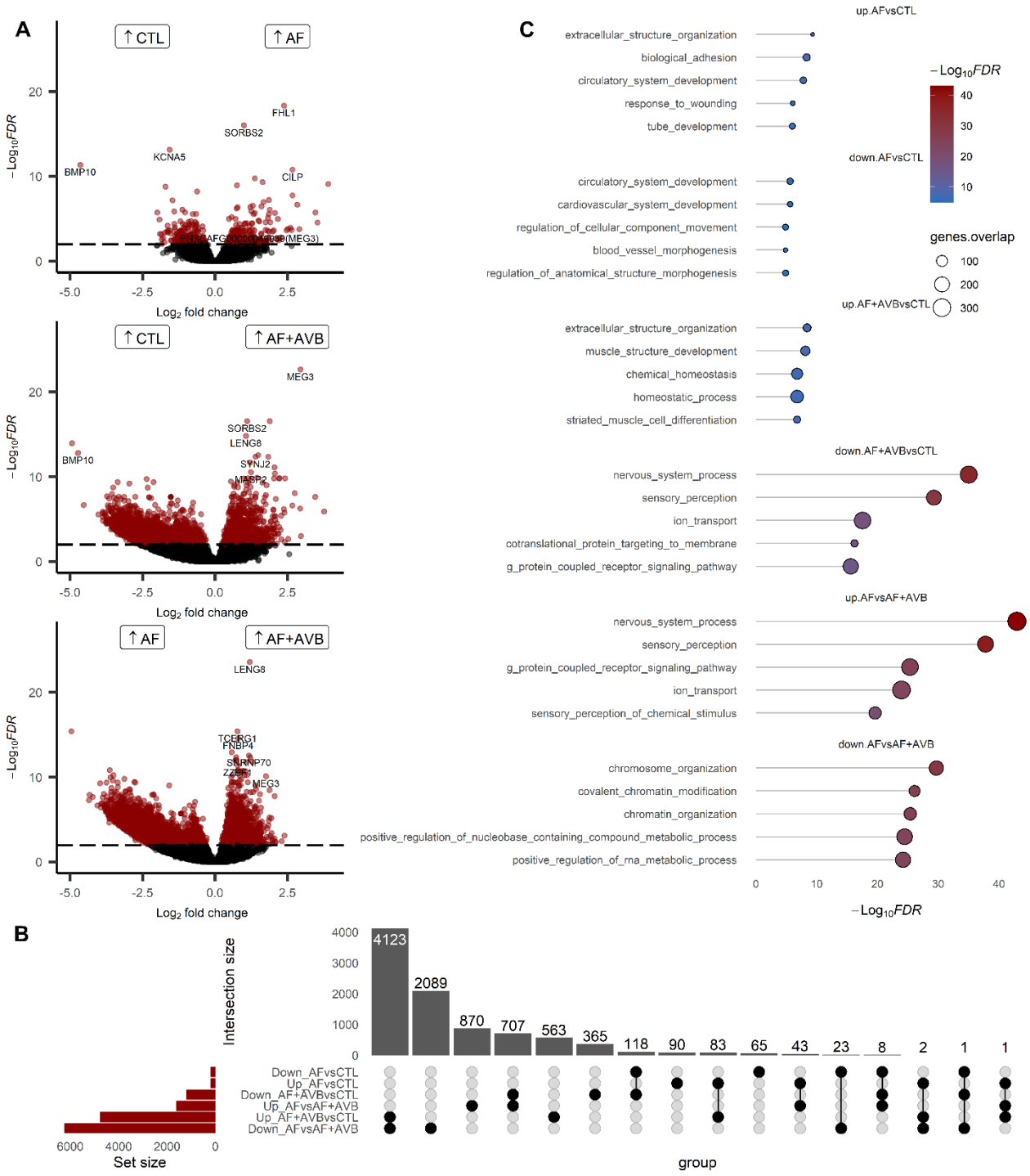


Figure 3. Analyses of differentially expressed atrial genes identify many biological pathways that are dysregulated in atrial fibrillation dog models.

(**A**) Volcano plots of all transcripts that we analyzed in this study. Transcripts in red have a false discovery rate (FDR)<0.01. We found 434, 5971 and 7867 genes that were DE in the AFvsCTL, AF+AVBvsCTL, and AFvsAF+AVB analyses, respectively. The full DE results are available in **Supplemental Table II**. (**B**) Upset plot showing the intersection of up -and down-regulated DE

genes (FDR<0.01) in each analysis. (**C**) The five most significant biological pathways identified using gene-set enrichment analyses (GSEA) for each set of DE genes (FDR <0.01). Full results are available in **Supplemental Table IV**.

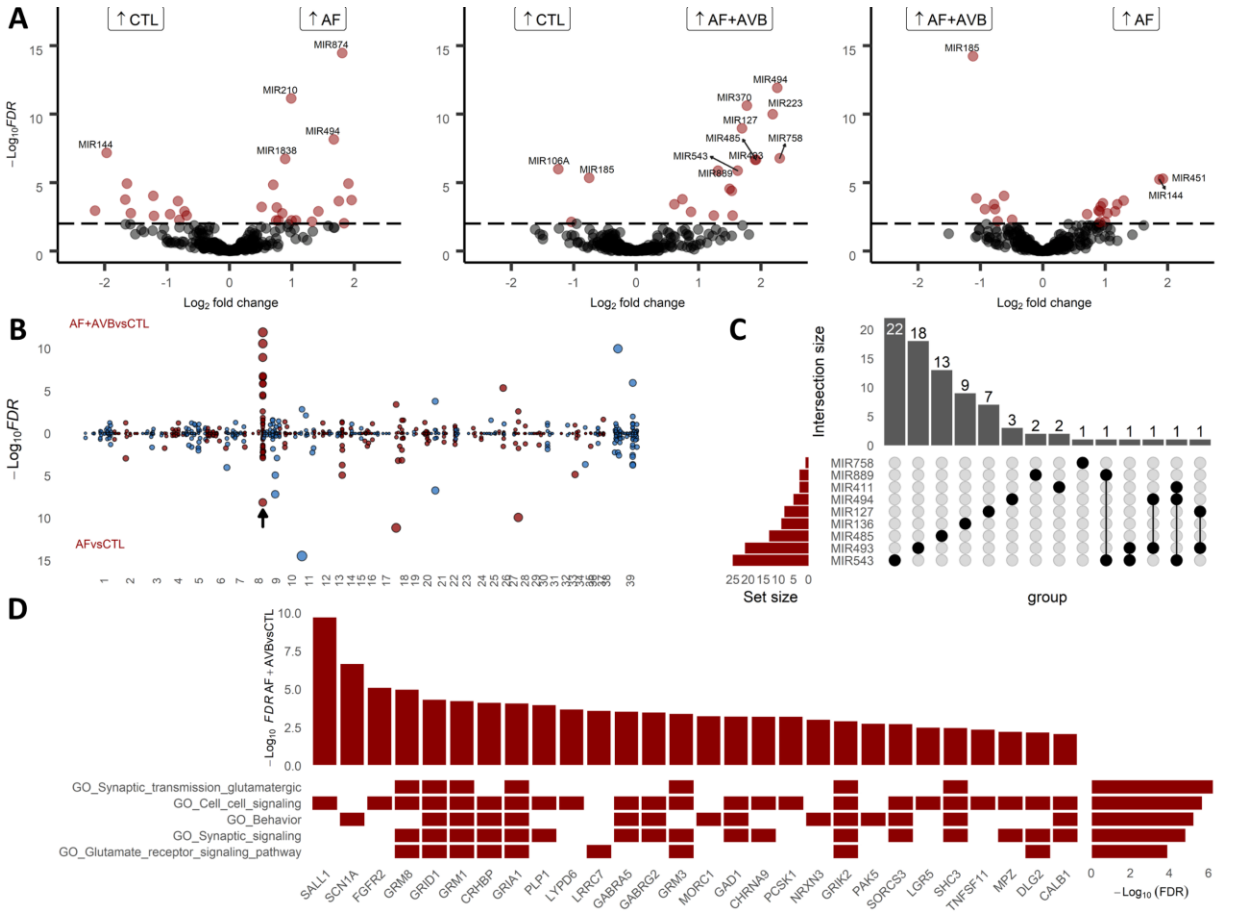


Figure 4. Eleven differentially expressed microRNA*s* (miRNA*s*) map to a canine chromosome 8 region that is syntenic to human DLK1-DIO3.

(**A**) Volcano plots of all miRNA that we measured in our experiments. We identified 31, 19 and 20 miRNA that are differentially expressed (false discovery rate (FDR) <0.01) in the AFvsCTL, AF+AVBvsCTL and AFvsAF+AVB analyses, respectively. (**B**) Miami plots of miRNA and their corresponding statistical significance (y-axis) for the AF+AVBvsCTL (top) and ATvsCTL (bottom) analyses. An arrow indicates the miRNA cluster located on the canine chromosome 8 region that is syntenic to human DLK1-DIO3. The odd and even chromosomes FDR values are in blue and red respectively. (**C**) Upset plot showing the DE miRNA targets located in the syntenic DLK1-DIO3locus and their corresponding number of potential target RNA. We identified potential targets with the MiRNAtap package (predicted by ≥ 3 databases) from DE miRNA (FDR <0.01) and DE mRNA (FDR<0.01). (**D**) Gene-set enrichment analyses (GSEA) with the potential gene targets (x-axis) of the DE miRNA located at the syntenic DLK1-DIO3 locus. We only present the top five pathways enriched in this analysis. A red square in the heatmap indicates membership of a given target gene to the biological pathways located on the left (empty columns were removed for clarity). GSEA FDR and AF+AVBvsCTL DE FDR are on the right and top of the heatmap, respectively.

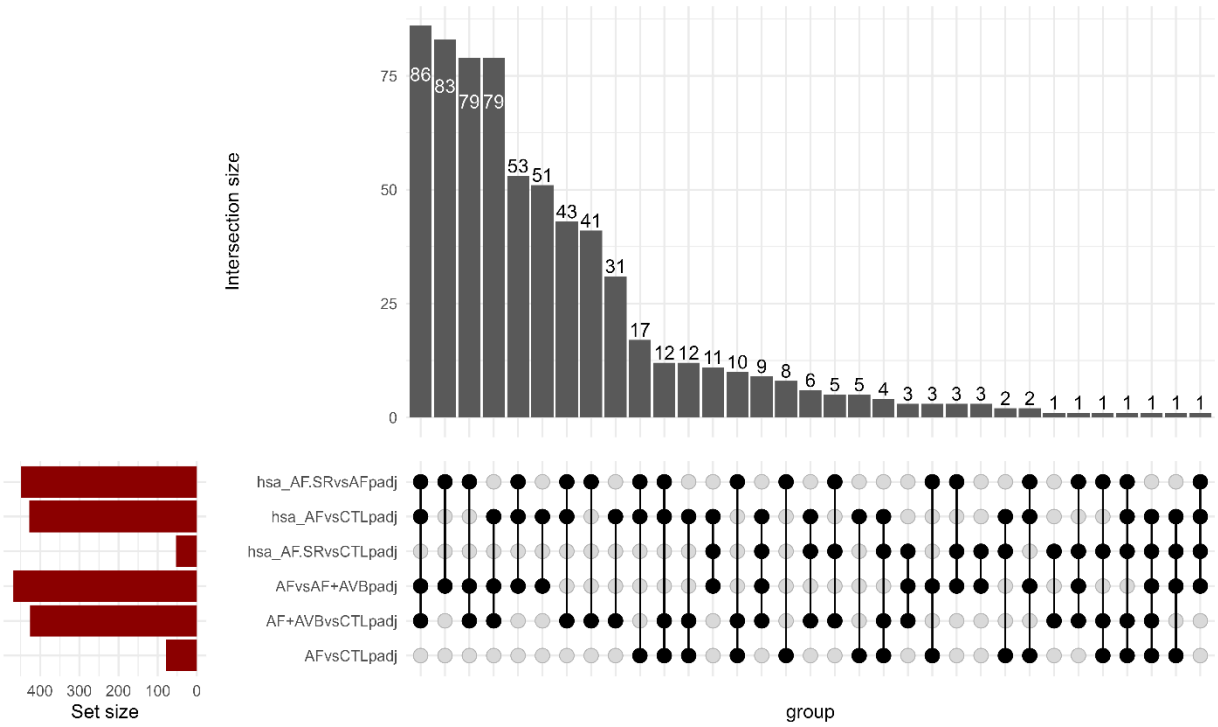


Figure 5. Overlaps in genes differentially expressed in canine AF models and human AF patients.

We compared differentially expressed genes in our canine AF models with annotated human orthologues that are DE in human AF left atrial appendages[216]. Homo Sapiens; hsa, hsa_AF; AF in AF rhythm, hsa_AF.SR; AF in sinus rhythm, hsa_CTL; no AF.

2.8 Supplementary material

Supplementary **Tables S2** to **S6** are provided in the attached zipped folder.

Table S1. Supplemental Table I. Dogs estimated age, weight and sex by treatment.

Phenotypic data			
Condition	Estimated age (Years), mean[min,max]	Weight (Kg), mean[min,max]	Sex (n)
AF (n=6)	3 [1 ,4]	27.1 [22.2 ,31.8]	F=3, M=3
AF+AVB (n=6)	4 [1 ,9]	25 [21 ,31.8]	F=2, M=4
CTL (n=6)	1.8 [1 ,3]	21.6 [18.8 ,24.2]	F=3, M=3
Weight Anova, p=0.055 Estimated age Anova, p=0.247			

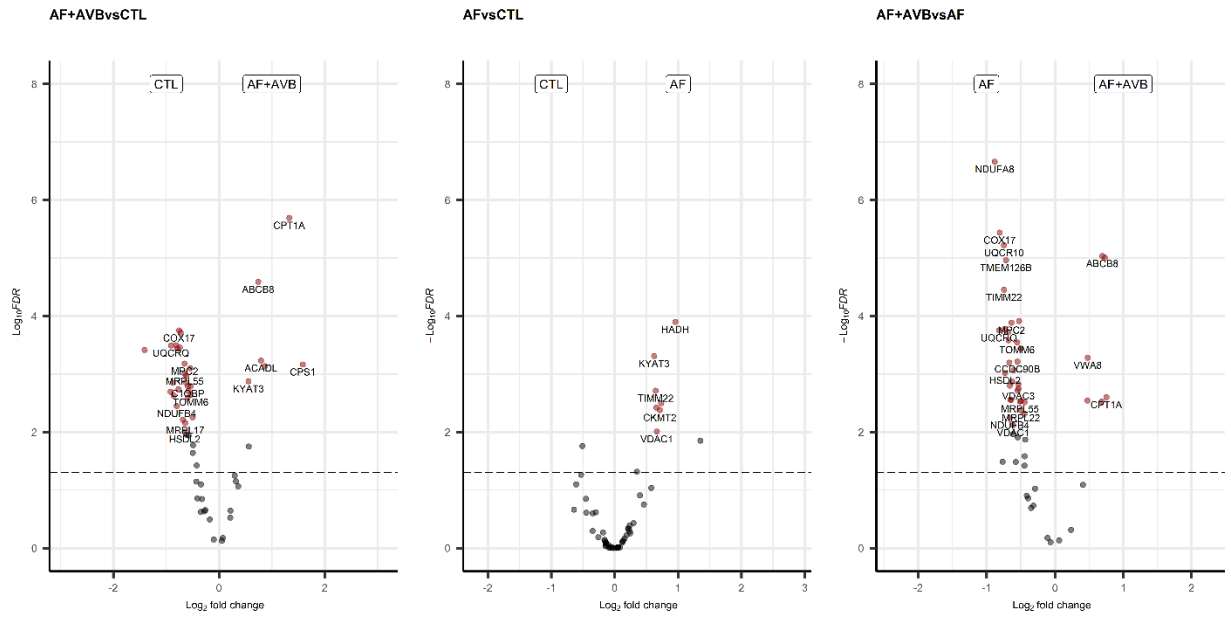


Figure S1. Differentially expressed mitochondrial genes.

Genes with non-zero peak intensity in the heart from the Human MitoCarta3.0[346] were used as reference of mitochondrial genes. Within that list of 539 genes, we find 54 DEG in our AF models represented as volcano plots.

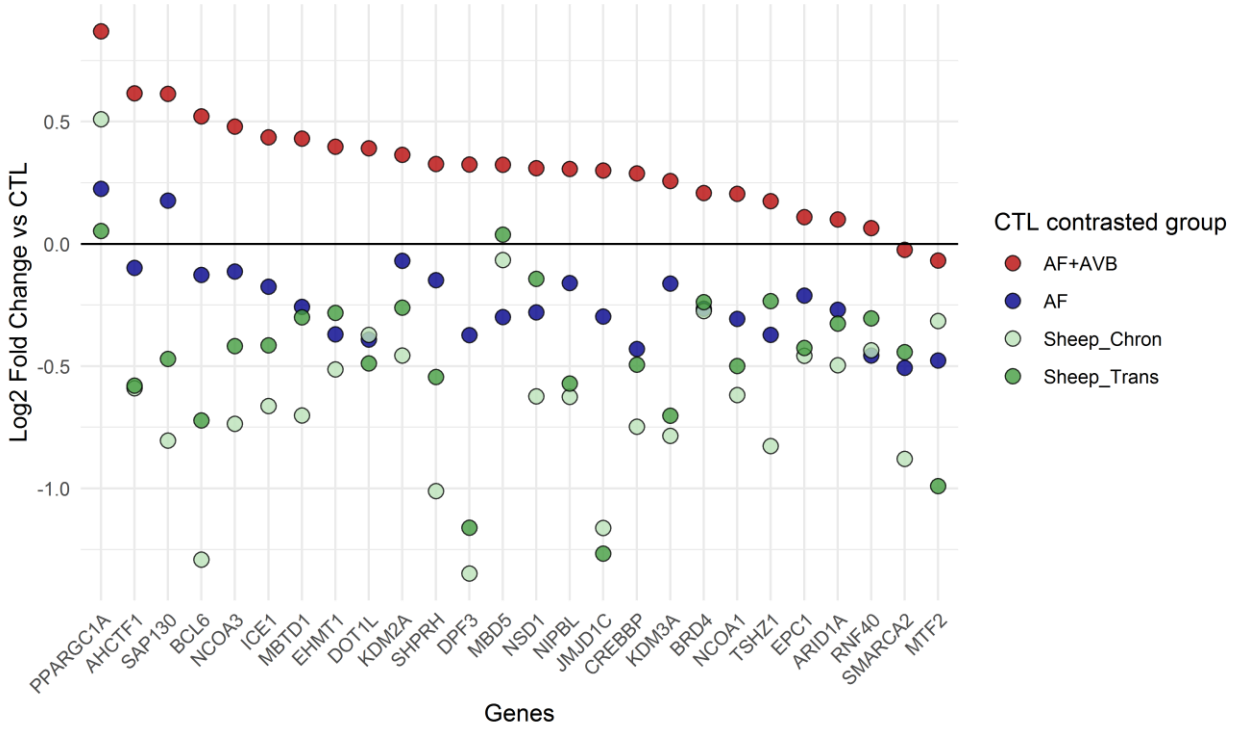


Figure S2. Overlaps in genes differentially expressed in canine and sheep AF models.

Log2 fold changes (L2FC) with controls of chromatin related genes DE in one of our canine models and one of a transition or chronic sheep AF models. The L2FC of canine groups reflect the mRNA changes against CTL group and while the L2FC of the transition and chronic sheep AF models reflect the mRNA changes against the left atrial appendage cardiomyocyte enriched sheep controls[354].

Chapter 3: Major cell-types in multiomic single-nucleus datasets impact statistical modeling of links between regulatory sequences and target genes

Francis J.A. Leblanc[1,2], Guillaume Lettre[1,2,*]

[1]Université de Montréal, Montreal, QC, Canada

[2]Montreal Heart Institute, Montreal, QC, Canada

[*]Corresponding author (guillaume.lettre@umontreal.ca)

Reference: Leblanc, Francis JA, and Guillaume Lettre. "Major cell-types in multiomic single-nucleus datasets impact statistical modeling of links between regulatory sequences and target genes." *Scientific Reports* 13.1 (2023): 3924.

3.1 ABSTRACT

Epigenomic profiling, including ATACseq, is one of the main tools used to define enhancers. Because enhancers are overwhelmingly cell-type specific, inference of their activity is greatly limited in complex tissues. Multiomic assays that probe in the same nucleus both the open chromatin landscape and gene expression levels enable the study of correlations (links) between these two modalities. Current best practices to infer the regulatory effect of candidate *cis*-regulatory elements (cCREs) in multiomic data involve removing biases associated with GC content by generating null distributions of matched ATACseq peaks drawn from different chromosomes. This strategy has been broadly adopted by popular single-nucleus multiomic workflows such as Signac. Here, we uncovered limitations and confounders of this approach. We found a strong loss of power to detect a regulatory effect for cCREs with high read counts in the dominant cell-type. We showed that this is largely due to cell-type-specific *trans*-ATACseq peak correlations creating bimodal null distributions. We tested alternative models and concluded that physical distance and/or the raw Pearson correlation coefficients are the best predictors for peak-gene links when compared to predictions from Epimap (e.g. CD14 area under the curve [AUC] =

0.51 with the method implemented in Signac vs 0.71 with the Pearson correlation coefficients) or validation by CRISPR perturbations (AUC = 0.63 vs 0.73).

3.2 INTROCUCTION

Understanding how the non-coding genome regulates gene expression is paramount to attribute functions to noncoding variants identified by genome-wide association studies (GWAS). We can gain insights into the regulatory potential of non-coding regions through epigenetic mark assessments (CHIPseq), chromatin conformation capture methods (3C, Hi-C, ChIA-PET), expression quantitative loci analysis (eQTL), and open chromatin sequencing (ATACseq and DNase-seq). Extensive databases that collate and summarize these methods' results in a broad range of cell lines and tissues are now available (i.e. ENCODE[368], GENECARD[369]).

Leveraging on this data, statistical models such as those derived by the activity by contact[137] (ABC) method and the Epimap[370] project, EMERGE[371] and others[372-374] have generated strong predictions about the regulatory potential of many non-coding regions. Some of these predictions have been experimentally validated by CRISPR screens, often carried out in cancer cell lines[137,375].

Direct measurement of both open chromatin regions and gene expression can be done concomitantly within a single nucleus using multiomic methods. At the single-nucleus resolution, the correlation of ATACseq peaks and RNAseq genes read counts (henceforth defined as links) provides highly specific hypotheses about the regulatory potential of non-coding candidate *cis*-regulatory elements (cCRE). Current best practices to analyze multiomic datasets and infer the regulatory effect of cCRE involve removing biases associated with ATACseq peak coverage and GC content. As proposed by Ma et al. 2020[299], that method builds a null distribution of gene-peak correlations using ATACseq peaks of matching coverage and GC content drawn from chromosomes excluding the one hosting the tested gene (*trans*-links). The resulting distribution of Pearson correlation coefficients is then scaled, providing Z-scores for the *cis*- and each matched *trans*-links (**Fig. 1A**). This is done under the assumption that these *trans*-ATACseq peaks should not have a regulatory effect on the tested gene. This strategy has been broadly adopted by popular single-nucleus multiomic workflows such as Signac[289].

Here, by analyzing a publicly available multiomic peripheral blood mononuclear cells (PBMC) dataset (**Methods**), we uncovered limitations and confounders associated with this approach (termed the Z-scores method below). We found that the Z-scores method results in a strong loss of power to detect the regulatory effect of cCREs with high read counts in the most abundant cell-type(s). We tested various alternative models and concluded that the simplest approach, that is the raw Pearson correlation coefficients (this method is termed Pearson R below) and/or physical distance is computationally advantageous and provides the best predictions of "ATACseq peak-target gene" links when compared to results from Epimap or CRISPR perturbation screens.

3.3 RESULTS

3.3.1 The number of cells in each cell-type biases the null distributions and statistics of the Z-scores method

In this study, we refer to Z-score as the scaled Pearson R value of a *cis*-link between an ATACseq peak and a nearby gene against its matched *trans*-link null distribution (the Z-scores method, **Fig. 1A**). After processing the PBMC multiomic data with Signac (**Fig. S1** and **Methods**), we noticed striking differences in terms of statistical significance for many peak-gene links when comparing the Pearson R coefficients and the Z-scores. For instance, the ATACseq peak chr16-50684843-50685984, upstream of *NOD2*, contains a *NOD2* eQTL (rs9302752) in whole blood, liver, tibial nerve, spleen and brain based on data from GTEx[376] that is also associated with leprosy and Crohn's disease by GWAS (**Fig. 1B)**[377]. In the PBMC dataset, this ATACseq peak and *NOD2* expression are relatively specific to monocytes and are correlated (R=0.12), although no significant links are identified using the Z-scores method (nominal P-value=0.07) (**Fig. 1B**). However, a significant link is identified between this peak and *SNX20* (nominal P-value=0.02), a gene with an expression density that poorly overlaps the ATACseq signal at chr16-50684843-50685984 (**Fig. 1B**, right column). Importantly, we also noted that the null distributions for the *trans*-peaks matched with three ATACseq peaks at the *NOD2* locus are bimodal, making their corresponding peak-gene link Z-score statistics inaccurate (**Fig. 1C**). When we exclude from the dataset ATACseq peaks that are specific to the cell-type in which the ATACseq peak is mostly

accessible and then create a null distribution with the remaining *trans*-peaks, the distribution is unimodal (**Fig. 1D**). This suggests that the choice of which ATACseq peaks are included in the null distribution has a huge impact on the Z-scores method results (see below).

Next, we performed several analyses to better understand how cell-type composition affect the identification of ATACseq peak-gene links from single-nucleus multiomic experiments. In the PBMC dataset, CD14 monocytes is the dominant cell-type (n=3075 cells [27%])(**Fig. 2A)** and clusters with CD16 monocytes and classical dendritic cells 2 (cDC2) both on ATACseq and RNAseq UMAP (**Fig. S1**). We refer to this cell archetype as mononuclear phagocytes (MP). We found that ATACseq peaks with specific accessibility in MP had lower median peak-gene link statistics as calculated with the Z-scores method (**Fig. 2A**), and that rarer cell-types had more extreme Z-score statistics (**Fig. S2** and **S3A-B**). Thus, the links between ATACseq peaks and genes have less significant statistics when identified in the major cell-types of this PBMC dataset.

To evaluate if the number of MP influenced the calculated link statistics with the Z-scores method, we down-sampled cells from the MP clusters from 3,788 to 500 cells, and repeated the analyses. The down-sampling increased the Z-scores of cells from the MP clusters (*t*-test P-value$_{CD14}$=1.7x10^{-81}, P-value$_{CD16}$=4.5x10^{-7}, P-value$_{cDC2}$=3.2x10^{-14}), and reduced the peak-gene link Z-scores for all other cell-types except for the ones which had few cell-type-specific ATACseq peaks (**Fig. 2B**). These results suggest that the cell-type composition of the dataset has a strong influence on the statistics calculated using the Z-scores method.

3.3.2 More abundant cell-types have more power to identify correlated ATACseq peaks in *trans*

As highlighted for the *NOD2* locus, we found that the Z-scores method often generates bimodal null distributions (**Fig. 1C-D** and **Fig. S4**), and that these bimodal distributions are more frequent in more abundant cell-types (**Fig. S3C**). We hypothesized that the co-accessibility of cell-type-specific *trans*-open chromatin regions – for instance due to the activity of a common transcription factors giving rise to a co-regulatory network – could cause the emergence of a second mode in the null distributions. In support of this hypothesis, removing cell-type-specific *trans*-peaks from the null distributions generally eliminated the bimodality and increased the Z-score statistics (**Fig. 1C-D** and **Fig. S4**). To better understand the effect of the bimodality on the

Z-scores method, we tested each null distributions for multiple modes (P-value<0.05 for >1 modes [**Methods**][378]) and compared the Z-scores and Pearson R coefficients of peak-gene links obtained for multimodal and unimodal null distributions. Whereas statistics from the Z-scores method were significantly lower for peak-gene links with multimodal null distributions (Wilcoxon P-value $<1x10^{-300}$), we found that the simple Pearson R statistics were higher (Wilcoxon P-value=$1.93x10^{-7}$)(**Fig. 2C**). Consistently, we found that peak-gene links with multimodal null distributions were more likely to have non-significant Z-score statistics (near 0), even when the corresponding Pearson R coefficients were relatively high (**Fig. 2D**). Additionally, links between ATACseq peaks and target genes in MP were more likely to have multimodal null distributions when compared to other rarer cell-types (**Fig. 2E**). Together, these results suggest that the popular Z-scores method used to infer a regulatory effect between ATACseq peaks and target genes in multiomic data is biased, counter-intuitively, towards lower abundant cell-types. Our analyses show that this bias arises, at least in part, from the production of bimodal null distributions when matching the tested peak-gene links with links found in *trans* for abundant cell-types, presumably because of increased power to detect co-regulated *trans* ATACseq peaks.

3.3.3 Read coverage, but not GC content, impacts peak-gene link statistics

Beside the Pearson R and Z-score methods, we considered two additional approaches. First, because of the inherent sparsity of single-nucleus ATACseq data, we tested a zero-inflated negative binomial (ZINB) model, allowing to independently account for the zero component of a peak-gene link. Second, we also tested a new method – scREG – that is reported to outperform the simple Pearson R model on CD14 monocytes peak-gene link predictions based on eQTL data[298,379]. Of note, scREG output link scores for peak-gene within each cell-type. We compared the peak-gene links identified by each of these four models to the Epimap predictions of cCREs and target genes for CD14 monocytes, B cells and NK cells (**Methods**).

In the Z-scores method, the rationale for generating null distributions is to account for possible confounders such as the number of mapped reads (i.e. coverage) and GC bias. We decided to explore the impact of these two factors on the Z-scores, Pearson R, $scREG_{CD14}$, and ZINB statistics. In the Signac workflow, initial filtering is done separately at the gene expression and ATACseq peak level, removing genes or peaks with <10 cells with non-zero counts. Thus, it is possible to have peak-gene links defined by a single cell that has counts for both RNAseq and

ATACseq modalities. We found that the Z-scores method was particularly sensitive to the number of cells with non-zero counts, with extreme Z-scores associated with links identified in a small number of cells (**Fig. S5**). Although the same effect was less striking for the Pearson R, $scREG_{CD14}$, and ZINB methods (with high statistics being positively correlated with high number of cells with non-zero counts), we still observed some extreme statistics for links identified in a small number of cells (**Fig. S5**).

We further characterized the impact of GC content on the Z-scores method, as this is the only method that consider this variable. *Trans*-peak matching for an ATACseq peak is done through the attribution of weights dependent on the input variables (here GC and counts). We compared peak-gene link Z-scores from two analyses with identical parameters (i.e. null distributions generated independently twice for the same peak-gene links), and also for analyses with and without GC as a matching criteria. We found that the correlation value of Z-scores for two analyses with identical matching parameters is 0.97 (**Fig. S6A**), while that of a model matching for counts only vs one that matches for both GC content and counts is 0.95 (**Fig. S6B**), suggesting that the GC content does not strongly influence the identification of peak-gene links (even for the stronger links, see the right-hand tail of the distributions in **Fig. S6**).

3.3.4 The raw Pearson R coefficients and/or physical distance provide better statistics to capture predicted or functionally validated links between ATACseq peaks and target genes

We next turned to independent datasets that have predicted or experimentally ascertained links between cCREs and target genes to address the limitations of the Z-score method and propose new strategies. ZINB is a computationally expensive method compared to other methods tested (**Fig. S7**). Further, scREG currently limits its output to 100,000 links. For these reasons, we started by comparing with Epimap predictions the accuracy of the Pearson R and the Z-scores methods when considering a very large number of peak-gene links (|Pearson R| >0.01, n=590,842 links). Our Receiver Operating Characteristics (ROC) curve analysis showed that the Pearson R method outperformed the Z-scores method in these three cell-types (**Fig. 3A**, **S8A** and **S9A**). For instance, the area under the curve (AUC) were 0.71 and 0.51 in CD14 monocytes when applying the Pearson

R and Z-scores methods, respectively (**Fig. 3A**). We further compared the predictive value of the Pearson R and Z-scores against links found in promoter capture Hi-C (PCHi-C) consisting of 17 human primary blood cell-types[380] (**Fig. S10**). The results were consistent with the Epimap validation, where the Pearson R (AUC = 0.57) outperforms the Z-scores method (AUC = 0.49). Using a smaller set of peak-gene links with a more stringent threshold for inclusion (|Pearson R| > 0.1, n=15,113 links), we then applied the four methods and compared results with predictions from Epimap. scREG outperformed other models in all cell-types, and the Z-scores method performed worse (**Fig. 3B**, **S8B** and **S9B**).

The physical distance between regulatory sequences and gene transcription start sites has been found to be a strong predictor of cCRE's effects on nearby genes[137,381]. Because the scREG model weights the peak-gene link scores with physical distance ($e^{(-distance/200kb)}$), we reasoned that weighting the Pearson R coefficients by the distance between the ATACseq peaks and the target genes could improve its accuracy. Physical distance-weighted Pearson R coefficients resulted in AUC that were similar to those obtained when using distance alone, and remarkedly better than with scREG on all three Epimap cell-type predictions (**Fig. 3B**, **S8B** and **S9B**).

As described above, scREG calculates peak-gene link scores per cell-type[298]. We repeated the analyses of the Epimap predictions with the Z-scores, Pearson R and ZINB methods but focusing on single cell-type. For instance, for the Epimap CD14 predictions, we only analyzed peak-gene links identified in the CD14 subset of the PBMC multiomic dataset. This approach had a minimal impact on the AUC statistics for all three cell-types analyzed (compare panels **B** and **C** in **Fig. 3** (CD14 cells), **Fig. S8** (B cells) and **Fig. S9** (NK cells)), but severely reduced the number of detected peak-gene links (**Fig. 3D**, **Fig. S8D** and **Fig. S9D**). One major drawback of this single cell-type approach is that by reducing the number of cells in the analyses, we significantly reduced power to detect peak-gene links. For instance, our analysis of the whole PBMC dataset (i.e., using all PBMC to compute statistics) yielded 15,113 peak-gene links with |Pearson R| > 0.1, including 1611 links (10.7%) that overlap with Epimap predictions for CD14 cells. In contrast, when we restricted our analysis to CD14 cells from the PBMC multiomic dataset, we found 2,499 links, including 143 (5.7%) also predicted by Epimap in CD14 cells (**Fig. 3D**). From these observations, we conclude that using all available cells from multiomic experiments to detect links between

ATACseq peaks and target genes is a more powerful strategy than limiting the analyses to single cell-type.

Finally, we compared the peak-gene links identified in the PBMC multiomic datasets with 644 cCRE-gene pairs that were functionally validated using CRISPR perturbations in different cell models[382]. We found that distance alone (or in combination with the Pearson R coefficient) was the best predictor of links between ATACseq peaks and genes that were consistently validated by CRISPR perturbations (**Fig. 4**). We also noted that the simple Pearson R statistic, even without being weighted by physical distance (AUC=0.73), outperformed all other metrics, including the distance-weighted scREG scores (AUC=0.65-0.67)(**Fig. 4**).

3.4 DISCUSSION

Motivated by the absence of several strong candidate links between regulatory sequences and target genes in our analysis of multiomic PBMC data using a common bioinformatic pipeline, we investigated several factors that could impact these results. We found that cell-type composition in single-nucleus data can have a dramatic effect on the ability to detect peak-gene links. Indeed, we showed that null distributions matched on ATACseq peak coverage and GC content are often bimodal, especially when the peaks are specific to (or enriched in) the major cell-types. Our analyses suggest that this second mode arise because of the following reasons: First, the number of ATACseq peaks detected in a given cell-type increases with the number of cells, thus increasing the chances to draw *trans*-ATACseq peaks that are opened in that cell-type when building the null distributions. Second, as cells within a given cell-type share transcription factors, their open chromatin regions tend to also be more correlated. Together, this creates two modes: one coming from the *trans*-ATACseq peaks of the dominant cell-type and a second from other cell-types (generally less correlated). We illustrate this conclusion by showing that if a peak-gene link is cell-type-specific, removing ATACseq peaks detected in this cell-type from the null distribution generally removes the mode most associated with the tested link (**Fig. 1C-D**) and drastically increases its Z-score (**Fig. 2B**). This approach also has the consequence to inflate the Z-scores of links mostly found in less abundant cell-types, because the null distributions will contain fewer ATACseq peaks from the same, rare cell-types (**Fig. 2B**). One apparent solution to

this problem is to detect peak-gene links within specific cell-types, although we showed that this method is sub-optimal because of the loss in power to detect peaks and genes when fewer cells are analyzed.

The rationale behind the null distributions implemented in the Z-scores model comes from reports that the Tn5 transposase used in the ATACseq protocol has a GC bias[372]. However, a recent comprehensive study specifically addressing this issue did not detect such bias[383]. Furthermore, our own analyses found minimal (if any) effect of GC content the detection of links between ATACseq peaks and target genes, while the number of cells with non-zero counts in both the ATACseq peak and the gene is a major determinant. We recommend not analyzing links if at least 15 cells do not have both non-zero counts in the ATACseq and RNAseq modalities.

There are no perfect datasets to validate links between ATACseq peaks and the promoter of target genes that are inferred from single-nucleus multiomic experiments. In our study, we used predictions from Epimap, PCHi-C and published perturbations using CRISPR tools. Surprisingly, we found that simply considering physical distance and/or the Pearson correlation coefficients provide optimal concordance with these datasets. These methods also have the advantage to be computationally scalable, something that can become problematic for the ZINB method (and to some extent the scREG and Z-scores methods as well). It is obvious that larger multiomic experiments as well as true "gold-standard" datasets of *bona fide* peak-gene links will enable the development of more sophisticated statistical methods. In the meantime, we recommend to carefully consider "ATACseq peaks-target genes" links inferred from single-nucleus multiomic analyses, and to validate them using orthogonal approaches such as 3D chromatin conformation analyses, expression quantitative trait loci (eQTL) results, and *in silico* predictions.

3.5 METHODS

3.5.1 Multiomic PBMC data

We analyzed the PBMC multiomic dataset from 10X Genomics (https://www.10xgenomics.com/resources/datasets/pbmc-from-a-healthy-donor-granulocytes-

removed-through-cell-sorting-10-k-1-standard-1-0-0). The data was processed according to the Signac tutorial (https://satijalab.org/signac/articles/pbmc_multiomic.html), which uses the same dataset. For the 11,331 cells identified, the workflow annotated 30 cell-types, of which 17 have more than 50 cells (cell-types represented in the dendrogram in **Fig. S1A**). We restricted our analyses to those 17 cell-types for power purposes. For all *cis*-links, we only analyzed ATACseq peaks located within 500kb of a gene transcription start site.

3.5.2 Links cell-type marker ATACseq peaks

The accessibility specificity of ATACseq peaks was tested using the Presto package (*wilcoxauc.Seurat()* function). A marker ATACseq peak was attributed to a cell-type using the highest area under the curve (AUC). ATACseq peaks with AUCs < 0.55 and an FDR > 10^{-5} in all cell-types with more than 50 cells were attributed to the non-specific peak group. We used these labels to attribute links to each cell-type.

3.5.3 Down-sampling mononuclear phagocytes

Cell-types clustering together, both in RNAseq and ATACseq UMAPs, were categorised as part of mononuclear phagocytes (MP; CD14 Mono, CD16 Mono, cDC2). 500 out of 3,782 MP were randomly drawn and reprocessed with the rest of the PBMC cells. We compared Z-scores of peak-gene links with overlapping peaks and identical genes between the full dataset (n=11,331 cells) and the down-sampled MP dataset (n =8,049 cells). Overlapping links with |Pearson R| > 0.1 in the full dataset are shown by cell-type in Fig. 2.

3.5.4 Removing co-regulated peaks from null distributions

To assess the co-accessibility effect of cell-type-specific *trans*-open chromatin regions on Z-scores we used the output of Signac's *CallPeaks()* function to retrieve from which cell-type a peak was called by MACS2[384] (implemented in Signac). The cell-types were categorised in 4 broader classes representative of the UMAP and dendrogram:

1. **Lymphoid**; CD8 Naive, CD4 Naive, CD4 TCM, CD8 TEM, CD8 TCM, CD4 TEM, MAIT, Treg

2. **NK cells**; gdT, NK, CD8 TEM, MAIT

3. **Monocytes**; CD14 Mono, CD16 Mono, cDC2, pDC

4. **B cells**; B intermediate, B memory, B naive

For ATACseq peaks that were called in all 4 broad cell-type classes, no filtering was done. For ATACseq peaks with some specificity (i.e., not called in all 4 broad cell-type), we removed all *trans*-peaks from the *trans*-peak pool to match the *cis*-peak that were also called in the same broad cell-type class. Therefore, a tested ATACseq peaks called only in B cells and Monocytes by MACS2 would have a null distribution composed of *trans*-peaks called in lymphoid and/or NK cells.

3.5.5 Multimodal test

To better assess the bimodality of the null distributions, for each link we drew 1000 GC and log(ATACseq peak sum of counts +1) matched *trans*-peaks using Signac's function *MatchRegionStats()* (instead of the default 200), computed their Pearson R and scaled them. To establish if the resulting null distributions were multimodal, we used the mixtools package expectation–maximization function *normalmixEM()* with k=2 and epsilon = 1e-03 as described in Ameijeiras-Alonso *et al.*, 2019[385]. Null distributions with nominal p-values < 0.05 were categorised as multimodal.

3.5.6 Pearson R and Z-score models

Links using all PBMC

We used the R package Signac function's *LinkPeaks()* with a 500kb window for a gene's transcription starting site and a null distribution of 200 *trans*-ATACseq peaks to obtain Z-scores and Pearson R as described in the package tutorial (https://satijalab.org/signac/articles/pbmc_multiomic.html) on all PBMC passing the quality-control steps. We used the log(ATACseq peak sum of counts +1) instead of the counts to match peaks. Peak-gene links were filtered for |Pearson R| > 0.01 to remove cells with zero count in both the ATACseq peak and the gene tested or > 0.1 to limit the number of tests as mentioned in the text and the figure legends. The PBMC RNAseq annotation has 36,601 genes, of which 29,613 were detected in at least 1 cell and 21,878 were detected in at least 10 cells (Signac's default threshold for genes to test). Using |Pearson R| > 0.01 as threshold, we obtained 590,842 links for 15,011 genes, while a threshold of 0.1 resulted in 15,113 links for 2,088 genes.

Cell-type subsetted

The same strategy mentioned in *Links using all PBMC,* was applied independently after subsetting the Epimap matching cell-types; CD14 mono cells, B cells (B intermediate, B memory, B naive) and NK cells (NK, NK Proliferating, NK_CD56bright). This resulted in 2501 links with |Pearson R| > 0.1 using 3,096 CD14 Mono cells, 11095 links using 934 B cells and 11959 links using 522 NK cells.

3.5.7 ZINB model

We tested the ZINB model below using the R package pscl function *zeroinfl()*[386].

ZINB model: $Gene \sim ATACseq + cdr.ATAC + cdr.RNA \mid cdr.ATAC + cdr.RNA$

The zero-inflated component was modeled with cellular detection rates (cdr) for both ATACseq peaks and genes (proportion of features with 0 counts). For genes with no 0 counts across all cells, we used the negative binomial generalized linear model (GLMNB**)** implemented with the R package MASS function *glm.nb()* given that no zero component could be modeled.

GLMNB model: $Gene \sim cCRE + cdr.ATAC + cdr.RNA$

The |Z-value| were used as predictive value for each tested peak-gene links.

3.5.8 scREG implementation

We initially tested an exact implementation of the package tutorial (https://github.com/Durenlab/RegNMF). The current software returns a prioritized list of links (10,000 per identified clusters) with very low overlap with our validation data, which made the comparison with Epimap and CRISPRi validation uninformative. We found 2 likely explanations. First, the output from *SplitGroup()* are the 10,000 peak-gene pairs with lowest scores values for that cell-type, as opposed to the 10,000 highest scores. Second, the ATACseq data is log10 transformed while the RNAseq data is log2 transformed. This creates stronger weights for the gene expression component of the links matrix and skews results towards highly expressed genes (i.e. ATACseq peaks linked to *MALAT1* were the top 20 links for all clusters). To have comparable results to the other tested models, we used as inputs the genes and peaks from the 15,113 peak-gene links with |Pearson R| > 0.1 as well as the 644 peak-gene links kept from CRISPR validations.

3.5.9 Peak-gene link models comparison with Epimap

We retrieved the Epimap peak-gene link predictions for the PBMC matching cell-types (CD14 MONOCYTE, B CELL, NK CELL) from https://personal.broadinstitute.org/cboix/epimap/links/links_corr_only/. We chose these cell-types because their cluster showed greater homogeneity and boundaries in the PBMC multiome dataset (in contrast to the lymphoid cells, see **Fig. S1**). For each Epimap cell-type, we kept peak-gene links found in all replicates to insure reproducibility. We recovered the hg38 positions using the AnnotationHub package (Annotationhub chain: hg19ToHg38.over.chain.gz). For these analyses, peak-gene links from the PBMC multiomic dataset were considered positive when the ATACseq peak overlapped at least partly the Epimap enhancer position and the linked gene was the same. ROC curves were calculated with the ROCR package by increasing the thresholds of the model's statistic. For the 15,113 peak-gene links with |Pearson R| > 0.1, 1630, 984 and 1538 were considered positive (found in Epimap) for CD14 MONOCYTE, B CELL, NK CELL respectively. For the 590,842 peak-gene links with |Pearson R| > 0.01, 12848, 7562 and 13084 were considered positive (found in Epimap) for CD14 MONOCYTE, B CELL, NK CELL respectively.

3.5.10 Peak-gene link models comparison with PCHi-C

We used the PCHi-C data from *BM Javierre et al., Cell, 2016,* (Data S1) consisting of 17 human primary blood cell-types. We used all links with a CHICAGO score >5 in at least one cell-type (n = 728,838 links). We first filtered out genes that were not found in the PBMC RNA matrix, then we filtered links to keep only bait regions that overlaps 1 gene promotor, restricted the link to 500kb distance and removed bait-bait links. This resulted in a list of 273,208 PCHi-C links. Liftover to hg38, overlap with links passing |Pearson R| > 0.01 and ROC curve analysis were done as in the Epimap comparison described in the section above. Lastly, we filtered out links for which the gene promotor was not found in PCHi-C, resulting in ROC curves for 330,224 links with 51,602 positives (link found in PCHi-C) and 278,622 negatives (link not found in PCHi-C).

3.5.11 Peak-gene links validation with CRISPR perturbation results

Given the modest number of CRISPR-based validated links, we expanded the number of tested links to include all links with non-zero read counts for both the gene and the ATACseq peak tested (|Pearson R| > 0.01 (n=590,842)). ROC curves were calculated using the ROCR package in

a set of 664 CRISPR validations from *J. Nasser et al. 2021* (Table S5) overlapping PBMC links of which 51 were tagged as significant (used as positive peak-gene links). The original CRISPR validation data (n=5755) was filtered to include CRISPR targets overlapping an ATACseq peak with a corresponding gene expression readout. We also excluded duplicates (same link tested in multiple cell lines), links that showed divergent results across cell lines and those that were excluded by the author of the study for various reasons, denoted by the *IncludeInModel* column (power insufficient, overlapping promotor and others).

3.6 DECLARATIONS

3.6.1 Data availability

All results presented here were generated with our code available on GitHub: https://github.com/lebf3/Links_models_multomic .

3.6.2 Acknowledgements

This work was funded by the Canadian Institutes of Health Research (MOP #136979), the Canada Research Chair Program, the Foundation Joseph C. Edwards and the Montreal Heart Institute Foundation (G. Lettre). F. JA Leblanc was supported by the Fonds de Recherche en Santé du Québec (FRQS) and Université de Montréal.

3.6.3 Author contributions

All analyses and figures were done by FL. The manuscript was written by GL and FL.

3.6.4 Competing interests

The authors declare no competing interests.

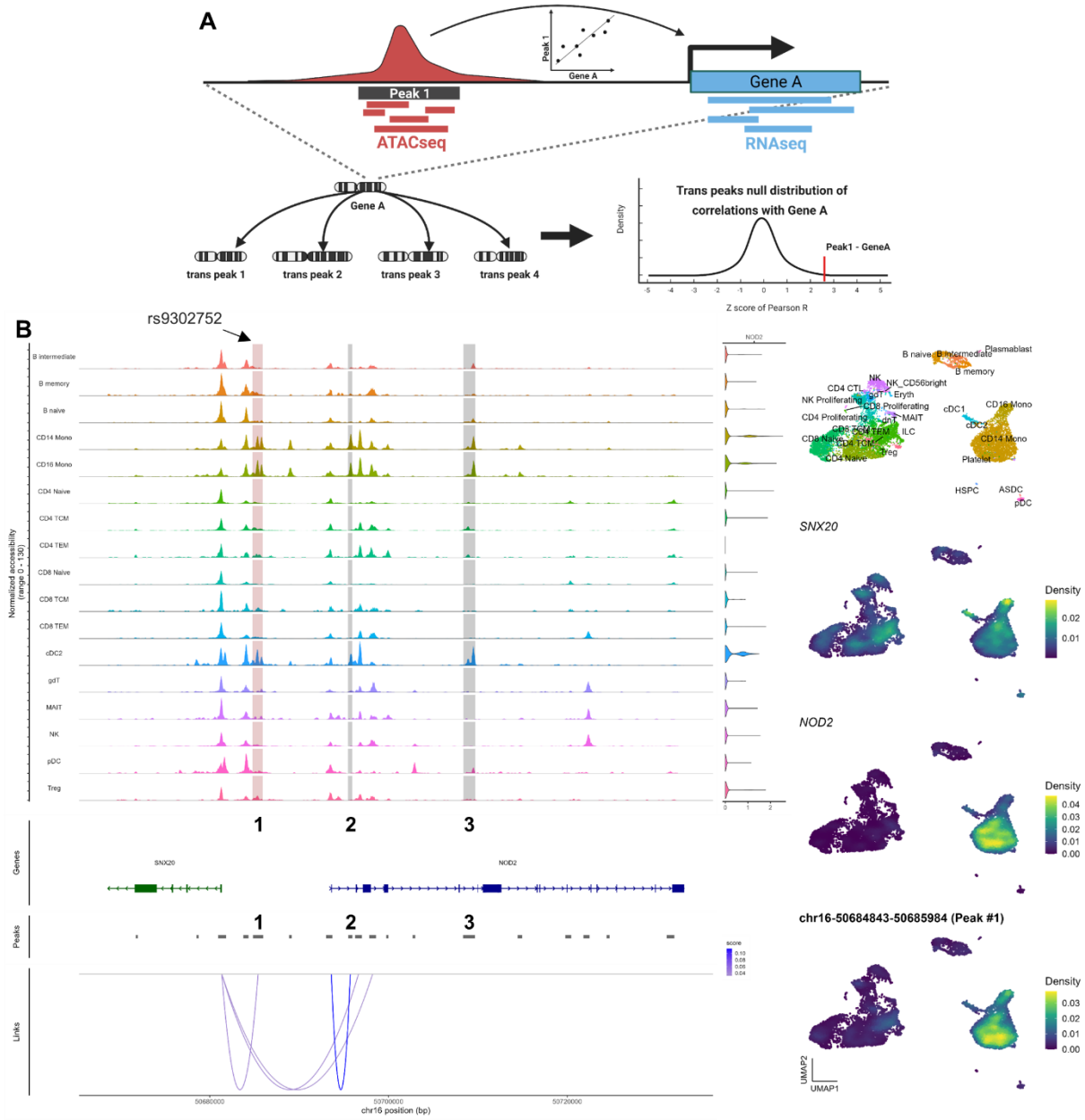

A
Peak 1
Gene A
ATACseq
RNAseq
Peak 1
Gene A
Gene A
trans peak 1
trans peak 2
trans peak 3
trans peak 4
Trans peaks null distribution of correlations with Gene A
Density
Peak1 - GeneA
Z score of Pearson R
B
rs9302752
Normalized accessibility (range 0 - 150)
B intermediate
B memory
B naive
CD14 Mono
CD16 Mono
CD4 Naive
CD4 TCM
CD4 TEM
CD8 Naive
CD8 TCM
CD8 TEM
cDC2
gdT
MAIT
NK
pDC
Treg
1
2
3
Genes
SNX20
NOD2
Peaks
Links
chr16 position (bp)
Plasmablast
NK_CD56bright
Eryth
CD4 CTL
NK Proliferating
CD8 Proliferating
CD4 Proliferating
MAIT
ILC
Treg
cDC1
Platelet
HSPC
ASDC
pDC
SNX20
NOD2
chr16-50684843-50685984 (Peak #1)
Density
UMAP1
UMAP2
C
All trans-peaks in null distribution
3 NOD2_chr16-50708347-50709667
2 NOD2_chr16-50695493-50695963
1 NOD2_chr16-50684843-50685984
1.282
1.688
1.488
Z-scores of Pearson R
D
No cell-type specific trans-peak in null distribution
3 NOD2_chr16-50708347-50709667
2 NOD2_chr16-50695493-50695963
1 NOD2_chr16-50684843-50685984
7.942
8.146
9.125
Z-scores of Pearson R

Figure 1. The Z-scores method misses candidate regulatory sequences linked to *NOD2* expression in peripheral blood mononuclear cells (PBMC).

(**A**) The Z-scores model matches an ATACseq peak for GC content and coverage with ATACseq peaks in *trans* to create a scaled null distribution, producing Z-scores for each *trans*-links and the tested peak. (**B**) ATACseq tracks at the *NOD2* locus identified in PBMC. The grey areas (labeled 1-2-3) highlight the top three ATACseq peaks correlations with *NOD2* expression using the simple Pearson R method. Peak #1 (chr16-50684843-50685984) includes an eQTL for *NOD2* that is also associated with leprosy and Crohn's disease by GWAS. The loops highlighted in the "Links" row are identified using the Z-score method (P-value <0.05); we note that there is no significant link between peak chr16-50684843-50685984 (peak #1) and *NOD2*. Loops are drawn from the middle of the ATACseq peaks to the transcription start site of the correlated gene(s). In the right column, we showed (top to bottom) RNAseq UMAP of cell-type annotations, *SNX20* expression density, *NOD2* expression density, and chr16-50684843-50685984 ATACseq accessibility density. The violin plots represent *NOD2* expression levels in each cell-type. (**C**) Three GC- and coverage-matched null distributions for ATACseq peaks (peaks #1-2-3) at the *NOD2* locus generated using the Z-scores method. Labeled boxes represent the corresponding Z-score statistics for the peaks tested against *NOD2* expression. Only peak #2 is significant using this approach (nominal P-value = 0.04). (**D**) As in **C**, but we generated the null distributions after excluding ATACseq peaks specific to the same cell-type as the tested ATACseq peak (see **Methods** for details). With this strategy, the three peaks (#1-2-3) are significantly linked with *NOD2* expression (P-value $<1x10^{-15}$).

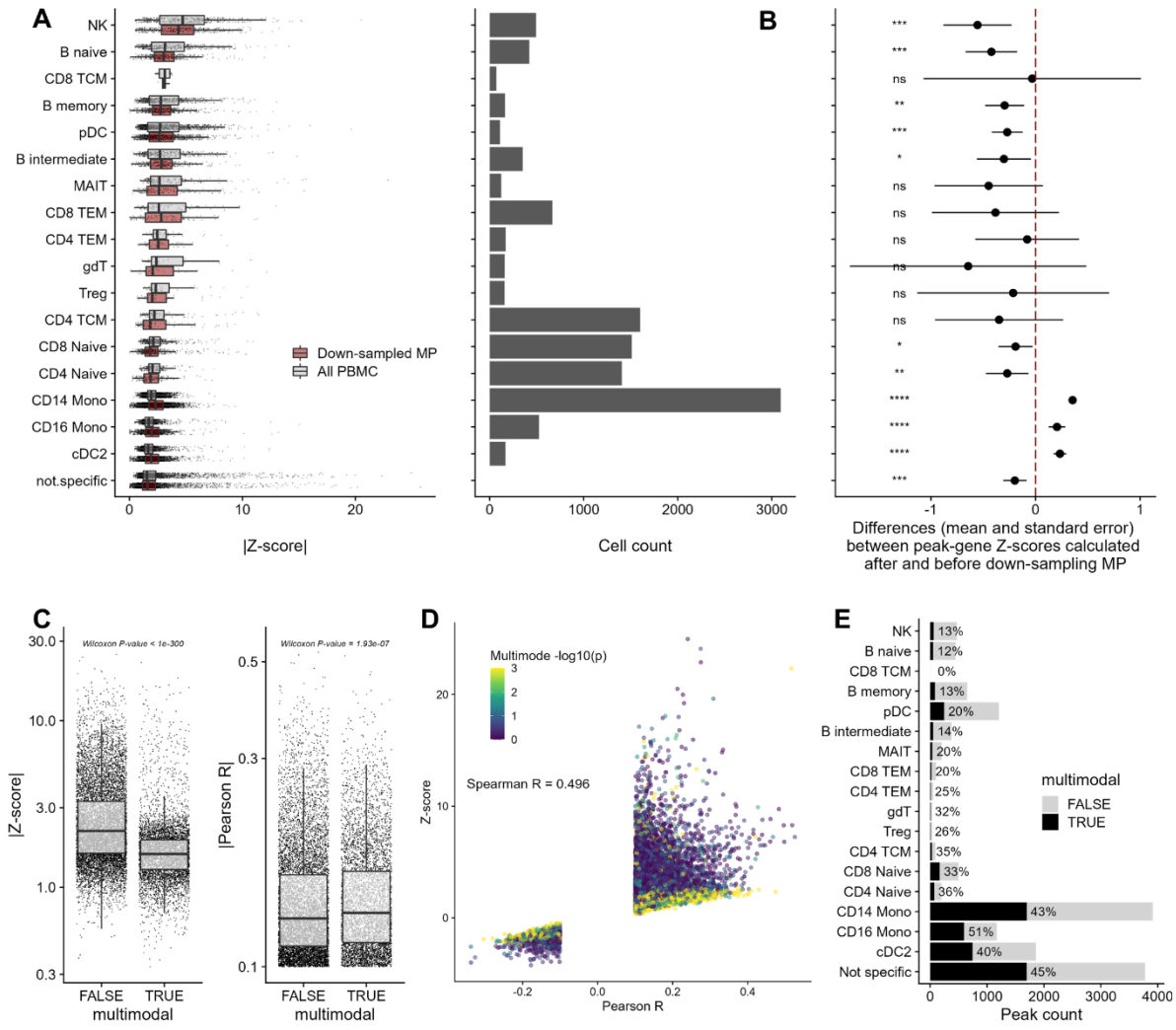


Figure 2. The cell-type composition of the PBMC single-nucleus multiomic dataset impacts the identification of gene-peak links using the Z-scores method.

(**A**) Our analyses included 15,113 gene-peak links with |Pearson R| >0.1 (Signac's default parameter) identified in 30 cell-types. The left column shows the number of cells in each cell-type. In the right column, we show the boxplots of the statistics calculated using the Z-scores method. For this analysis, we assigned gene-peak links to specific cell-type using sensitivity and specificity metrics (**Methods**). Links that could not be unambiguously assigned are grouped in the "not.specific" category. We repeated the analyses by down-sampling the number of mononuclear phagocytes (MP) to n=500. (**B**) Effect of down-sampling mononuclear phagocytes (MP) from 3,788 to 500 cells on peak-gene link statistics calculated with the Z-scores method. Positive values indicate higher Z-scores (i.e. more significant) after down-sampling. ns; not significant, *; P-value<0.05, **; P-value<0.01, ***; P-value<0.001, ****; P-value<0.0001. (**C**) The Z-score statistics and Pearson R coefficients for links between ATACseq peaks and target genes that generated uni- or multimodal null distributions (with the Z-scores method). (**D**) Scatterplot of the Pearson R coefficients (*x*-axis) and statistics calculated with the Z-scores method (*y*-axis) for all links between ATACseq peaks and target genes. Each point is color-coded based on the P-value of the multimode test. Peak-gene links that generated multimode null distributions (in yellow) tend to have Z-score statistics ~ 0 despite many having high Pearson R coefficients. (**E**) Proportion of

multimodal null distributions by cell-type generated by the Z-scores method for the tested links between ATACseq peaks and target genes. Mono; Monocytes, cDC; classical Dendritic cells, NK; Natural killer cells, pDC; progenitor Dendritic cells, TEM; T effector memory cells, TCM; T central memory cells, gdT; Gamma delta (γδ) T cells, MAIT; mucosal-associated invariant T cells, Treg; regulatory T cells.

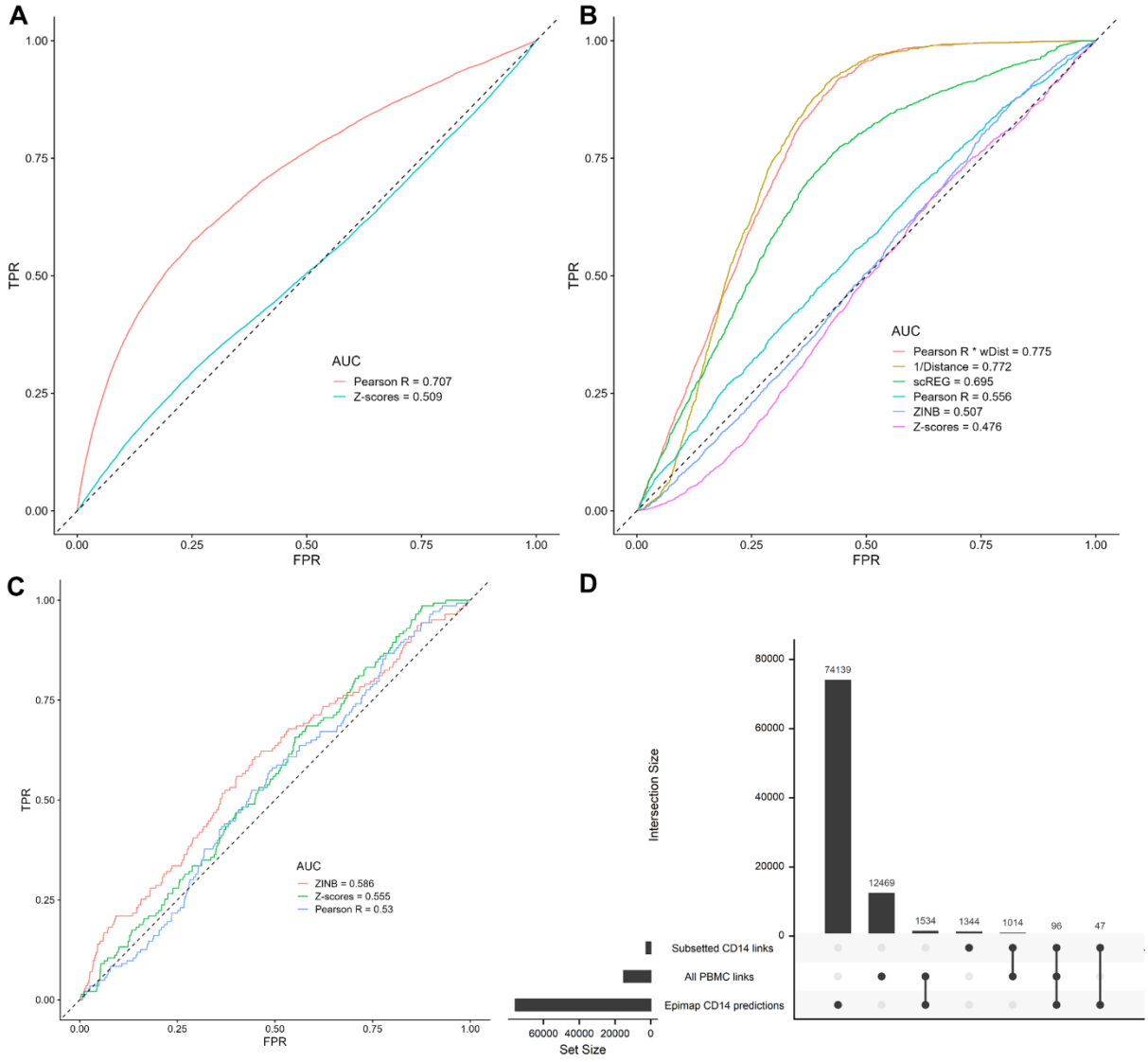


Figure 3. The Pearson R method more accurately validates Epimap-predicted links between cCRE and target genes in CD14 cells.

(**A**) We used the Pearson R and Z-scores methods to detect links between ATACseq peaks and target genes (590,842 links with |Pearson R| >0.01) in the complete (i.e., using all PBMC to compute statistics) PBMC multiomic dataset. Then, we performed Receiving Operating Curves (ROC) analyses to compare the identified peak-gene links from the multiomic data with regulatory links in CD14 cells predicted by the Epimap Project. (**B**) As in **A**, but using a smaller set of links defined using a more stringent statistical threshold (15,113 links with |Pearson R| >0.1). All cell-types are used to identify links, except for scREG which by design output link scores by cell-type (in this case, CD14 cells). (**C**) As in **B**, but limiting these ROC analyses to links between ATACseq

peaks and target genes with |Pearson R| >0.1 that were found in the CD14 cells subset of the PBMC multiomic dataset. (**D**) Upset plot that shows the intersections of links identified between ATACseq peaks and target genes using either the full PBMC multiomic dataset or only the CD14 cells subset with cCRE-gene regulatory links in CD14 cells as predicted by the Epimap Project. ZINB; zero-inflated negative binomial, wDist; weighted distance ($e^{(-distance/200kb)}$), TPR, true positive rate; FPR, false positive rate.

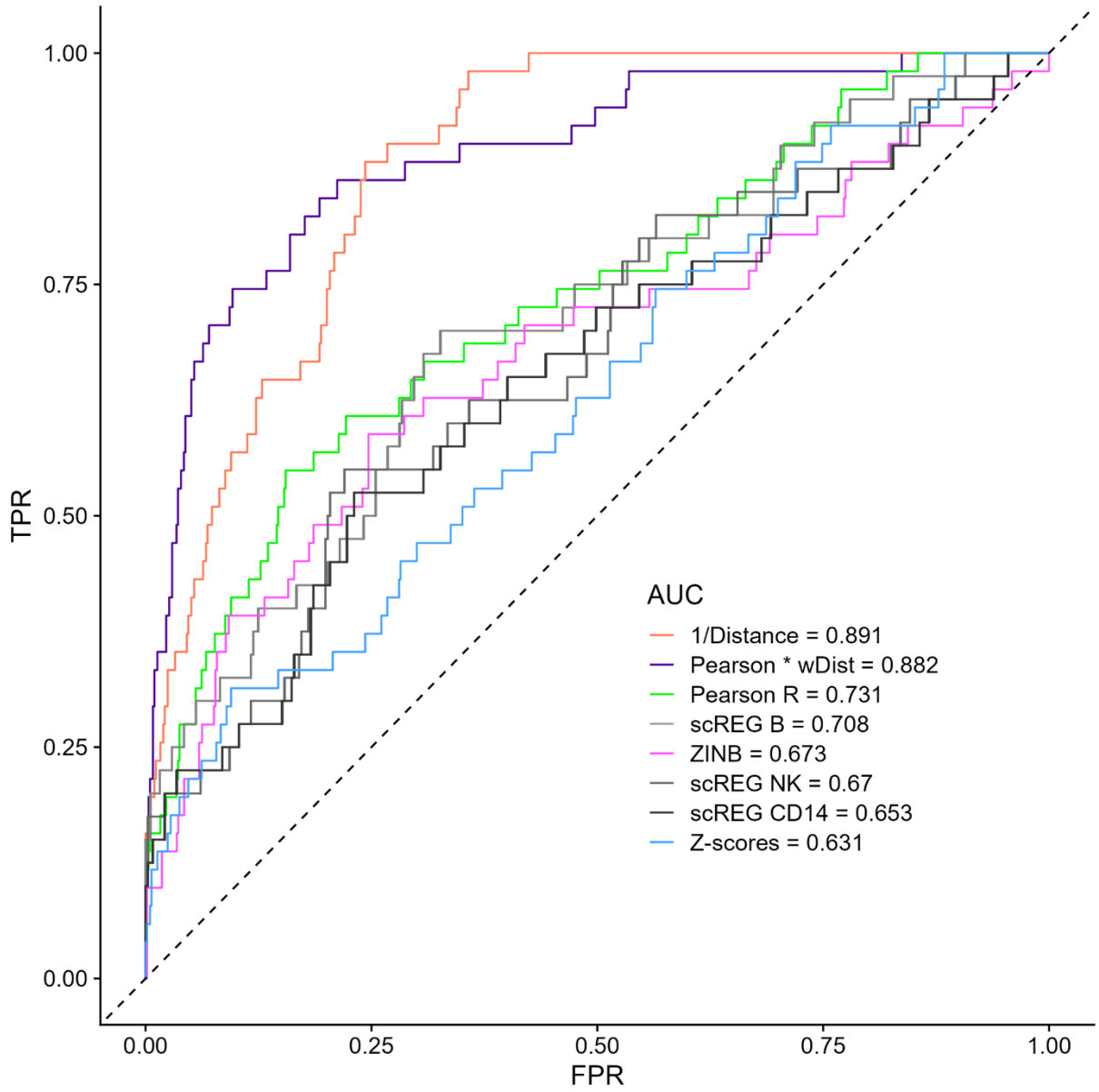


Figure 4. Physical distance and the Pearson R coefficient best capture cCRE-gene pairs identified by CRISPR perturbations.

We identified 644 CRISPR-validated cCRE-gene pairs that had corresponding links (defined using |Pearson R| > 0.01) in the PBMC multiomic dataset. Distance-alone or distance-weighted Pearson R coefficients are the best predictors, with the Z-score method (implemented in Signac) performing worst. ZINB; zero-inflated negative binomial, wDist; weighted distance ($e^{(-distance/200kb)}$), TPR, true positive rate; FPR, false positive rate.

3.7 Supplementary material

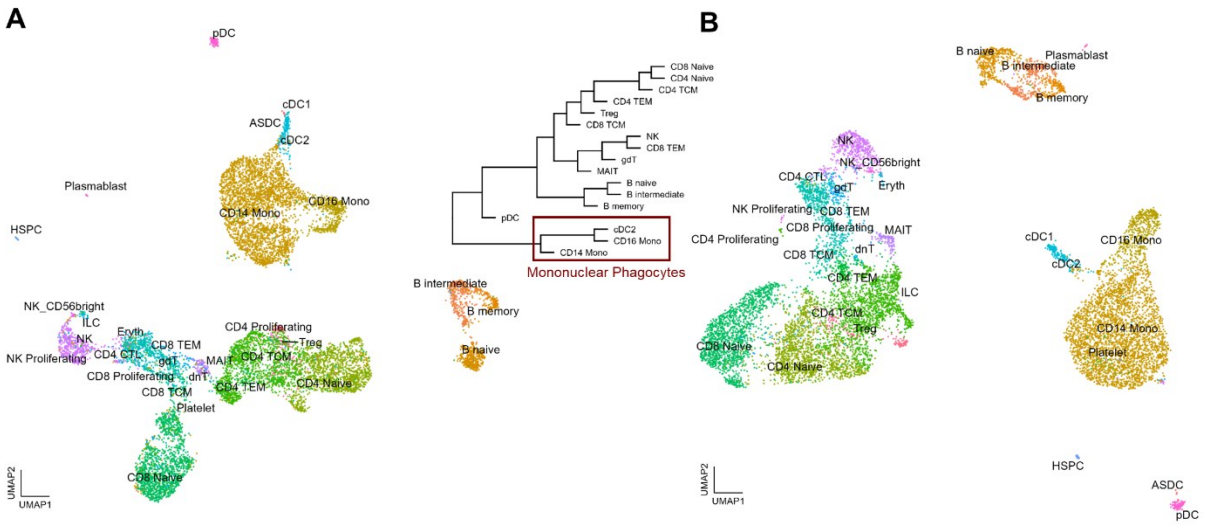


Figure S1. Clustering of 11,331 PBMC.

(**A**) ATACseq and (**B**) RNAseq data. The data is embedded using uniform manifold approximation and projection (UMAP). The Euclidean distance of mean expression by cell-type represented as dendrogram shows mononuclear phagocytes as a distinct cell archetype (CD14, CD16 and cDC2).

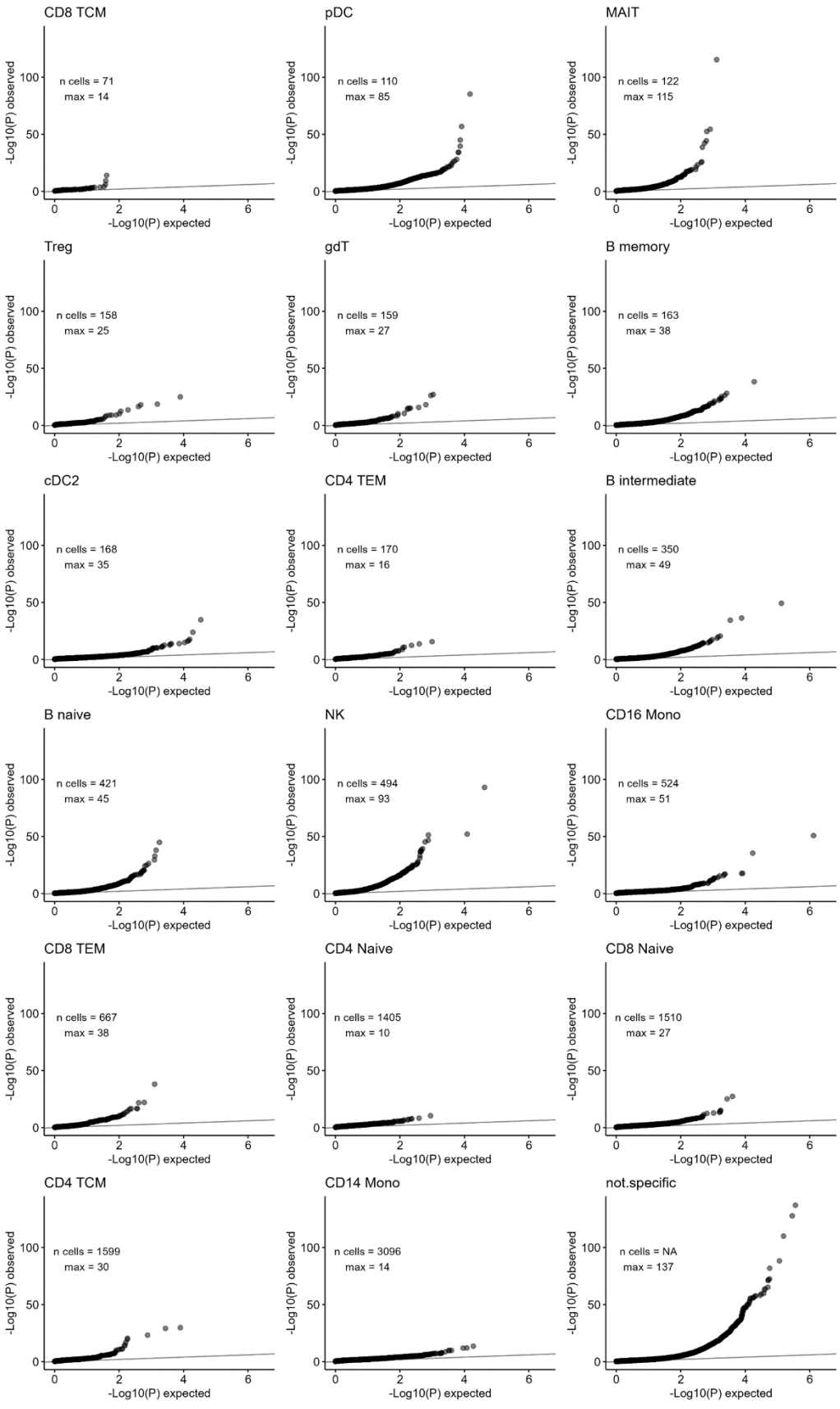

CD8 TCM
n cells = 71
max = 14
pDC
n cells = 110
max = 85
MAIT
n cells = 122
max = 115
Treg
n cells = 158
max = 25
gdT
n cells = 159
max = 27
B memory
n cells = 163
max = 38
cDC2
n cells = 168
max = 35
CD4 TEM
n cells = 170
max = 16
B intermediate
n cells = 350
max = 49
B naive
n cells = 421
max = 45
NK
n cells = 494
max = 93
CD16 Mono
n cells = 524
max = 51
CD8 TEM
n cells = 667
max = 38
CD4 Naive
n cells = 1405
max = 10
CD8 Naive
n cells = 1510
max = 27
CD4 TCM
n cells = 1599
max = 30
CD14 Mono
n cells = 3096
max = 14
not.specific
n cells = NA
max = 137
-Log10(P) observed
-Log10(P) expected

Figure S2. Distributions of ATACseq peaks-gene link statistics calculated using the Z-scores method as implemented in Signac.

The peak-gene links with |Pearson R > 0.01| were attributed to specific cell-type (cell-types with n>50 cells) using the peak's specificity of accessibility (**Methods**). Mono; Monocytes, cDC; classical Dendritic cells, NK; Natural killer cells, pDC; progenitor Dendritic cells, TEM; T effector memory cells, TCM; T central memory cells, gdT; Gamma delta (γδ) T cells, MAIT; mucosal-associated invariant T cells, Treg; regulatory T cells, Max; maximal $-\log_{10}$(P-value) calculated for a given cell-type.

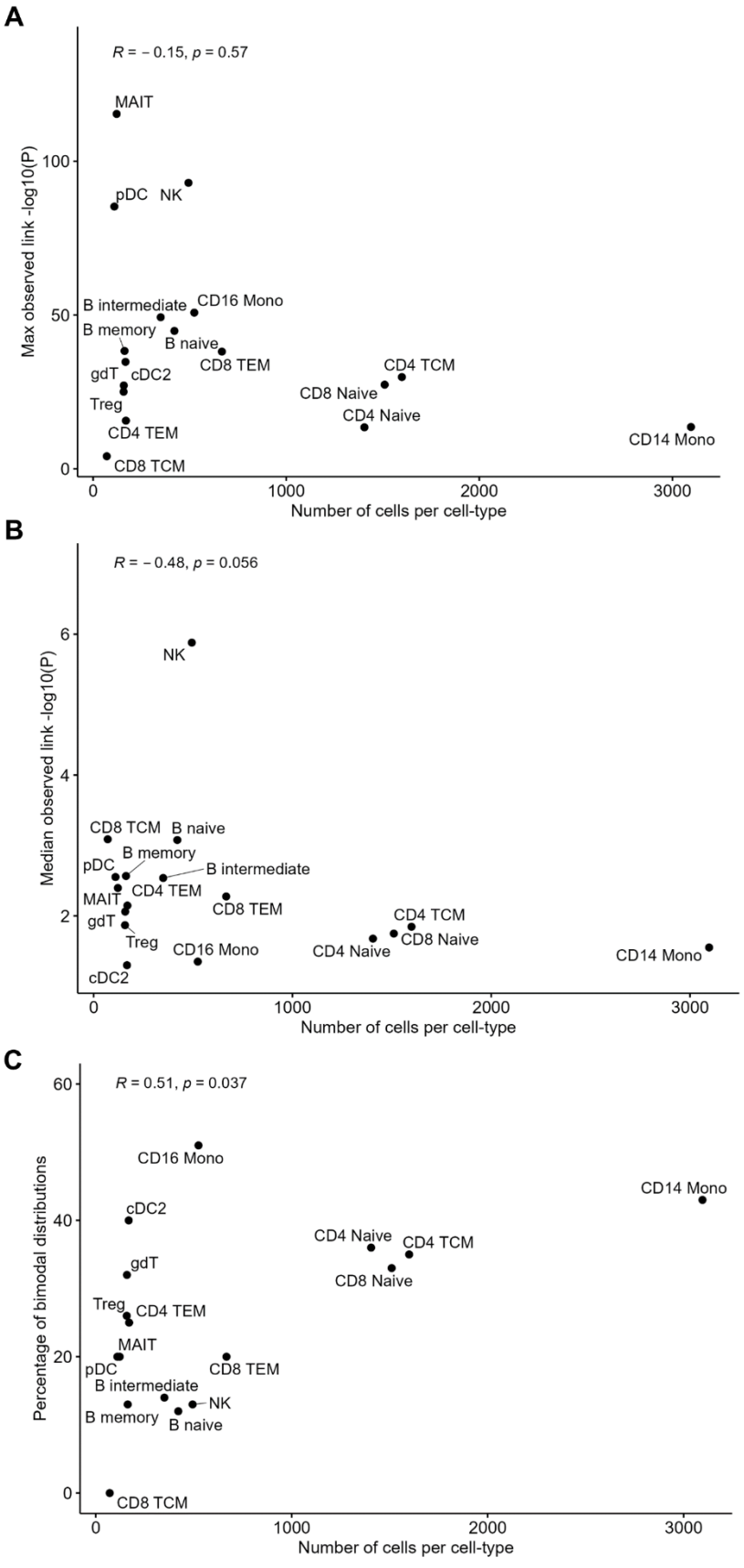

A
R = − 0.15, p = 0.57
MAIT
pDC
NK
B intermediate
CD16 Mono
B memory
B naive
CD8 TEM
CD4 TCM
gdT
cDC2
CD8 Naive
Treg
CD4 Naive
CD4 TEM
CD14 Mono
CD8 TCM
Max observed link -log10(P)
Number of cells per cell-type
B
R = − 0.48, p = 0.056
NK
CD8 TCM
B naive
B memory
pDC
B intermediate
MAIT
CD4 TEM
CD8 TEM
gdT
CD4 TCM
Treg
CD16 Mono
CD4 Naive
CD8 Naive
CD14 Mono
cDC2
Median observed link -log10(P)
Number of cells per cell-type
C
R = 0.51, p = 0.037
CD16 Mono
CD14 Mono
cDC2
CD4 Naive
CD4 TCM
gdT
CD8 Naive
Treg
CD4 TEM
MAIT
pDC
CD8 TEM
B intermediate
NK
B memory
B naive
CD8 TCM
Percentage of bimodal distributions
Number of cells per cell-type

Figure S3. The impact of cell-type counts on properties of the Z-scores method implemented in Signac.

For cell-type-specific ATACseq peaks, we identified peak-gene links and compared (**A**) the most extreme P-value from the Z-score method, (**B**) the median P-value from the Z-score method, or (**C**) the percentage of bimodal null distributions with the number of cells in that cell-type. P-values are from the Spearman correlation test. Mono; Monocytes, cDC; classical Dendritic cells, NK; Natural killer cells, pDC; progenitor Dendritic cells, TEM; T effector memory cells, TCM; T central memory cells, gdT; Gamma delta (γδ) T cells, MAIT; mucosal-associated invariant T cells, Treg; regulatory T cells.

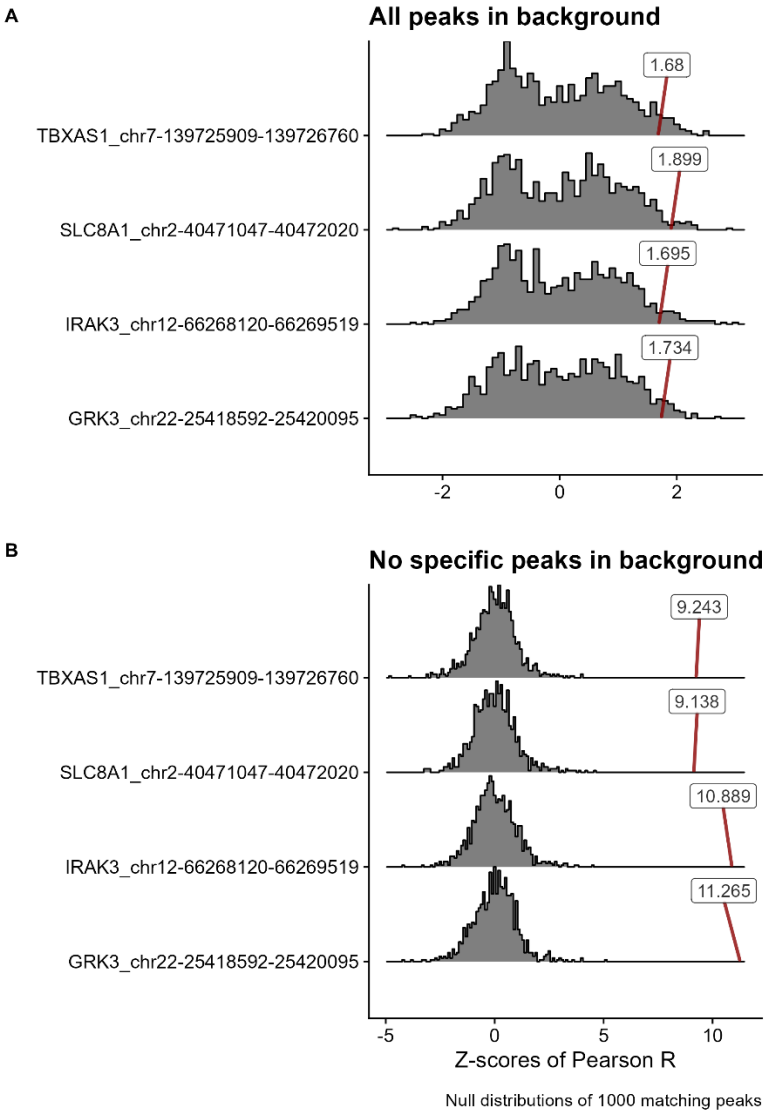


Figure S4. Other examples of bimodal null distributions generated by the Z-scores method.

Four GC- and count-matched null distributions with high Pearson R coefficients and low Z-scores. Labeled boxes represent the Z-scores for the tested cCRE and its linked gene (*y*-axis names: gene_peak) against the null distributions (**A**) before and (**B**) after removing from the dataset *trans* ATACseq peaks that are specific to the cell-type in which the cCRE is mostly accessible.

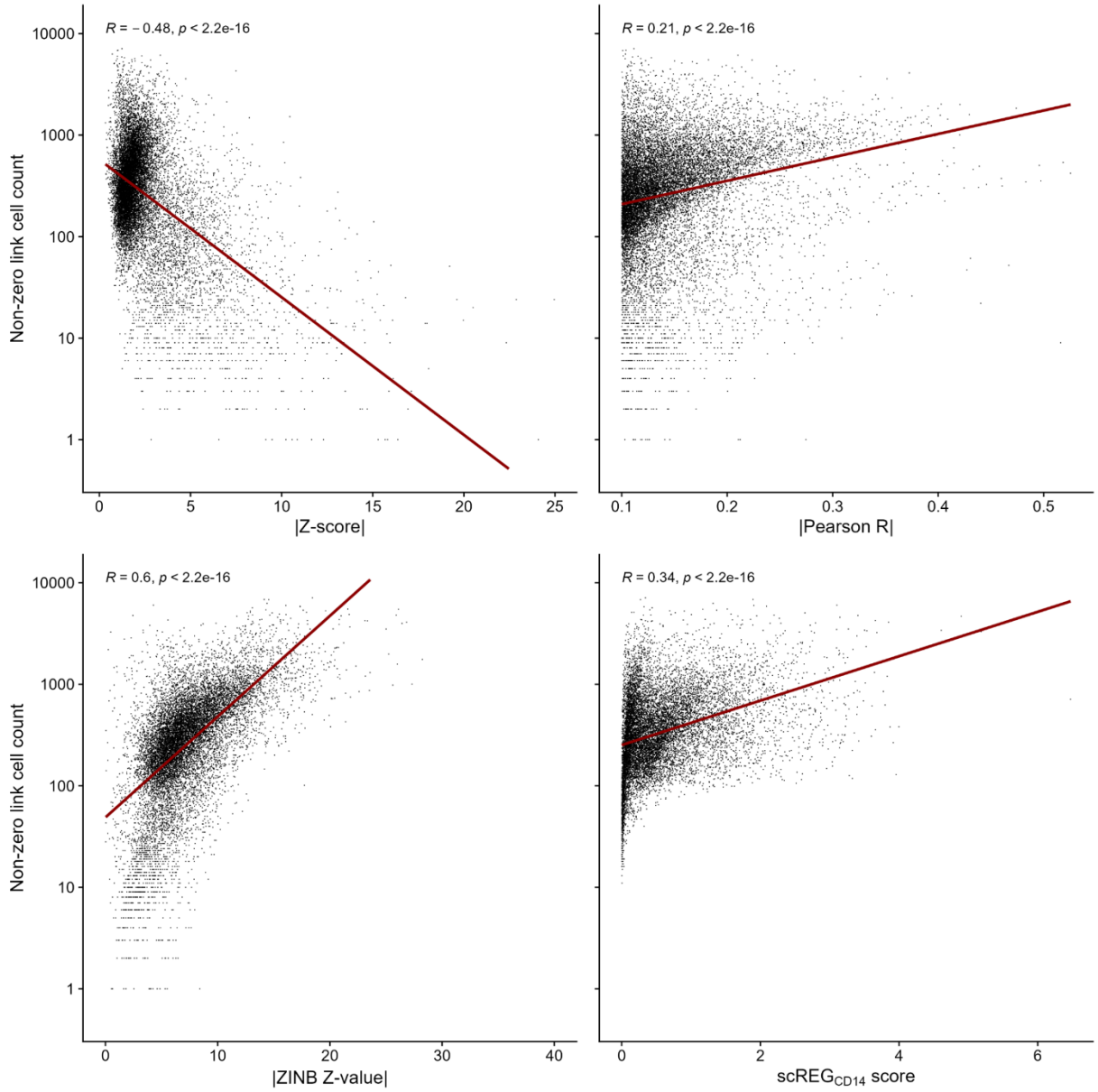


Figure S5. The Z-scores method tends to output extreme statistics for peak-gene links that are identified in a few cells.

In contrast, the statistics for the Pearson R, ZINB and scREGCD14 methods are higher when there are more cells with non-zero counts (as expected given higher power to detect links). Zscores, ZINB Z-values and Pearson R from links with |Preason R| > 0.1 were compared against the number of cells for which both the gene and the peak from that link had a non-zero read count. We found one outlier using ZINB which was removed for visualisation. On each plot, we added the Pearson R coefficient and corresponding P-value.

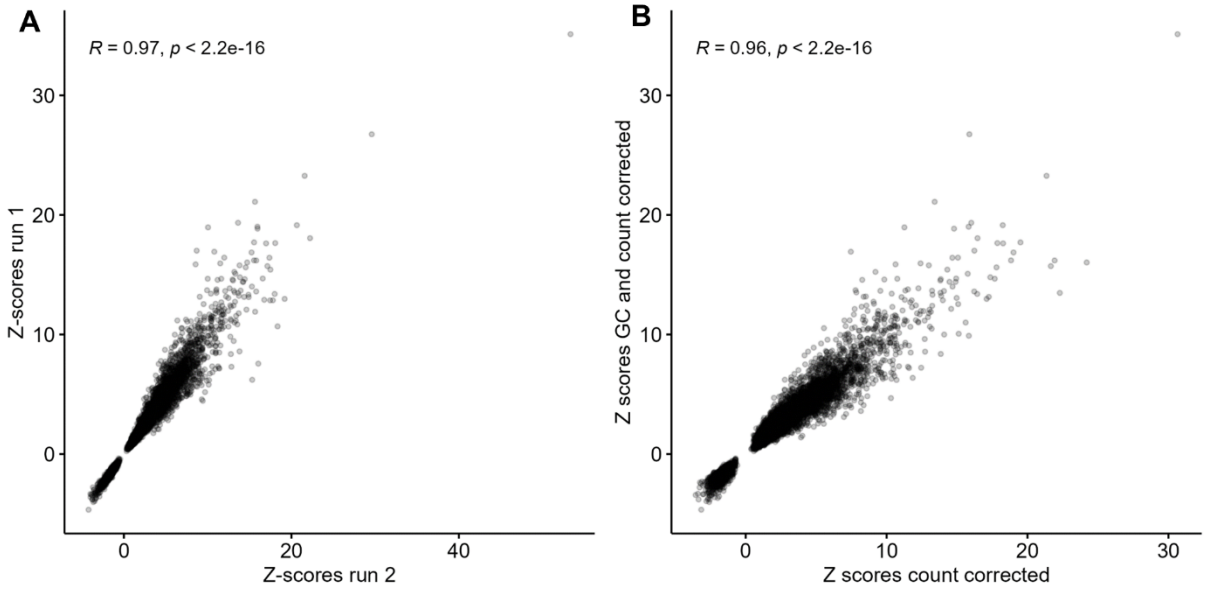


Figure S6. Accounting for GC content has minimal impact on the peak-gene link statistics.

(**A**) Distribution of Z-scores for 2 analyses of all peak-gene links with |Pearson R| > 0.1 using the same model. The variability is due to the stochastic sampling of peaks to create null distributions. (**B**) Distribution of statistics comparing the Z-scores model matching peaks for both GC percent and counts (y-axis) or counts only (x-axis). On each plot, we added the Pearson R coefficient and corresponding P-value.

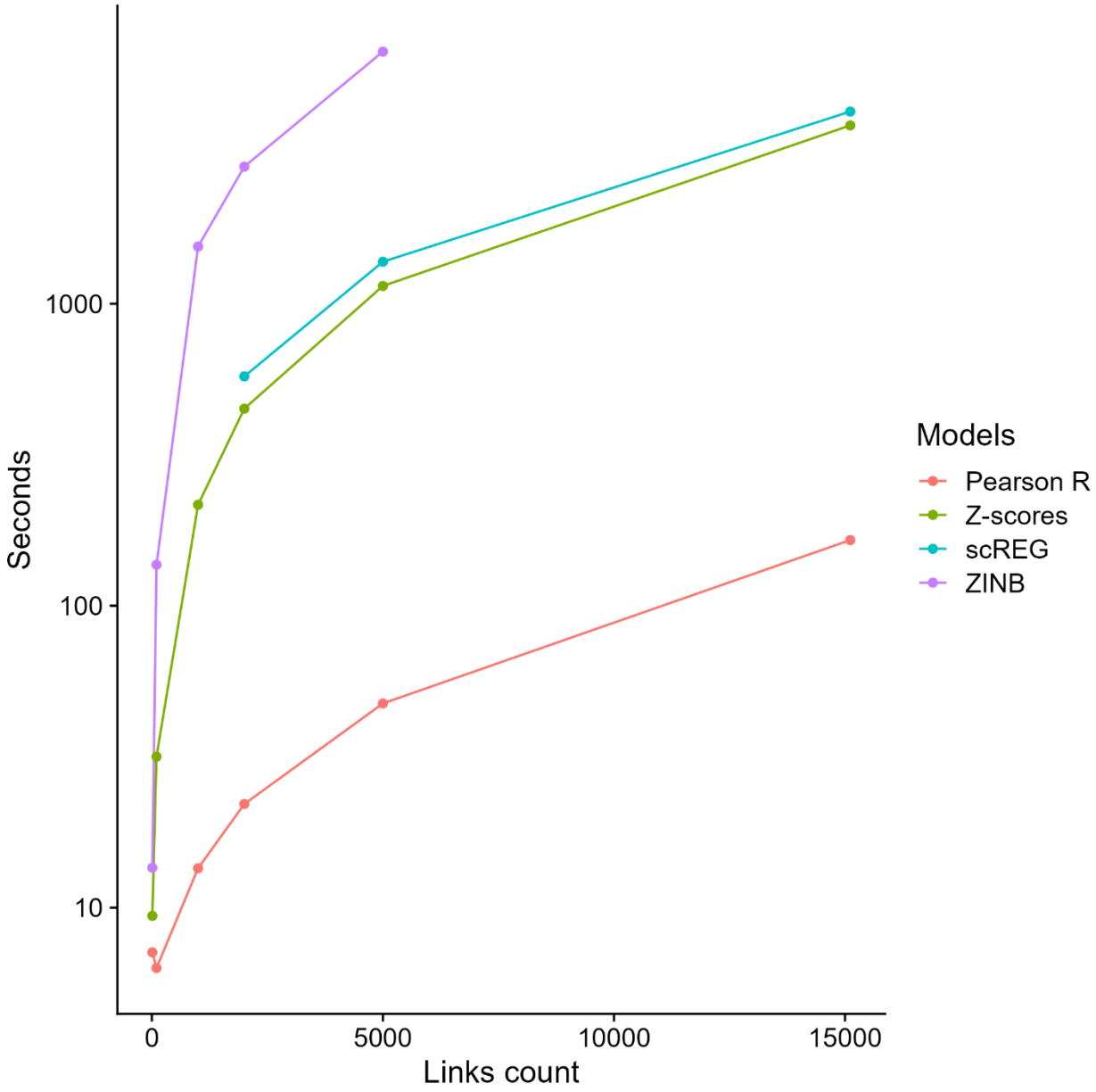


Figure S7. The Pearson R provides an important scalability advantage.

We benchmarked times to run each of the 4 models tested using 1 core with an AMD Ryzen 7 5800X 3,8GHz processor. For each model we tested 10, 100, 1000, 5000 and 15000 links except for ZINB which shows poor scalability. scREG returned errors for inputs with < 2000 links.

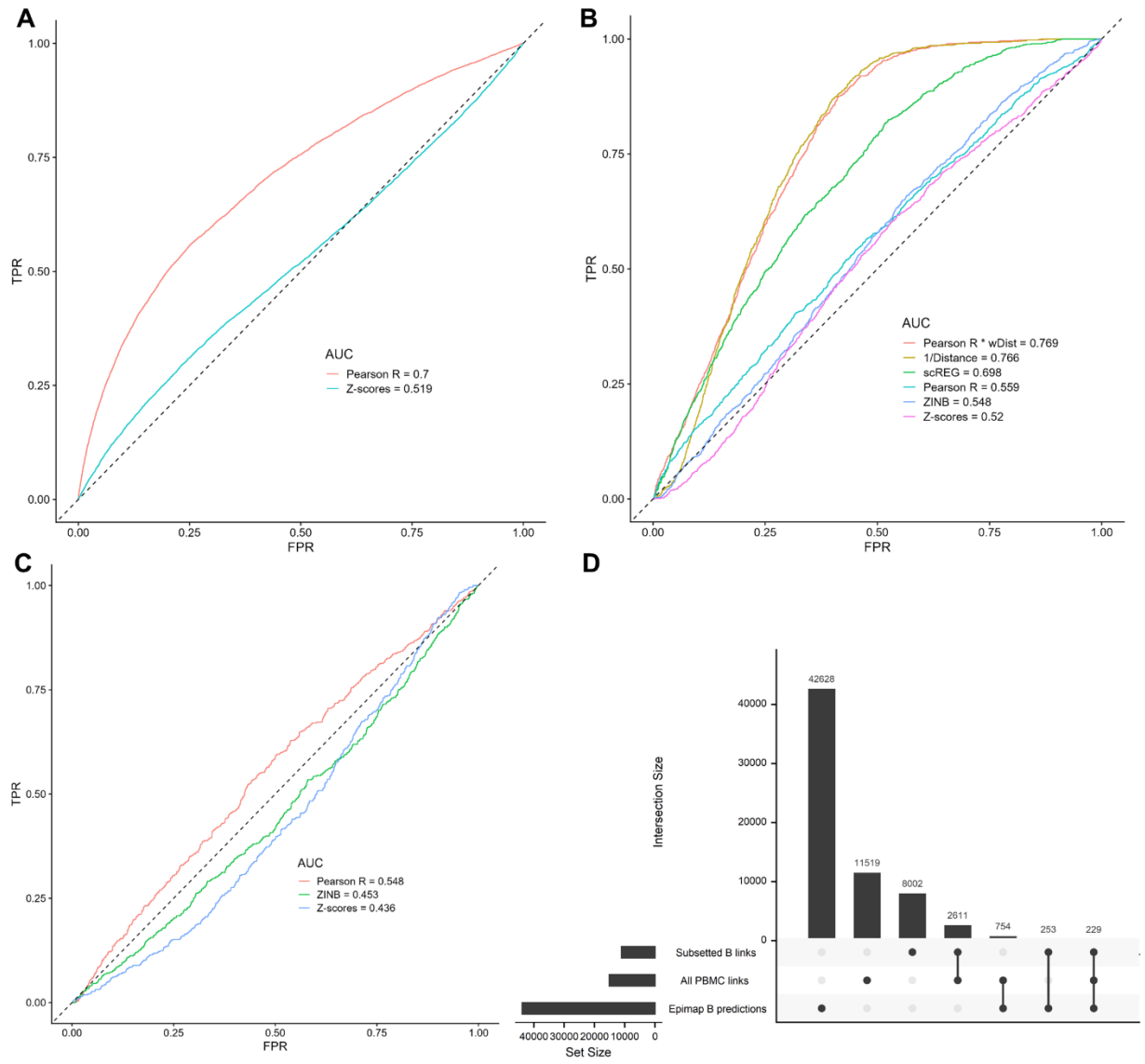


Figure S8. The Pearson R method more accurately validates Epimap-predicted links between cCRE and target genes in B cells.

(**A**) We used the Pearson R and Z-scores methods to detect links between ATACseq peaks and target genes (590,842 links with |Pearson R| >0.01) in the complete (i.e., using all PBMC to compute statistics) PBMC multiomic dataset. Then, we performed Receiving Operating Curves (ROC) analyses to compare the identified peak-gene links from the multiomic data with regulatory links in B cells predicted by the Epimap Project. (**B**) As in A, but using a smaller set of links defined using a more stringent statistical threshold (15,113 links with |Pearson R| >0.1). All cell-types are used to identify links, except for scREG which by design output link scores by cell-type (in this case, B cells). (**C**) As in B, but limiting these ROC analyses to links between ATACseq peaks and target genes with |Pearson R| >0.1 that were found in the B cells subset of the PBMC multiomic dataset. (**D**) Upset plot that shows the intersections of links identified between ATACseq peaks and target genes using either the full PBMC multiomic dataset or only the B cells subset with cCRE-gene regulatory links in B cells as predicted by the Epimap Project. ZINB; zero-inflated negative binomial, wDist; weighted distance (e(-distance/200kb)), TPR, true positive rate; FPR, false positive rate.

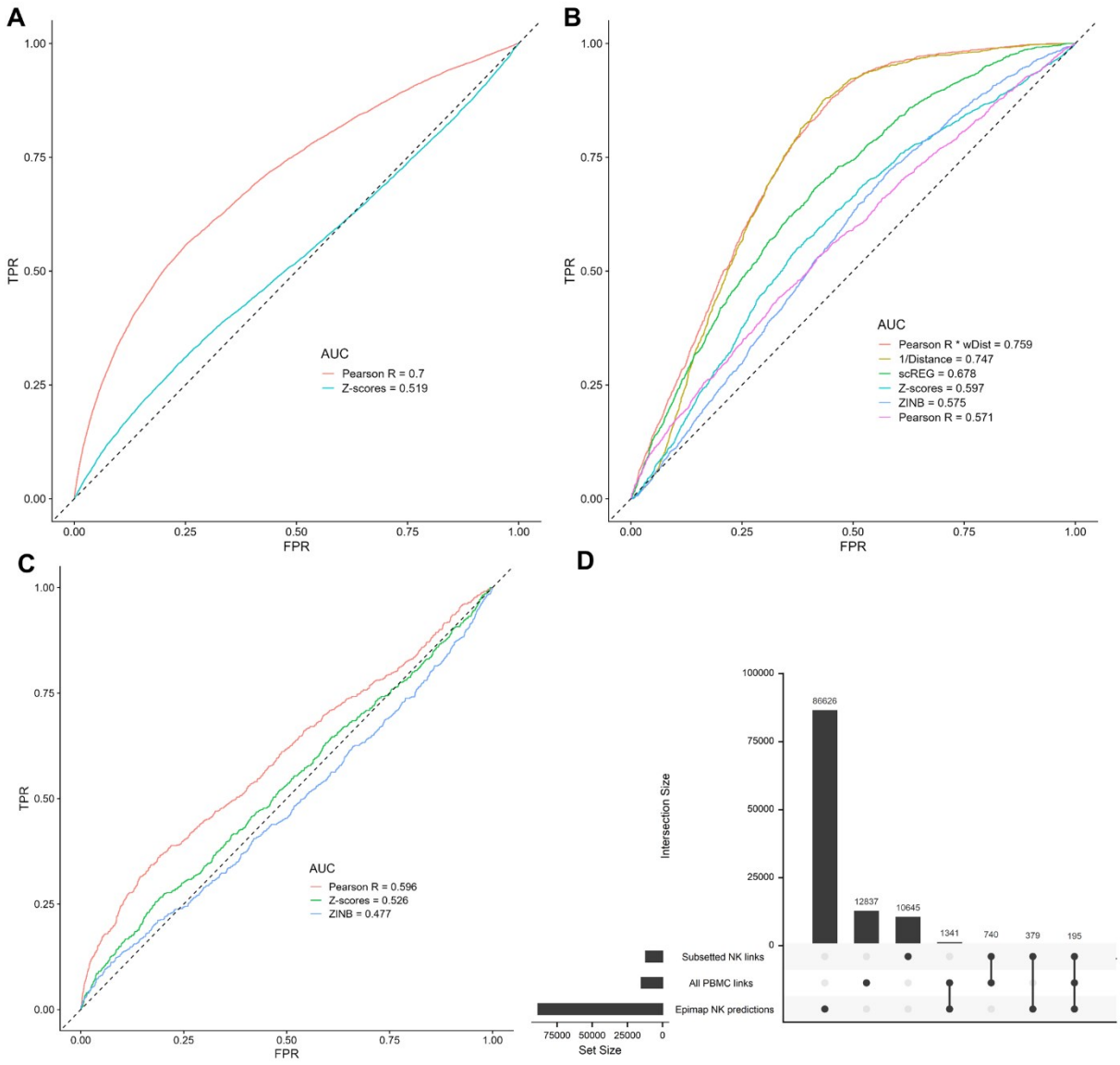


Figure S9. The Pearson R method more accurately validates Epimap-predicted links between cCRE and target genes in NK cells.

(**A**) We used the Pearson R and Z-scores methods to detect links between ATACseq peaks and target genes (590,842 links with |Pearson R| >0.01) in the complete (i.e., using all PBMC to compute statistics) PBMC multiomic dataset. Then, we performed Receiving Operating Curves (ROC) analyses to compare the identified peak-gene links from the multiomic data with regulatory links in NK cells predicted by the Epimap Project. (**B**) As in A, but using a smaller set of links defined using a more stringent statistical threshold (15,113 links with |Pearson R| >0.1). All cell-types are used to identify links, except for scREG which by design output link scores by cell-type (in this case, NK cells). (**C**) As in B, but limiting these ROC analyses to links between ATACseq peaks and target genes with |Pearson R| >0.1 that were found in the NK cells subset of the PBMC multiomic dataset. (**D**) Upset plot that shows the intersections of links identified between ATACseq peaks and target genes using either the full PBMC multiomic dataset or only the NK cells subset with cCRE-gene regulatory links in NK cells as predicted by the Epimap Project. ZINB; zero-inflated negative binomial, wDist; weighted distance (e(-distance/200kb)), TPR, true positive rate; FPR, false positive rate.

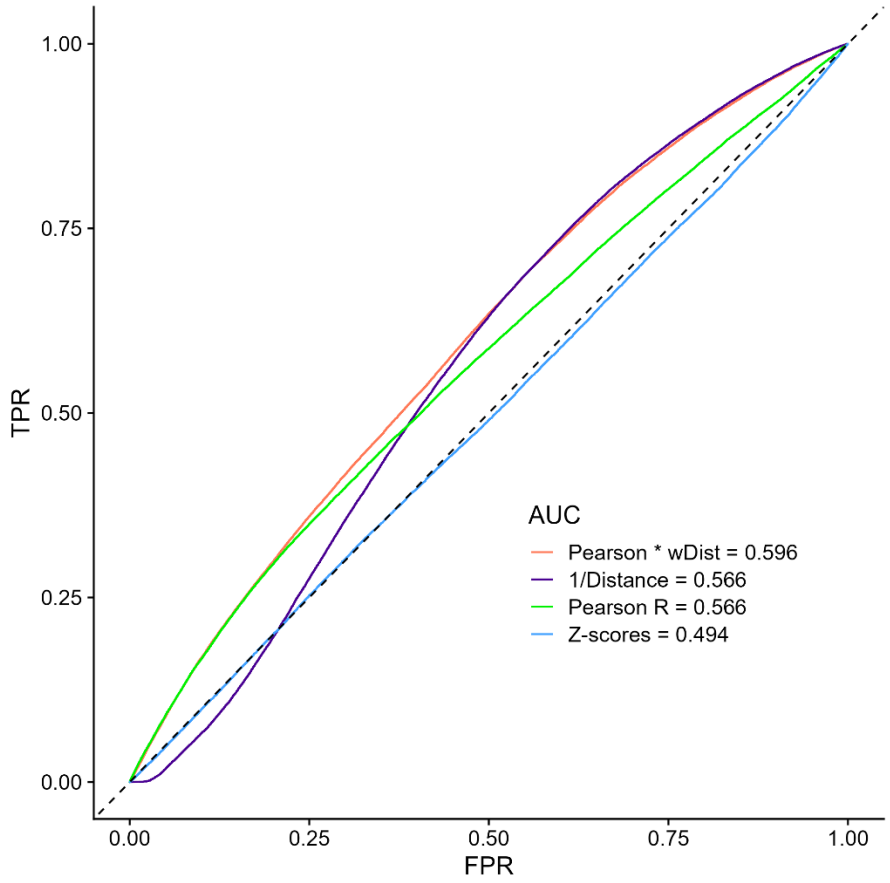


Figure S10. The Pearson R method more accurately validates PCHi-C-predicted links.

We used the Pearson R and Z-scores methods to detect links between ATACseq peaks and target genes (590,842 links with |Pearson R| >0.01) in the PBMC multiomic dataset. Then, we performed Receiving Operating Curves (ROC) analyses to compare the identified peak-gene links from the multiomic data with links found in PCHi-C (see Methods). wDist; weighted distance (e(-distance/200kb)), TPR, true positive rate; FPR, false positive rate.

Chapter 4: Atrial fibrillation variant-to-gene prioritization through cross-ancestry eQTL and single-nucleus multiomic analyses

Francis J.A. Leblanc[1,2], Xuexin Jin[3,4], Kai Kang[5], Chang Jie Mick Lee[6], Juan Xu[7], Lina Xuan[3], Wenbo Ma[3], Hicham Belhaj[1], Neelam Mehta[8], Roger Sik Yin Foo[6], Svetlana Reilly[8], Chukwuemeka George Anene-Nzelu[1,2,6], Zhenwei Pa[3], Stanley Nattel[1,2,9,10,11,*], Baofeng Yang[3,*], Guillaume Lettre[1,2,*]

[1]Montreal Heart Institute, Montreal, Quebec, Canada; [2]Department of Medicine, Université de Montréal, Montréal, Québec, Canada; [3]Department of Pharmacology (State Key Laboratory of Frigid Zone Cardiovascular Disease, Key Laboratory of Cardiovascular Research, Ministry of Education), College of Pharmacy, Harbin Medical University, Harbin, Heilongjiang 150086, P. R. China; [4]Department of Cardiology, The First Affiliated Hospital, Harbin Medical University, Harbin 150001, China; [5]Department of Cardiovascular Surgery, The First Affiliated Hospital, Harbin Medical University, Harbin 150001, China; [6]Cardiovascular Disease Translational Research Programme, Yong Loo Lin School of Medicine, National University of Singapore, Singapore; [7]College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150081, China; [8]Division of Cardiovascular Medicine, Radcliffe Department of Medicine, University of Oxford, John Radcliffe Hospital, Oxford, UK; [9]Department of Pharmacology and Therapeutics, McGill University, Montreal, Quebec, Canada; [10]IHU Liryc and Fondation Bordeaux Université, Bordeaux, France; [11]Institute of Pharmacology, West German Heart and Vascular Center, Faculty of Medicine, University Duisburg-Essen, Essen, Germany.

*Equal co-senior authors.

Correspondence to: Guillaume Lettre, Montreal Heart Institute, 5000 Belanger St, Montreal, Quebec, H1T 1C8, Canada. Email guillaume.lettre@umontreal.ca

3085 words

4 figures and 2 tables

13 supplemental figures and 8 supplemental tables

4.1 ABSTRACT

Atrial fibrillation (AF) is the most common arrhythmia in the world, and is linked to significant morbidity and mortality. Despite advances in the treatment and management of AF, important challenges remain for patients. Human genetics can provide strong therapeutic candidates, but the identification of the causal genes and their functions is difficult. Here, we applied an AF fine-mapping strategy that leverages results from a cross-ancestry genome-wide association study (GWAS), expression quantitative trait loci (eQTLs) from left atrial appendages (LAAs) obtained from two cohorts with distinct ancestry (European and East Asian), and a paired RNAseq and ATACseq LAA single-nucleus assay (sn-multiome). We found that AF-associated LAA eQTLs are largely consistent across ancestries. At 20 AF loci, our co-localization and fine-mapping analyses implicated 25 genes. Furthermore, by integrating our LAA sn-multiome data and other epigenomic datasets with our fine-mapping results, we identified several primary candidate causal AF variants, including rs7612445 at *GNB4* and rs242557 at *MAPT*, for which we propose molecular mechanisms of AF-association at the cellular level. Finally, we showed that the repression of the strongest AF-associated eQTL gene, *LINC01629,* in human embryonic stem cell-derived cardiomyocytes using CRISPR inhibition results in the dysregulation of pathways linked to genes involved in the development of atrial tissue and the cardiac conduction system (*e.g. HCN4*, *PITX2* and *TBX5*).

Keywords: atrial fibrillation, genome-wide association study (GWAS), expression quantitative trait loci (eQTLs), cross-ancestry, single-nucleus multiome, *LINC01629*.

4.2 INTRODUCTION

Atrial fibrillation (AF) is the most common cardiac arrhythmia, significantly impacting health outcomes. It increases the risk of death by 1.5 to 3.5 times and the risk of stroke by ~5 times[1]. In the United States, AF affects a sizable portion of the population: 1 in 3 European Americans and 1 in 5 African Americans are projected to develop AF during their lifetime[387]. AF onset is strongly associated with age, increasing rapidly after age 50: below the age of 50, prevalence is less than 0.5%, whereas by age 80, it exceeds 10%[7]. Given the current aging global population trend, the prevalence of AF is projected to more than double from 2010 to 2030[388].

Substantial advances have been made in our understanding of AF in the last two decades, leading to innovations such as catheter ablation, and improvements in prevention and stroke management. However, current therapies have important limitations. Invasive treatments such as catheter ablation have significant risks, with a complication rate of $4-14\%$[1]. Post-ablation AF recurrence is also common, with a 2-year recurrence rate of 44% in paroxysmal patients[389] and a 1-year recurrence rate of 70% for persistent patients[1]. As a result, repeated procedures are often required. Additionally, pharmacological treatments for this disease remain largely ineffective, failing to reduce onset or progression in up to 85% of patients[53]. A better understanding of the mechanisms of AF is imperative to improve prediction, prevention and the development of new, more specific pharmacological therapies.

Rare mutations in 50 genes have been reported in familial AF[53]. These mutations predominantly occur in ion channels such as *HCN4* and the potassium channels (KCN) group, but also in cardiac transcription factors such as *NKX2-5*, *PITX2* and *TBX5*, and cytoskeletal proteins. AF is also a complex disease with an important genetic component (heritability ~22%)[125]. Recent large-scale genome-wide association studies (GWAS) have identified 150 single nucleotide polymorphisms (SNPs) associated with AF[126,127,390]. However, most of these genetic associations have not yet been functionally characterized. Expression quantitative trait locus (eQTL) analysis can provide a mechanistic interpretation for AF-associated SNPs beyond the closest gene approach. With the advent of single-nucleus transposase-accessible chromatin with sequencing (snATACseq), investigators have shown enrichment of AF-associated SNPs in cardiomyocyte-specific open chromatin regions[287,305,307]. Thus, it is expected that most AF-associated SNPs mediate their risk through cardiomyocyte-specific non-coding regulatory sequences. Yet, linkage

disequilibrium (LD) remains a barrier in the identification of causal variants. This is exacerbated by the strong European-ancestry bias observed in large GWAS and eQTL cohorts[127,391]. Furthermore, studies have revealed that most genes (95% protein-coding and 67% long non-coding RNA [lncRNA]) have one or more eQTL[391]. Therefore, more information is generally required to accurately fine-map GWAS associations.

To fine-map AF GWAS signals and identify causal genes and variants, we performed a cross-ancestry eQTL study using left atrial appendages (LAAs) from AF patients and controls in normal sinus rhythm (SR). We combined GWAS-eQTL co-localization and Bayesian fine-mapping analyses to prioritize candidate causal variants, and leveraged a new LAA single-nucleus multiome dataset (snATACseq + snRNA-sequencig [snRNAseq]) to link regulatory sequences and AF genes. Finally, we performed a CRISPR inactivation (CRISPRi) experiment in human embryonic stem cells-derived cardiomyocytes (hESC-CM) to explore how the most strongly associated lncRNA gene implicated by our eQTL study can modulate AF risk.

4.3 RESULTS

4.3.1 AF-associated cis-eQTLs are concordant across European and East Asian ancestries

To prioritize causal AF variants and genes using a cross-ancestry approach, we profiled two cohorts of participants with or without persistent AF that were recruited on two different continents. We genotyped participants from the Cardiothoracic Surgical Trials Network (CTSN, N=84), a cohort of patients recruited in North America, and a cohort recruited at the University of Harbin in China (Harbin, N=67). We imputed genotypes using TOPMed reference haplotypes and obtained 17,649,215 and 10,537,217 variants in the CTSN and Harbin cohorts, respectively. We projected the genotype data from the CTSN and Harbin cohorts against populations from the 1000 Genomes Project (1000G). As expected, most participants from the CTSN cohort are of European ancestry, although we identified a few participants of admixed African ancestry and one individual of East Asian ancestry (**Fig. S1A-B**). All participants from the Harbin cohort clustered with the Han Chinese in Beijing population from the 1000G dataset (**Fig. S1A-C**). For bulk RNAseq analyses, we obtained left atrial appendages (LAAs) from the same CTSN and Harbin participants. LAA is an ideal tissue to study AF as it is easily accessible during open heart surgery, and a

previous study showed that open chromatin sites found in atrial cardiomyocytes capture most of the AF heritability.[287] After quality-control, we obtained paired genotype and RNAseq data for 31 AF and 31 sinus rhythm (SR, controls) individuals in the CTSN, and 28 AF and 37 SR individuals in the Harbin cohort (**Table S1**).

We focused our cis-eQTL analyses on the 150 sentinel variants recently identified in a cross-ancestry meta-analysis of AF GWAS data (which included 77,690 European and 9,826 Japanese AF cases, and 1,167,040 European and 140,446 Japanese controls)[390]. We restricted our analyses to genes located within one megabase (Mb) from the sentinel AF variants. We found 25 and 17 significant cis-eQTLs (false discovery rate [FDR] <5%) in the CTSN and Harbin cohorts, respectively, of which 11 were significant in both (**Fig. 1A-C**, and **Table S2**). We found evidence of co-localization between the AF GWAS and eQTL signals (posterior probability [PP] that both AF and gene expression share a single causal variant; H4 ≥0.4) at 20 loci (**Table 1**).

While our downstream analyses focus on our LAA eQTL findings, we further confirmed our results against cis-eQTLs from right atrial appendages (RAAs) from the Genotype-Tissue Expression (GTEx) dataset (**Table S3**)[391]. Overall, eQTL results were very concordant when comparing these datasets: this strong replication validates our experiment, but also alleviates concerns of false positive genetic associations due to the multi-ancestry component of the CTSN cohort (**Fig. 1C-E**). We found seven novel AF SNP-eGene pairs that were not identified in GTEx (**Table 1**). When we compared allele frequencies and effect sizes for both gene expression and AF in datasets of European- or East Asian-ancestry, we saw little evidence of heterogeneity (**Table 1**). Lastly, in the CTSN and Harbin cohorts, most of the genes implicated by the eQTL studies were not differentially expressed between AF cases and SR controls (**Table S4**), and we did not find significant AFxSNP interactions for the tested cis-eQTLs, although we acknowledge limited power given our small sample size (**Fig. S2)**.

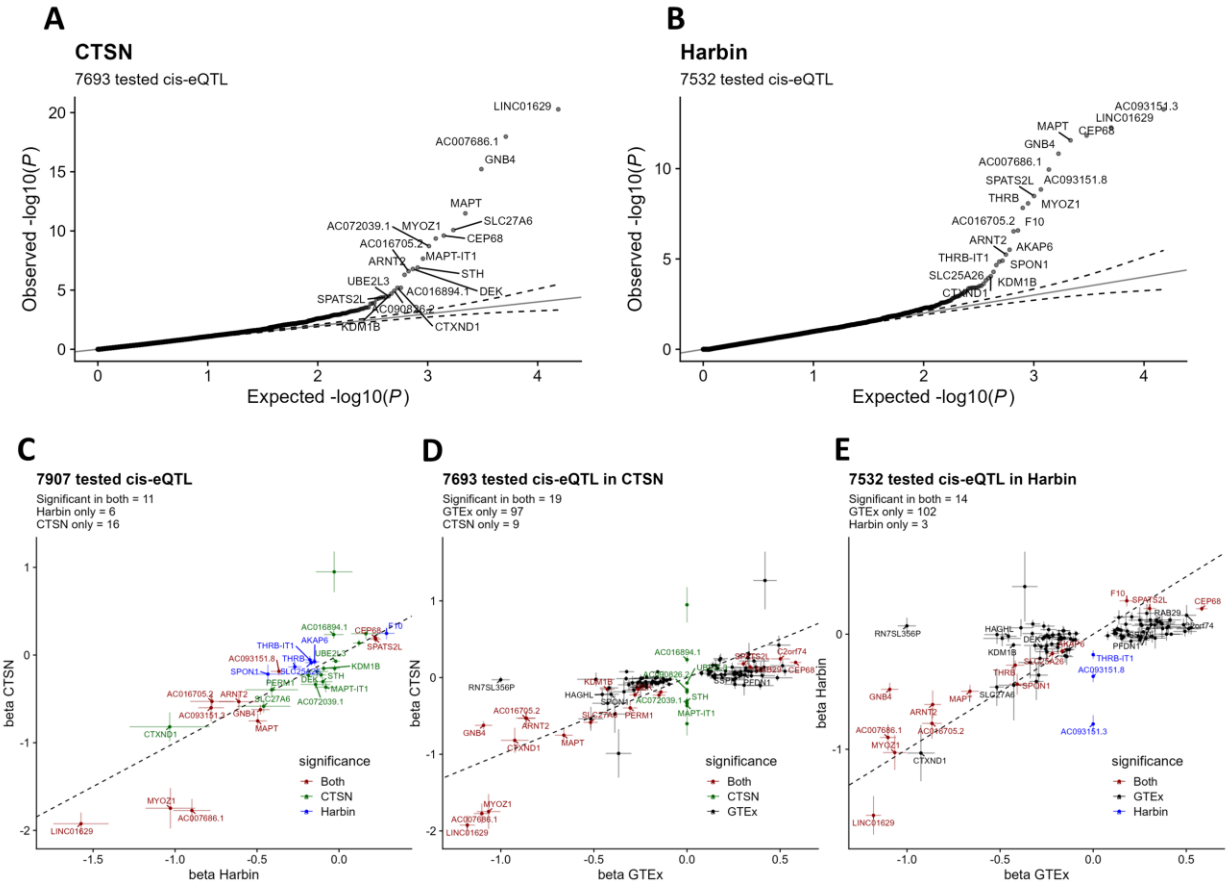


Figure 1. Cross-ancestry LAA eQTLs at AF loci.

(**A-B**) Quantile-quantile (QQ) plots of cis-eQTL –log10(p-values) for 150 sentinel atrial fibrillation (AF) SNPs and nearby genes (<1 Mb) in the (**A**) CTSN and (**B**) Harbin datasets. Dotted lines represent the 95% confidence interval of randomly generated normally distributed p-values. (**C-E**) Scatter plots comparing the eQTLs betas and their standard deviations from left atrial appendages (CTSN and Harbin) and right atrial appendages (Genotype-Tissue Expression; GTEx). (**C**) CTSN vs Harbin, (**D**) CTSN vs GTEx, (**E**) Harbin vs GTEx. We attributed the value 0 to the eQTL if it was not tested in that cohort. (**C-E**) Colors indicate in which dataset(s) the eQTL is significant.eQTLs that were not significant in either of the paired cohorts are not shown for clarity. We added gene labels for the 20 most significant eQTLs in each cohort. Dotted lines represent x=y coordinates.

Table 1. Expression quantitative trait loci (eQTLs) for atrial fibrillation (AF)-associated variants in the CTSN and Harbin cohorts.

We report eQTL results for AF variants that are associated (false discovery rate [FDR] <0.05) with the expression of nearby genes in left atrial appendages from 62 and 65 participants recruited in the CTSN (mostly European-ancestry) and Harbin (East Asian-ancestry) cohorts, respectively, that also show evidence of co-localization with the AF GWAS signals (coloc H4 posterior probability >0.4). To compare allele frequencies and effect sizes (beta) between our eQTL study and genome-wide association (GWAS) results for AF, we also report AF summary statistics from the study by Nielsen et al. (N= 1,030,836, European-ancestry) and Biobank Japan (BBJ, N= 150,272, East Asian-ancestry). Genomic coordinates are on build hg38. Allele frequencies and effect sizes are for the alternate allele. We shaded significant results (FDR <0.05 for the eQTL studies and P <5E-8 for the AF GWAS). eGenes in bold are novel cis-eQTL when compared to GTEx results from right atrial appendages.

				eQTL						AF GWAS					
				CTSN			Harbin			Nielsen et *al.*			BBJ		
rsID	**Chr:Pos**	**REF/ ALT**	**eGene**	**Alt.AF**	**BETA**	**FDR**	**Alt.AF**	**BETA**	**FDR**	**Alt.AF**	**BETA**	**P**	**Alt.AF**	**BETA**	**P**
rs4970418	1:983237	G/A	*PERM1*	0.15	0.40	0.0086	0.07	0.41	0.82	0.17	0.044	7.5E-6	0.076	0.062	0.029
rs2885697	1:41078607	G/T	***AC093151.3***	0.64	-0.60	0.050	0.74	-0.78	2.5E-9	0.65	-0.044	2.9E-10	0.69	-0.005	0.75
			AC093151.8		-0.19	0.012		-0.37	1.5E-5						
rs4951258	1:205722188	G/A	*RAB29*	0.33	-0.14	0.047	0.27	-0.12	0.40	0.42	0.038	2.1E-8	0.36	0.02	0.20
rs2540949	2:65057097	A/T	*CEP68*	0.45	-0.20	9.1E-8	0.32	-0.22	4.5E-8	0.39	-0.066	3.0E-22	0.33	-0.089	3.1E-8
rs3820888	2:200315300	T/C	*SPATS2L*	0.46	-0.18	0.0081	0.74	-0.22	2.8E-5	0.39	0.068	5.8E-24	0.69	0.093	1.4E-8
rs34080181	3:66403767	G/A	*SLC25A26*	0.38	0.09	0.44	0.19	0.17	0.027	0.38	-0.045	1.3E-10	0.22	-0.062	9.1E-4
rs1278493	3:136095167	G/A	***AC072039.1***	0.57	-0.33	6.9E-7	0.14	-0.14	1.00	0.57	-0.039	8.8E-9	0.20	-0.04	0.048
rs7612445	3:179455191	G/T	*GNB4*	0.28	0.62	6.7E-13	0.12	0.48	3.0E-7	0.19	0.049	4.8E-9	0.18	0.068	9.2E-4
rs223449	4:102791180	A/T	*BDH2*	0.44	0.15	0.023	0.41	0.03	1.00	0.49	0.036	7.1E-8	0.46	0.039	9.9E-3
rs2012809	5:128854670	A/G	*SLC27A6*	0.82	-0.58	3.1E-8	0.93	-0.46	0.70	0.79	0.058	4.9E-10	0.92	0.039	0.16
rs3756687	5:137866004	A/G	*FAM13B*	0.17	0.21	9.0E-5	0.060	-0.05	1.00	0.19	0.099	1.1E-31	0.017	0.036	0.52
rs34969716	6:18209878	G/A	*KDM1B*	0.27	0.15	0.0044	0.31	0.09	0.052	0.31	0.07	1.6E-19	0.35	0.071	2.6E-5
rs60212594	10:73654586	G/C	*MYOZ1*	0.15	1.75	1.6E-7	0.19	1.03	6.2E-5	0.14	-0.118	9.2E-35	0.16	-0.065	0.0014
rs2316443	13:113210523	G/A	*F10*	0.22	-0.25	0.073	0.36	-0.29	8.3E-4	0.23	-0.045	2.4E-8	0.44	-0.021	0.16
rs11156751	14:32521231	T/C	*AKAP6*	0.28	0.08	0.81	0.41	0.15	0.0055	0.29	0.072	6.9E-21	0.35	0.099	7.6E-10
rs10873298	14:76960182	C/T	*AC007686.1*	0.53	-1.77	5.0E-15	0.69	-0.90	1.7E-6	0.63	-0.04	7.1E-9	0.66	-0.05	0.0014
			LINC01629		-1.92	2.5E-17		-1.57	1.9E-8						
rs12908004	15:80384583	A/G	***AC016705.2***	0.17	0.53	7.4E-5	0.05	0.77	9.2E-4	0.16	0.073	4.1E-16	0.063	0.142	1.5E-6
			ARNT2		0.53	1.5E-4		0.61	0.0093						
			CTXND1		0.82	0.0015		1.03	0.074						
rs242557	17:45942346	G/A	*MAPT*	0.38	0.75	1.4E-9	0.61	0.50	7.1E-8	0.38	-0.031	1.4E-5	0.46	-0.078	2.0E-7
			MAPT-IT1		0.37	7.8E-6		0.08	1.00						
			STH		0.30	3.8E-5		0.10	1.00						
rs6089752	20:62557186	C/T	*MIR1-1HG-AS1*	0.60	-0.23	0.0089	0.51	-0.19	0.50	0.52	0.033	2.2E-6	0.49	0.071	4.0E-6
rs5754508	22:21644940	C/G	***UBE2L3***	0.17	0.07	0.0071	0.31	0.02	1.00	0.19	0.036	1.0E-4	0.36	0.069	1.4E-4

4.3.2 Multi-ancestry fine-mapping of AF-associated loci

Statistical fine-mapping should help prioritize candidate causal variants at the AF-associated loci with evidence of co-localization (**Table 1**). We derived 95% credible sets using association summary statistics from the AF GWAS and the CTSN/Harbin eQTL studies. We further filtered this list and focused on nine AF loci with GWAS-eQTL co-localization and at least one strong candidate causal variant (posterior inclusion probability [PIP] >0.1 [**Table 2** and **Table S5**]). These nine loci implicate 14 different genes; we discuss their biology and potential link to AF pathophysiology in **Table S6**.

Because the eQTL studies are small in comparison to the sample size of the AF GWAS, the 95% credible sets for the eQTL signals were often larger than the 95% credible sets for the AF signals (**Table S5**). However, because there is evidence of co-localization and since the two eQTL studies have different ancestries (and thus likely different linkage disequilibrium patterns), we reasoned that intersecting the 95% credible sets would enrich for candidate causal AF variants. Using this approach, we were able to limit to a maximum of six the number of potential causal variants at the 9 AF loci mentioned above (**Table 2** and **Table S5**), including three loci with only one candidate variant (*KDM1B*, *ARNT2*/*CTXND1*/*AC016705.2*, *MAPT*/*STH*/*MAPT-ITI*).

4.3.3 Variant-to-gene (V2G) prioritization using single-nucleus multiomic data

To gain insights into the regulatory mechanisms by which AF-associated variants modulate disease risk, we integrated fine-mapped AF variants described above with our single-nucleus (sn) multiome dataset (paired ATAC and RNAseq in the same nuclei) generated from LAAs obtained from three AF and four SR human donors (**Methods**). For completeness, we also queried publicly available data from the *cis*-element Atlas (CATlas; 1,323,041 nuclei ATACseq data from 30 adults and 15 fetal human tissue types)[287], EpiMap (harmonized and imputed missing epigenomic [18 marks/assays from ENCODE, Roadmap and GGR] in 859 biological samples [n=3030 observed and n=14952 imputed datasets])[370], and ENCODE[392]. For this annotation, we focused on variants with fine-mapping posterior inclusion probability (PIP) >0.1 in the AF GWAS and at least one of the two eQTL studies; this represented 15 variants at nine loci (**Table 2** and **Table S7**). For all these loci, the graphical representation of the association results and functional annotations is in **Figures S3-S12**. Below, we emphasize two compelling examples of V2G prioritization for AF.

Table 2. Functional annotation of AF-associated and eQTL variants prioritized by Bayesian fine-mapping.

For variants with posterior inclusion probability (PIP) >0.1 in the AF GWAS and at least one of the eQTL study, we retrieved functional annotations from our multiome single-cell RNA-sequencing and ATAC-sequencing experiment, the *cis*-element ATLAS (CATlas), Epimap and ENCODE. ND; not determined because the association was not significant in this dataset, NC; not in the 95% credible set, enhD; distal enhancer, enhP; proximal enhancer.

Sentinel GWAS variant	eGene	Prioritized variant	CHR:POS (hg38)	PIP AF GWAS	PIP CTSN	PIP Harbin	Left atrial appendage multiome		*Cis*-element ATLAS (CATlas)	Epimap	ENCODE
							Prioritized variant in cardiomyocyte or any cell-type ATAC peak	Link between ATAC peak and eGene promoter (in cardiomyocyte or any cell-types)	Prioritized variant in cardiomyocyte ATAC peak	Prioritized variant in element linked to eGene in heart	
rs4970418	*PERM1*	rs74045046	1:976536	0.141	0.282	ND					enhD
		rs56028034	1:981282	0.1	0.202	ND					
rs7612445	*GNB4*	rs7612445	3:179455191	0.422	0.5	0.333	Yes	Yes	Yes	Yes	
		rs7634416	3:179455436	0.348	0.5	0.333					
rs3756687	*FAM13B*	rs3756687	5:137866004	0.791	0.333	ND					
		rs7722600	5:137859073	0.105	0.333	ND			Yes		enhD
rs34969716	*KDM1B*	rs34969716	6:18209878	0.999	0.137	ND	Yes	Yes	Yes		enhD
rs11156751	*AKAP6*	rs7140396	14:32514611	0.283	ND	0.307					
rs10873298	*AC007686.1*	rs12889775	14:76959734	0.134	0.25	NC					
	LINC01629	rs12889775	14:76959734	0.134	0.25	NC					
	AC007686.1	rs10873298	14:76960182	0.264	0.25	NC					enhP
	LINC01629	rs10873298	14:76960182	0.264	0.25	NC					enhP
	AC007686.1	rs10873299	14:76960368	0.21	0.25	NC					enhP
	LINC01629	rs10873299	14:76960368	0.21	0.25	NC					enhP
	AC007686.1	rs8181996	14:77427469	0.264	0.25	NC					
	LINC01629	rs8181996	14:77427469	0.264	0.25	NC					
rs12908004	*AC016705.2*	rs12908004	15:80384583	1	0.989	0.792					DNase and H3K4me3 mark
	ARNT2	rs12908004	15:80384583	1	0.984	0.887					DNase and H3K4me3 mark
	CTXND1	rs12908004	15:80384583	1	0.222	ND					DNase and H3K4me3 mark
rs242557	*MAPT*	rs242557	17:45942346	0.988	1	1	Yes	Yes	Yes	Yes	enhD
	MAPT-IT1	rs242557	17:45942346	0.988	0.997	ND	Yes		Yes	Yes	enhD
	STH	rs242557	17:45942346	0.988	0.991	ND	Yes		Yes	Yes	enhD
rs6089752	*MIR1-1HG-AS1*	rs6089753	20:62556900	0.139	0.393	ND					

GNB4, which encodes G Protein Subunit Beta 4, is one of the strongest cis-eQTL that we detected and it co-localizes with the AF GWAS signal in both the CTSN and Harbin cohorts. In particular, intersection of the GWAS and eQTL fine-mapped results at this locus prioritized two variants, rs7612445 and rs7634416, with PIP >0.1 (**Fig. 2A**). In our LAA snRNAseq data, *GNB4* is expressed in most cell-types, with high expression in pericytes and endothelial, endocardial and myeloid cells, but relatively low levels in cardiomyocytes (**Fig. 2B**). However, one of the fine-mapped variants, rs7612445, overlaps with a cardiomyocyte-specific ATACseq peak (**Fig. 2C**), and the ATACseq signal at this peak is correlated with *GNB4* expression in cardiomyocytes (Pearson's R=0.28), but not when we considered all cell-types (**Fig. 2D**)[322]. The same variant is also prioritized in the CATlas and EpiMap databases (**Table 2** and **Table S7**).

When we genotyped rs7612445 in six out of seven donors who provided LAAs for the sn-multiome experiment, all but one individual were homozygous for the G-allele. Interestingly, the one individual with the GT genotype had increased chromatin accessibility and higher *GNB4* expression in cardiomyocytes (**Fig. 2D-E**), consistent with the cis-eQTL effect detected by bulk RNAseq in the CTSN and Harbin cohorts (**Fig. 2F**). While rs7612445 and rs7634416 are in strong LD (r^2~1 in European and East Asian populations), our results suggest that rs7612445 is the more likely AF causal variant. Our finding is also consistent with a previous report that used electrophoretic mobility shift assay in cardiomyocytes derived from induced pluripotent stem cells to show that the rs7612445-T allele increases binding with the NKX2-5 transcription factor[219]. Thus, we posit that rs7612445 is the causal variant at this AF locus and mediates its effect on disease through the regulation of *GNB4* in cardiomyocytes.

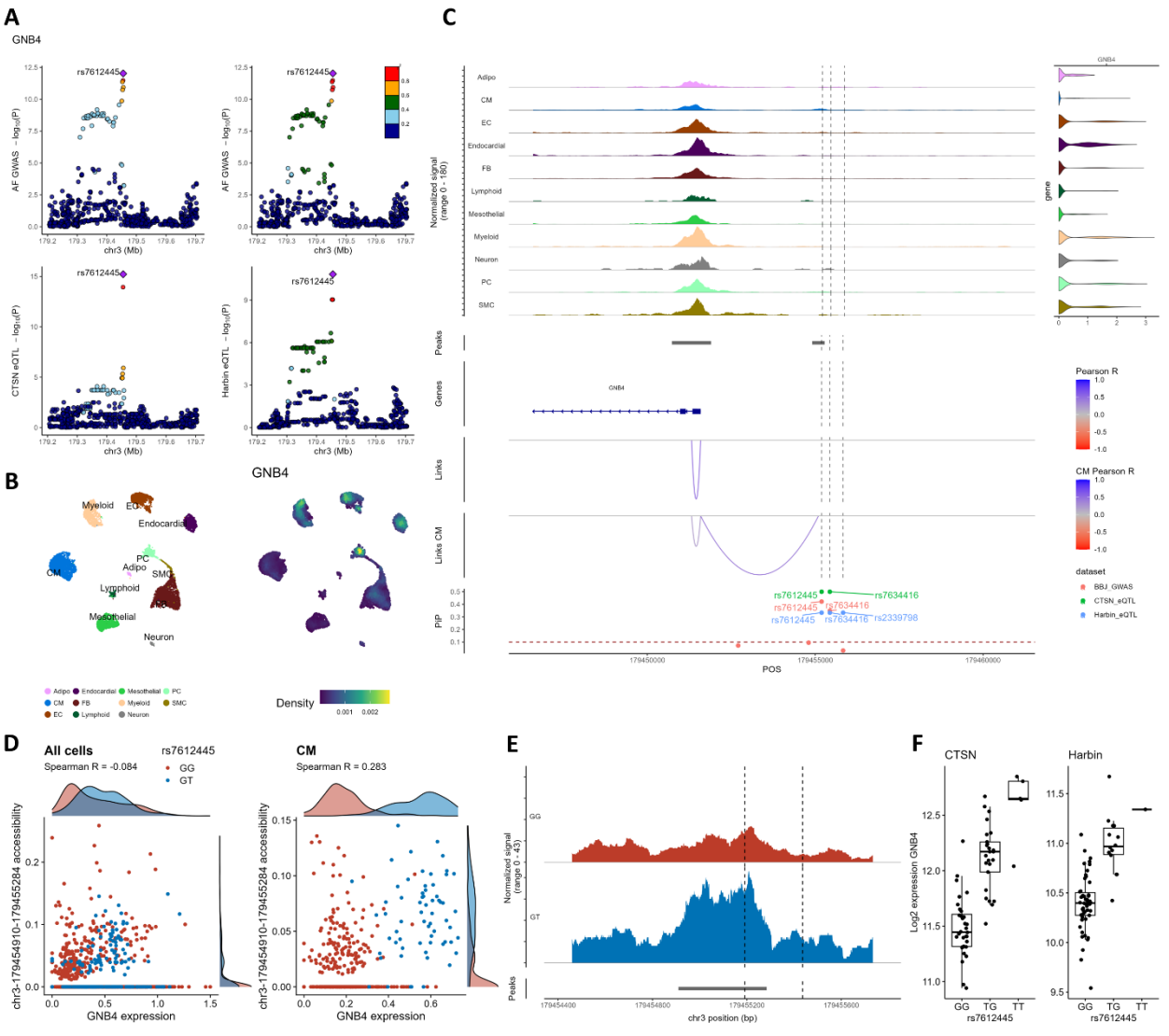


Figure 2. Fine-mapping and annotation of the *GNB4* locus.

(**A**) The top panels show –log10(p-values) from the cross-ancestry atrial fibrillation (AF) GWAS (y-axis) against genomic coordinates (x-axis, hg38) for a 500kb window centered on the sentinel AF SNP. SNPs are colored based on the 1000 Genome Project linkage disequilibrium (LD) r^2 with the lead SNP in the European (left) and East Asian (right) super-populations. In the bottom panels, we report the eQTL –log10(p-values) in the CTSN and Harbin cohorts for *GNB4* expression in left atrial appendages (LAAs). LD is based on the European super-population for CTSN (left) and the East Asian super-population for Harbin (right). (**B**) Uniform manifold approximation and projection (UMAP) of left atrial appendage (LAA) single-nucleus multiome cell-types (left) and *GNB4* expression density (right). (**C**) LAA single-nucleus multiome genomic context of the prioritized AF-associated variants. From top to bottom: The first track shows ATACseq read coverage at the *GNB4* locus (left) paired with violin plots of *GNB4* expression aggregated by cell-type (right). The second track shows ATACseq peaks. The third track shows the gene annotation (exon, intron) for the genes found at the locus. The fourth and fifth tracks highlight links between ATACseq peaks and gene promoters identified in either all cell-types (Links) or specifically in cardiomyocytes (Links CM). We use Pearson correlation tests between ATACseq peak accessibility and gene expression (in the same nucleus) to derive links. Only links with |Pearson R| > 0.2

are shown. Link heights are proportional to their |Pearson R| in the range indicated in the legend. In the final track, we report AF GWAS and eQTL fine-mapping posterior inclusion probabilities (PIP). (**D**) Scatter plots of the chr3:179454910-179455284 peak accessibility against *GNB4* expression colored by the genotype of the prioritized variant (rs7612445) in the six genotyped individuals of our single nucleus multiome LAA data (one GT and five GG. (**E**) ATACseq read coverage of the chr3:179454910-179455284 peak (across all cell-types) aggregated by genotype showing greater accessibility for the individual with the T-allele. (**F**) rs7612445-*GNB4* eQTL boxplots in the CTSN and Harbin bulk RNAseq cohorts. Adipo; Adipocytes, CM; Cardiomyocytes, EC; Endothelial cells, FB; Fibroblasts, PC; Pericytes, SMC; Smooth muscle cells.

At the *MAPT* locus on chromosome 17 (encoding the microtubule-associated protein tau), we detected a strong co-localization between the AF GWAS signal and the expression of three genes – *MAPT*, *MAPT-IT1*, and *STH* – in the CTSN and Harbin datasets (H4 PP >0.9, **Table S5**). We focused our downstream analyses on *MAPT* given that *MAPT-IT1* and *STH* are expressed at very low levels in our sn-multiome data (**Figures S9-10**) and have not been implicated in cardiac phenotypes in the past (**Table S6**). Statistical fine-mapping of the AF GWAS and eQTL datasets prioritized a single variant in the 95% credible sets (rs242557) with high confidence (PIP >0.95, **Fig. 3A** and **Table 2**). In the LAA sn-multiome data, *MAPT* is highly expressed in cardiomyocytes (**Fig. 3B**). rs242557 intersects with an ATACseq peak correlated with the expression of *MAPT* when we considered all cell-types (**Fig. 3C**), as well as with an ATACseq peak opened in a broad range of cell-types in CATlas (including cardiomyocytes) and a distal enhancer element in ENCODE (**Table 2** and **Table S7**). Among the six donors who provided LAA for the sn-multiome experiment and that we could genotype, four were heterozygous and two were homozygous for the reference G-allele. GA heterozygous donors showed higher *MAPT* expression (in all cell-types or considering only cardiomyocytes, **Fig. 3D**), consistent with the bulk eQTL results in the CTSN and Harbin cohorts. However, we found no difference in chromatin accessibility for this ATACseq peak based on genotypes at rs242557 (**Fig. 3D-F**), maybe because our sample size is too small or because the variant affects gene expression without modulating chromatin accessibility.

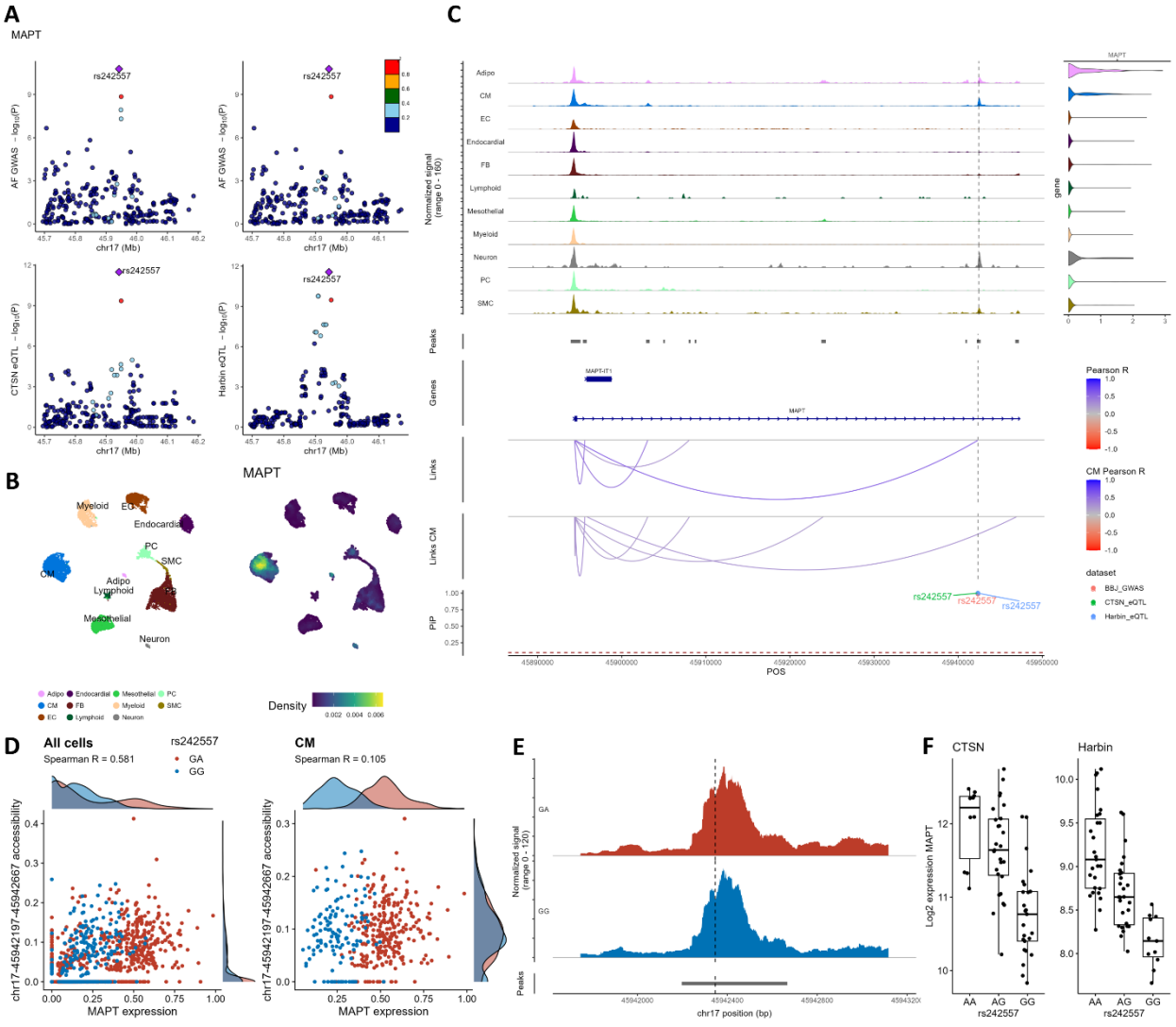


Figure 3. Fine-mapping and annotation of the MAPT locus.

(**A**) The top panels show –log10(p-values) from the cross-ancestry atrial fibrillation (AF) GWAS (y-axis) against genomic coordinates (x-axis, hg38) for a 500kb window centered on the sentinel AF SNP. SNPs are colored based on the 1000 Genome Project linkage disequilibrium (LD) r^2 with the lead SNP in the European (left) and East Asian (right) super-populations. In the bottom panels, we report the eQTL –log10(p-values) in the CTSN and Harbin cohorts for *MAPT* expression in left atrial appendages (LAAs). LD is based on the European super-population for CTSN (left) and the East Asian super-population for Harbin (right). (**B**) Uniform manifold approximation and projection (UMAP) of left atrial appendage (LAA) single-nucleus multiome cell-types (left) and *MAPT* expression density (right). (**C**) LAA single-nucleus multiome genomic context of the prioritized AF-associated variants. From top to bottom: The first track shows ATACseq read coverage at the *MAPT* locus (left) paired with violin plots of *MAPT* expression aggregated by cell-type (right). The second track shows ATACseq peaks. The third track shows the gene annotation (exon, intron) for the genes found at the locus. The fourth and fifth tracks highlight links between ATACseq peaks and gene promoters identified in either all cell-types (Links) or specifically in cardiomyocytes (Links CM). We used Pearson correlation tests between ATACseq peak accessibility and gene expression (in the same nucleus) to derive links. Only links with |Pearson R| > 0.2

are shown. Link heights are proportional to their |Pearson R| in the range indicated in the legend. In the final track, we report AF GWAS and eQTL fine-mapping posterior inclusion probabilities (PIP). (**D**) Scatter plots of the chr17:45942197-45942667 ATACseq peak accessibility against *MAPT* expression colored by the genotype of the prioritized variant (rs242557) in the six genotyped individuals of our single-nucleus multiome LAA data (two GT and four GG). (**E**) ATACseq read coverage of the chr17:45942197-45942667 ATACseq peak aggregated by genotype (across all cell-types). (**F**) rs242557-MAPT eQTL boxplots in the CTSN and Harbin bulk RNAseq datasets. Adipo; Adipocytes, CM; Cardiomyocytes, EC; Endothelial cells, FB; Fibroblasts, PC; Pericytes, SMC; Smooth muscle cells.

4.3.4 LINC01629 repression alters key AF genes expression in hESC-CMs

Intriguingly, the strongest eQTL in both the CTSN and Harbin datasets implicated genotypes at the sentinel SNP rs10873298 and the expression of an uncharacterized lncRNA (*LINC01629*) and a pseudo-gene (*AC007686.1*) in LAAs (**Fig. 1** and **Table 1**). Expression of both genes are co-localized with the AF GWAS signal at the locus (H4 >0.9, **Table S5**) and we could resolve the 95% credible set to four variants in CTSN (each with PIP >0.1, **Table 2**). We could not functionally annotate any of these four variants using the sn-multiome LAA, CATlas, or EpiMap data, although two of them (rs10873298 and rs10873299) overlap with a generic proximal enhancer catalogued by ENCODE (**Table 2**). A third prioritized variant, rs12889775, is also of interest for two reasons: First, it is located just next to a CM-specific ATACseq peak and could therefore directly affects the expression of *LINC01629* (**Fig. S5B**). And second, it also maps to a *LINC01629* exon so that it might influence the stability of this lncRNA.

We were able to detect the expression of *LINC01629* in cardiomyocytes in the LAA sn-multiome data (**Fig. S5B**). To gain molecular insights into the role that *LINC01629* can play in AF etiology, we knocked down its expression in hESC-CMs using CRISPRi and performed bulk RNAseq on three biological replicates (**Methods** and **Fig. S13**). For these CRISPRi experiments, we used a guide RNA (gRNA) that maps to the *LINC01629* promoter. We confirmed that CRISPRi repressed *LINC01629* expression when compared to a non-targeting gRNA (**Fig. 4A**). Differential gene expression analysis identified 217 up- and 299 down-regulated genes (FDR <0.1) upon CRISPRi on the *LINC01629* promoter (**Fig. 4B** and **Table S8**). Many of these genes play key roles in CM functions and/or have already been implicated in AF by GWAS (*e.g. TBX5*, *PITX2*, *HCN4*, *SCN5A*)[393-395]. Notably, the most down-regulated gene in this *LINC01629* CRISPRi experiment is *FOXP2*, a transcription factor that was recently implicated in the control of regulatory networks found in pacemaker cells[396]. Our pathway analyses confirmed these observations, for instance highlighting genes implicated in cardiac conduction among the CRISPRi down-regulated genes

(**Fig. 4C**). The expression of *AC007686.1* was not significantly modulated in the CRISPRi experiment (log_2 fold-change = -0.036, nominal P=0.16), although we cannot formerly exclude that this pseudogene also contributes to the AF GWAS signal.

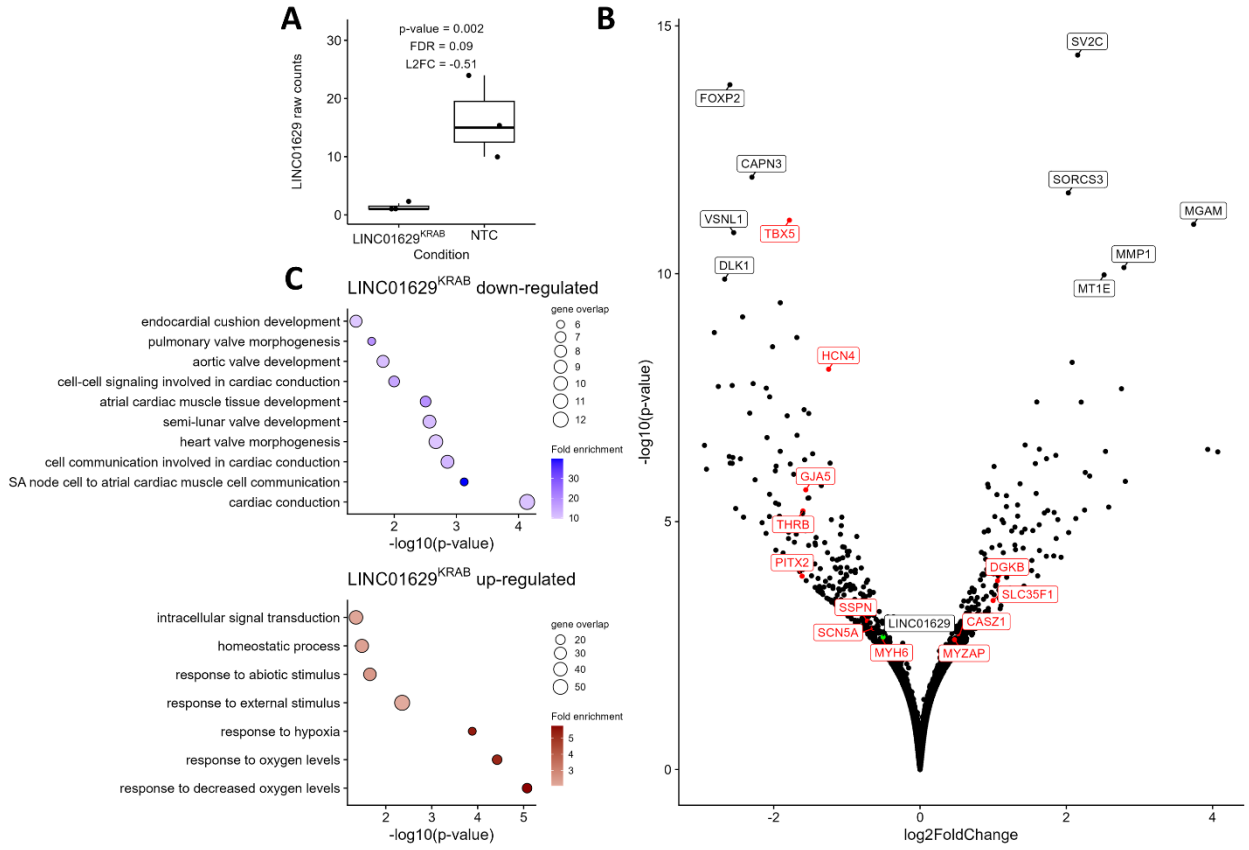


Figure 4. In vitro validation of *LINC01629*.

(**A**) Boxplot of raw read counts for *LINC01629* in human embryonic stem cells-derived cardiomyocytes (hESC-CMs) with CRISPRi targeting the promoter of *LINC01629* (LINC01629^{KRAB}) *vs* a non-targeting control (NTC) guide RNA (gRNA). (**B**) Volcano plot of the differential gene expression analysis carried out by comparing the transcriptome of hESC-CMs treated with a gRNA against the promoter of *LINC01629* vs a negative control (NTC gRNA). We labeled the top 10 genes (black dot and label), the AF prioritized genes by Open Targets (red dot and label) as well as *LINC01629* (green dot and black label). **(C)** Pathway analyses with genes down-regulated (top panel) or up-regulated (down panel)(false discovery rate <0.1) in the CRISPRi experiment.

4.4 DISCUSSION

One strength of our study is that we focused on AF-associated variants identified in a large cross-ancestry GWAS study and that we included LAA samples from different ancestries. Our initial conclusion is that there is little evidence of heterogeneity when comparing the effect of AF-associated variants between European- and East Asian-ancestry populations (**Table 1**). This observation is largely consistent with recent studies that suggest that the genetic architecture of common human diseases, including AF, is concordant between populations (at least when considering common variants)[397]. However, it is important to nuance these conclusions with two important considerations. First, most cross-ancestry GWAS published to date are still very much European-ancestry-centric, and are therefore less likely to yield genetic associations specific to non-European samples (which would increase heterogeneity). Second, the limited sample size of our eQTL datasets (CTSN, N=62; Harbin, N=65) would not allow us to detect small differences in effect sizes on gene expression for the AF-associated variants. Thus, there is still great value in continuing to increase the sample size of non-European-ancestry GWAS and eQTL studies for complex human diseases, including AF, to discover population-specific biology and also to enable more powerful fine-mapping experiments.

Many GWAS variants are eQTLs, yet this overlap does not necessarily imply that the corresponding eGenes are involved in the diseases. To increase the specificity of our strategy, we only considered: (1) loci with strong evidence of co-localization between the AF GWAS signals and gene expression in LAAs and (2) variants with high PIP (**Table 2**). We acknowledge that these stringent criteria could make us miss interesting loci with more complex genetic architecture. Nonetheless, we prioritized many interesting genes for a role in AF (**Table S6**). Some of these genes have previously been linked to AF (*e.g. GNB4*, *MAPT*, *FAM13B*, *ARNT2*) because of roles in heart rhythm, cardiac conduction, or other aspects of cardiomyocyte biology. But many of these genes are implicated in AF for the first time, highlighting once again the power of human genetics to gain novel insights into clinically-relevant human phenotypic variation.

One interesting finding of our study is the discovery that down-regulating the expression of *LINC01629* in hESC-CMs using CRISPRi modifies the expression of many key AF genes, including candidate causal genes implicated in AF GWAS (like *PITX2*, **Fig. 4B**). Pathway analyses confirms this observation by revealing that many of the down-regulated genes impact cardiac

conduction. The T-allele at rs10873298, which is associated with lower *LINC01629* expression in LAAs, is protective against AF (**Table 1**). This observation, combined with results from our CRISPRi experiment suggests that the down-regulation of key genes like *PITX2, GJA5* and *TBX5* in CMs might prevent AF. While it is known that the up-regulation of these genes can promote arrhythmia and AF[394,398-401], loss-of-function and haploinsufficiency of these genes has also been associated with AF[401-404]. This observation underscores the intricate balance necessary for normal conduction in the heart, and that too little or too much of a given gene can lead to heart diseases. It is also important to emphasize that while our acute treatment of CMs with CRISPRi links *LINC01629* to pathways of genes involved in atrial development and cardiac conduction, it does not recapitulate the chronic impact of *LINC01629* over (or under)-expression during a lifetime. Additional work is needed to clarify whether this long non-coding RNA promotes or protects against AF, and how molecularly it modulates the expression of key AF genes.

As for most other complex human diseases tackled by GWAS, progress towards a better understanding of AF pathophysiology has been hampered by challenges to move from genetic associations to genes and variants. In this study, we combined statistical methodologies (co-localization, Bayesian fine-mapping), eQTL analyses in a disease-relevant tissue obtained from donors of different ancestries, and CRISPRi in hESC-CMs to prioritize new variants, regulatory sequences and genes that modulate the risk of developing AF. While we uncovered strong variant and gene candidates for further downstream analyses, we recognize that larger eQTL studies, potentially including other tissues, are required to functionally dissect the ~150 GWAS loci associated with AF. Because we made these discoveries by studying human genetic and phenotypic variation, they promise to yield insights into the causes of AF in humans. Given the unmet need to develop and characterize new molecules to treat (or prevent) AF, this is particularly exciting since candidate drug targets that are supported by genetic evidence are twice as likely to yield effective therapies[108,405].

4.5 METHODS

4.5.1 Participants

All participants provided written informed consent and the project was approved by the Montreal Heart Institute ethical committee (project number: #2011-209). Studies involving human participants recruited at the Oxford site were approved by the local Research Ethics Committee (South Central - Berkshire B Research Ethics Committee, UK; ref: 18/SC/0404). All participants provided written informed consent.

4.5.2 RNA extraction and sequencing

Harbin

RNA isolation

We extracted total RNA with the Qiagen kit. We then processed the total RNA as follows: (1) We tested the RNA samples for possible contamination and degradation using 1% agarose gel electrophoresis; (2) We examined RNA purity and concentration using the NanoPhotometer® spectrophotometer; and (3) We measured RNA integrity and quantity using the RNA Nano 6000 Assay Kit on the Bioanalyzer 2100 system.

Library preparation and sequencing.

We used the Nugen kit with the ribosomal RNA (rRNA) depletion and stranded method to construct the RNA libraries for RNA-seq. Briefly, we depleted rRNA from total RNA using the rRNA Removal Kit as per the manufacturer's instructions. Next, we fragmented the RNA into ~250-~300 bp fragments and reverse transcribed the first strand cDNA using fragmented RNA and dNTPs (dATP, dTTP, dCTP and dGTP). We degraded the RNA using RNase H and synthesized the second strand cDNA using DNA polymerase I and dNTPs (dATP, dUTP, dCTP and dGTP). We then converted the remaining overhangs of double-stranded cDNA into blunt ends using exonuclease/polymerase activities. After adenylation of the 3' ends of DNA fragments, we ligated sequencing adaptors to the cDNA. To select cDNA fragments of preferentially ~250-~300 bp in length, we purified the library fragments using the AMPure XP system. We performed uridine digestion using Uracil-N-Glycosylase, followed by cDNA amplification using PCR. After library construction, we measured the concentration of the library using the Qubit fluorometer and adjusted

it to 1 ng/µL. We used the Agilent 2100 Bioanalyzer to measure the insert size of the acquired library. Finally, we examined the accurate concentration of the cDNA library using qPCR. Finally, we subjected the samples to sequencing on the Illumina NovaSeq 6000 S4 using a paired-ends 150 bp protocol.

CTSN

RNA isolation

We blended samples with the Bullet Blender Storm method using Green RINO Lysis tubes and added 200uL of Quiazol per tube for ~50 mg of tissue, following the manufacturer's protocol for heart tissue. We then extracted RNA using the miRNeasy Qiagen kit (Cat No./ID: 217004), following the manufacturer's protocol.

Library preparation and sequencing

We quantified total RNA using a NanoDrop Spectrophotometer ND-1000 (NanoDrop Technologies, Inc.) and assessed its integrity on a 2100 Bioanalyzer (Agilent Technologies). We depleted rRNA from 250 ng of total RNA using the QIAseq FastSelect - rRNA HMR Kit. We performed cDNA synthesis using the NEBNext RNA First Strand Synthesis and NEBNext Ultra Directional RNA Second Strand Synthesis Modules (New England BioLabs). We carried out the remaining steps of library preparation using the NEBNext Ultra II DNA Library Prep Kit for Illumina (New England BioLabs), and we purchased adapters and PCR primers from New England BioLabs. We quantified libraries using the Kapa Illumina GA with Revised Primers-SYBR Fast Universal kit (Kapa Biosystems) and determined the average size fragment using a LabChip GX (PerkinElmer) instrument. We conducted sequencing on an Illumina NovaSeq 6000 S4 paired-ends 100-bp with a minimum aimed 100M reads per sample.

4.5.3 DNA extraction and genotyping

Harbin

We isolated DNA using a Qiagen kit. We measured the DNA concentration (≥60 ng/µl) and volume (≥30µl) using the Qubit® DNA Assay Kit on a Qubit® 3.0 Fluorometer (Invitrogen, USA). We performed genotyping using the Illumina GSA v3 or ASA genotyping arrays.

CTSN

We isolated DNA from blood, when available, using the Qiagen DNaeasy kit (Cat No./ID: 69504), following the manufacturer's protocol. We quantified the genomic DNA using a 2100 Bioanalyzer (Agilent Technologies) and performed genotyping using the Illumina GSA-24v3 genotyping array.

4.5.4 Genotype quality-control and imputation

We used PLINK 1.9 to handle the genotype files. In the Harbin cohort, we removed one individual with >5% missingness. In the Harbin cohort, we also removed monomorphic SNPs and those with a Hardy-Weinberg equilibrium exact test P-value <1E-6, resulting in 417,862 genotyped SNPs. In the CTSN cohort, we removed monomorphic SNPs, resulting in 531,541 genotyped SNPs (we did not filter based on Hardy-Weinberg equilibrium in CTSN due to the multi-ancestry nature of the cohort). Next, we flipped alleles to match hg19 using Snpflip (https://github.com/biocore-ntnu/snpflip). We performed genotype imputation on the TOPMed Imputation Server using the TOPMed-r2 reference. We retained variants with an imputation R2 >0.3 for downstream analyses. In total, we obtained 10,537,217 and 17,649,215 imputed variants in the Harbin and CTSN cohorts, respectively.

4.5.5 Genetically-defined continental ancestry

We used 2,722 common ancestry informative markers (https://genome.sph.umich.edu/wiki/Exome_Chip_Design) to perform principal component analyses (PCA) on a combined dataset that included samples from the Harbin and CTSN cohorts, as well as populations from the 1000 Genomes Project. We visualize participants using axes of variation obtained by PCA and Uniform Manifold Approximation and Projection (UMAP) calculated with the first 10 principal components (PC).

4.5.6 RNAseq processing and differential expression analysis

We performed the following steps independently for the CTSN and Harbin datasets. We pseudoaligned RNAseq reads to Gencode v32 using Kallisto with the options quant -b 100 --rf-stranded. We aggregated transcripts by genes and quantified them with DESeq2[166]. We removed genes with less than 10 reads. We used DESeq2's Wald test with sex as a covariate, followed by a log2 shrunken transformation (ashr shrinkage estimator[406]) to compare AF differentially expressed

genes between Harbin and CTSN. Lastly, we determined the number of PCs to use as hidden variables (covariates) with the runElbow() function from the PCAForQTL package[407].

4.5.7 eQTL calling

For the eQTL datasets, we obtained 65 and 62 samples in Harbin and CTSN, respectively, for which we had genotypic and transcriptomic data. We first tested eQTLs for 150 sentinel variants recently associated with AF in a cross-ancestry meta-analysis[390]. Later, for co-localization analyses, we retrieved all genetic variants in a 500 kilobases (kb) window centered on the sentinel variants that were significant eQTLs (FDR <0.05). At the *FAM13B* locus, we discovered that rs529526 was miss-annotated as sentinel variant (the GWAS meta-analysis P-value was inferior to rs3756687 and rs7722600) and changed it to rs3756687 for all downstream analyses. To compute eQTLs, we used transformed gene expression matrices derived from the vst() function of DESeq2. We accounted for hidden variables in the RNAseq datasets using PCs (shown to outperform PEER factors[407]) with the PCAForQTL package function runElbow(). For both CTSN and Harbin the elbow for proportion of variance explained occurred at 7 PCs. We then used the MatrixEQTL package[198] with sex and 7 PCs as covariates to find eQTLs with less than one megabase (Mb) between the gene and the SNP. To test for statistical interaction between genotype and disease status, we used the following linear model for sentinel variants only: gene expression levels ~ sex + disease_status + SNP + PC1 + PC2 + PC3 + PC4 + PC5 + PC6 + PC7 + SNP:disease_status.

4.5.8 Co-localization

We retrieved the hg38 positions using the AnnotationHub package (Annotationhub chain: hg19ToHg38.over.chain.gz). For CTSN and Harbin, we merged eQTL summary statistics overlapping with the GWAS summary statistics and ran coloc (https://github.com/chr1swallace/coloc) on each locus with a significant eQTL. We used 87,516 as the number of cases and 1,395,002 as the sample size for the GWAS[390].

4.5.9 Single-nucleus multiome

Sample collection from LAA.

A total of eight patients were initially included in the study; all patients underwent cardiac surgery (coronary artery bypass grafting) in the John Radcliffe hospital at Oxford. Left atrial

appendage biopsies were collected before cardiopulmonary bypass and immediately, rinsed of blood, towel-dried and snap-frozen in liquid nitrogen until use for subsequent experiments.

Single nuclei sample preparation and sequencing.

To isolate nuclei from LAA samples we used a modified version of the 10X multiome nuclei isolation protocol. Unless specifically mentioned in our description bellow, we re-suspended nuclei by pipette mixing slowly 10 times and we used the buffer described by 10X here: https://www.10xgenomics.com/support/single-cell-multiome-atac-plus-gene-expression/documentation/steps/sample-prep/nuclei-isolation-from-complex-tissues-for-single-cell-multiome-atac-plus-gene-expression-sequencing. All steps were performed on ice or maintained at 4°C. We extracted nuclei from tissues using a Singulator™100 with the nuclei manufacturer protocol (fast extraction setting). For each sample, we loaded approximately 25mg of tissue with 1.5ml of nuclei lysis buffer (Tris-HCL pH 7.4 10mM, NaCl 10mM, $MgCl_2$ 3mM, Tween-20 0.1%, Nonidet P40 Substitute 0.1%, Digitonin 0.01%, BSA 2%, DTT 1mM, 0.5U/uL Protector RNase inhibitor [Sigma catalog no. 3335402001], in nuclease free water) in the Singulator™100. Once the run completed, we rinsed the cartridge with an additional 1ml of lysis buffer, passed the solution in two 20um filter (Miltenyi catalogue no. 130-101-812) sequentially, and performed a first round of centrifugation at 500g and 4°C for 5 min in a swing bucket centrifuge. We then removed the supernatant, added 1ml of suspension buffer (PBS, 2% BSA, 0.5U/uL Protector RNase inhibitor), waited 5 min for buffer exchange, re-suspended the nuclei, passed the solution through a 20um filter, rinsed with an additional 1mL of suspension buffer and performed a second centrifugation at 500g and 4°C for 5 min. We then removed the supernatant and re-suspended the nuclei in 100uL of 0.1X lysis buffer (1U/uL Protector RNase inhibitor) by pipette mixing 5 times, waited 2min, added 1mL of wash buffer, pipette mixed 5 times and centrifuged at 500g and 4°C for 5 min. Finally, we re-suspended the nuclei in diluted nuclei buffer, quantified them on a Countess® II FL, and proceeded to loading the chip on the Chromium controller and downstream steps from the manufacturer protocol. We sequenced libraries on a Novaseq 6000 S4 PE100 with a targeted 30,000 paired-reads per cells for the RNA libraries and 60,000 paired-reads per cells for the ATAC libraries.

Alignment and pre-processing.

We aligned FASTQ to Cellranger's GRCh38-2020-A reference (https://support.10xgenomics.com/single-cell-multiome-atac-gex/software/downloads/latest) using the **count** function of cellranger-arc-2.0.0 for each sample. We then aggregated all samples using cellranger-arc **aggr** function. We removed one sample having a low number of detected genes, linked genes and number of cells. For downstream analyses, we used Seurat v4 and Signac[289]. We kept ATAC peaks present in at least 10 cells. We removed low quality cells using the following thresholds: > 200 detected genes, > 400 detected peaks, <10% mitochondrial reads, > 2 transcription start site enrichment score and > 10% ATAC reads in peaks. We then annotated cells by co-embedding our data with the heart atlas left atrial nuclei[306]. We used SCTransform for normalization and regressing out mitochondrial reads percentages. We integrated the data using Harmony on 30 PCs. We used the heart cell atlas labels to assign cell-types to the resulting clusters. We removed doublets using both scDblFinder[408] scores and manual sub-clustering curation. We first calculated a doublet score using scDblFinder on both the RNA and ATAC data. We removed cells for which the product of the 2 scores was greater than 0.5 (scDblFinder RNA score * scDblFinder ATAC score, labeled high confidence doublets). We then sub-clustered each cell-type and removed sub-clusters that showed an enrichment of the scDblFinder scores and the top marker gene of another cell-type (labeled sub-cell-type doublets). We then re-clustered cells using seurat's function FindMultiModalNeighbors() with the first 20 PCs of the RNAseq data and PCs 2-20 of the ATACseq data. Finally, we refined ATACseq peaks with cell-type labels using Signac's CallPeaks function.

4.5.10 Visualization of fine-mapped AF-associated variants in single-nucleus multiome data

We calculated approximate Bayes factors (aBF) for each dataset using a previously described fine-mapping algorithm[409]. Briefly, aBF was calculated with summary statistics of the GWAS meta-analysis and both eQTL datasets using the following equation;

$$aBF = \sqrt{\frac{2SE^2}{2SE^2 + \omega}} \exp\left(\frac{\omega\beta^2}{2SE^2(2SE^2 + \omega)}\right)$$

where β and SE are the variant's effect size and standard error, respectively, and ω is the prior variance in allelic effects, taken here to be 0.04. We calculated PIP for variants in the 95%

credible sets for each dataset. We report credible set sizes and their overlap using the eQTL datasets in which the eQTL was significant and where the GWAS and eQTL signals co-localized (defined as H4 PP >0.4). When both Harbin and CTSN credible sets were included, we overlapped their union with the GWAS credible sets. In LocusZoom panels, we plotted P-values for each locus with 1000 Genomes European (EUR) and East Asian (EAS) populations linkage disequilibrium patterns using the locuscomparer package for CTSN and Harbin.

To evaluate the genomic context of each each locus, we retrieved the overlapping ATACseq peaks and coverage, peak-gene links and the eQTL gene expression from our sn-multiome dataset. To calculate peak-gene links, we first created MetaCells within each sample using the hdWGCNA package[410]. We calculated MetaCells genes and peaks counts by averaging the genes/cells and peaks/cells matrices using 30 neighbor cells in the gene expression harmony space and limited the number of overlapping cells to 15. We then calculated 2 Pearson's correlation scores between the eQTL genes and peaks within 1Mb of the gene. The first correlation score was calculated using all cells. The second was calculated using cardiomyocytes only. In both cases, we kept peak-gene links with a |Pearson R| > 0.2. For clarity, we only display links for eQTL genes found in **Table 1**.

4.5.11 Prioritized variants overlap with other genomic datasets

For each annotation, when necessary, we recovered hg38 positions as mentioned above. We obtained ENCODE regulatory elements from the genome.ucsc.edu table browser, table encodeCcreCombined (https://genome.ucsc.edu/cgi-bin/hgTables?hgta_doMainPage=1&hgta_group=regulation&hgta_track=encodeCcreCombined&hgta_table=encodeCcreCombined&hgsid=1439910105_RsimqAdh3sPECjdmse1QPtYFPY3c). We obtained CATlas ATAC peaks from http://catlas.org/catlas_downloads/humantissues/cCRE_by_cell_type. We obtained EpiMap links in the heart from https://personal.broadinstitute.org/cboix/epimap/links/pergroup/links_by_group.heart.tsv.gz.

4.5.12 LINC01629 CRISPRi

Cloning of gRNA plasmid and lentivirus production

Non-Targeting Control (NTC) gRNA sequence: Forward (5'-CACCGAAAACAGGACGATGTGCGGC-3') and Reverse (5'-AAACGCCGCACATCGTCCTGTTTTC-3'); *LINC01629* gRNA sequence: Forward (5'-CACCGTAGAAAAAGACACTTCCAA-3') and Reverse (5'-AAACTTGGAAGTGTCTTTTTCTAC-3'). The above gRNA oligonucleotides were annealed at a final concentration of 0.4uM and were cloned into 500ng of Esp3I-digested LentiGuide-Puro plasmid using T4 DNA Ligase (Catalog #M0202, NEB). Ligated plasmids were transformed into OneShot Stbl3 E. coli competent cells (Catalog #C737303, Invitrogen) as per the manufacturer's protocol. Successful plasmid clones were screened using Sanger sequencing and plasmids were extracted using FavorPrep plasmid miniprep kit (Catalog #FAPDE 300). For lentivirus production, lentiviral particles were produced in a 10cm plate format using HEK293T cells cultured in DMEM with 10% FBS. Briefly, 10 ug of Lentiguide-puro gRNA plasmid for *LINC01629* or NTC (Addgene #52963), 7.5 ug of pMDLg/pRRE, 2.5 ug of pRSV-REV, and 2.5 ug pMD2.G (Addgene #12251, #12253 & #12259) were co-transfected using 50 ul of PEI (1mg/ml) and 3ml of Opti-MEM I Reduced Serum Medium (Catalog #31985070, ThermoFisher Scientific). Following overnight incubation, the media was changed to DMEM with 5% FBS. Viral supernatant for the next 48 hrs was collected, pooled, and filtered through 0.22um PES filter. Viral particles were concentrated using Lenti-Pac Lentivirus Concentration Solution (Catalog #LPR-LCS-01, GeneCopoeia™), according to the manufacturer's instructions. Final concentrated viral particles were then resuspended in 100 uL of PBS solution.

Cell culture and CRISPRi knockdown of LINC01629

The H1-dCas9-KRAB hESC line was engineered by lentiviral infection of the Lenti-dCas9-KRAB-blast plasmid (Addgene #89567), and monoclones with stable constitutive dCas9-KRAB expression were isolated and expanded for targeting. All stem cell cultures were maintained in mTesR1 (Catalog #85857, STEMCELL Technologies), seeded in Geltrex (Catalog #A1413202, ThermoFisher Scientific) coated plates at 37 °C with 5% CO2 in the incubator. Cells were routinely tested for mycoplasma prior to culture. For the lentiviral knockdown of *LINC01629*, the H1-dCas9-KRAB hESC were cultured in 12-well plates and treated with 30ul of lentiviral particles in 8ug/ml polybrene, per well. 24 hours later, infected cells were positively selected with both 1ug/ml puromycin (to select for the gRNA) and 10ug/ml blasticidin (to ensure only the H1-dCas9-KRAB

cells). The knockdown of *LINC01629* was validated by qPCR, and three independent experiments per group were performed.

Cardiomyocyte differentiation

We performed cardiomyocyte differentiation using the GiWi protocol as previously described[411]. Briefly, both H1-dCas9-KRAB hESC lines infected with NTC and *LINC01629* gRNAs were dissociated with Accutase (Catalog: #07920, STEMCELL Technology), and seeded in 12-well plates containing mTesR1 with Y27632 (Catalog #72307, STEMCELL Technology) until 80% confluency. Following 48 hours, cells were induced with CHIR99021 for 24 hours (Catalog #72054, STEMCELL Technology) and media was subsequently changed to RPMI+B27 without insulin (Catalog A1895601, ThermoFisher). Three days after CHIR99021 induction, media was changed to RPMI without insulin (Catalog #A1895601, ThermoFisher) and 5uM IWP2 (Catalog #72122, STEMCELL Technology) for another 48 hours. On Day 5 post-induction, media was refreshed in RPMI+B27 without insulin. RPMI+B27 supplement with insulin (Catalog 17504044, ThermoFisher) was added only from day 7 thereafter when there were beating cardiomyocyte clusters. Cardiomyocytes were cultured and matured for 60 days before harvesting for downstream experiments.

RNA Extraction and RNAseq

Cardiomyocytes targeted with both NTC and *LINC01629* gRNA were isolated and harvested in 400ul Trizol reagent (ThermoFisher, 15596026). Briefly, RNA from three independent biological replicates were isolated using the Direct-Zol RNA Miniprep kit (Zymo Research, R2050). RNA quality and yield were assessed using Agilent RNA 6000 Pico kit (Agilent, 50167-1513) for quality control. Total RNA library preparations were prepared using TruSeq Stranded Total RNA Library Prep HMR kit (Illumina, 20020596) and respective cDNA libraries were prepared by Macrogen Asia Pacific Pte. Ltd. RNA libraries were sequenced on HiSeq 4000 Illumina sequencing platform to achieve a sequencing depth of at least 50 million paired-end reads per biological sample.

RNAseq processing and differential expression analysis

We pseudoaligned RNAseq reads to Gencode v32 using Kallisto with the options quant -b 100 --rf-stranded. We aggregated transcripts by genes and quantified them with DESeq2. We

calculated principle components using the 500 most variable genes. We used DESeq2's Wald test for condition (CRISPRi targeting the promoter of *LINC01629* vs NTC), followed by a log2 shrunken transformation (ashr shrinkage estimator). We performed an over-representation analysis on Gene Ontology Biological Processes using the rbioapi package[412] rba_panther_enrich() function. We used all genes with a baseMean value above 1 as background and corrected for multiple testing using the Bonferroni method. We report gene set enrichments for up-regulated and down-regulated genes (FDR <0.1 and log2-fold change > 0 and < 0, respectively).

4.6 DECLARATIONS

4.6.1 Acknowledgments

We thank all participants who contributed biosamples to this study.

4.6.2 Data availability

The sn-multiome data discussed in this publication have been deposited in NCBI's Gene Expression Omnibus and are accessible through GEO Series accession number GSE238242 (token: GSE238242). The bulk RNAseq data for the CTSN and Harbin cohorts are available at: http://www.mhi-humangenetics.org/en/resources/.

4.6.3 Funding

This work was funded by the Fonds de Recherche en Santé du Québec (FRQS), the Canada Research Chair Program and the Montreal Heart Institute Foundation (to S.N. and G.L.), and by the National Natural Science Foundation of China (81861128022, U21A20339 to B.Y.). S.R was funded by the British Heart Foundation Intermediate and Senior Fellowships, the British Research Council (BRC4) NIHR (Oxford) grant, the Wellcome Trust Institutional Strategic Individual Career Support grant and the John Fell Foundation Fund (Oxford). This research was enabled in part by support provided by Calcul Quebec (https://www.calculquebec.ca/en/) and Compute Canada (www.computecanada.ca). We thank Génome Québec for performing next-generation DNA sequencing for this project.

4.6.4 Competing interests

The authors declare that they have no competing interests.

4.6.5 Author contributions

Conceived and designed the analyses: F.J.A.L. and G.L.; Collected the data: X.J., K.K., and N.M.; Contributed data: X.J., K.K., M.L., J.X., L.X., W.M., H.B., N.M., R.S.-Y.F., S.R., C.G.A.-N., Z.P., S.N., and B.Y.; Performed analyses: F.J.A.L., M.L., J.X., C.G.A.-N., and G.L.; Secured funding and supervised the work: S.R., C.G.A.-N., Z.P., S.N., B.Y., and G.L.; Wrote the manuscript: F.J.A.L. and G.L., with contributions from all authors.

4.7 Supplementary material

For **tables S2** and **S8** see the attached zipped folder.

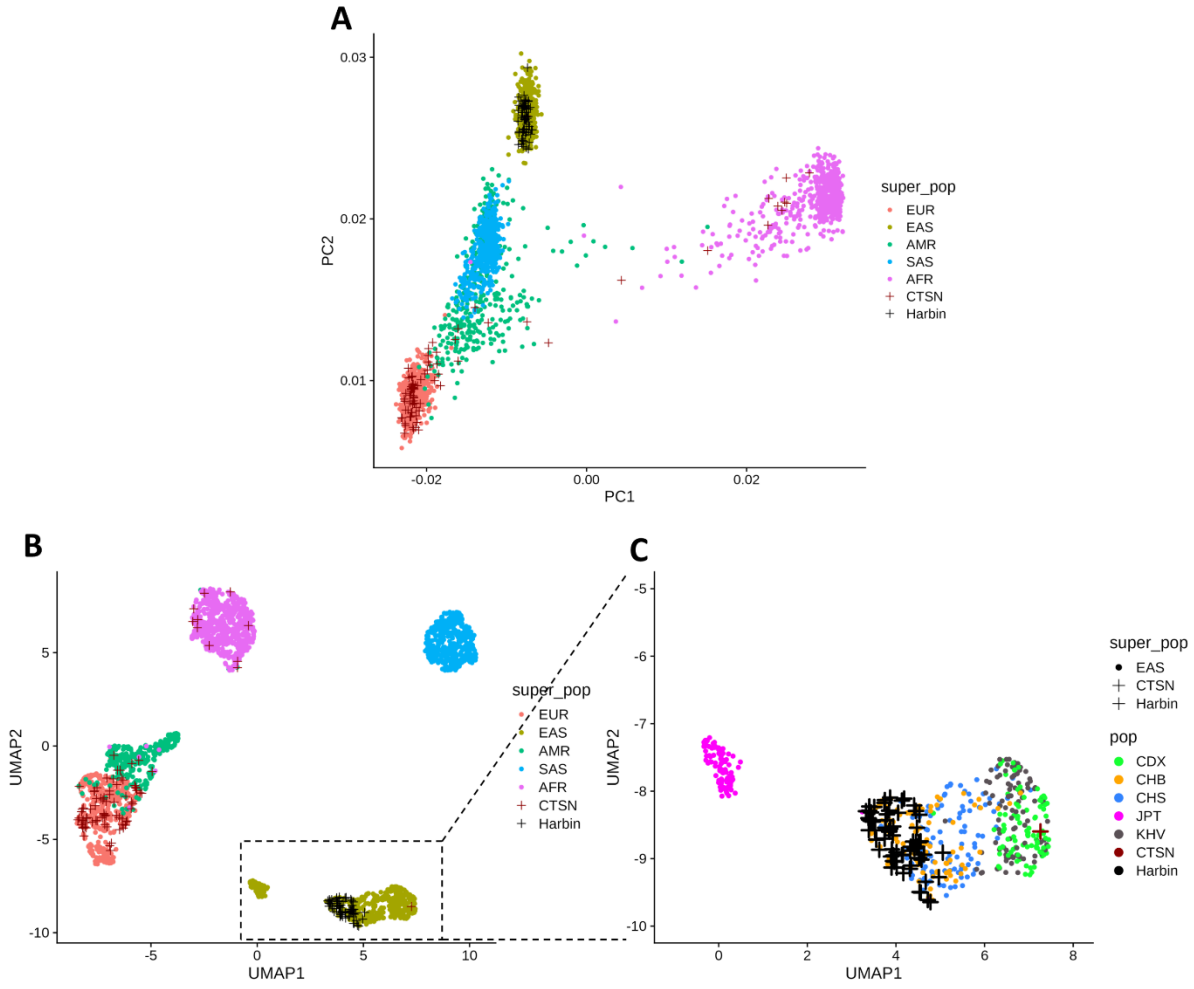


Figure S1. Cohort ancestry against 1000 Genomes Project.

(**A**) Principal component analysis of genotype data from the 1000 Genomes Project, CTSN and Harbin cohorts using the European/African ancestry informative markers list variants present in all cohorts. (**B**) Uniform Manifold Approximation and Projection (UMAP) generated using the first 10 principal components of the data presented in (**A**). (**C**) Zoom-in of the UMAP region selected in (**B**). The CTSN and Harbin participants are labeled with red and black "+" sign, respectively. The 1000 Genomes Project participants are labeled with dots, color-coded based on their population of origin. AFR, African-ancestry; EUR, European-ancestry; EAS, East Asian-ancestry; AMR, Admixed Americans; SAS, South Asian-ancestry; CDX, Chinese Dai in Xishuangbanna, China; CHB, Han Chinese in Beijing, China; CHS, Han Chinese South; JPT, Japanese in Tokyo, Japan; KHV, Kinh in Ho Chi Minh City, Vietnam.

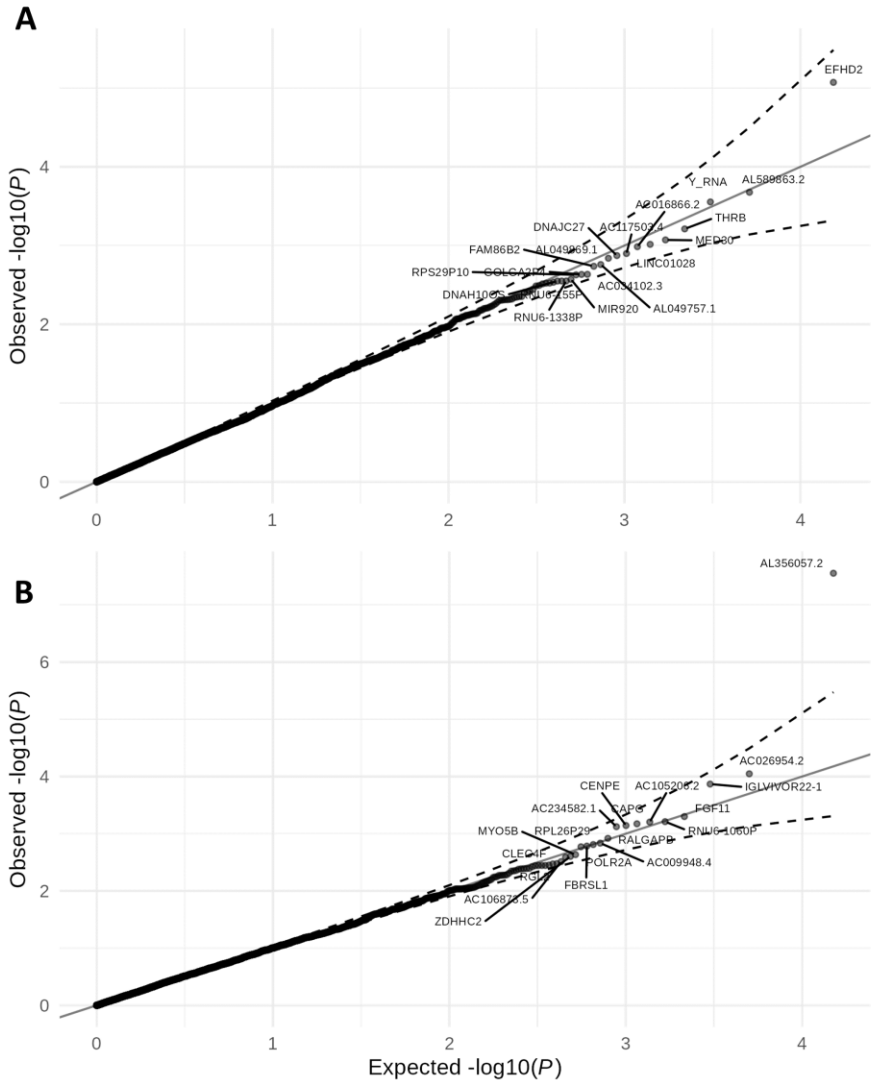


Figure S2. eQTL AF interactions.

Quantile-quantile-plots of the eQTL interaction term (Genotype:Rhythm) -log10(p-values) for the 150 atrial fibrillation SNPs associated in the multi-ancestry GWAS in (**A**) CTSN and (**B**) Harbin. While the interaction term is significant for rs12209223 and *AL356057.2* in Harbin, it does not replicate in CTSN (false discovery rate = 0.99).

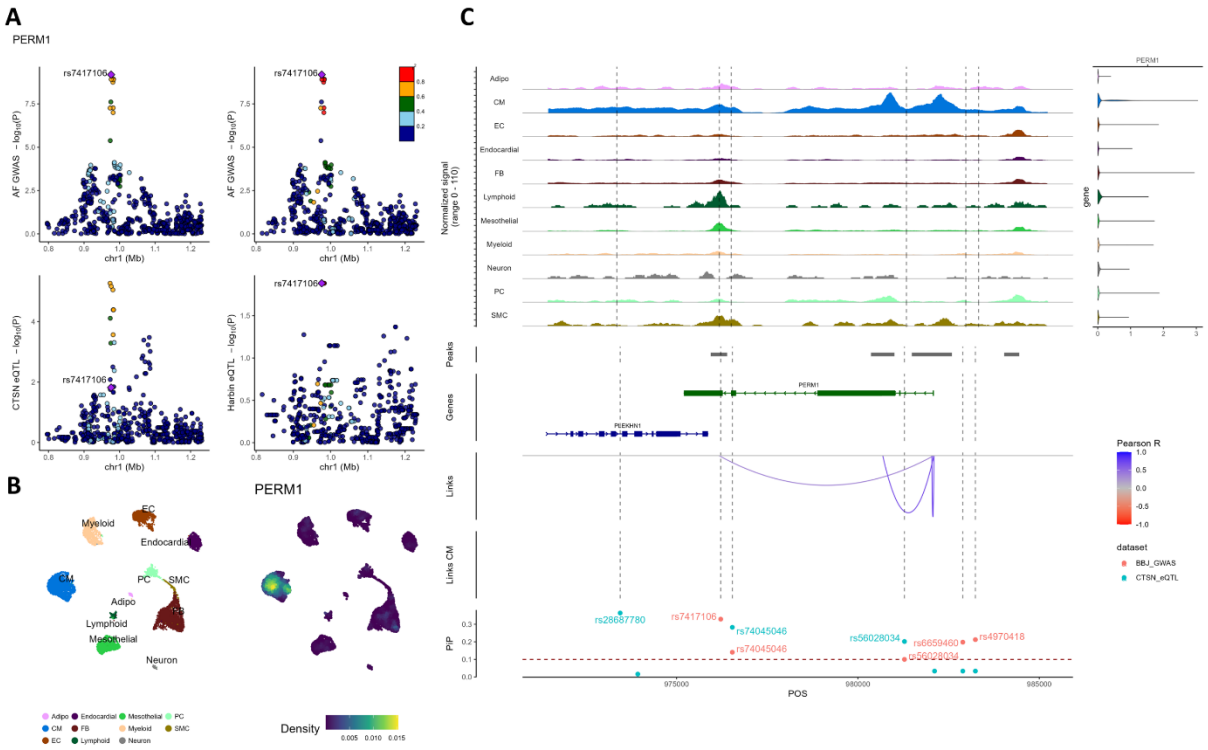


Figure S3. Fine-mapping and annotation of the *PERM1* locus.

(**A**) The top panels show –log10(p-values) from the cross-ancestry AF GWAS (y-axis) against genomic coordinates (x-axis, hg38) for a 500kb window centered on the sentinel AF SNP. SNPs are colored based on the 1000 Genome Project linkage disequilibrium (LD) r^2 with the lead SNP in the European (left) and East Asian (right) super-populations. In the bottom panels, we report the eQTL –log10(p-values) in the CTSN and Harbin cohorts for *PERM1* expression in left atrial appendages (LAAs). LD is based on the European super-population for CTSN (left) and the East Asian super-population for Harbin (right).

(**B**) Uniform manifold approximation and projection (UMAP) of LAA single-nucleus multiome cell-types (left) and *PERM1* expression density (right). *PERM1* expression is enriched in cardiomyocytes (CMs).

(**C**) LAA single-nucleus multiome genomic context of the prioritized AF-associated variants. From top to bottom: The first track shows ATACseq read coverage at the *PERM1* locus (left) paired with violin plots of *PERM1* expression aggregated by cell-type (right). The second track shows ATACseq peaks. The third track shows the gene annotation (exon, intron) for the genes found at the locus. The fourth and fifth tracks highlight links between ATACseq peaks and gene promoters identified in either all cell-types (Links) or specifically in cardiomyocytes (CM). We use Pearson correlation tests between ATACseq peak accessibility and gene expression (in the same nucleus) to derive links. Only links with |Pearson R| > 0.2 are shown. Link heights are proportional to their |Pearson R| in the range indicated in the legend. In the final track, we report AF GWAS and eQTL fine-mapping posterior inclusion probabilities (PIP). Two variants (rs56028034, rs74045046) have a PIP >0.1 in both the cross-ancestry AF GWAS and the CTSN

eQTL study (*PERM1* was not an eQTL in the Harbin cohort). These SNPs do not overlap with an ATACseq peak in our LAA multiome data. However, rs74045046 overlaps a distal enhancer identified by ENCODE.

Adipo; Adipocytes, CM; Cardiomyocytes, EC; Endothelial cells, FB; Fibroblasts, PC; Pericytes, SMC; Smooth muscle cells.

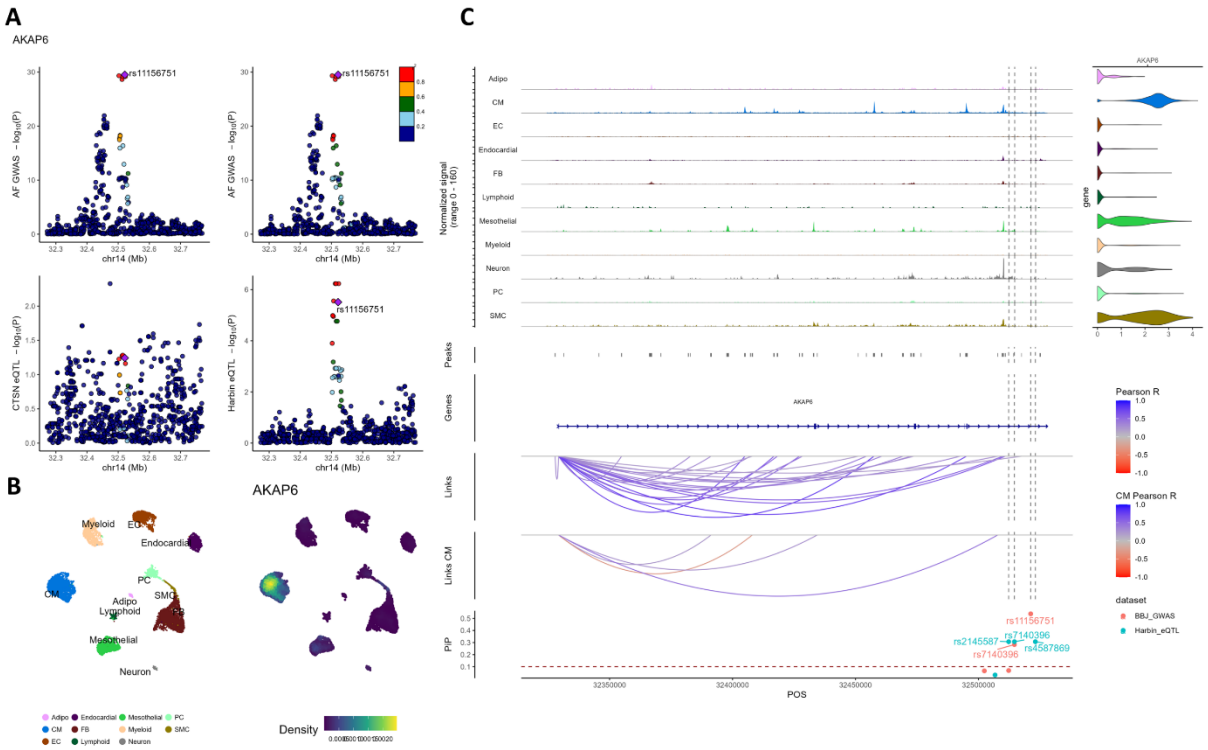


Figure S4. Fine-mapping and annotation of the *AKAP6* locus.

(**A**) The top panels show –log10(p-values) from the cross-ancestry AF GWAS (y-axis) against genomic coordinates (x-axis, hg38) for a 500kb window centered on the sentinel AF SNP. SNPs are colored based on the 1000 Genome Project linkage disequilibrium (LD) r^2 with the lead SNP in the European (left) and East Asian (right) super-populations. In the bottom panels, we report the eQTL –log10(p-values) in the CTSN and Harbin cohorts for *AKAP6* expression in left atrial appendages (LAAs). LD is based on the European super-population for CTSN (left) and the East Asian super-population for Harbin (right).

(**B**) Uniform manifold approximation and projection (UMAP) of LAA single-nucleus multiome cell-types (left) and *AKAP6* expression density (right). *AKAP6* expression is enriched in cardiomyocytes (CMs).

(**C**) LAA single-nucleus multiome genomic context of the prioritized AF-associated variants. From top to bottom: The first track shows ATACseq read coverage at the *AKAP6* locus (left) paired with violin plots of *AKAP6* expression aggregated by cell-type (right). The second track shows ATACseq peaks. The third track shows the gene annotation (exon, intron) for the genes found at the locus. The fourth and fifth tracks highlight links between ATACseq peaks and gene promoters identified in either all cell-types (Links) or specifically in cardiomyocytes (CM). We use Pearson correlation tests between ATACseq peak accessibility and gene expression (in the same nucleus) to derive links. Only links with |Pearson R| > 0.2 are shown. Link heights are proportional to their |Pearson R| in the range indicated in the legend. In the final track, we report AF GWAS and eQTL fine-mapping posterior inclusion probabilities (PIP). One variant (rs7140396) has a PIP >0.1 in both the cross-ancestry AF GWAS and Harbin eQTLs (AKAP6 was

not an eQTL in the CTSN cohort). This SNP does not overlap with an ATACseq peak in our LAA multiome data, but does overlap with an ATACpeak identified in adrenal cells in CATlas.

Adipo; Adipocytes, CM; Cardiomyocytes, EC; Endothelial cells, FB; Fibroblasts, PC; Pericytes, SMC; Smooth muscle cells.

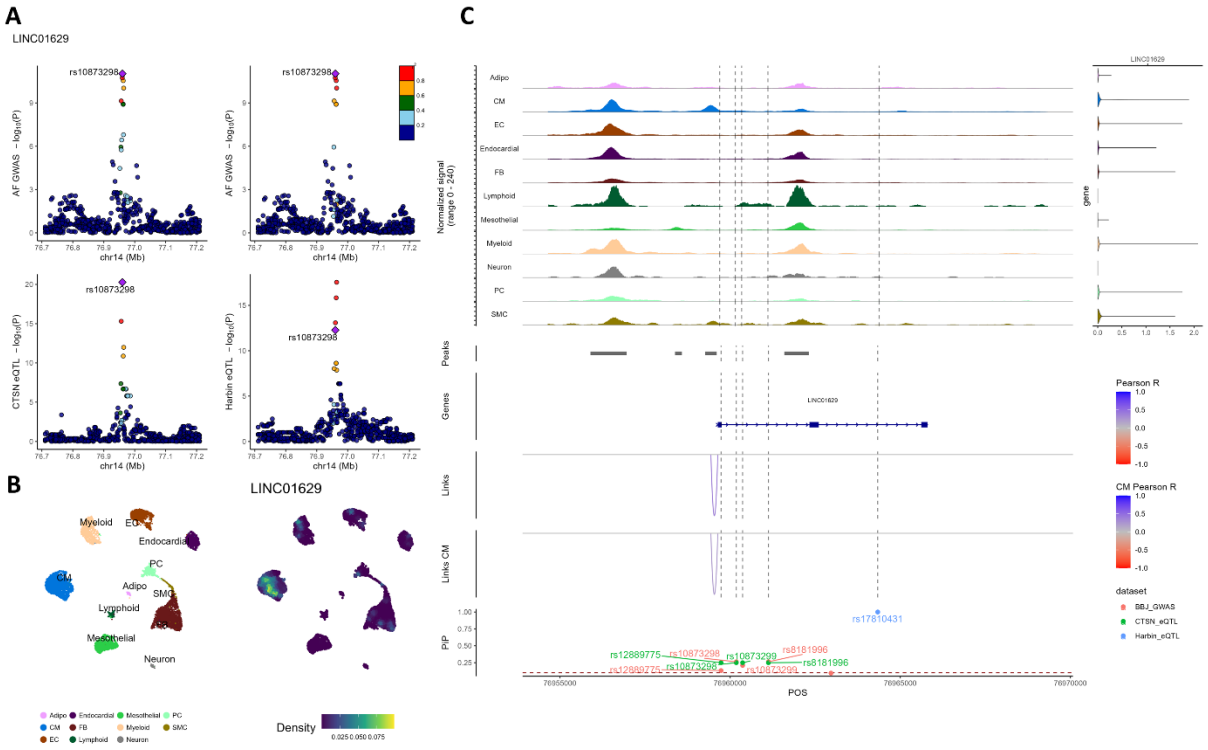


Figure S5. Fine-mapping and annotation of the *LINC01629* locus.

(**A**) The top panels show –log10(p-values) from the cross-ancestry AF GWAS (y-axis) against genomic coordinates (x-axis, hg38) for a 500kb window centered on the sentinel AF SNP. SNPs are colored based on the 1000 Genome Project linkage disequilibrium (LD) r^2 with the lead SNP in the European (left) and East Asian (right) super-populations. In the bottom panels, we report the eQTL –log10(p-values) in the CTSN and Harbin cohorts for *LINC01629* expression in left atrial appendages (LAAs). LD is based on the European super-population for CTSN (left) and the East Asian super-population for Harbin (right).

(**B**) Uniform manifold approximation and projection (UMAP) of LAA single-nucleus multiome cell-types (left) and *LINC01629* expression density (right). *LINC01629* expression is enriched in cardiomyocytes (CMs).

(**C**) LAA single-nucleus multiome genomic context of the prioritized AF-associated variants. From top to bottom: The first track shows ATACseq read coverage at the *LINC01629* locus (left) paired with violin plots of *LINC01629* expression aggregated by cell-type (right). The second track shows ATACseq peaks. The third track shows the gene annotation (exon, intron) for the genes found at the locus. The fourth and fifth tracks highlight links between ATACseq peaks and gene promoters identified in either all cell-types (Links) or specifically in cardiomyocytes (CM). We use Pearson correlation tests between ATACseq peak accessibility and gene expression (in the same nucleus) to derive links. Only links with |Pearson R| > 0.2 are shown. Link heights are proportional to their |Pearson R| in the range indicated in the legend. In the final track, we report AF GWAS and eQTL fine-mapping posterior

inclusion probabilities (PIP). Four variants (rs10873298, rs10873299, rs12889775, rs8181996) have a PIP > 0.1 in both the cross-ancestry AF GWAS and the two eQTL studies. These SNPs do not overlap with an ATACseq peak in our LAA multiome data. However, rs10873298 and rs10873299 overlap a proximal enhancers defined by ENCODE.

Adipo; Adipocytes, CM; Cardiomyocytes, EC; Endothelial cells, FB; Fibroblasts, PC; Pericytes, SMC; Smooth muscle cells.

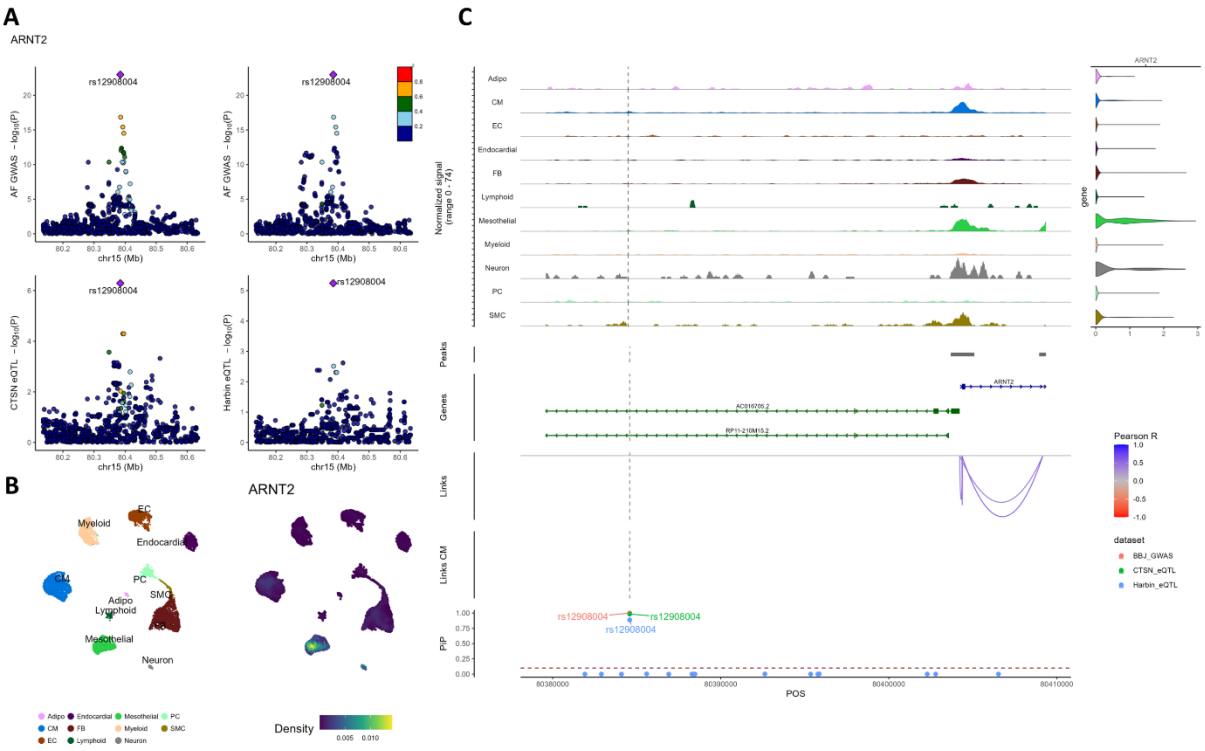


Figure S6. Fine-mapping and annotation of the *ARNT2* locus.

(**A**) The top panels show –log10(p-values) from the cross-ancestry AF GWAS (y-axis) against genomic coordinates (x-axis, hg38) for a 500kb window centered on the sentinel AF SNP. SNPs are colored based on the 1000 Genome Project linkage disequilibrium (LD) r^2 with the lead SNP in the European (left) and East Asian (right) super-populations. In the bottom panels, we report the eQTL –log10(p-values) in the CTSN and Harbin cohorts for *ARNT2* expression in left atrial appendages (LAAs). LD is based on the European super-population for CTSN (left) and the East Asian super-population for Harbin (right).

(**B**) Uniform manifold approximation and projection (UMAP) of LAA single-nucleus multiome cell-types (left) and *ARNT2* expression density (right). *ARNT2* expression is enriched in mesothelial cells.

(**C**) LAA single-nucleus multiome genomic context of the prioritized AF-associated variants. From top to bottom: The first track shows ATACseq read coverage at the *ARNT2* locus (left) paired with violin plots of *ARNT2* expression aggregated by cell-type (right). The second track shows ATACseq peaks. The third track shows the gene annotation (exon, intron) for the genes found at the locus. The fourth and fifth tracks highlight links between ATACseq peaks and gene promoters identified in either all cell-types (Links) or specifically in cardiomyocytes (CM). We use Pearson correlation tests between ATACseq peak accessibility and gene expression (in the same nucleus) to derive links. Only links with |Pearson R| > 0.2 are shown. Link heights are proportional to their

|Pearson R| in the range indicated in the legend. In the final track, we report AF GWAS and eQTL fine-mapping posterior inclusion probabilities (PIP). One variant (rs12908004) has a PIP >0.1 in the AF GWAS, CTSN and Harbin datasets. This SNP does not overlap with an ATACseq peak in our LAA multiome data, but does overlap with an ATACseq peak identified in multiple cell-types (accessible in pericytes, liver fibroblasts, exocrine endothelial cells, and type II skeletal myocyte) in CATlas, as well as in a DNase hypersensitive site and a H3K4me3-marked element annotated in ENCODE.

Adipo; Adipocytes, CM; Cardiomyocytes, EC; Endothelial cells, FB; Fibroblasts, PC; Pericytes, SMC; Smooth muscle cells.

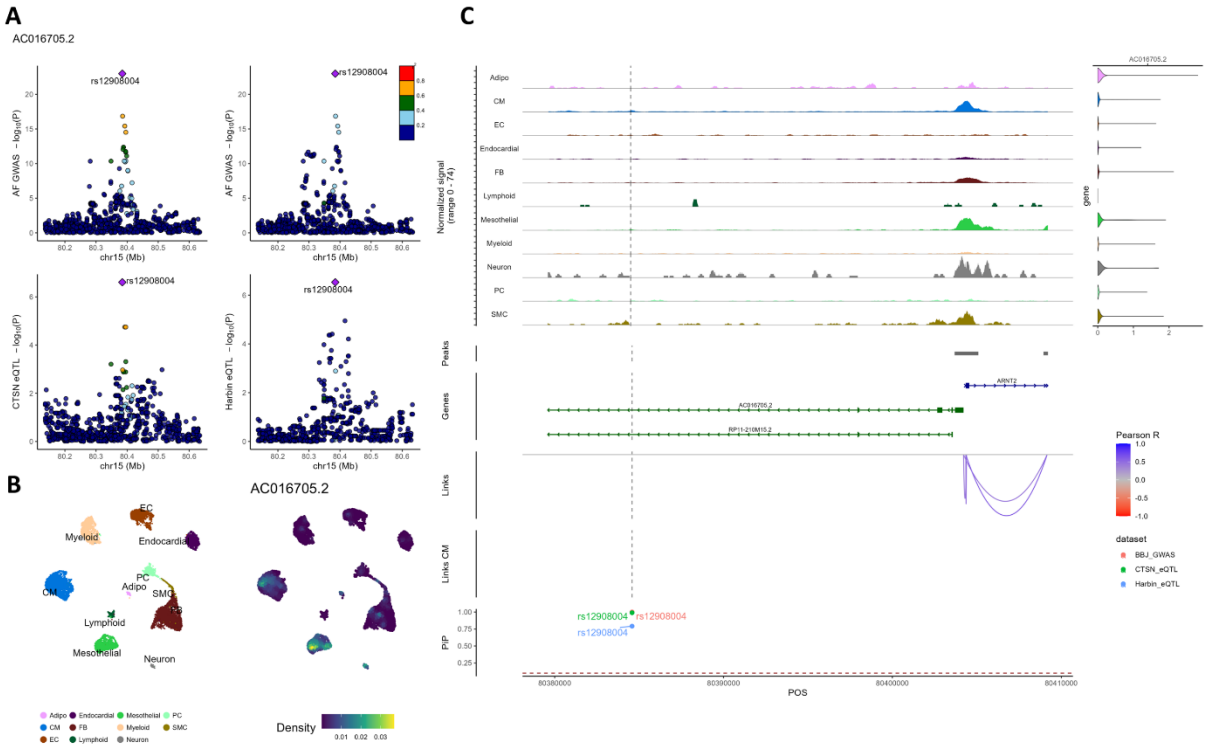


Figure S7. Fine-mapping and annotation of the AC016705.2 locus.

(**A**) The top panels show –log10(p-values) from the cross-ancestry AF GWAS (y-axis) against genomic coordinates (x-axis, hg38) for a 500kb window centered on the sentinel AF SNP. SNPs are colored based on the 1000 Genome Project linkage disequilibrium (LD) r^2 with the lead SNP in the European (left) and East Asian (right) super-populations. In the bottom panels, we report the eQTL –log10(p-values) in the CTSN and Harbin cohorts for *AC016705.2* expression in left atrial appendages (LAAs). LD is based on the European super-population for CTSN (left) and the East Asian super-population for Harbin (right).

(**B**) Uniform manifold approximation and projection (UMAP) of LAA single-nucleus multiome cell-types (left) and *AC016705.2* expression density (right). *AC016705.2* expression is enriched in mesothelial cells.

(**C**) LAA single-nucleus multiome genomic context of the prioritized AF-associated variants. From top to bottom: The first track shows ATACseq read coverage at the *AC016705.2* locus (left) paired with violin plots of *AC016705.2* expression aggregated by cell-type (right). The second track shows ATACseq peaks. The third track shows the gene annotation (exon, intron) for the genes found at the locus. The fourth and fifth tracks highlight links between ATACseq peaks and gene promoters identified in either all cell-types (Links) or specifically in cardiomyocytes (CM). We use Pearson correlation tests between ATACseq peak accessibility and gene expression (in the same nucleus) to derive links. Only links with |Pearson R| > 0.2 are shown. Link heights are proportional

to their |Pearson R| in the range indicated in the legend. In the final track, we report AF GWAS and eQTL fine-mapping posterior inclusion probabilities (PIP). One variant (rs12908004) has a PIP >0.1 in the AF GWAS, CTSN and Harbin datasets. This SNP does not overlap with an ATACseq peak in our LAA multiome data, but does overlap with an ATACseq peak identified in multiple cell-types (accessible in pericytes, liver fibroblasts, exocrine endothelial cells, and type II skeletal myocyte) in CATlas, as well as in a DNase hypersensitive site and a H3K4me3-marked element annotated in ENCODE.

Adipo; Adipocytes, CM; Cardiomyocytes, EC; Endothelial cells, FB; Fibroblasts, PC; Pericytes, SMC; Smooth muscle cells.

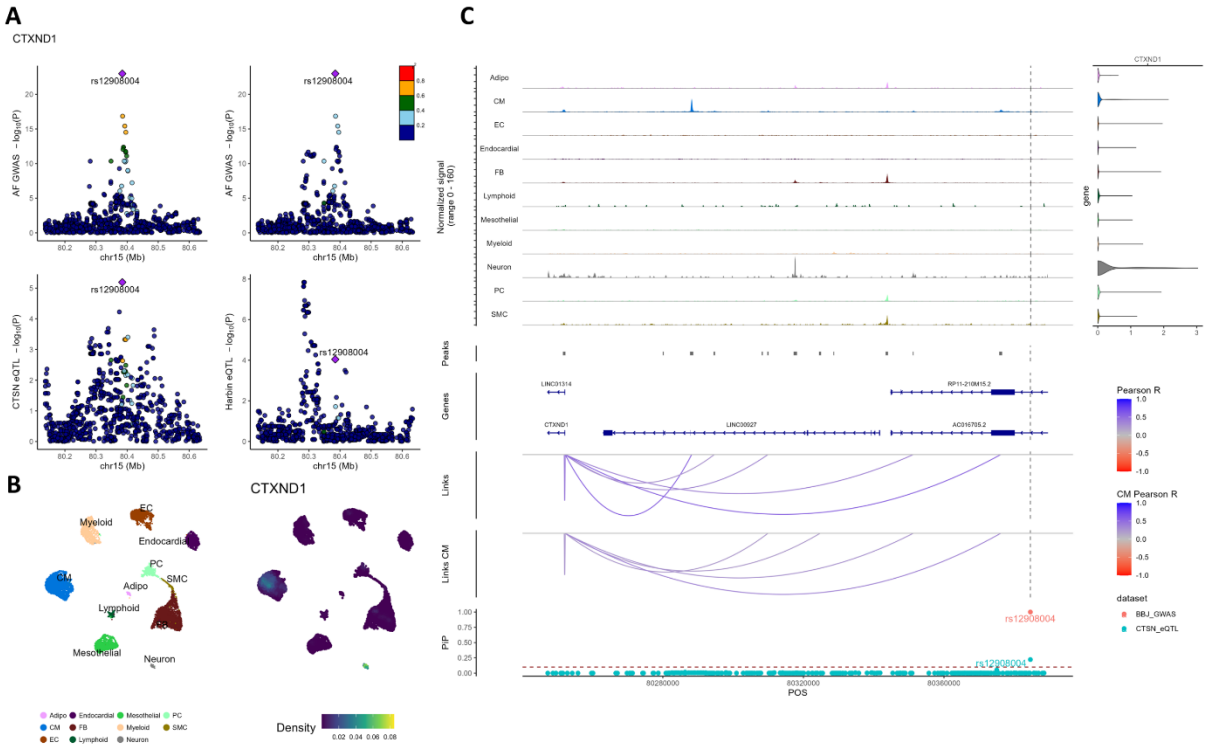


Figure S8. Fine-mapping and annotation of the *CTXND1* locus.

(**A**) The top panels show –log10(p-values) from the cross-ancestry AF GWAS (y-axis) against genomic coordinates (x-axis, hg38) for a 500kb window centered on the sentinel AF SNP. SNPs are colored based on the 1000 Genome Project linkage disequilibrium (LD) r^2 with the lead SNP in the European (left) and East Asian (right) super-populations. In the bottom panels, we report the eQTL –log10(p-values) in the CTSN and Harbin cohorts for *CTXND1* expression in left atrial appendages (LAAs). LD is based on the European super-population for CTSN (left) and the East Asian super-population for Harbin (right).

(**B**) Uniform manifold approximation and projection (UMAP) of LAA single-nucleus multiome cell-types (left) and *CTXND1* expression density (right). *CTXND1* expression is enriched in neurons and cardiomyocytes.

(**C**) LAA single-nucleus multiome genomic context of the prioritized AF-associated variants. From top to bottom: The first track shows ATACseq read coverage at the *CTXND1* locus (left) paired with violin plots of *CTXND1* expression aggregated by cell-type (right). The second track shows ATACseq peaks. The third track shows the gene annotation (exon, intron) for the genes found at the locus. The fourth and fifth tracks highlight links between ATACseq peaks and gene promoters identified in either all cell-types (Links) or specifically in cardiomyocytes (CM). We use Pearson correlation tests between ATACseq peak accessibility and gene expression (in the same nucleus) to derive links. Only links with |Pearson R| > 0.2 are shown. Link heights are proportional to their

|Pearson R| in the range indicated in the legend. In the final track, we report AF GWAS and eQTL fine-mapping posterior inclusion probabilities (PIP). One variant (rs12908004) has a PIP >0.1 in the AF GWAS, CTSN and Harbin datasets. This SNP does not overlap with an ATACseq peak in our LAA multiome data, but does overlap with an ATACseq peak identified in multiple cell-types (accessible in pericytes, liver fibroblasts, exocrine endothelial cells, and type II skeletal myocyte) in CATlas, as well as in a DNase hypersensitive site and a H3K4me3-marked element annotated in ENCODE.

Adipo; Adipocytes, CM; Cardiomyocytes, EC; Endothelial cells, FB; Fibroblasts, PC; Pericytes, SMC; Smooth muscle cells.

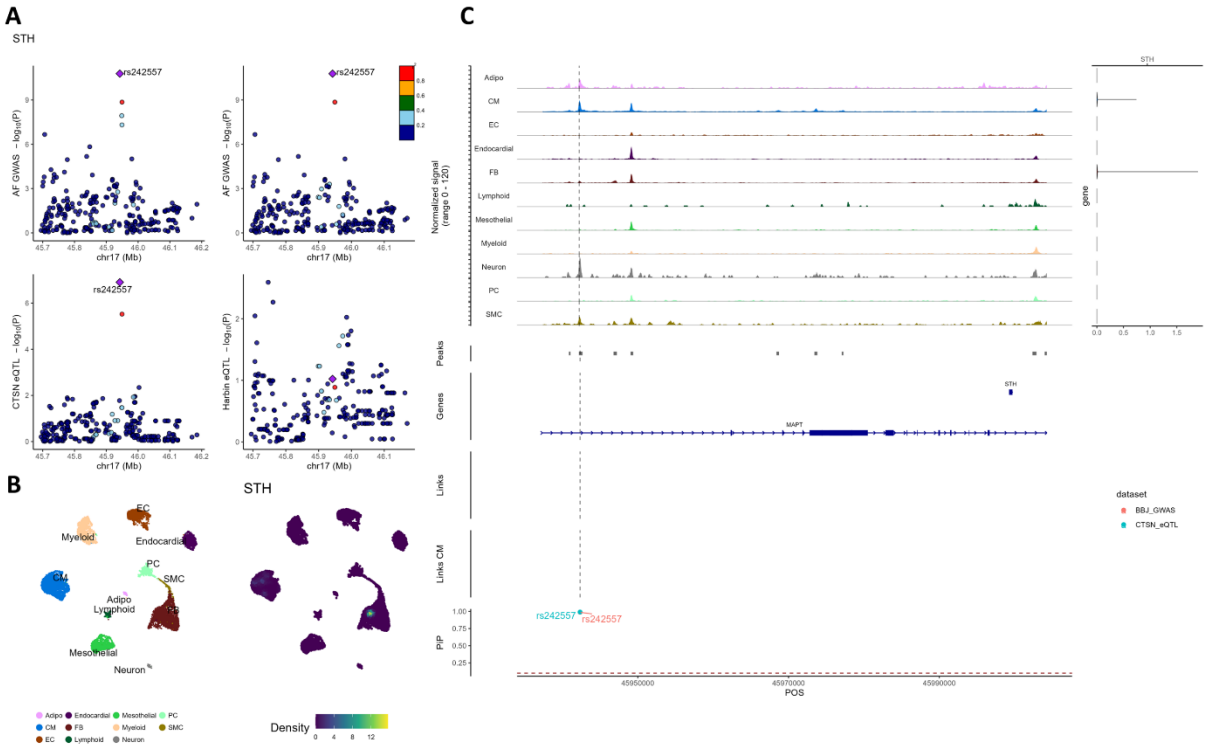


Figure S9. Fine-mapping and annotation of the *STH* locus.

(**A**) The top panels show –log10(p-values) from the cross-ancestry AF GWAS (y-axis) against genomic coordinates (x-axis, hg38) for a 500kb window centered on the sentinel AF SNP. SNPs are colored based on the 1000 Genome Project linkage disequilibrium (LD) r^2 with the lead SNP in the European (left) and East Asian (right) super-populations. In the bottom panels, we report the eQTL –log10(p-values) in the CTSN and Harbin cohorts for *STH* expression in left atrial appendages (LAAs). LD is based on the European super-population for CTSN (left) and the East Asian super-population for Harbin (right).

(**B**) Uniform manifold approximation and projection (UMAP) of LAA single-nucleus multiome cell-types (left) and *STH* expression density (right). *STH* expression is very low in all cell-types.

(**C**) LAA single-nucleus multiome genomic context of the prioritized AF-associated variants. From top to bottom: The first track shows ATACseq read coverage at the *STH* locus (left) paired with violin plots of *STH* expression aggregated by cell-type (right). The second track shows ATACseq peaks. The third track shows the gene annotation (exon, intron) for the genes found at the locus. The fourth and fifth tracks highlight links between ATACseq peaks and gene promoters identified in either all cell-types (Links) or specifically in cardiomyocytes (CM). We use Pearson correlation tests between ATACseq peak accessibility and gene expression (in the same nucleus) to derive links. Only links with |Pearson R| > 0.2 are shown. Link heights are proportional to their |Pearson

R| in the range indicated in the legend. In the final track, we report AF GWAS and eQTL fine-mapping posterior inclusion probabilities (PIP). One variant (rs242557) has a PIP >0.1 in the AF GWAS and CTSN dataset (*STH* was not an eQTL in the Harbin cohort). This SNP overlaps with the ATACseq peak chr17-45942197-45942667 in our LAA multiome data, overlaps with an ATACseq peak identified in multiple cell-types in CATlas and is in a distal enhancer element annotated in ENCODE.

Adipo; Adipocytes, CM; Cardiomyocytes, EC; Endothelial cells, FB; Fibroblasts, PC; Pericytes, SMC; Smooth muscle cells.

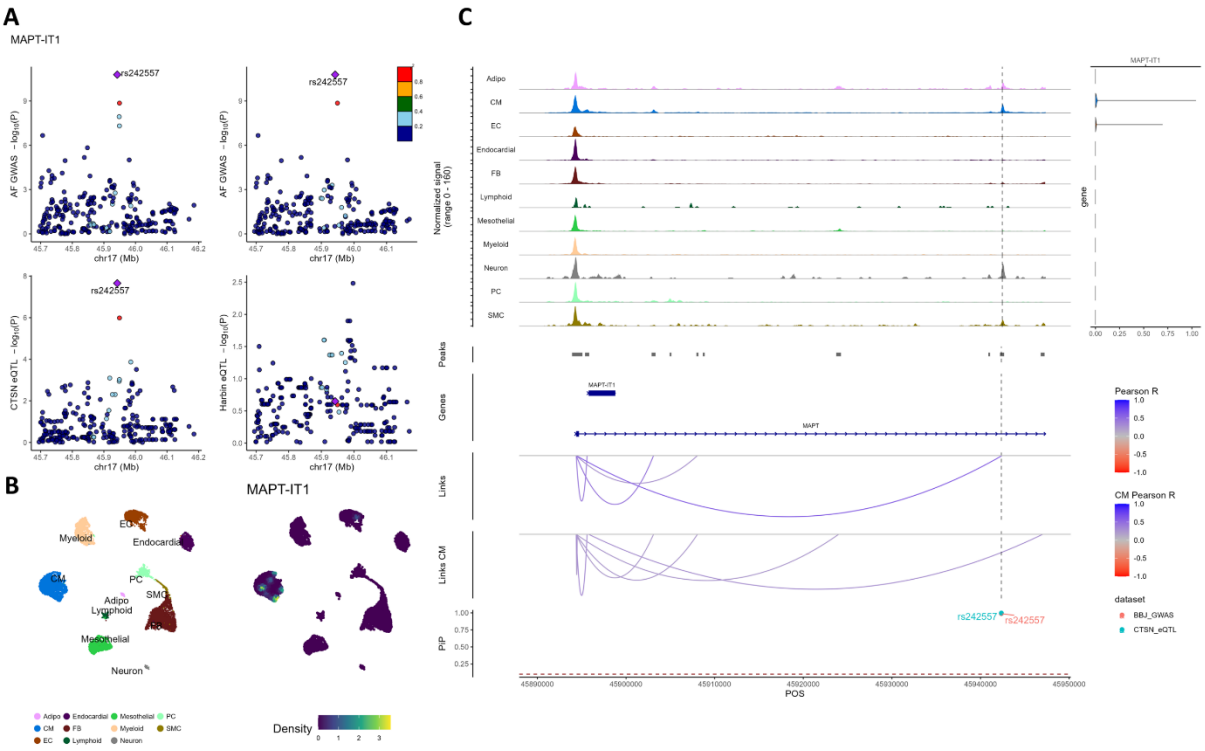


Figure S10. Fine-mapping and annotation of the *MAPT-IT1* locus.

(**A**) The top panels show –log10(p-values) from the cross-ancestry AF GWAS (y-axis) against genomic coordinates (x-axis, hg38) for a 500kb window centered on the sentinel AF SNP. SNPs are colored based on the 1000 Genome Project linkage disequilibrium (LD) r^2 with the lead SNP in the European (left) and East Asian (right) super-populations. In the bottom panels, we report the eQTL –log10(p-values) in the CTSN and Harbin cohorts for *MAPT-IT1* expression in left atrial appendages (LAAs). LD is based on the European super-population for CTSN (left) and the East Asian super-population for Harbin (right).

(**B**) Uniform manifold approximation and projection (UMAP) of LAA single-nucleus multiome cell-types (left) and *MAPT-IT1* expression density (right). *MAPT-IT1* expression is very low in all cell-types.

(**C**) LAA single-nucleus multiome genomic context of the prioritized AF-associated variants. From top to bottom: The first track shows ATACseq read coverage at the *MAPT-IT1* locus (left) paired with violin plots of *MAPT-IT1* expression aggregated by cell-type (right). The second track shows ATACseq peaks. The third track shows the gene annotation (exon, intron) for the genes found at the locus. The fourth and fifth tracks highlight links between ATACseq peaks and gene promoters identified in either all cell-types (Links) or specifically in cardiomyocytes (CM). We use Pearson correlation tests between ATACseq peak accessibility and

gene expression (in the same nucleus) to derive links. Only links with |Pearson R| > 0.2 are shown. Link heights are proportional to their |Pearson R| in the range indicated in the legend. In the final track, we report AF GWAS and eQTL fine-mapping posterior inclusion probabilities (PIP). One variant (rs242557) has a PIP >0.1 in the AF GWAS and CTSN dataset (*MAPT-IT1* was not an eQTL in the Harbin cohort). This SNP overlaps with the ATACseq peak chr17-45942197-45942667 in our LAA multiome data, overlaps with an ATACseq peak identified in multiple cell-types in CATlas and is in a distal enhancer element annotated in ENCODE.

Adipo; Adipocytes, CM; Cardiomyocytes, EC; Endothelial cells, FB; Fibroblasts, PC; Pericytes, SMC; Smooth muscle cells.

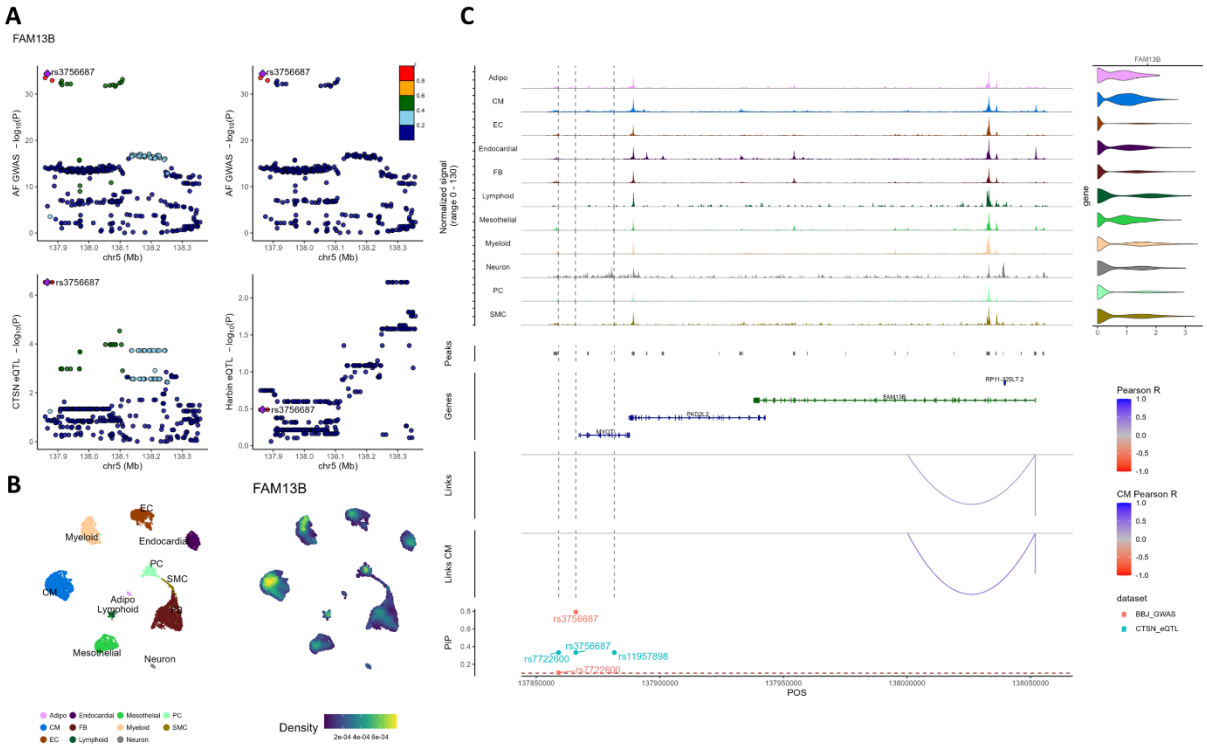


Figure S11. Fine-mapping and annotation of the *FAM13B* locus.

(**A**) The top panels show –log10(p-values) from the cross-ancestry AF GWAS (y-axis) against genomic coordinates (x-axis, hg38) for a 500kb window centered on the sentinel AF SNP. SNPs are colored based on the 1000 Genome Project linkage disequilibrium (LD) r^2 with the lead SNP in the European (left) and East Asian (right) super-populations. In the bottom panels, we report the eQTL –log10(p-values) in the CTSN and Harbin cohorts for *FAM13B* expression in left atrial appendages (LAAs). LD is based on the European super-population for CTSN (left) and the East Asian super-population for Harbin (right).

(**B**) Uniform manifold approximation and projection (UMAP) of LAA single-nucleus multiome cell-types (left) and *FAM13B* expression density (right). *FAM13B* is expressed in all cell-types.

(**C**) LAA single-nucleus multiome genomic context of the prioritized AF-associated variants. From top to bottom: The first track shows ATACseq read coverage at the *FAM13B* locus (left) paired with violin plots of *FAM13B* expression aggregated by cell-type (right). The second track shows ATACseq peaks. The third track shows the gene annotation (exon, intron) for the genes found at the locus. The fourth and fifth tracks highlight links between ATACseq peaks and gene promoters identified in either all cell-types (Links) or specifically in cardiomyocytes (CM). We use Pearson correlation tests between ATACseq peak accessibility and gene

expression (in the same nucleus) to derive links. Only links with |Pearson R| > 0.2 are shown. Link heights are proportional to their |Pearson R| in the range indicated in the legend. In the final track, we report AF GWAS and eQTL fine-mapping posterior inclusion probabilities (PIP). Two variants (rs3756687 and rs7722600) have PIP >0.1 in the AF GWAS and CTSN datasets (*FAM13B* was not an eQTL in the Harbin cohort). These SNPs do not overlap with an ATACseq peak in our LAA multiome data. However, rs7722600 overlap with an ATACseq peak identified in multiple cell-types (including cardiomyocytes) in CATlas, as well as a distal enhancer annotated in ENCODE.

Adipo; Adipocytes, CM; Cardiomyocytes, EC; Endothelial cells, FB; Fibroblasts, PC; Pericytes, SMC; Smooth muscle cells.

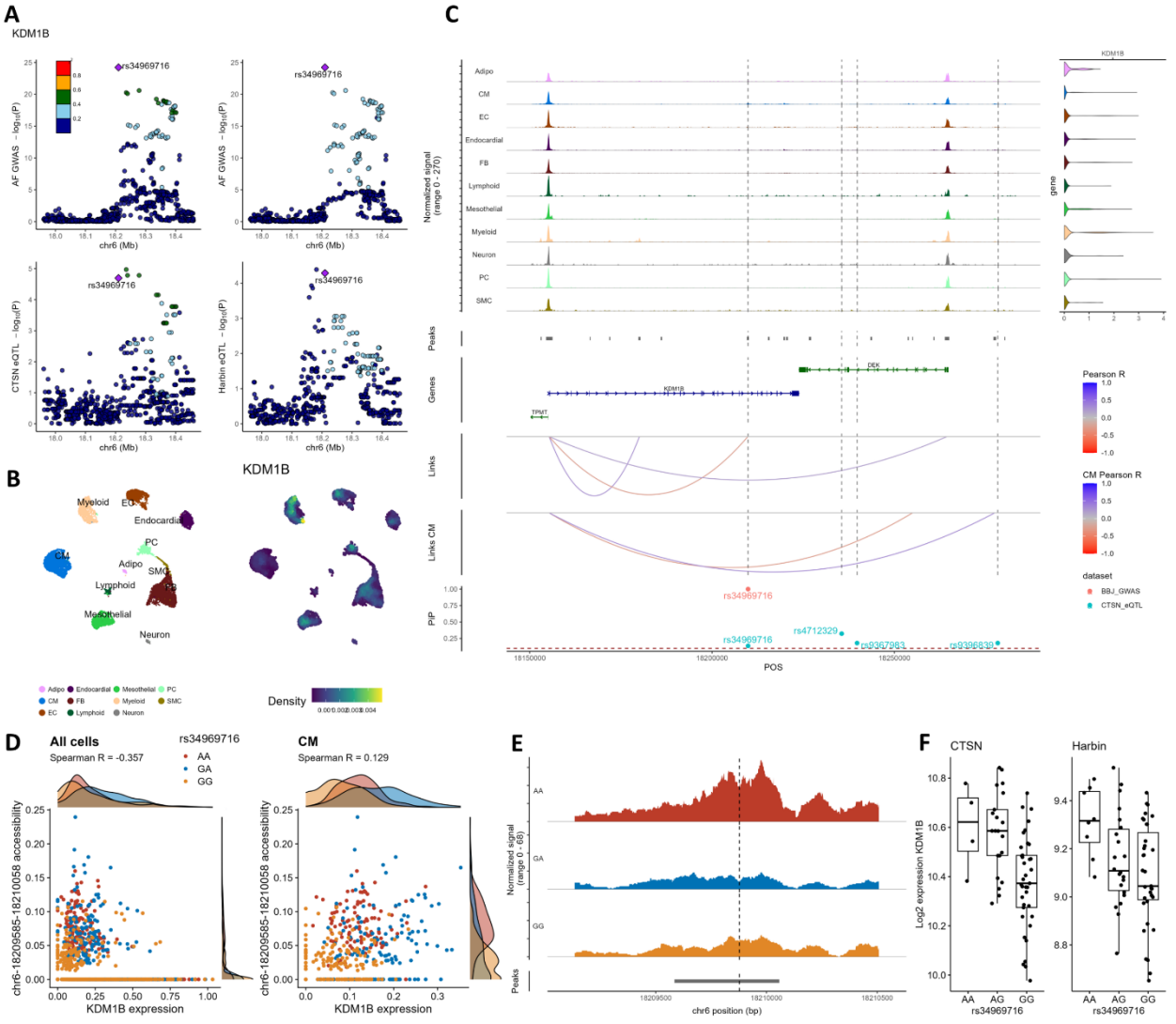


Figure S12. Fine-mapping and annotation of the *KDM1B* locus.

(**A**) The top panels show –log10(p-values) from the cross-ancestry AF GWAS (y-axis) against genomic coordinates (x-axis, hg38) for a 500kb window centered on the sentinel AF SNP. SNPs are colored based on the 1000 Genome Project linkage disequilibrium (LD) r^2 with the lead SNP in the European (left) and East Asian (right) super-populations. In the bottom panels, we report the eQTL

–log10(p-values) in the CTSN and Harbin cohorts for *FAM13B* expression in left atrial appendages (LAAs). LD is based on the European super-population for CTSN (left) and the East Asian super-population for Harbin (right).

(**B**) Uniform manifold approximation and projection (UMAP) of LAA single-nucleus multiome cell-types (left) and *KDM1B* expression density (right). *KDM1B* is expressed in all cells.

(**C**) LAA single-nucleus multiome genomic context of the prioritized AF-associated variants. From top to bottom: The first track shows ATACseq read coverage at the *KDM1B* locus (left) paired with violin plots of *KDM1B* expression aggregated by cell-type (right). The second track shows ATACseq peaks. The third track shows the gene annotation (exon, intron) for the genes found at the locus. The fourth and fifth tracks highlight links between ATACseq peaks and gene promoters identified in either all cell-types (Links) or specifically in cardiomyocytes (CM). We use Pearson correlation tests between ATACseq peak accessibility and gene expression (in the same nucleus) to derive links. Only links with |Pearson R| > 0.2 are shown. Link heights are proportional to their |Pearson R| in the range indicated in the legend. In the final track, we report AF GWAS and eQTL fine-mapping posterior inclusion probabilities (PIP). One variant (rs34969716) has a PIP >0.1 in the AF GWAS and CTSN datasets (*KDM1B* was not an eQTL in the Harbin cohort). This variant overlaps an ATACseq peak in our LAA multiome data, and accessibility to this peak is anti-correlated with the expression of *KDM1B* (when considering all cell-types). rs34969716 also overlaps with an ATACseq peak identified in multiple cell-types (including cardiomyocytes) in CATlas, as well as a distal enhancer annotated in ENCODE.

(**D**) Scatter plots of the chr6:18209585-18210058 peak accessibility against *KDM1B* expression colored by the genotype of the prioritized variant (rs34969716) in the six genotyped individuals of our single-nucleus multiome data. Each point represents a metacell (**Methods**). We found one individual with the AA genotype, three with the GA genotype and two with the GG genotype. The relatively weak effect of the A-allele on *KDM1B* expression and open chromatin of chr6:18209585-18210058 is consistent with the weak eQTL effect observed in (**F**) (Harbin eQTL FDR = 0.052 and CTSN eQTL FDR = 0.0044).

(**E**) ATACseq read coverage of the chr6:18209585-18210058 ATACseq peak aggregated by genotype showing greater accessibility for the AA genotype (across all cell-types).

(**F**) rs34969716-*KDM1B* eQTL boxplots in the CTSN and Harbin bulk RNAseq datasets.

Adipo; Adipocytes, CM; Cardiomyocytes, EC; Endothelial cells, FB; Fibroblasts, PC; Pericytes, SMC; Smooth muscle cells.

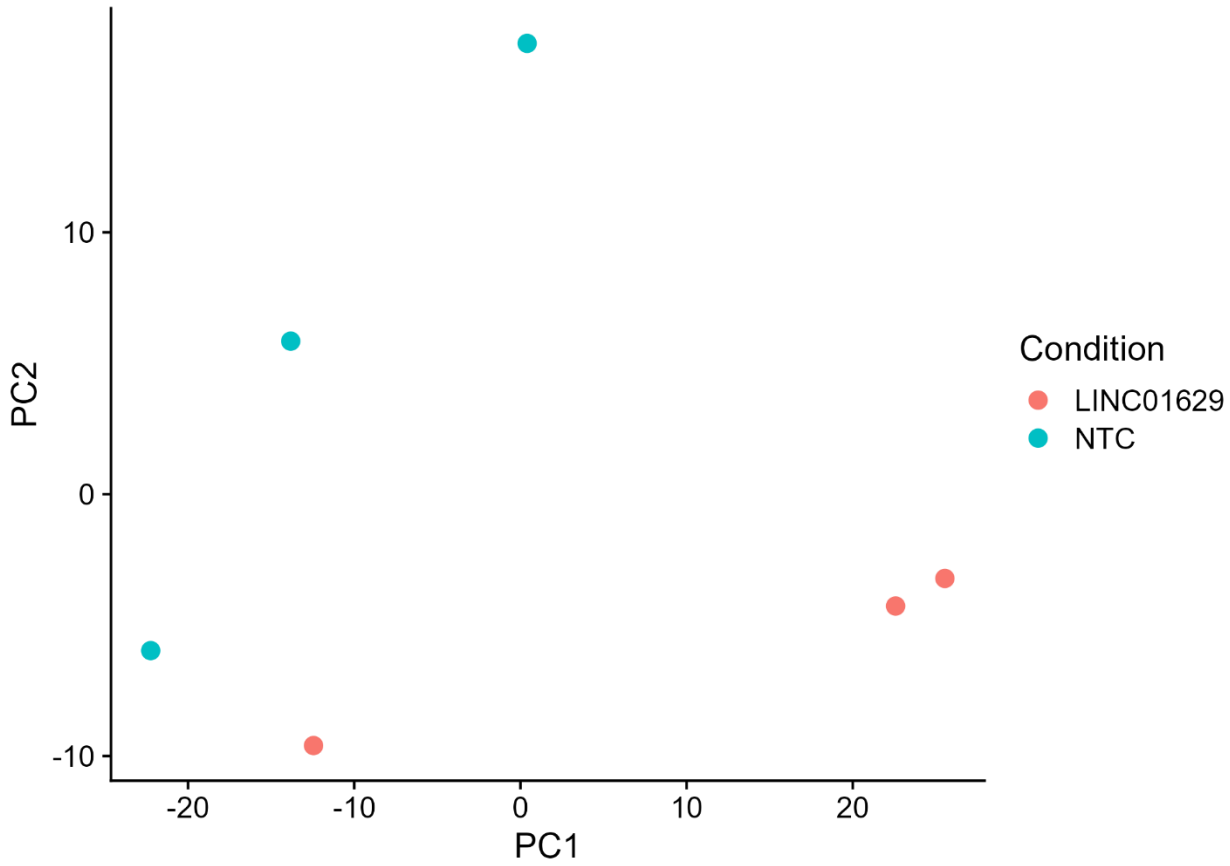


Figure S13. PCA of in vitro validation of *LINC01629* RNAseq.

Principal component analysis of the 500 genes with most variance in human induced pluripotent stem cells derived cardiomyocytes CRISPR interference with a guide RNA targeting the promoter of *LINC01629* vs a non-targeting control (NTC) guide RNA.

Table S1. Demographics and clinical information of the CTSN and Harbin cohorts.

To assess the difference between atrial fibrillation (AF) cases and sinus rhythm (SR) controls, we used a t-test for age and Fisher's exact test for all other characteristics. CABG; coronary artery bypass graft surgery.

	CTSN			***Harbin***		
Patient Characteristics	**AF n=31**	**SR n=31**	**P-value**	**AF n=28**	**SR n=37**	**P-value**
Mean age, y	69.3	65.6	0.075	59.25	58.19	0.62
Sex, female	10	14	0.43	15	9	0.02
Diabetes mellitus	3	16	0.0007	2	11	0.03
History of MI	4	23	0.000002	0	11	0.002
Hypertension	23	25	0.76	4	11	0.23
Surgical indication						
Isolated CABG	0	1		1	17	
CABG + Valve repair	2	21		0	0	
CABG + Valve replacement	5	4		0	1	
Valve repair alone	17	1		0	4	
Valve replacement alone	7	4		23	11	

Table S3. GTEx V8 right atrial appendage (RAA) eQTL results for the atrial fibrillation-associated variants and eGenes presented in Table 1.

Effects (slope) are reported for the alternate allele (ALT). ND; not detected in GTEx V8, NS; not significant in GTEx V8 RAA, MAF; minor allele frequency, slope_se; slope standard error, pval_nominal; nominal p-value, eGene; eQTL gene.

rsID	eGene	ALT	MAF	pval_nominal	slope	slope_se
rs4951258	*RAB29*	A	0.38	1.32E-18	-0.34	0.036
rs2885697	*AC093151.3*		ND	ND	ND	ND
rs2885697	*AC093151.8*		ND	ND	ND	ND
rs4970418	*PERM1*	A	0.16	4.30E-14	0.31	0.038
rs60212594	*MYOZ1*	C	0.13	1.39E-45	1.07	0.063
rs2316443	*F10*	A	0.23	5.69E-06	-0.18	0.039
rs11156751	*AKAP6*	C	0.27	4.25E-08	0.17	0.030
rs10873298	*AC007686.1*	T	0.44	1.16E-78	-1.10	0.043
rs10873298	*LINC01629*	T	0.44	5.34E-89	-1.18	0.041
rs12908004	*ARNT2*	G	0.16	1.77E-54	0.86	0.045
rs12908004	*CTXND1*	G	0.16	6.43E-29	0.93	0.074
rs12908004	*AC016705.2*	G	0.16	4.79E-43	0.87	0.053
rs242557	*MAPT*	A	0.35	1.79E-34	0.66	0.047
rs242557	*STH*		NS	NS	NS	NS
rs242557	*MAPT-IT1*		NS	NS	NS	NS
rs3820888	*SPATS2L*	C	0.40	4.34E-18	-0.30	0.033
rs2540949	*CEP68*	T	0.40	1.63E-54	-0.58	0.030
rs6089752	*MIR1-1HG-AS1*	T	0.46	1.74E-05	-0.15	0.034
rs5754508	*UBE2L3*		NS	NS	NS	NS
rs1278493	*AC072039.1*		ND	ND	ND	ND
rs7612445	*GNB4*	T	0.19	2.01E-70	1.09	0.046
rs34080181	*SLC25A26*	A	0.35	2.43E-09	0.22	0.036
rs223449	*BDH2*	T	0.49	4.65E-08	0.26	0.047
rs2012809	*SLC27A6*	G	0.18	6.59E-33	-0.52	0.038
rs3756687	*FAM13B*	G	0.17	2.10e-09	0.24	0.039
rs34969716	*KDM1B*	A	0.30	5.07E-23	0.43	0.040

Table S4. Differential expression of genes implicated by eQTL studies in left atrial appendages of normal (sinus rhythm) participants and atrial fibrillation patients.

L2FC, log2 fold-change; FDR_DEG, false discovery rate for differential expression of genes in the first column. We shaded significant results (FDR <0.05).

eGene	L2FC_CTSN	FDR_DEG_CTSN	L2FC_Harbin	FDR_DEG_Harbin
AC007686.1	-0.056	0.400	-0.0003	0.996
AC016705.2	0.120	0.171	-0.070	0.373
AC072039.1	-0.016	0.857	-0.002	0.986
AC093151.3	0.005	0.954	-0.013	0.879
AC093151.8	0.011	0.905	0.023	0.831
AKAP6	-0.026	0.773	0.100	0.201
ARNT2	0.063	0.427	-0.067	0.420
BDH2	0.052	0.485	0.006	0.962
CEP68	-0.060	0.355	0.013	0.904
CTXND1	0.286	0.043	-0.046	0.550
F10	-0.019	0.841	-0.060	0.486
GNB4	0.020	0.829	0.123	0.109
KDM1B	-0.096	0.099	-0.047	0.549
LINC01629	-0.073	0.300	0.005	0.925
MAPT	-0.645	0.0003	-0.144	0.084
MAPT-IT1	-0.316	0.023	0.023	0.765
MIR1-1HG-AS1	-0.052	0.519	-0.115	0.148
MYOZ1	-0.062	0.406	0.025	0.767
PERM1	-0.029	0.737	0.031	0.719
RAB29	0.020	0.822	0.038	0.667
SLC25A26	0.013	0.878	-0.129	0.082
SLC27A6	-0.718	0.00002	-0.235	0.031
SPATS2L	-0.143	0.049	0.027	0.793
STH	-0.067	0.396	-0.022	NA
UBE2L3	-0.082	0.083	-0.062	0.420

Table S5. Co-localization and Bayesian fine-mapping of the atrial fibrillation (AF) genome-wide association study (GWAS) and expression quantitative trait loci (eQTL) results.

eQTL significance: we report if the eQTL signal is significant (FDR <0.05) in the CTSN or Harbin cohorts, or both. We used *coloc* to test if the AF GWAS and eQTL signals are co-localized and report the posterior probability (PP) of hypothesis 4 (H4), that is that the GWAS and eQTL signals are co-localized (defined as H4 >0.4, shaded). We used approximate Bayesian fine-mapping to determine the size of the 95% credible sets (cs) for the AF GWAS, CTSN eQTL and Harbin eQTL signals. The "Overlap 95%.cs.size" and "Variants in the overlapping sets" include variants found in the AF GWAS cs and in at least one of the eQTL study cs. ND, not determined because the eQTL signal is not significant for this SNP and gene in this cohort.

			Co-localization		Fine-mapping 95% credible sets (cs)				
rsID	**CHR:POS:REF:ALT:eGene**	**eQTL significance**	**CTSN PP.H4**	**Harbin PP.H4**	**GWAS 95%.cs.size**	**CTSN 95%.cs.size**	**Harbin 95%.cs.size**	**Overlap 95%.cs.size**	**Variants in the overlapping sets**
rs4970418	chr1:983237:G:A_PERM1	CTSN	0.43	ND	5	7	ND	4	rs4970418, rs6659460 ,rs74045046 ,rs56028034
rs7612445	chr3:179455191:G:T_GNB4	Both	>0.99	0.93	5	2	3	3	rs7612445, rs7634416, rs2339798
rs3756687	chr5:137866004:A:G_FAM13B	CTSN	0.78	ND	3	3	ND	2	rs3756687 ,rs7722600
rs34969716	chr6:18209878:G:A_KDM1B	CTSN	0.66	ND	1	17	ND	1	rs34969716
rs11156751	chr14:32521231:T:C_AKAP6	Harbin	ND	0.88	4	ND	4	2	rs7140396 ,rs2145587
rs10873298	chr14:76960182:C:T_AC007686.1	Both	>0.99	0.98	5	4	1	4	rs10873298 ,rs8181996 ,rs10873299 ,rs12889775
rs10873298	chr14:76960182:C:T_LINC01629	Both	>0.99	0.98	5	4	1	4	rs10873298 ,rs8181996 ,rs10873299 ,rs12889775
rs12908004	chr15:80384583:A:G_ARNT2	Both	0.75	0.24	1	1	274	1	rs12908004
rs12908004	chr15:80384583:A:G_CTXND1	CTSN	0.59	ND	1	717	ND	1	rs12908004
rs12908004	chr15:80384583:A:G_AC016705.2	Both	0.79	0.31	1	1	6	1	rs12908004
rs242557	chr17:45942346:G:A_MAPT	Both	>0.99	>0.99	1	1	1	1	rs242557
rs242557	chr17:45942346:G:A_STH	CTSN	0.96	ND	1	1	ND	1	rs242557
rs242557	chr17:45942346:G:A_MAPT-IT1	CTSN	0.98	ND	1	1	ND	1	rs242557

rs6089752	chr20:62557186:C:T_MIR1-1HG-AS1	CTSN	0.68	ND	7	17	ND	6	rs6089752 ,rs6089741 ,rs6089753 ,rs6089750 ,rs12624794 ,rs4637207

Table S6. Description of the expression genes (eGenes) prioritized in our expression quantitative trait loci (eQTL) experiment to identify modulators of atrial fibrillation (AF) risk.

Gene	Full name	Genecards	Literature
PERM1	PPARGC1 And ESRR Induced Regulator, Muscle 1	Involved in response to muscle activity.	• In skeletal muscle, *PERM1* is an exercise-induced gene (DOI: 10.1016/j.molmet.2019.02.009) and regulates oxidative capacity (DOI: 10.1074/jbc.M113.489674). • Perm1 KO mice ejection fraction is reduced. PERM1 binds to ERRα in cardiomyocytes, bind and activate ERR promotor (DOI: 10.3389/fcvm.2022.1033457). • Lower cardiac *PERM1* expression has been observed during pressure overload and hypertrophic stress (DOI: 10.1371/journal.pone.0234913).
GNB4	G Protein Subunit Beta 4	This gene encodes a beta subunit. Beta subunits are important regulators of alpha subunits, as well as of certain signal transduction receptors and effectors.	• *GNB4* is associated to heart rate by GWAS (DOI: 10.1038/ng.2610). *GNB4* mutations is a cause of dominant intermediate Charcot-Marie-tooth disease (DOI: 10.1016/j.ajhg.2013.01.014). • Gnb4 KO mice have enlarged hearts. The encoded Gβ4 subunit regulates heart rhythm through the GIRK channel/M2 receptors. (DOI: 10.3390/cells8121567).
FAM13B	Family With Sequence Similarity 13 Member B	Predicted to enable GTPase activator activity	• The AF risk allele of rs17171731 reduce its host enhancer activity. Knockdown of *FAM13B* modified the sodium current of a human cardiomyocyte model. FAM13B may localize to the plasma membrane and at the Z-disk. (DOI: 10.1101/719914)
KDM1B	Lysine Demethylase 1B	Flavin-dependent histone demethylases, regulate histone lysine methylation	• *KDM1B* is a regulator of cellular reprogramming (DOI: 10.1016/j.yexcr.2022.113339).
AKAP6	A-Kinase Anchoring Protein 6	Involved in anchoring PKA to the nuclear membrane or sarcoplasmic reticulum	• Rare variants in *AKAP6* alter cAMP/PKA signaling (DOI: 10.1152/ajpheart.00034.2018). • May regulates *RYR2*, the sodium calcium exchanger and calcineurin/MEF2 regulatory complex. Cardiomyocyte-specific *AKAP6* KO are resistant to pressure overload. (DOI: 10.1016/j.yjmcc.2016.12.006)
AC007686.1	Ribosomal Protein, Large, P1 (RPLP1) Pseudogene	-	-
LINC01629	Long Intergenic Non-Protein Coding RNA 1629	-	• *LINC01629* is also a methylation QTL in human cardiac tissue (DOI: 10.1186/s12863-021-00975-2)
AC016705.2/ARNT2-DT	ARNT2 Divergent Transcript	-	-
ARNT2	Aryl Hydrocarbon Receptor Nuclear Translocator 2	Under hypoxic conditions, the encoded protein complexes with hypoxia-inducible factor 1alpha in the nucleus and this complex	• Arnt2 KO zebrafish have enlarged ventricles, decreased wall thickness, bradycardia and cardiac arrhythmia characterized

		binds to hypoxia-responsive elements in enhancers and promoters of oxygen-responsive genes	by missing heartbeats and over inflation of the atrium (DOI: 10.1089/zeb.2008.0536).
CTXND1/LINC01314	Cortexin Domain Containing 1	Predicted to be integral component of membrane	• Is differentially expressed in hypertrophic cardiomyopathy (DOI: 10.3390/ijms232315280). • Is involved in multiple cancers (DOI: 10.1186/s12935-019-0799-9)
MAPT	Microtubule Associated Protein Tau	MAPT transcripts are differentially expressed in the nervous system, depending on stage of neuronal maturation and neuron type. Promotes microtubule assembly and stability and might be involved in the establishment and maintenance of neuronal polarity.	• Mapt KO mice have systolic and diastolic dysfunction, decreased heart rate and increased heart rate variability (DOI: 10.1096/fasebj.2020.34.s1.02885). • Tau aggregates in the myocardium is associated with diastolic dysfunction (DOI: 10.1093/eurheartj/ehad205). • rs242557 appears to alter *MAPT* and other genes in microglial cell line (DOI: 10.1002/alz.052360).
MAPT-IT1	MAPT Intronic Transcript 1	-	-
STH	Saitohin	Involved in positive regulation of mRNA splicing, via spliceosome	• May interact with Tau. Is genetically associated with Alzheimer disease, Parkinson disease and schizophrenia. (DOI: 10.1002/jcb.23279).
MIR1-1HG-AS1/CRMA	Cardiomyocyte Maturation Associated LncRNA	-	• Its expression and the expression miRNAs are negatively correlated. Its knockdown increases *MIR1-1* and *MIR-133a2* expression (DOI: 10.1093/cvr/cvab281).

Table S7. Functional annotation of likely causal variants (posterior inclusion probability >0.1 for atrial fibrillation and eQTL).

All coordinated are on build hg38 of the human genome. cCRE, candidate *cis*-regulatory elements; enhD, distal enhancer; dELS, distal enhancer-like signature; enhP, proximal enhancer; dELS, distal enhancer-like signature.

				Left atrial appendage multiome		Cis-element atlas (CATlas)		Epimap	ENCODE	
Sentinel GWAS variant	**eGene**	**Prioritized variant**	**CHR:POS (hg38)**	**ATAC peak coordinates (hg38)**	**Link between ATAC peak and gene promoter**	**ATAC peak coordinates (hg38)**	**Cell-types**	**Coordinates (hg38) of regulatory element linked to gene in heart**	**UCSC label**	**cCRE**
rs4970418	*PERM1*	rs74045046	1:976536						enhD	dELS,CTCF-bound
rs4970418	*PERM1*	rs56028034	1:981282							
rs7612445	*GNB4*	rs7612445	3:179455191	chr3-179454910-179455284	CM	3:179454891-179455290	V Cardiomyocyte,A Cardiomyocyte	chr3-179455080-179455300		
rs7612445	*GNB4*	rs7634416	3:179455436			3:179455335-179455734	Keratinocyte 1,Hepatocyte			
rs3756687	*FAM13B*	rs3756687	5:137866004							
rs3756687	*FAM13B*	rs7722600	5:137859073			5:137858983-137859382	Cardiac Pericyte 1,Pericyte Muscularis,Parietal,Glomerulosa,V Cardiomyocyte,Endothelial Myocardial,Type I Skeletal Myocyte,Cardiac Fibroblast,Fetal Adrenal Cortical,Fetal Adrenal Neuron		enhD	dELS
rs34969716	*KDM1B*	rs34969716	6:18209878	chr6-18209585-18210058	All cells	6:18209767-18210166	Follicular,Beta 1,Alpha 1,Beta 2,Delta+Gamma,Alpha 2,Gastric Neuroendocrine,Alveolar Type 2,Chief,Enteric Neuron,Hepatocyte,V Cardiomyocyte,A Cardiomyocyte,Sm Ms Colon		enhD	dELS,CTCF-bound

							1,Sm Ms GE junction,Sm Ms GI,Fetal Macrophage Hepatic 3,Fetal Megakaryocyte,Fetal Gastri Goblet,Fetal Hepatoblast,Fetal Fibro GI,Fetal Cilliated			
rs11156751	*AKAP6*	rs7140396	14:32514611			14:32514264-32514663	Transitional Cortical,Glomerulosa,Enterocyte, Fetal Adrenal Cortical,Fetal Adrenal Neuron			
rs10873298	*AC007686.1/ LINC01629*	rs12889775	14:76959734							
rs10873298	*AC007686.1/ LINC01629*	rs10873298	14:76960182						enhP	pELS
rs10873298	*AC007686.1/ LINC01629*	rs10873299	14:76960368						enhP	pELS
rs10873298	*AC007686.1/ LINC01629*	rs8181996	14:77427469							
rs12908004	*AC016705.2/ ARNT2/ CTXND1*	rs12908004	15:80384583			15:80384470-80384869	Pericyte General 4,Fibro Liver Adrenal,Endothelial Exocrine,Type II Skeletal Myocyte		K4m3	DNase-H3K4me3,CTCF-bound
rs242557	*MAPT/ MAPT-IT1/ STH*	rs242557	17:45942346	chr17-45942197-45942667	All cells	17:45942191-45942590	Fibro Epithelial,Pericyte General 1,Pericyte General 2,Pericyte General 3,Cardiac Pericyte 2,Pericyte Muscularis,Adipocyte,Beta 1,Beta 2,Gastric Neuroendocrine,Mammary Luminal Epi 1,Mammary Basal Epi,Mammary Luminal Epi 2,Satellite,Schwann General,Melanocyte,Enteric Neuron,Cortical Epithelial,Hepatocyte,Oligo Precursor,Astrocyte 1,Oligodendrocyte,V Cardiomyocyte,A Cardiomyocyte,Sm Ms Vaginal,Sm Ms Mucosal,Sm Ms	17:45942297-45942483 17:45942320-45942640	enhD	dELS,CTCF-bound

							Muscularis 1,Sm Ms Colon 1,Sm Ms Muscularis 2,Sm Ms GE junction,Sm Ms Colon 2,Sm Ms Muscularis 3,Type I Skeletal Myocyte,Type II Skeletal Myocyte,Vasc Sm Muscle 1,Vasc Sm Muscle 2,Cardiac Fibroblast,Paneth,Fetal Chromaffin,Fetal Enteric Neuron,Fetal Enteric Glia,Fetal Schwann General,Fetal Skeletal Myocyte 1,Fetal Satellite 1,Fetal Satellite 2,Fetal Skeletal Myocyte 2,Fetal Excitatory Neuron 9,Fetal Excitatory Neuron 10,Fetal Astrocyte 1,Fetal Astrocyte 2,Fetal Astrocyte 3,Fetal Astrocyte 4,Fetal Oligo Progenitor 2,Fetal Fibro Placental 1,Fetal Astrocyte 5,Fetal V Cardiomyocyte,Fetal Fibro General 3,Fetal Excitatory Neuron 1,Fetal Excitatory Neuron 2			
rs6089752	*MIR1-1HG-AS1*	rs6089753	20:62556900							

Chapter 5: Revealing cell-type specific gene dysregulation under persistent atrial fibrillation with single-nucleus multiomics

Francis J.A. Leblanc[1,2], Neelam Mehta[3], Svetlana Reilly[3], Stanley Nattel[1,2,4,5,6], Guillaume Lettre[1,2]

[1]Montreal Heart Institute, Montreal, Quebec, Canada; [2]Department of Medicine, Université de Montréal, Montréal, Québec, Canada; [3]Division of Cardiovascular Medicine, Radcliffe Department of Medicine, University of Oxford, John Radcliffe Hospital, Oxford, UK, UK; [4]IHU Liryc and Fondation Bordeaux Université, Bordeaux, France; [5]Institute of Pharmacology, West German Heart and Vascular Center, Faculty of Medicine, University Duisburg-Essen, Essen, Germany, [6]Department of Pharmacology and Therapeutics, McGill University, Montreal, Quebec, Canada.

Correspondence to: Guillaume Lettre, Montreal Heart Institute, 5000 Belanger St, Montreal, Quebec, H1T 1C8, Canada. Email guillaume.lettre@umontreal.ca

4 figures

16 supplemental figures and 11 supplemental tables

5.1 ABSTRACT

The etiology of persistent atrial fibrillation (AF), the most common supraventricular arrhythmia, remains incompletely understood. Gene dysregulation during AF has been exhaustively studied, yet few genes have been consistently replicated. Moreover, the cellular origin of this dysregulation remains undetermined. Here we comprehensively characterize the cellular landscape of the left atrial appendages (LAA) of AF and sinus rhythm patients using the 10X multiome assay (paired single nucleus RNAseq (snRNAseq) and open chromatin (snATACseq)). Our differential expression analysis reveals 755 robust differentially expressed genes (DEGs) validated in two independent large-sample-size bulk RNAseq datasets. By integrating bulk RNAseq and snRNAseq data, we identify multiple non-coding genes at the *IFNG* locus (*LINC01479* and *IFNG-AS1*) that stand out as cardiomyocytes-specific and as the strongest transcriptional signals in AF. We further identify cell-type-specific gene modules suggesting an increase in T-cell and decrease in adipocyte and neuronal cells gene expression in AF. Additionally, we detect gene modules enriched in distinct fibroblast states and suggest transcription factors that regulate these modules. Lastly, we identify the androgen receptor as repressor of AF DEG signature in cardiomyocytes and highlight novel gene targets that show high cardiomyocytes and AF specificity including *SYNPR*, *COLQ*, *CHRNE*, *PDE8B, LINC01479* and *IFNG-AS1*.

5.2 INTRODUCTION

Atrial fibrillation (AF), already the most common arrhythmia, is expected to see its prevalence more than double in the US population by 2050[1]. This is likely attributable to the world's aging populations, given that age is the strongest risk factor for AF. To date, the genetics of AF have almost exclusively implicated cardiomyocytes (CM) as causal driver of the disease. Rare mutations in familial AF patients are largely found in ion channels, cardiac transcription factors or cytoskeleton-associated proteins[53]. AF genome wide association studies (GWAS) were shown to specifically enrich for CM open chromatin regions[287,307]. On the other hand, an important body of research has implicated other processes such as fibrosis and inflammation, pointing towards the contributions of other cell-types in AF[42,53,64]. Furthermore, AF heritability is estimated to be 22% in Europeans[125], underlying the importance of environmental factors which may alter disease risks through different mechanisms.

To identify novel therapeutic targets, many groups conducted differential gene expression studies of the atria from AF and sinus rhythm (SR) patients. Most of these studies, however, were either small, limited to some coding genes by microarray probes or focused on differences across atrial chambers[172]. Together, this likely explains the relatively low replication rate of differentially expressed genes (DEG) reported across these studies[176]. Surprisingly, in a census of AF DEG by Victorino et al., none of the most frequently replicated DEGs were ion channels[176]. This underlines the need to establish reproducible AF DEGs in human atria. Moreover, the cell-types and transcription factors (TFs) causing the observed DEGs in bulk RNAseq from atria remains elusive.

Here we sought to establish and characterize robust DEG in AF at the cellular level and uncover their potential TFs causing their dysregulation. Using the 10X multiome assay, we profiled paired single nucleus RNAseq (snRNAseq) and open chromatin (snATACseq) in the same nuclei from left atrial appendages (LAA) of AF and SR patients. By integrating cell-type specific insights with a robust list of LAA AF DEGs replicating in two large-scale bulk RNAseq datasets, we leveraged the statistical power of large-sample-size bulk RNAseq and the resolution of our multiome dataset to refine AF DEGs by cell-type. We identify cell-type specific DEG modules in rare cardiac cell-types and fibroblasts. Focusing on CM, we find that our robust CM specific AF DEG signature is associated with the androgen receptor motif activity and its expression. Lastly,

we confirm the specificity of this signature against other cardiomyopathies and suggest multiple potential AF specific gene targets.

5.3 RESULTS

5.3.1 The cellular landscape of LAA

To dissect persistent AF dysregulated genes and their TF regulators at the cellular level we used the 10X multiome assay, profiling both RNA and ATACseq in the same nuclei of 4 AF and SR LAA. After stringent quality control and doublet removal, we obtained a dataset composed of 7 samples (3 AF [CF93, CF97, CF102] and 4 SR [CF69, CF77, CF89, CF91]) and 11986 nuclei (**Fig S1-S4**). Hereafter we refer to this dataset as "scAF". We annotated cell-types by co-embedding our LAA nuclei with the human heart atlas left atrial (LA) nuclei[306], identifying 12 major cell-types (**Fig. S3C-D**). Using dimensionality reduction with either RNA or ATAC modalities independently (**Fig. S4C-D**) or both combined (**Fig. 1A**), we show that clustering the scAF dataset without integration with the heart atlas recapitulates the same cell-type specific clusters. To the exception of mesothelial cells, the cell-types generally were well distributed across all samples (**Fig. 1B**). Similarly to what was reported in other human cardiac datasets[306,314], CM and fibroblasts (FB) were the two most abundant cell-types, accounting for 25% and 23% of nuclei respectively (**Fig. 1B**). We further validated cell-type identities showing the enrichment of known cell-type specific gene expression (**Fig. 1C, S3** and **Table S1**), such as *FGF12*, *TTN* and *RYR2* in CM, *DCN* in FB and *PTPRM* and *PECAM1* in endothelial cells (EC).

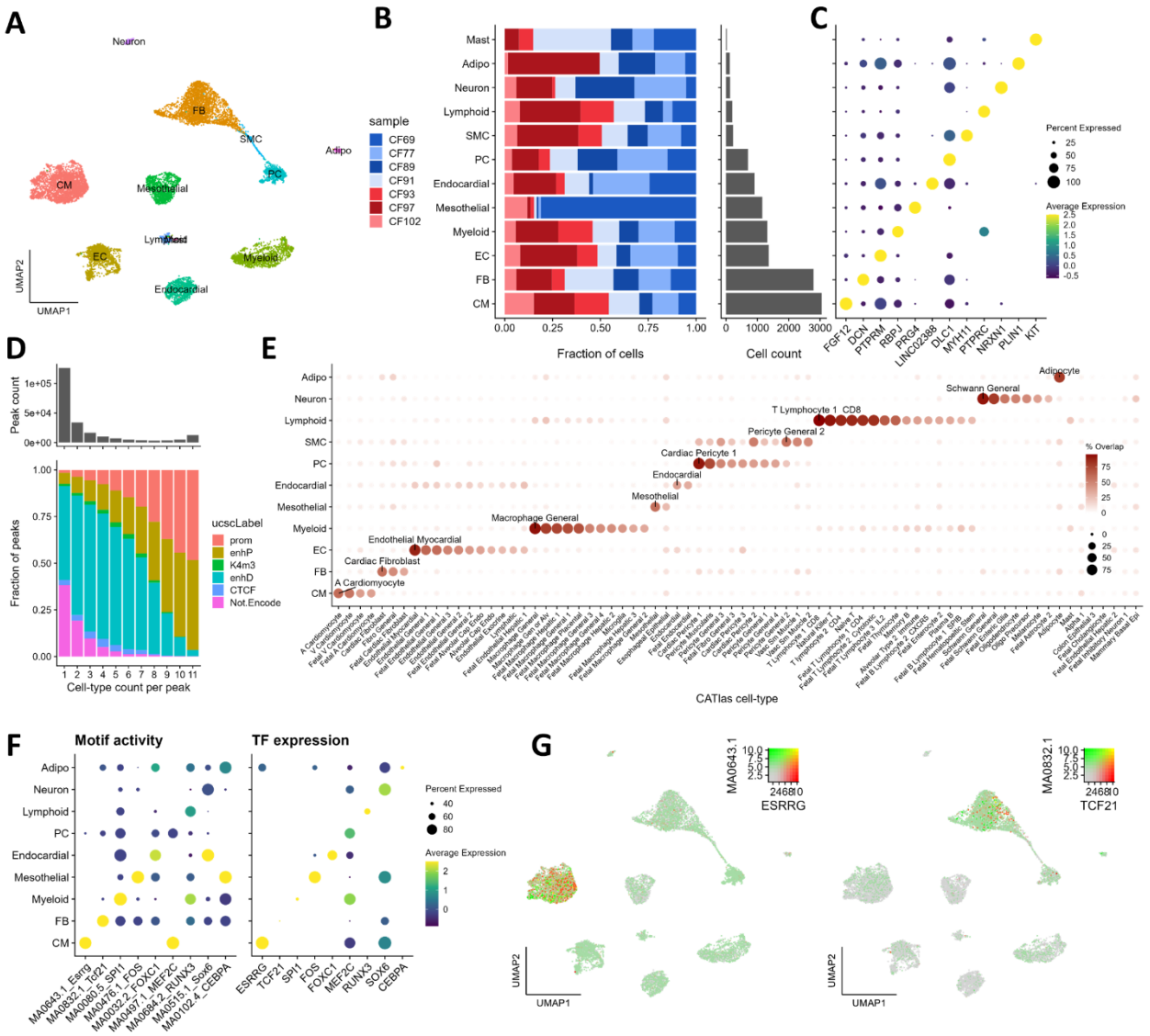


Figure 1. Cellular landscape of LAA.

A) Single nuclei uniform manifold approximation and projection (UMAP) colored by cell-type. **B)** Bar plot showing (left) the proportion of each sample by cell-type and (right) nuclei count per cell-type. Atrial fibrillation and sinus rhythm samples are colored in red and blue respectively. **C)** Dot plot, showing the strongest marker genes for each cell-type. **D)** (top) Histogram of peaks count by number cell-types in which the peak is detected and (bottom) bar plot showing the proportion peak overlapping each ENCODE cis-regulatory element (cCRE) types by number cell-types in which the peak is detected. **E)** Dot plot comparing the percentage of overlap of cell-type specific peaks between the left atrial appendage scAF dataset and the human enhancer atlas cell-types. We added labels for the human enhancer atlas cell-types with the strongest overlap for each of the scAF cell-types. **F)** (left) Dot plots showing the most enriched motif activity scores and (right) their transcription factor (TF) expression by cell-type. **G)** Single nuclei UMAP colored by motif activity (green) and TF expression (red) for (left) ESRRG and (right) TCF21. Prom; promoter, enhP; proximal enhancer, K4m3; lysine 4 tri-methyl mark, enhD; distal enhancer, CTCF; CCCTC-binding factor mark, not.Encode; peak found in the scAF dataset without any overlapping ENCODE cCRE, Adipo; Adipocytes, CM; Cardiomyocytes, EC; Endothelial cells, FB; Fibroblasts, PC; Pericytes, SMC; Smooth muscle cells.

5.3.2 The chromatin landscape of LAA

To validate our ATACseq data, we then compared the ATACseq peaks called in our LAA dataset (n=212,084, hereafter referred simply as "peaks") to the cross-tissue/cell-type ENDODE[413] candidate *cis*-regulatory elements (cCREs). Most peaks overlapped ENCODE cCREs (154,801, 73%) and were found in only one cell-type (124,843, 59%). We found the number of cell-types in which a peak was detected to be strongly associated with the ENCODE cCRE types (**Fig. 1D**). For instance, promoters were generally accessible in multiple cell-types, while distal enhancers had greater cell-type specificity, an observation previously reported by others[414]. Furthermore, peaks that did not overlap with any ENCODE cCRE, most likely because of their specificity to adult atrial tissue, were more often cell-type specific and displayed characteristics analog to distal elements i.e., higher cell-type specificity, lower read counts, GC content and length (**Fig. S5**). To rule out that these peaks were false positives, we compared them to the human enhancer atlas[287], which is composed of 222 cell-types from 30 adult and 15 fetal tissues. We found that the largest overlap systematically corresponded to an analogous cell-type. For example, our CM specific peaks had the highest overlap with adult atrial cardiomyocytes of the human enhancer atlas, our FB specific peaks with adult cardiac fibroblasts and our EC specific peaks with endothelial myocardial cells (**Fig. 1E**).

Lastly, we assessed TF motifs activity by cell-type using ChromVar scores derived with the JASPAR 2020 database. Motifs of the same family often produce similar scores given the similarity of their motifs. This makes the identification of the causal TF difficult. To overcome this limitation, we leveraged the bimodality of our scAF dataset and filtered the TFs expression in conjuncture with their corresponding motif activities to refine their selection (**Table S2**). Using this strategy, we selected candidate TFs with known cell-type specific functions and novel ones. For instance we selected both ESRRG and TBX5 in CM, which are involved in CM maturation[415-417], while TCF21[418] showed similar enrichment in FB (**Fig. 1F-G** and **Table S2**).Our analysis also suggests potentially novel specific TF actions such as FOXC1 in endocardial cells.

5.3.3 Upregulation of the IFNG locus in CM is the strongest transcriptomic feature of persistent AF

While some large-scale studies have investigated the changes in gene expression that occur during AF at the tissue level[173,174], these changes have yet to be dissected at the cellular level. Our

scAF dataset has a relatively low number of samples, which provides limited power to detect DEGs. To address this, we chose to rely on large-sample-size LAA bulk RNAseq datasets. Thus, we first selected genes of interest based on our previously published bulk RNAseq dataset (labeled CTSN hereafter) and that of J.Hsu et al. (labeled J. Hsu hereafter)[216]. Afterwards, we investigate their pattern of expression in single nuclei from the same tissue type to decipher cell-type specific DEGs and their TF regulators.

The two bulk RNAseq datasets (**methods**) were composed of 31 AF and 31 SR samples (CTSN dataset), and 130 AF and 50 SR samples (J. Hsu dataset[216]). First, using principal component analysis (PCA), we found a dominant effect of sex in both datasets (**Fig. S6A**). We also found that mesothelial cell marker genes constituted features with the strongest contribution for the first principal component (PC1) (**Fig. S6B-D**). While this may be due to an epicardial sampling bias, there is evidence of a strong inter-individual variability in epicardium thickness and composition as well as growing evidence for its involvement in AF[419,420]. Furthermore, this inter-individual variability is consistent with the overrepresentation of mesothelial cells in two of our samples in our scAF dataset (**Fig. 1B**).

We then conducted differential expression analyses in both bulk RNAseq datasets and our scAF dataset with sex as covariate. Bulk differential expression analyses resulted in 3531 and 3619 DEGs (false discovery rate [FDR] <0.05) in our CTSN dataset and the J. Hsu dataset respectively (**Table S3**). We found a strong overlap and concordance of effect direction between the 2 datasets (Fisher exact test p-value = 1.7×10^{-163}, **Fig. S7A)**. Eight hundred DEGs were common to both CTSN and J. Hsu of which 755 (hereafter referred to as robust DEGs) and 45 had concordant and discordant direction of effect respectively (**Fig. S7A-B**). In our scAF dataset, we found 118, 77, 48, 38, 10, 10, 1 and 1 DEGs in FB, CM, pericytes (PC), myeloid cells, EC, endocardial cells, adipocytes and smooth muscle cells respectively (**Fig. 2A** and **Table S4**).

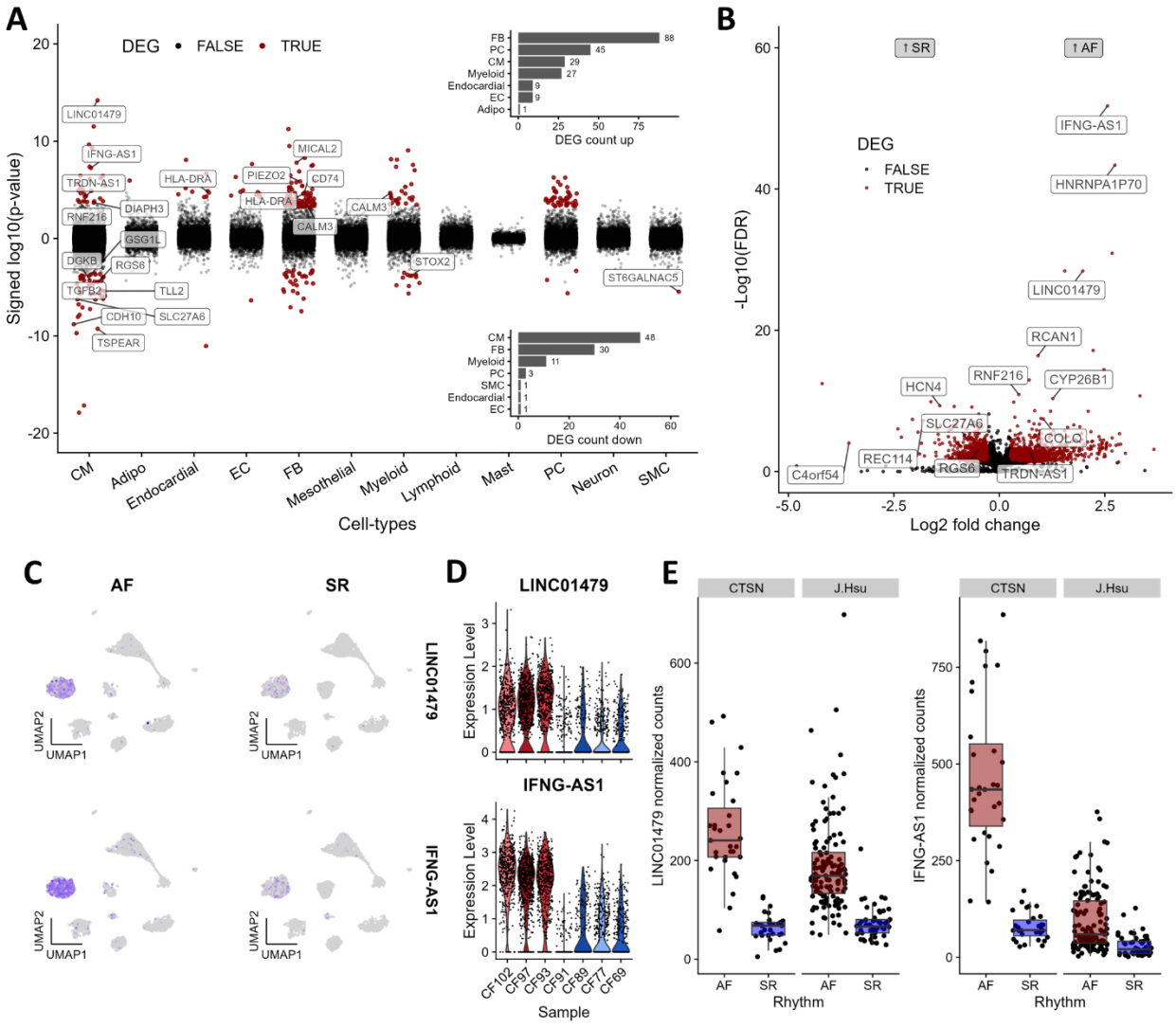


Figure 2. Cell-type specific contributions to bulk AF dysregulated genes.

A) Scatter plot showing the signed log10(p-value) for pseudo-bulk differential gene expression analysis of each cell-type between atrial fibrillation (AF) patients and sinus rhythm (SR) patients. The sign value was attributed based on the sign of the log2 fold change. Gene colored in red are differentially expressed (false discovery rate [FDR] < 0.05). We labeled genes that were also differentially expressed in bulk RNAseq. The inset histograms show the number of (top) upregulated differentially expressed genes (DEGs) and (bottom) downregulated DEGs in each cell-type. **B)** Volcano plot showing the results of differential expression analysis in the CTSN cohort. Genes colored in red have an FDR < 0.05 and an |log2 fold change| > 0.25. **C)** Single nuclei uniform manifold approximation and projection (UMAP) colored based on the level of normalized expression of (top) *LINC01479* and (bottom) *IFNG-AS1* in (left) AF and (right) SR patients. This shows increased expression of both genes (blue) in cardiomyocytes of AF patients. **D)** Violin plots showing the cardiomyocytes normalized expression of (top) *LINC01479* and (bottom) *IFNG-AS1* in each sample. Violin distributions of samples colored in red and blue are AF and SR respectively. **E)** Boxplot showing the normalized counts of (left) *LINC01479* and (right) *IFNG-AS1* in the two bulk RNAseq cohort used in this study. Adipo; Adipocytes, CM; Cardiomyocytes, EC; Endothelial cells, FB; Fibroblasts, PC; Pericytes, SMC; Smooth muscle cells.

Across both bulk datasets and our scAF dataset we found consistently that genes at the *IFNG* locus produced the strongest signals (**Fig. 2A-B** and **Fig. S8**). We show that the contiguous genes *LINC01479* and *IFNG-AS1* were both CM specific (**Fig. 2C**). Furthermore, they were highly discriminant of AF CM for all scAF samples (**Fig. 2D**), with *LINC01479* showing higher consistency across bulk datasets (**Fig. 2E**). While investigating this locus, we found that transcription sharply increased and diverged between AF and SR patients at the *LINC01479* promotor and gradually decreased towards *MDM1* (**Fig. S8B**). We also found divergent DEGs results at this locus for genes on the negative strand between bulk RNAseq datasets. Our stranded RNAseq suggests a signal specific to the positive strand (*LINC01479*, *IFNG-AS1*, *HNRNPA1P70* and *AC007458.1*) as opposed to both strands in J. Hsu unstranded data (**Table S3**). To assess the impact of the assay strandedness, we realigned our data without strand specification and compared p-values using the same dataset. The alignment without strand specification produced signals on both strands (**Fig. S8B**), suggesting that DEGs on the negative strand at this locus may be due to the unstranded nature of the assay. For instance, the FDR of *IL26* (located on the negative strand) goes from 10^{-47} to 1 when strand information is provided for read mapping in our CTSN dataset. We further annotated 66 DEG in the J. Hsu dataset that were dependent on strand information in CTSN (**Fig. S8A, Table S3** and **methods**). Namely, *GRM8*, the third strongest hit in J. Hsu, was not significant upon inclusion of strand information in the CTSN cohort. Instead, our data suggests the 2 anti-sense long non-coding RNAs (lncRNA) *AC002057.1* and *AC002057.2* to be the source of signal at this locus.

5.3.4 CMs produce more reproducible DEGs

Across the 755 robust DEGs identified by bulk RNAseq, we find 22 and 6 DEGs in our scAF with concordant and discordant direction of effect respectively (**Fig. S9A**). All 13 overlapping robust DEGs in CM had concordant effects, while FB DEGs diverged in 3 out 8 DEGs. Given the relatively similar abundance of CM and FB (25% and 23% respectively, **Fig. 1B**), we sought to explain the greater DEGs concordance in CM. Comparing the proportion of unique molecular identifier (UMI) and proportion of cells in each cell-type, we found that CM appeared transcriptionally more active (**Fig. S9B, D**). For instance, CM UMIs accounted for 48% of all UMIs while representing only 25% of cells. This relationship was inverted in FBs, with their UMIs accounting for 13% of all UMIs and while only accounting for 23% of cells. The ratios describing transcriptional activity were correlated (correlation of ratios: Spearman R = 0.69, p-value = 0.017)

and largely concordant in the heart atlas (**Fig. S9C, E**). While AF heritability has been shown to be mostly explained by CM cCREs[287], their seemingly higher transcriptional activity may also contribute to the greater number and reproducibility of CM DEGs that we found.

5.3.5 Gene module analysis identifies shared, and cell-type specific programs dysregulated in AF

Considering these results, we reasoned that partitioning genes into modules could identify AF gene programs that would be otherwise masked by the prevailing CM transcriptomes. To this end, we used WCGNA on both bulk RNAseq combined. This analysis partitioned 7970 genes into 16 modules of variable sizes ranging from 74 to 2077 genes (**Fig. 3A, S10A, Table S5** and **methods**). We found that all modules, except for the uncorrelated gene module (labeled grey), strongly enriched for either upregulated or downregulated AF genes (**Fig. 3A-B, S10B** and **Table S5**).

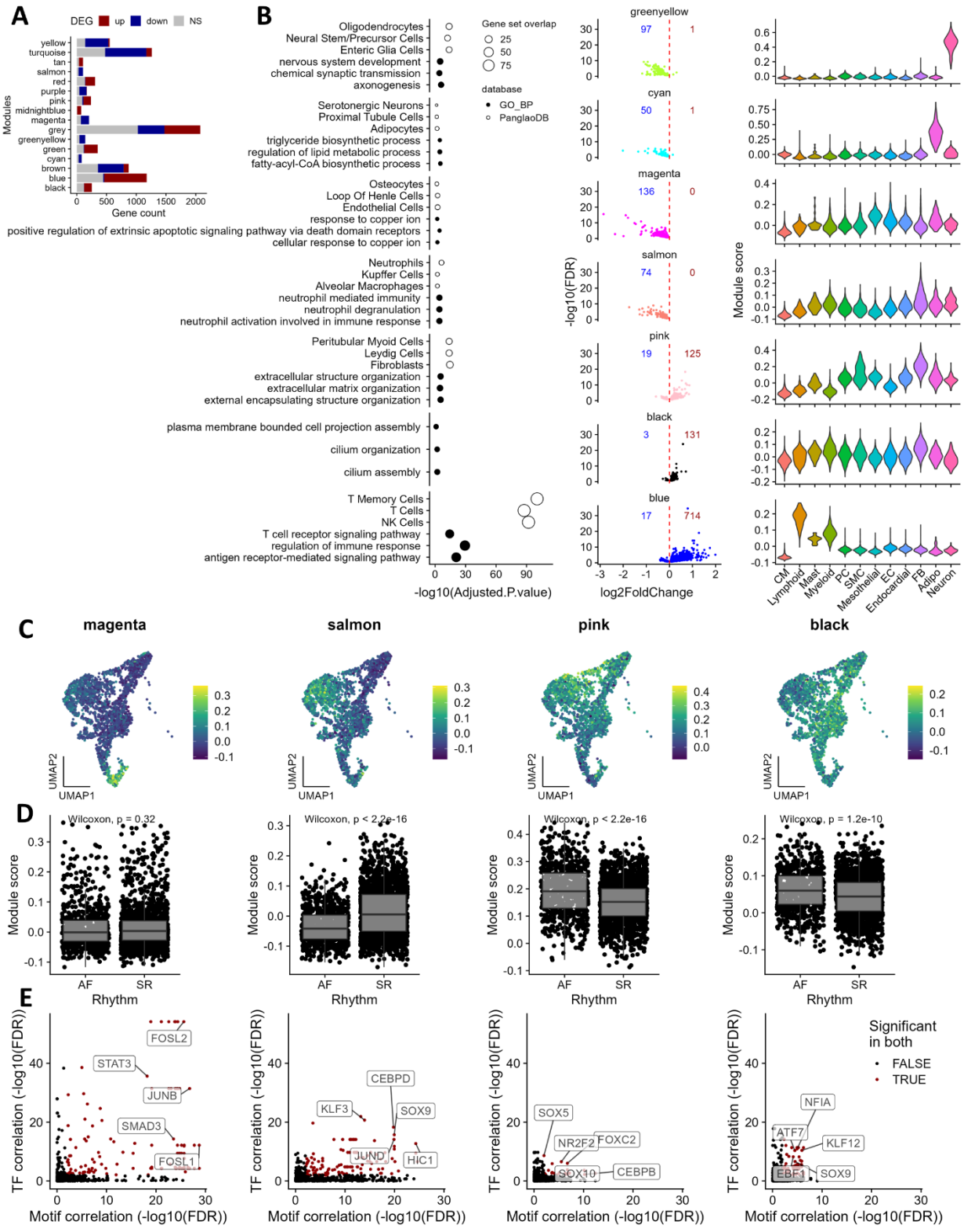


Figure 3. AF gene modules and their regulators.

A) Bar plot showing the number of genes per bulk RNAseq module (**methods**). The bars are colored in red, blue and grey for atrial fibrillation (AF) vs sinus rhythm (SR) bulk RNAseq upregulated (up) differentially expressed genes (DEG), downregulated (down) DEG and non-significant (NS) genes respectively. **B)** (left) Dot plots sowing the top 3 gene sets from an overrepresentation analysis of the DEGs in each module (**methods**). For this analysis we used the two gene set libraries PanglaoDB and gene ontology biological process (GO BP). (center) Volcano plots showing the log2 fold change and log10(false discovery rate [FDR]) statistics from the DEG analysis, stratified by modules. Red and blue integers indicate the number of AF upregulated and downregulated genes respectively in each module. (right) Violin plot showing the module scores (**methods**) in each cell-type of the scAF dataset. **C)** Uniform manifold approximation and projection (UMAP) of fibroblasts colored by module scores. From left to right, the scores are shown for the magenta, salmon pink and black modules. The same order is kept in D and E. **D)** Boxplot showing the distribution of module scores between AF and SR fibroblasts. **E)** Scatter plot showing the -log10(FDR) of the Pearson correlation of motif activity (x-axis) and its corresponding transcription factor (TF, y-axis) against the module score. Red dots indicate an FDR < 0.01 for both the correlation of the TF and its motif with the module score. Adipo; Adipocytes, CM; Cardiomyocytes, EC; Endothelial cells, FB; Fibroblasts, PC; Pericytes, SMC; Smooth muscle cells.

Using gene set analysis and single cell gene set score calculated with Seurat (hereafter referred as single cell enrichment scores, **methods**), we found shared and cell-type specific modules. For instance, we show that single cell enrichment scores for the greenyellow and cyan modules (mostly downregulated genes in AF) were highly specific to neurons and adipocytes respectively, while the scores for the blue and green modules (mostly upregulated genes in AF) were highly lymphoid and mesothelial specific (**Fig. 3B, S10B-C** and **S11**). Moreover, for these cell-type specific modules, we found concordant gene set enrichments, namely lipid metabolism, nervous system development, T-cell and epithelial cell pathways in the greenyellow, cyan, blue, and green modules respectively (**Fig. 3B, S10B** and **Table S6**). Despite the strong gene set enrichment results for the tan and midnighblue modules, we found no enrichment for their single cell scores in any specific cell-type in our scAF dataset (**Fig. S10B-C**). Their gene set enrichments suggest that cell division and plasma cells are more prevalent in AF. Three modules with some of the highest number of DEG had the highest single cell enrichment scores in CM, while the magenta, salmon, black and pink modules had low cell-type specificity.

Fibrosis is known to increase incidence and progression of AF. Interestingly, we observed that the non-specific modules (magenta, salmon, black and pink) appeared higher in a subset of the nuclei in the FB cluster (**Fig. S11**), which prompted us to investigate their effect specifically in FB. Our sub-cluster analysis of FBs resulted in 3 FB states with similar gene profiles to FB states previously reported in ventricles of dilated and arrhythmogenic cardiomyopathy patients[314], which we labeled myofibroblast (MFB), resident fibroblasts (RFB) and pro-inflammatory (pIFB) (**Fig. S12A-C**). RFB were enriched for canonical FB markers *DCN*, *GSN* and *TCF21*, MFB for pro-

fibrotic markers and smooth muscle contraction markers such as *TNC, COL1A1* and *ACTA2*, and pIFB for *NR4A1*, *THBS1* and *CCL2* (**Table S7**). We found the distribution of the magenta, salmon and pink modules scores to be higher in the pIFB, both RFB and pIFB, and MFB respectively (**Fig. 3C** and **S12D**). We asked whether the trend in bulk DEG direction for these modules could also be observed in our scAF FBs. Concordantly, the module scores were higher in AF FB for the pink and black modules and lower in AF for the salmon module (**Fig. 3D**), but we found no difference for the magenta module scores, which may reflect a lack of power given the low sample size of our scAF dataset.

Given the likely relevance of these modules in AF, we sought to identify TFs that may govern their expression. To this end, we correlated TFs expression and motif activities with the modules single cell scores in FB metacells (**methods**). We found strong candidate TFs regulators for the two modules with downregulated DEG in AF (magenta and salmon), while the modules with upregulated DEG in AF (pink and black) showed weaker associations (**Fig. 3E, S12E** and **Table S8**). TFs expression and motif activity for FOSL2 and JUNB (or their joint motif FOSL2::JUNB) showed the strongest correlation with the magenta module. In the salmon module, we find CEBPD, KLF3 and JUND as strongest signals which have jointly been implicated in adipocyte differentiation and fibroblast quiescence[421,422].

5.3.6 The androgen receptor as regulator of AF upregulated genes

Next, we sought to identify CM states and possible TFs associated with AF. We first sub-clustered CM, yielding 2 clusters which showed a strong composition bias for one SR individual (**Fig. S12A, C-D**). To our knowledge, this individual was the only one who suffered from an MI (less than a year before tissue collection), possibly explaining the strong difference of its CM transcriptome. We found that genes enriched in cluster 1 were associated with familial isolated hypertrophic cardiomyopathy (**Fig. S12B, E** and **Table S9**). To our dismay, we could not identify CM states specific to AF using sub-clustering analysis (**Fig. S12F**). Instead, we leveraged our robust AF DEG signature and scored CM either with robustly AF upregulated genes (AF signature UP) or downregulated genes (AF signature DOWN, **Fig. 4A**). With both signatures, we show a clear CM segregation between AF and SR individuals (**Fig. 4B**). Hereinafter we evaluate gene expression and motifs activities that change along this continuum.

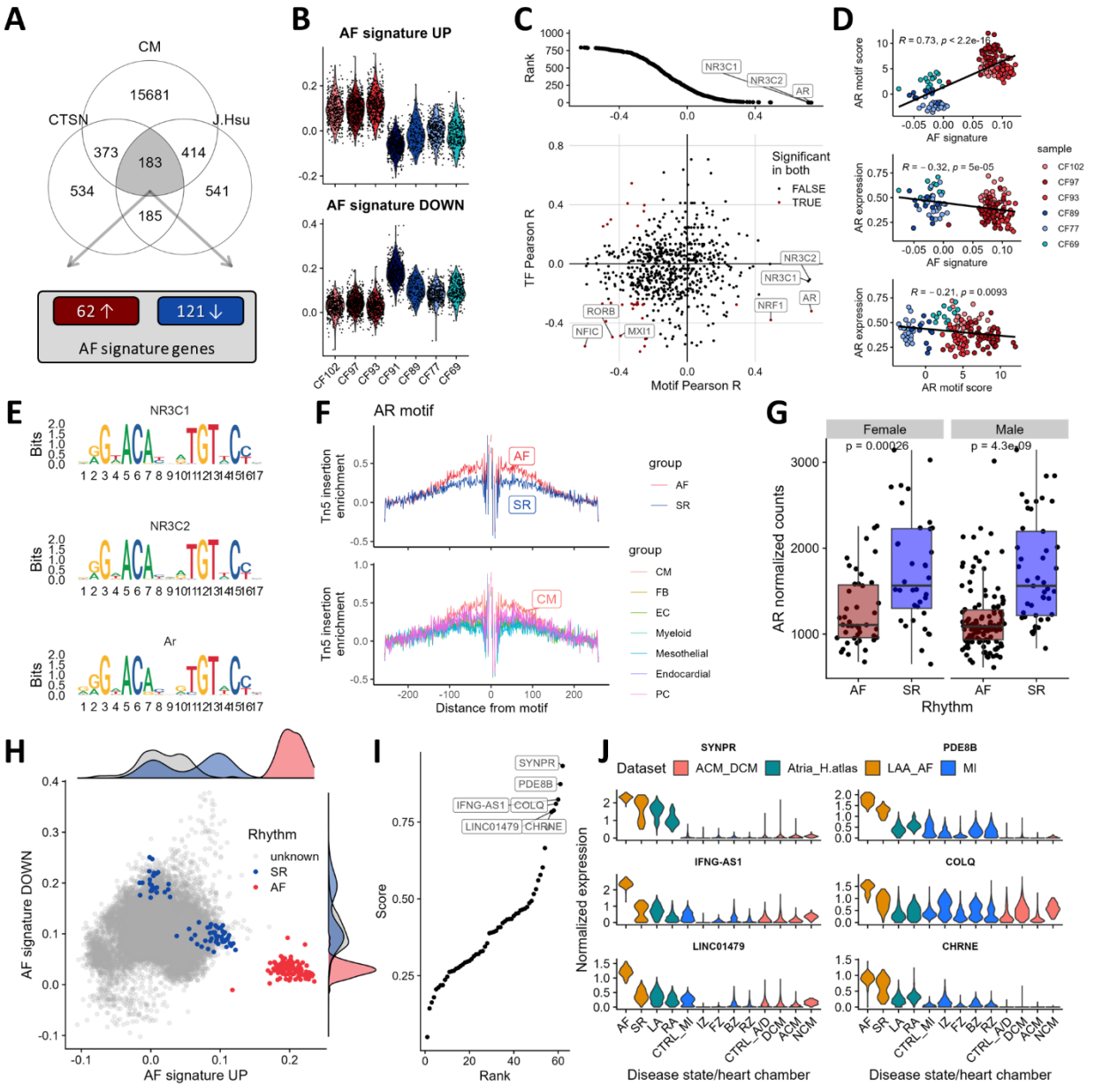


Figure 4. The androgen receptor regulates AF's cardiomyocyte specific gene signature.

A) (**top**) Venn diagram showing the number of genes intersecting between; 1-the differentially expressed genes in the CTSN cohort, 2-the differentially expressed genes in the J.Hsu cohort and 3-the genes with higher expression in cardiomyocytes (CM). (**bottom**) Number of upregulated and downregulated genes found in the 183 intersecting genes that constitute the atrial fibrillation (AF) signatures UP (in red) and DOWN (in blue). **B)** Violin plots of AF signatures in cardiomyocytes for each sample. Red and blue samples indicate AF and sinus rhythm (SR) samples respectively. **C)** (**top**) Rank plot showing the Pearson R for the motif activities correlation with the AF signature UP scores in CM metacells (**methods**). (**bottom**) Scatter plot of the transcription factors (TF) and their motif activities correlation with the AF signature UP scores in CM metacells. Red dot represent TFs for which the expression and their motif activity is significantly correlated (false discovery rate < 0.01) with the AF signature UP scores. **D)** From top to bottom respectively; scatter plots showing 1-the androgen receptor (AR) motif activity against the AF signature UP, 2-*AR* expression

against the AF signature UP and 3- *AR* expression against its motif activity, in CM metacells. For each scatter plot, we show the Pearson R and its nominal p-value. **E)** Motifs logos of NR3C1, NR3C2 and AR from the JASPAR 2020 database. **F)** Footprinting enrichments of the AR motif in AF vs SR (top) and in the most prevalent cell-types. We represent all cell-types individually in **fig. S13** for clarity because the lowly abundant cell-types produced noisy footprints. **G)** Boxplot showing the normalized counts of *AR* in (**left**) males and (**right**) females with their nominal Wilcoxon p-values in both bulk RNAseq cohorts combined. **H)** Scatter and density plot showing AF signatures scores in CM metacells from four adult cardiac single nuclei RNAseq dataset (**methods**). Red, blue and gray dots represent CM metacells from the scAF AF samples, scAF SR samples and other datasets samples respectively. **I)** Rank plot showing the scores for gene's cell-type and diseases/chamber specificity of the AF signature UP genes (**methods**). **J)** Violin plot of the 6 genes with the highest scores in I split by disease or atrial chamber. Colors indicate the dataset of origin. CM; Cardiomyocytes, EC; Endothelial cells, FB; Fibroblasts, PC; Pericytes. LAA_AF; this study scAF dataset, MI; myocardial infarction dataset, Atrial_H.Atlas; atrial heart atlas nuclei dataset, ACM_DCM; arrhythmogenic and dilated cardiomyopathy dataset, LA; left atria, RA; right atria, AF; atrial fibrillation, SR; sinus rhythm, CTRL_MI; control samples from the MI dataset, FZ; fibrotic zone, BZ; boarder zone, IZ; ischemic zone, RZ; remote zone, DCM; dilated cardiomyopathy, ACM; arrhythmogenic cardiomyopathy, CTRL_A/D; control samples from the ACM_DCM dataset, NCM; non-compaction cardiomyopathy.

As mentioned above, one of the SR samples behaved as an outlier and confounded some of our downstream analyses (**Fig. 4B, S12D** and **G**), therefore, we chose to remove this sample for subsequent TF analyses. To identify AF-related TFs in CM, we correlated motif activities and their TF expression with both AF signatures in CM metacells (**methods**). We find that both the AF signature UP and DOWN capture similar features i.e., correlated motifs with the UP signature were generally anti-correlated with the DOWN signature (**Fig. 4C, S14** and **Table S10**). We chose to focus on the AF signature UP, which contained genes at the *IFNG* locus discussed above. We identified 3 closely related motifs as the strongest signal; from the androgen receptor (AR) and nuclear receptor subfamily 3 group C members 1 & 2 (NR3C1 and NR3C2) (**Fig. 4C, E** and **Table S10**). Among these three, we find that only *AR* expression was correlated with the signature (negative correlation, Pearson R = -0.33, $p = 5x10^{-5}$), as well as with its motif (positive correlation, Pearson R = 0.73, $p = 5 x10^{-27}$) (**Fig. 4D** and **Table S10**). We corroborated these results using TF footprinting analysis, showing higher Tn5 insertions around AR motifs in AF samples compared to SR as well as in CM compared to other cell-types (**Fig. 4F** and **S13**). The decreased AR motif accessibility and increased expression of *AR* in SR samples suggests a repressor activity of AR. Furthermore, among *NR3C1*, *NR3C2* and *AR*, *AR* was the only robust differentially expressed gene (**Table S3**) which remained true in both males and females (**Fig. 4G**). We also find other strong candidates (for which we find a correlation of both TF expression and motif accessibility with the AF signature UP) including NFIC, RORB and MXI1, all anti-correlated with the AF signature UP

and NRF1 which had a similar profile to AR (**Fig. 4B** and **S14**). Among these additional TFs, only *RORB*, and *NFIC* were DEG in both bulk RNAseq datasets (both downregulated in AF, **Table S3**).

5.3.7 AF CM signature is specific across AF co-morbidities

We then examined AF DEG signature's specificity in other publicly available snRNAseq datasets of cardiac diseases with an associated increased risk of AF, namely dilated cardiomyopathies, myocardial infarction and arrhythmogenic right ventricular cardiomyopathy. We also included atrial CMs nuclei from the heart atlas to identify differences in left vs right atrium. Altogether, we scored 32417 CM metacells from 4 datasets [306,312,314] and 117 samples with our two AF signatures (**Fig. S15A**). Scores for the AF signature UP, as opposed to the AF signature DOWN, showed a strong specificity for AF (**Fig. 4H** and **S15B**).

To identify novel potential therapeutic targets, we sought the most cell-type and disease specific genes in the AF signature UP. We ranked each gene based on their cell-type specificity and their AF specificity (**methods**). Beyond the lncRNAs of the *IFNG* locus, we found that *SYNPR* (Synaptoporin), *COLQ* (Collagen Like Tail Subunit Of Asymmetric Acetylcholinesterase), *CHRNE* (Cholinergic Receptor Nicotinic Epsilon Subunit) and *PDE8B* (Phosphodiesterase 8B) stood out as specifically expressed in LAA CM and in AF (**Fig. 4I-J** and **Table S11**).

5.4 DISCUSSION

Contrary to the low number of previously reported replicating AF DEGs, our analysis of large-sample-size bulk RNAseq datasets identified an extensive list of replicating DEGs. Among the top DEGs in this list we found multiple genes with known effects in AF associated with CM functions i.e., *HCN4*[423], *RCAN1*[424], *CALM3*[425] and *RGS6*[426] (**Table S3**). We also identified other strong and replicating AF DEG that likely have important CM functions, such as *ANGPTL2*, *REC114*, *RNF216* and *C4orf54*. Most noticeably, we identified multiple genes at the *IFNG* locus (*LINC01479*, *IFNG-AS1*, *HNRNPA1P70* and *AC007458.1*) which were consistently found as the strongest signal, including in single nuclei (**Fig. 2, S8A** and **S9A**). Furthermore, both *LINC01479* and *IFNG-AS1* were highly CM-specific. Notably, other researchers have identified a similar transcriptional genomic hot spot at the *IFNG* locus in AF patients[174]. While our results confirm its importance, they are in contradiction to their conclusion that *MDM1* (located on the negative strand) is the target gene at this locus. We argue that our analysis contrasting the inclusion and

exclusion of strand information during the read mapping process provides a sensible explanation for this discordance. Little is known about the function of the positive strand genes at this locus in the heart. *LINC01479* is highly enriched in the right atrial appendage, skeletal muscle and tibial artery of GTEx tissues. *IFNG-AS1* is generally associated with immune cell *IFNG* regulation[427] but its function in CM is unknown. More work is needed to decipher the function of this locus specifically in human CM.

For the first time, we characterized the cellular and open chromatin landscape in the human LAA, a tissue that has been broadly used to assess AF gene expression in bulk RNAseq. Multiple deconvolution methods exist to infer cell-type proportions in bulk RNAseq but these methods perform poorly on lowly abundant cell-types and on correlated cell-states[282]. Instead, we opted to partition gene expression into modules based on their LAA bulk RNAseq co-expression and interpret these modules using both our single cell data and gene set libraries. In line with the literature, our results confirm the increased immune cells proportions or activity in AF[428]. Specifically, T-cells appeared to best explain one of the immune gene modules enriched for upregulated AF DEG (blue module). A small module showed strong enrichment in B-cell specific genes, but we did not detect B cells in our scAF dataset. Both T- and B-cells have been reported to be more prevalent in LAA of AF patients compared to SR[429] (albeit B-cell being rarely observed in the heart[306]). Cardiac mesothelial cells are epithelial cells from the epicardium with progenitor properties. These cells can undergo mesothelial-to-mesenchymal transition and become multipotent. This can lead to increased fibroblast proliferation around the epicardium (potentially leading to the formation of re-entrant circuits[420]), a process that may be exacerbated by immune infiltration[419]. Among all datasets analyzed here, the mesothelial cells component explained a large proportion of the variance across samples and appeared increased in AF. We also found two modules composed almost exclusively of downregulated AF DEG that strongly enriched for neuronal cells and adipocytes in our scAF dataset, suggesting that these cell-types may be depleted in AF patients. Notably, AF progression has been associated with an increasing loss of sub-epicardial adipocytes which anti-correlated with sub-epicardial fibrosis[419]. Moreover, in the same study, cytotoxic T-cells appeared as the predominant infiltrating inflammatory cells precluding this remodeling, also in line with our results. An alternative interpretation for these cell-type specific gene modules could also be a shift in cell-state instead of cell abundance. For instance, brown, beige and white adipocytes have all been observed in the heart, exerting different effects on cardiac

function, but current cardiac single cell data has not yet differentiated them, likely due to their low abundance[430].

We found four non-cell-type-specific modules which showed FB state specificity. Among the three FB states that we identified, the iPFB state showed striking similarity to a state described by Reichart et al. in ventricular FB (labeled vFB3)[314]. In agreement with our interpretation that this cell-state is depleted in AF, (as evidenced the downregulation of DEG found in the salmon and magenta modules) these investigators also found a depletion of this FB state in dilated and arrhythmogenic cardiomyopathy patients. They further associated this cell state with genes suppressing fibrosis and facilitating myeloid recruitment. Additionally, multiple genes enriched in our iPFB were also found to be downregulated in scRNAseq of isolated FB from AF patients[303].

One of the key strengths of our study rests in the paired modality of our multiome assay, which enabled us to decipher convoluted TF motifs activity signals by further filtering by their expression levels. Using motif activity and gene expression we identified FOSL2 and JUNB as likely regulators of the iPFB enriched magenta module and CEBPD, KLF3 and JUND for the salmon module. While *Fosl2* overexpression in mice has been associated with increased cardiac fibrosis[431], *JUNB* is associated with suppressed proliferation[432]. Our results suggest an enrichment for their dimer motif (JUNB::FOSL2). Enrichment for this motif has also been reported upon glucocorticoid treatment[433], which is known to reduce FB proliferation[434]. *JUND* is known to induce FB quiescence, in agreement with the higher salmon module score in in RFB. FB cultured with the adipocyte media was found to increase *CEBPD* expression[435]. Furthermore, it was shown that *KLF3* KO fibroblasts more readily differentiate into adipocytes[436]. Therefore, FB with the salmon gene signature could be depleted in AF because of a reduction in adipocyte-FB signaling. More research is needed to decipher the role of these gene programs in cardiac FBs.

We identified AR as likely regulator of our AF signature UP in CM as opposed to NR3C2 (same motif), which have been found to mediate a stressed CM state in myocardial infartion[312]. There is convincing evidence linking the androgen signaling to AF. Low dihydrotestosterone has been reported to increase AF risk in older men[437]. During hormone cancer therapies, androgen deprivation therapy was associated with increased QT interval duration[438]. Also, Ar knock out (KO) mice were shown to have impaired Ca^{2+} homeostasis[439]. Moreover, AR has been shown to act as repressor under multiple conditions, in agreement with our results. However, the effect of

testosterone replacement therapy in clinical trials has shown mixed outcomes for AF[440,441]. Together, this supports our results suggesting that AR acts as repressor on upregulated AF DEG but the role of androgen therapy for AF treatment needs further investigation.

Lastly, we propose interesting AF gene targets that show high disease and cell-type specificity. Phosphodiesterases hydrolyze the cyclic secondary messenger cAMP and cGMP, directly impacting a host of CM functions such as contractility, stress response or gene transcription[442]. Interestingly, *PDE8B* has recently been shown to alter L-type calcium current of AF patients[443], supporting our candidate gene selection approach. Interestingly, CM have been shown to have age-dependent intrinsic acetylcholine (ACh) synthesis, storage and transport properties[444], which, when altered, impact CM size[445]. *COLQ* is one of the most replicated DEG in AF[176]. It is known to anchor acetylcholinesterase (AChE, which hydrolyzes ACh) at neuromuscular junctions[446]. Mutations in this gene have been associated with AChE deficiency. *CHRNE* encodes an ACh receptor (AChR) subunit also found at neuromuscular junctions[446]. *CHRNE* was found to be enriched in the atria in 3 studies[447-449]. Mutations in this gene have been associated with congenital myasthenic syndrome, causing post-synaptic Ca^{2+} overload[450]. Together, both *COLQ* and *CHRNE* increased expression would lead to increased ACh signaling, the former through increased extracellular AChE anchoring, the latter through increased AChR formation. Finally, *SYNPR* encodes a synaptic vesicular membrane component and has been suggestively associated to left-sided cardiac malformation through a GWAS *SYNPR* intronic variant[451]. It was also one of the most atrial CM specific genes in the heart atlas[306]. Its role in the heart remains to be determined.

5.4.1 Limitations

While most modules could be confidently attributed to a cell-type based on concordant pathway and single cell enrichment scores, the pathways of modules that we assessed in FB were not specific to this cell-type. Therefore, other cell-types or cell-states which we did not detect in our scAF dataset, may also explain the co-regulation of genes observed in bulk for these modules. For instance, gene sets enriched in the salmon module suggests an enrichment for neutrophils which are rarely detected in cardiac snRNAseq[306]. Another limitation of our study is that our scAF dataset had a relatively low number of samples and did not allow for robust differential expression analysis in less abundant cell-types. In the futures, larger single cell datasets on AF patients could be especially informative to identify DEGs and cell-state transitions from rare cell-types such as

neuronal cells and adipocytes. Additionally, our selection of TFs is partly based on their transcriptional level, which does not account for post-translational regulation mechanisms. Therefore, motifs enrichments due to translocation event can be missed. Lastly, a possible epicardial sampling bias cannot be ruled out as the cause of the observed increased mesothelial gene expression in AF despite the mechanistic plausibility.

5.5 CONCLUSION

In this study we profile for the first-time single nuclei of the LAA using paired ATAC and RNA modalities. We identify strong TFs candidates for cell-type identity based on their combined specific activity and expression. Moreover, we establish a robust list of AF DEGs found in the two largest human LAA bulk RNAseq datasets to date. We report that non-coding genes at the *IFNG* locus consistently show the strongest signals in bulk and CM nuclei RNAseq of LAA from AF patients. Additionally, we identify cell-type specific DEG modules suggesting a loss of rare cardiac cell-types, such as neurons and adipocytes, and a gain of inflammatory cells, such as T- and B-cells. We further identify likely TF regulators of FB gene modules downregulated in AF. Finally, in a CM centric analysis, we identify a highly AF specific gene signature for which AR is the most likely regulator and suggest novel CM and AF-specific target DEG. Our results provide a valuable resource to orient future research aiming at deciphering the effect of AF DEG in specific in vitro models and cell-type specific KO mice.

5.6 METHODS

Data analyses were done in R version 4.2.2.

5.6.1 Multiome sample preparation, raw data processing and pre-processing steps

These steps are detailed in the methods section of chapter 4.

5.6.2 ATACseq peak comparison with ENCODE and human enhancer atlas

We compared the scAF peaks against other annotations using the *findOverlaps()* function from the GenomicRanges package. We labeled peaks based on any overlap with other ranges. To attribute cell-type identity to our peaks, we used the output of Signac's *CallPeaks()* function from the cell-type specific peak calling step. We retrieved ENCODE hg38 cCREs (track named encodeCcreCombined) from the UCSC genome browser (https://genome.ucsc.edu).

To validate our cell-type specific peaks we filtered peaks that were uniquely called in one cell-type. We retrieved cCREs from the human enhancer atlas[285] and then created Granges from files at downloaded from http://catlas.org/catlas_downloads/humantissues/cCRE_by_cell_type/. We then calculated the percentage of scAF cell-type specific peaks overlapping cCREs from each human enhancer atlas cell-type. We truncated the matrix for overlaps below 25% for clarity.

5.6.3 TF activity scores and selection of cell-type specific TF

To select cell-type specific TFs, we ranked them based on both motif activity and gene expression. First, we calculated TF motif scores in each nucleus using chromVAR[294] (implemented in Signac[289] with the function *RunChromVAR()*) and the JASPAR2020 database[452]. Then, for each cell-type we calculated the area under the receiver operating characteristic curve (AUC) using the presto package function *wilcoxauc.Seurat()* for both gene expression and motif activity. Finally, we ranked each TF based on the product of its gene AUC and motif AUC.

5.6.4 CTSN bulk RNAseq sample preparation, sequencing and raw data processing

Please refer to the methods section of chapter 4.

5.6.5 Comparison of bulk RNAseq DEG

We downloaded the GSE69890 (J. Hsu[216]) raw count data from GEO and analyzed it using the same DESeq2[166] steps used for our CTSN dataset. We first compared studies using PCAs for the 500 most variable genes and the contribution of sex. We used variance stabilizing transformation (vst; DESeq2 function *vst()*) values. We further assessed which cell-type was likely most contributing to the top PC1 genes by using the package factoextra to retrieve genes with the

highest contribution to PC1 and then looked for their expression levels in each cell-types of our scAF dataset.

Differential expression analysis was done using the *DESeq()* function with sex as covariate followed by log fold change shrinkage using *lfcShrink()*. For the J. Hsu dataset, we used AF patients in AF rhythm against no-AF patients[173]. We used an FDR of 0.05 as significance threshold in both datasets and compared the overlapping upregulated and downregulated DEG in both sets as well as their signed log10(FDR).

To assess the effect of the strand, we did a second pseudocount alignment using kallisto[156] omitting the **`--rf-stranded`** flag. We labeled genes that lost their significance when strand information was provided as probable false positives in the J. Hsu unstranded data.

5.6.6 Single nuclei differential expression analysis

We conducted differential expression analysis in the scAF dataset using pseudobulk. We aggregated counts for each cell-type and sample using Seurat's[266] *AggregateExpression()* function, keeping only genes found in more than 5% of nuclei. The same DESeq model used in bulk was used here to the exception of the shrinkage model used which was *ashr*[453] instead of the default. We then compared robust bulk DEGs to each cell-type DEGs.

5.6.7 Cell-type transcriptional activity comparison

To infer transcriptional activity, we created a ratio of UMI fraction per cell-type (the sum of all counts for that cell-type divided by all counts of the gene by nuclei matrix) divided by its cell-type contribution to all nuclei in the dataset (cell-type nuclei count / total nuclei count).

5.6.8 WCGNA gene modules analysis

We created gene modules using the combination of both bulk RNAseq dataset (CTSN and J. Hsu) and the package WGCNA[454]. We ran a joint differential expression analysis using both datasets with the same parameters described in the *Comparison of bulk RNAseq DEG* section with further addition of the dataset as covariate. We then filtered out genes with low expression and small variability across conditions with the following filters: base mean expression >1, |log2 fold change| > 0.05 and p-value < 0.05. In total, we used 7,970 genes and 323 samples for module analysis. We used the vst expression values corrected for sex and dataset using the limma function

removeBatchEffect() to avoid creating modules that capture the effect of these variables. We created modules using the WGCNA *blockwiseModules()* function with the following parameters: power = 9, minModuleSize = 50, reassignThreshold = 0, mergeCutHeight = 0.25, networkType = "signed" and the rest as default. For each module, we ran pathway analyses in the GO_Biological_Process_2021 and PanglaoDB_Augmented_2021 libraries using the enrichR[455] package on genes with an FDR < 0.05. We used the same sets of genes within each modules to score each nucleus with Seurat's function *AddModuleScore()*.

5.6.9 Sub-clustering analyses

For both FB and CM sub-clustering we used 10 PCs from the RNA modality. Otherwise, we used the same standard Seurat process that was used for the whole dataset i.e., SCTransform, Harmony, FindNeighbors and FindClusters.

5.6.10 TF motif and expression correlations

To select likely regulatory TFs for specific gene signatures we addressed data sparsity by creating metacells within cell-types of interest (FB and CM). This was accomplished similarly as described previously (see methods of chapter 4). Briefly, we used the hdWGCNA[456] package *MetacellsByGroups()* function. We called metacells within cell-types and samples using the RNA harmony reduction to aggregate 30 neighbor nuclei with a maximum of 10 overlapping nuclei per metacell. Both RNA and ATAC counts were aggregated using the same neighbors. We then correlated each TF expression and motif activities (calculated using ChromVAR on metacells as mentioned in the *TF activity scores and selection of cell-type specific TF* method section) with the gene signature scores (calculated with Seurat's function *AddModuleScore()*) in metacells using the psych *corr.test()* function with FDR adjustment. To validate TF motifs, we used Signac's *Footprint()* function on the scAF dataset.

5.6.11 AF CM signature specificity across AF co-morbidities

To create a strong gene signature specific to AF and CM, we selected genes based on bulk differential expression and CM specificity. We calculated gene specificity using the presto package function *wilcoxauc.Seurat()* and filtered genes with an AUC > 0.5 and FDR < 0.05. For bulk RNAseq, we used an FDR < 0.05 and |log2 fold change| > 0.25 in both CTSN and J. Hsu

independent DEG. The overlap of these 3 sets was used to create the upregulated and downregulated AF signatures.

We compared gene signature scores in CM from 4 datasets. We downloaded two ventricular cardiomyocytes datasets[312,314] from the cellxgene portal (https://cellxgene.cziscience.com) and atrial CM from the heart atlas[306] at (https://www.heartcellatlas.org). We used the same parameters to create metacells described in the previous method section and scored cells for the AF CM signatures using *AddModuleScore()*.

5.6.12 Robust AF CM target genes selection

We used gene expression from AF co-morbidities CM metacells and the scAF cell-types to try to identify therapeutic gene targets with the highest specificity. Thus, we used the presto package to calculate AUCs across all scAF cell-types and metacell groups included in the 4 single nuclei datasets:

- scAF; LAA AF, LAA SR
- heart atlas; left atrium, right atrium
- Kuppe et al. (CM from various myocardial infarction ventricular zones); control, fibrotic, ischemic, boarder, remote
- and Reichart et al. (dilated and arrhythmogenic cardiomyopathy ventricular CM); control, pathogenic variant negative, pathogenic variant positive

We then used the product of these AUCs to rank genes from the AF signature UP

5.7 DECLARATIONS

5.7.1 Acknowledgments

We thank all participants who contributed bio samples to this study.

5.7.2 Data availability

The data availability is disclosed in chapter 4. All code to analyze the data and reproduce the results of this manuscript will be made available on GitHub upon publication (https://github.com/lebf3/scAF).

5.7.3 Funding

This work was funded by the Fonds de Recherche en Santé du Québec (FRQS), the Canada Research Chair Program and the Montreal Heart Institute Foundation (to S.N. and G.L.). S.R was funded by the British Heart Foundation Intermediate and Senior Fellowships, the British Research Council (BRC4) NIHR (Oxford) grant, the Wellcome Trust Institutional Strategic Individual Career Support grant and the John Fell Foundation Fund (Oxford). This research was enabled in part by support provided by Calcul Quebec (https://www.calculquebec.ca/en/) and Compute Canada (www.computecanada.ca). We thank Génome Québec for performing next-generation DNA sequencing for this project.

5.7.4 Competing interests

The authors declare that they have no competing interests.

5.7.5 Author contributions

Conceived and designed the analyses: F.J.A.L. and G.L.; Collected the data: F.J.A.L. and N.M.; Contributed data: F.J.A.L., N.M., S.R. and S.N.; Performed analyses: F.J.A.L.; Secured funding and supervised the work: S.R., S.N. and G.L.; Wrote the manuscript: F.J.A.L.

5.8 Supplementary materials

All supplementary tables are included in the attached zipper folder.

A

sample	condition	sex	age	Estimated number of cells	Linked genes	GEX Sequenced read pairs	GEX Median genes per cell
CF69	1	F	51	3497	10865	800842000	3406.0
CF77	1	M	78	3129	7937	641025981	2386.0
CF89	1	F	69	1960	5013	608280552	1956.0
CF91	1	M	57	2587	10700	867534711	2113.0
CF93	0	M	80	3117	5736	551655924	2806.0
CF94	0	F	81	1783	1384	833761813	1664.0
CF97	0	M	58	3572	10985	640489870	2341.5
CF102	0	M	75	2511	4567	791705278	3717.0

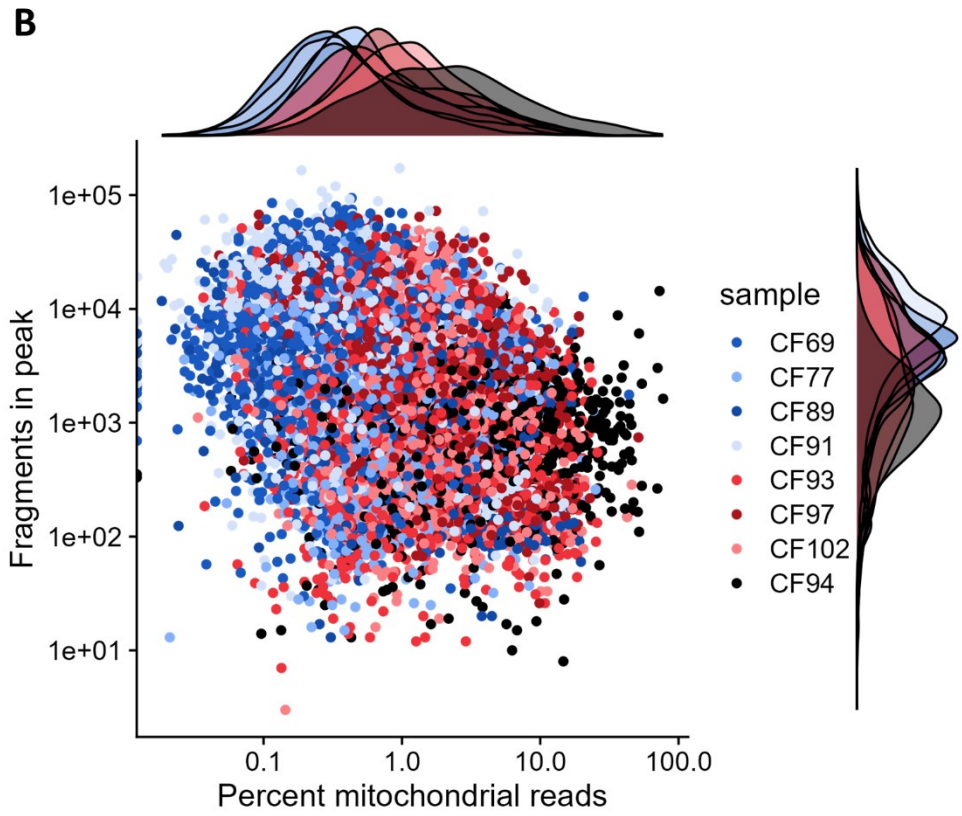


Figure S1. Sample quality control.

A) Sample metric output from CellRanger.

B) Scatter plot showing the percentage of mitochondrial read count and number of fragments in peak. Red and blue dots are from atrial fibrillation and sinus rhythm samples respectively. We highlight in black the sample that was removed based on the poor-quality control metrics shown in **A** and **B**.

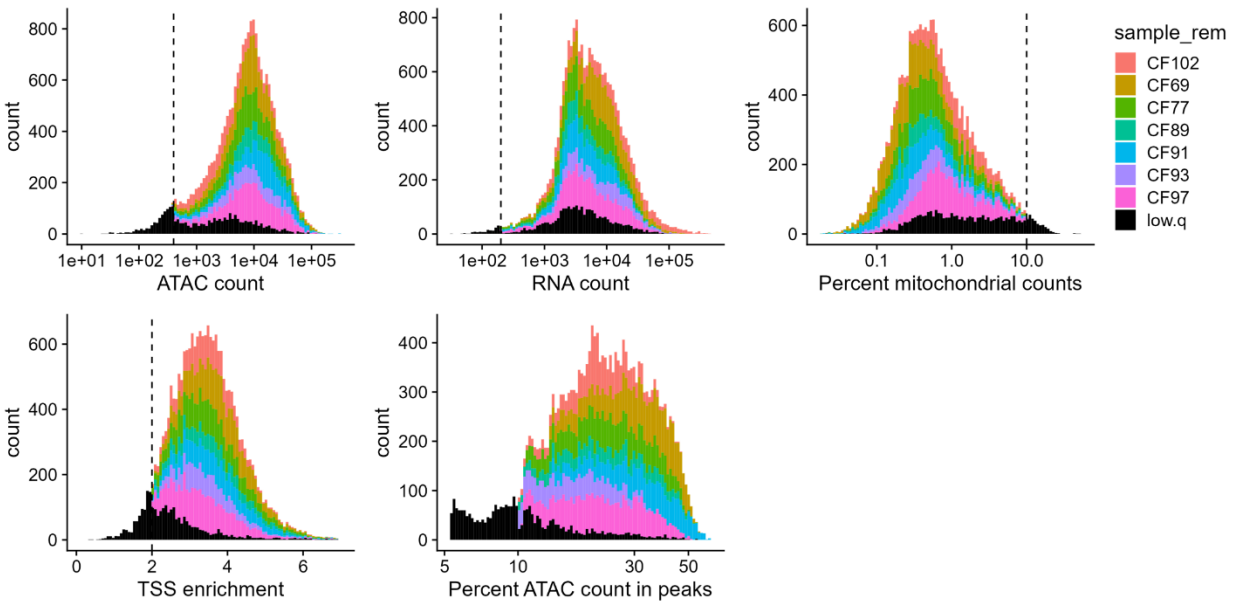


Figure S2. Nuclei quality control.

Histograms showing the thresholds used (dashed vertical lines) for initial filtering of low-quality nuclei. Nuclei in black are the aggregate of cells from all samples that do not pass one of the filters. sample_rem; nuclei grouped by sample or binned in the low-quality group, TSS; transcription starting site.

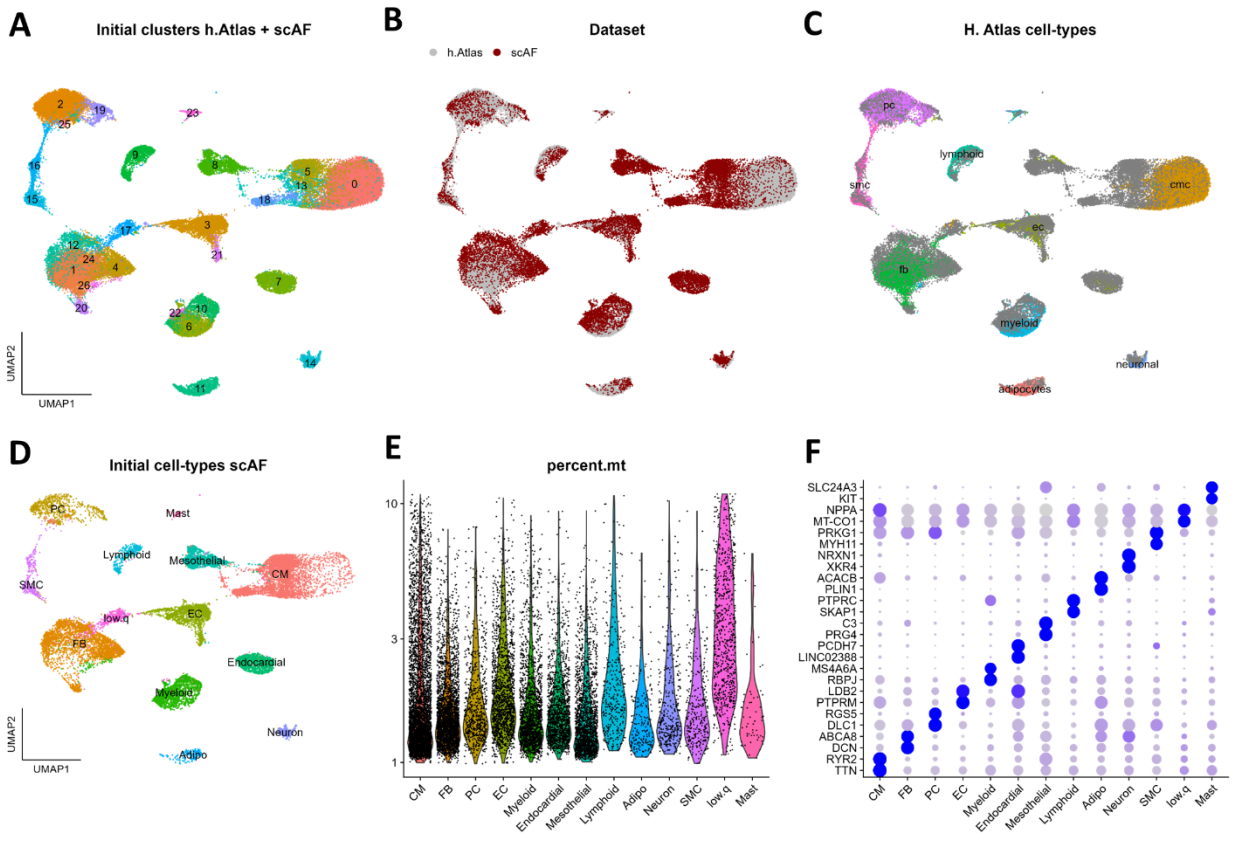


Figure S3. Doublet calling and cell-type annotation

A-D) Single nuclei uniform manifold approximation and projection (UMAP) of the integrated nuclei from the heart atlas left atria and our scAF dataset colored by; **A**) Seurat clusters, **B**) dataset, **C**) cell-type labels from the heart atlas and **D**) attributed cell-type labels to the combined datasets.

E) Violin plot showing the percentage of mitochondrial counts in each cell-type of the scAF dataset.

F) Dot plot showing the top 2 marker genes for each cell-type in the scAF dataset.

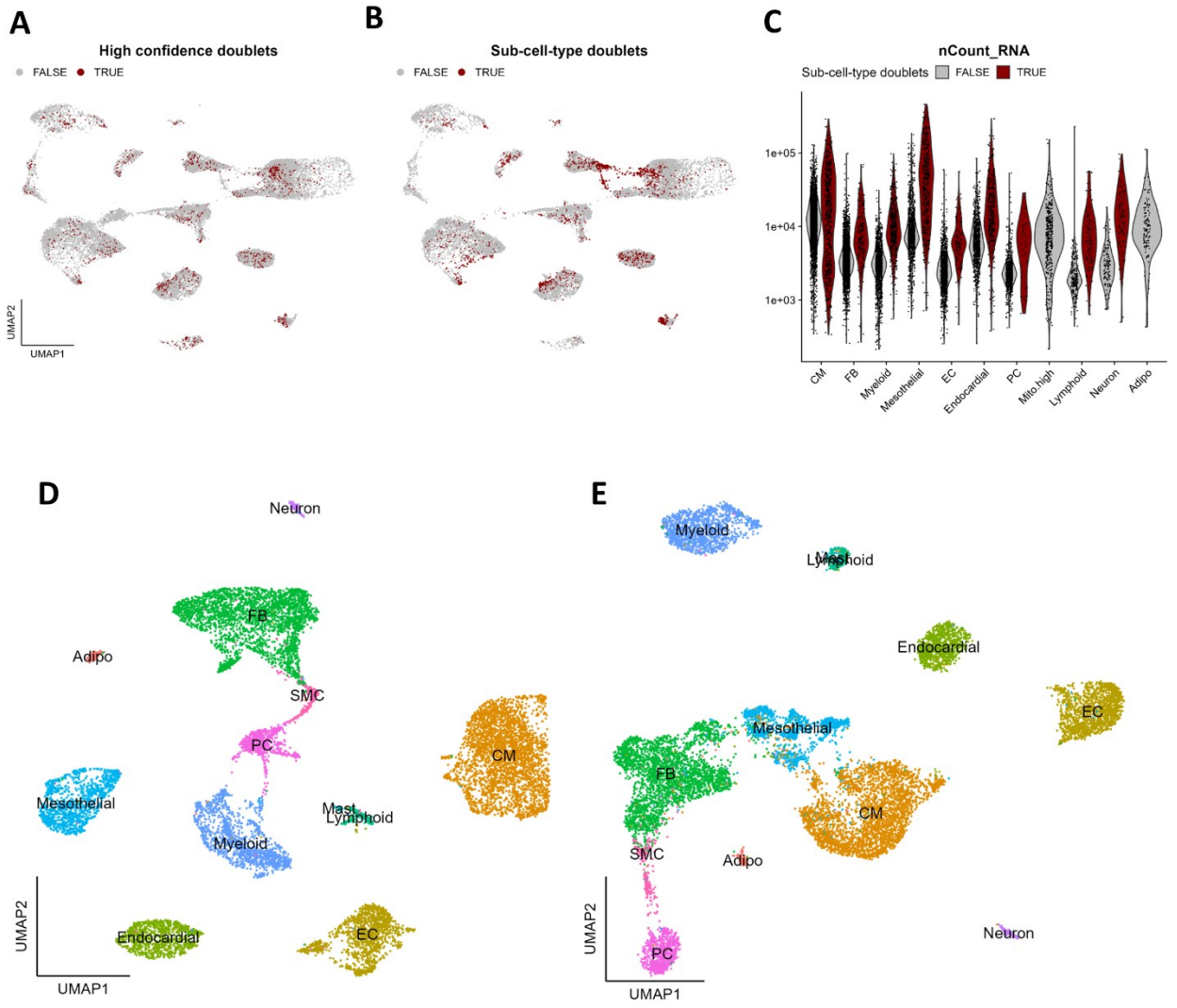


Figure S4. Final clustering and manual doublet curation

A-B) Single nuclei multiome dataset uniform manifold approximation and projection (UMAP) colored by; **A**) doublet labels attributed by scDblFinder and **B**) additional doublets manually attributed based on cell-type markers and scDblFinder scores during sub-clustering analyses.

C) Violin plot showing the total RNA read counts per nucleus in each cell-type for doublets and singlets labeled in **B**.

D-E) Single nuclei UMAP colored by cell-type using the **D**) RNA matrix and **E**) ATAC matrix.

Adipo; Adipocytes, CM; Cardiomyocytes, EC; Endothelial cells, FB; Fibroblasts, PC; Pericytes, SMC; Smooth muscle cells.

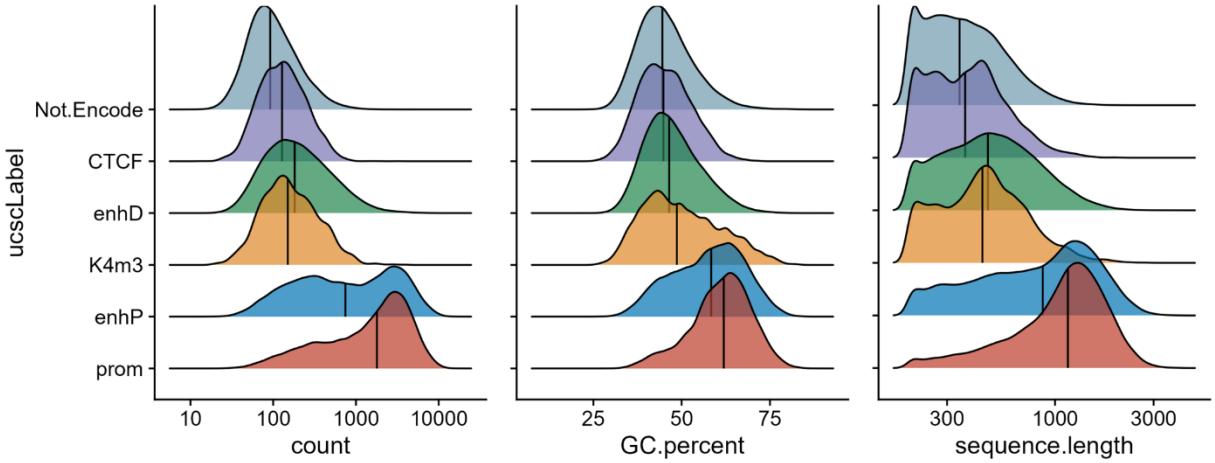


Figure S5. scAF Peak characteristics

Density plots of peak types based on their overlap with ENCODE candidate *cis*-regulatory elements (cCREs). From left to right, we show the distribution for each type of; 1-number of fragments per peak, 2-the percentage of GC per peak and 3-the length per peak.

Prom; promoter, enhP; proximal enhancer, K4m3; lysine 4 tri-methyl mark, enhD; distal enhancer, CTCF; CCCTC-binding factor mark, not.Encode; peak found in the scAF dataset without any overlapping ENCODE peak.

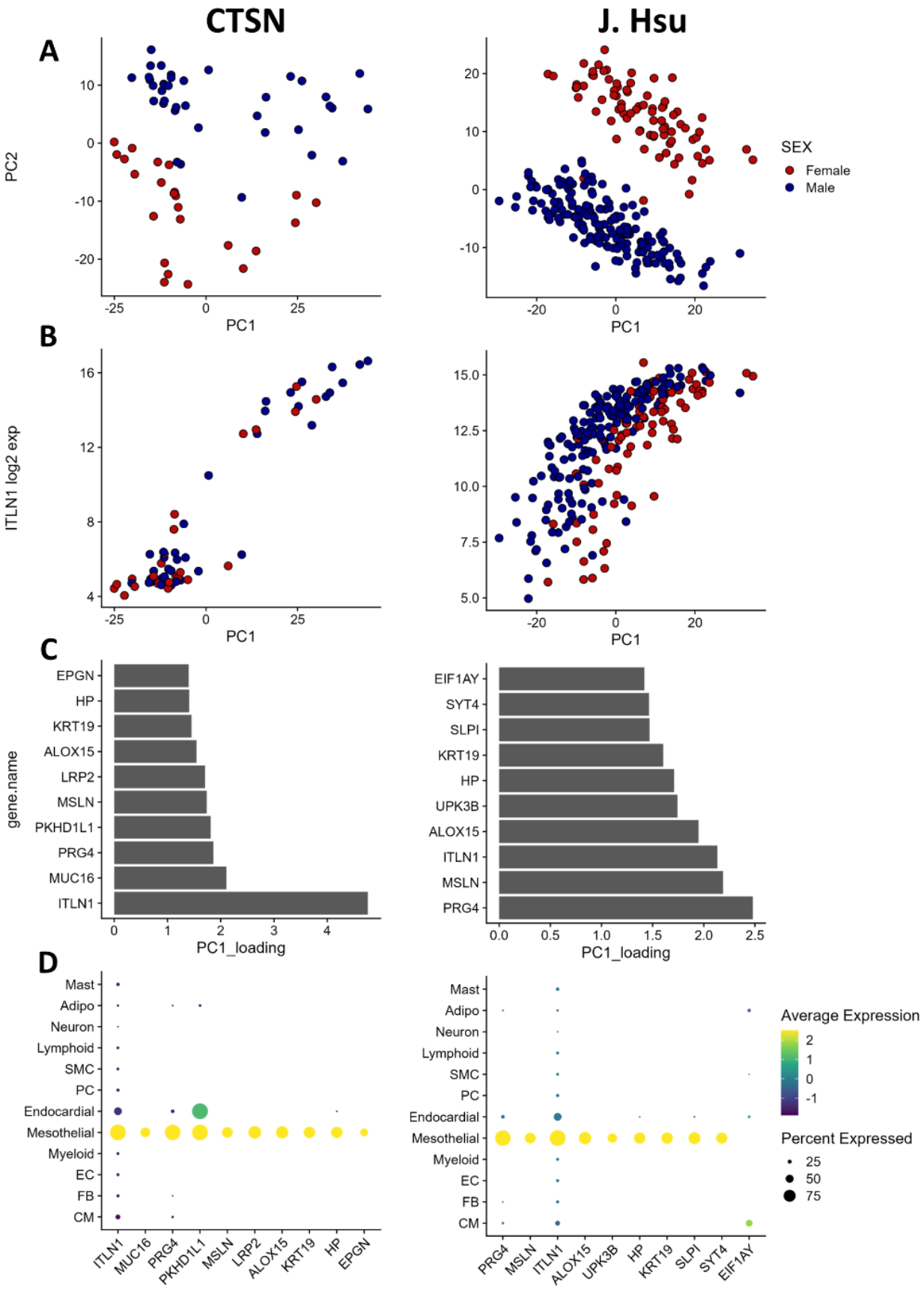

CTSN
J. Hsu
A
B
C
D
PC2
PC1
ITLN1 log2 exp
gene.name
PC1_loading
SEX
Female
Male
Average Expression
Percent Expressed
EPGN
HP
KRT19
ALOX15
LRP2
MSLN
PKHD1L1
PRG4
MUC16
ITLN1
EIF1AY
SYT4
SLPI
UPK3B
Mast
Adipo
Neuron
Lymphoid
SMC
PC
Endocardial
Mesothelial
Myeloid
EC
FB
CM

Figure S6. Bulk RNAseq QC

Left and right panels show results from left atrial appendages bulk RNAseq datasets from the CTSN cohort (this study) and J. Hsu respectively.

A) Principal component analysis (PCA) of the 500 most variable genes.

B) Scatter plot showing the log2 transformed expression of the epicardium marker *ITLN1* against the first principal component's (PC1) coordinates. A-B) Colors show the sex for each sample.

C) Top 10 loadings for PC1 in **A** and **B** with their **D**) levels of normalized expression by cell-type in single nuclei RNAseq from left atrial appendages.

Adipo; Adipocytes, CM; Cardiomyocytes, EC; Endothelial cells, FB; Fibroblasts, PC; Pericytes, SMC; Smooth muscle cells.

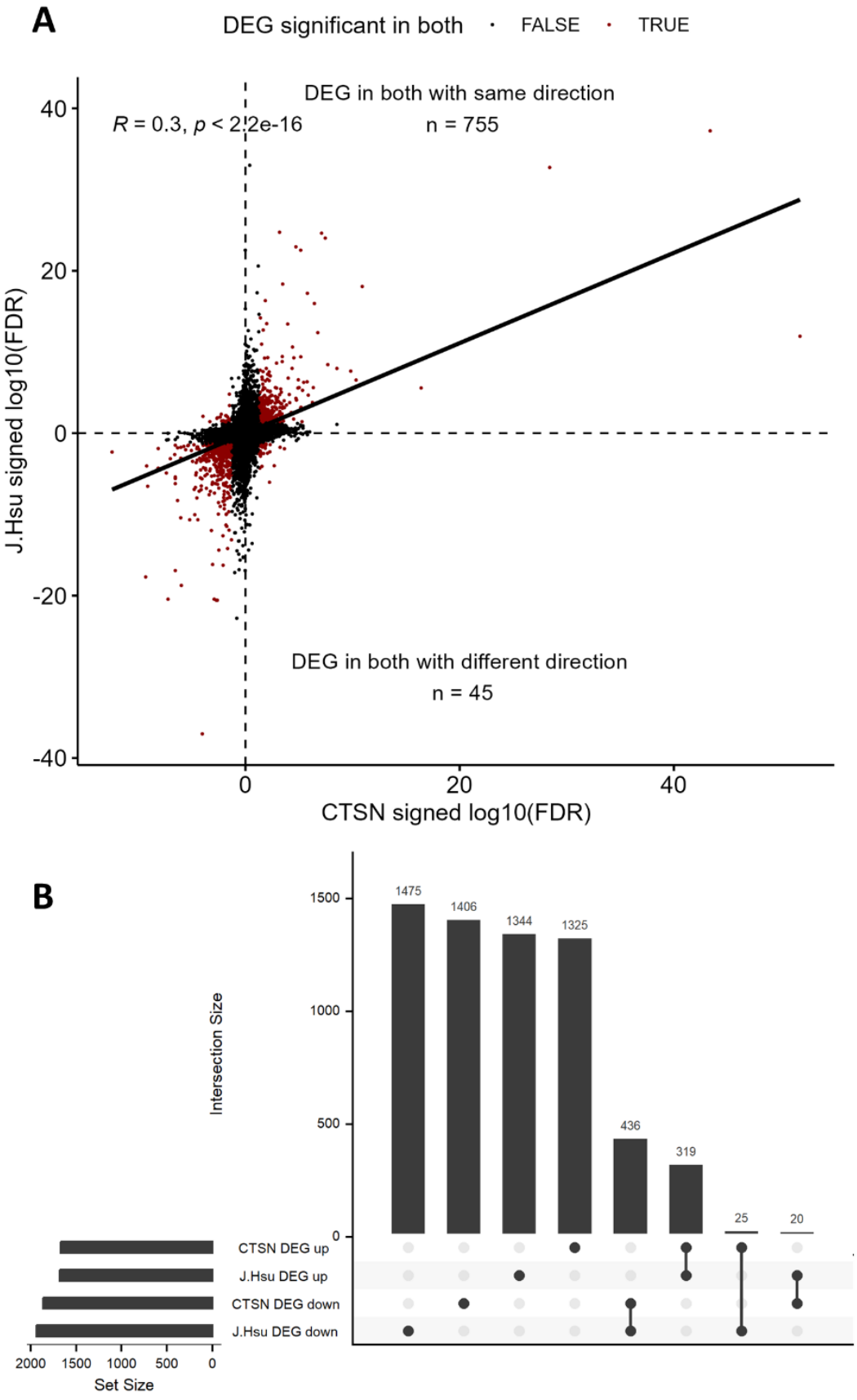

A
DEG significant in both
FALSE
TRUE
DEG in both with same direction
R = 0.3, p < 2.2e-16
n = 755
DEG in both with different direction
n = 45
J.Hsu signed log10(FDR)
CTSN signed log10(FDR)
40
20
0
-20
-40
B
Intersection Size
1475
1406
1344
1325
436
319
25
20
1500
1000
500
0
CTSN DEG up
J.Hsu DEG up
CTSN DEG down
J.Hsu DEG down
2000
1500
1000
500
0
Set Size

Figure S7. DEG overlap across bulk studies

A) Scatter plot comparing the bulk RNAseq differential expression results between atrial fibrillation and sinus rhythm samples, using signed log10(false discovery rates [FDR]), in the CTSN and J. Hsu dataset. The sign was attributed based on the log2 fold change, with a positive indicating upregulation in atrial fibrillation. Red dots show genes qualified as differentially expressed genes (DEGs), with an FDR < 0.05, in both datasets.

B) Upset plot showing the number genes in each set of intersecting and non-intersecting DEGs in each dataset and direction of effect. Up and down indicate upregulated and downregulated genes in atrial fibrillation.

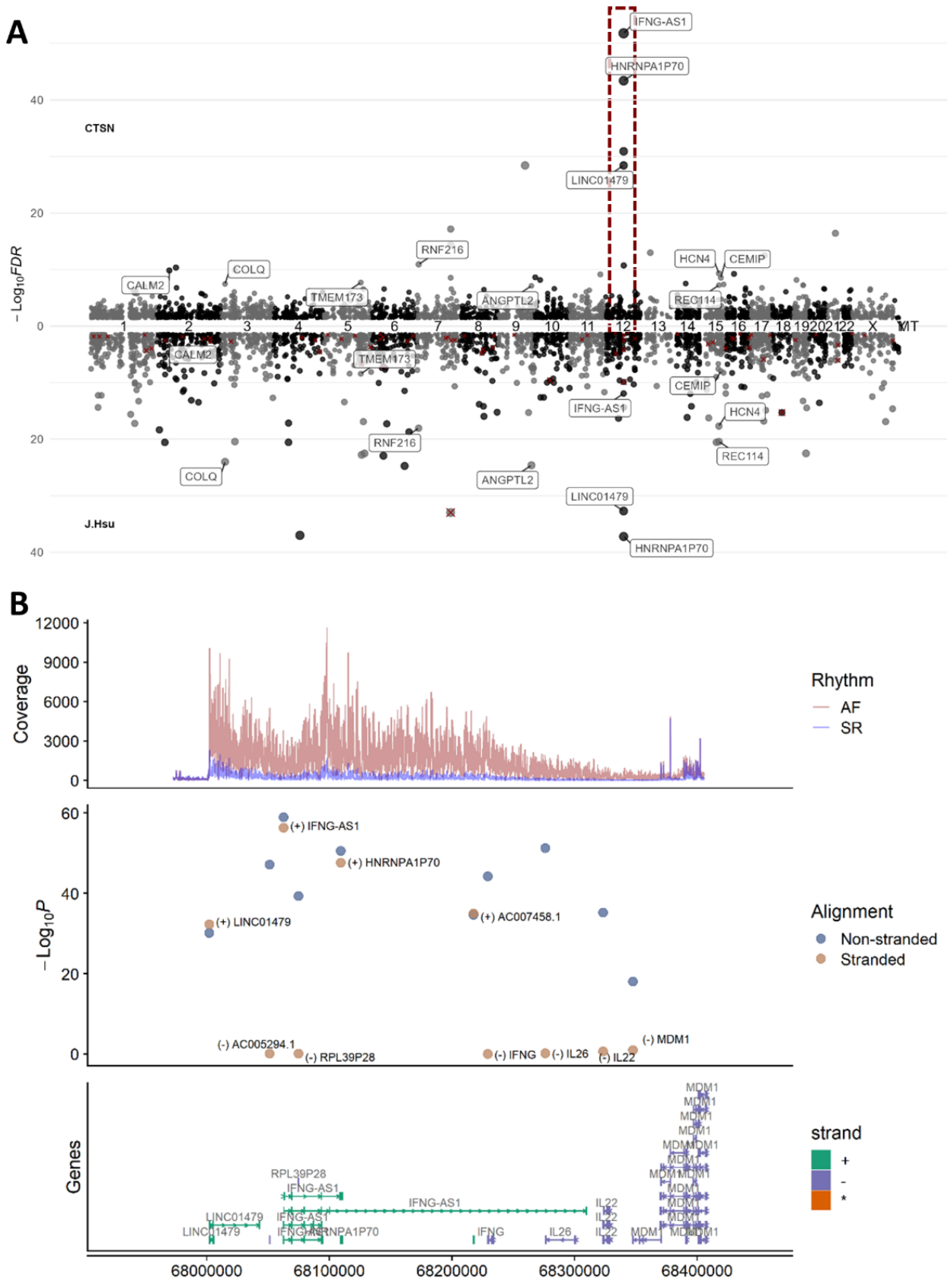

A
IFNG-AS1
HNRNPA1P70
CTSN
LINC01479
RNF216
HCN4
CEMIP
COLQ
CALM2
TMEM173
ANGPTL2
REC114
$-\mathrm{Log}_{10}FDR$
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X MT
J.Hsu
B
Coverage
Rhythm
AF
SR
$-\mathrm{Log}_{10}P$
(+) IFNG-AS1
(+) HNRNPA1P70
(+) LINC01479
(+) AC007458.1
(-) AC005294.1
(-) RPL39P28
(-) IFNG
(-) IL26
(-) IL22
(-) MDM1
Alignment
Non-stranded
Stranded
Genes
strand
+
-
*
68000000
68100000
68200000
68300000
68400000

Figure S8. Effect of strand specific alignment

A) Miami plot comparing the bulk RNAseq differential expression between atrial fibrillation and sinus rhythm samples -log10(false discovery rates [FDR]) in the CTSN and J. Hsu dataset. Colors and numbers on the x-axis show the chromosomal position of each gene. Red crossed genes denote genes that were only differentially expressed when omitting strand information in the CTSN dataset (method). The dashed red box highlights the *IFNG* locus genes shown in B.

B) Gene expression profiles at the *IFNG* locus in the CTSN cohort. (top) Coverage plot of all atrial fibrillation and sinus rhythm samples combined. (center) -log10(false discovery rates [FDR]) with (orange) and without (blue) provision of the strand information during read alignment (method). Sings in parentheses show the strand for each gene. (bottom) Transcript annotation at this locus colored by strand.

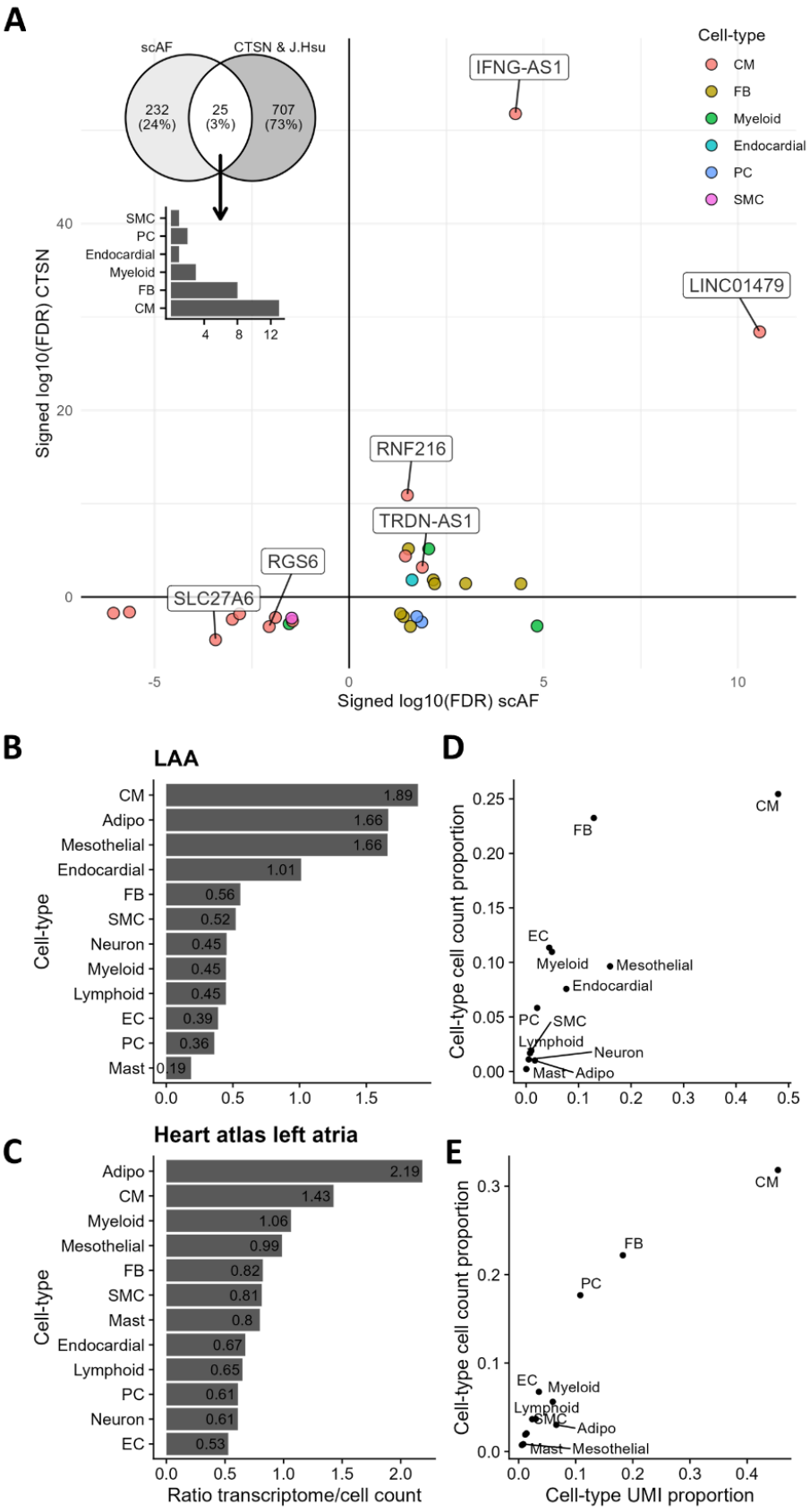

A
scAF
CTSN & J.Hsu
232 (24%)
25 (3%)
707 (73%)
SMC
PC
Endocardial
Myeloid
FB
CM
4
8
12
Cell-type
CM
FB
Myeloid
Endocardial
PC
SMC
IFNG-AS1
LINC01479
RNF216
TRDN-AS1
RGS6
SLC27A6
Signed log10(FDR) CTSN
Signed log10(FDR) scAF
40
20
0
-5
0
5
10
B
LAA
CM 1.89
Adipo 1.66
Mesothelial 1.66
Endocardial 1.01
FB 0.56
SMC 0.52
Neuron 0.45
Myeloid 0.45
Lymphoid 0.45
EC 0.39
PC 0.36
Mast 0.19
Cell-type
0.0
0.5
1.0
1.5
C
Heart atlas left atria
Adipo 2.19
CM 1.43
Myeloid 1.06
Mesothelial 0.99
FB 0.82
SMC 0.81
Mast 0.8
Endocardial 0.67
Lymphoid 0.65
PC 0.61
Neuron 0.61
EC 0.53
Cell-type
0.0
0.5
1.0
1.5
2.0
Ratio transcriptome/cell count
D
Cell-type cell count proportion
0.25
0.20
0.15
0.10
0.05
0.00
CM
FB
EC
Myeloid
Mesothelial
Endocardial
PC
SMC
Lymphoid
Neuron
Mast
Adipo
0.0
0.1
0.2
0.3
0.4
0.5
E
Cell-type cell count proportion
0.3
0.2
0.1
0.0
CM
FB
PC
EC
Myeloid
Lymphoid
SMC
Adipo
Mast
Mesothelial
0.0
0.1
0.2
0.3
0.4
Cell-type UMI proportion

Figure S9. Enhanced reproducibility of CM DEG

A) Scatter plot comparing differential expression between atrial fibrillation and sinus rhythm samples in the CTSN and scAF datasets. Only genes that are differentially expressed in both bulk RNAseq dataset and in the scAF datasets are shown. The sign of the log10(false discovery rates [FDR]) is attributed based on the log2 fold change, with a positive value indicating upregulation in atrial fibrillation. Colors show the cell-type in which the gene was found to be differentially expressed in the scAF datasets. The inset Venn diagram shows the number of differentially expressed genes found in both the CTSN and J. Hsu dataset and in the scAF dataset (scAF) with the number of intersecting genes (25). The bar plot below shows the number of differentially expressed genes found in this intersection by cell-types.

B-C) Bar plot showing the ratio of fraction of total RNA unique molecular identifiers (UMIs) divided by the fraction of total number of nuclei by cell-type in **B**) the scAF dataset and **C**) the heart atlas left atrial nuclei. The ratios above 1 indicate that the proportion of all RNA counts explained by this cell-type is higher than its contributing proportion of nuclei in the whole dataset.

D-E) Scatter plot showing the fraction of total RNA UMIs and the fraction of total number of nuclei by cell-types.

CM; Cardiomyocytes, FB; Fibroblasts, PC; Pericytes, SMC; Smooth muscle cells.

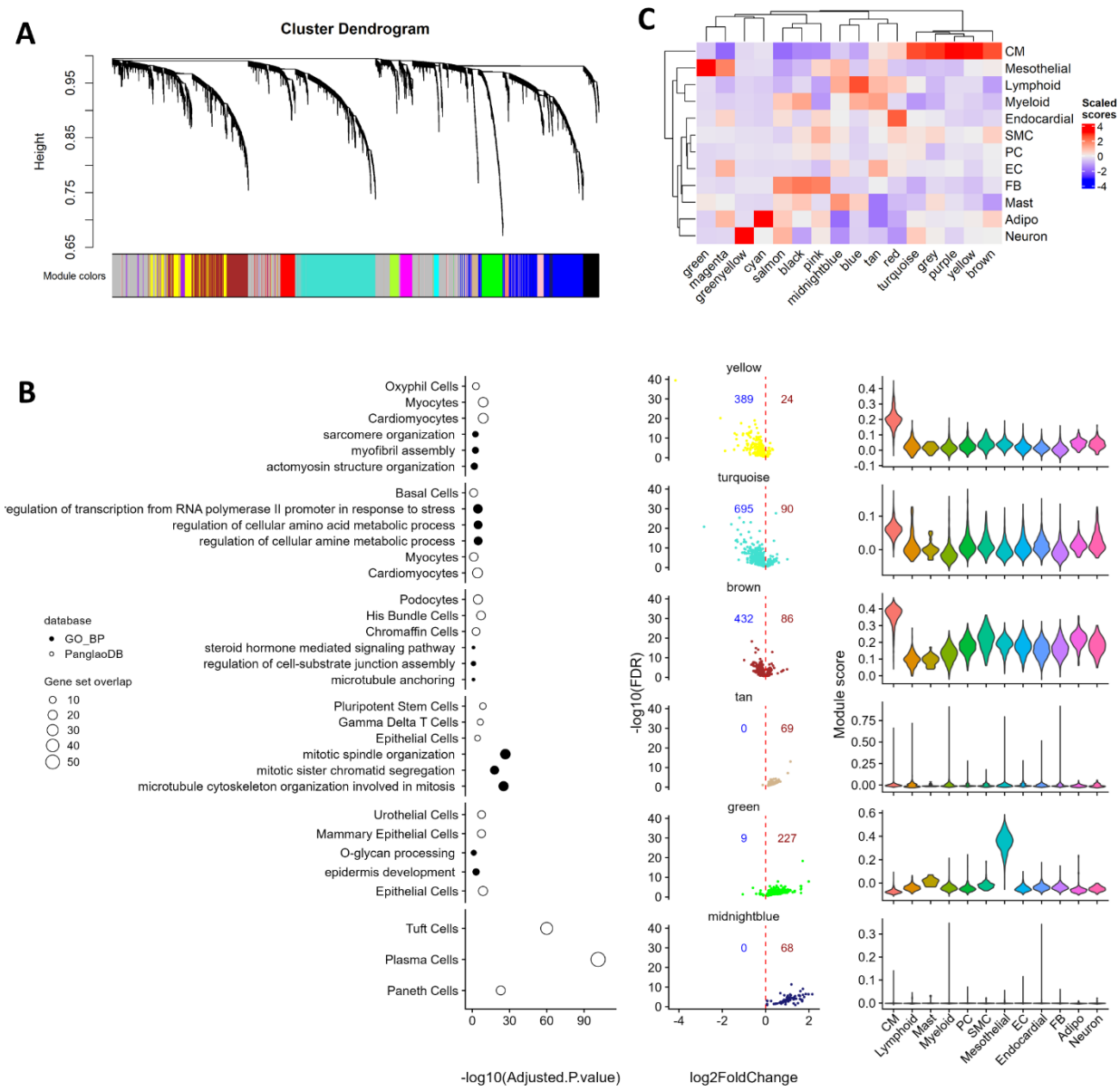


Figure S10. WGCNA gene modules

A) Dendrogram of gene x gene similarity based on their expression in left atrial appendage bulk RNAseq datasets (**methods**). Each gene's module attribution by WGCNA is labeled by colors at the bottom.

B) Remaining modules not included in figure 3 for which at least one gene set was significant (adjusted p-value < 0.1, none were found for the red and purple modules). (**left**) Dot plots showing the top 3 gene sets from a gene set overrepresentation analysis of the DEGs in each module (**methods**). For this analysis we used the two gene set libraries PanglaoDB and gene ontology biological process (GO BP). (**center**) Volcano plots showing the log2 fold change and log10(false discovery rate [FDR]) statistics from the DEG analysis, stratified by modules. Red and blue integers indicate the number of AF upregulated and downregulated genes respectively in each module. (**right**) Violin plot showing the module scores (**methods**) in each cell-type of the scAF dataset.

C) Heatmap showing the mean module score by cell-type shown in **B**, scaled by module. The margin dendrograms show the similarity of each module and cell-type.

Adipo; Adipocytes, CM; Cardiomyocytes, EC; Endothelial cells, FB; Fibroblasts, PC; Pericytes, SMC; Smooth muscle cells.

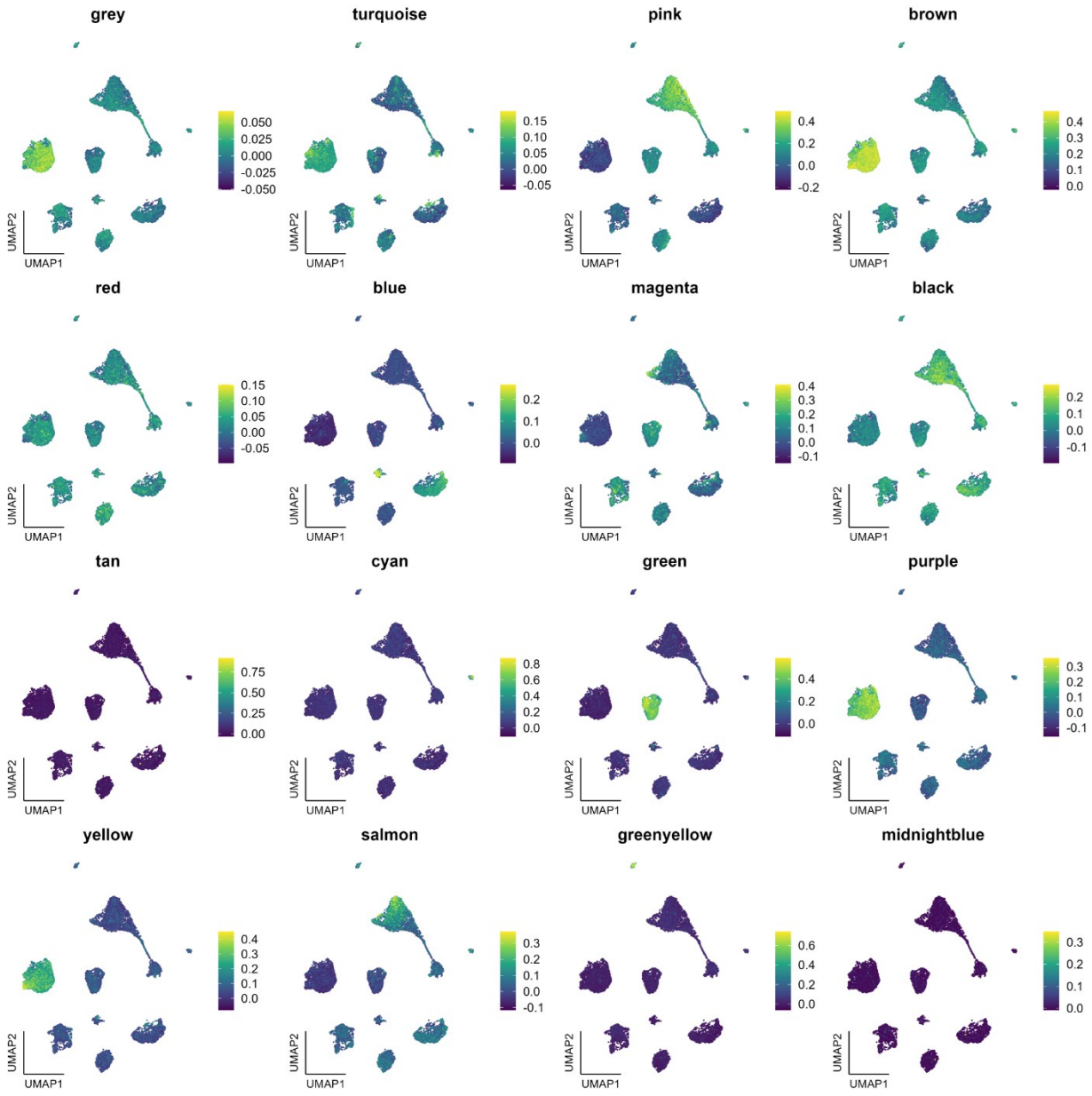


Figure S11. Single cell enrichment scores of WGCNA gene modules

Single nuclei uniform manifold approximation and projection (UMAP) colored by module scores (**methods**).

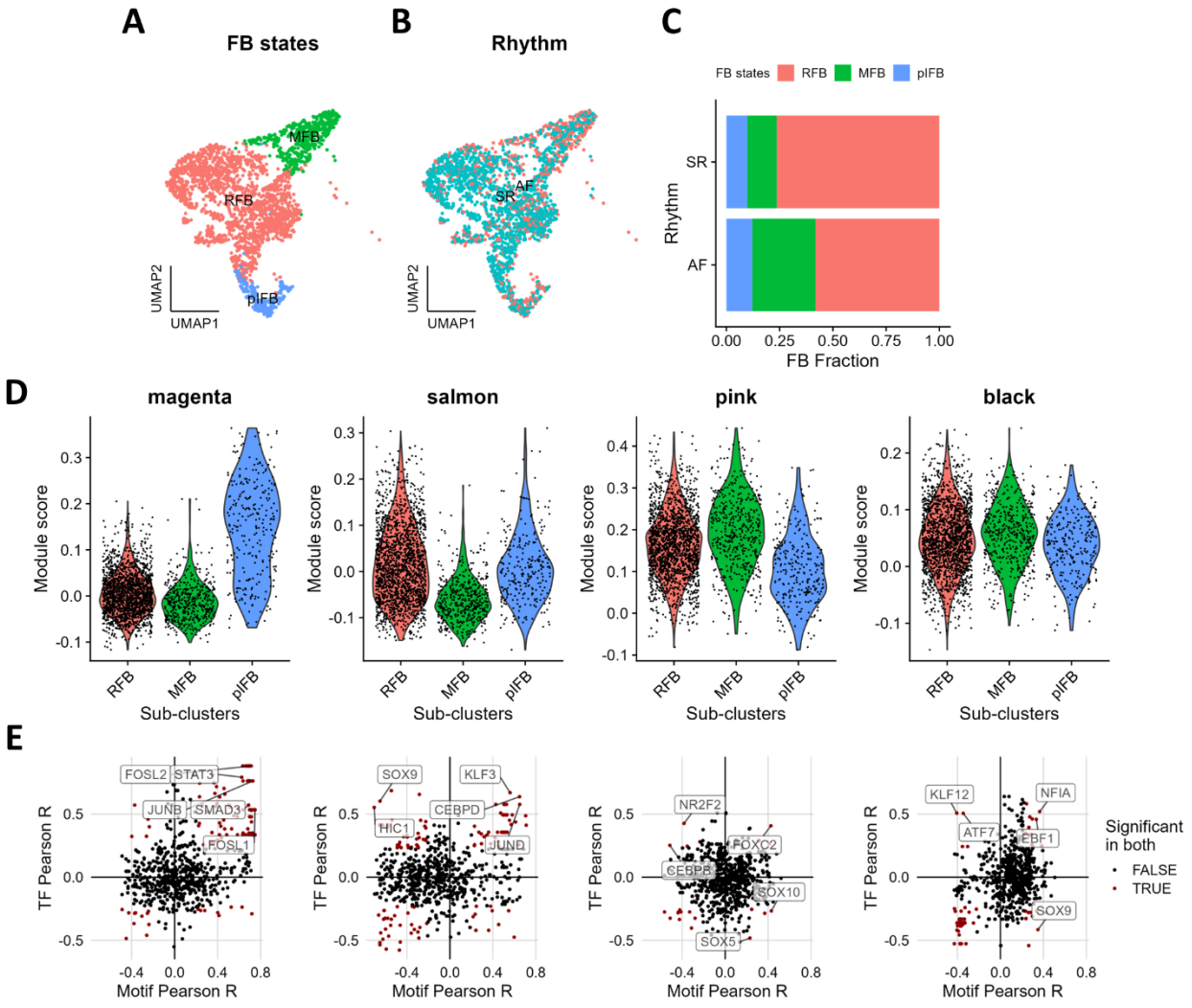


Figure S12. Fibroblast sub-clustering

A-B) Fibroblasts (FB) uniform manifold approximation and projection (UMAP) colored by **A**) state and **B**) rhythm.

C) Bar plot showing the FB state proportion by rhythm.

D) Violin plot showing module scores by FB states.

E) Scatter plot of the transcription factors (TF) and their motif activities correlation with each module in FB metacells. Red dots represent TFs for which the expression and motif activity is significantly correlated (false discovery rate < 0.01) with the module scores.

RFB; resident fibroblasts, MFB; myofibroblasts, pIFB; pro-inflammatory FB.

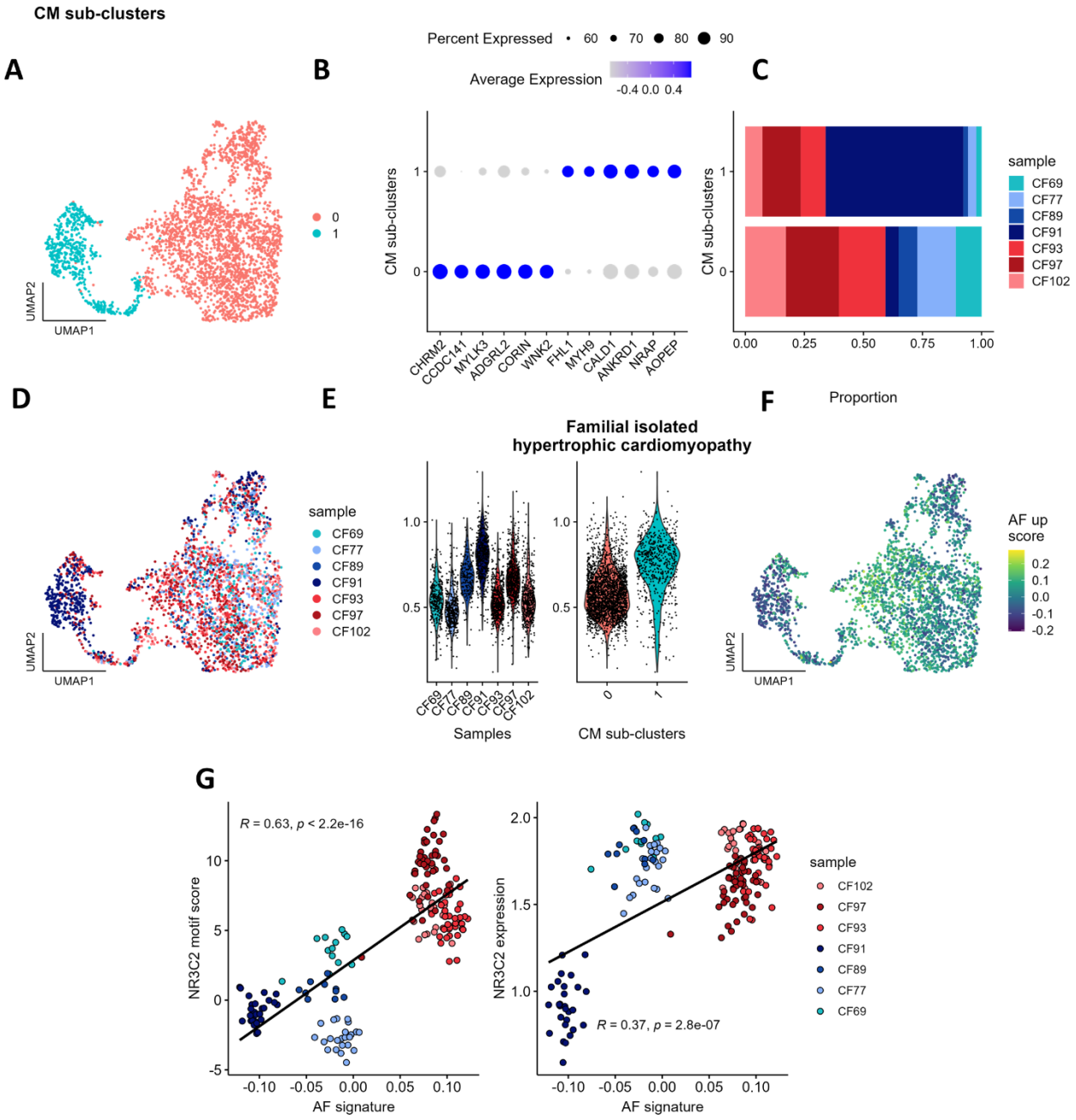


Figure S13. Cardiomyocyte sub-clustering

A, D, F) Cardiomyocyte (CM) uniform manifold approximation and projection (UMAP) colored by **A**) Seurat clusters, **D**) sample and F) the AF signature UP scores (**methods**). Red and blue samples represent atrial fibrillation and sinus rhythm patients respectively.

B) Dot plot showing the top 6 marker genes for each CM sub-cluster.

C) Bar plot showing the sample portions in each CM sub-cluster. Red and blue samples represent atrial fibrillation and sinus rhythm patients respectively.

E) Violin plot showing scores for the familial isolated hypertrophic cardiomyopathy gene set in (left) each sample and (right) each CM sub-cluster. Red and blue samples represent atrial fibrillation and sinus rhythm patients respectively.

G) Scatter plot of (left) NR3C2 motif activity and (right) expression correlations with the AF signature UP scores in CM metacells. Red and blue samples represent atrial fibrillation and sinus rhythm patients respectively.

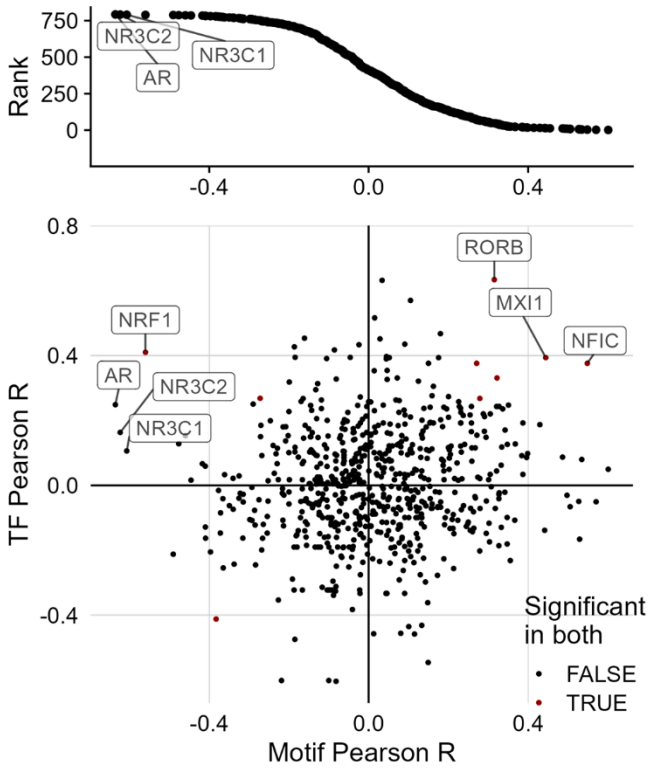


Figure S14. AF gene signature DOWN TF selection

(**top**) Rank plot showing the Pearson R for the motif activities correlation with the AF signature DOWN scores in CM metacells (**methods**). (**bottom**) Scatter plot of the transcription factors (TF) and their motif activities correlation with the AF signature DOWN scores in CM metacells. Red dots represent TFs for which the expression and their motif activity is significantly correlated (false discovery rate < 0.01) with the AF signature DOWN scores.

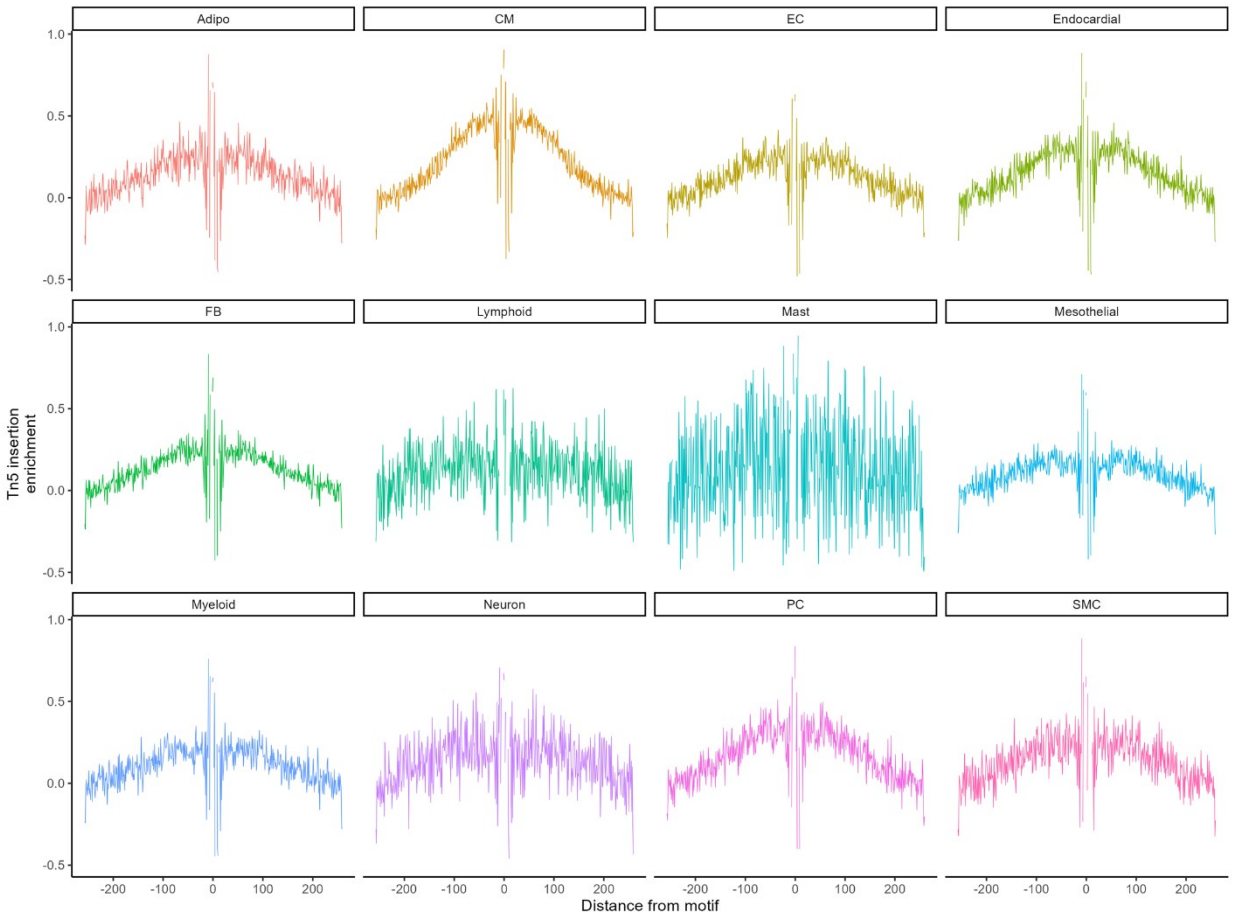


Figure S15. AR footprinting

Footprinting enrichments of the *AR* motif in each cell-type.

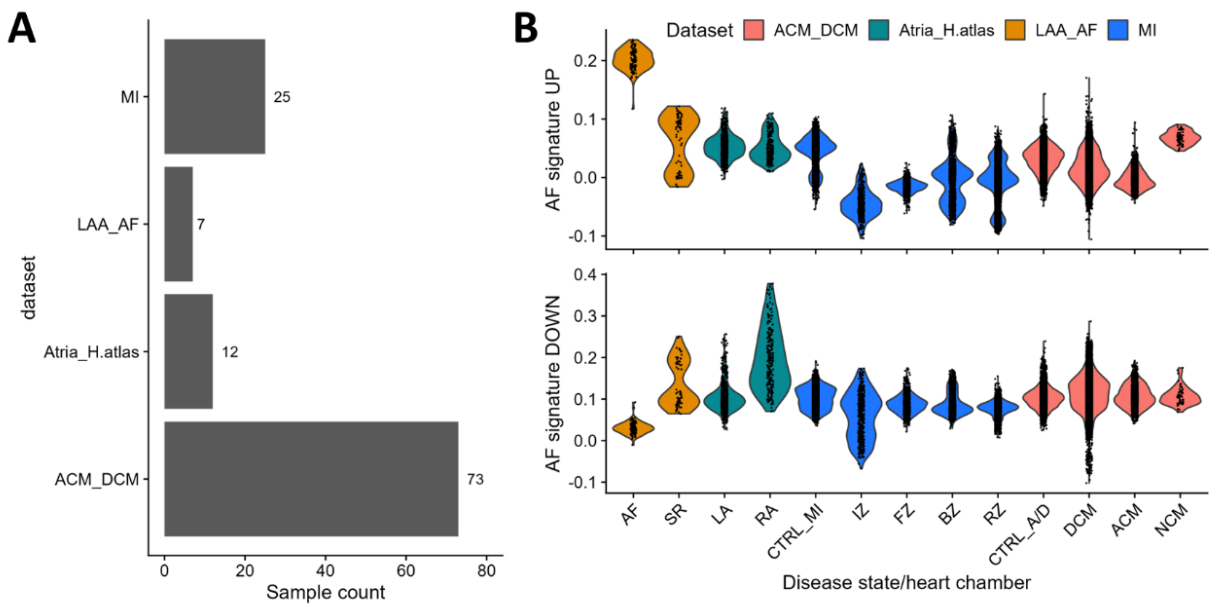


Figure S16. AF CM signature specificity across co-morbidities

A) Bar plot showing the number of samples found in each dataset used to compare the specificity of the AF signatures in **B**.

B) Violin plot showing the AF signature scores (**methods**) of cardiomyocytes metacells in each disease and/or cardiac chamber found in the combined datasets.

LAA_AF; this study scAF dataset, MI; myocardial infarction dataset, Atrial_H.Atlas; atrial heart atlas nuclei dataset, ACM_DCM; arrhythmogenic and dilated cardiomyopathy dataset, LA; left atria, RA; right atria, AF; atrial fibrillation, SR; sinus rhythm, CTRL_MI; control samples from the MI dataset, FZ; fibrotic zone, BZ; boarder zone, IZ; ischemic zone, RZ; remote zone, DCM; dilated cardiomyopathy, ACM; arrhythmogenic cardiomyopathy, CTRL_A/D; control samples from the ACM_DCM dataset, NCM; non-compaction cardiomyopathy.

Chapter 6: Discussion

6.1 Implications

6.1.1 Unappreciated neuron-like characteristics of CMs

Both Chapter 2 and Chapter 5 implicated CM dysregulated genes with functions generally associated with neuron functions. Glutamate, the predominant neurotransmitter, has been exhaustively studied in the central and peripheral nervous system. Released in the synapses through exocytosis, it can activate ion channels opening (ionotropic glutamate receptors; iGluR) and cellular excitability through G-protein receptors (metabotropic GluR; mGluR)[457]. These receptors have been detected in multiple tissues including the heart[363]. Furthermore, they have been detected on CM membranes, with higher prevalence in the atria[458,459]. Importantly, Xi et al. have demonstrated that CM are sensitive to glutamate, N-methyl-D-aspartate (NMDA) receptors and α-amino-3-hydroxy-5 methylisoxazole-4-propionate (AMPA), and that iGluR inhibition reduced conduction velocity, excitability and AF inducibility[367]. Our results in tachypaced canine models showing the upregulation of miRNAs from the DLK1-DIO3 locus predicted to target glutamate signaling, substantiate the evidence of an intrinsic CM gene program dedicated to control glutamate signaling. The regulation of this pathway in our AF models would suggest an adaptative response to AF, reducing CM excitability to restore SR. It should be noted, however, that while our CM preparations strongly enrich for CM, we cannot rule out the presence of other cell-types in these samples. It is therefore a possibility that some of the DEGs in this study come from neuronal cells. Still, the very low proportion of neuronal cells in cardiac tissue without prior CM enrichment (~2.3% in adult human atria[306]) makes the likelihood of a significant neuronal contribution to our 18 CM enriched bulk RNAseq samples unlikely.

The relevance of autonomic imbalance in AF patients has been recognized for decades[460]. Low level vagal stimulation therapy was shown to improve autonomic balance, reduce inflammation and remodeling[461]. On the other hand, high doses of ACh reproducibility has potent arrhythmogenic effects[461], reducing APD and promoting the formation of rotors[60]. The recent discovery of a non-neuronal cholinergic system in CM[462] adds further complexity to our current understanding of this system. CM were shown to possess the machinery responsible for ACh synthesis (choline acetyltransferase; ChAT), storage and transport (transport proteins choline

transporter-1; CHT-1 and the vesicular ACh transporter; VAChT)[444]. Disruption of this system in CM, as evidenced by impaired heart rate recovery post-exercise in VAChT KO mice[463] underscores its physiological significance. Putting our results into perspective, I provided multiple candidate CM target genes implicated in ACh signaling (*CHNRE*), ACh regulation (*COLQ*) and presynaptic vesicle formation (*SYNPR*). Compared to snRNAseq of normal atrial CM, dilated and arrhythmogenic cardiomyopathy ventricular CM, myocardial infarction ventricular CM, normal ventricular CM, and all other cell-types of the LAA of AF patients, these genes were amongst the most specific to CM of AF patients. Contrarily to VAChT which appears to have important ventricular hemodynamic repercussions when targeted in CM specific KO[463], these genes may provide a way to specifically target the CM intrinsic ACh system, without affecting neuronal cells or ventricular CM. Together, our results support the emerging significance of both the glutamate and ACh CM intrinsic systems and their relevance in AF.

With the advent of single cell technologies, the cardiac cellular landscape continues to grow in complexity. Furthermore, cell-states and cellular abundance can readily be associated to diseases. As yet another example of newly discovered cardiac property akin to the nervous system, glial cells have recently been discovered in the heart. In the central nervous system, glial cells support neuron function and homeostasis. Multiomic and spatial omic analyses of the SN have shown the interaction of glial cells with pacemaker cells[308]. These glial cells exhibited astrocyte-like glutamate to glutamine recycling machinery and pacemaker cells were shown to express the necessary genes for its transport, storage and signaling. Astrocytes are also known to be sensitive[464] and have regulating functions on ACh[465]. Importantly, these cells were also found near atrial CM-nerve connections[317], indicating that they are also likely involved in regulating CMs neural connections in the atria. Another group showed that AF recurrence after catheter ablation was correlated with concentrations of S100B, a marker of neuronal damage secreted by glial cells[317]. Given that we found evidence of both the glutamate and ACh systems being disrupted in atrial CM of AF patients and canine models compared to SR, it is reasonable to speculate that such glial cells may be implicated in AF through the regulation of neuro-cardia junctions. It also suggests that CM may be an active participant in regulating neural connections and glial phenotype. The effect of those connections and their nature (sympathetic vs parasympathetic) for AF patients is a matter that would benefit from additional study given the mixed results obtained from targeted ganglionated plexi ablation[317,466-468].

6.1.2 Rare cell-types and cell-states involved in AF

In Chapter 5, I showed enrichments of AF DEG modules in specific cell-types, including modules suggesting an increased proportion of T-cells and mesothelial cells, and modules suggesting a depletion of neurons and adipocytes in AF patients. While this does not provide direct evidence of a change in cell-type abundance, it is the most likely cause for these gene expression shifts since cell-type composition is usually the dominant factor explaining bulk RNAseq variance[469]. These results are also in line with changes that have been described in the epicardium of AF patients. Specifically, fibrosis appeared to antagonise the sub-epicardial adipose tissue (so called fibro-fatty infiltration) and to correlate with increased T-cell infiltration in the epicardium[419]. Expert consensus has posited that sub-epicardial adipogenesis contributes to AF substrate formation[470]. It was argued that this process may occur because of a metabolic shift of AF CM. Conversely, it was also shown that the adult epicardium can lead to fibroblast or adipocyte differentiation through distinct signals[471] and that mesothelial cells of AF patients were more reactive to pro-fibrotic signaling (TGF-β1)[472]. Furthermore, when corrected for body mass index, the PRS for epicardial adipose tissue was not predictive of AF, suggesting that the systemic metabolic dysfunction associated with obesity may be the driver of this association rather than a local effect on conduction heterogeneity[473]. Together, this suggests that sub-epicardial adipogenesis likely precludes fibro-fatty infiltration. Whether the fibro-fatty infiltration correlation with AF progression constitute a more arrhythmogenic substrate than adipose tissue or is merely a consequence remains to be determined. There is, however, little evidence that pertains to a reduction of neural cells in the atria of AF patients. Nonetheless, it is reasonable to assume that the same fibro-fatty infiltration process disrupts myocardial innervations leading to neuronal, adipocyte and CM death. Interestingly, I also identified a module dominated by downregulated genes found to be enriched in a subset of fibroblasts with adipogenic features. CEBPD and KLF3 were both the strongest TF regulatory candidates for this module in fibroblasts which have been associated with responses to adipocyte signaling[435,436]. Further characterization of this fibroblast state could help clarify its possible interaction with adipocytes and its role in regulating fibrosis. Given the low abundance of these cell-type and the large variability of AF presentations in humans, to validate these changes in cell-type proportions, large-sample-size studies will likely be required.

6.1.3 Mechanistically substantiating eQTLs using single nuclei multiome

In chapter 3, I put an emphasis on the identification of peak-gene links. It is important, however, to note that while identifying such links supports a mechanistic interpretation of these peaks on gene expression i.e., repression or activation through chromatin remodeling, it does not exclude the possibility that genes in *cis* may also be regulated by a peak if no links are found. Discordance of these readouts may in fact be the norm. In prefrontal cortex of 272 individuals, only 23% of caQTLs overlapped eQTLs, 63.6% of which had a concordant direction of effect[474].

There are multiple steps required to modify the repressed heterochromatin towards high transcriptional activity. Pioneer factors are a subclass of TFs that can bind nucleosomal motifs and initiate chromatin remodeling from heterochromatin. These include *TBX5, NKX2-5* and *GATA4* for CM commitment[475]. Their initial binding promotes the recruitment of chromatin modifiers and a primed chromatin state, characterized by a weak H3K4me1 ChiPseq mark and a weak ATACseq peak[476]. This state enables the recruitment of other co-activators such as p300–CREB-binding protein (CBP) which adds the active enhancer H3K27ac mark and contributes to nucleosomal remodeling, creating a strong ATACseq peak. Conversely, repressors can reverse this process and reduce chromatin accessibility. In a very dynamic system, such as dedifferentiation of B-cell into pluripotent stem cells, this process is very deterministic on gene expression, where peaks accessibility precedes gene expression in *cis*[477]. However, the changes occurring in an adult heart are generally relatively minor compared to cell reprogramming. Hence, multiple mechanisms leading to adjustments in gene expression may be independent of changes in chromatin accessibility. For instance, disruption of TF motifs involved in the last steps of polymerase recruitment, after the establishment of the active enhancer state, may not necessarily change peak accessibility but would change gene expression. Conversely, it can change topologically associated domains through motif disruption of insulators such as CTCF[478]. It can also directly affect post-transcriptional regulation such as RNA stability[479]. Lastly, some variants can influence splicing, which itself can influence RNA stability and total gene abundance[480]. Therefore, establishing the presence of links is most useful to reinforce the connection of an eQTL through plausible mechanisms and help prioritize causal SNP(s) through their peak proximity, but to completely rule out the regulatory potential of an eQTL containing peak, more in depth analyses of the eGene transcripts and ChiP are necessary.

These matters are especially relevant to the results presented in chapter 4. There I showed that multiple fine mapped AF eQTL SNPs overlapped with ATACseq peaks of our LAA multiome data. Specifically, these involved peaks around the eGenes *MAPT*, *GNB4* and *KDM1B*. Despite identifying some correlations between the peak and the gene (links) for all these loci, the peaks accessibilities did not systematically mirror the eQTLs. For instance, the *MAPT* eQTL SNP rs242557 replicated in our two cohorts of different ancestries, in GTEx LAA and showed concordant increased expression with the A allele in our multiome CMs but we did not observe concordant modifications in peak accessibility in the same CMs. This contrasts with *GNB4* which showed concordant peak and gene increase with the rs7612445 T allele. The latter fits the canonical active enhancer model while the former suggests that the change in *MAPT* expression is not mediated by modulating peak accessibility. This exemplifies the current difficulty regarding the interpretation of multiome data discussed above. Additionally, the *MAPT* eQTL hosting peak was also present in LAA neurons. Given their rare abundance (~ 10 nuclei per sample) and the high sparsity of snATACseq data, we could not assess the QTL effect on neuron peak accessibility. Instead, we fall back on the assumption that CMs explain most of AF heritability, as demonstrated by GWAS enrichments in CM peaks[287,305,307]. Yet, rs242557 was shown to regulate gene expression at the *MAPT* locus in microglia[481]. Also, the A allele is protective for AF[129] but appears to be deleterious for Alzheimer disease by increasing tau aggregates[482]. If these phenotypes from the central nervous system replicate in atrial neurons or glial cells and that this mechanism is causal for AF, it would entail that their dysfunction be protective for AF. This appears less plausible than a mediation through a CM phenotype. Further mystifying the *MAPT* association, tau aggregates clearance in a heart failure mice model improved diastolic function[84]. Hence, functional studies are needed to reconcile this discordant information. Specifically, confirming the eQTL effect of rs242557, evaluating the effect of *MAPT* expression on CM electrophysiology and if rs242557 impacts specific *MAPT* isoforms.

6.2 Limitations

There are several limitations that apply to this research. In addition to what was previously discussed in chapter 2 (potential cell contamination in our CM preparation), the identification of miRNA gene targets is based on predictions and will need to be validated. In chapter 3, an important

limitation of our links validation was that we used data from a single individual. Some of these conclusions may differ for inter-individual analyses sur as eQTLs. Specifically, within cell-type links may turn out to provide better sensitivity if a link is only active within a cell-type which would otherwise be diluted (e.g., the *GNB4* locus in chapter 4), similarly to cell-type specific eQTLs. This aspect is further discussed in the outlook section 6.3. In addition, this underlines the need for larger multiome datasets to improve these models.

Concurrently, our multiome dataset used in chapter 4 and 5 had a relatively low number of nuclei which limited some of our analyses to the more abundant cell-types. One example of this is our identification of gene modules that did not perfectly recapitulate fibroblast states (i.e., the salmon module). More nuclei/cells may provide finer resolution of cell-states with more specific enrichment of such modules. The low number of nuclei in our multiome dataset is in part due to the nature of the tissue and the disease. Because AF is rare in young individuals our LAA samples were from individuals with a mean age of 67 years. LAA from older individuals are more fibrotic, producing more extracellular debris in the nuclei preparation and increasing the number of filtrations required. Moreover, recruitment of patients is a limiting factor, and even more so for female patients. Hence, our multiome data only contained seven individuals, of which two were female. We did not evaluate the effect of sex on eQTLs or DGE. Given that sex is an important risk factor for AF, it will be key component to evaluate in future studies. Another sample-size limiting factor was in relation to our bulk eQTL analyses which we limited to common variants (minor allele frequencies > 5%). Therefore, we may miss the causal variants in our fine-mapping analysis.

A more general limitation of snRNAseq from cardiac tissue is that cardiomyocytes are often binucleated (estimated between 25% and 63% in adults[483]), which may increase the chance of CM doublet formation, especially given the important ECM component of this tissue. The formation of CM homotypic doublets could be favored under incomplete cell membranes lysis during tissue dissociation protocols. The detection of homotypic doublets is difficult due to the high gene expression homology of the nuclei. Together, this may explain why we observed a CM increased transcriptional activity in cardiac scRNAseq datasets.

6.3 Outlook

6.3.1 Single cell QTLs

Currently, single cell eQTL studies are rare because large sample sizes are required and expensive. Unsurprisingly, most scRNAseq eQTL studies to date are from peripheral blood mononuclear cells (PBMCs) given their accessibility and user-friendliness. Nevertheless, they provide a glimpse of the information that ought to be gained. The replication rate of snRNAseq eQTLs in the same bulk tissue has been reported to be between 41–79%, suggesting that a significant number is missed in bulk[484]. Most interestingly, given the resolution that single cell analyses provides, eQTLs can be assessed along a continuum in a dynamic system and uncover SNP effects that only occur in narrow widow of differentiation stages. For instance, 66% of eQTLs identified in naïve-to-memory B-cells transition states could only be identified through this transition as opposed to using fixed cell-type analyses[208]. This suggests that many potentially impactful transitional eQTLs have yet to be discovered. This is important because the strongest eQTLs may not necessarily be the ones causing the phenotypes of interest. First, the strongest eQTLs are easier to detect and are therefore the first ones detected in small cohorts, but this is often caused by higher allele frequencies, which implies lower selective pressure on the SNP and a likely weak consequence on the trait. Case in point, some bottleneck genes appear to have evolved a more robust regulatory landscape through increasing the redundancy of their enhancers, effectively buffering the impact of genetic variations[485]. Moreover, many cases of functional redundancy within cellular pathways[486] can mitigate variants' consequences, which may need to occur in conjunction to impact the trait, oncogenesis being the most obvious example[487]. Then again, weak bulk eQTLs may in fact cause strong gene expression changes in specific cell-types when they are rare in the tissue. Large cohort snRNAseq eQTL studies will help decipher rare cell-type-specific QTLs and condition specific QTLs missed by bulk studies. This may answer if eQTLs common to multiple cell-types are more or less likely to impact traits than specific ones. It will also provide new insights that may lead to resolve some GWAS associations and explain the low proportion of GWAS SNPs captured by current eQTLs[488].

Despite the vast knowledge that has yet to be gained, snRNAseq has important limitations. It has an important 3' bias which limits the detection of splicing events and isoform quantification. In a more distant future, long read sequencing in single cells are expected to eventually provide a

solution[249,250]. The snRNAseq method is also sparse compared to bulk RNAseq complicating eQTL modeling. Importantly, it does not account for post-translational modifications and regulation. Other single cell molecular QTLs such as pQTLs[238] and mQTLs[237] will likely improve in scalability but the coming years will surely be dominated by single cell eQTLs given its accessibility.

Still, even in a completely resolved map of eQTLs and other molecular QTLs, multiple challenges will remain to link gene variant to phenotypes. Chiefly, identification of causal SNP(s) will require more diverse cohorts, especially from African ancestry, to reduce LD bloc sizes and improve fine-mapping. Moreover, the multiple eGenes for a given SNP conundrum may require functional validation of each gene and possibly the combination of multiple KO to truly elucidate the SNP's effect.

Profiling QTLs with the multiome assay may provide profound insight into genomic regulatory processes. Some groups have touted having sequenced hundreds of thousands of nuclei from hundreds of individuals using the multiome platform, but those studies have yet to be published. Such large scale paired single nuclei caQTL and eQTL studies will allow to unambiguously partition cCREs into different categories; correlated, anti-correlated, eQTL only and caQTL only. Each of these categories likely have distinct rules that can be learned and applied in future models to improve genomic regulation predictions. In the process, this will also provide invaluable datasets to design better computational methodologies to identify significant links. Much remains to be done in this domain. The chief challenge remains the data sparsity. Some solutions have been proposed to mitigate this effect, such as the creation of metacells or imputation of missing values. Metacells are aggregates of neighboring cells (or nuclei) in the reduced dimensionality feature space. This reduces the impact of "drop outs" and was shown to improve the detection of TF activities of known importance in erythroid differentiation[301]. For imputation, multiple methods have been developed[300,489,490], which are often used to "denoise" snATACseq data. One of them (scBasset) was shown to improve TF activity compared to ChromVAR and outperform other imputation methods on predicting gene expression from snATACseq data[300]. The effect of these methods on calling links will be important to evaluate in future studies. However, the paucity of ground truth to calibrate these models is another important challenge. Given that gene expression changes can plausibly occur without chromatin accessibility changes, RNA abundance is an imperfect tool to calibrate on. Further refinement of peaks that show ChiPseq mark

modifications associated with chromatin modifications or do not modify TADs in exhaustively characterised model such as hematopoietic differentiation may help refine those ground truths by identifying elements expected to have combined caQTL and eQTL effects.

6.3.2 High throughput CM screens

Functional validation of the candidate genes identified in this research is an important future aim. In chapter 5, I showed that the lncRNAs *LINC01479* and *IFNG-AS1* at the *IFNG* locus are reproducibly the strongest DEGs in persistent AF and are specific to CMs. In chapter 4, the strongest eQTL in LAA was with the eGene *LINC01629*. Lower *LINC01629* is associated with lower AF risk[129]. Its repression in human embryonic stem-cell-derived CMs reduced *FOXP2*, *TBX5* and *PITX2*. *FOXP2* was recently identified as a key regulator of gene expression in pacemaker cells[308]. Hence, it is reasonable to hypothesise that the allele(s) reducing *LINC01629* expression may prevent the activation of pacemaker gene programs in myocardial cells and reduce the occurrence of ectopic activities.

Further functional validation of *LINC01629* and the lncRNAs at the IFNG locus in CM will be necessary to confirm these results, as well as the four SNPs from this locus to identify the causal SNP(s). Importantly, lncRNA genes are generally more cell-type specific[205], an significant advantage for their use as a therapeutic targets[491]. However, lncRNAs are poorly conserved across species, which limits the usability of non-human cells for their investigation. Important progress has recently been made on the differentiation of CM from human induced pluripotent stem cells (hiPSCs-CM). Both ventricular and atrial CM specific markers have now been identified under different growth factor exposition (atrial lineage being induced by retinoic acid)[492,493]. While these cells display lineage specificity, they generally have fetal phenotypes, i.e., absence of striations indicating incomplete myofibril formation, altered Ca^{2+} handling and higher resting membrane potential[494]. Longer culture times, electrical stimulation, low glucose and high lipid cultures have shown promising results to further mature these cells and obtain phenotypes reflecting adult electrophysiology[494,495].

Recently, variations of the clustered regularly interspaced short palindromic repeats (CRISPR)-Cas9 system have been developed to promote or repress gene expression. An inactive Cas9 (dead Cas9; dCas9) has been coupled to the transcriptional activator VP64 (CRIPSRa) or the repressor KRAB (CRISPRi). Using appropriate single guide RNAs (sgRNA) infected cells,

specific regions of the genome can be activated with CRISPRa or repressed with CRISPRi. This is a method of choice to probe non-coding regions of interest for their effect on gene expression and cellular phenotypes without introducing cuts to the DNA (less toxic to cells). Gain and loss of function CRISPR screens are dependent on the stable expression of CRISPRi and/or CIRSPRa for sustained gene activation or silencing[496]. This limitation enforces the use of stable cell lines, further highlighting the importance of hiPSCs-CMs in cardiovascular research.

Parallel phenotypic assessment can now be done using multi-well instruments such as the Nanion CardioExcyte 96[497]. This allows combined impedance and field potential to be recorded simultaneously on a 96 well plate for contractility and electrorheological readouts respectively at various time points. Furthermore, CMs can be paced at different frequencies to improve maturation. Combined with improvements in atrial CM differentiation, maturation and stable hiPSCs CRISPRi cell lines can be used to screen for a set of candidate genes in parallel by targeting their promotor regions with sgRNAs or for the effect of SNPs on whole transcriptomes and CM phenotypes with increasing feasibility. Alternatively, pooled screens assessing atrial differentiation efficiency or other traceable phenotypes through trans-gene fluorescence is another appealing method to evaluate an even higher number of targets[498,499]. Together, this promises to tremendously accelerate throughput of candidate gene validation as well as drug discovery[500-502].

6.3.3 Improving early detection and tailoring AF treatments

As I've discussed in the introduction and discussion, AF is not a homogeneous condition. The etiology can differ importantly between patients warranting different therapeutic approaches. Yet, the sub-categorisation of AF beyond its progression spectrum remains a challenge and tailored treatments are rarely envisioned[503]. Among current efforts to personalize treatments, targeted catheter ablation aided with an artificial intelligence algorithm is currently the object of a clinical trial (TAILOR-AF; NCT05169320). Other considerations include bleeding risk, history of heart failure and symptoms which may warrant specific therapies such as LAA occlusion[1]. Improvement in AF mechanism characterization may provide insight to further stratify patients with adequate indications for targeted therapy. Its myriad of contributing factors i.e., inflammation, metabolic syndrome, blood pressure, autonomic function, age, sex and genetic risk all contribute in varying proportions across individuals and likely warrant different therapeutic approaches. For example, in some patients, AF appears to be triggered by heightened sympathetic activity, occurring

predominantly during exercise or periods of stress, while for others, vagal tone may trigger nocturnal events[504]. Differentiating these patients may show that vagus nerve stimulation is only adequate for individuals with sympathetic tone driven AF. Similarly, testosterone replacement therapy may preferentially benefit patients in whom CM have atrophied, such as those also presenting with dilated cardiomyopathies[505].

To better stratify patients and devise appropriate interventions for each stratum, we must first improve discrimination methods. Blood biomarkers, wearable ECG and genetic risk factors provide new ways to discriminate AF patients into groups that may have different indications, which is the mission of the MAESTRIA consortium. For instance, metabolomic analyses of LAA of AF patients showed that both glutamate and choline were increased[364]. These biomarkers may help segment AF patients which suffer from autonomic imbalance. Likewise, the robust LAA AF gene signature described in Chapter 5 may serve as blood biomarker if it is detectable in the blood of patients. Paired with detailed cardiac mapping and other metrics, specific etiologies could then be associated to proxies of increasing accessibility and be scaled to larger cohorts. Ultimately, combining these measurements with low-cost wearable ECG in a large cohort may also provide enough training data for a machine learning model to identify distinct ECG patterns that would, in the future, reduce the necessity of more cumbersome blood tests or cardiac imaging. The number of cardiovascular and non-cardiovascular phenotypes that are predictable from twelve lead ECG and very often also from wearables is astounding and growing quickly. This includes AF, ventricular arrhythmias, left ventricular systolic dysfunction, heart failure, dilated cardiomyopathy, dyskalemia, hyperthyroidism, anemia and others[506]. Twelve-lead ECG can also be used to predict new onset AF[507], allowing for earlier intervention and possibly avoiding progression from undetected AF. Together, these technologies promise to improve personalization of AF interventions for diverse patient subsets.

6.4 Conclusion

Here, I demonstrate the strengths of integrative multiomic technologies, most importantly, in revealing novel AF molecular and cellular associations through innovative single-cell approaches. First, I implicated a predicted glutamate regulatory region at the DLK1-DIO3 locus showing increased activity in early-stage AF canine models. I then focused on improving the predictive power of statistical models used to link gene expression to open chromatin regions for the emerging multimodal single nuclei multiome assay. Building upon this, I combined single nuclei multiome and eQTL data from multiple ancestries to refine fine-mapping results and mechanistically substantiate eQTL associations with cell-type specificity. Finally, I comprehensively characterised AF dysregulated genes, identifying robust and cell-type specific DEG signatures and their regulatory TFs.

These results involved underappreciated neuro-like functions of CMs in the context of AF including intrinsic glutamate and ACh systems. This suggests that CMs are not mere receivers of neuronal inputs but might actively participate in neuromodulation. Taken with recent evidence of glial cells located at atrial CM and nerve fiber junctions, this proposes new biology with potential high impact in AF. Thus, we identified multiple candidate target genes that may specifically target these systems in atrial CM, which may further the development of novel AF therapeutic opportunities.

With up-and-coming multiplexing single-cell technologies, greater sample sizes will be facilitated providing a more complete understanding of cell-type specific and transitional eQTLs. This in turn will lead to novel causal associations and potential therapeutic targets. Expanding the current applications of CRISPR technologies to novel hiPSC-CM models will be instrumental in validating these novel associations. Together, this progress will broaden our understanding of genomic regulation, enable personalized medicine and management of AF and other pathologies.

References

1 Hindricks, G. *et al.* 2020 ESC Guidelines for the diagnosis and management of atrial fibrillation developed in collaboration with the European Association for Cardio-Thoracic Surgery (EACTS) The Task Force for the diagnosis and management of atrial fibrillation of the European Society of Cardiology (ESC) Developed with the special contribution of the European Heart Rhythm Association (EHRA) of the ESC. *European heart journal* **42**, 373-498 (2021).

2 Heijman, J. *et al.* The value of basic research insights into atrial fibrillation mechanisms as a guide to therapeutic innovation: a critical analysis. *Cardiovascular research* **109**, 467-479 (2016).

3 Mou, L. *et al.* Lifetime risk of atrial fibrillation by race and socioeconomic status: ARIC Study (Atherosclerosis Risk in Communities). *Circulation: Arrhythmia and Electrophysiology* **11**, e006350 (2018).

4 Chugh, S. S. *et al.* Worldwide epidemiology of atrial fibrillation: a Global Burden of Disease 2010 Study. *Circulation* **129**, 837-847 (2014).

5 Benjamin, E. J. *et al.* Heart disease and stroke statistics—2019 update: a report from the American Heart Association. *Circulation* **139**, e56-e528 (2019).

6 Dewland, T. A., Olgin, J. E., Vittinghoff, E. & Marcus, G. M. Incident atrial fibrillation among Asians, Hispanics, blacks, and whites. *Circulation* **128**, 2470-2477 (2013).

7 Feinberg, W. M., Blackshear, J. L., Laupacis, A., Kronmal, R. & Hart, R. G. Prevalence, age distribution, and gender of patients with atrial fibrillation: analysis and implications. *Archives of internal medicine* **155**, 469-473 (1995).

8 Chung, M. K. *et al.* Lifestyle and risk factor modification for reduction of atrial fibrillation: a scientific statement from the American Heart Association. *Circulation* **141**, e750-e772 (2020).

9 Flegal, K. M., Kruszon-Moran, D., Carroll, M. D., Fryar, C. D. & Ogden, C. L. Trends in obesity among adults in the United States, 2005 to 2014. *Jama* **315**, 2284-2291 (2016).

10 Guyton, A. C. *Text book of medical physiology*. (China, 2006).

11 András, V. *et al.* Cardiac transmembrane ion channels and action potentials: cellular physiology and arrhythmogenic behavior. *Physiological reviews* (2021).

12 Lane, J. D. & Tinker, A. Have the findings from clinical risk prediction and trials any key messages for safety pharmacology? *Frontiers in Physiology* **8**, 890 (2017).

13 Bers, D. M. Cardiac excitation–contraction coupling. *Nature* **415**, 198-205 (2002).

14 Jose, A. D. & Collison, D. The normal range and determinants of the intrinsic heart rate in man. *Cardiovascular research* **4**, 160-167 (1970).

15 Wink, J. *et al.* Human adult cardiac autonomic innervation: Controversies in anatomical knowledge and relevance for cardiac neuromodulation. *Autonomic Neuroscience* **227**, 102674 (2020).

16 Coote, J. Myths and realities of the cardiac vagus. *The Journal of physiology* **591**, 4073-4085 (2013).

17 Campos, I. D., Pinto, V., Sousa, N. & Pereira, V. H. A brain within the heart: A review on the intracardiac nervous system. *Journal of molecular and cellular cardiology* **119**, 1-9 (2018).

18 Mehra, R. *et al.* Research opportunities in autonomic neural mechanisms of cardiopulmonary regulation: a report from the national heart, lung, and blood institute and the national institutes of health office of the director workshop. *Basic to Translational Science* **7**, 265-293 (2022).
19 Baruscotti, M., Bucchi, A. & DiFrancesco, D. Physiology and pharmacology of the cardiac pacemaker ("funny") current. *Pharmacology & therapeutics* **107**, 59-79 (2005).
20 Issa, Z., Miller, J. & Zipes, D. Electrophysiological mechanisms of cardiac arrhythmias. *Clinical Arrhythmology and Electrophysiology, 3rd ed.; Issa, ZF, Miller, JM, Zipes, DP, Eds*, 51-80 (2012).
21 Tse, G. Mechanisms of cardiac arrhythmias. *Journal of arrhythmia* **32**, 75-81 (2016).
22 Ono, K. *et al.* Japanese Circulation Society and Japanese Heart Rhythm Society Joint Working Group. JCS/JHRS 2020 guideline on pharmacotherapy of cardiac arrhythmias. *Circ J* (2022).
23 Antzelevitch, C. & Burashnikov, A. Overview of basic mechanisms of cardiac arrhythmia. *Cardiac electrophysiology clinics* **3**, 23-45 (2011).
24 Gourine, A. V. & Ackland, G. L. Cardiac vagus and exercise. *Physiology* **34**, 71-80 (2019).
25 Mangoni, M. E. & Nargeot, J. Genesis and regulation of the heart automaticity. *Physiological reviews* **88**, 919-982 (2008).
26 Weiss, J. N., Qu, Z. & Shivkumar, K. Electrophysiology of hypokalemia and hyperkalemia. *Circulation: arrhythmia and electrophysiology* **10**, e004667 (2017).
27 Klein, I. & Danzi, S. Thyroid disease and the heart. *Circulation* **116**, 1725-1735 (2007).
28 Yamakawa, H. *et al.* Thyroid hormone plays an important role in cardiac function: from bench to bedside. *Frontiers in physiology* **12**, 606931 (2021).
29 Maruyama, M. *et al.* Genesis of phase 3 early afterdepolarizations and triggered activity in acquired long-QT syndrome. *Circulation: Arrhythmia and Electrophysiology* **4**, 103-111 (2011).
30 Zhang, Z. & Qu, Z. Mechanisms of phase-3 early afterdepolarizations and triggered activities in ventricular myocyte models. *Physiological Reports* **9**, e14883 (2021).
31 Tsutsui, H. *et al.* Alterations in sarcoplasmic reticulum calcium-storing proteins in pressure-overload cardiac hypertrophy. *American Journal of Physiology-Heart and Circulatory Physiology* **272**, H168-H175 (1997).
32 Kho, C., Lee, A. & Hajjar, R. J. Altered sarcoplasmic reticulum calcium cycling—targets for heart failure therapy. *Nature Reviews Cardiology* **9**, 717-733 (2012).
33 Tazmini, K. *et al.* Hypokalemia promotes arrhythmia by distinct mechanisms in atrial and ventricular myocytes. *Biophysical Journal* **118**, 103a (2020).
34 Sleiman, Y., Lacampagne, A. & Meli, A. C. "Ryanopathies" and RyR2 dysfunctions: can we further decipher them using in vitro human disease models? *Cell Death & Disease* **12**, 1041 (2021).
35 Liu, M. B., de Lange, E., Garfinkel, A., Weiss, J. N. & Qu, Z. Delayed afterdepolarizations generate both triggers and a vulnerable substrate promoting reentry in cardiac tissue. *Heart rhythm* **12**, 2115-2124 (2015).
36 Arbustini, E. *et al.* in *Left ventricular noncompaction* (Oxford University Press, 2018).
37 Cosío, F. G., MARTÍN-PEÑATO, A., Pastor, A., Nuñez, A. & Goicolea, A. Atypical flutter: a review. *Pacing and clinical electrophysiology* **26**, 2157-2169 (2003).
38 Pandit, S. V. & Jalife, J. Rotors and the dynamics of cardiac fibrillation. *Circulation research* **112**, 849-862 (2013).

39 Nattel, S. & Dobrev, D. Controversies about atrial fibrillation mechanisms: aiming for order in chaos and whether it matters. *Circulation research* **120**, 1396-1398 (2017).
40 Allessie, M. A., Bonke, F. & Schopman, F. Circus movement in rabbit atrial muscle as a mechanism of tachycardia. II. The role of nonuniform recovery of excitability in the occurrence of unidirectional block, as studied with multiple microelectrodes. *Circulation Research* **39**, 168-177 (1976).
41 Valderrábano, M. Influence of anisotropic conduction properties in the propagation of the cardiac action potential. *Progress in biophysics and molecular biology* **94**, 144-168 (2007).
42 Nguyen, T. P., Qu, Z. & Weiss, J. N. Cardiac fibrosis and arrhythmogenesis: the road to repair is paved with perils. *Journal of molecular and cellular cardiology* **70**, 83-91 (2014).
43 Kotadia, I. *et al.* Anisotropic cardiac conduction. *Arrhythmia & Electrophysiology Review* **9**, 202 (2020).
44 Issa, Z., Miller, J. & Zipes, D. Ventricular arrhythmias in inherited channelopathies. *Clinical Arrhythmology and Electrophysiology: A Companion to Braunwald's Heart Disease. 2ª ed. Philadelphia: Elsevier Saunders*, 645-684 (2012).
45 Maoz, A., Christini, D. J. & Krogh-Madsen, T. Dependence of phase-2 reentry and repolarization dispersion on epicardial and transmural ionic heterogeneity: a simulation study. *Europace* **16**, 458-465 (2014).
46 Comtois, P., Kneller, J. & Nattel, S. Of circles and spirals: bridging the gap between the leading circle and spiral wave concepts of cardiac reentry. *EP Europace* **7**, S10-S20 (2005).
47 Nattel, S., Shiroshita-Takeshita, A., Brundel, B. J. & Rivard, L. Mechanisms of atrial fibrillation: lessons from animal models. *Progress in cardiovascular diseases* **48**, 9-28 (2005).
48 Jai¨ s, P. *et al.* A focal source of atrial fibrillation treated by discrete radiofrequency ablation. *Circulation* **95**, 572-576 (1997).
49 Oral, H. *et al.* Pulmonary vein isolation for paroxysmal and persistent atrial fibrillation. *Circulation* **105**, 1077-1081 (2002).
50 Spitzer, S. G. *et al.* Randomized evaluation of redo ablation procedures of atrial fibrillation with focal impulse and rotor modulation-guided procedures: the REDO-FIRM study. *Europace* **25**, 74-82 (2023).
51 Miller, J. M. *et al.* Initial independent outcomes from focal impulse and rotor modulation ablation for atrial fibrillation: multicenter FIRM registry. *Journal of cardiovascular electrophysiology* **25**, 921-929 (2014).
52 Buch, E. *et al.* Long-term clinical outcomes of focal impulse and rotor modulation for treatment of atrial fibrillation: A multicenter experience. *Heart Rhythm* **13**, 636-641 (2016).
53 JJM, B. B. *et al.* Atrial fibrillation (Primer). *Nature Reviews: Disease Primers* **8** (2022).
54 Walden, A., Dibb, K. & Trafford, A. Differences in intracellular calcium homeostasis between atrial and ventricular myocytes. *Journal of molecular and cellular cardiology* **46**, 463-473 (2009).
55 Voigt, N. *et al.* Enhanced sarcoplasmic reticulum Ca2+ leak and increased Na+-Ca2+ exchanger function underlie delayed afterdepolarizations in patients with chronic atrial fibrillation. *Circulation* **125**, 2059-2070 (2012).
56 Qi, X. Y. *et al.* Cellular signaling underlying atrial tachycardia remodeling of L-type calcium current. *Circulation research* **103**, 845-854 (2008).
57 Calloe, K., Goodrow, R., Olesen, S.-P., Antzelevitch, C. & Cordeiro, J. M. Tissue-specific effects of acetylcholine in the canine heart. *American Journal of Physiology-Heart and Circulatory Physiology* **305**, H66-H75 (2013).

58 Huang, C.-X. *et al.* Differential densities of muscarinic acetylcholine receptor and IK, ACh in canine supraventricular tissues and the effect of amiodarone on cholinergic atrial fibrillation and IK, ACh. *Cardiology* **106**, 36-43 (2006).
59 Dobrev, D. *et al.* The G protein–gated potassium current IK, ACh is constitutively active in patients with chronic atrial fibrillation. *Circulation* **112**, 3697-3706 (2005).
60 Kneller, J. *et al.* Cholinergic atrial fibrillation in a computer model of a two-dimensional sheet of canine atrial cells with realistic ionic properties. *Circulation research* **90**, e73-e87 (2002).
61 Mahida, S. *et al.* Science linking pulmonary veins and atrial fibrillation. *Arrhythmia & electrophysiology review* **4**, 40 (2015).
62 Nattel, S. & Dobrev, D. Electrophysiological and molecular mechanisms of paroxysmal atrial fibrillation. *Nature Reviews Cardiology* **13**, 575-590 (2016).
63 Chen, Y.-J. *et al.* Effects of rapid atrial pacing on the arrhythmogenic activity of single cardiomyocytes from pulmonary veins: implication in initiation of atrial fibrillation. *Circulation* **104**, 2849-2854 (2001).
64 Burstein, B. & Nattel, S. Atrial fibrosis: mechanisms and clinical relevance in atrial fibrillation. *Journal of the American College of Cardiology* **51**, 802-809 (2008).
65 Verdecchia, P., Angeli, F. & Reboldi, G. Hypertension and atrial fibrillation: doubts and certainties from basic and clinical studies. *Circulation Research* **122**, 352-368 (2018).
66 Huang, G. *et al.* Angiotensin-converting enzyme inhibitors and angiotensin receptor blockers decrease the incidence of atrial fibrillation: a meta-analysis. *European journal of clinical investigation* **41**, 719-733 (2011).
67 Li, X., Garcia-Elias, A., Benito, B. & Nattel, S. The effects of cardiac stretch on atrial fibroblasts: analysis of the evidence and potential role in atrial fibrillation. *Cardiovascular Research* **118**, 440-460 (2022).
68 Camelliti, P., Green, C. R., LeGrice, I. & Kohl, P. Fibroblast network in rabbit sinoatrial node: structural and functional identification of homogeneous and heterogeneous cell coupling. *Circulation research* **94**, 828-835 (2004).
69 Yue, L., Xie, J. & Nattel, S. Molecular determinants of cardiac fibroblast electrical function and therapeutic implications for atrial fibrillation. *Cardiovascular research* **89**, 744-753 (2011).
70 Miragoli, M., Salvarani, N. & Rohr, S. Myofibroblasts induce ectopic activity in cardiac tissue. *Circulation research* **101**, 755-758 (2007).
71 Pool, L., Wijdeveld, L. F., de Groot, N. M. & Brundel, B. J. The role of mitochondrial dysfunction in atrial fibrillation: Translation to druggable target and biomarker discovery. *International journal of molecular sciences* **22**, 8463 (2021).
72 Harada, M., Melka, J., Sobue, Y. & Nattel, S. Metabolic considerations in atrial fibrillation—mechanistic insights and therapeutic opportunities—. *Circulation Journal* **81**, 1749-1757 (2017).
73 Ozcan, C., Zhenping, L., Kim, G., Jeevanandam, V. & Uriel, N. Molecular mechanism of the association between atrial fibrillation and heart failure includes energy metabolic dysregulation due to mitochondrial dysfunction. *Journal of Cardiac Failure* **25**, 911-920 (2019).
74 Li, W. & Sauve, A. A. NAD+ content and its role in mitochondria. *Mitochondrial Regulation: Methods and Protocols*, 39-48 (2015).

75 Zhang, D. *et al.* DNA damage-induced PARP1 activation confers cardiomyocyte dysfunction through NAD+ depletion in experimental atrial fibrillation. *Nature communications* **10**, 1307 (2019).
76 Xie, W. *et al.* Mitochondrial oxidative stress promotes atrial fibrillation. *Scientific reports* **5**, 11427 (2015).
77 Scott Jr, L., Li, N. & Dobrev, D. Role of inflammatory signaling in atrial fibrillation. *International journal of cardiology* **287**, 195-200 (2019).
78 Li, N. & Brundel, B. J. Inflammasomes and proteostasis novel molecular mechanisms associated with atrial fibrillation. *Circulation research* **127**, 73-90 (2020).
79 Ren, M., Li, X., Hao, L. & Zhong, J. Role of tumor necrosis factor alpha in the pathogenesis of atrial fibrillation: a novel potential therapeutic target? *Annals of medicine* **47**, 316-324 (2015).
80 Xu, Q. *et al.* High mobility group box 1 was associated with thrombosis in patients with atrial fibrillation. *Medicine* **97** (2018).
81 Guo, H., Callaway, J. B. & Ting, J. P. Inflammasomes: mechanism of action, role in disease, and therapeutics. *Nature medicine* **21**, 677-687 (2015).
82 Yao, C. *et al.* Enhanced cardiomyocyte NLRP3 inflammasome signaling promotes atrial fibrillation. *Circulation* **138**, 2227-2242 (2018).
83 Wiersma, M. *et al.* Endoplasmic reticulum stress is associated with autophagy and cardiomyocyte remodeling in experimental and human atrial fibrillation. *Journal of the American Heart Association* **6**, e006458 (2017).
84 Luciani, M. *et al.* Big tau aggregation disrupts microtubule tyrosination and causes myocardial diastolic dysfunction: from discovery to therapy. *European Heart Journal* **44**, 1560-1570 (2023).
85 Lip, G. Y. The ABC pathway: an integrated approach to improve AF management. *Nature Reviews Cardiology* **14**, 627-628 (2017).
86 Ansell, J. *et al.* Pharmacology and management of the vitamin K antagonists: American College of Chest Physicians evidence-based clinical practice guidelines. *Chest* **133**, 160S-198S (2008).
87 Dean, L. Warfarin therapy and VKORC1 and CYP genotype. (2018).
88 Ruff, C. T. *et al.* Comparison of the efficacy and safety of new oral anticoagulants with warfarin in patients with atrial fibrillation: a meta-analysis of randomised trials. *The Lancet* **383**, 955-962 (2014).
89 Smith, T. W. Digitalis: mechanisms of action and clinical use. *New England Journal of Medicine* **318**, 358-365 (1988).
90 Lei, M., Wu, L., Terrar, D. A. & Huang, C. L.-H. Modernized classification of cardiac antiarrhythmic drugs. *Circulation* **138**, 1879-1896 (2018).
91 Kodama, I., Kamiya, K. & Toyama, J. Amiodarone: ionic and cellular mechanisms of action of the most promising class III agent. *The American journal of cardiology* **84**, 20-28 (1999).
92 Merino, J. L. & Perez de Isla, L. Treatment with amiodarone: how to avoid complications. *E-Journal of the European Society of Cardiology Council for Cardiology Practice* **10**, 0 (2011).
93 Investigators, A. F. F.-u. I. o. R. M. A comparison of rate control and rhythm control in patients with atrial fibrillation. *New England Journal of Medicine* **347**, 1825-1833 (2002).
94 Kirchhof, P. *et al.* Early rhythm-control therapy in patients with atrial fibrillation. *New England Journal of Medicine* **383**, 1305-1316 (2020).

95 Willems, S. *et al.* Systematic, early rhythm control strategy for atrial fibrillation in patients with or without symptoms: the EAST-AFNET 4 trial. *European Heart Journal* **43**, 1219-1230 (2022).
96 Blomström-Lundqvist, C. *et al.* Effect of catheter ablation vs antiarrhythmic medication on quality of life in patients with atrial fibrillation: the CAPTAF randomized clinical trial. *Jama* **321**, 1059-1068 (2019).
97 Haissaguerre, M. *et al.* Spontaneous initiation of atrial fibrillation by ectopic beats originating in the pulmonary veins. *New England Journal of Medicine* **339**, 659-666 (1998).
98 Proietti, R. *et al.* Comparative effectiveness of wide antral versus ostial pulmonary vein isolation: a systematic review and meta-analysis. *Circulation: Arrhythmia and Electrophysiology* **7**, 39-45 (2014).
99 Chen, C. F., Gao, X. F., Liu, M. J., Jin, C. L. & Xu, Y. Z. Safety and efficacy of the ThermoCool SmartTouch SurroundFlow catheter for atrial fibrillation ablation: a meta-analysis. *Clinical cardiology* **43**, 267-274 (2020).
100 Kuck, K.-H. *et al.* Cryoballoon or radiofrequency ablation for paroxysmal atrial fibrillation. *New England Journal of Medicine* **374**, 2235-2245 (2016).
101 Hindricks, G., Dagres, N., Sommer, P. & Bollmann, A. in *The ESC Textbook of Cardiovascular Medicine* (eds A. John Camm, Thomas F. Lüscher, Gerald Maurer, & Patrick W. Serruys) 0 (Oxford University Press, 2018).
102 Dörr, M. *et al.* The WATCH AF trial: SmartWATCHes for detection of atrial fibrillation. *JACC: Clinical Electrophysiology* **5**, 199-208 (2019).
103 Lau, D. H., Nattel, S., Kalman, J. M. & Sanders, P. Modifiable risk factors and atrial fibrillation. *Circulation* **136**, 583-596 (2017).
104 Wetterstrand, K. A. *DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP)*, <www.genome.gov/sequencingcostsdata> (2023).
105 Mukherjee, S. *The Gene: An Intimate History* (Scribner, 2016).
106 Darwin, C. *The variation of animals and plants under domestication*. Vol. 2 (J. murray, 1868).
107 Mendel, G. Experiments in plant hybridization (1865). *Verhandlungen des naturforschenden Vereins Brünn.) Available online: www. mendelweb. org/Mendel. html (accessed on 1 January 2013)* (1996).
108 King, E. A., Davis, J. W. & Degner, J. F. Are drug targets with genetic support twice as likely to be approved? Revised estimates of the impact of genetic support for drug mechanisms on the probability of drug approval. *PLoS genetics* **15**, e1008489 (2019).
109 Botstein, D., White, R. L., Skolnick, M. & Davis, R. W. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American journal of human genetics* **32**, 314 (1980).
110 Burg, S. & Attali, B. Targeting of potassium channels in cardiac arrhythmias. *Trends in Pharmacological Sciences* **42**, 491-506 (2021).
111 Savio-Galimberti, E. & Darbar, D. Atrial fibrillation and SCN5A variants. *Cardiac electrophysiology clinics* **6**, 741-748 (2014).
112 Chalazan, B. *et al.* Association of rare genetic variants and early-onset atrial fibrillation in ethnic minority individuals. *JAMA cardiology* **6**, 811-819 (2021).
113 Galton, F. The average contribution of each several ancestor to the total heritage of the offspring. *Proceedings of the Royal Society of London* **61**, 401-413 (1897).
114 Yengo, L. *et al.* A saturated map of common genetic variants associated with human height. *Nature* **610**, 704-712 (2022).

115 LaFramboise, T. Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic acids research* **37**, 4181-4193 (2009).
116 Consortium, G. P. A global reference for human genetic variation. *Nature* **526**, 68 (2015).
117 uk, I. H. C. A. D. a. m. m. h. e. D. P. d. s. o. a. A haplotype map of the human genome. *Nature* **437**, 1299-1320 (2005).
118 Sirugo, G., Williams, S. M. & Tishkoff, S. A. The missing diversity in human genetic studies. *Cell* **177**, 26-31 (2019).
119 Schaid, D. J., Chen, W. & Larson, N. B. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nature Reviews Genetics* **19**, 491-504 (2018).
120 Weissbrod, O. *et al.* Functionally informed fine-mapping and polygenic localization of complex trait heritability. *Nature genetics* **52**, 1355-1363 (2020).
121 Chen, W., McDonnell, S. K., Thibodeau, S. N., Tillmans, L. S. & Schaid, D. J. Incorporating functional annotations for fine-mapping causal variants in a Bayesian framework using summary statistics. *Genetics* **204**, 933-958 (2016).
122 Kichaev, G. *et al.* Improved methods for multi-trait fine mapping of pleiotropic risk loci. *Bioinformatics* **33**, 248-255 (2017).
123 Hernández, N. *et al.* The flashfm approach for fine-mapping multiple quantitative traits. *Nature communications* **12**, 1-14 (2021).
124 LaPierre, N. *et al.* Identifying causal variants by fine mapping across multiple studies. *PLoS genetics* **17**, e1009733 (2021).
125 Weng, L.-C. *et al.* Heritability of atrial fibrillation. *Circulation: Cardiovascular Genetics* **10**, e001838 (2017).
126 Roselli, C. *et al.* Multi-ethnic genome-wide association study for atrial fibrillation. *Nature genetics* **50**, 1225-1233 (2018).
127 Nielsen, J. B. *et al.* Biobank-driven genomic discovery yields new insight into atrial fibrillation biology. *Nature genetics* **50**, 1234-1239 (2018).
128 Roselli, C., Rienstra, M. & Ellinor, P. T. Genetics of atrial fibrillation in 2020: GWAS, genome sequencing, polygenic risk, and beyond. *Circulation research* **127**, 21-33 (2020).
129 Miyazawa, K. *et al.* Cross-ancestry genome-wide analysis of atrial fibrillation unveils disease biology and enables cardioembolic risk prediction. *Nature genetics* **55**, 187-197 (2023).
130 Ryan, A. K. *et al.* Pitx2 determines left–right asymmetry of internal organs in vertebrates. *Nature* **394**, 545-551 (1998).
131 Nattel, S., Heijman, J., Zhou, L. & Dobrev, D. Molecular basis of atrial fibrillation pathophysiology and therapy: a translational perspective. *Circulation research* **127**, 51-72 (2020).
132 Torkamani, A., Wineinger, N. E. & Topol, E. J. The personal and clinical utility of polygenic risk scores. *Nature Reviews Genetics* **19**, 581-590 (2018).
133 Weng, L.-C. *et al.* Genetic predisposition, clinical risk factor burden, and lifetime risk of atrial fibrillation. *Circulation* **137**, 1027-1038 (2018).
134 Khera, A. V. *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature genetics* **50**, 1219-1224 (2018).
135 Choe, W.-S. *et al.* A genetic risk score for atrial fibrillation predicts the response to catheter ablation. *Korean Circulation Journal* **49**, 338-349 (2019).
136 Mountjoy, E. *et al.* An open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci. *Nature genetics* **53**, 1527-1533 (2021).

137 Fulco, C. P. *et al.* Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. *Nature genetics* **51**, 1664-1669 (2019).
138 Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proceedings of the national academy of sciences* **74**, 5463-5467 (1977).
139 Mardis, E. R. Next-generation sequencing platforms. *Annual review of analytical chemistry* **6**, 287-303 (2013).
140 Kchouk, M., Gibrat, J.-F. & Elloumi, M. Generations of sequencing technologies: from first to next generation. *Biology and Medicine* **9** (2017).
141 Marx, V. Method of the year: long-read sequencing. *Nature Methods* **20**, 6-11 (2023).
142 Logsdon, G. A., Vollger, M. R. & Eichler, E. E. Long-read human genome sequencing and its applications. *Nature Reviews Genetics* **21**, 597-614 (2020).
143 Nurk, S. *et al.* The complete sequence of a human genome. *Science* **376**, 44-53 (2022).
144 Leinonen, R., Sugawara, H., Shumway, M. & Collaboration, I. N. S. D. The sequence read archive. *Nucleic acids research* **39**, D19-D21 (2010).
145 Stark, R., Grzelak, M. & Hadfield, J. RNA sequencing: the teenage years. *Nature Reviews Genetics* **20**, 631-656 (2019).
146 Burden, C. J., Qureshi, S. E. & Wilson, S. R. Error estimates for the analysis of differential expression from RNA-seq count data. *PeerJ* **2**, e576 (2014).
147 Schurch, N. J. *et al.* How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *Rna* **22**, 839-851 (2016).
148 Lamarre, S. *et al.* Optimization of an RNA-Seq differential gene expression analysis depending on biological replicate number and library size. *Frontiers in plant science* **9**, 108 (2018).
149 Deng, Z.-L., Münch, P. C., Mreches, R. & McHardy, A. C. Rapid and accurate identification of ribosomal RNA sequences via deep learning. *Nucleic acids research* **50**, e60-e60 (2022).
150 Shanker, S. *et al.* Evaluation of commercially available RNA amplification kits for RNA sequencing using very low input amounts of total RNA. *Journal of biomolecular techniques: JBT* **26**, 4 (2015).
151 Ma, F. *et al.* A comparison between whole transcript and 3'RNA sequencing methods using Kapa and Lexogen library preparation methods. *BMC genomics* **20**, 1-12 (2019).
152 Zhao, S. *et al.* Comparison of stranded and non-stranded RNA-seq transcriptome profiling and investigation of gene overlap. *BMC genomics* **16**, 1-14 (2015).
153 Conesa, A. *et al.* A survey of best practices for RNA-seq data analysis. *Genome biology* **17**, 1-19 (2016).
154 Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).
155 Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature biotechnology* **37**, 907-915 (2019).
156 Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nature biotechnology* **34**, 525-527 (2016).
157 Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nature methods* **14**, 417-419 (2017).
158 Sahraeian, S. M. E. *et al.* Gaining comprehensive biological insight into the transcriptome by performing a broad-spectrum RNA-seq analysis. *Nature communications* **8**, 59 (2017).

159 Schaarschmidt, S., Fischer, A., Zuther, E. & Hincha, D. K. Evaluation of seven different RNA-seq alignment tools based on experimental data from the model plant Arabidopsis thaliana. *International journal of molecular sciences* **21**, 1720 (2020).
160 Kovaka, S. *et al.* Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome biology* **20**, 1-13 (2019).
161 Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols* **7**, 562-578 (2012).
162 Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics* **12**, 1-16 (2011).
163 Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923-930 (2014).
164 Zheng, H., Brennan, K., Hernaez, M. & Gevaert, O. Benchmark of long non-coding RNA quantification for RNA sequencing of cancer samples. *GigaScience* **8**, giz145 (2019).
165 Wu, D. C., Yao, J., Ho, K. S., Lambowitz, A. M. & Wilke, C. O. Limitations of alignment-free tools in total RNA-seq quantification. *BMC genomics* **19**, 1-14 (2018).
166 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology* **15**, 1-21 (2014).
167 Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *bioinformatics* **26**, 139-140 (2010).
168 Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome biology* **15**, 1-17 (2014).
169 Baik, B., Yoon, S. & Nam, D. Benchmarking RNA-seq differential expression analysis methods using spike-in and simulation data. *PLoS One* **15**, e0232271 (2020).
170 Maciejewski, H. Gene set analysis methods: statistical models and methodological differences. *Briefings in bioinformatics* **15**, 504-518 (2014).
171 Geistlinger, L. *et al.* Toward a gold standard for benchmarking gene set enrichment analysis. *Briefings in bioinformatics* **22**, 545-556 (2021).
172 Steenman, M. Insight into atrial fibrillation through analysis of the coding transcriptome in humans. *Biophysical Reviews* **12**, 817-826 (2020).
173 Deshmukh, A. *et al.* Left atrial transcriptional changes associated with atrial fibrillation susceptibility and persistence. *Circulation: Arrhythmia and Electrophysiology* **8**, 32-41 (2015).
174 Zeemering, S. *et al.* Atrial fibrillation in the presence and absence of heart failure enhances expression of genes involved in cardiomyocyte structure, conduction properties, fibrosis, inflammation, and endothelial dysfunction. *Heart Rhythm* **19**, 2115-2124 (2022).
175 Yeh, Y.-H. *et al.* Region-specific gene expression profiles in the left atria of patients with valvular atrial fibrillation. *Heart rhythm* **10**, 383-391 (2013).
176 Victorino, J., Alvarez-Franco, A. & Manzanares, M. Functional genomics and epigenomics of atrial fibrillation. *Journal of Molecular and Cellular Cardiology* **157**, 45-55 (2021).
177 Man, J., Barnett, P. & Christoffels, V. M. Structure and function of the Nppa–Nppb cluster locus during heart development and disease. *Cellular and Molecular Life Sciences* **75**, 1435-1444 (2018).
178 Yang, J. *et al.* RGS6, a modulator of parasympathetic activation in heart. *Circulation research* **107**, 1345-1349 (2010).
179 Günbey, C. *et al.* Cardiac autonomic function evaluation in pediatric and adult patients with congenital myasthenic syndromes. *Neuromuscular Disorders* **29**, 290-295 (2019).

180 Ohno, S. in *Brookhaven Symposium in Biology.* 366-370.
181 Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489**, 101-108 (2012).
182 Thum, T. & Condorelli, G. Long noncoding RNAs and microRNAs in cardiovascular pathophysiology. *Circulation research* **116**, 751-762 (2015).
183 Kozomara, A. & Griffiths-Jones, S. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic acids research* **39**, D152-D157 (2010).
184 Kozomara, A., Birgaoanu, M. & Griffiths-Jones, S. miRBase: from microRNA sequences to function. *Nucleic acids research* **47**, D155-D162 (2019).
185 Friedman, R. C., Farh, K. K.-H., Burge, C. B. & Bartel, D. P. Most mammalian mRNAs are conserved targets of microRNAs. *Genome research* **19**, 92-105 (2009).
186 Pajak, M., Simpson, T. I. & Pajak, M. M. Package 'miRNAtap'. (2015).
187 Romaine, S. P., Tomaszewski, M., Condorelli, G. & Samani, N. J. MicroRNAs in cardiovascular disease: an introduction for clinicians. *Heart* **101**, 921-928 (2015).
188 Luo, X., Yang, B. & Nattel, S. MicroRNAs and atrial fibrillation: mechanisms and translational potential. *Nature Reviews Cardiology* **12**, 80-90 (2015).
189 Viereck, J. & Thum, T. Circulating noncoding RNAs as biomarkers of cardiovascular disease and injury. *Circulation research* **120**, 381-399 (2017).
190 Liu, S. J. *et al.* CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. *Science* **355**, eaah7111 (2017).
191 Shen, C. *et al.* YY1-induced upregulation of lncRNA KCNQ1OT1 regulates angiotensin II-induced atrial fibrillation by modulating miR-384b/CACNA1C axis. *Biochemical and biophysical research communications* **505**, 134-140 (2018).
192 Wang, Y. *et al.* LncRNA NRON alleviates atrial fibrosis via promoting NFATc3 phosphorylation. *Molecular and cellular biochemistry* **457**, 169-177 (2019).
193 van Heesch, S. *et al.* The translational landscape of the human heart. *Cell* **178**, 242-260. e229 (2019).
194 Donovan, M. K., D'Antonio-Chronowska, A., D'Antonio, M. & Frazer, K. A. Cellular deconvolution of GTEx tissues powers discovery of disease and cell-type associated regulatory variants. *Nature communications* **11**, 955 (2020).
195 Assum, I. *et al.* Tissue-specific multi-omics analysis of atrial fibrillation. *Nature Communications* **13**, 441 (2022).
196 Buccitelli, C. & Selbach, M. mRNAs, proteins and the emerging principles of gene expression control. *Nature Reviews Genetics* **21**, 630-644 (2020).
197 Arun, G., Diermeier, S. D. & Spector, D. L. Therapeutic targeting of long non-coding RNAs in cancer. *Trends in molecular medicine* **24**, 257-277 (2018).
198 Shabalin, A. A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**, 1353-1358 (2012).
199 Ongen, H., Buil, A., Brown, A. A., Dermitzakis, E. T. & Delaneau, O. Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics* **32**, 1479-1485 (2016).
200 Ye, Y., Zhang, Z., Liu, Y., Diao, L. & Han, L. A multi-omics perspective of quantitative trait loci in precision medicine. *Trends in Genetics* **36**, 318-336 (2020).
201 Võsa, U. *et al.* Large-scale cis-and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nature genetics* **53**, 1300-1310 (2021).
202 Fauman, E. B. & Hyde, C. An optimal variant to gene distance window derived from an empirical definition of cis and trans protein QTLs. *BMC bioinformatics* **23**, 1-11 (2022).

203 Battle, A., Brown, C. D., Engelhardt, B. E. & Montgomery, S. B. Genetic effects on gene expression across human tissues. *Nature* **550**, 204-213 (2017).
204 Kerimov, N. *et al.* A compendium of uniformly processed human gene expression and splicing quantitative trait loci. *Nature genetics* **53**, 1290-1299 (2021).
205 de Goede, O. M. *et al.* Population-scale tissue transcriptomics maps long non-coding RNAs to complex disease. *Cell* **184**, 2633-2648. e2619 (2021).
206 Donovan, M. K., D'Antonio-Chronowska, A., D'Antonio, M. & Frazer, K. A. Cellular deconvolution of GTEx tissues powers discovery of disease and cell-type associated regulatory variants. *Nature communications* **11**, 1-14 (2020).
207 Selewa, A. *et al.* Single-cell genomics improves the discovery of risk variants and genes of cardiac traits. *medRxiv* (2022).
208 Yazar, S. *et al.* Single-cell eQTL mapping identifies cell type–specific genetic control of autoimmune disease. *Science* **376**, eabf3041 (2022).
209 Gamazon, E. R. *et al.* Using an atlas of gene regulation across 44 human tissues to inform complex disease-and trait-associated variation. *Nature genetics* **50**, 956-967 (2018).
210 Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS genetics* **10**, e1004383 (2014).
211 Giambartolomei, C. *et al.* A Bayesian framework for multiple trait colocalization from summary association statistics. *Bioinformatics* **34**, 2538-2545 (2018).
212 Hormozdiari, F. *et al.* Colocalization of GWAS and eQTL signals detects target genes. *The American Journal of Human Genetics* **99**, 1245-1260 (2016).
213 Foley, C. N. *et al.* A fast and efficient colocalization algorithm for identifying shared genetic risk factors across multiple traits. *Nature communications* **12**, 764 (2021).
214 Wainberg, M. *et al.* Opportunities and challenges for transcriptome-wide association studies. *Nature genetics* **51**, 592-599 (2019).
215 Sinner, M. F. *et al.* Integrating genetic, transcriptional, and functional analyses to identify 5 novel genes for atrial fibrillation. *Circulation* **130**, 1225-1235 (2014).
216 Hsu, J. *et al.* Genetic control of left atrial gene expression yields insights into the genetic susceptibility for atrial fibrillation. *Circulation: Genomic and Precision Medicine* **11**, e002107 (2018).
217 Martin, R. I. *et al.* Genetic variants associated with risk of atrial fibrillation regulate expression of PITX2, CAV1, MYOZ1, C9orf3 and FANCC. *Journal of molecular and cellular cardiology* **85**, 207-214 (2015).
218 Sigurdsson, M. I. *et al.* Post-operative atrial fibrillation examined using whole-genome RNA sequencing in human left atrial tissue. *BMC Medical Genomics* **10**, 1-11 (2017).
219 Benaglio, P. *et al.* Allele-specific NKX2-5 binding underlies multiple genetic associations with human electrocardiographic traits. *Nature genetics* **51**, 1506-1517 (2019).
220 Feingold, E. *et al.* The ENCODE (ENCyclopedia of DNA elements) project. *Science* **306**, 636-640 (2004).
221 Bernstein, B. E. *et al.* The NIH roadmap epigenomics mapping consortium. *Nature biotechnology* **28**, 1045-1048 (2010).
222 Snyder, M. P. *et al.* Perspectives on ENCODE. *Nature* **583**, 693-698 (2020).
223 Ecker, J. R. *et al.* ENCODE explained. *Nature* **489**, 52-54 (2012).
224 Skene, P. J. & Henikoff, S. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *elife* **6**, e21856 (2017).
225 Mifsud, B. *et al.* Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nature genetics* **47**, 598-606 (2015).

226 Yan, F., Powell, D. R., Curtis, D. J. & Wong, N. C. From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. *Genome biology* **21**, 1-16 (2020).
227 Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature methods* **10**, 1213-1218 (2013).
228 Grandi, F. C., Modi, H., Kampman, L. & Corces, M. R. Chromatin accessibility profiling by ATAC-seq. *Nature protocols* **17**, 1518-1552 (2022).
229 Hall, A. W. *et al.* Epigenetic analyses of human left atrial tissue identifies gene networks underlying atrial fibrillation. *Circulation: Genomic and precision medicine* **13**, e003085 (2020).
230 van Ouwerkerk, A. F. *et al.* Identification of atrial fibrillation associated genes and functional non-coding variants. *Nature communications* **10**, 4755 (2019).
231 Li, D., Nie, J., Han, Y. & Ni, L. Epigenetic mechanism and therapeutic implications of atrial fibrillation. *Frontiers in Cardiovascular Medicine* **8**, 763824 (2022).
232 Ferreira, J. P., Pitt, B. & Zannad, F. Histone deacetylase inhibitors for cardiovascular conditions and healthy longevity. *The Lancet Healthy Longevity* **2**, e371-e379 (2021).
233 Seki, M. *et al.* Class I histone deacetylase inhibition for the treatment of sustained atrial fibrillation. *Journal of Pharmacology and Experimental Therapeutics* **358**, 441-449 (2016).
234 Heumos, L. *et al.* Best practices for single-cell analysis across modalities. *Nature Reviews Genetics*, 1-23 (2023).
235 Ramani, V. *et al.* Massively multiplex single-cell Hi-C. *Nature methods* **14**, 263-266 (2017).
236 Wu, S. J. *et al.* Single-cell CUT&Tag analysis of chromatin modifications in differentiation and tumor progression. *Nature biotechnology* **39**, 819-824 (2021).
237 Mulqueen, R. M. *et al.* Highly scalable generation of DNA methylation profiles in single cells. *Nature biotechnology* **36**, 428-431 (2018).
238 Bennett, H. M., Stephenson, W., Rose, C. M. & Darmanis, S. Single-cell proteomics enabled by next-generation sequencing or mass spectrometry. *Nature Methods* **20**, 363-374 (2023).
239 Zhu, C. *et al.* An ultra high-throughput method for single-cell joint analysis of open chromatin and transcriptome. *Nature structural & molecular biology* **26**, 1063-1070 (2019).
240 Chen, S., Lake, B. B. & Zhang, K. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nature biotechnology* **37**, 1452-1457 (2019).
241 Mimitou, E. P. *et al.* Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells. *Nature biotechnology* **39**, 1246-1258 (2021).
242 Svensson, V., da Veiga Beltrame, E. & Pachter, L. A curated database reveals trends in single-cell transcriptomics. *Database* **2020**, baaa073 (2020).
243 Li, H. & Humphreys, B. D. Single cell technologies: Beyond microfluidics. *Kidney360* **2**, 1196 (2021).
244 Bakken, T. E. *et al.* Single-nucleus and single-cell transcriptomes compared in matched cortical cell types. *PloS one* **13**, e0209648 (2018).
245 Slyper, M. *et al.* A single-cell and single-nucleus RNA-Seq toolbox for fresh and frozen human tumors. *Nature medicine* **26**, 792-802 (2020).
246 Oh, J.-M. *et al.* Comparison of cell type distribution between single-cell and single-nucleus RNA sequencing: enrichment of adherent cell types in single-nucleus RNA sequencing. *Experimental & Molecular Medicine* **54**, 2128-2134 (2022).

247 Lafzi, A., Moutinho, C., Picelli, S. & Heyn, H. Tutorial: guidelines for the experimental design of single-cell RNA sequencing studies. *Nature protocols* **13**, 2742-2757 (2018).
248 Mylka, V. *et al.* Comparative analysis of antibody-and lipid-based multiplexing methods for single-cell RNA-seq. *Genome Biology* **23**, 1-21 (2022).
249 You, Y. *et al.* Identification of cell barcodes from long-read single-cell RNA-seq with BLAZE. *Genome Biology* **24**, 1-23 (2023).
250 Tian, L. *et al.* Comprehensive characterization of single-cell full-length isoforms in human and mouse with long-read sequencing. *Genome biology* **22**, 1-24 (2021).
251 Haque, A., Engel, J., Teichmann, S. A. & Lönnberg, T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome medicine* **9**, 1-12 (2017).
252 Zheng, G. X. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nature communications* **8**, 14049 (2017).
253 Kaminow, B., Yunusov, D. & Dobin, A. STARsolo: accurate, fast and versatile mapping/quantification of single-cell and single-nucleus RNA-seq data. *Biorxiv*, 2021.2005. 2005.442755 (2021).
254 Melsted, P. *et al.* Modular, efficient and constant-memory single-cell RNA-seq preprocessing. *Nature biotechnology* **39**, 813-818 (2021).
255 Lun, A. T., Riesenfeld, S., Andrews, T., Gomes, T. & Marioni, J. C. EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *Genome biology* **20**, 1-9 (2019).
256 Germain, P.-L., Sonrel, A. & Robinson, M. D. pipeComp, a general framework for the evaluation of computational pipelines, reveals performant single cell RNA-seq preprocessing tools. *Genome biology* **21**, 1-28 (2020).
257 Hippen, A. A. *et al.* miQC: An adaptive probabilistic framework for quality control of single-cell RNA-sequencing data. *PLoS computational biology* **17**, e1009290 (2021).
258 Xi, N. M. & Li, J. J. Benchmarking computational doublet-detection methods for single-cell RNA sequencing data. *Cell systems* **12**, 176-194. e176 (2021).
259 Xi, N. M. & Li, J. J. Protocol for executing and benchmarking eight computational doublet-detection methods in single-cell RNA sequencing data analysis. *STAR protocols* **2**, 100699 (2021).
260 Janssen, P. *et al.* The effect of background noise and its removal on the analysis of single-cell expression data. *Genome Biology* **24**, 140 (2023).
261 Young, M. D. & Behjati, S. SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. *Gigascience* **9**, giaa151 (2020).
262 Yang, S. *et al.* Decontamination of ambient RNA in single-cell RNA-seq with DecontX. *Genome biology* **21**, 1-15 (2020).
263 Fleming, S. J., Marioni, J. C. & Babadi, M. CellBender remove-background: a deep generative model for unsupervised removal of background noise from scRNA-seq datasets. *BioRxiv* **791699** (2019).
264 Caglayan, E., Liu, Y. & Konopka, G. Neuronal ambient RNA contamination causes misinterpreted and masked cell types in brain single-nuclei datasets. *Neuron* **110**, 4043-4056. e4045 (2022).
265 Ahlmann-Eltze, C. & Huber, W. Comparison of transformations for single-cell RNA-seq data. *Nature Methods*, 1-8 (2023).
266 Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587. e3529 (2021).

267 Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome biology* **19**, 1-5 (2018).
268 Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nature methods* **15**, 1053-1058 (2018).
269 Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature methods* **16**, 1289-1296 (2019).
270 Luecken, M. D. *et al.* Benchmarking atlas-level data integration in single-cell genomics. *Nature methods* **19**, 41-50 (2022).
271 Lotfollahi, M. *et al.* Mapping single-cell data to reference atlases by transfer learning. *Nature biotechnology* **40**, 121-130 (2022).
272 McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).
273 Van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *Journal of machine learning research* **9** (2008).
274 Wang, T., Li, B., Nelson, C. E. & Nabavi, S. Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. *BMC bioinformatics* **20**, 1-16 (2019).
275 Nguyen, H. C., Baik, B., Yoon, S., Park, T. & Nam, D. Benchmarking integration of single-cell differential expression. *Nature Communications* **14**, 1570 (2023).
276 Gagnon, J. *et al.* Recommendations of scRNA-seq differential gene expression analysis based on comprehensive benchmarking. *Life* **12**, 850 (2022).
277 Junttila, S., Smolander, J. & Elo, L. L. Benchmarking methods for detecting differential states between conditions from multi-subject single-cell RNA-seq data. *Briefings in bioinformatics* **23**, bbac286 (2022).
278 Finak, G. *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome biology* **16**, 1-13 (2015).
279 Squair, J. W. *et al.* Confronting false discoveries in single-cell differential expression. *Nature communications* **12**, 5692 (2021).
280 Newman, A. M. *et al.* Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nature biotechnology* **37**, 773-782 (2019).
281 Wang, X., Park, J., Susztak, K., Zhang, N. R. & Li, M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nature communications* **10**, 380 (2019).
282 Jin, H. & Liu, Z. A benchmark for RNA-seq deconvolution analysis under dynamic testing environments. *Genome biology* **22**, 1-23 (2021).
283 Luca, B. A. *et al.* Atlas of clinically distinct cell states and ecosystems across human solid tumors. *Cell* **184**, 5482-5496. e5428 (2021).
284 Frishberg, A. *et al.* Cell composition analysis of bulk genomics using single-cell data. *Nature methods* **16**, 327-332 (2019).
285 Jiang, S. *et al.* Single-cell chromatin accessibility and transcriptome atlas of mouse embryos. *Cell Reports* **42** (2023).
286 Calderon, D. *et al.* The continuum of Drosophila embryonic development at single-cell resolution. *Science* **377**, eabn5800 (2022).
287 Zhang, K. *et al.* A single-cell atlas of chromatin accessibility in the human genome. *Cell* **184**, 5985-6001. e5919 (2021).
288 Granja, J. M. *et al.* ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nature genetics* **53**, 403-411 (2021).

289 Stuart, T., Srivastava, A., Madad, S., Lareau, C. A. & Satija, R. Single-cell chromatin state analysis with Signac. *Nature methods* **18**, 1333-1341 (2021).
290 Chen, H. *et al.* Assessment of computational methods for the analysis of single-cell ATAC-seq data. *Genome biology* **20**, 1-25 (2019).
291 Pliner, H. A. *et al.* Cicero predicts cis-regulatory DNA interactions from single-cell chromatin accessibility data. *Molecular cell* **71**, 858-871. e858 (2018).
292 Castro-Mondragon, J. A. *et al.* JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic acids research* **50**, D165-D173 (2022).
293 Weirauch, M. T. *et al.* Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**, 1431-1443 (2014).
294 Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nature methods* **14**, 975-978 (2017).
295 Hao, Y. *et al.* Dictionary learning for integrative, multimodal, and scalable single-cell analysis (preprint). (2022).
296 Fleck, J. S. *et al.* Inferring and perturbing cell fate regulomes in human brain organoids. *Nature*, 1-8 (2022).
297 Kartha, V. K. *et al.* Functional inference of gene regulation using single-cell multi-omics. *Cell genomics* **2** (2022).
298 Duren, Z. *et al.* Regulatory analysis of single cell multiome gene expression and chromatin accessibility data with scREG. *Genome biology* **23**, 1-19 (2022).
299 Ma, S. *et al.* Chromatin potential identified by shared single-cell profiling of RNA and chromatin. *Cell* **183**, 1103-1116. e1120 (2020).
300 Yuan, H. & Kelley, D. R. scBasset: sequence-based modeling of single-cell ATAC-seq using convolutional neural networks. *Nature Methods* **19**, 1088-1096 (2022).
301 Persad, S. *et al.* SEACells infers transcriptional and epigenomic cellular states from single-cell genomics data. *Nature Biotechnology*, 1-12 (2023).
302 Feregrino, C. & Tschopp, P. Assessing evolutionary and developmental transcriptome dynamics in homologous cell types. *Developmental Dynamics* **251**, 1472-1489 (2022).
303 Moreira, L. M. *et al.* Paracrine signalling by cardiac calcitonin controls atrial fibrogenesis and arrhythmia. *Nature* **587**, 460-465 (2020).
304 Steimle, J. D. *et al.* Decoding the PITX2-controlled genetic network in atrial fibrillation. *Jci Insight* **7** (2022).
305 Selewa, A. *et al.* Single-cell genomics improves the discovery of risk variants and genes of cardiac traits. *MedRxiv*, 2022.2002. 2002.22270312 (2022).
306 Litviňuková, M. *et al.* Cells of the adult human heart. *Nature* **588**, 466-472 (2020).
307 Hocker, J. D. *et al.* Cardiac cell type–specific gene regulatory programs and disease risk association. *Science advances* **7**, eabf1444 (2021).
308 Kanemaru, K. *et al.* Spatially resolved multiomics of human cardiac niches. *Nature*, 1-10 (2023).
309 Goodyer, W. R. *et al.* Transcriptomic profiling of the developing cardiac conduction system at single-cell resolution. *Circulation research* **125**, 379-397 (2019).
310 Koenig, A. L. *et al.* Single-cell transcriptomics reveals cell-type-specific diversification in human heart failure. *Nature cardiovascular research* **1**, 263-280 (2022).
311 Wang, Z. *et al.* Cell-type-specific gene regulatory networks underlying murine neonatal heart regeneration at single-cell resolution. *Cell reports* **33** (2020).

312 Kuppe, C. *et al.* Spatial multi-omic map of human myocardial infarction. *Nature* **608**, 766-777 (2022).
313 Chaffin, M. *et al.* Single-nucleus profiling of human dilated and hypertrophic cardiomyopathy. *Nature* **608**, 174-180 (2022).
314 Reichart, D. *et al.* Pathogenic variants damage cell composition and single cell transcription in cardiomyopathies. *Science* **377**, eabo1984 (2022).
315 Simonson, B. *et al.* Single-nucleus RNA sequencing in ischemic cardiomyopathy reveals common transcriptional profile underlying end-stage heart failure. *Cell Reports* **42** (2023).
316 Kikel-Coury, N. L. *et al.* Identification of astroglia-like cardiac nexus glia that are critical regulators of cardiac development and function. *PLoS Biology* **19**, e3001444 (2021).
317 Scherschel, K. *et al.* Cardiac glial cells release neurotrophic S100B upon catheter-based treatment of atrial fibrillation. *Science translational medicine* **11**, eaav7770 (2019).
318 Calcagno, D. *et al.* Single-cell and spatial transcriptomics of the infarcted heart define the dynamic onset of the border zone in response to mechanical destabilization. *Nature Cardiovascular Research* **1**, 1039-1055 (2022).
319 Piroddi, N. *et al.* Myocardial overexpression of ANKRD1 causes sinus venosus defects and progressive diastolic dysfunction. *Cardiovascular research* **116**, 1458-1472 (2020).
320 Huang, L. *et al.* Critical roles of Xirp proteins in cardiac conduction and their rare variants identified in sudden unexplained nocturnal death syndrome and Brugada syndrome in Chinese Han population. *Journal of the American Heart Association* **7**, e006320 (2018).
321 Leblanc, F. J. *et al.* Transcriptomic profiling of canine atrial fibrillation models after one week of sustained arrhythmia. *Circulation: Arrhythmia and Electrophysiology* **14**, e009887 (2021).
322 Leblanc, F. J. & Lettre, G. Major cell-types in multiomic single-nucleus datasets impact statistical modeling of links between regulatory sequences and target genes. *Scientific Reports* **13**, 3924 (2023).
323 Andrade, J., Khairy, P., Dobrev, D. & Nattel, S. The clinical profile and pathophysiology of atrial fibrillation: relationships among clinical features, epidemiology, and mechanisms. *Circ Res* **114**, 1453-1468, doi:10.1161/CIRCRESAHA.114.303211 (2014).
324 Heijman, J. *et al.* The value of basic research insights into atrial fibrillation mechanisms as a guide to therapeutic innovation: a critical analysis. *Cardiovasc Res* **109**, 467-479, doi:10.1093/cvr/cvv275 (2016).
325 Wijffels, M. C., Kirchhof, C. J., Dorland, R. & Allessie, M. A. Atrial fibrillation begets atrial fibrillation: a study in awake chronically instrumented goats. *Circulation* **92**, 1954-1968 (1995).
326 Wakili, R., Voigt, N., Kaab, S., Dobrev, D. & Nattel, S. Recent advances in the molecular pathophysiology of atrial fibrillation. *J Clin Invest* **121**, 2955-2968, doi:10.1172/JCI46315 (2011).
327 Nattel, S. & Harada, M. Atrial remodeling and atrial fibrillation: recent advances and translational perspectives. *J Am Coll Cardiol* **63**, 2335-2345, doi:10.1016/j.jacc.2014.02.555 (2014).
328 Ma, N. *et al.* Left Atrial Appendage Fibrosis and 3-Year Clinical Outcomes in Atrial Fibrillation After Endoscopic Ablation: A Histologic Analysis. *Ann Thorac Surg* **109**, 69-76, doi:10.1016/j.athoracsur.2019.05.055 (2020).
329 Burstein, B., Qi, X. Y., Yeh, Y. H., Calderone, A. & Nattel, S. Atrial cardiomyocyte tachycardia alters cardiac fibroblast function: a novel consideration in atrial remodeling. *Cardiovasc Res* **76**, 442-452, doi:10.1016/j.cardiores.2007.07.013 (2007).

330 Raymond-Paquin, A., Nattel, S., Wakili, R. & Tadros, R. Mechanisms and Clinical Significance of Arrhythmia-Induced Cardiomyopathy. *Can J Cardiol* **34**, 1449-1460, doi:10.1016/j.cjca.2018.07.475 (2018).
331 DiMarco, J. P. Atrial fibrillation and acute decompensated heart failure. *Circ Heart Fail* **2**, 72-73, doi:10.1161/CIRCHEARTFAILURE.108.830349 (2009).
332 Guichard, J. B. *et al.* Role of atrial arrhythmia and ventricular response in atrial fibrillation induced atrial remodeling. *Cardiovasc Res*, doi:10.1093/cvr/cvaa007 (2020).
333 Wood, M. A., Brown-Mahoney, C., Kay, G. N. & Ellenbogen, K. A. Clinical outcomes after ablation and pacing therapy for atrial fibrillation: a meta-analysis. *Circulation* **101**, 1138-1144 (2000).
334 Harada, M. *et al.* Transient receptor potential canonical-3 channel-dependent fibroblast regulation in atrial fibrillation. *Circulation* **126**, 2051-2064, doi:10.1161/CIRCULATIONAHA.112.121830 (2012).
335 Zhang, D. *et al.* Activation of histone deacetylase-6 induces contractile dysfunction through derailment of α-tubulin proteostasis in experimental and human atrial fibrillation. *Circulation* **129**, 346-358 (2014).
336 Brundel, B. J. *et al.* Induction of heat shock response protects the heart against atrial fibrillation. *Circulation research* **99**, 1394-1402 (2006).
337 Qi, X.-Y. *et al.* Role of small-conductance calcium-activated potassium channels in atrial electrophysiology and fibrillation in the dog. *Circulation* **129**, 430-440 (2014).
338 Soneson, C., Love, M. I. & Robinson, M. D. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research* **4** (2015).
339 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550, doi:10.1186/s13059-014-0550-8 (2014).
340 Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research* **43**, e47-e47 (2015).
341 Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884-i890 (2018).
342 ENCODE. *ENCODE_miRNA-seq_STAR_parameters* <https://github.com/rm2011/ENCODE_miRNA-seq_STAR_parameters> (
343 Donovan, M. K. R., D'Antonio-Chronowska, A., D'Antonio, M. & Frazer, K. A. Cellular deconvolution of GTEx tissues powers discovery of disease and cell-type associated regulatory variants. *Nat Commun* **11**, 955, doi:10.1038/s41467-020-14561-0 (2020).
344 Newman, A. M. *et al.* Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat Biotechnol* **37**, 773-782, doi:10.1038/s41587-019-0114-2 (2019).
345 Shen, N.-N. *et al.* Identification of microRNA biomarkers in atrial fibrillation: A protocol for systematic review and bioinformatics analysis. *Medicine* **98** (2019).
346 Rath, S. *et al.* MitoCarta3. 0: an updated mitochondrial proteome now with sub-organelle localization and pathway annotations. *Nucleic Acids Research* **49**, D1541-D1547 (2021).
347 Cox, J. *et al.* Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol Cell Proteomics* **13**, 2513-2526, doi:10.1074/mcp.M113.031591 (2014).
348 Chen, C.-L. *et al.* Altered expression of FHL1, CARP, TSC-22 and P311 provide insights into complex transcriptional regulation in pacing-induced atrial fibrillation. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease* **1772**, 317-329 (2007).

349 Ding, Y. *et al.* Knockout of SORBS2 Protein Disrupts the Structural Integrity of Intercalated Disc and Manifests Features of Arrhythmogenic Cardiomyopathy. *J Am Heart Assoc* **9**, e017055, doi:10.1161/JAHA.119.017055 (2020).
350 Olson, T. M. *et al.* Kv1. 5 channelopathy due to KCNA5 loss-of-function mutation causes human atrial fibrillation. *Human molecular genetics* **15**, 2185-2191 (2006).
351 Lin, H. *et al.* Methylome-wide association study of atrial fibrillation in Framingham Heart Study. *Scientific reports* **7**, 40377 (2017).
352 Luo, X., Yang, B. & Nattel, S. MicroRNAs and atrial fibrillation: mechanisms and translational potential. *Nature Reviews Cardiology* **12**, 80 (2015).
353 Dill, T. L. & Naya, F. J. A hearty dose of noncoding RNAs: The imprinted DLK1-DIO3 locus in cardiac development and disease. *Journal of cardiovascular development and disease* **5**, 37 (2018).
354 Alvarez-Franco, A. *et al.* Transcriptome and proteome mapping in the sheep atria reveal molecular features of atrial fibrillation progression. *Cardiovascular Research* (2020).
355 Deelen, P. *et al.* Improving the diagnostic yield of exome-sequencing by predicting gene–phenotype associations using large-scale gene expression analysis. *Nature communications* **10**, 1-13 (2019).
356 Shalom-Barak, T. *et al.* Ligand-dependent corepressor (LCoR) is a rexinoid-inhibited peroxisome proliferator-activated receptor γ–retinoid X receptor α coactivator. *Molecular and cellular biology* **38** (2018).
357 Calderon, M. R. *et al.* Ligand-dependent Corepressor (LCoR) Recruitment by Krüppel-like Factor 6 (KLF6) regulates expression of the cyclin-dependent kinase inhibitor CDKN1A gene. *Journal of Biological Chemistry* **287**, 8662-8674 (2012).
358 Zhou, Y., Zhang, X. & Klibanski, A. MEG3 noncoding RNA: a tumor suppressor. *Journal of molecular endocrinology* **48**, R45-R53 (2012).
359 Kumar, S., Williams, D., Sur, S., Wang, J.-Y. & Jo, H. Role of flow-sensitive microRNAs and long noncoding RNAs in vascular dysfunction and atherosclerosis. *Vascular pharmacology* **114**, 76-92 (2019).
360 Kaneko, S. *et al.* Interactions between JARID2 and noncoding RNAs regulate PRC2 recruitment to chromatin. *Molecular cell* **53**, 290-300 (2014).
361 Mondal, T. *et al.* MEG3 long noncoding RNA regulates the TGF-β pathway genes through formation of RNA–DNA triplex structures. *Nature communications* **6**, 7743 (2015).
362 Song, S. *et al.* EZH2 as a novel therapeutic target for atrial fibrosis and atrial fibrillation. *Journal of molecular and cellular cardiology* **135**, 119-133 (2019).
363 Gill, S., Veinot, J., Kavanagh, M. & Pulido, O. Human heart glutamate receptors—implications for toxicology, food safety, and drug discovery. *Toxicologic pathology* **35**, 411-417 (2007).
364 Lai, S. *et al.* Combinational Biomarkers for Atrial Fibrillation Derived from Atrial Appendage and Plasma Metabolomics Analysis. *Scientific reports* **8**, 1-11 (2018).
365 de Lartigue, G. Putative roles of neuropeptides in vagal afferent signaling. *Physiology & behavior* **136**, 155-169 (2014).
366 Agarwal, S. K. *et al.* Cardiac autonomic dysfunction and incidence of atrial fibrillation in a large population-based cohort. *Journal of the American College of Cardiology* **69**, 291 (2017).
367 Xie, D. *et al.* Identification of an endogenous glutamatergic transmitter system controlling excitability and conductivity of atrial cardiomyocytes. *Cell Research*, 1-14 (2021).

368 Luo, Y. *et al.* New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic acids research* **48**, D882-D889 (2020).
369 Stelzer, G. *et al.* The GeneCards suite: from gene data mining to disease genome sequence analyses. *Current protocols in bioinformatics* **54**, 1.30. 31-31.30. 33 (2016).
370 Boix, C. A., James, B. T., Park, Y. P., Meuleman, W. & Kellis, M. Regulatory genomic circuitry of human disease loci by integrative epigenomics. *Nature* **590**, 300-307 (2021).
371 van Duijvenboden, K., de Boer, B. A., Capon, N., Ruijter, J. M. & Christoffels, V. M. EMERGE: a flexible modelling framework to predict genomic regulatory elements from genomic signatures. *Nucleic acids research* **44**, e42-e42 (2016).
372 Yan, F., Powell, D. R., Curtis, D. J. & Wong, N. C. From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. *Genome biology* **21**, 1-16 (2020).
373 Ramirez, R. N. *et al.* Dynamic gene regulatory networks of human myeloid differentiation. *Cell systems* **4**, 416-429. e413 (2017).
374 Duren, Z., Chen, X., Jiang, R., Wang, Y. & Wong, W. H. Modeling gene regulation from paired expression and chromatin accessibility data. *Proceedings of the National Academy of Sciences* **114**, E4914-E4923 (2017).
375 Li, K. *et al.* Interrogation of enhancer function by enhancer-targeting CRISPR epigenetic editing. *Nature communications* **11**, 1-16 (2020).
376 Consortium, G. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318-1330 (2020).
377 Jung, S. *et al.* Identification of shared loci associated with both Crohn's disease and leprosy in East Asians. *Human Molecular Genetics* (2022).
378 Benaglia, T., Chauveau, D., Hunter, D. R. & Young, D. S. mixtools: an R package for analyzing mixture models. *Journal of statistical software* **32**, 1-29 (2010).
379 Fairfax, B. P. *et al.* Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science* **343**, 1246949 (2014).
380 Javierre, B. M. *et al.* Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters. *Cell* **167**, 1369-1384. e1319 (2016).
381 Võsa, U. *et al.* Large-scale cis-and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nature genetics* **53**, 1300-1310 (2021).
382 Nasser, J. *et al.* Genome-wide enhancer maps link risk variants to disease genes. *Nature* **593**, 238-243 (2021).
383 Zhang, H. *et al.* Comprehensive understanding of Tn5 insertion preference improves transcription regulatory element identification. *NAR genomics and bioinformatics* **3**, lqab094 (2021).
384 Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome biology* **9**, 1-9 (2008).
385 Ameijeiras-Alonso, J., Crujeiras, R. M. & Rodríguez-Casal, A. Mode testing, critical bandwidth and excess mass. *Test* **28**, 900-919 (2019).
386 Jackman, S. pscl: Classes and methods for R. Developed in the Political Science Computational Laboratory, Stanford University. Department of Political Science, Stanford University, Stanford, CA. R package version 1.03. 5. *http://www. pscl. stanford. edu/* (2010).
387 Delling *et al.* Heart Disease and Stroke Statistics—2020 Update: A Report From the American Heart Association. doi:10.1161/CIR.0000000000000757 (2020).

388 Colilla, S. *et al.* Estimates of current and future incidence and prevalence of atrial fibrillation in the US adult population. *The American journal of cardiology* **112**, 1142-1147 (2013).
389 Ayzenberg, O. *et al.* Atrial Fibrillation Ablation Success Rate-A Retrospective Multicenter Study. *Current Problems in Cardiology*, 101161 (2022).
390 Miyazawa, K. *et al.* Cross-ancestry genome-wide analysis of atrial fibrillation unveils disease biology and enables cardioembolic risk prediction. *Nature Genetics*, 1-11 (2023).
391 Consortium, G. T. E. The GTEx Consortium atlas of genetic regulatory effects across human tissues The Genotype Tissue Expression Consortium. *Science* **369**, 1318-1330 (2019).
392 Epstein, C. B. *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699 (2020).
393 Ghoussaini, M. *et al.* Open Targets Genetics: systematic identification of trait-associated genes using large-scale genetics and functional genomics. *Nucleic acids research* **49**, D1311-D1320 (2021).
394 Bosada, F. M. *et al.* An atrial fibrillation-associated regulatory region modulates cardiac Tbx5 levels and arrhythmia susceptibility. *Elife* **12**, e80317 (2023).
395 Postma, A. V. *et al.* A gain-of-function TBX5 mutation is associated with atypical Holt–Oram syndrome and paroxysmal atrial fibrillation. *Circulation research* **102**, 1433-1442 (2008).
396 Kanemaru, K. *et al.* Spatially resolved multiomics of human cardiac niches. *Nature*, doi:10.1038/s41586-023-06311-1 (2023).
397 Hou, K. *et al.* Causal effects on complex traits are similar for common variants across segments of different continental ancestries within admixed individuals. *Nature genetics*, 1-10 (2023).
398 Bai, J., Lu, Y., Lo, A., Zhao, J. & Zhang, H. PITX2 upregulation increases the risk of chronic atrial fibrillation in a dose-dependent manner by modulating IKs and ICaL—insights from human atrial modelling. *Annals of translational medicine* **8** (2020).
399 Wirka, R. C. *et al.* A common connexin-40 gene promoter variant affects connexin-40 expression in human atria and is associated with atrial fibrillation. *Circulation: Arrhythmia and Electrophysiology* **4**, 87-93 (2011).
400 Perez-Hernandez, M. *et al.* Pitx2c increases in atrial myocytes from chronic atrial fibrillation patients enhancing I Ks and decreasing I Ca, L. *Cardiovascular research* **109**, 431-441 (2016).
401 Syeda, F., Kirchhof, P. & Fabritz, L. PITX2-dependent gene regulation in atrial fibrillation and rhythm control. *The Journal of Physiology* **595**, 4019-4026 (2017).
402 Guo, D. F. *et al.* TBX5 loss-of-function mutation contributes to atrial fibrillation and atypical Holt-Oram syndrome. *Molecular medicine reports* **13**, 4349-4356 (2016).
403 Kirchhoff, S. *et al.* Reduced cardiac conduction velocity and predisposition to arrhythmias in connexin40-deficient mice. *Current Biology* **8**, 299-302 (1998).
404 Gollob, M. H. *et al.* Somatic mutations in the connexin 40 gene (GJA5) in atrial fibrillation. *New England Journal of Medicine* **354**, 2677-2688 (2006).
405 Nelson, M. R. *et al.* The support of human genetic evidence for approved drug indications. *Nature genetics* **47**, 856-860 (2015).
406 Stephens, M. False discovery rates: a new deal. *Biostatistics* **18**, 275-294 (2017).
407 Zhou, H. J., Li, L., Li, Y., Li, W. & Li, J. J. PCA outperforms popular hidden variable inference methods for molecular QTL mapping. *Genome Biology* **23**, 1-17 (2022).

408 Germain, P.-L., Lun, A., Meixide, C. G., Macnair, W. & Robinson, M. D. Doublet identification in single-cell sequencing data using scDblFinder. *F1000Research* **10** (2021).
409 Chen, M.-H. *et al.* Trans-ethnic and ancestry-specific blood-cell genetics in 746,667 individuals from 5 global populations. *Cell* **182**, 1198-1213. e1114 (2020).
410 Morabito, S., Reese, F., Rahimzadeh, N., Miyoshi, E. & Swarup, V. High dimensional co-expression networks enable discovery of transcriptomic drivers in complex biological systems. *bioRxiv*, 2022.2009. 2022.509094 (2022).
411 Lian, X. *et al.* Robust cardiomyocyte differentiation from human pluripotent stem cells via temporal modulation of canonical Wnt signaling. *Proceedings of the National Academy of Sciences* **109**, E1848-E1857 (2012).
412 Rezwani, M., Pourfathollah, A. A. & Noorbakhsh, F. rbioapi: user-friendly R interface to biologic web services' API. *Bioinformatics* **38**, 2952-2953 (2022).
413 Moore, J. *et al.* & Weng, Z.(2020). Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699-710.
414 Ernst, J. & Kellis, M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nature biotechnology* **28**, 817-825 (2010).
415 Sakamoto, T. *et al.* A critical role for estrogen-related receptor signaling in cardiac maturation. *Circulation research* **126**, 1685-1702 (2020).
416 Sakamoto, T. *et al.* The nuclear receptor ERR cooperates with the cardiogenic factor GATA4 to orchestrate cardiomyocyte maturation. *Nature communications* **13**, 1-20 (2022).
417 Steimle, J. & Moskowitz, I. TBX5: a key regulator of heart development. *Current topics in developmental biology* **122**, 195-221 (2017).
418 Acharya, A. *et al.* The bHLH transcription factor Tcf21 is required for lineage-specific EMT of cardiac fibroblast progenitors. *Development* **139**, 2139-2149 (2012).
419 Haemers, P. *et al.* Atrial fibrillation is associated with the fibrotic remodelling of adipose tissue in the subepicardium of human and sheep atria. *European heart journal* **38**, 53-61 (2017).
420 Chaumont, C. *et al.* Epicardial origin of cardiac arrhythmias: clinical evidences and pathophysiology. *Cardiovascular Research* **118**, 1693-1702 (2022).
421 Meixner, A., Karreth, F., Kenner, L., Penninger, J. M. & Wagner, E. F. Jun and JunD-dependent functions in cell proliferation and stress response. *Cell Death & Differentiation* **17**, 1409-1419 (2010).
422 Pfarr, C. M. *et al.* Mouse JunD negatively regulates fibroblast growth and antagonizes transformation by ras. *Cell* **76**, 747-760 (1994).
423 DiFrancesco, D. HCN4, sinus bradycardia and atrial fibrillation. *Arrhythmia & electrophysiology review* **4**, 9 (2015).
424 Corbalan, J. J. & Kitsis, R. N. Vol. 122 796-798 (Am Heart Assoc, 2018).
425 Kato, K. *et al.* Novel CALM3 variant causing calmodulinopathy with variable expressivity in a 4-generation family. *Circulation: Arrhythmia and Electrophysiology* **15**, e010572 (2022).
426 Posokhova, E., Wydeven, N., Allen, K. L., Wickman, K. & Martemyanov, K. A. RGS6/Gβ5 complex accelerates I KACh gating kinetics in atrial myocytes and modulates parasympathetic regulation of heart rate. *Circulation research* **107**, 1350-1354 (2010).
427 Stein, N. *et al.* IFNG-AS1 enhances interferon gamma production in human natural killer cells. *Iscience* **11**, 466-473 (2019).
428 Liu, Y., Shi, Q., Ma, Y. & Liu, Q. The role of immune cells in atrial fibrillation. *Journal of Molecular and Cellular Cardiology* **123**, 198-208 (2018).

429 Hohmann, C. *et al.* Inflammatory cell infiltration in left atrial appendageal tissues of patients with atrial fibrillation and sinus rhythm. *Scientific Reports* **10**, 1685 (2020).
430 Thoonen, R., Hindle, A. G. & Scherrer-Crosbie, M. Brown adipose tissue: The heat is on the heart. *American Journal of Physiology-Heart and Circulatory Physiology* **310**, H1592-H1605 (2016).
431 Stellato, M. *et al.* The AP-1 transcription factor Fosl-2 drives cardiac fibrosis and arrhythmias under immunofibrotic conditions. *Communications Biology* **6**, 161 (2023).
432 Passegué, E. & Wagner, E. F. JunB suppresses cell proliferation by transcriptional activation of p16INK4a expression. *The EMBO journal* **19**, 2969-2979 (2000).
433 Seo, J. *et al.* AP-1 subunits converge promiscuously at enhancers to potentiate transcription. *Genome research* **31**, 538-550 (2021).
434 Ramalingam, A., Hirai, A. & Thompson, E. A. Glucocorticoid inhibition of fibroblast proliferation and regulation of the cyclin kinase inhibitor p21Cip1. *Molecular Endocrinology* **11**, 577-586 (1997).
435 Faust, H. *et al.* Adipocytes regulate fibroblast function, and their loss contributes to fibroblast dysfunction in inflammatory diseases. *bioRxiv*, 2023.2005. 2016.540975 (2023).
436 Sue, N. *et al.* Targeted disruption of the basic Krüppel-like factor gene (Klf3) reveals a role in adipogenesis. *Molecular and cellular biology* **28**, 3967-3978 (2008).
437 Rosenberg, M. A. *et al.* Serum androgens and risk of atrial fibrillation in older men: The Cardiovascular Health Study. *Clinical cardiology* **41**, 830-836 (2018).
438 Barber, M. *et al.* Cardiac arrhythmia considerations of hormone cancer therapies. *Cardiovascular Research* **115**, 878-894 (2019).
439 Tsai, W.-C. *et al.* Ablation of the androgen receptor gene modulates atrial electrophysiology and arrhythmogenesis with calcium protein dysregulation. *Endocrinology* **154**, 2833-2842 (2013).
440 Sharma, R. *et al.* Normalization of testosterone levels after testosterone replacement therapy is associated with decreased incidence of atrial fibrillation. *Journal of the American Heart Association* **6**, e004880 (2017).
441 Lincoff, A. M. *et al.* Cardiovascular Safety of Testosterone-Replacement Therapy. *New England Journal of Medicine* (2023).
442 Kim, G. E. & Kass, D. A. Cardiac phosphodiesterases and their modulation for treating heart disease. *Heart Failure*, 249-269 (2017).
443 Pavlidou, N. G. *et al.* Phosphodiesterase 8 governs cAMP/PKA-dependent reduction of L-type calcium current in human atrial fibrillation: a novel arrhythmogenic mechanism. *European heart journal* (2023).
444 Rana, O. R. *et al.* Acetylcholine as an age-dependent non-neuronal source in the heart. *Autonomic Neuroscience* **156**, 82-89 (2010).
445 Rocha-Resende, C. *et al.* Non-neuronal cholinergic machinery present in cardiomyocytes offsets hypertrophic signals. *Journal of molecular and cellular cardiology* **53**, 206-216 (2012).
446 Rodríguez Cruz, P. M., Palace, J. & Beeson, D. The neuromuscular junction and wide heterogeneity of congenital myasthenic syndromes. *International journal of molecular sciences* **19**, 1677 (2018).
447 Darkow, E. *et al.* Meta-Analysis of Mechano-Sensitive Ion Channels in Human Hearts: Chamber-and Disease-Preferential mRNA Expression. *International Journal of Molecular Sciences* **24**, 10961 (2023).

448 Ahn, J., Wu, H. & Lee, K. Integrative analysis revealing human heart-specific genes and consolidating heart-related phenotypes. *Frontiers in genetics* **11**, 777 (2020).
449 Darkow, E. *et al.* TREK-1 knock-down in human atrial fibroblasts leads to a myofibroblastic phenotype: a role in phenoconversion and overview of mechano-sensitive channel mRNA expression in cardiac diseases. *bioRxiv*, 2022.2007. 2021.492231 (2022).
450 Zayas, R., Groshong, J. S. & Gomez, C. M. Inositol-1, 4, 5-triphosphate receptors mediate activity-induced synaptic Ca2+ signals in muscle fibers and Ca2+ overload in slow-channel syndrome. *Cell calcium* **41**, 343-352 (2007).
451 Mitchell, L. E. *et al.* Genome-wide association study of maternal and inherited effects on left-sided cardiac malformations. *Human molecular genetics* **24**, 265-273 (2015).
452 Fornes, O. *et al.* JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic acids research* **48**, D87-D92 (2020).
453 Zhu, A., Ibrahim, J. G. & Love, M. I. Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences. *Bioinformatics* **35**, 2084-2092 (2019).
454 Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics* **9**, 1-13 (2008).
455 Chen, E. Y. *et al.* Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC bioinformatics* **14**, 1-14 (2013).
456 Morabito, S., Reese, F., Rahimzadeh, N., Miyoshi, E. & Swarup, V. hdWGCNA identifies co-expression networks in high-dimensional transcriptomics data. *Cell Reports Methods* (2023).
457 Reiner, A. & Levitz, J. Glutamatergic signaling in the central nervous system: ionotropic and metabotropic receptors in concert. *Neuron* **98**, 1080-1098 (2018).
458 Gill, S. S., Pulido, O. M., Mueller, R. W. & McGuire, P. F. Molecular and immunochemical characterization of the ionotropic glutamate receptors in the rat heart. *Brain research bulletin* **46**, 429-434 (1998).
459 Gill, S. S., Pulido, O. M., Mueller, R. W. & McGuire, P. F. Immunochemical localization of the metabotropic glutamate receptors in the rat heart. *Brain research bulletin* **48**, 143-146 (1999).
460 Coumel, P. Paroxysmal atrial fibrillation: a disorder of autonomic tone? *European heart journal* **15**, 9-16 (1994).
461 Kharbanda, R. K. *et al.* Vagus nerve stimulation and atrial fibrillation: Revealing the paradox. *Neuromodulation: Technology at the Neural Interface* **25**, 356-365 (2022).
462 Roy, A., Guatimosim, S., Prado, V. F., Gros, R. & Prado, M. A. Cholinergic activity as a new target in diseases of the heart. *Molecular Medicine* **20**, 527-537 (2014).
463 Roy, A. *et al.* Cardiomyocyte-secreted acetylcholine is required for maintenance of homeostasis in the heart. *The FASEB Journal* **27**, 5072 (2013).
464 Ozawa, A., Kadowaki, E., Horio, T. & Sakaue, M. Acetylcholine suppresses the increase of glia fibrillary acidic protein expression via acetylcholine receptors in cAMP-induced astrocytic differentiation of rat C6 glioma cells. *Neuroscience Letters* **698**, 146-153 (2019).
465 Smit, A. B. *et al.* A glia-derived acetylcholine-binding protein that modulates synaptic transmission. *Nature* **411**, 261-268 (2001).
466 Katritsis, D. G. *et al.* Autonomic denervation added to pulmonary vein isolation for paroxysmal atrial fibrillation: a randomized clinical trial. *Journal of the American College of Cardiology* **62**, 2318-2325 (2013).

467 Mao, J. *et al.* Ablation of epicardial ganglionated plexi increases atrial vulnerability to arrhythmias in dogs. *Circulation: Arrhythmia and Electrophysiology* **7**, 711-717 (2014).
468 Driessen, A. H. *et al.* Ganglion plexus ablation in advanced atrial fibrillation: the AFACT study. *Journal of the American College of Cardiology* **68**, 1155-1165 (2016).
469 Kim-Hellmuth, S. *et al.* Cell type–specific genetic regulation of gene expression across human tissues. *Science* **369**, eaaz8528, doi:doi:10.1126/science.aaz8528 (2020).
470 Goette, A. *et al.* EHRA/HRS/APHRS/SOLAECE expert consensus on atrial cardiomyopathies: definition, characterization, and clinical implication. *Ep Europace* **18**, 1455-1490 (2016).
471 Suffee, N. *et al.* Reactivation of the epicardium at the origin of myocardial fibro-fatty infiltration during the atrial cardiomyopathy. *Circulation research* **126**, 1330-1342 (2020).
472 Gambini, E. *et al.* Preferential myofibroblast differentiation of cardiac mesenchymal progenitor cells in the presence of atrial fibrillation. *Translational Research* **192**, 54-67 (2018).
473 Ramo, J. T. *et al.* The Cardiovascular Impact and Genetics of Pericardial Adiposity. *medRxiv*, 2023.2007. 2016.23292729 (2023).
474 Bryois, J. *et al.* Evaluation of chromatin accessibility in prefrontal cortex of individuals with schizophrenia. *Nature communications* **9**, 3121 (2018).
475 Bruneau, B. G. Signaling and transcriptional networks in heart development and regeneration. *Cold Spring Harbor perspectives in biology* **5**, a008292 (2013).
476 Balsalobre, A. & Drouin, J. Pioneer factors as master regulators of the epigenome and cell fate. *Nature Reviews Molecular Cell Biology* **23**, 449-464 (2022).
477 Stadhouders, R. *et al.* Transcription factors orchestrate dynamic interplay between genome topology and gene regulation during cell reprogramming. *Nature genetics* **50**, 238-249 (2018).
478 Rowley, M. J. & Corces, V. G. Organizational principles of 3D genome architecture. *Nature Reviews Genetics* **19**, 789-800 (2018).
479 Pai, A. A. *et al.* The contribution of RNA decay quantitative trait loci to inter-individual variation in steady-state gene expression levels. (2012).
480 Qi, T. *et al.* Genetic control of RNA splicing and its distinct role in complex trait variation. *Nature Genetics* **54**, 1355-1363 (2022).
481 Das, R. G., Choi, E., Jiang, M., Dombroski, B. A. & Schellenberg, G. D. MAPT enhancer containing rs242557 regulates multiple neighboring genes in human microglial cell line. *Alzheimer's & Dementia* **17**, e052360 (2021).
482 Shen, X.-N. *et al.* MAPT rs242557 variant is associated with hippocampus tau uptake on 18F-AV-1451 PET in non-demented elders. *Aging (Albany NY)* **11**, 874 (2019).
483 Botting, K. J. *et al.* Early origins of heart disease: low birth weight and determinants of cardiomyocyte endowment. *Clinical and Experimental Pharmacology and Physiology* **39**, 814-823 (2012).
484 Maria, M., Pouyanfar, N., Örd, T. & Kaikkonen, M. U. The power of single-cell RNA sequencing in eQTL discovery. *Genes* **13**, 502 (2022).
485 Wang, X. & Goldstein, D. B. Enhancer domains predict gene pathogenicity and inform gene discovery in complex disease. *The American Journal of Human Genetics* **106**, 215-233 (2020).
486 Nowak, M. A., Boerlijst, M. C., Cooke, J. & Smith, J. M. Evolution of genetic redundancy. *Nature* **388**, 167-171 (1997).

487 Martincorena, I. *et al.* Universal patterns of selection in cancer and somatic tissues. *Cell* **171**, 1029-1041. e1021 (2017).
488 Connally, N. J. *et al.* The missing link between genetic association and regulatory function. *Elife* **11**, e74970 (2022).
489 Li, Z. *et al.* Chromatin-accessibility estimation from single-cell ATAC-seq data with scOpen. *Nature communications* **12**, 6386 (2021).
490 Van Dijk, D. *et al.* Recovering gene interactions from single-cell data using data diffusion. *Cell* **174**, 716-729. e727 (2018).
491 Winkle, M., El-Daly, S. M., Fabbri, M. & Calin, G. A. Noncoding RNA therapeutics—Challenges and potential solutions. *Nature reviews Drug discovery* **20**, 629-651 (2021).
492 Goldfracht, I. *et al.* Generating ring-shaped engineered heart tissues from ventricular and atrial human pluripotent stem cell-derived cardiomyocytes. *Nature communications* **11**, 75 (2020).
493 Thorpe, J. *et al.* Development of a robust induced pluripotent stem cell atrial cardiomyocyte differentiation protocol to model atrial arrhythmia. (2023).
494 Guo, Y. & Pu, W. T. Cardiomyocyte maturation: new phase in development. *Circulation research* **126**, 1086-1106 (2020).
495 Madsen, A. *et al.* An important role for DNMT3A-mediated DNA methylation in cardiomyocyte metabolism and contractility. *Circulation* **142**, 1562-1578 (2020).
496 Bock, C. *et al.* High-content CRISPR screening. *Nature Reviews Methods Primers* **2**, 8 (2022).
497 Juhasz, K. *et al.* Combined impedance and extracellular field potential recordings on iPS cardiomyocytes.
498 Sontayananon, N., Redwood, C., Davies, B. & Gehmlich, K. Fluorescent PSC-derived cardiomyocyte reporter lines: generation approaches and their applications in cardiovascular medicine. *Biology* **9**, 402 (2020).
499 Chirikian, O. *et al.* CRISPR/Cas9-based targeting of fluorescent reporters to human iPSCs to isolate atrial and ventricular-specific cardiomyocytes. *Scientific reports* **11**, 3026 (2021).
500 Grafton, F. *et al.* Deep learning detects cardiotoxicity in a high-content screen with induced pluripotent stem cell-derived cardiomyocytes. *Elife* **10**, e68714 (2021).
501 Williams, T. L. *et al.* Human embryonic stem cell-derived cardiomyocyte platform screens inhibitors of SARS-CoV-2 infection. *Communications Biology* **4**, 926 (2021).
502 Sapp, V. *et al.* Genome-wide CRISPR/Cas9 screening in human iPS derived cardiomyocytes uncovers novel mediators of doxorubicin cardiotoxicity. *Scientific Reports* **11**, 13866 (2021).
503 Deb, B., Ganesan, P., Feng, R. & Narayan, S. M. Identifying atrial fibrillation mechanisms for personalized medicine. *Journal of Clinical Medicine* **10**, 5679 (2021).
504 Rebecchi, M. *et al.* Atrial fibrillation and sympatho–vagal imbalance: from the choice of the antiarrhythmic treatment to patients with syncope and ganglionated plexi ablation. *European Heart Journal Supplements* **25**, C1-C6 (2023).
505 Diaconu, R., Donoiu, I., Mirea, O. & Bălşeanu, T. A. Testosterone, cardiomyopathies, and heart failure: A narrative review. *Asian journal of andrology* **23**, 348 (2021).
506 Martínez-Sellés, M. & Marina-Breysse, M. Current and future use of artificial intelligence in electrocardiography. *Journal of Cardiovascular Development and Disease* **10**, 175 (2023).

507 Raghunath, S. *et al.* Deep neural networks can predict new-onset atrial fibrillation from the 12-lead ECG and help identify those at risk of atrial fibrillation–related stroke. *Circulation* **143**, 1287-1298 (2021).

Appendix