



**UNIVERSIDAD DE BUENOS AIRES**

*Facultad de Ciencias Exactas y Naturales*

*Departamento de Fisiología, Biología Molecular y Celular*

# **Efectos fenotípicos de variantes génicas en genes asociados a cardiopatías congénitas**

*Tesis presentada para optar por el título de Doctor de la Universidad de Buenos Aires  
en el área Ciencias Biológicas*

***Lic. Jorge Emilio Kolomenski***

Director de Tesis: Dr. Alejandro Nadra

Co-Directora de Tesis: Dra. Liliana Dain

Consejero de Estudios: Dr. Norberto Iusem

Lugar de trabajo: Instituto de Biociencias, Biotecnología y Biología traslacional (IB3)

*Buenos Aires, 2024*

# **Efectos fenotípicos de variantes génicas en genes asociados a cardiopatías congénitas**

Las cardiopatías congénitas (CC) abarcan un amplio espectro de anomalías en la estructura del corazón y los grandes vasos. Tienen una prevalencia mundial que oscila entre 6 y 9 de cada 1000 nacimientos y se destacan como causa de muerte prenatal, neonatal e infantil y de morbilidad en edad pediátrica.

El gen *NKX2-5* codifica para un factor de transcripción crucial para el desarrollo del corazón. Es el primer factor de transcripción para el cual se observó una asociación entre sus variantes génicas (VGs) y el desarrollo de CC. Desde entonces, se han descrito múltiples variantes en *NKX2-5* potencialmente causales de CC.

En este trabajo nos propusimos desarrollar un flujo de trabajo para VGs de *NKX2-5* que permita el estudio predictivo de su patogenicidad y su interacción con otras proteínas que se asocien funcionalmente en el desarrollo del corazón.

Se recopilaron todas las VGs de *NKX2-5* de bases de datos, publicaciones científicas y trabajos previos incluyendo variantes de pacientes del Centro Nacional de Genética Médica. También se caracterizaron todas las variantes de cambio de sentido posibles de la región estructurada de *NKX2-5* (el homeodominio, HD). Por último, se caracterizaron todos los sitios de unión de *NKX2-5* en el genoma humano y de ratón,

así como los sitios de unión cercanos de 6 factores de transcripción relacionados, y se estudiaron los *Gene Ontology terms* (GO-terms) de genes cercanos.

De esta forma, se recopilaron 1389 VGs de *NKX2-5*, de las cuales 144 fueron clasificadas como patogénicas. Estudiando su distribución, observamos una concentración de variantes clasificadas como patogénicas en la tercera alfa hélice del HD, la cual está en contacto con el ADN. Haciendo foco en fenotipos específicos, analizamos los fenotipos más comúnmente asociados a VGs patogénicas de *NKX2-5* y su distribución a lo largo del gen. Observamos una abundancia de fenotipos asociados con comunicación interauricular, comunicación interventricular y bloqueo atrioventricular. Otros fenotipos presentaron frecuencias menores al 20%. En particular, se destacó la sobrerrepresentación de bloqueo atrioventricular en VGs de gran efecto en *NKX2-5* y la distribución de VGs asociadas a Tetralogía de Fallot principalmente fuera del homeodominio.

Se complementó el estudio realizado con el análisis de las 337 VGs de cambio de sentido posibles en el HD, para las cuales definimos una jerarquía de potenciales efectos en la funcionalidad de *NKX2-5*. Con estos resultados, estimamos una explicación funcional para 29 de las 36 VGs patogénicas de cambio de sentido del HD, y propusimos los potenciales efectos para 186 de las 301 VGs restantes, basándonos en predicciones *in silico*.

Por otro lado, se recopilaron sitios de unión de *NKX2-5* a partir de estudios de ChIP-seq. Para el genoma humano se obtuvieron 7730 potenciales sitios y 3130 en ratón, con 1021 y 1257 genes cercanos probablemente regulados por *NKX2-5*, respectivamente. Para cada uno de estos sitios, se analizó también la co-ocurrencia de sitios de unión de otros 6 factores de transcripción (*TBX20*, *MEIS1*, *TGIF1*, *GATA4*, *TAL1*, *MAF*) y se registraron GO-terms de genes cercanos. De este análisis se seleccionaron 21 genes de interés para los cuales se investigó su asociación a CC.

Con el estudio de VGs patogénicas, los distintos genes seleccionados y sus fenotipos patogénicos asociados, se sugirieron genes y vías funcionales de interés para

estudios a futuro, entre ellos, *TBX5* para el estudio de fenotipos relacionados a bloqueo atrioventricular y *ZIC3* para el estudio de fenotipos relacionados a heterotaxia.

Por último, se creó un flujo de trabajo con la potencialidad de ser aplicado a otros genes y se definió una lista de genes sugeridos para estudiar de esta manera.

En resumen, mediante un análisis exhaustivo de sitios de unión al genoma, genes potencialmente regulados, así como la contribución biológica de todas las posibles variantes de cambio de sentido en el HD, nuestros resultados contribuyen a reducir la brecha en el entendimiento de la relación genotipo-fenotipo entre VGs en NKX2-5 y CC. Asimismo, fue posible definir posibles mecanismos implicados en el desarrollo de CC a partir de interacciones con otros factores de transcripción y proponer un flujo de trabajo y genes de interés para aplicarlo y direccionar futuros estudios *in vitro*, así como de casos clínicos.

**Palabras claves: NKX2-5; Cardiopatías congénitas; Variantes génicas; Sitios de unión; Factores de transcripción.**

# **Phenotypic effect of genetic variants in genes associated with congenital heart disease**

Congenital heart defects (CHD) encompass a broad spectrum of anomalies in the structure of the heart and great vessels. They affect between 6 and 9 births worldwide and stand out as a cause of prenatal, neonatal and infant death and pediatric morbidity.

The *NKX2-5* gene encodes for a transcription factor that is a key factor in heart development. It's the first transcription factor for which an association between its genetic variants (GVs) and CHD development was observed. Since then, multiple *NKX2-5* GV's have been observed to potentially cause CHD.

In this work, we developed a pipeline to study *NKX2-5* GV's to allow the predictive study of its pathogenicity and its interaction with other proteins that are also functionally associated with heart development.

In order to do this, we gathered all of *NKX2-5*'s GV's from public databases, scientific papers and previous works including variants from Centro Nacional de Genética Médica patients. We also characterized all the possible missense variants in the structured region of *NKX2-5*, the homeodomain (HD). Lastly, we compiled all of *NKX2-5*'s binding sites in the human and mouse genomes, as well as nearby binding

sites from other 6 transcription factors that are closely associated with *NKX2-5*, and we studied the nearby genes' Gene Ontology (GO) terms.

With this analysis, we compiled 1389 *NKX2-5* GVs, of which 144 were classified as pathogenic. Studying their distribution, we observed that pathogenic variants were concentrated in the third alpha-helix of the HD, which is in contact with the DNA. Regarding specific phenotypes, we were able to describe the phenotypes most commonly associated with pathogenic GVs in *NKX2-5* and their distribution. In this way, we were able to observe an abundance of atrial septal defects, ventricular septal defects and atrioventricular block, with other phenotypes being associated with less than 20% of pathogenic GVs. Of note, we found an overrepresentation of atrioventricular block in GVs with big effects on the *NKX2-5* protein. We also observed that most pathogenic GVs associated with Tetralogy of Fallot were located outside of the homeodomain.

We further complemented the study with the analysis of the 337 possible missense GVs in the *NKX2-5* HD, for which we defined a hierarchy of potential effects on normal *NKX2-5* functionality. With these results, we were able to estimate a functional explanation for 29 of the 36 pathogenic missense GVs in the HD, and we proposed potential effects for 186 of the 301 remaining GVs, going by our *in silico* predictors.

In addition, we compiled *NKX2-5* binding sites in human and mouse genomes from ChIP-seq studies. We obtained 7730 binding sites in human cells and 3130 in mice, with 1021 and 1257 nearby genes that are probably regulated by *NKX2-5*, respectively. For each one of these sites, we also analyzed co-occurrence of binding sites of 6 other transcription factors (*TBX20*, *MEIS1*, *TGIF1*, *GATA4*, *TAL1*, *MAF*), recorded the GO-terms of nearby genes and selected 21 genes to study their association with CHD.

With the study of pathogenic GVs, selected genes and their associated pathogenic phenotypes, we proposed genes and functional pathways of interest to

future studies. These are *TBX5* for the study of atrioventricular block associated phenotypes and *ZIC3* for the study of heterotaxy associated phenotypes.

Lastly, we created a workflow that could be applied to other genes and defined a list of suggested genes to be studied with it.

In summary, our results help to narrow the gap in our understanding of the genotype-phenotype between NKX2-5 GV's and CHD, contributing with an exhaustive analysis of genome binding sites, potentially regulated genes, as well as the biological contribution of all possible missense variants in the HD, accessible for future clinical cases. On the other hand, defining CHD development pathways from interaction with other transcription factors and proposing a workflow and interesting genes to apply it on to direct future studies.

**Keywords: NKX2-5; Congenital heart defects; Genetic variants; Binding sites; Transcription factors.**

## **Agradecimientos**

Gracias, de todo corazón, a todas las personas que con su apoyo y sus consejos me ayudaron a completar este trabajo.

Al Dr. Alejandro Daniel Nadra y a la Dra. Liliana Dain, mis directores de tesis, por guiarme (y soportarme) con una pandemia en el medio.

A Ceci, mi futura esposa, por el infinito cariño (y paciencia) durante todos estos años. No podría haber completado este trabajo sin vos, tu compañía me dio felicidad en los momentos difíciles y siempre fuiste la primera en festejar mis éxitos.

A mis compañerxs del IB3: Maru, Eze, Javi, Flor, Elliioth, Iani, Laura, y al grupo de trabajo del Centro Nacional de Genética Médica: Aldi, Meli, Juli, Marisol, Lucía, Noemí, Tania. Su compañía y ayuda hicieron de las oficinas y laboratorios un segundo hogar.

Al Dr. David Bruque y al Dr. Leandro Simonetti, quienes me dieron el puntapié inicial en la bioinformática y siempre estuvieron presentes como mentores.

A mi madre, por sus palabras de aliento que siempre estuvieron presentes y al resto de mi familia, quienes a distintas distancias siempre me apoyaron para continuar, sin importar lo difícil que fuera.

A mis amigxs por comprender mis ausencias y motivarme a seguir con mi esfuerzo.

Al CONICET, por el otorgamiento de la beca que proveyó el sustento económico y el apoyo académico que me permitieron realizar esta tesis de doctorado.

A la Educación Pública, y en especial a la Universidad de Buenos Aires, por permitirme formarme y trabajar en un espacio y con unos equipos apropiados, y a lxs docentes de los numerosos cursos y materias necesarios para completar mi formación.

*A mi papá, Pablo  
Que siempre me empujó a hacer realidad todo lo que me imaginaba.*

# Índice

|          |                                                                                     |    |
|----------|-------------------------------------------------------------------------------------|----|
| 1.       | Introducción .....                                                                  | 1  |
| 1.1.     | Cardiopatías congénitas .....                                                       | 2  |
| 1.1.1.   | Generalidades y epidemiología .....                                                 | 2  |
| 1.1.2.   | Clasificaciones de las cardiopatías congénitas .....                                | 2  |
| 1.1.3.   | Etiología de cardiopatías congénitas .....                                          | 5  |
| 1.2.     | Gen <i>NKX2-5</i> .....                                                             | 6  |
| 1.2.1.   | Origen evolutivo .....                                                              | 7  |
| 1.2.2.   | Estructura génica .....                                                             | 8  |
| 1.2.3.   | Estructura proteica .....                                                           | 10 |
| 1.2.4.   | Asociación con defectos congénitos .....                                            | 11 |
| 1.3.     | Abordaje estructural para el estudio de <i>NKX2-5</i> .....                         | 12 |
| 1.3.1.   | Estructuras tridimensionales de proteínas (RX/RMN) .....                            | 12 |
| 1.3.2.   | Estabilidad de estructuras proteicas .....                                          | 14 |
| 1.3.3.   | Proteínas intrínsecamente desestructuradas .....                                    | 16 |
| 1.3.4.   | Predictores de efecto de variantes génicas<br>en la estructura proteica .....       | 19 |
| 1.4.     | Estudio de variantes génicas en <i>NKX2-5</i> y<br>su potencial patogenicidad ..... | 19 |
| 1.4.1.   | Formato HGVS .....                                                                  | 20 |
| 1.4.2.   | Criterio ACMG .....                                                                 | 21 |
| 1.5.     | Interacción de factores de transcripción con otras macromoléculas ...               | 25 |
| 1.5.1.   | Interacción de factores de transcripción con el ADN .....                           | 25 |
| 1.5.1.1. | Búsqueda de sitios de unión .....                                                   | 25 |
| 1.5.1.2. | Confirmación funcional de sitios de unión .....                                     | 26 |
| 1.5.1.3. | Contexto génico de los sitios de unión .....                                        | 28 |
| 1.5.2.   | Interacción de factores de transcripción con otras proteínas ...                    | 30 |
| 1.6.     | Marco teórico .....                                                                 | 31 |

|          |                                                                                     |    |
|----------|-------------------------------------------------------------------------------------|----|
| 2.       | Hipótesis y objetivos .....                                                         | 32 |
| 2.1.     | Antecedentes del laboratorio .....                                                  | 33 |
| 2.2.     | Hipótesis .....                                                                     | 33 |
| 2.3.     | Objetivos .....                                                                     | 34 |
| 2.3.1.   | Objetivo general .....                                                              | 34 |
| 2.3.2.   | Objetivos específicos .....                                                         | 34 |
| 3.       | Materiales y métodos .....                                                          | 35 |
| 3.1.     | Generación de modelo estructural de NKX2-5 .....                                    | 36 |
| 3.1.1.   | Selección de estructuras tridimensionales<br>correspondientes a NKX2-5 .....        | 38 |
| 3.1.2.   | Estudio de regiones con estructura tridimensional .....                             | 38 |
| 3.1.2.1. | Conservación evolutiva de residuos de región<br>con estructura tridimensional ..... | 40 |
| 3.1.3.   | Búsqueda de motivos lineales en regiones accesibles .....                           | 40 |
| 3.1.4.   | Potenciales interacciones con secuencias de ADN .....                               | 41 |
| 3.2.     | Recopilación y análisis de todas las variantes génicas<br>conocidas de NKX2-5 ..... | 42 |
| 3.2.1.   | Actualización de lista de variantes .....                                           | 43 |
| 3.2.1.1. | Actualización de VGs en publicaciones científicas .....                             | 44 |
| 3.2.1.2. | Actualización de VGs en bases de datos públicas .....                               | 44 |
| 3.2.2.   | Procesamiento y agregado de información de VGs .....                                | 44 |
| 3.2.2.1. | Definición de variantes que afectan motivos lineales .....                          | 45 |
| 3.2.2.2. | Frecuencias poblacionales .....                                                     | 45 |
| 3.2.3.   | Evaluación de patogenicidad de variantes génicas .....                              | 46 |
| 3.2.3.1. | Patologías específicas asociadas con<br>VGs de NKX2-5 .....                         | 46 |
| 3.2.4.   | Formato de la información en la lista de VGs .....                                  | 47 |
| 3.2.5.   | Carga de VGs a bases de datos .....                                                 | 48 |
| 3.2.6.   | Generación de scripts para visualizar VGs sobre NKX2-5 .....                        | 49 |

|          |                                                                                                        |    |
|----------|--------------------------------------------------------------------------------------------------------|----|
| 3.3.     | Análisis de regiones específicas de NKX2-5 .....                                                       | 50 |
| 3.3.1.   | Creación de todas las VGs missense posibles<br>en la región seleccionada de <i>NKX2-5</i> ... ..       | 52 |
| 3.3.2.   | Estudios <i>in silico</i> para cada VG generada .....                                                  | 52 |
| 3.3.3.   | Clasificación de potenciales efectos de las variantes<br>generadas en la funcionalidad de NKX2-5 ..... | 53 |
| 3.4.     | Recopilación y estudio de todos los sitios de unión<br>de NKX2-5 en el genoma humano y de ratón .....  | 55 |
| 3.4.1.   | Selección de resultados .....                                                                          | 57 |
| 3.4.2.   | Recopilación de sitios de unión de NKX2-5 .....                                                        | 57 |
| 3.4.3.   | Estudio de asociación de sitios de unión de NKX2-5<br>cerca de genes con expresión diferencial .....   | 58 |
| 3.5.     | Estudio de sitios de unión cercanos de otros factores<br>de transcripción y genes transactivados ..... | 60 |
| 3.5.1.   | Sitios de unión de otros factores de transcripción<br>cercanos a sitios de NKX2-5 .....                | 62 |
| 3.5.2.   | Análisis de genes cercanos a sitios de unión seleccionados ....                                        | 64 |
| 3.5.3.   | Análisis de genes seleccionados en relación<br>a subtipos de CC .....                                  | 66 |
| 4.       | Resultados .....                                                                                       | 68 |
| 4.1.     | Modelo estructural de NKX2-5 .....                                                                     | 70 |
| 4.1.1.   | Selección de estructuras tridimensionales<br>correspondientes a NKX2-5 .....                           | 70 |
| 4.1.2.   | Estudio de regiones con estructura terciaria .....                                                     | 71 |
| 4.1.2.1. | Conservación evolutiva de residuos de región<br>con estructura tridimensional .....                    | 75 |
| 4.1.3.   | Búsqueda de motivos lineales en regiones accesibles .....                                              | 76 |
| 4.1.4.   | Potenciales interacciones con secuencias de ADN .....                                                  | 78 |

|          |                                                                                                             |     |
|----------|-------------------------------------------------------------------------------------------------------------|-----|
| 4.2.     | Recopilación y análisis de todas las variantes génicas conocidas de <i>NKX2-5</i> .....                     | 78  |
| 4.2.1.   | Distribución de variantes totales y patogénicas a lo largo de <i>NKX2-5</i> .....                           | 79  |
| 4.2.1.1. | Distribución de variantes que afectan motivos lineales ...                                                  | 82  |
| 4.2.1.2. | Distribución de variantes con información de frecuencia poblacional .....                                   | 83  |
| 4.2.1.3. | Distribución de variantes patogénicas en <i>NKX2-5</i> .....                                                | 84  |
| 4.2.2.   | Evaluación de patogenicidad de variantes génicas .....                                                      | 85  |
| 4.2.3.   | Distribución de subtipos de CC .....                                                                        | 86  |
| 4.3.     | Análisis de todas las variantes de cambio de sentido en el homeodominio de <i>NKX2-5</i> .....              | 90  |
| 4.3.1.   | Generación de todas las VGs de cambio de sentido posibles en la región del homeodominio .....               | 90  |
| 4.3.2.   | Estudios <i>in silico</i> para cada VG generada .....                                                       | 91  |
| 4.3.3.   | Clasificación de potenciales efectos de VGs de cambio de sentido en la funcionalidad de <i>NKX2-5</i> ..... | 92  |
| 4.4.     | Recopilación y estudio de todos los sitios de unión de <i>NKX2-5</i> en el genoma humano y de ratón .....   | 96  |
| 4.4.1.   | Selección de resultados .....                                                                               | 96  |
| 4.4.2.   | Recopilación de sitios de unión de <i>NKX2-5</i> .....                                                      | 98  |
| 4.4.3.   | Estudio de asociación de sitios de unión de <i>NKX2-5</i> con regulación río abajo .....                    | 100 |
| 4.5.     | Estudio de sitios de unión cercanos de otros factores de transcripción y genes transactivados .....         | 104 |
| 4.5.1.   | Sitios de unión de otros factores de transcripción cercanos a sitios de <i>NKX2-5</i> .....                 | 104 |
| 4.5.2.   | Análisis de genes cercanos a sitios de unión seleccionados ...                                              | 106 |
| 4.5.3.   | Análisis de genes seleccionados en relación a subtipos de CC .....                                          | 112 |
| 4.6.     | Flujo de trabajo replicable .....                                                                           | 114 |

|      |                                                                 |     |
|------|-----------------------------------------------------------------|-----|
| 5.   | Discusión .....                                                 | 119 |
| 5.1. | Estudio de VGs de <i>NKX2-5</i> .....                           | 121 |
| 5.2. | Interacción de <i>NKX2-5</i> con el ADN y otras proteínas ..... | 129 |
| 6.   | Conclusiones .....                                              | 146 |
| 7.   | Bibliografía ..                                                 | 149 |
| 8.   | Anexo .....                                                     | 198 |
|      | Anexo I. Enlaces externos .....                                 | 199 |
|      | Anexo II. Tablas y figuras suplementarias .....                 | 201 |

# **1. Introducción**

## **1.1. Cardiopatías congénitas**

### **1.1.1. Generalidades y epidemiología**

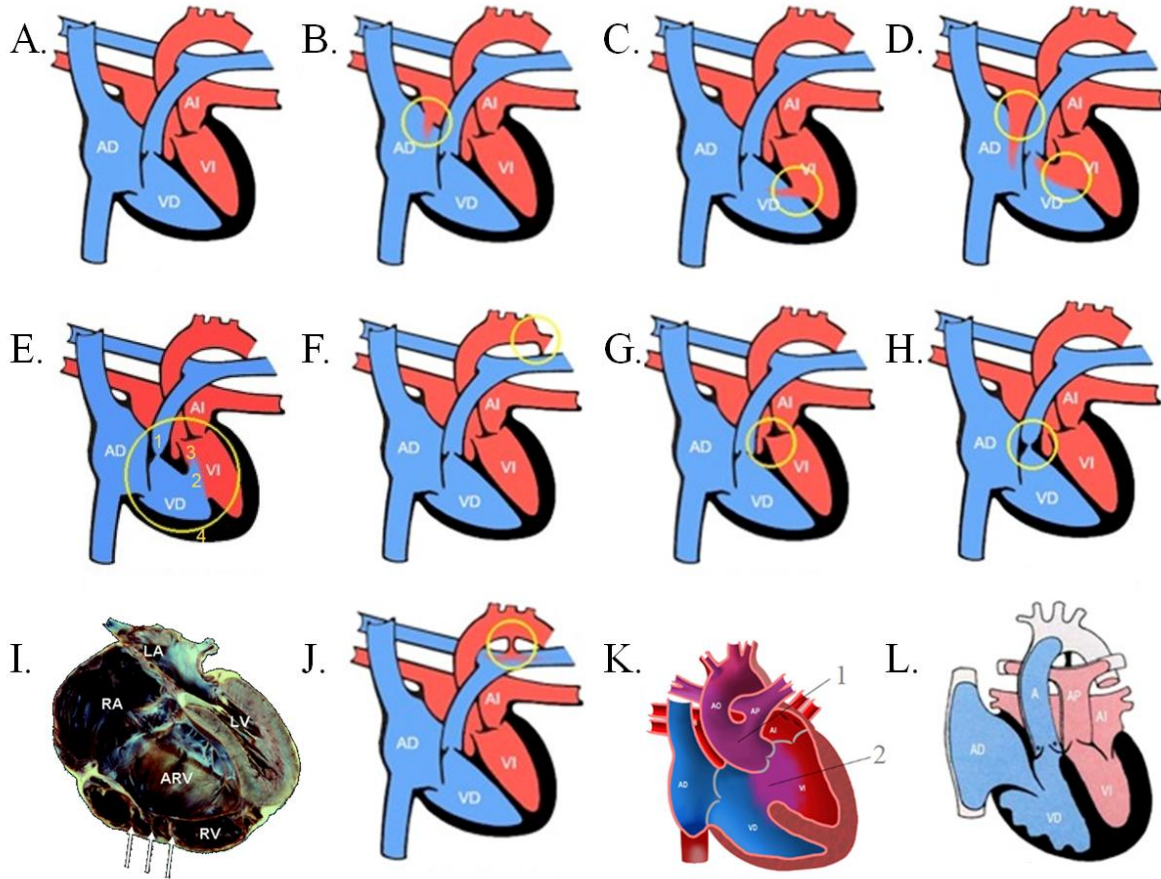
Las cardiopatías congénitas (CC) son anomalías que se manifiestan desde el nacimiento, se caracterizan por el desarrollo irregular del corazón en el período embrionario-fetal y afectan entre 6 y 9 de cada 1000 nacimientos [1-3]. Abarcan un amplio espectro de defectos de cavidades cardíacas y/o grandes vasos que incluyen tanto defectos sin significancia clínica como defectos complejos que conllevan un riesgo para la vida. La mayoría son toleradas intrauterinamente y sólo se manifiestan después del nacimiento [3], por lo que es de gran interés su detección temprana [4].

El tipo más común de CC es la comunicación intraventricular (CIV), seguido de la comunicación interauricular (CIA) y el ductus arterioso persistente (DAP). La prevalencia de estas lesiones, comúnmente consideradas “leves”, ha aumentado en las últimas décadas. Esto posiblemente se deba a avances que permiten una detección más precisa, mientras que las lesiones graves mantienen una prevalencia relativamente constante [2,5].

Las CC pueden ocurrir como un defecto de nacimiento aislado o en conjunto con anomalías extracardíacas, incluyendo malformaciones craneofaciales, otras anomalías congénitas, anomalías en el desarrollo y/o retraso en el crecimiento [6,7]. Diversos estudios estiman que el riesgo general de recurrencia es del 2 al 10% [1], con una variación según la malformación en cuestión, lo que indicaría que algunas anomalías poseen un mayor componente hereditario que otras [1].

### **1.1.2. Clasificaciones de las cardiopatías congénitas**

Existen diversas clasificaciones de las CC que tienen en cuenta distintos factores y utilizan distintos criterios de clasificación [8]. En la Figura 1 se pueden ver esquematizados corazones con algunos subtipos de CC.



**Figura 1. Esquemas de distintas cardiopatías congénitas (CC).** **A.** Corazón normal. **B.** Comunicación interauricular (CIA). **C.** Comunicación interventricular (CIV). **D.** Canal auriculoventricular (Canal AV). **E.** Tetralogía de Fallot; (1) Estenosis pulmonar, (2) CIV, (3) Dextroposición de la aorta; (4) Hipertrofia ventricular derecha. **F.** Coartación de la aorta. **G.** Estenosis aórtica. **H.** Estenosis pulmonar. **I.** Anomalia de Ebstein. **J.** Ductus arterioso persistente (DAP). **K.** Truncus arterioso persistente (TAP); (1) Continuación de la arteria pulmonar y la aorta, (2) Abertura entre los ventrículos. **L.** Transposición de grandes vasos (TGV). Fuentes: **A-H, J y L,** <http://www.dra-amor.com/styled-10/index.html>; **I,** Jost y col., 2007; **K,** Stanford Children's Hospital (<http://www.stanfordchildrens.org>).

Desde el punto de vista clínico, una de las clasificaciones más usadas es aquella que considera si la CC es acianótica o cianótica [9]. Este tipo de clasificación, propuesta por la *Society of Thoracic Surgeons National Congenital Heart Surgery Database Committee*, 2000 [10], se muestra en la Tabla 1.

**Tabla 1. Clasificación de las cardiopatías congénitas según la *Society of Thoracic Surgeons National, Congenital Heart Surgery Database Committee*, 2000.**

|                                            |                                                                                                   |                                                       |                                             |
|--------------------------------------------|---------------------------------------------------------------------------------------------------|-------------------------------------------------------|---------------------------------------------|
| <b>Cardiopatías congénitas acianóticas</b> | Con desviaciones del flujo sanguíneo de izquierda a derecha                                       | Comunicación interauricular (CIA)                     |                                             |
|                                            |                                                                                                   | Comunicación interventricular (CIV)                   |                                             |
|                                            |                                                                                                   | Defecto del canal auriculoventricular completo (CAVC) |                                             |
|                                            |                                                                                                   | Ventana aortopulmonar (Ventana AP)                    |                                             |
|                                            |                                                                                                   | Ductus arterioso persistente (DAP)                    |                                             |
|                                            | Con lesiones obstructivas del lado izquierdo                                                      | Coartación de la aorta (CoA)                          |                                             |
|                                            |                                                                                                   | Estenosis aórtica congénita                           |                                             |
|                                            |                                                                                                   | Interrupción del arco aórtico (IAA)                   |                                             |
|                                            |                                                                                                   | Estenosis mitral                                      |                                             |
| <b>Cardiopatías congénitas cianóticas</b>  | Lesiones asociadas con disminución de la circulación pulmonar (desviación de derecha a izquierda) | Tetralogía de Fallot (TDF)                            |                                             |
|                                            |                                                                                                   | Estenosis de la válvula pulmonar (EP)                 |                                             |
|                                            |                                                                                                   | Atresia pulmonar (AP)                                 | Con septo ventricular intacto (AP/SVI)      |
|                                            |                                                                                                   |                                                       | Con comunicación interventricular (AP/CIV)  |
|                                            |                                                                                                   | Atresia tricuspídea                                   |                                             |
|                                            |                                                                                                   | Anomalía de Ebstein (AE)                              |                                             |
|                                            | Lesiones asociadas con incremento de circulación pulmonar (mezcla de sangre completa)             | Transposición de grandes vasos (TGV)                  | Con septo ventricular intacto               |
|                                            |                                                                                                   |                                                       | Con comunicación interventricular (TGV/CIV) |
|                                            |                                                                                                   | Doble salida del ventrículo derecho (DSVD)            |                                             |
|                                            |                                                                                                   | Conexión venosa pulmonar anómala                      |                                             |
|                                            |                                                                                                   | Trunco arterioso persistente (TAP)                    |                                             |
|                                            | Fisiología de un ventrículo                                                                       | Síndrome del corazón izquierdo hipoplásico (SCIH)     |                                             |
|                                            |                                                                                                   | Doble entrada del ventrículo izquierdo (DEVI)         |                                             |

Otra forma de clasificar las CC es teniendo en cuenta su origen [11]. En esta clasificación se identifican 4 factores principales como los responsables de la mayoría de las cardiopatías congénitas: problemas en la migración del tejido mesenquimal, problemas de hemodinámica, problemas en la regulación de la muerte celular, anomalías en la matriz celular.

Las CC también se pueden clasificar según la parte del corazón que afectan, y los dos tipos más comunes de CC en esta clasificación son las cardiopatías septales y las cardiopatías conotruncales. Los defectos septales son las anomalías más frecuentes, constituyendo el 50% de las CC. Estos defectos incluyen: la comunicación interventricular, la comunicación interauricular y el canal aurículo-ventricular. Pueden estar asociados a síndromes de origen génico y cromosómico, pero constituyen la mayor proporción de los defectos cardíacos aislados [12]. Las anomalías conotruncales afectan la formación de los tractos de salida y el arco aórtico, entre las que se incluyen la Tetralogía de Fallot (TDF), la atresia pulmonar con comunicación interventricular, la transposición de grandes vasos (TGV), el ventrículo derecho con doble salida, el tronco arterioso persistente y la interrupción del arco aórtico. Las más frecuentes son TDF y TGV. Constituyen del 20 al 30% de las CC y son la primera causa de cianosis de origen cardíaco durante el primer año de vida [13].

#### 1.1.3. Etiología de cardiopatías congénitas

En la mayoría de los casos de CC no se puede establecer causalidad directa. Alrededor de 60% de los casos son considerados de etiología desconocida o multifactorial, probablemente afectados por factores tanto genéticos como ambientales [14,15].

Las causas no genéticas asociadas a CC incluyen teratógenos ambientales, exposición a ciertas drogas y/o medicamentos, agentes infecciosos y otros factores de la salud personal [15-22].

Por otro lado, aún en casos sin causalidad directa establecida, se ha observado recurrencia familiar en el desarrollo de CC, por lo que se estima que existe gran influencia de factores genéticos [1,14,23].

En los casos en los que se puede asociar factores genéticos al desarrollo de CC, las anomalías cromosómicas se detectan entre el 6% y el 14% de los pacientes, principalmente en casos que presentan alguna otra malformación además de la cardíaca [24]. Otros pacientes poseen microdeleciones o microduplicaciones, comúnmente denominadas CNVs (siglas en inglés para *Copy Number Variations*), como causa de su

afección. Diferentes estudios estiman una prevalencia de CNVs en CC que oscila entre el 5 y el 30% de los afectados [15,25-27].

En aproximadamente 5-10% de los pacientes, las CC serían el producto de la alteración en un único gen. Se estima que hay alrededor de 400 genes asociados a CC y que este número puede seguir en aumento en futuros estudios [15,22]. En general, a los genes implicados se los relaciona a una de las siguientes tres categorías: regulación de la transcripción, transducción de señales o proteínas estructurales cardíacas. Algunos ejemplos incluyen a los genes *NKX2-5*, *NKX2-6*, *GATA4*, *GATA5*, *GATA6*, *IRX4*, *TBX20*, *TBX5*, *NOTCH1*, *NOTCH2*, *ELN*, entre otros [22,28-33].

Es destacable que diversos estudios en distintos genes con variantes génicas (VGs) asociadas a CC mostraron que una misma variante se asocia con diferentes malformaciones cardíacas en distintos individuos y que variantes de distintos genes se asocian a un mismo tipo de malformación cardíaca [29]. Aún así, la distribución de subtipos de CC asociados a factores de transcripción en particular tienen diferencias y tendencias de asociación a distintos subtipos [34-37].

En particular, el gen *NKX2-5* (OMIM: \*600584; UniProt\_ID: P52952) fue el primer factor de transcripción para el cual se observó una asociación entre variantes génicas y CC [32,38,39]. Desde entonces y hasta el día de hoy se hallaron numerosas variantes en *NKX2-5* asociadas a CC [40].

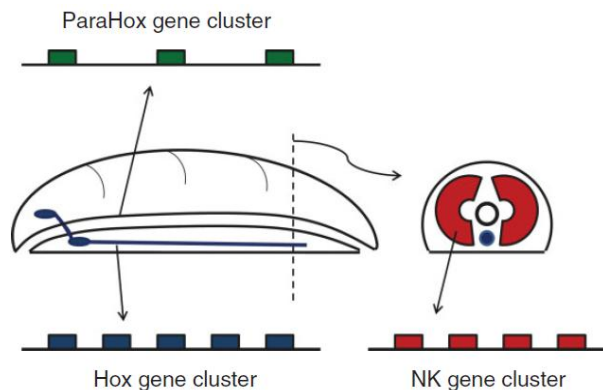
## **1.2. Gen *NKX2-5***

El gen *NKX2-5*, también conocido como *CSX*, se encuentra ubicado en el cromosoma 5, en la región cromosómica 5q34 [41]. Se expresa mayoritariamente en tejido cardíaco desde el período fetal y juega un rol significativo en el desarrollo del nodo auriculoventricular, así como en el mantenimiento de su función a lo largo de la vida adulta [41,42]. También se expresa en tiroides [43] y el bazo [44,45] durante su desarrollo.

### 1.2.1. Origen evolutivo

*NKX2-5* pertenece a la familia NK de genes homeobox. Está altamente conservado en diferentes especies y se encuentra presente desde moscas a humanos [46,47]. En particular, el homeodominio, su región más conservada, tiene un 100% de identidad de secuencia proteica y 94% de identidad para su secuencia de ADN correspondiente entre *NKX2-5* humano y *Nkx2-5* de ratón [47,48].

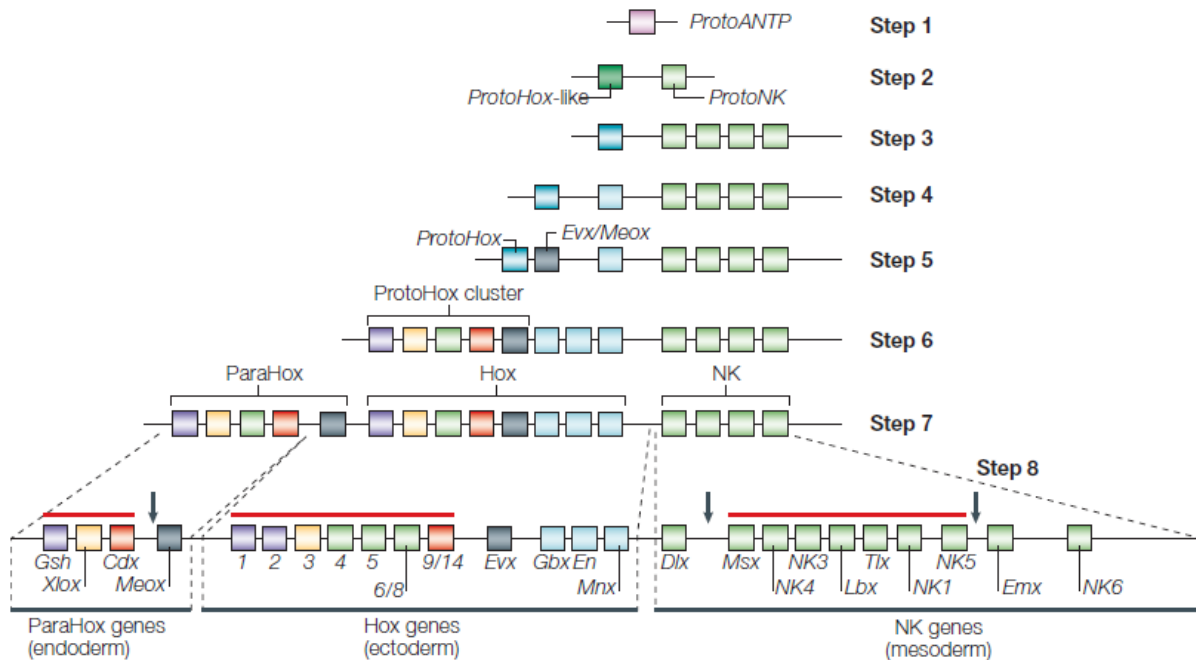
La familia NK comprende un grupo de genes asociados a la diferenciación celular y regulación del tejido mesodérmico. Los genes de regulación del ectodermo (Genes HOX), mesodermo (Genes NK) y endodermo (Genes ParaHox) tienen un origen evolutivo común en el linaje de los animales (Figura 2) [49].



**Figura 2. Papel hipotético de los *clusters* Hox, ParaHox y NK homeobox en el patrón de tejidos neurales, intestinales y mesodérmicos en un animal bilateral temprano.** Cluster Hox en azul, cluster NK en rojo, cluster ParaHox en verde.

Se cree que el primer gen precursor de las tres familias génicas fue un gen de la clase homeobox-ANTP (Figura 3). En un primer paso, el gen *ANTP* ancestral se duplicó, dando lugar a un gen similar a *ProtoHox* y a un gen *ProtoNK*. A través de una serie de duplicaciones en tándem, cada familia se expandió para formar un clúster NK (Figura 3, color verde) y una serie de genes similares a Hox (Figura 3, pasos 3-6, varios colores) que conducen finalmente (Figura 3, paso 7) a la generación de los *clústeres* Hox y ParaHox, incluidos *Evx* y *Meox* (Figura 3, color gris), así como los genes Hox extendidos (Figura 3, color azul claro).

Se estima que en el último ancestro común entre protostomados y deuterostomados se produjo la composición particular de un *megacluster*. Mediante las rupturas cromosómicas, el *megacluster* quedó separado en al menos tres regiones, dando tres bloques principales que se conocen como las regiones ParaHox, Hox y NK (flechas de la Figura 3) [50].



**Figura 3. El *megacluster* homeobox.**

Modelo para la génesis del *megacluster* homeobox clase ANTP.

A su vez, a lo largo de la historia evolutiva de los deuterostomados, los genes NK fueron cambiando su ubicación cromosómica. En la actualidad, nuestra especie tiene los genes NK distribuidos en 5 cromosomas, entre los que se incluye el cromosoma 5, donde se encuentra *NKX2-5*.

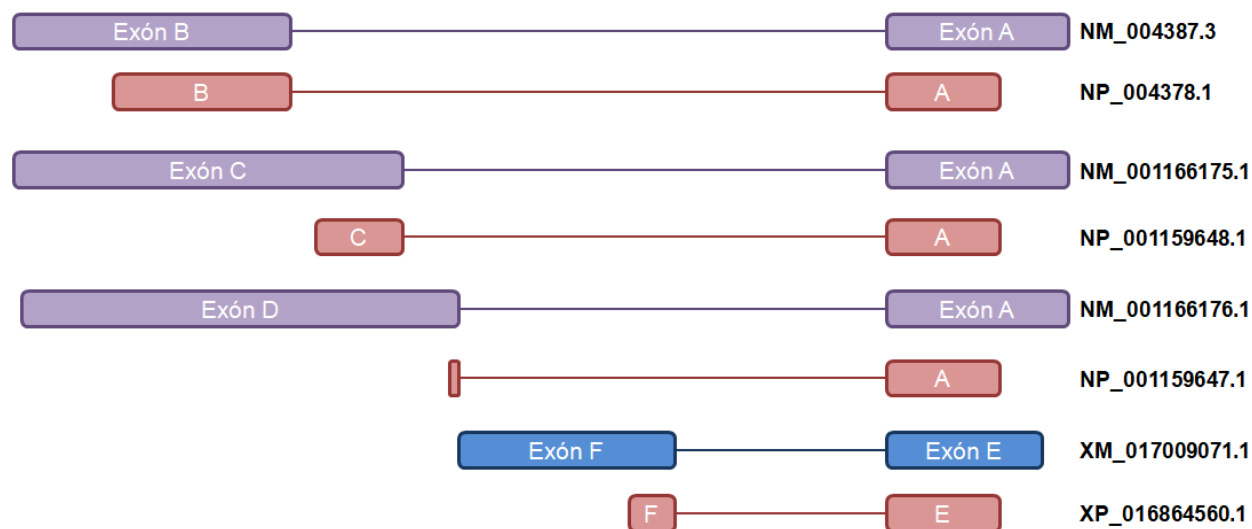
### 1.2.2. Estructura génica

El gen *NKX2-5* contiene cuatro exones diferentes que codifican para tres proteínas distintas [51,52], más dos exones predichos codificantes para una proteína (XM\_017009071.1 y XP\_016864560.1). Cada una de las isoformas consiste en dos exones separados por un intrón (Figura 4). El gen tiene un complejo conjunto de regiones reguladoras, incluyendo *enhancers* en la región promotora, proximales y

distales, con regulación tanto de otros genes (familia de proteínas GATA, proteínas SMAD) como con autorregulación por NKX2-5 [53-59].

De las cuatro isoformas, las observadas experimentalmente corresponden a la misma secuencia transcripta pero tienen variación en la región de *splicing* (NM\_004387.3, NM\_001166175.1 y NM\_001166176.1), mientras que la isoforma predicha por métodos computacionales corresponde a un marco de transcripción más corto (XM\_017009071.1). Las cuatro isoformas comparten los primeros 111 aminoácidos.

La isoforma con más evidencia experimental de su funcionalidad y contexto de expresión es NM\_004387.3, que codifica para la proteína más larga de las 4 isoformas (secuencia proteica NP\_004378.1) y es la considerada “canónica”. Las otras tres isoformas codifican para proteínas sustancialmente más cortas (NM\_001166175.1, NM\_001166176.1 y XM\_017009071.1) y tienen poca o nula evidencia de funcionalidad y contexto de expresión (<https://www.uniprot.org/uniprotkb/P52952/entry#sequences>).



**Figura 4. Esquema de los exones del gen NKX2-5 (violeta y azul) y las proteínas codificadas por cada grupo de exones (en rojo, debajo de sus respectivos exones).** En recuadros, regiones del ARN mensajero, en líneas, regiones intrónicas eliminadas por *splicing*. Los tamaños son relativos y no proporcionales a las distancias en el genoma. Los exones en violeta corresponden a variantes de *splicing* sin cambios en inicio ni finalización de transcripción. Los exones en azul corresponden a una variante de transcripción predicha por modelos computacionales. El exón E y el exón A codifican para la misma secuencia aminoacídica. Fuente: <https://www.ncbi.nlm.nih.gov/variation/view/>.

### 1.2.3. Estructura proteica

La isoforma seleccionada para este trabajo (NM\_004387.3) codifica para una proteína que posee 324 aminoácidos y tiene 6 regiones características (Figura 5):

- Homeodominio (HD): El homeodominio de NKX2-5 se ubica entre los aminoácidos 138-197 y consiste de un motivo hélice-vuelta-hélice que reconoce una secuencia consenso 5' T(C/T)AAGTG 3' en el ADN [60,61]. También interactúa con diversas proteínas (GATA4, SRF, TBX5) y es esencial para la dimerización de NKX2-5 [61].
- Dominio específico de proteínas NK-2 (NK2-SD): El dominio específico de proteínas NK-2 en NKX2-5 se ubica entre los aminoácidos 212-234 y funciona como represor transcripcional, según lo que se observó en ensayos *in vitro* [61].
- Región rica en tirosinas (YRR): Región de NKX2-5 con abundancia de tirosinas que cumple la función de interfaz de dimerización [47,62,63].
- Dominio tinman (TN): El dominio tinman de NKX2-5 es una región conservada que se ubica entre los aminoácidos 10-21 y su función no fue elucidada [61].
- Dominio Nkx2-5 box: Dominio conservado que tiene evidencia de ser importante para la transactivación mediada por la región C-terminal de NKX2-5 [62,64].
- Motivo GIRAW: Es un motivo de 5 aminoácidos ubicados en el extremo C-terminal que puede tener un rol en interacciones con otras proteínas [64,65].



**Figura 5. Esquema de distribución de regiones características de NKX2-5.** TN: Dominio tinman. HD: Homeodominio. NK2-SD: Dominio específico de proteínas NK-2. YRR: Región rica en tirosinas.

También se han identificado varios sitios funcionalmente importantes. En particular:

- Una señal de localización nuclear [66]
- Un motivo de SUMOylación [67]
- Un motivo de unión de caseína quinasa 2 (CK2) [66]

Por último, en un trabajo previo de nuestro laboratorio [68] se confirmó que el homeodominio es la única región de NKX2-5 con estructura tridimensional (registrada en Protein Data Bank, [www.rcsb.org](http://www.rcsb.org)) [69] y se utilizaron programas *in silico* para determinar que toda la región fuera del homeodominio se predice como intrínsecamente desordenada (ver secciones 1.3.1. y 1.3.3.).

#### 1.2.4. Asociación con defectos congénitos

Como ya se mencionó anteriormente, *NKX2-5* juega un rol significativo en el desarrollo del nodo auriculoventricular, así como en el mantenimiento de su función a lo largo de la vida adulta, y es el primer factor de transcripción para el cual se observó una asociación entre variantes génicas y CC. También se expresa en el desarrollo de la tiroides y del bazo. Se ha visto que variantes de *NKX2-5* (o en otros genes que afectan la expresión de *NKX2-5* en el período fetal) se encuentran relacionadas con asplenia [70], hipotiroidismo [71,72], heterotaxia [73,74] y, mayoritariamente, con el desarrollo de distintos tipos de CC [22,40].

Particularmente para el caso de CC, se han visto variantes de *NKX2-5* asociadas con distintos subtipos de cardiopatías. Algunos de los defectos comúnmente vistos en asociación a variantes de *NKX2-5* son comunicación interauricular (CIA), bloqueo auriculoventricular (BAV), comunicación interventricular (CIV) y tetralogía de Fallot (TDF) [32,42].

En el trabajo previo de nuestro laboratorio [68] se registraron y procesaron 877 VGs, de las cuales 132 se encontraron en pacientes con CC, y se observó que la gran mayoría (n=115) afectan directamente a la secuencia de la proteína codificada. Estas observaciones son similares a lo consignado por otros autores [32,40], y sugiere una

gran importancia de la estructura proteica de NKX2-5 en el desarrollo de la patogénesis.

### **1.3. Abordaje estructural para el estudio de NKX2-5**

Una de las áreas donde la disponibilidad de información creció enormemente en los últimos años es en el análisis de estructuras tridimensionales de proteínas. Esto se debe tanto a mejoras en la obtención de estructuras de forma experimental (principalmente mediante difracción de rayos X, resonancia magnética nuclear y más recientemente crio-electro microscopía) como a avances en modelado molecular y predicción de plegamiento *in silico*. Este incremento en la cantidad de datos disponibles para el análisis de la estructura tridimensional conlleva un aumento de la potencial información que puede ser extraída, lo cual constantemente abre oportunidades para nuevos estudios.

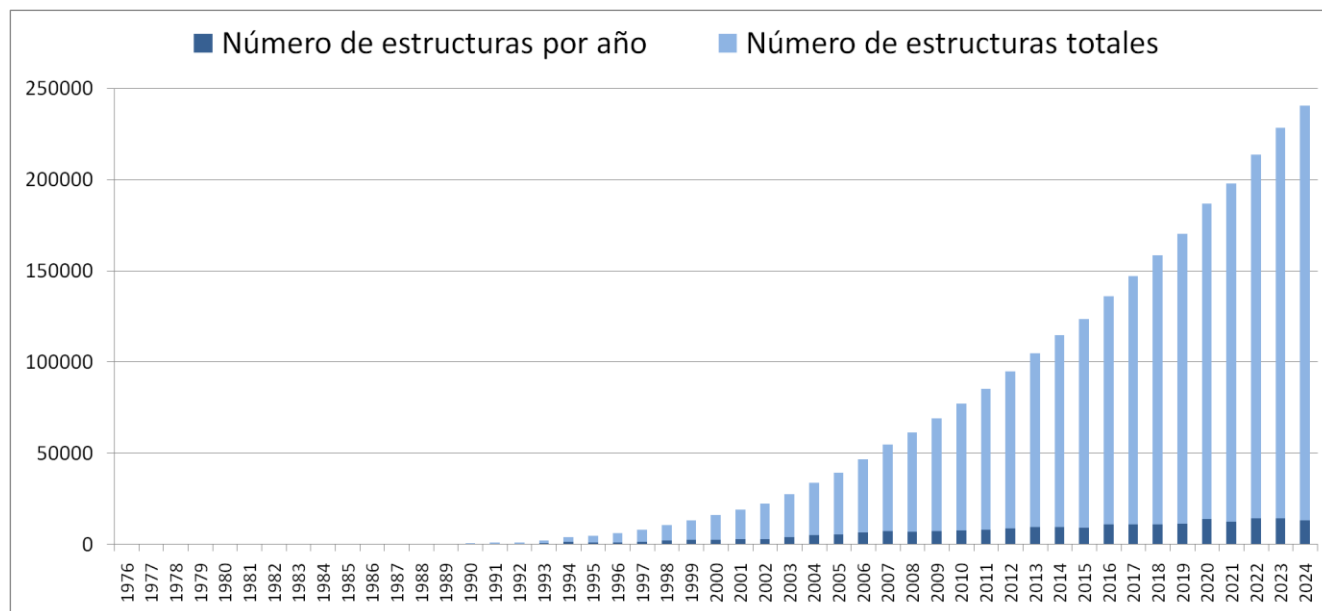
#### **1.3.1. Estructuras tridimensionales de proteínas (RX/RMN)**

Una de las principales bases de datos que recopila información sobre estructuras tridimensionales es Protein Data Bank (PDB, [www.rcsb.org](http://www.rcsb.org)) [69], y su evolución recapitula el incremento del volumen de información y las medidas utilizadas para sobrellevar las dificultades que este incremento conlleva.

PDB se creó en 1971 contando en un principio con sólo 7 estructuras. Con el tiempo, este número (y la complejidad de cada modelo) creció exponencialmente, ya que se desarrollaron nuevos métodos de producción y cristalización de proteínas aisladas, así como nuevos métodos de obtención de datos y análisis de estructuras (Figura 6). El desarrollo de resonancia magnética nuclear (RMN) para determinación de estructura proteica y la criomicroscopía de electrones son dos de los grandes avances que marcaron esta etapa [75].

En la década del 1980, muchas revistas científicas comenzaron a solicitar que se depositen las estructuras tridimensionales publicadas en PDB, lo cual incrementó aún más el caudal de estructuras ingresadas a la base de datos.

A principios de 2023, PDB alcanzó un total de 200.000 registros (Figura 6) y al día de hoy cuenta con más de 225.000 registros (<https://www.rcsb.org/stats/growth/growth-released-structures/>).



**Figura 6. Evolución de la cantidad de estructuras depositadas en PDB desde 1976 hasta noviembre de 2024.** En azul claro, número de estructuras total en cada año. En azul oscuro, número de estructuras agregadas en cada año. Fuente: <https://www.rcsb.org/stats/growth/growth-released-structures/>.

A pesar de que el número de estructuras nuevas agregadas a PDB incrementa cada año, la mayoría son similares a las ya depositadas [76]. Aunque esto vuelve imprácticos los estudios manuales a gran escala, abre la puerta a estudios más específicos sobre proteínas individuales, permitiendo, por ejemplo, estudiarlas en contacto con distintas macromoléculas y modelar estructuras no determinadas a partir de una gran cantidad de homólogos.

Más recientemente, los modelos de aprendizaje automático como AlphaFold [77] han permitido predecir la estructura tridimensional de proteínas con precisión muy superior a predictores anteriores, en parte gracias a esta gran cantidad de estructuras confirmadas experimentalmente. Esta herramienta daría una nueva perspectiva a la predicción de estructuras tridimensionales en regiones no cristalizadas, aunque su

publicación reciente implicó que no se pudo implementar en el flujo de trabajo. Aún así, su relevancia a los resultados de este trabajo y los futuros se discute en la sección 5.2.

### 1.3.2. Estabilidad de estructuras proteicas

La estabilidad estructural de las proteínas es extremadamente importante para su función [78]. De hecho, cambios relativamente pequeños en la estabilidad pueden favorecer el estado desestructurado de una proteína o afectar su funcionalidad aún si mantiene su estructura.

Por esto, las estructuras tridimensionales de regiones estructuradas de una proteína pueden dar información respecto de potenciales efectos funcionales de variantes génicas. Esta información se puede obtener, por ejemplo, mediante el cálculo de la diferencia de energía libre ( $\Delta\Delta G$ ) de plegado entre la proteína *wild type* y la proteína con aminoácidos mutados [78].

La estabilidad de las estructuras proteicas está estrechamente asociada con las interacciones intra- e intermoleculares que la proteína pueda tener. Ambas interacciones se pueden estudiar con mayor robustez dada la secuencia proteica y una estructura tridimensional asociada.

El correcto plegamiento de proteínas es comúnmente entendido como dependiente de una serie de interacciones covalentes (estructura primaria y eventuales puentes disulfuro) y no covalentes que dan estabilidad, en contra de la inestabilidad dada por interacciones desfavorables y la entropía conformacional de la proteína.

Las interacciones no covalentes comúnmente consideradas como las más relevantes comprenden: El efecto hidrofóbico, interacciones de van der Waals, interacciones mediadas por puentes de hidrógeno e interacciones de Coulomb [79,80].

El efecto hidrofóbico se produce por la tendencia de moléculas hidrofóbicas a agruparse al encontrarse en un medio hidrofílico [81]. Tiene una contribución importante a la estabilidad de la estructura estable de proteínas, siendo más importante mientras más grande es la región estructurada en cuestión [82]. El núcleo hidrofóbico de una estructura proteica es conocido como el *core* hidrofóbico de la proteína. Las

interacciones entre moléculas hidrofóbicas se conocen como interacciones de van der Waals, y suelen encontrarse dentro del *core* hidrofóbico de una proteína o en regiones de interacción con otras moléculas hidrofóbicas, contribuyendo a su estabilidad [80].

A diferencia de las interacciones entre moléculas hidrofóbicas, las interacciones mediadas por puentes de hidrógeno y las interacciones de Coulomb ocurren entre moléculas hidrofílicas. Esto no significa que no se encuentren presentes en el *core* hidrofóbico, pero son más comunes en la superficie de contacto de la proteína con el solvente, y no siempre es claro si contribuyen a la estabilidad del *core* hidrofóbico [80].

Las interacciones mediadas por puentes de hidrógeno son causadas por fuerzas electrostáticas entre dipolos (terminaciones -OH, -NH, -C=O) y suelen contribuir favorablemente a la estabilidad de la proteína. Tienen una distancia entre partes de 2,5-3,5 Angstroms y, por su naturaleza, pueden estar mediadas por moléculas de agua [83].

Por otro lado, las interacciones de Coulomb consisten en la atracción (o repulsión) de moléculas según su carga a un determinado pH, por lo que dependen de la presencia de aminoácidos que puedan tener carga a pH fisiológicos (Asp, Glu, His, Lys y Arg), o de aminoácidos con modificaciones post-traduccionales (fosforilación de Ser, Thr y Tyr y acetilación de Lys) [84]. Las cargas en la superficie de una proteína suelen ser principalmente atractivas a pH neutro [85,86], y cambios en estas cargas pueden desestabilizar la proteína con mucha facilidad. Pueden ocurrir entre dos moléculas cargadas o entre una molécula cargada y un dipolo, y su distancia de acción es mayor a la de las interacciones mediadas por puentes de hidrógeno, llegando a ser relevantes a 5-10 Angstroms de distancia [84].

En cuanto a interacciones covalentes, las proteínas con cisteínas pueden formar interacciones covalentes en forma de puentes disulfuro entre cisteínas. Estas interacciones pueden alterar significativamente la estabilidad y movilidad de la proteína, y nuevas cisteínas aumentan de forma multiplicativa la cantidad de puentes disulfuro que se pueden formar. Esto puede aumentar la estabilidad de la proteína por la

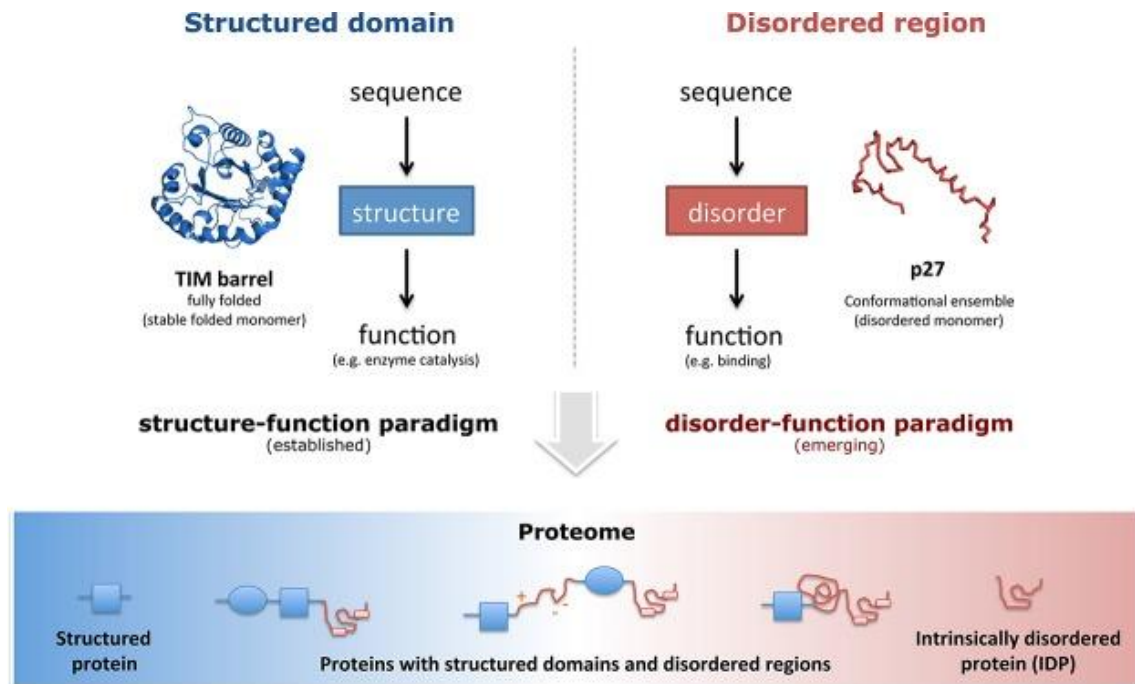
formación de los puentes disulfuro o disminuir la entropía conformacional del estado desplegado y favorecer conformaciones no funcionales [87,88].

Por último, la fuerza que más comúnmente actúa desestabilizando la estructura de una proteína es la entropía conformacional [80]. Esta aumenta significativamente con la introducción de residuos pequeños, particularmente glicina, y puede desestabilizar completamente a proteínas pequeñas.

### 1.3.3. Proteínas intrínsecamente desestructuradas

El paradigma clásico para explicar la funcionalidad de las proteínas consistía en asumir que estas únicamente actúan en un estado de plegado altamente ordenado [89]. Este plegado es el que definiría qué función cumple la proteína y se estipulaba que requieren de una estructura terciaria definida para que pudieran ser funcionales. Sin embargo, existen proteínas, o dominios dentro de estas proteínas, que no necesitan de una estructura terciaria estable para cumplir su función (Figura 7) [90].

Las proteínas intrínsecamente desestructuradas (PID) se caracterizan por la falta de una estructura terciaria estable cuando se encuentran en forma de cadena peptídica aislada bajo condiciones fisiológicas *in vitro* [91,92]. Poseen una función fisiológica sin una estructura terciaria definida, si bien en algunos casos adoptan una estructura al estar en interacción con otra proteína, ligando o por efecto de modificaciones postraduccionales. Las PID son frecuentes en genomas eucariotas [93] y suelen estar relacionadas con rutas regulatorias y de señalización celular [94]. En proteínas que contienen algunas regiones con estructura tridimensional y otras intrínsecamente desestructuradas, se habla de regiones intrínsecamente desestructuradas (RID). Se estima que, sino todos, la gran mayoría de los factores de transcripción del genoma humano contienen RID funcionales [95].



**Figura 7. Esquema de función según el paradigma de estructura-función (izquierda) y según el paradigma de desorden-función (derecha).** En la parte inferior se muestra una escala entre los dos paradigmas, con esquemas de proteínas con distintas proporciones de regiones estructuradas e intrínsecamente desestructuradas. En rojo, regiones intrínsecamente desestructuradas. En azul, regiones con estructura tridimensional estable. Imagen de [90].

Los motivos lineales (SLiMs, abreviatura de *Short Linear Motif*) son secuencias dentro de una proteína, típicamente de menos de 10 aminoácidos de largo y ubicadas en RIDs, que representan un tipo importante de interfaz de interacción y reconocimiento. Si bien cumplen un amplio rango de funciones, típicamente se los clasifica en dos tipos [96]:

A) Motivos de unión que median interacciones con dominios globulares de otras proteínas.

B) Sitios de modificación post-traducciona que son reconocidos y modificados por enzimas que los reconocen.

Se estima que en el proteoma humano hay alrededor de un millón de SLiMs [96,97]. Con el objetivo de estudiar las numerosas instancias donde se han predicho y encontrado experimentalmente motivos lineales, se desarrolló *Eukaryotic Linear Motif*

*resource* (base de datos ELM) cuyo objetivo es la anotación y detección de motivos lineales en proteínas de organismos eucariotas [97].

Los SLiMs en ELM se definen por medio de expresiones regulares (comúnmente llamadas RegEx, abreviatura de *Regular Expression*). Una expresión regular es un patrón utilizado para buscar una combinación de caracteres en cadenas de texto. Este patrón contiene una serie de posiciones predeterminadas, y cada una de ellas limita los caracteres que coinciden con el patrón (Figura 8).

En el caso de las expresiones regulares de ELM, cada posición del patrón corresponde a aminoácidos permitidos o prohibidos en esa ubicación en una secuencia buscada. Cada posición puede dividirse en una de dos categorías dependiendo de qué tan permisivo es el patrón en esa posición. Las posiciones “fijas” son aquellas que no toleran ningún cambio por otro aminoácido. Las posiciones “degeneradas” son aquellas que permiten algún nivel de variabilidad con una lista de aminoácidos permitidos, generalmente siendo aminoácidos similares. Las posiciones “fijas” y “degeneradas” son consideradas posiciones “definidas”. Las posiciones indefinidas o “abiertas” son aquellas donde cualquier aminoácido es permitido o donde una lista de aminoácidos están prohibidos, generalmente siendo unos pocos aminoácidos similares (Figura 8) [98].

**[KR]xLx{0,1}[^KR]**

|              |                                                                         |
|--------------|-------------------------------------------------------------------------|
| <b>L</b>     | Aminoácido específico (Leu)                                             |
| <b>[LI]</b>  | Distintos aminoácidos permitidos en una misma posición (Leu, Ile)       |
| <b>[^KR]</b> | Uno o distintos aminoácidos prohibidos en una misma posición (Lys, Arg) |
| <b>x / .</b> | Cualquier aminoácido permitido                                          |
| <b>{1,3}</b> | Cantidad variable de repeticiones                                       |

|      |   |   |        |       |
|------|---|---|--------|-------|
| 1    | 2 | 3 | 4      | 5     |
| [KR] | x | L | x{0,1} | [^KR] |

**Figura 8. Ejemplo de expresión regular.** Arriba, secuencia correspondiente a la expresión regular. Medio, explicación del formato. Abajo, separación de la expresión regular en cada uno de los componentes buscados. En azul, posiciones definidas. En rojo, posiciones abiertas.

#### 1.3.4. Predictores de efecto de variantes génicas en la estructura proteica

Existen numerosos programas computacionales desarrollados con el objetivo de estimar el efecto en una proteína causado por una dada VG. Ejemplos de programas utilizados actualmente son pyDock [99,100], FoldX [101,102], EvoEF1 [103], EvoEF2 [104], PRODIGY [105,106], BeAtMuSiC [107] y mCSM [108].

La mayoría de estos programas predicen el efecto de una VG desde información estructural de la proteína y/o aplicando principios moleculares (por ejemplo, calculando la diferencia en energías de interacción entre una estructura tridimensional *wild type* o la misma estructura con la variante en cuestión). Predictores más recientes, como RaSP [109], incluyen el uso parcial o total de algoritmos de aprendizaje automático (*machine learning*), que predicen si una variante tiene un efecto en la función proteica principalmente a partir de la secuencia proteica.

Estos programas son de utilidad para sugerir efectos de cambios en secuencias proteicas que serían muy difícil de confirmar *in vitro*, como ser el caso de estudiar el efecto en la estructura proteica de todas las VGs para uno o varios genes.

### 1.4. Estudio de variantes génicas en *NKX2-5* y su potencial patogenicidad

La asociación de VGs con patogenicidad ha sido estudiada en gran detalle, aunque las fuentes suelen tener datos discordantes y la información no siempre se encuentra unificada. El estudio de variantes en *NKX2-5* y su asociación con presentaciones de CC, en particular, tiene gran cantidad de ejemplos [32,40,110], incluyendo el trabajo realizado con anterioridad en nuestro laboratorio [68].

La información obtenida sobre VGs incluye desde estudios en pacientes hasta predicciones *in silico* del potencial efecto de una dada variante en distintas funciones de la proteína.

Dada la numerosa cantidad de fuentes de las cuales puede obtenerse esta información, es crucial unificar formatos de maneras escalables a futuro y comparables con la mayor cantidad de fuentes posible.

Estos formatos permiten comparar clasificaciones de distintas fuentes e incluso facilitar la actualización y comparación de fuentes en formatos compatibles. En el caso de la evaluación de patogenicidad de VGs, resulta de particular utilidad tener formatos estandarizados para:

- 1) Definir la posición y tipo de las VGs estudiadas de forma inequívoca.
- 2) Definir si distintas VGs en distintas condiciones serán clasificadas como patogénicas, benignas, de interés, etc.

Para la posición y tipo de VGs estudiados, el formato propuesto por la *Human Genome Variation Society* (HGVS, web: <http://www.hgvs.org/>) [111] es comúnmente reconocido como el estándar de nomenclatura para genética molecular [112,113]. Este formato define un identificador para variantes génicas con el objetivo de ser estable, significativo, fácil de recordar e inequívoco.

Por su parte, la clasificación de VGs como patogénicas, benignas o de interés cumple la doble función de proveer una potencial clasificación de patogenicidad, y permite comparar clasificaciones entre distintas variantes estudiadas por distintos equipos. El sistema de clasificación del *American College of Medical Genetics* (ACMG) [113] para definir la patogenicidad de las VGs recopiladas es reconocido internacionalmente, y es particularmente útil para determinar variantes de interés en futuros estudios.

#### 1.4.1. Formato HGVS

El identificador recomendado por HGVS es construido en tres partes: Una secuencia de referencia (*scaffold*), la posición de la variante dentro de esa referencia y la variación en un formato específico definido por el sistema de convenciones de HGVS (Figura 9).

|                |                             |
|----------------|-----------------------------|
| ADN            | NC_000005.10:g.173232992G>C |
| ADNc           | NM_004387.3:c.552C>G        |
| Proteína (CEE) | NP_004378.1:p.Ile184Met     |
| Proteína (SEE) | NP_004378.1:p.(Ile184Met)   |

**Figura 9. Ejemplos de una misma variante de NKX2-5 para distintas referencias.** En rojo, identificador de secuencias de referencia. En azul, posición de la variante dentro de la secuencia de referencia. En verde, nucleótidos o aminoácidos que varían. ADNc: ADN copia. CEE/SEE: Con/Sin evidencia experimental de expresión de la proteína.

Para la secuencia de referencia, se puede utilizar cualquier identificador de bases de datos reconocidas. Uno de los más comúnmente utilizados es el identificador RefSeq del *National Center for Biotechnology Information* (NCBI) [114], que incluye referencias para ADN, ADNc y proteínas, entre otros.

#### 1.4.2. Criterio ACMG

El sistema de clasificación de ACMG permite asignarle a cada VG una de 5 clasificaciones: “*pathogenic*”, “*likely pathogenic*”, “*uncertain significance*”, “*likely benign*” y “*benign*” [113]. Esta clasificación no tiene la función de reemplazar el diagnóstico clínico de la patogenidad de variantes, sino de resumir la información conocida para su interpretación en distintas áreas, y es comúnmente utilizada para la selección de variantes de interés.

El sistema incluye un proceso de clasificación basado en criterios utilizando evidencia típica de efecto de las variantes y se actualiza frecuentemente para reevaluar la significancia de los distintos tipos de evidencia disponible. Los tipos de evidencia utilizados en la versión de 2015 (la última versión durante el desarrollo de este trabajo) son los siguientes:

- **Información poblacional.** Incluye la frecuencia de la variante en bases de datos poblacionales y estudios que muestren estadísticamente su presencia en pacientes con una dada patogenidad.
- **Información computacional y de predictores *in silico*.** Incluye el análisis del tipo de variante, como variantes que truncan la proteína o variantes en un

aminoácido que ya tiene una variante patogénica asociada, y de una gran cantidad de predictores distintos que evalúen la variante a nivel conservacional, evolutivo y su efecto en secuencia y estructura proteica.

- **Datos de estudios funcionales.** Incluye ensayos *in vitro* o *in vivo* que muestren un efecto (o falta de efecto) de la variante en la función normal de la proteína, y también información de regiones de la proteína de importancia funcional sobre las que pueda haber caído la variante.
- **Segregación de la variante con la patogénesis.** Incluye estudios de co-segregación donde se vea que la variante se asocia con patogenicidad en distintos individuos, preferiblemente de un mismo árbol familiar, o donde se vea que no se asocia con ella. Su importancia aumenta mientras más estudios haya y más significativos sean.
- **Confirmación de variante *de novo*.** Se considera que una variante *de novo* en ciertos genes es evidencia de patogenicidad. Requiere secuenciación en ambos padres y se debe confirmar tanto maternidad como paternidad por estudios de filiación, caso contrario se establece como presunta *de novo*, dándole un nivel de evidencia menor.
- **Información alélica.** Para variantes en enfermedades recesivas, se registra evidencia de otras variantes patogénicas en *cis* y en *trans*, que pueden aportar evidencia de patogenicidad o de polimorfismo.
- **Otras bases de datos.** Se registra si la variante está anotada en una base de datos como benigna o patogénica. En la versión más reciente se recurre directamente a la evidencia de evaluación usada por la base de datos para evitar doble conteo de experimentos.
- **Otros datos.** Incluye registrar si se encontró en casos con evidencia alternativa de la patogénesis o si se encontró en un paciente con un fenotipo específico asociado al gen donde se encuentra la variante.

Con estos grupos de evidencia, se clasifican los distintos grupos en evidencia “de apoyo” o “fuerte” para que la variante sea benigna, o evidencia “de apoyo”, “moderada”, “fuerte” o “muy fuerte” para que la variante sea patogénica, de acuerdo a lo que se esquematiza en la Figura 10.

|                                        | Benigna                                            |                                                                              | Patogénica                                                                                                                                    |                                                                        |                                                   |               |
|----------------------------------------|----------------------------------------------------|------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------|---------------------------------------------------|---------------|
|                                        | Fuerte                                             | De apoyo                                                                     | De apoyo                                                                                                                                      | Moderada                                                               | Fuerte                                            | Muy fuerte    |
| Información poblacional                | Frecuencia poblacional<br>Prevalencia en controles |                                                                              |                                                                                                                                               | Ausente en bases de datos poblacionales                                | Prevalencia en pacientes                          |               |
| Información computacional y predictiva |                                                    | Evidencia computacional<br>Variante de efecto no asociado                    | Evidencia computacional                                                                                                                       | Residuo con otra variante patogénica<br>Cambia el largo de la proteína | Mismo cambio que una variante patogénica conocida | Variante nula |
| Información funcional                  | Estudios funcionales                               |                                                                              | Gen con pocas variantes benignas y muchas variantes patogénicas                                                                               | Hot spot de variantes patogénicas                                      | Estudios funcionales                              |               |
| Segregación                            | No segrega con la enfermedad                       |                                                                              | Co-segregación con la enfermedad en una familia (aumenta la fuerza de la evidencia mientras más información de segregación tenga la variante) |                                                                        |                                                   |               |
| Variante de novo                       |                                                    |                                                                              |                                                                                                                                               | Variante <i>de novo</i> (no confirmada)                                | Variante <i>de novo</i> (confirmada)              |               |
| Información alélica                    |                                                    | En trans con variante patogénica dominante<br>En cis con variante patogénica |                                                                                                                                               | En trans con variante patogénica para una enfermedad recesiva          |                                                   |               |
| Bases de datos                         |                                                    | Clasificada como benigna en base de datos                                    | Clasificada como patogénica en base de datos                                                                                                  |                                                                        |                                                   |               |
| Otra información                       |                                                    | Paciente con explicación alternativa de la patogénesis                       | Fenotipo de paciente altamente específico para el gen                                                                                         |                                                                        |                                                   |               |

**Figura 10. Esquema de las clasificaciones de la información y la forma en la que aportan evidencia para que una variante sea clasificada como benigna o patogénica.** Basado en datos de [113].

Con la evidencia recopilada, se determina la clasificación de la variante tal como se muestra en la Figura 11. Cada característica de las variantes se asocia a un código compuesto por 6 niveles (PVS, PS, PM, PP, BA y BS) que evalúan 28 tipos de evidencia diferente. La suma del puntaje determina la clasificación de la variante en: Patogénica, Probablemente patogénica, Probablemente benigna y Benigna. En caso de no superar el puntaje mínimo para ninguna de las clasificaciones, o de superar el puntaje para ambas, la variante se clasifica como de significado incierto (VUS, por “*variant of uncertain significance*”).

| Clasificación                   | Requisitos                                             |                                  |
|---------------------------------|--------------------------------------------------------|----------------------------------|
| Patogénica                      | 1 Muy fuerte                                           | 1 (o más) Fuertes                |
|                                 |                                                        | 2 (o más) Moderadas              |
|                                 |                                                        | 1 Moderada y 1 de apoyo          |
|                                 |                                                        | 2 (o más) de apoyo               |
|                                 | 2 (o más) Fuertes                                      |                                  |
|                                 | 1 Fuerte                                               | 3 (o más) Moderadas              |
|                                 |                                                        | 2 Moderadas y 2 (o más) de apoyo |
| 1 Moderada y 4 (o más) de apoyo |                                                        |                                  |
| Probablemente patogénica        | 1 Muy fuerte y 1 Moderada                              |                                  |
|                                 | 1 Fuerte                                               | 1 o 2 Moderadas                  |
|                                 |                                                        | 2 (o más) de apoyo               |
|                                 | 3 (o más) Moderadas                                    |                                  |
|                                 | 2 Moderadas y 2 (o más) de apoyo                       |                                  |
|                                 | 1 Moderada y 4 (o más) de apoyo                        |                                  |
| Benigna                         | 1 Muy fuerte                                           |                                  |
|                                 | 2 (o más) Fuertes                                      |                                  |
| Probablemente benigna           | 1 Fuerte y 1 de apoyo                                  |                                  |
|                                 | 2 (o más) de apoyo                                     |                                  |
| Significado incierto            | No clasificada con ninguno de los criterios anteriores |                                  |
|                                 | Clasificación como benigna y patogénica al mismo       |                                  |

**Figura 11. Esquema de la clasificación dada a una variante según la cantidad de cada tipo de evidencia que tenga.** La columna requisitos hace referencia a las columnas de la Figura 10. Basado en datos de [113].

Como se mencionó anteriormente, la clasificación dada por este criterio no está pensada para dar una clasificación definitiva de patogenicidad de una variante, y su uso es más propicio para direccionar futuros estudios. En particular, las variantes de significado incierto que necesitan poca evidencia para ser clasificadas como probablemente patogénicas son de interés para estudios que puedan ayudar a determinar su clasificación.

Una vez que una variante sea clasificada como benigna o patogénica, esto puede ayudar a la hora de un diagnóstico clínico. Aunque no sea una clasificación definitiva, la identificación de variantes patogénicas según este criterio es una forma efectiva de filtrar la gran cantidad de variantes disponibles a estudiar en ensayos genómicos y direccionar futuros estudios diagnósticos.

## **1.5. Interacción de factores de transcripción con otras macromoléculas**

Los factores de transcripción son proteínas que se unen al ADN con especificidad definida por su secuencia y que regulan su transcripción. Su función y sus interacciones con otras macromoléculas son de gran importancia en el desarrollo y diferenciación de tejidos embrionarios [115-117].

Por todo esto, la determinación tanto de los sitios de unión al ADN como de otras proteínas con las que interactúa un dado factor de transcripción es un elemento clave en el estudio de su funcionalidad normal.

### **1.5.1. Interacción de factores de transcripción con el ADN**

El estudio de los sitios de unión de factores de transcripción al ADN en el genoma da información tanto de genes transactivados (activados o inhibidos) río abajo como de genes que puedan ser co-regulados con el factor de transcripción estudiado.

Esta información tiene un aumento exponencial en su potencial utilidad al permitir estudiar los sitios de unión de factores de transcripción distintos en condiciones experimentales similares. De esta forma, se pueden identificar sitios de unión cercanos entre distintos factores de transcripción que pueden aportar información nueva que no se podría obtener de los estudios por separado.

#### **1.5.1.1. Búsqueda de sitios de unión**

El estudio de inmunoprecipitación de cromatina (ChIP, abreviatura de "*Chromatin ImmuNoPrecipitation*"), es un ensayo experimental que permite evaluar interacciones de proteínas con diferentes regiones del ADN genómico, utilizando anticuerpos específicos, enriqueciendo así las regiones de unión al ADN para su posterior secuenciación [118].

Con el advenimiento de las tecnologías de secuenciación masiva (NGS), se ha facilitado la secuenciación de fragmentos de ADN inmunoprecipitados con ChIP, dando lugar al ensayo de ChIP-seq, que permite estudios a nivel de todo el genoma.

A las regiones involucradas se las identifica como “picos” (*peaks*) de ChIP-seq, y suelen almacenarse en archivos con información de cromosoma, posición inicial/final y, a veces, puntajes de calidad y/o identificadores numerados [119-121].

La relativa facilidad de realización de este ensayo y la cantidad de información que deja disponible para análisis computacionales han permitido estudiar sitios de unión de múltiples genes. Esta información se encuentra disponible en bases de datos de experimentos, como *Gene Expression Omnibus* (GEO, <https://www.ncbi.nlm.nih.gov/geo/>) [122], las cuales facilitan el acceso para estudios informáticos.

#### 1.5.1.2. Confirmación funcional de sitios de unión

Los ensayos de ChIP-seq frecuentemente encuentran miles de sitios de unión a lo largo de un genoma y, aún considerando los controles experimentales realizados para evitar falsos positivos, es probable que varios de los sitios de unión no tengan asociación funcional a los genes cercanos y/o que no sea obvio a qué gen cerca del sitio de unión esté afectando el factor de transcripción.

Uno de los experimentos utilizados para obtener evidencia de que un factor de transcripción se une a un dado sitio de unión por razones funcionales es la secuenciación de ARN (*RNA-seq*) en condiciones de expresión diferencial de la proteína de interés.

El ensayo de RNA-seq se basa en secuenciar todos los fragmentos de ARN en una dada condición experimental. Se suele realizar para ARN mensajero, el cual se retrotranscribe a ADN copia (ADNc) antes de ser secuenciado con tecnologías NGS [123,124].

Para estudiar la expresión diferencial de ARN, se aproxima la expresión de cada uno de los genes del transcriptoma en dos condiciones experimentales distintas y se seleccionan aquellos genes con diferencia de expresión entre ambas condiciones [124-126]. En el contexto de confirmar picos de ChIP-seq, este ensayo se puede realizar comparando las condiciones utilizadas para la inmunoprecipitación del ensayo de ChIP-

seq con un *knock out* o *knock down* del gen de interés. De esta forma, se pueden seleccionar aquellos picos que se corresponden con genes que evidenciaron expresión diferencial.

Más allá de la predicción funcional, *in vitro* o *in vivo*, de sitios de unión de un factor de transcripción en condiciones experimentales dadas, también es posible generar sistemas *in silico* para predecir estos sitios de unión.

El sistema más simple de predicción es la selección de secuencias exactas, o una lista de secuencias consenso [127], las cuales pueden ser buscadas directamente a lo largo del genoma o, más convenientemente, dentro de picos de ChIP-seq.

Un sistema más elaborado es la utilización de matrices de pesos por posición (matriz de pesos o PWM, por sus siglas en inglés para "*Position-Weight Matrix*") [127]. En el contexto de sitios de unión de factores de transcripción, estas matrices se representan con tablas de 4 filas (una por nucleótido) y tantas columnas como posiciones tenga el sitio de unión (Figura 12, arriba).

Con esta tabla se puede asignar un puntaje a una secuencia de largo predeterminado, con el cual se puede determinar si la secuencia es un sitio de unión para un factor de transcripción específico. Para calcular el puntaje que una matriz de puntajes le asigna a una secuencia de ADN, se recorre la secuencia, viendo qué puntaje tiene cada nucleótido en la posición correspondiente de la matriz (Figura 12, abajo).

Para definir que una dada secuencia de ADN es un sitio de unión de un factor de transcripción en particular, el puntaje que le asigne la PWM para sitios de unión de ese factor de transcripción debe ser mayor a un valor de corte determinado con anterioridad. Este valor de corte puede calcularse de diversas formas, y la conclusión de cuál es el mejor en cada situación depende de qué tan tolerante a falsos negativos y/o falsos positivos sea el diseño experimental [128].

Las matrices de conteos y de pesos se pueden obtener de bases de datos que recopilan experimentos de distintos factores de transcripción y determinan sus matrices

de puntajes. Una de estas bases de datos es HOCOMOCO (<https://hocomoco11.autosome.org/>) [129], que utiliza resultados de experimentos de ChIP-seq para determinar sitios de unión, con los que se generan las matrices de puntajes. Esta incluye valores precomputados de corte para los puntajes de las matrices, con p-valores asociados correspondientes a la probabilidad de reconocer una secuencia al azar de ADN como sitio de unión.

|                          | 1      | 2      | 3      | 4      | 5      | 6      | 7      | 8      | 9      | 10     | 11     |        |
|--------------------------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| A                        | 0,399  | 0,942  | -0,923 | -6,456 | -5,411 | -5,977 | -1,006 | -1,656 | 1,183  | 0,685  | 0,188  |        |
| C                        | -0,375 | -1,549 | -0,242 | -4,489 | 0,095  | -6,823 | -1,227 | -1,351 | -1,809 | -0,891 | -0,237 |        |
| T                        | -0,147 | -0,827 | -1,597 | -6,823 | 1,057  | -6,823 | 1,149  | 0,131  | -1,531 | -0,181 | -0,134 |        |
| G                        | -0,041 | -0,241 | 0,961  | 1,383  | -3,966 | 1,385  | -1,677 | 0,88   | -1,037 | -0,258 | 0,122  |        |
| Secuencia:               | A      | A      | G      | G      | C      | G      | T      | G      | A      | A      | G      | Suma   |
| Puntaje de la secuencia: | 0,399  | 0,942  | 0,961  | 1,383  | 0,095  | 1,385  | 1,149  | 0,88   | 1,183  | 0,685  | 0,122  | 9,184  |
| Puntaje máximo:          | 0,399  | 0,942  | 0,961  | 1,383  | 1,057  | 1,385  | 1,149  | 0,88   | 1,183  | 0,685  | 0,188  | 10,212 |

**Figura 12. Ejemplo de matriz de pesos por posición (PWM) de TBX20 y cálculo de puntaje para una secuencia de ADN.** Arriba, PWM con puntajes para cada uno de los nucleótidos posibles en cada posición (nucleótidos de la secuencia recuadrados en negro, puntajes máximos por posición en rojo). Abajo, ejemplo de cálculo de puntaje para una secuencia de ADN, incluyendo cálculo de puntaje máximo de la matriz.

Tanto las secuencias consenso como las matrices de pesos pueden representarse como LOGOs [130], utilizando las secuencias de referencia de la matriz PCM como lista de secuencias consenso.

### 1.5.1.3. Contexto génico de los sitios de unión

Los sitios de unión por sí solos no proveen información funcional, y es necesario analizar el contexto en el cual se ubican. Como se explicó anteriormente en la sección 1.5.1.1., el análisis de la funcionalidad de sitios de unión por ensayos como RNA-seq con expresión diferencial aportan evidencia para sitios de unión en picos de ChIP-seq. Pero, además, estos resultados proveen una lista de genes que pueden ser estudiados para ver patrones en su funcionalidad. El estudio de una lista de genes no es trivial y, para ello, una clasificación estandarizada es de gran ayuda en la búsqueda de patrones en su función.

En el año 2000, tras las primeras secuenciaciones de genomas completos en eucariotas, se vio un gran nivel de homología y similitud entre genes tanto de distintos genomas como dentro de cada genoma. Este nivel de homología y similitud entre distintos genes sugirió un nivel similar de solapamiento funcional entre ellos. Teniendo en cuenta que era de público conocimiento la proximidad de la secuenciación del genoma humano [131,132] se creó el Gene Ontology (GO) Consortium [133].

El objetivo de este grupo fue crear un lenguaje común para hablar de genes y sus funciones. De esta forma, se busca aprovechar las similitudes entre los distintos genes y asignarles clasificaciones (GO-terms) que permitan estudiarlos de formas más significativas y asociadas a sus características funcionales y/o estructurales.

Los GO-terms representan características funcionales que pueden asociarse a los productos de un gen [134]. Estas se dividen en 3 grupos:

- Función molecular (“*Molecular function*”), que hacen referencia a la función de los productos del gen a nivel molecular (por ejemplo, “*carbohydrate metabolic process*” para genes cuyos productos participan en el metabolismo de carbohidratos).
- Componente celular (“*Cellular component*”), que refieren a la ubicación donde se llevan a cabo las funciones moleculares (por ejemplo, “*plasma membrane*” para genes cuyos productos pertenecen o interactúan con la membrana plasmática).
- Procesos biológicos (“*Biological process*”), que representan funciones más complejas que emergen de las funciones moleculares (por ejemplo, “*axonogenesis*” para hacer referencia a los genes cuyos productos están involucrados en el desarrollo de los axones).

Estos GO-terms se organizan en jerarquías y redes, donde algunos términos incluyen a otros (por ejemplo, “*carbohydrate metabolic process*”, correspondiente al metabolismo de carbohidratos, está incluido por “*metabolic process*”, que es un término más general para todos los procesos metabólicos), y términos de distintos grupos se relacionan entre sí (por ejemplo, “*carbohydrate metabolic process*” puede asociarse al

componente celular “*mitochondria*” para un gen del metabolismo de carbohidratos que realiza su función en la mitocondria).

Desde su creación, la base de datos de GO (<https://geneontology.org/>) se ha vuelto un componente crucial del estudio de sistemas biológicos [135] y en noviembre de 2022 contenía evidencia experimental de más de 170.000 artículos científicos [134].

#### 1.5.2. Interacción de factores de transcripción con otras proteínas

La interacción de factores de transcripción individuales con el ADN no es suficiente para explicar la complejidad de su función. Para esto, es común que interactúen con otras proteínas, mayoritariamente con otros factores de transcripción, con los cuales crean complejos de transactivación y redes de interacción asociadas a funciones particulares [116]. En los casos más extremos, varios factores de transcripción pueden unirse a un sólo sitio (conocido como *super enhancer*), formar un complejo estable y activar el gen río abajo mucho más que en sitios de unión individuales [136,137]. Estos grupos suelen incluir genes de diferenciación temprana y, particularmente, genes que definen la identidad celular [138].

La interacción entre factores de transcripción da lugar a redes de interacción que permiten estudiar numerosas funciones del desarrollo en mucho más detalle que mediante el estudio de cada uno de ellos por separado [139]. Estas interacciones se encuentran registradas en bases de datos que compilan evidencia de distintas fuentes para la interacción entre factores de transcripción y otras proteínas y productos génicos. Algunos ejemplos son GeneMANIA (<https://genemania.org/>) [140], STRING (<https://string-db.org/>) [141], IntAct (<https://www.ebi.ac.uk/intact/>) [142] y BioGRID (<https://thebiogrid.org/>) [143].

La información de las interacciones entre factores de transcripción puede provenir de distintas fuentes, las cuales incluyen estudios experimentales de interacción y transactivación [144,145] y registro de estructuras tridimensionales entre proteínas [146].

## 1.6. Marco teórico

La asociación funcional entre el gen *NKX2-5* y el desarrollo del corazón es entendida en sus generalidades y se conocen numerosas variantes en *NKX2-5* asociadas, principalmente, con CC. Aún así, no se comprende en profundidad el mecanismo que relaciona a estas variantes y CC. Específicamente, no se conoce el mecanismo por el cual distintas variantes génicas (o una misma variante con variabilidad fenotípica) se asocian con distintos subtipos de cardiopatías.

Esta falencia puede conceptualizarse como una brecha en el conocimiento con, por un lado, las variantes génicas conocidas de *NKX2-5* y, por el otro, los subtipos de CC a los que se encuentran asociadas. La eliminación (o reducción) de esta brecha implicaría completar información o sistematizar la información existente así como sistematizar un entramado de conocimientos que puedan ayudar a asociar la información entre genotipos y fenotipos. Este entramado necesariamente incluye, para un factor de transcripción, información de interacción con otras proteínas, sitios de unión al ADN y genes transactivados río abajo.

En cualquiera de los casos, la información disponible es abundante, y su sistematización y procesamiento en el contexto de asociar variantes en *NKX2-5* con subtipos de CC ayudaría a entender mejor el desarrollo de CC.

## **2. Hipótesis y objetivos**

## 2.1. Antecedentes del laboratorio

Como parte de trabajos previos, nuestro grupo ha estudiado las causas genéticas asociadas al desarrollo de CC. En particular, se analizaron 236 muestras de pacientes con diferentes tipos de CC mediante un abordaje integral que incluyó análisis masivos de genes que permitieron hallar variantes de significancia clínica en el 67% de los pacientes analizados y describir variantes novedales en individuos de nuestra población [33]. Dentro de esta búsqueda, el gen *NKX2-5* fue estudiado en forma particular en un grupo seleccionado de 64 pacientes [147].

Por otro lado, este trabajo continúa la línea de [68], donde se creó una lista exhaustiva (hasta 2017) de variantes génicas de *NKX2-5* y se analizaron implicancias estructurales, funcionales y de estabilidad de un grupo de variantes seleccionadas, a partir de una estructura resuelta por difracción de rayos X.

## 2.2. Hipótesis

Las dos hipótesis propuestas en este trabajo son:

1. Es posible desarrollar un modelo computacional predictivo de la patogenicidad potencial de VGs en *NKX2-5* no caracterizadas funcionalmente en relación a CC. Este modelo implica una comprensión profunda de la secuencia, estructura y función proteica con relación funcional al desarrollo del corazón, e incrementa su poder predictivo mientras más se estudien estas características.
2. La función proteica de *NKX2-5* en relación a CC se puede comprender estudiando sus sitios de unión en el genoma y la red de interacciones con otros factores de transcripción y genes que transactiva.

## 2.3. Objetivos

### 2.3.1. Objetivo general

Desarrollar un flujo de trabajo para VGs de *NKX2-5* que permita el estudio predictivo de su patogenicidad y su interacción con otras proteínas que se asocien funcionalmente en el desarrollo del corazón y su posible asociación con CC.

### 2.3.2. Objetivos específicos

1. Generar un modelo estructural de *NKX2-5*.
2. Recopilar y analizar todas las variantes génicas conocidas en *NKX2-5*.
3. Realizar un análisis específico y en profundidad de variantes en regiones de *NKX2-5* con estructura tridimensional definida.
4. Recopilar y estudiar todos los sitios de unión de *NKX2-5* en el genoma humano y otros posibles factores de transcripción que puedan estar relacionados con estos sitios.
5. Utilizar toda la información recopilada para evaluar las posibles causas de CC asociada a *NKX2-5* en base a los procesos biológicos que puedan estar involucrados en el desarrollo de la patología.

### **3. Materiales y métodos**

Para cumplir el objetivo propuesto, se dividió el trabajo en 5 partes:

1. Desarrollo de un modelo estructural de *NKX2-5*. Este modelo definió la dirección de los demás estudios.
2. Recopilación de variantes génicas de *NKX2-5* y análisis de su relación con CC, con foco en su posible efecto en la función de *NKX2-5*.
3. Análisis específico de regiones estructuradas de la proteína codificada por *NKX2-5*, estudiando todas las variantes de cambio de sentido posibles en esta/s región/es.
4. Recopilación de sitios de unión de *NKX2-5* en genomas relevantes y análisis de su potencial asociación con genes transactivados o inhibidos por *NKX2-5*.
5. Estudio de sitios de unión cercanos de otros factores de transcripción y análisis de la función de genes transactivados o inhibidos por *NKX2-5*.

Cada una de estas partes tiene un flujo de trabajo asociado. En cada uno de ellos se utiliza la numeración de las secciones de la tesis para correlacionar los pasos del flujo de trabajo con las secciones de texto correspondientes.

### **3.1. Generación de modelo estructural de *NKX2-5***

En primer lugar, se recopiló información correspondiente a características estructurales de la proteína *NKX2-5*. Esta información fue luego utilizada durante el resto de este trabajo. El flujo general de trabajo para la generación del modelo estructural de *NKX2-5* se esquematiza en la Figura 13. El modelo obtenido representa un refinamiento del modelo creado y estudiado en un trabajo anterior [68].

El modelo obtenido incluye:

- Regiones con estructura tridimensional. Se hizo énfasis en la función de el/los dominio/s (si tuviera/n función asociada) y se estudió la posible relevancia de cada aminoácido en particular.

- Regiones intrínsecamente desordenadas. Se hizo énfasis en motivos lineales que puedan representar funcionalidad importante en la región desordenada. También se analizaron regiones conservadas que puedan ser asociadas a una función específica en el futuro.



**Figura 13. Flujo de trabajo general para la generación del modelo estructural de NKX2-5.** Los números en recuadros celestes se corresponden con las secciones de este trabajo donde se explica el paso correspondiente. PDB: Protein Data Bank. ELM: Eukaryotic Linear Motif. SLiMs: Motivos lineales (*Short Linear Motifs*).

### 3.1.1. Selección de estructuras tridimensionales correspondientes a NKX2-5

El primer paso para la definición del modelo estructural fue buscar estructuras tridimensionales de NKX2-5 obtenidas por cristalografía de rayos X o por resonancia magnética nuclear (RMN) en la base de datos Protein Data Bank (PDB) de la RCSB [69,148,149]. Esta búsqueda se hizo específicamente para estructuras asociadas al gen *NKX2-5*, no para la secuencia de la proteína NKX2-5.

Con la finalidad de comparar y seleccionar la mejor estructura que cubra cada región estructurada para los estudios *in silico*, se registraron las siguientes características en forma de tabla:

- Método de obtención de la estructura (cristalografía rayos X o RMN)
- Resolución (en Angstroms)
- Posición de la proteína correspondiente a la estructura tridimensional (región de homología)
- Macromoléculas de la estructura tridimensional (proteínas y ADN)

Para los estudios de generación de variantes de NKX2-5 se utilizó la/s estructura/s con mejor resolución, mientras que para estudios de interacción con otras proteínas se seleccionaron estructuras en interacción con otros factores de transcripción.

### 3.1.2. Estudio de regiones con estructura tridimensional

Se definió la región estructurada de NKX2-5 como aquella cubierta por alguna estructura tridimensional, y fue la elegida como foco de estudios específicos. Dentro de esta región, se asignaron clasificaciones para aminoácidos que puedan ser relevantes para variantes que afectan a dichos aminoácidos.

Para este fin, se definieron listas de aminoácidos que probablemente formen parte de un *core* hidrofóbico, estructuras secundarias (hélices alfa y hojas beta), que estén en regiones accesibles sin estructura secundaria (*loops*) y que estén en contacto

con el ADN o en contacto con otras proteínas (en estructuras donde NKX2-5 estuviese en contacto con ADN y/o proteínas).

El **core hidrofóbico** se definió con ayuda de las publicaciones asociadas a las estructuras seleccionadas y el programa Chimera [150]. En primer lugar, se ordenaron los residuos de acuerdo a su accesibilidad al solvente (*areaSAS*) con la función “*Render by attribute*” de Chimera. Con este valor, se clasificaron aquellos residuos con un área de accesibilidad de menos de 6 Angstroms<sup>2</sup> como “inaccesibles por el solvente”, por tener menor área de accesibilidad que un círculo con el diámetro de una molécula de agua, que es de 2,8 Angstroms [151]. Estos residuos formaron la base del *core* hidrofóbico, y también se clasificaron todos los residuos con cadenas laterales a menos de 5 Angstroms de los residuos inaccesibles como parte del *core* hidrofóbico ya que, al ser una estructura pequeña, cambios en estos residuos podrían afectar a la totalidad del *core* hidrofóbico.

El **contacto con el ADN y/u otras proteínas** se definió por predicción de interacciones entre átomos de NKX2-5 y átomos de ADN y/o proteínas de la estructura. Para esto, se definieron residuos que estuvieran a menos de 5 Angstroms de átomos de ADN, otras proteínas y/o moléculas de agua cercanas si la estructura las tuviera. Las interacciones buscadas fueron puentes de hidrógeno, interacciones electrostáticas e interacciones hidrofóbicas:

- Los puentes de hidrógeno se definieron por medidas de distancia y ángulos con la función “*FindHBond*” de Chimera, teniendo en cuenta la posibilidad de la presencia de moléculas de agua mediando la interacción.
- Las interacciones de Coulomb (entre átomos polarizados) se definieron para pares de grupos funcionales cargados positiva y negativamente a menos de 3 Angstroms de distancia, con la posibilidad de mediar la interacción con moléculas de agua.
- Las interacciones hidrofóbicas se consideraron para cadenas laterales hidrofóbicas que provinieran de distintas cadenas y estuvieran a menos de 3 Angstroms de distancia entre sí en regiones no accesibles (definidas por residuos del *core* hidrofóbico) de las estructuras tridimensionales.

### 3.1.2.1. Conservación evolutiva de residuos de región con estructura tridimensional

Por otro lado, se utilizaron los predictores de conservación evolutiva ConSurf [152] y SIFT [153] para evaluar la conservación de cada residuo de la región con estructura tridimensional. ConSurf estima la conservación de cada posición aminoacídica para determinar posiciones altamente variables o altamente conservadas, mientras que SIFT clasifica cada uno de los posibles cambios de un aminoácido como “tolerado” o “no tolerado”, en base a información de alineamientos múltiples con secuencias homólogas.

Tomando en cuenta los valores obtenidos con los predictores, se consideró a un residuo como altamente conservado por ConSurf si posee un *score* menor a -0,5, y como medianamente conservado si posee un *score* entre 0 y -0,5. Por su parte, utilizando SIFT, consideramos un residuo como altamente conservado si posee 1-2 residuos tolerados y 18-19 no tolerados, y como medianamente conservado si tiene 3-5 residuos tolerados y 15-17 no tolerados. El corte de 2 cambios tolerados se eligió para incluir posiciones que no toleran ningún cambio o sólo cambios por un aminoácido parecido al *wild type*, y el corte de 5 cambios tolerados, por incluir posiciones que toleran cambios por aminoácidos en grupos funcionales similares (positivos, negativos, no polares, pequeños).

Para este trabajo, se clasificó que un residuo tiene conservación “alta” si ambos predictores lo clasifican como tal y conservación “media” si ambos predictores lo clasifican como medianamente conservado, o si un predictor lo clasifica como altamente conservado y el otro como medianamente conservado. El resto de los residuos se consideraron de conservación “baja”.

### 3.1.3. Búsqueda de motivos lineales en regiones accesibles

Para el estudio de residuos sin limitarse a la región estructurada, se realizó una búsqueda de motivos lineales (SLiMs, acrónimo de *Short Linear Motif*) utilizando Eukaryotic Linear Motif (ELM, <http://elm.eu.org/>) [154]. Esta búsqueda fue complementada con una búsqueda bibliográfica.

La búsqueda en ELM se realizó con la secuencia aminoacídica completa de la isoforma canónica de NKX2-5, abarcando tanto la región estructurada como la que no lo está. La plataforma filtra regiones no accesibles automáticamente, pero esta función fue desactivada y se realizó un filtrado con nuestra información de regiones desestructuradas y loops accesibles obtenida previamente (sección 3.1.2.).

Para la búsqueda bibliográfica, se tuvo en cuenta la presencia de modificaciones post-traduccionales, señales de localización nuclear (NLS, acrónimo de *Nuclear Localization Signal*), interacciones secuencia-específicas con otras proteínas y, más particularmente, interacciones asociadas a SLiMs predichos por ELM. De manera acorde, se realizaron 4 tipos de búsquedas de referencias bibliográficas en PubMed (<https://pubmed.ncbi.nlm.nih.gov/>). Se buscó:

- <NKX2-5 post-translational modification> para buscar modificaciones post-traduccionales.
- <NKX2-5 NLS> para buscar señales de localización nuclear.
- <NKX2-5 protein interaction> para buscar interacciones secuencia-específicas con otras proteínas.
- <NKX2-5 [nombre del SLiM]>, donde [nombre del SLiM] corresponde al nombre de motivos lineales particulares que hayan sido encontrado en las búsquedas anteriores.

En aquellos SLiMs en los que se halló que existía información experimental que los avalara como tales, los motivos lineales fueron clasificados como “motivos lineales confirmados”, mientras que los motivos encontrados por ELM en los que no existía confirmación experimental fueron considerados “motivos lineales predichos”.

#### 3.1.4. Potenciales interacciones con secuencias de ADN

Para finalizar la recopilación de información para el modelo estructural, se buscó en la bibliografía evidencia experimental de interacción entre NKX2-5 y secuencias específicas de ADN. Se incluyó en este análisis secuencias que se encontraran en contacto con NKX2-5 en estructuras de PDB, secuencias específicas utilizadas en

ensayos de transactivación y estudios que recopilaran secuencias de unión al ADN de NKX2-5.

Con estas fuentes, se registró una lista de sitios de unión de NKX2-5, la cual fue utilizada para la búsqueda de sitios de unión en el genoma de la sección 3.4.

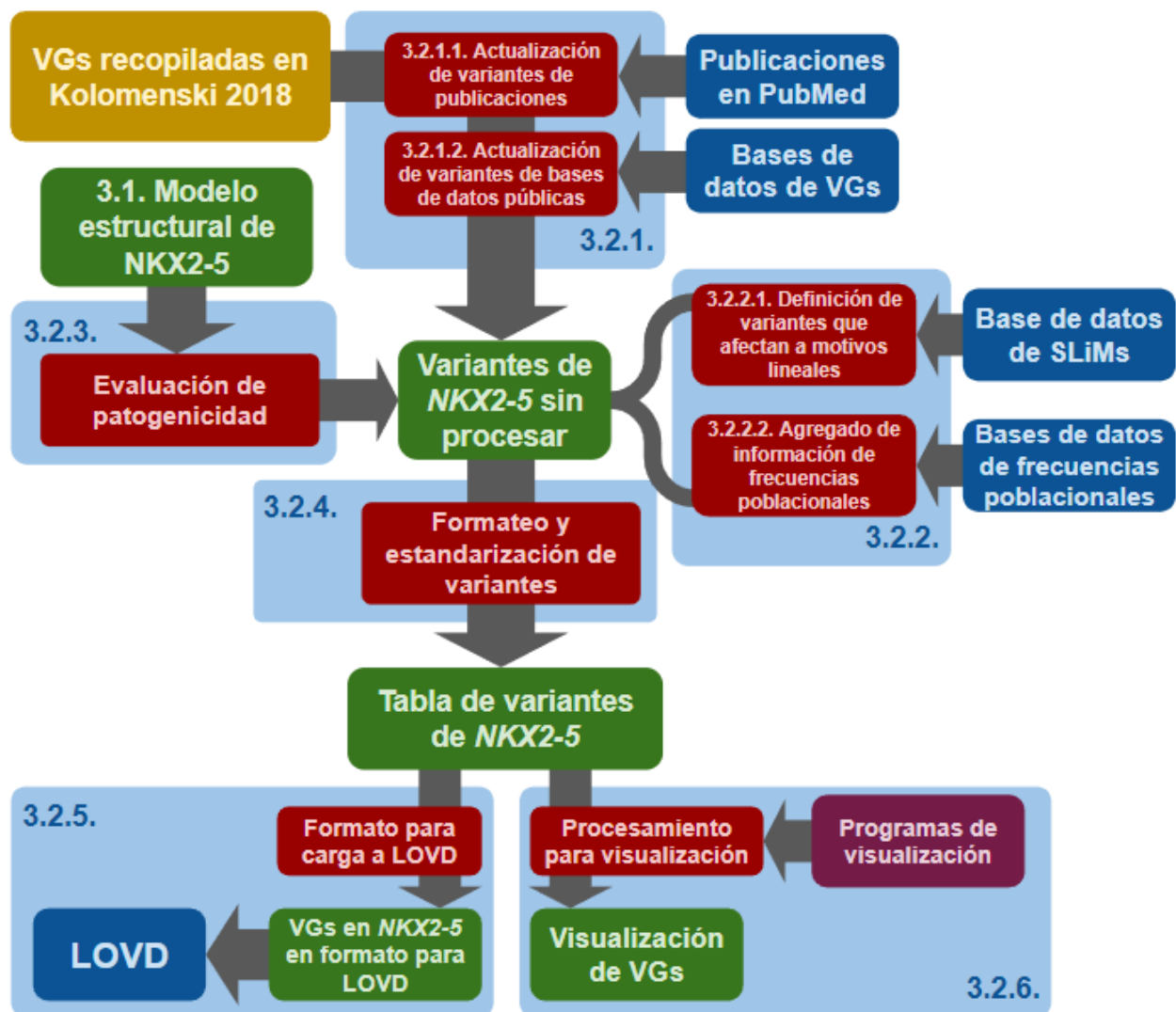
### **3.2. Recopilación y análisis de todas las variantes génicas conocidas de *NKX2-5***

Con el modelo estructural definido y actualizado, se recopilaron todas las variantes génicas (VGs) conocidas de *NKX2-5*, a las cuales se les agregó información de funcionalidad y potencial patogenicidad, haciendo uso del modelo estructural en varias etapas.

Se comenzó el trabajo tomando como fuente de datos una recopilación de todas las VGs conocidas de *NKX2-5* hasta julio de 2017. Provenientes de bases de datos públicas, de 52 publicaciones científicas y de 64 pacientes que acudieron al Centro Nacional de Genética Médica (CNGM) en los que se analizó la presencia de VGs en el gen *NKX2-5* como parte de un estudio de análisis de genes candidatos relacionados a cardiopatías congénitas [68].

Estas variantes se actualizaron, agregando variantes de bases de datos y publicaciones científicas. Luego, se agregó información funcional de motivos lineales, frecuencias poblacionales y predictores de efecto en estabilidad de ARN mensajero. Con la información recopilada, más la información del modelo estructural de la sección 3.1., se evaluó la patogenicidad de las variantes génicas y se registró de acuerdo a distintos criterios. Por último, se cargaron variantes en la base de datos LOVD y se desarrollaron programas específicos para visualizar la distribución de las variantes sobre el gen *NKX2-5* y/o la proteína codificada por el gen.

El flujo general de trabajo para la recopilación y análisis de todas las VGs en *NKX2-5* se esquematiza en la Figura 14.



**Figura 14. Flujo de trabajo general para la recopilación y análisis de todas las variantes génicas en *NKX2-5*.** Los números en recuadros celestes se corresponden con las secciones de este trabajo donde se explica el paso correspondiente. VGs: Variantes génicas. SLiMs: Motivos lineales (*Short Linear Motifs*). ACMG: American College of Medical Genetics.

### 3.2.1. Actualización de lista de variantes

Lo primero que se hizo con la lista de variantes de [68] fue actualizarla, incorporando VGs nuevas mediante una búsqueda de publicaciones científicas y consultando las bases de datos de acceso público.

### 3.2.1.1. Actualización de VGs en publicaciones científicas

Para la actualización de VGs en publicaciones científicas se tuvieron en cuenta las encontradas en la base de datos PubMed (<https://www.ncbi.nlm.nih.gov/pubmed/>) que incluyeran VGs en NKX2-5 hasta diciembre de 2019, seleccionando todas las publicaciones que contuvieran alguno de los nombres de NKX2-5 que figuran en GeneCards (<https://www.genecards.org/cgi-bin/carddisp.pl?gene=NKX2-5>) y “*mutation*” o “*genetic variant*” como palabras claves.

En caso de que una publicación clasificara a una variante como patogénica, benigna/polimorfismo o de significado incierto, se registró esta información.

### 3.2.1.2. Actualización de VGs en bases de datos públicas

Para el caso de las variantes recopiladas de las bases de datos públicas, la actualización se realizó en julio de 2019, por lo que se recopilaron VGs hasta esa fecha. Las bases de datos consultadas fueron NCBI's dbSNP (<https://www.ncbi.nlm.nih.gov/SNP/>) [155], NHLBI- ESP's EVS (<https://evs.gs.washington.edu/EVS/>), ExPASy's SwissVar (<https://swissvar.expasy.org/cgi-bin/swissvar/home/>) [156], ExAC (<https://exac.broadinstitute.org/>) [157] y LOVD (<http://www.lovd.nl/3.0/home/>) [158].

También se incluyeron variantes de ClinVar (<https://www.ncbi.nlm.nih.gov/clinvar/>) [159], que es una base de datos de clasificación de variantes según patogenicidad. Esta base de datos no había sido consultada en el trabajo previo [68].

En particular, se registró específicamente información de patogenicidad de LOVD, SwissVar y ClinVar.

### 3.2.2. Procesamiento y agregado de información de VGs

Con el objetivo de poder comparar VGs de distintas fuentes con distintos tipos de información, se unificaron las VGs extraídas de cada una de las bases de datos en

una sola lista. De esta forma se eliminan VGs duplicadas y se estandarizan los formatos, lo cual simplifica los análisis posteriores.

También se las complementó con información de motivos lineales (SLiMs) que pudieran ser afectados por cada variante y sus frecuencias en bases de datos poblacionales.

#### 3.2.2.1. Definición de variantes que afectan motivos lineales

Se utilizaron los SLiMs confirmados experimentalmente y predichos de acuerdo a lo consignado previamente (sección 3.1.3) para definir variantes que pudieran afectar su funcionalidad. Se registraron sólo las variantes que afectan específicamente a residuos dentro del motivo lineal (variantes de cambio de sentido o inserciones/delecciones sin cambio de marco de lectura), ya que su efecto en la funcionalidad de NKX2-5 es específico de la región correspondiente al SLiM. Las variantes que causan inserciones con cambio de marco de lectura o que introduzcan un codón *stop* que ocurran sobre un SLiM no se consideraron ya que su efecto probablemente se deba a que cambian el largo de la proteína, no por su ubicación en un SLiM.

Para los SLiMs predichos, se dividieron las variantes entre aquellas que se predice que afectarían al motivo lineal según el criterio de ELM y aquellas que no. El criterio de selección de ELM usa expresiones regulares, las cuales definen una lista de aminoácidos específicos permitidos en cada posición (Figura 8). No se realizó la misma distinción con los SLiMs confirmados experimentalmente por tener confirmación por otros métodos.

#### 3.2.2.2. Frecuencias poblacionales

Entre las bases de datos estudiadas, la información de frecuencia poblacional fue obtenida de dbSNP, EVS y ExAC. Para registrar la frecuencia como un valor único, se obtuvieron los valores de frecuencia separados en cantidad de individuos totales y cantidad de individuos con la variante, y se sumaron para las distintas fuentes,

obteniendo una frecuencia final. La información de la frecuencia para cada variante se expresó como su frecuencia cada 1000 individuos.

### 3.2.3. Evaluación de patogenicidad de variantes génicas

Con toda la información recopilada, se procedió a evaluar la patogenicidad de cada variante.

En primer lugar, se unificó la información de patogenicidad encontrada en publicaciones y bases de datos, registrando cada variante que estuviera clasificada como “benigna” o “patogénica”. Una misma variante clasificada de distintas formas por distintas publicaciones se registró como “evidencia conflictiva”.

La información de patogenicidad de las variantes registrada en publicaciones y en bases de datos se complementó con el criterio del *American College of Medical Genetics* (ACMG) [113] para la clasificación de variantes en uno de 5 posibles niveles de patogenicidad: Benigna (“*Benign*” según criterio de ACMG), Posiblemente benigna (“*Probably benign*” según criterio de ACMG), VSI (Variante de significado incierto, “*Variant of uncertain significance*” según criterio de ACMG), Probablemente patogénica (“*Possibly pathogenic*” según criterio de ACMG) y Patogénica (“*Pathogenic*” según criterio de ACMG).

La clasificación por los criterios de ACMG sólo se realizó para variantes que tuvieran algún tipo de evidencia que agregara información relevante para la clasificación. Esta evidencia pudo ser:

- Variantes con ensayos experimentales intentando medir su efecto biológico
- Variantes observadas en pacientes
- Variantes en el mismo aminoácido que otra variante patogénica o benigna
- Variantes reportadas en bases de datos poblacionales

#### 3.2.3.1. Patologías específicas asociadas con VGs de *NKX2-5*

Como es de particular interés el estudio de la presentación de subtipos de CC (y otras patologías asociadas a VGs en *NKX2-5*), se hizo foco en el análisis de las

patologías específicas asociadas a variantes patogénicas. Para esta parte del análisis, se utilizaron las variantes clasificadas como patogénicas en publicaciones y bases de datos.

Estas variantes se separaron en variantes de efecto en un aminoácido (VGs de cambio de sentido, inserciones/deleciones de un aminoácido sin cambio de marco de lectura), variantes con grandes cambios en la proteína (inserciones/deleciones de más de un aminoácido o con cambio de marco de lectura, variantes en sitios de *splicing*) y variantes sin efecto conocido en la proteína (variantes sinónimas o en regiones no codificantes).

Para cada uno de estos grupos, se registraron las patologías más representadas y, en el caso de las variantes de efecto en un aminoácido, se registró su distribución a lo largo del gen *NKX2-5* y/o la proteína NKX2-5.

#### 3.2.4. Formato de la información en la lista de VGs

Toda la información recopilada y procesada para las distintas VGs fue presentada en forma de tabla con cada fila siendo una variante única en el ADN y con columnas representando la información asociada a la variante.

La información presentada en esta tabla se presentó en 24 columnas, explicadas en detalle en la Tabla 2. Estas 24 columnas se dividen en 6 grupos:

- Identificadores estandarizados: “ADN (NC\_000005.10)”, “ADNc (NM\_004387.3)”, “Proteína (NP\_004378.1)”, “rs#”
- Información desglosada de posición y tipo de variante: “Región del gen”, “Tipo de variante por efecto en el ADN”, “Tipo de variante por efecto en la proteína”
- Fenotipos asociados: “Fenotipo asociado (cardíaco)”, “Fenotipo asociado (no cardíaco)”, “Con pacientes aparentemente sanos”, “Frecuencia (cada 1000 individuos)”, “Experimentos asociados”, “Resultado de experimentos asociados”
- Información asociada a la clasificación de la variante: “Conservación de residuo (sólo VGs de cambio de sentido)”, “SLiMs (sólo VGs de cambio de sentido)”

- Clasificación de la variante: Clasificación según ACMG, Clasificación por publicaciones y bases de datos
- Referencias: Referencias (PMID), Encontrada en dbSNP, Encontrada en SwissVar, Encontrada en LOVD, Encontrada en ExAC, Encontrada en EVS, Encontrada en pacientes del CNMG

**Tabla 2. Columnas utilizadas para registrar los datos correspondientes a las variantes génicas recopiladas.**

|                                |                                                         |                                                                                                              |
|--------------------------------|---------------------------------------------------------|--------------------------------------------------------------------------------------------------------------|
| Identificadores                | ADN (NC_000005.10)                                      | Variante en ADN (formato HGVS) sobre referencia hg19 del genoma humano                                       |
|                                | ADNc (NM_004387.3)                                      | Variante en ADNc (formato HGVS) sobre referencia de transcripto principal de <i>NKX2-5</i>                   |
|                                | Proteína (NP_004378.1)                                  | Variante en proteína (formato HGVS) sobre referencia de proteína isoforma 1 de <i>NKX2-5</i>                 |
|                                | rs#                                                     | Identificador estandarizado de HGVS                                                                          |
| Modelo estructural             | Región del gen                                          | Región genómica (UTR, Exón o Intrón)                                                                         |
|                                | Tipo de variante por efecto en el ADN                   | Efecto de la variante en el ADN (transposición, transversión, inserción, etc.)                               |
|                                | Tipo de variante por efecto en la proteína              | Efecto de la variante en la proteína (missense, nonsense, indel, etc.)                                       |
| Fenotipos asociados            | Fenotipo asociado (cardíaco)                            | Fenotipo cardíaco asociado a la variante                                                                     |
|                                | Fenotipo asociado (no cardíaco)                         | Fenotipo no cardíaco asociado a la variante                                                                  |
|                                | Con pacientes aparentemente sanos                       | Define si se encontró un individuo sano con la variante                                                      |
| Caracterización de la variante | SLIMs (sólo VGs de cambio de sentido)                   | Motivo lineal confirmado sobre el que se encuentra la variante                                               |
|                                | Conservación de residuo (sólo VGs de cambio de sentido) | Conservación del residuo predicha por ConSurf y SIFT, con criterio propio (Alta, Media o Leve)               |
|                                | Frecuencia (cada 1000 individuos)                       | Frecuencia de la variante en bases de datos poblacionales buscadas                                           |
|                                | Experimentos asociados                                  | Lista de experimentos realizados con la variante                                                             |
|                                | Resultado de experimentos asociados                     | Resultado de la lista de experimentos realizados con la variante (se corresponde con Experimentos asociados) |
| Clasificaciones                | Clasificación según ACMG                                | Clasificación de acuerdo al criterio de ACMG                                                                 |
|                                | Clasificación por publicaciones y bases de datos        | Clasificación de acuerdo a las referencias encontradas (criterio utilizado en Kolomenski et. al.)            |
| Referencias                    | Referencias (PMID)                                      | Lista de PMIDs de publicaciones que registran la variante                                                    |
|                                | Encontrada en dbSNP                                     | Marca si la variante se encontró en dbSNP                                                                    |
|                                | Encontrada en SwissVar                                  | Marca si la variante se encontró en SwissVar                                                                 |
|                                | Encontrada en LOVD                                      | Marca si la variante se encontró en LOVD                                                                     |
|                                | Encontrada en ExAC                                      | Marca si la variante se encontró en ExAC                                                                     |
|                                | Encontrada en EVS                                       | Marca si la variante se encontró en EVS                                                                      |
|                                | Encontrada en pacientes del CNMG                        | Marca si la variante se encontró en pacientes del CNMG                                                       |

VGs: Variantes génicas.

### 3.2.5. Carga de VGs a bases de datos

La clasificación de variantes según los criterios de ACMG y la compilación de variantes realizada de la literatura hizo evidente lo incompleto de dicha información en bases de datos públicas, por lo que se consolidó y se puso a disposición de forma abierta. En particular, se depositaron todas las variantes de nuestra lista a la base de

datos LOVD, proceso que requirió adaptar nuestra información al formato requerido por LOVD y completar información con el agregado de algunas columnas. En particular:

- Se explicitó el origen de cada variante (de publicación y/o de base de datos pública) con el formato requerido.
- Se duplicaron variantes que aparecieran en más de una publicación como nuevas entradas.
- Se asignó un identificador por paciente para variantes provenientes de publicaciones.
- Se modificaron fenotipos, efectos en patogenicidad, frecuencias y variantes a los en concordancia con los formatos aceptados por LOVD.

### 3.2.6. Generación de programas para visualizar VGs sobre *NKX2-5*

Con el objetivo de visualizar las VGs de cambio de sentido sobre el gen *NKX2-5* y sobre la proteína que codifica se desarrolló un programa en Python (<https://github.com/EmilioKolo/GraficoPyGame/>), con la biblioteca PyGame (<https://github.com/pygame/pygame/>). Este programa permitió ver, a grandes rasgos, la abundancia de distintos tipos de variantes en cada región de *NKX2-5*.

Se desarrolló un segundo programa en Python, con la biblioteca PyPlot de Matplotlib [160], para visualizar con mayor precisión a lo largo de la proteína codificada por *NKX2-5* la abundancia de aquellas variantes que se observan en población general y aquellas clasificadas como patogénicas de cambio de sentido. Para tal fin, se tomó una ventana de 3 aminoácidos (9 nucleótidos) alrededor de cada posición de la proteína y se registró la cantidad de variantes ubicadas en esta ventana (<https://github.com/EmilioKolo/GraficoVentanasGV/>).

Las variantes seleccionadas como potencialmente poblacionales fueron aquellas con frecuencia poblacional mayor a 1/10.000. Este corte se definió teniendo en cuenta que se estima un subdiagnóstico de CC en el 10% de los casos [23,161,162] y por la existencia de penetrancia incompleta [163-166], por lo que variantes con una frecuencia menor a 1/10.000 podrían representar variantes patogénicas no diagnosticadas o con penetrancia incompleta.

Por último, se utilizó la biblioteca JoyPy de Python (<https://github.com/leotac/joypy>) para visualizar la distribución de variantes que cambian uno o pocos aminoácidos y su asociación a subtipos específicos de patologías. Se utilizaron las variantes patogénicas asociadas a las patologías más comunes vistas en la sección 3.2.3.1. y se graficó un histograma de distribución de VGs por patología seleccionada. El conteo de variantes se realizó para cambios aminoacídicos únicos. Es decir, dos VGs distintas asociadas a la misma patología que causan el mismo cambio en aminoácidos de NKX2-5 fueron registradas como una sola variante. De esta forma, se pueden visualizar los cambios aminoacídicos que se han visto asociados a cada una de las patologías registradas, lo cual es de particular interés para el estudio de subtipos de CC.

### **3.3. Análisis de regiones específicas de NKX2-5**

Luego del análisis de VGs reportadas en la literatura y en bases de datos públicas, se estudió el potencial efecto sobre la funcionalidad de NKX2-5 de todas las posibles variantes de secuencia que producen un cambio en un aminoácido (variantes *missense*) en regiones de la proteína que poseen estructura tridimensional. Esto se realizó específicamente en la región con estructura tridimensional por aportar la información necesaria para los análisis *in silico* realizados.

El flujo de trabajo general correspondiente a estos estudios se esquematiza en la Figura 15.

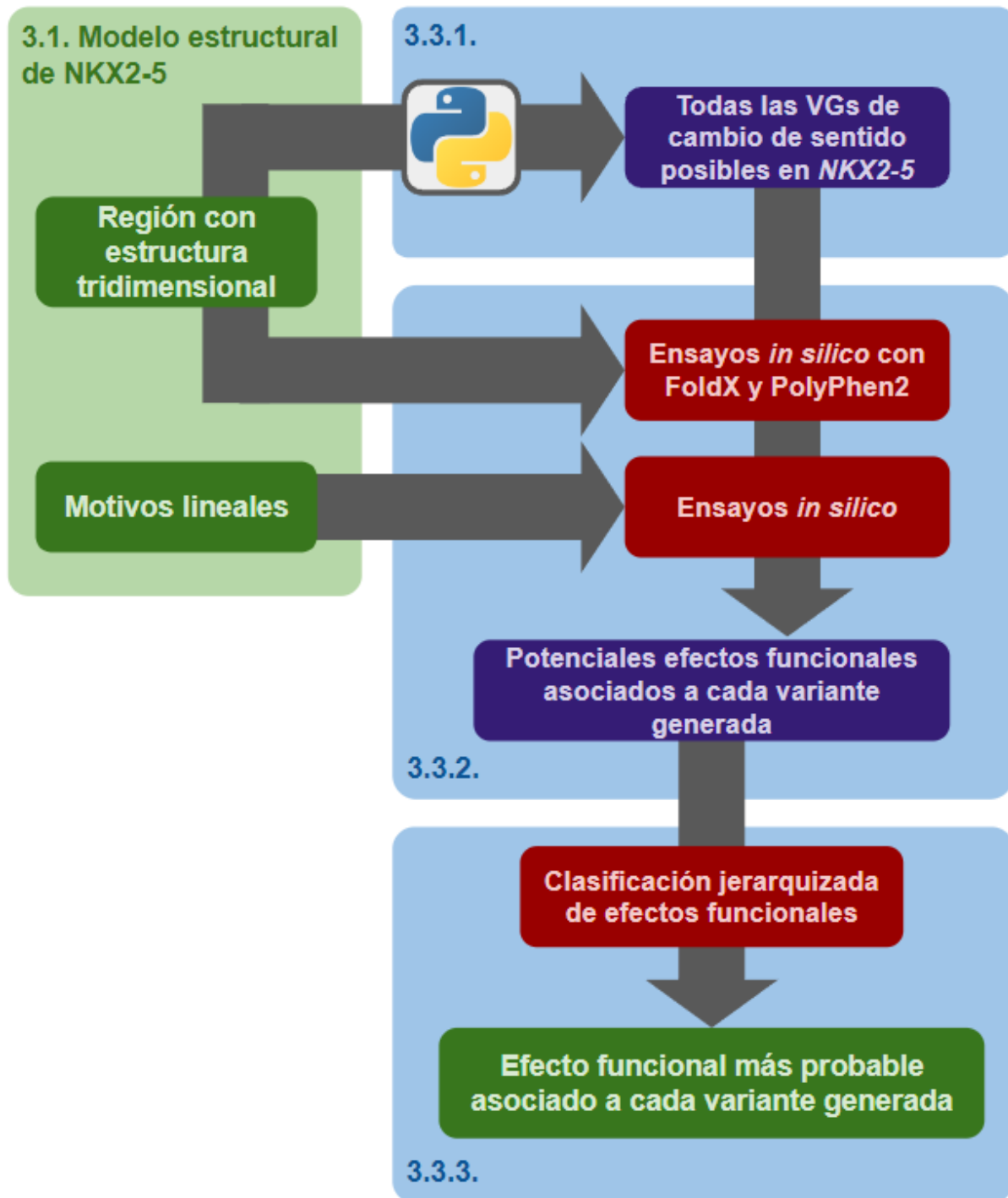


Figura 15. Flujo de trabajo general para el estudio del potencial efecto en la funcionalidad de NKX2-5 de todas las posibles variantes de cambio de un aminoácido (variantes *missense*). Los números en recuadros celestes se corresponden con las secciones de este trabajo donde se explica el paso correspondiente. VGs: Variantes génicas.

### 3.3.1. Creación de todas las VGs de cambio de sentido posibles en la región seleccionada de *NKX2-5*

En lugar de evaluar el efecto de cada uno de los 20 aminoácidos a lo largo de la secuencia, se desarrolló un programa en Python (<https://github.com/EmilioKolo/GeneracionVGs/>) con el objetivo de limitar la exploración a todas las variantes posibles que cambian un aminoácido en la secuencia de *NKX2-5* mediante el cambio en un nucleótido en la secuencia del gen (variantes *missense*).

La estrategia consistió en generar una lista de todas las variantes que cambian un aminoácido por triplete y seleccionar las posiciones correspondientes a la estructura tridimensional con mejor resolución para *NKX2-5*. Los cambios se realizaron a partir de la secuencia codificante de los exones de la isoforma canónica del gen *NKX2-5*.

### 3.3.2. Estudios *in silico* para cada VG generada

Para cada una de las variantes generadas se realizaron ensayos *in silico* para predecir su potencial efecto en la funcionalidad de *NKX2-5*. Estos ensayos incluyen predictores de conservación evolutiva (Polyphen2, ConSurf, SIFT) y funcionalidad (FoldX, ELM).

Se utilizó PolyPhen2 [167] para predecir el posible efecto de las variantes generadas en la funcionalidad de *NKX2-5* en base a información de conservación evolutiva. Esto se realizó por medio de un servicio web que permite correr PolyPhen2 para variantes en lote (<http://genetics.bwh.harvard.edu/pph2/bgi.shtml>). Esta clasificación se prefirió sobre la de ConSurf y SIFT por ser más específica al efecto de la conservación evolutiva en variantes en su asociación con patogénesis.

En segundo lugar, se utilizó FoldX [102] para identificar las variantes que potencialmente afecten la estructura de *NKX2-5*, su interacción con el ADN y/o su interacción con otras proteínas. Se recurrió a la función *BuildModel* para estimar el cambio de la energía de desplegado para cada variante con respecto a la proteína de referencia ( $\Delta\Delta G$ ) y a la función *AnalyseComplex* para el cambio de energía de interacción con el ADN y/u otras proteínas de cada variante. Se consideró que una

variante probablemente afectaría la funcionalidad de NKX2-5 con diferencias de energía mayores a 1,6 kcal/mol [101]. Esto se consideró para todas las variantes con *BuildModel* y sólo para variantes en contacto con el ADN y/u otras proteínas con *AnalyseComplex*.

Por último, se agregó información específica asociada a los cambios de secuencias causados por las variantes.

Por un lado, se utilizaron los SLiMs recopilados previamente para determinar variantes que afecten motivos lineales predichos y/o confirmados experimentalmente, los cuales están asociados a funciones específicas de una proteína.

Por otro lado, se registraron variantes que introducen cisteínas, ya que el agregado de cisteínas probablemente introduzca puentes disulfuro, los cuales desestabilizarían la estructura proteica de forma significativa. En el caso que la cisteína se introdujera en un residuo sin exposición al solvente, sólo se consideró la variante como potencialmente causal de patogénesis si además introdujera un cambio significativo de energía de *unfolding* en las predicciones de FoldX *BuildModel*.

### 3.3.3. Clasificación de potenciales efectos de las variantes generadas en la funcionalidad de NKX2-5

Con los resultados de la sección anterior, se obtuvo información de 4 potenciales explicaciones de desarrollo de patogenicidad para las variantes generadas:

- Desestabilización de la estructura terciaria. Predicha por FoldX *BuildModel*.
- Efecto en la interacción con el ADN y/u otras proteínas y, por lo tanto, en la función como factor de transcripción. Predicha por cambios de energía en FoldX *AnalyseComplex* y observaciones de residuos que interactúan en el modelo estructural.
- Eliminación de un SLiM, lo cual afecta a la función específica asociada a este motivo. Esto se realizó utilizando la información ya obtenida del modelo estructural.

- Introducción de una cisteína, la cual no permitiría el correcto plegado de la proteína por la introducción de puentes disulfuro. Esto se realizó por selección de las variantes generadas.

La predicción de PolyPhen2 no está asociada a una vía funcional, por lo que no aporta una vía de desarrollo de patogénesis, y sólo sugiere el riesgo asociado a una variante en particular.

Estas clasificaciones no son mutuamente excluyentes, y algunas variantes pueden tener múltiples causas asociadas, por lo que se desarrolló una jerarquía de potenciales efectos en la función de NKX2-5. De esta forma, si una variante tiene asociados dos o más potenciales efectos en la función de NKX2-5, sólo se considera el efecto más alto en la jerarquía, y se simplifica el análisis de la distribución de potenciales efectos de las variantes generadas.

En primer lugar, se consideraron aquellas variantes que afectan la interacción con el ADN. Estas fueron consideradas en primer lugar por consistir de la función principal de un factor de transcripción, como lo es NKX2-5.

En segundo lugar, se consideraron variantes que introducen una cisteína a la secuencia de la proteína. Estas variantes casi seguramente desestabilizarían la proteína de forma que no le permitirían realizar su función.

En tercer lugar, se consideraron variantes que afectan residuos que participan en la interacción con otros factores de transcripción o que interrumpieran motivos lineales confirmados experimentalmente. Esto se corresponde con funciones específicas que pueden o no ser cruciales en la función general de la proteína.

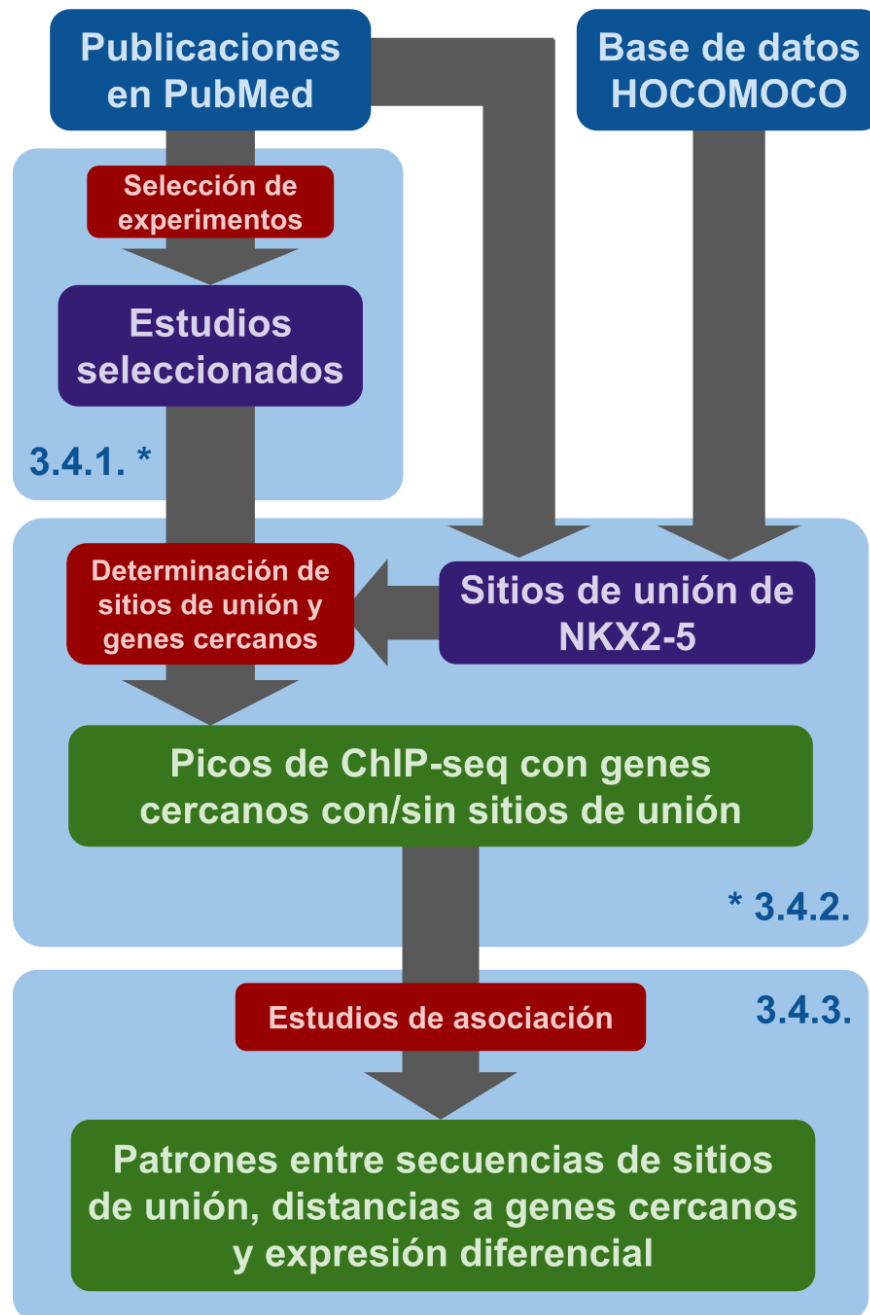
Por último, se registraron variantes que tuvieran un efecto sobre la estabilidad de NKX2-5 predicho por FoldX o que interrumpieran un motivo lineal predicho por ELM. Estas se ubican en último lugar por representar efectos posiblemente leves en la función proteica y por tener la mayor probabilidad de representar falsos positivos en los predictores.

### **3.4. Recopilación y estudio de todos los sitios de unión de NKX2-5 en el genoma humano y de ratón**

En esta parte del trabajo, y con el objetivo de analizar sitios de unión de NKX2-5 a nivel genómico, se desarrollaron programas para procesar resultados provenientes de estudios de inmunoprecipitación de la cromatina seguida de secuenciación masiva (ChIP-seq) en formatos que puedan usarse para estudiar con otras plataformas (MEME-ChIP, GREAT).

Esta estrategia se aplicó para los sitios de unión de NKX2-5 en genomas de humano y/o ratón, según la evidencia que se encontrara.

El flujo de trabajo general correspondiente a la recopilación y estudio de todos los sitios de unión de NKX2-5 se esquematiza en la Figura 16.



**Figura 16. Flujo de trabajo general para la recopilación y estudio de todos los sitios de unión de NKX2-5.** Los números en recuadros celestes se corresponden con las secciones de este trabajo donde se explica el paso correspondiente. El genoma de referencia utilizado fue GRCh37.p13 (hg19). Asteriscos: Datos que se obtuvieron para ratón además de humano.

### 3.4.1. Selección de resultados

En primer lugar, se realizó una búsqueda de experimentos de ChIP-seq en PubMed (<https://pubmed.ncbi.nlm.nih.gov/>), haciendo foco en que estén relacionados a estructuras del corazón. Se registraron los archivos .bed con sitios de unión encontrados. Asimismo, se recopilaron las tablas con resultados de estudios de expresión diferencial de RNA-seq entre la condición *wild type* y la condición *knock out* o *knock down* para NKX2-5 que acompañaron los estudios de ChIP-seq.

### 3.4.2. Recopilación de sitios de unión de NKX2-5

De los resultados de experimentos de ChIP-seq, se extrajeron las posiciones en el genoma registradas en los archivos .bed (picos o *peaks*). Luego, se buscaron los genes cerca de estos picos, definidos como genes a cierta distancia de cada extremo del pico. Esta selección se repitió para cuatro distancias máximas: 1500 pb, 100.000 pb, 500.000 pb y 1.000.000 pb. Asimismo, se utilizaron 2 criterios para la selección de los genes: seleccionar todos aquellos que estuvieran contenidos hasta la distancia definida y seleccionar el gen más cercano a cada lado del sitio de unión. Para seleccionar todos los genes se usó un programa de Python personalizado (<https://github.com/EmilioKolo/ChIP-seq-analysis-scripts/>, función *pipeline\_generador*), mientras que para seleccionar los genes más cercanos se usó el programa GREAT (<https://great.stanford.edu/great/public/html/>) [168,169].

Se definieron como picos “confirmados” a aquellos con genes a menos de 100.000 pares de bases de distancia para los cuales se observó una expresión diferencial significativa.

Los genes se buscaron a una distancia máxima de 100.000 pb para incluir a los genes conocidos que son transactivados por NKX2-5 a mayor distancia [170-172].

Dentro de estos picos se buscaron sitios de unión de NKX2-5. Para tal fin, se definió una lista de secuencias consenso de sitios de unión utilizados experimentalmente en ensayos de transactivación.

Además, se descargó una matriz de puntaje de secuencias que potencialmente sean sitios de unión de NKX2-5 de la base de datos HOCOMOCO versión 11 (<https://hocomoco11.autosome.org/>) [129]. La matriz de puntajes utilizada fue la matriz de nombre "NKX25\_HUMAN.H11MO.0.B" ([https://hocomoco11.autosome.org/motif/NKX25\\_HUMAN.H11MO.0.B](https://hocomoco11.autosome.org/motif/NKX25_HUMAN.H11MO.0.B)). Se utilizó un puntaje de corte para identificar sitios de unión del 90% del puntaje máximo posible de la matriz.

### 3.4.3. Estudio de asociación de sitios de unión de NKX2-5 cerca de genes con expresión diferencial

Con el objetivo de encontrar una asociación entre la condición de regulación de genes a menos de 100.000 pb de sitios de unión de NKX2-5 y el sitio de unión en particular, se realizaron distintas búsquedas de patrones entre los datos asociados a sitios de unión. Para esto, se trabajó sobre sitios de unión de NKX2-5 encontrados en picos de ChIP-seq confirmados por tener genes con expresión diferencial significativa en los ensayos de RNA-seq a menos de 100.000 pb de distancia. Los datos registrados fueron:

- La distancia del sitio de unión al gen con expresión diferencial significativa en los ensayos de RNA-seq
- La secuencia nucleotídica del sitio de unión
- El valor de aumento o disminución de expresión de cada gen con expresión diferencial significativa en los ensayos de RNA-seq

En primer lugar, se buscaron patrones simples graficando dos combinaciones de valores con la biblioteca Matplotlib de Python [160].

Para cada secuencia posible de sitios de unión se analizó la cantidad de genes con expresión diferencial significativa en los ensayos de RNA-seq, los cuales se dividieron según si eran genes con expresión que aumentaba o disminuía en presencia de NKX2-5. El objetivo fue analizar si alguna de las secuencias específicas se encuentra asociada con activación o inhibición de genes cercanos.

También se analizó la cantidad de genes con expresión diferencial significativa en los ensayos de RNA-seq a distintas distancias de sitios de unión de NKX2-5. En este caso, también se agruparon según la proporción de genes que aumentan o disminuyen su expresión en presencia de NKX2-5. Para esto, se incluyeron sitios de unión en picos de ChIP-seq confirmados por tener genes con expresión diferencial significativa en los ensayos de RNA-seq a menos de 100.000 pb.

Por último, se realizó un análisis de componentes principales con el módulo Cluster de BioPython [173], y se buscaron los componentes que explican mayor variabilidad para analizar sus elementos principales. Las variables analizadas fueron las secuencias de los sitios de unión, las distancias al gen con expresión diferencial más cercano y el aumento o disminución de la expresión del gen cercano por NKX2-5.

Se utilizaron los sitios de unión con genes confirmados cercanos. Los sitios de unión con dos genes cercanos confirmados se procesaron como dos sitios distintos. Todos los valores fueron normalizados y las variables analizadas se pre-procesaron de formas particulares, como se ejemplifica a continuación.

Las distancias se procesaron de 4 formas distintas:

- Con el valor de distancia sin modificar
- Como logaritmo de la distancia
- Como inversa (1/distancia)
- Como inversa del logaritmo de la distancia

El aumento o disminución de expresión por NKX2-5 se procesó de 2 formas distintas:

- Con el valor de *fold change* del gen cercano (obtenido en el estudio RNA-seq para la condición *wild type*/condición *knock out*) sin modificar
- Como -1 si el gen aumenta su expresión por NKX2-5 y como 1 si la disminuye (correspondiente a la función signo aplicada sobre el valor *fold change*)

Las secuencias de los sitios de unión se procesaron como una lista de variables, teniendo en cuenta la cantidad de secuencias únicas encontradas, donde cada

secuencia en particular tiene valor 1 en una de las variables y 0 en las demás (codificación *one-hot*).

Esto da un total de 8 conjuntos diferentes de datos utilizados para el análisis, de los cuales se seleccionó uno para estudiar asociaciones que pudieran haber entre los distintos valores estudiados. El objetivo del conjunto seleccionado fue separar lo mejor posible las secuencias de sitios de unión y sus distancias por genes según su transactivación utilizando un máximo de 3 componentes principales.

### **3.5. Estudio de sitios de unión cercanos de otros factores de transcripción y genes transactivados**

Con los sitios de unión confirmados en la sección anterior, se estudiaron los genes cercanos a estos sitios y se definieron otros factores de transcripción que pudieran interactuar con NKX2-5 para, potencialmente, mediar en la transactivación o inhibición de estos genes.

Los genes se estudiaron por medio de sus Gene Ontology terms (GO-terms) en relación a los sitios de unión de los distintos factores de transcripción definidos que pudieran estar transactivándolos o inhibiéndolos.

El flujo de trabajo general correspondiente al estudio de factores de transcripción que interactúan con NKX2-5 y sus genes transactivados se esquematiza en la Figura 17.



**Figura 17. Flujo de trabajo general para el estudio de sitios de unión cercanos de otros factores de transcripción y genes transactivados.** Los números en recuadros celestes se corresponden con las secciones de este trabajo donde se explica el paso correspondiente. El genoma de referencia utilizado fue GRCh37.p13 (hg19). CC: Cardiopatías congénitas. VGs: Variantes génicas.

### 3.5.1. Sitios de unión de otros factores de transcripción cercanos a sitios de NKX2-5

Para interpretar la información asociada a otros factores de transcripción que posiblemente interactúen con NKX2-5, se comenzó por definir la lista de estos factores de transcripción y se buscaron sus sitios de unión cercanos a los sitios de unión confirmados de NKX2-5.

En primer lugar, se hizo una búsqueda exploratoria de factores de transcripción que tuvieran sitios de unión sobrerrepresentados en los picos de ChIP-seq. Esto se hizo con ayuda del programa MEME-ChIP versión 5.4.1 (<https://meme-suite.org/meme/tools/meme-chip/>) [174], que busca regiones sobrerrepresentadas en una lista de secuencias y trata de asociarlas con sitios de unión conocidos. Para esto, el programa requiere una lista de secuencias del mismo largo, las cuales se obtuvieron de los archivos .bed con picos de ChIP-seq utilizando un programa de Python (<https://github.com/EmilioKolo/ChIP-seq-analysis-scripts/>, función *pipeline\_meme\_chip*). Con este programa, se definió un rango de 500 pares de bases alrededor del centro de cada pico de ChIP-seq (250 pares de bases a cada lado) y se obtuvo su secuencia de referencia en el genoma humano. El largo de 500 pares de bases utilizado para estandarizar el largo de las secuencias es el recomendado por MEME-ChIP.

Una vez obtenidas las secuencias, éstas se procesaron con el servidor web de MEME-ChIP. Se utilizó un archivo .fasta con las secuencias seleccionadas y se cambiaron dos de los valores que tiene por defecto el programa:

- Se utilizó “*Human and Mouse (HOCOMOCO v11 full)*” como lista de motivos de búsqueda. De esta forma, se trabajó específicamente con las matrices de HOCOMOCO.
- En “*MEME options*”, se seleccionó “*Any number of repetitions*” como distribución esperada de los motivos. Esta opción puede encontrar motivos únicos, pero aumenta la sensibilidad para motivos repetidos.

Con estos parámetros, se obtuvo una tabla de sitios sobrerrepresentados y los factores de transcripción que los reconocerían, si los hubieran. El programa lista los sitios ordenados por E-valor de menor a mayor. El E-valor aproxima qué tan probable

sea que un dado motivo se haya encontrado por azar en la lista de secuencias, por lo que, mientras menor sea, más probable es que el motivo esté presente por razones funcionales.

Para establecer un criterio de selección de la lista de factores de transcripción que devuelve MEME-ChIP, se definió que el punto de corte fuera el E-valor de un factor de transcripción con evidencia experimental de interacción con NKX2-5, o el E-valor de NKX2-5, si este fuera un valor mayor.

La lista de factores de transcripción que pudieran interactuar con NKX2-5 se definió con una búsqueda en bases de datos de interacciones entre proteínas. Las bases de datos utilizadas fueron STRING (<https://string-db.org/>) [141] y GeneMANIA (<https://genemania.org/>) [140].

Con la lista de factores de transcripción que pudieran interactuar con NKX2-5, se buscó aquellos que se encontraran entre los factores de transcripción que devuelve MEME-ChIP.

Entre los factores de transcripción con evidencia experimental de interacción con NKX2-5 que se encontraron en los resultados de MEME-ChIP, se seleccionó el que obtuvo mayor E-valor, y se utilizó ese número como criterio de corte, registrando todos los factores de transcripción con menor puntaje de los resultados de MEME-ChIP.

Dentro de los factores de transcripción con E-valor menor, se descartaron aquellos que pertenecieran a la familia NK2 (de NKX2-5) o a las familias de los factores de transcripción con evidencia experimental de interacción con NKX2-5. Esto es para eliminar falsos positivos debido a que factores de transcripción de la misma familia tienen sitios de unión similares. En los casos que se encontraran varios factores de transcripción de una misma familia, o que hubieran varias matrices asociadas a un mismo factor de transcripción, se seleccionó el que tuviera menor E-valor. Esto se hizo para simplificar el análisis, ya que los sitios de unión de factores de transcripción de la misma familia suelen ser similares entre sí.

Con los factores de transcripción seleccionados y filtrados, se descargaron matrices de puntaje de HOCOMOCO versión 11 (<https://hocomoco11.autosome.org/>) [129], las cuales se utilizaron para definir sitios de unión de cada factor de transcripción. Con estas matrices, se utilizó el mismo criterio de selección que para NKX2-5 en la sección 3.4.2., utilizando un 90% del puntaje máximo posible de la matriz como corte.

Utilizando estas matrices, se buscaron todos sus sitios de unión a menos de cierta distancia de sitios de unión de NKX2-5 confirmados experimentalmente. Esto se hizo con un programa personalizado de Python desarrollado para este trabajo (<https://github.com/EmilioKolo/ChIP-seq-analysis-scripts/>, función *pipeline\_otros\_tf*). Se utilizó una distancia de 250 pares de bases por ser la mitad del largo establecido como estándar por MEME-ChIP, en las cuales asumimos que el sitio de unión de NKX2-5 está cerca del centro de los picos de ChIP-seq.

Los sitios de unión de estos factores de transcripción fueron registrados junto con los genes confirmados cercanos, su condición de regulación por NKX2-5 y las distancias a estos genes.

### 3.5.2. Análisis de genes cercanos a sitios de unión seleccionados

Además del análisis de otros factores de transcripción, se hizo un análisis de los genes a menos de 100.000 pb de distancia de los sitios de unión seleccionados. Con estos genes, se definieron distintas listas de genes cerca de sitios de unión de NKX2-5 y de los factores de transcripción seleccionados en la sección anterior. Estas listas representan a los genes a menos de 100.000 pb de un sitio de unión de NKX2-5 que esté a menos de 250 pb de un sitio de unión de otro factor de transcripción.

Cada una de estas listas se utilizó 3 veces, generando listas con:

- Genes con expresión diferencial aumentada en presencia de NKX2-5 (regulados positivamente o *upregulados*)
- Genes con expresión diferencial disminuida en presencia de NKX2-5 (regulados negativamente o *downregulados*)

- Todos los genes en los que se observó expresión diferencial y representarían los genes potencialmente regulados por *NKX2-5* (totales)

Esto dio un total de 3 listas de genes por factor de transcripción seleccionado, más 3 listas para el total de sitios de unión de *NKX2-5*.

Los genes analizados se obtuvieron de los archivos de sitios de unión generados por la función de Python utilizada en la sección 3.5.1. (<https://github.com/EmilioKolo/ChIP-seq-analysis-scripts/>, función *pipeline\_otros\_tf*) en formato apto para su procesamiento por el servicio web de ShinyGO (<http://bioinformatics.sdstate.edu/go/>) [175]. Este servicio busca sobrerrepresentación de funciones y vías metabólicas, representadas por GO-terms, y devuelve listas de GO-terms con sus E-valores correspondientes. Se buscó específicamente “*GO Biological Process*” en la opción “*Pathway database*”, para obtener GO-terms de funciones en procesos biológicos.

Con cada una de las listas de genes analizadas por ShinyGO se obtuvieron GO-terms que fueron analizados en el resto del trabajo. De la lista de GO-terms obtenida, el servicio web de ShinyGO selecciona un máximo de GO-terms de acuerdo a su enriquecimiento y la probabilidad de falsos positivos (FDR, por sus siglas en inglés para “*False Discovery Rate*”). Se registró un máximo de 10 GO-terms por lista de genes, clasificando sus FDR en 3 rangos:

- $FDR > 0,001$ : Representa GO-terms no significativos, generalmente muy específicos y con pocos genes, donde estos pocos genes representan un aumento significativo de enriquecimiento relativo. Para estos GO-terms, se hizo foco en genes específicos asociados a funciones específicas.
- $0,001 \geq FDR > 10^{-5}$  (1/100.000): Representa GO-terms no significativos, pero que se encuentran representados por numerosos genes.
- $FDR \leq 10^{-5}$  (1/100.000): Representa GO-terms significativos según el criterio recomendado por ShinyGO. Estos son los GO-terms que muy probablemente representen una función relevante en común para la lista de genes estudiada.

En primer lugar, se estudiaron las listas de todos los genes confirmados experimentalmente sin filtrar por sitios de unión de otros factores de transcripción. Sus GO-terms sobrerrepresentados se agruparon en categorías para simplificar el análisis. Esto se repitió para las listas de genes con expresión disminuida o aumentada en los ensayos de expresión diferencial.

En una segunda etapa, se analizaron los GO-terms que se obtienen para las distintas listas filtradas por sitios de unión de los factores de transcripción seleccionados. Se registró si se agrupaban en las mismas categorías o si algún GO-term representaba una categoría distinta no vista en la lista sin filtrar, y se estudió su distribución relativa para los distintos factores de transcripción.

Finalmente, se registraron genes de interés tanto los correspondientes a GO-terms con pocos genes asociados, como genes cerca de sitios de unión para múltiples de los factores de transcripción seleccionados.

### 3.5.3. Análisis de genes seleccionados en relación a subtipos de CC

Para finalizar, se estudiaron los genes seleccionados en las secciones anteriores y su asociación a subtipos de CC.

Los genes seleccionados fueron genes co-cristalizados con NKX2-5 (sección 3.1.1.), genes con interacción confirmada específicamente con NKX2-5 (sección 3.5.1.), factores de transcripción seleccionados entre los predichos por MEME-ChIP (sección 3.5.1.) y genes de interés transactivados o inhibidos por *NKX2-5* (sección 3.5.2.).

Para cada uno de estos genes, se registró su potencial asociación al desarrollo de CC y al desarrollo de otras patogenicidades asociadas a VGs de *NKX2-5* vistas en este trabajo (sección 3.2.3.1.).

De esta forma se obtiene una compilación de genes con distintos tipos de relaciones a NKX2-5 y a patogénesis específicas, la cual puede ser comparada con la relación entre estas patogénesis específicas y VGs de *NKX2-5*. Esta información, a su vez, puede ser comparada con VGs de la región estructurada de *NKX2-5* que potencialmente interactúen con otros factores de transcripción (sección 3.3.3.).

De esta forma, esperamos achicar la brecha entre la información de genotipos de *NKX2-5* en pacientes y sus fenotipos asociados, haciendo foco en la búsqueda de vías funcionales en alguna de las patogénesis específicas asociadas a VGs en *NKX2-5*.

## **4. Resultados**

El objetivo de este trabajo consiste en desarrollar y ejecutar un flujo de trabajo para el estudio de *NKX2-5* y su relación con el desarrollo de cardiopatías congénitas. Tal como los materiales y métodos, los resultados de este trabajo se dividen, a grandes rasgos, en 5 partes:

1. Desarrollo de un modelo estructural de *NKX2-5*. Este modelo sirvió de base para los demás estudios.
2. Recopilación de variantes génicas de *NKX2-5* y análisis de su relación con CC. Se estudiaron motivos lineales afectados, frecuencias poblacionales y distribución de las variantes en relación a sus cardiopatías congénitas asociadas.
3. Análisis específico de la región estructurada de la proteína codificada por *NKX2-5* (el homeodominio), estudiando todas las variantes de cambio de sentido posibles en esta región.
4. Recopilación de sitios de unión de *NKX2-5* en el genoma humano y de ratón y análisis de su potencial asociación con genes transactivados o inhibidos por *NKX2-5*.
5. Estudio de sitios de unión cercanos de otros factores de transcripción y análisis de la función de genes transactivados o inhibidos por *NKX2-5*.

Toda esta información apunta a dilucidar mecanismos potencialmente involucrados en el desarrollo de cardiopatías congénitas y, en muchos casos, abren puertas para futuros estudios.

Al final de estas 5 secciones, se agrega una sexta, donde se propone un flujo de trabajo que pueda ser potencialmente utilizado para estudiar otros genes de la misma forma que se estudió *NKX2-5*.

## 4.1. Modelo estructural de NKX2-5

El modelo estructural obtenido incluye la identificación de una región estructurada (el homeodominio) entre dos regiones desestructuradas. Dentro de la región estructurada, se hizo foco en la clasificación de cada aminoácido según contactos, conservación evolutiva y accesibilidad al solvente. En la región desestructurada (y regiones accesibles de la región estructurada) se buscaron y confirmaron motivos lineales, que aportan información funcional sin necesidad de una estructura asociada.

Se utilizó, a grandes rasgos, el modelo estructural creado y estudiado en [68], que sentó bases sobre las que se realizó este trabajo.

### 4.1.1. Selección de estructuras tridimensionales correspondientes a NKX2-5

A partir de la búsqueda de estructuras tridimensionales de NKX2-5 realizada en Protein Data Bank (PDB), se hallaron 5 estructuras, todas ellas correspondientes al homeodominio y obtenidas por cristalografía de rayos X (Tabla 3). En todas las estructuras, el factor de transcripción se encuentra en conjunto con cadenas de ADN y como homo- o heterodímero con otros factores de transcripción, y se reemplaza la cisteína 193 por una serina.

**Tabla 3. Comparación entre estructuras tridimensionales de NKX2-5 encontradas en PDB.**

| PDB ID                     | 3RKQ                      | 4S0H                      | 5FLV                       | 6WC2                               | 6WC5                      |
|----------------------------|---------------------------|---------------------------|----------------------------|------------------------------------|---------------------------|
| Método de obtención        | Cristalografía de rayos X | Cristalografía de rayos X | Cristalografía de rayos X  | Cristalografía de rayos X          | Cristalografía de rayos X |
| Resolución (Å)             | 1,7                       | 2,817                     | 3,00                       | 2,1                                | 2,9                       |
| Región correspondiente     | Homeodominio (138-194)    | Homeodominio (142-194)    | Homeodominio (134-197)     | Homeodominio (137-197)             | Homeodominio (140-196)    |
| Moléculas de la estructura | NKX2-5, ADN               | NKX2-5, TBX5, ADN         | Quimera NKX2-5 + TBX5, ADN | NKX2-5, Quimera MEF2B + MEF2A, ADN | NKX2-5, MEF2B, ADN        |
| Bibliografía               | [176]                     | [146]                     | [177]                      | [178]                              | [178]                     |

Para los estudios posteriores, se hizo énfasis en 3 estructuras en particular:

- 3RKQ como estructura con mejor resolución de NKX2-5 en general.
- 4S0H como estructura con mejor resolución de NKX2-5 y TBX5.
- 6WC5 como estructura no quimérica de NKX2-5 y MEF2B.

Entre las estructuras seleccionadas, se estudiaron 58 aminoácidos del homeodominio correspondientes a las posiciones 137-194 de NKX2-5. La definición de residuos que forman parte del *core* hidrofóbico se hizo con la estructura tridimensional de PDB ID 3RKQ, la definición de residuos que interactúan con TBX5 o MEF2B se hicieron con las estructuras PDB ID 4S0H y 6WC5, respectivamente, y la definición de residuos que interactúan con el ADN se hizo con las 3 estructuras.

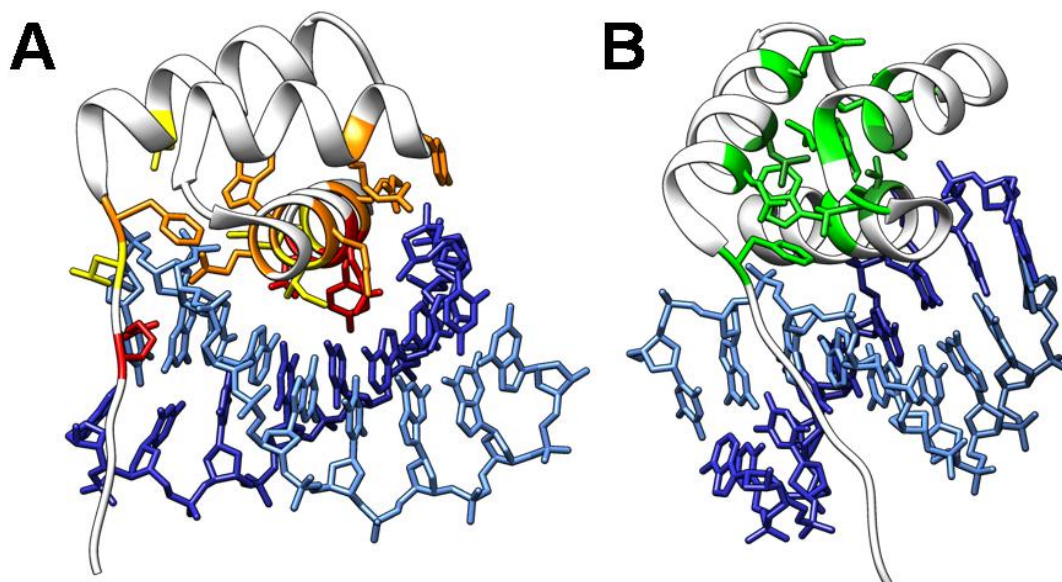
Los estudios de residuos conservados evolutivamente y la predicción de motivos lineales se realizan considerando únicamente la secuencia, aunque se confirmó la accesibilidad al solvente de residuos con motivos lineales utilizando la estructura tridimensional de PDB ID 3RKQ. Estos análisis se hicieron para los 58 aminoácidos definidos anteriormente (residuos 137 a 194).

#### 4.1.2. Estudio de regiones con estructura terciaria

Entre los 58 residuos estudiados, se definieron 13 en contacto con el ADN (más 4 potenciales contactos), 12 formando parte del *core* hidrofóbico y 5 que potencialmente interactúan con otras proteínas. Las 3 estructuras estudiadas aportaron los mismos residuos en contacto con el ADN.

De los 17 residuos potencialmente en contacto con el ADN, en 4 se observó contacto con una base nitrogenada (residuos 142, 187, 188 y 191), 9 están en contacto con grupos fosfatos (residuos 143, 145, 162, 163, 168, 181, 183, 185 y 190) y 4 tienen cadenas laterales a menos de 5 Angstroms del ADN (residuos 144, 150, 184, 192), por lo que algunas VGs podrían introducir contactos con el ADN (Figura 18A).

De los 12 residuos que forman parte del *core* hidrofóbico, 4 se encuentran inaccesibles al agua (residuos 153, 172, 182, 186) y 8 están en contacto con estos residuos (residuos 145, 149, 156, 163, 171, 175, 177 y 185) (Figura 18B).



**Figura 18. Esquema del homeodominio de NKX2-5 en contacto con el ADN, correspondiente a aminoácidos 137-194 del cristal PDB-ID 3RKQ, coloreado según el resultado de la clasificación de distintos residuos de acuerdo a su ubicación. En azul, el ADN. A) Se resaltan 17 residuos en contacto con el ADN. En rojo, residuos en contacto con las bases nitrogenadas. En naranja, residuos en contacto con grupos fosfato. En amarillo, residuos a menos de 5 Angstroms del ADN. B) Se resaltan en verde 12 residuos que forman parte del core hidrofóbico.**

Se identificaron 5 residuos de NKX2-5 que potencialmente interactúan con otras proteínas: 2 residuos en contacto con TBX5 (residuos 158 y 193) (Figura 19), de los cuales Lys158 posee contacto mediante su cadena lateral (Figura 19A) y 3 que potencialmente estarían en contacto con MEF2B (residuos 144, 191 y 194) (Figura 20), de los cuales Leu144 y Tyr191 tendrían contacto mediante su cadena lateral (Figura 20A y 20B). No se observó contacto entre distintas cadenas de NKX2-5 en ninguno de los cristales.

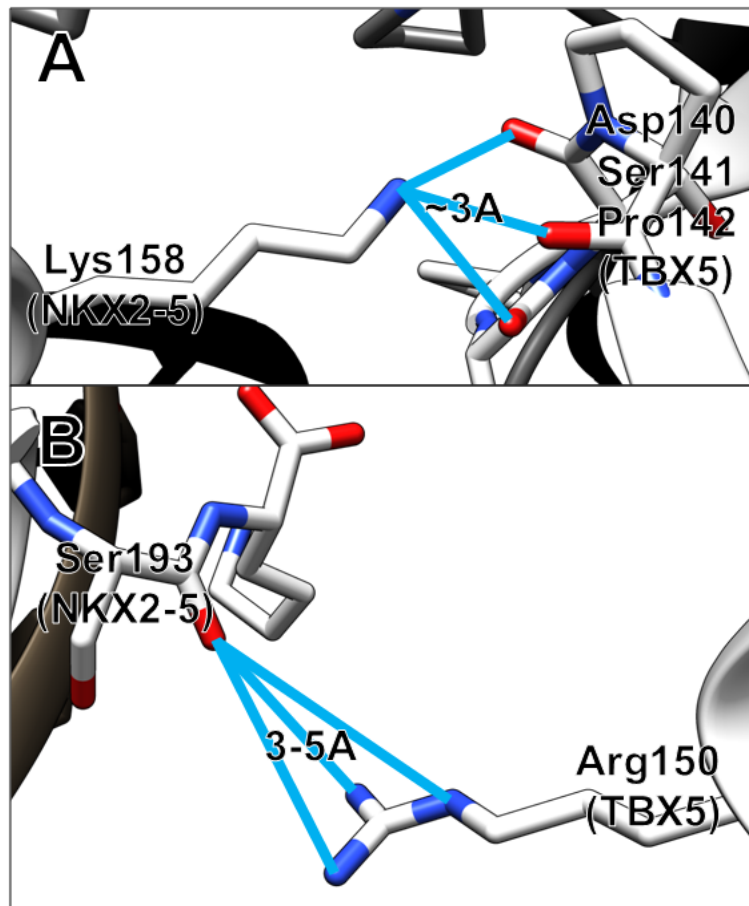
Sólo para 3 residuos se identificó una posible interacción específica:

- El potencial contacto de Lys 158 de NKX2-5 ocurre entre el grupo amino de la lisina y grupos carboxilos de aminoácidos de TBX5, con una distancia de alrededor de 3 Angstroms (Figura 19A), por lo que podría representar una interacción de Coulomb.
- El potencial contacto de la Lys 194 de NKX2-5 ocurre entre su grupo carboxilo y el grupo carboxilo de la cadena lateral del Asp 13 de MEF2B, con una distancia

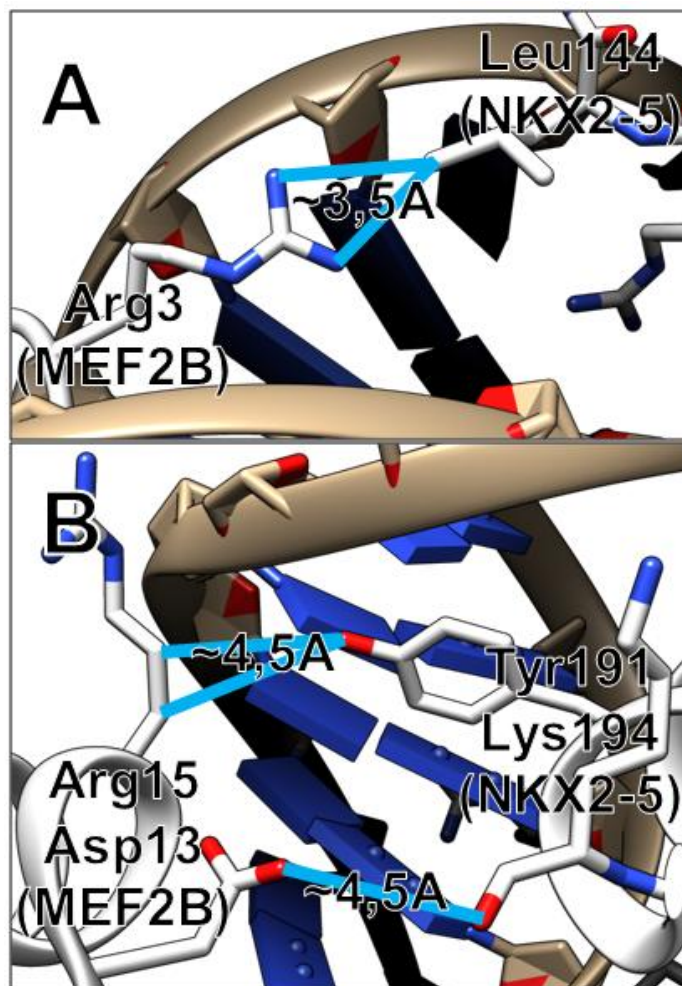
de 4,5 Angstroms (Figura 20B, abajo), por lo que podría representar una interacción de puente de hidrógeno mediada por moléculas de agua.

- El potencial contacto de la Ser 193 de NKX2-5 ocurre entre su grupo carboxilo y los 3 grupos amino de la cadena lateral de la Arg 150 de TBX5, con distancias entre 3 y 5 Angstroms (Figura 19B), por lo que podría representar interacciones de puente de hidrógeno mediadas por moléculas de agua.

Los otros 2 potenciales contactos ocurrirían entre átomos no cargados y átomos cargados, a distancias mayores a 3 Angstroms (Figura 19A y Figura 19B, arriba), por lo que podrían no representar interacciones en su heterodímero.



**Figura 19. Esquema de aminoácidos de NKX2-5 a menos de 5 Angstroms de aminoácidos de TBX5.** En azul, átomos de nitrógeno (carga positiva). En rojo, átomos de oxígeno (carga negativa). A) Distancia entre el nitrógeno de la cadena lateral de la lisina 158 y los oxígenos de los grupos carboxilo de los residuos 140, 141 y 142 de TBX5. B) Distancia entre el grupo carboxilo de la serina 193 de NKX2-5 (variante de la cisteína *wild type* usada en los cristales) y los nitrógenos de la cadena lateral de la arginina 150 de TBX5.



**Figura 20. Esquema de aminoácidos de NKX2-5 a menos de 5 Angstroms de aminoácidos de MEF2B.** En azul, átomos de nitrógeno (carga positiva). En rojo, átomos de oxígeno (carga negativa). En dorado, cadena de ADN. A) Distancia entre la cadena lateral de la leucina 144 de NKX2-5 y la cadena lateral de la arginina 3 de MEF2B. B) Distancias entre las cadenas laterales de la tirosina 191 y el grupo carboxilo de la lisina 194 de NKX2-5 y la cadena lateral de la arginina 15 y el ácido aspártico 13 de MEF2B, respectivamente.

Cabe destacar que las potenciales interacciones entre átomos que no forman parte de la cadena lateral de los residuos correspondientes podrían no verse afectadas por variantes en esos residuos. Por otro lado, los potenciales contactos sin interacciones específicas podrían ser relevantes si el residuo correspondiente tuviera una variante con una cadena lateral que sí incluyera una interacción.

Todas las interacciones consideradas relevantes fueron entre residuos de NKX2-5 y residuos de otras proteínas de un dímero en *cis* (ambas proteínas en contacto con la misma hebra de ADN), pero en las estructuras tridimensionales de heterodímeros

con TBX5 y MEF2 se pueden apreciar interacciones entre residuos de NKX2-5 con otros factores de transcripción en *trans* (en contacto con distintas hebras de ADN). Estas interacciones estuvieron mediadas por los residuos 158, 196 y 197 para TBX5 y 155, 156, 162, 167 y 170 para MEF2B. Sin embargo, no fueron consideradas como relevantes por haber una mayor posibilidad de que sean forzadas por las condiciones de formación del cristal y por no encontrarse evidencia experimental de que estas interacciones existan *in vivo* o *in vitro*.

#### 4.1.2.1. Conservación evolutiva de residuos de región con estructura tridimensional

Analizando los 58 residuos seleccionados del homeodominio para estudiar su conservación evolutiva, se observó que 25 tuvieron conservación alta, 20 conservación media y 13 conservación baja de acuerdo a nuestro criterio basado en las predicciones de SIFT y ConSurf (Tabla 4).

A destacar de lo visto en los residuos clasificados según su conservación:

- De los 16 residuos de la tercera alfa hélice, 13 tuvieron conservación alta.
- Los residuos con conservación alta fuera de la tercera alfa hélice apuntan mayoritariamente hacia el interior de la estructura, mientras que los residuos con conservación baja mayoritariamente apuntan hacia el exterior.

**Tabla 4. Clasificación de residuos del homeodominio por conservación evolutiva.**

| AA  | Conservación evolutiva | AA  | Conservación evolutiva | AA  | Conservación evolutiva |
|-----|------------------------|-----|------------------------|-----|------------------------|
| 137 | Media                  | 157 | Alta                   | 177 | Alta                   |
| 138 | Baja                   | 158 | Media                  | 178 | Media                  |
| 139 | Media                  | 159 | Media                  | 179 | Baja                   |
| 140 | Media                  | 160 | Media                  | 180 | Media                  |
| 141 | Baja                   | 161 | Media                  | 181 | Alta                   |
| 142 | Alta                   | 162 | Alta                   | 182 | Alta                   |
| 143 | Alta                   | 163 | Alta                   | 183 | Alta                   |
| 144 | Media                  | 164 | Media                  | 184 | Alta                   |
| 145 | Alta                   | 165 | Baja                   | 185 | Alta                   |
| 146 | Media                  | 166 | Media                  | 186 | Alta                   |
| 147 | Alta                   | 167 | Media                  | 187 | Alta                   |
| 148 | Media                  | 168 | Alta                   | 188 | Alta                   |
| 149 | Alta                   | 169 | Baja                   | 189 | Media                  |
| 150 | Media                  | 170 | Baja                   | 190 | Alta                   |
| 151 | Baja                   | 171 | Media                  | 191 | Alta                   |
| 152 | Baja                   | 172 | Alta                   | 192 | Alta                   |
| 153 | Alta                   | 173 | Baja                   | 193 | Baja                   |
| 154 | Alta                   | 174 | Baja                   | 194 | Alta                   |
| 155 | Media                  | 175 | Alta                   |     |                        |
| 156 | Media                  | 176 | Baja                   |     |                        |

En celeste, residuos que forman parte de la primera o segunda alfa hélice. En violeta, residuos que forman parte de la tercera alfa hélice. AA: Aminoácido.

#### 4.1.3. Búsqueda de motivos lineales en regiones accesibles

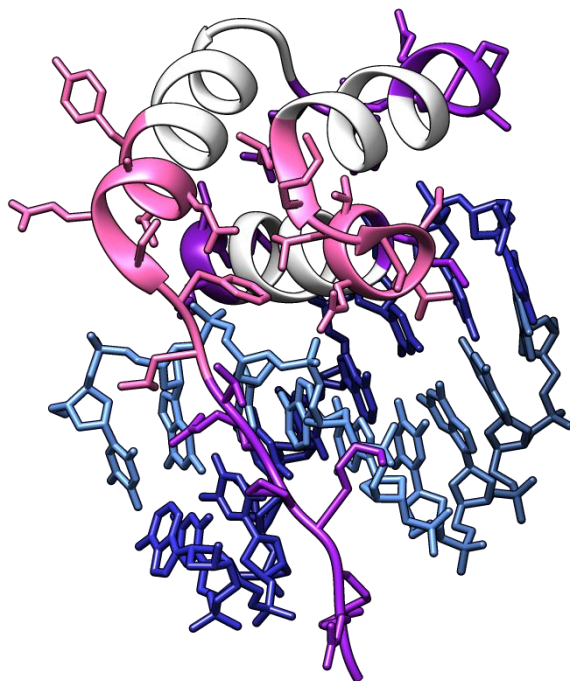
Para el estudio de residuos sin limitarse a la región estructurada de NKX2-5, se hizo la búsqueda de motivos lineales confirmados en la bibliografía y motivos lineales predichos utilizando ELM.

A partir de la búsqueda bibliográfica realizada, se compilaron 4 motivos lineales: 2 se corresponden con señales de localización nuclear ubicadas entre los residuos 136 a 143 y 192 a 198 [179,180], 1 con una señal de SUMOylación, entre los residuos 51 y 54 [181,182], y 1 con un sitio modificado por caseín-kinasa 2 (CK2), entre los residuos

161 y 167 [179]. Además, se compiló 1 sitio con modificación post-traducciona, que se corresponde con un sitio de acetilación en el residuo Lys183 [183,184].

Con ELM se predijeron 97 motivos lineales, incluyendo los 4 motivos lineales confirmados experimentalmente (Anexo II. Tabla S1). No se encontró un motivo lineal correspondiente al sitio de acetilación de Lys183. En el homeodominio se encontraron 27 motivos lineales predichos, de los cuales 3 se corresponden con las 2 señales de localización nuclear y el sitio modificado por CK2 confirmados experimentalmente.

Los motivos lineales del homeodominio se ubicaron en 34 residuos de NKX2-5, todos fuera de hélices alfa y consistiendo principalmente de aminoácidos en *loops* y en contacto con el solvente (Figura 21).

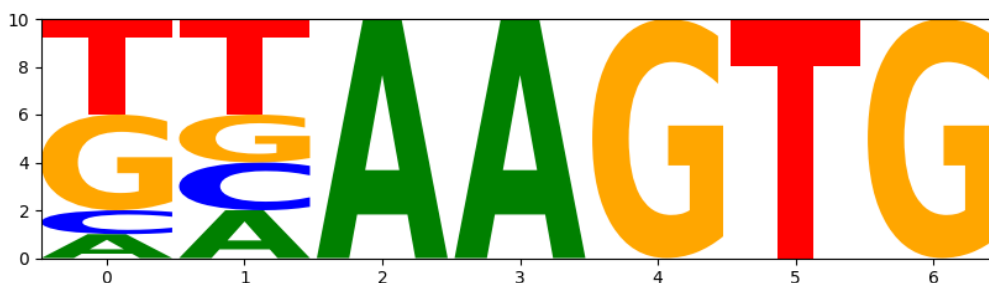


**Figura 21. Esquema del homeodominio de NKX2-5, correspondiente a aminoácidos 137-194 de cristal PDB-ID 3RKQ, en contacto con el ADN coloreado según el resultado de la clasificación de distintos residuos de acuerdo a su ubicación. En azul, el ADN. En violeta, 18 residuos que forman parte de motivos lineales confirmados. En rosa, 16 residuos que forman parte de motivos lineales predichos.**

#### 4.1.4. Potenciales interacciones con secuencias de ADN

Para completar lo respectivo al modelo estructural, se realizó la búsqueda bibliográfica para definir interacciones de NKX2-5 con secuencias específicas de ADN.

A partir de las publicaciones de los cristales para NKX2-5 [146,176-178], la bibliografía recopilada de ensayos de unión al ADN [40,60,185] y de ensayos de transactivación [70,186-205], se compilaron 10 secuencias de unión para NKX2-5: GCAAGTG, GGAAGTG, GAAAGTG, ATAAGTG, GTAAGTG, CTAAGTG, TCAAGTG, TGAAGTG, TAAAGTG y TTAAGTG. El LOGO correspondiente a estas 10 secuencias de unión se presenta en la Figura 22.



**Figura 22. LOGO correspondiente a las 10 secuencias de unión confirmadas experimentalmente para NKX2-5.** La altura de los nucleótidos en cada posición corresponden con la frecuencia con la que se vieron en las 10 secuencias seleccionadas.

## 4.2. Recopilación y análisis de todas las variantes génicas conocidas de *NKX2-5*

Una vez definido el modelo estructural, se recopilaron todas las VGs conocidas de *NKX2-5*, las cuales fueron complementadas con información de funcionalidad y potencial patogenicidad, haciendo uso del modelo estructural en varias secciones.

Se actualizó la base de datos que poseíamos con anterioridad con 877 VGs [68] con el agregado de 512 VGs nuevas, de las cuales 97 provinieron de publicaciones, dando un total de **1389 VGs**. Estas 1389 variantes fueron depositadas como 1547 registros (las variantes que se encontraron en múltiples individuos se cargaron una vez

por cada uno que la tuviera) a la base de datos correspondiente a *NKX2-5* en LOVD (<https://www.lovd.nl/NKX2-5/>). El enlace para la tabla completa con las VGs compiladas se encuentra disponible en la sección Anexo I.1.

#### 4.2.1. Distribución de variantes totales y patogénicas a lo largo de *NKX2-5*

Del total de 1389 variantes, 629 se encontraron en la región traducida (TR), y 760 en la región no traducida (NTR, incluye intrones y regiones UTR 3' y 5') (Tabla 5).

Las variantes totales se encontraron distribuidas a lo largo de todo el gen, aunque se encontró casi el doble de variantes por nucleótido en la región traducida (64 VGs cada 100 nucleótidos en la región traducida, 34 VGs cada 100 nucleótidos en la región no traducida). Cabe notar que se observó una acumulación de variantes en una región pequeña del intrón (chr5:173.233.506-173.233.550, RefSeq: NC\_000005.10), la cual consiste de numerosas repeticiones de adenina y timina y coincide con una agrupación de variantes con frecuencias poblacionales mayores a 1/10.000.

Del total de variantes recopiladas, 146 fueron clasificadas como patogénicas por alguna publicación y/o base de datos. El enlace para la tabla conteniendo únicamente estas VGs patogénicas se encuentra disponible en la sección Anexo I.2.

La Figura 23 muestra un esquema de la ubicación de las variantes a lo largo del gen así como aquellas variantes que se corresponden con la secuencia proteica.

Del total de las 629 variantes ubicadas en la región traducida, 391 (62.2%) fueron de cambio de sentido (*missense*), de las cuales 103 fueron clasificadas como patogénicas. También encontramos 170 (27%) variantes sinónimas, ninguna de las cuales se asoció con patogénesis. También encontramos 20 (3,2%) variantes que introducen un codón stop (*stop loss*), 33 (5,2%) variantes de inserciones con cambio de sentido (*frameshift*) y 1 (0,2%) variante de pérdida de codón stop. De estas 54, 41 fueron clasificadas como patogénicas. Por último, encontramos un total de 14 (2,2%) variantes de inserciones sin cambio de sentido (*in-frame*): 7 deleciones, 6 duplicaciones y 1 inserción-delección. Ninguna de las variantes de inserción/delección sin cambio de marco de lectura fue clasificada como patogénica (Tabla 5).

De las 760 variantes ubicadas en la región no traducida, 2 fueron clasificadas como patogénicas (Tabla 5): c.(334+1G>T) y c.(335-1G>T). Ambas se ubican en sitios de *splicing* de *NKX2-5* (Figura 23).

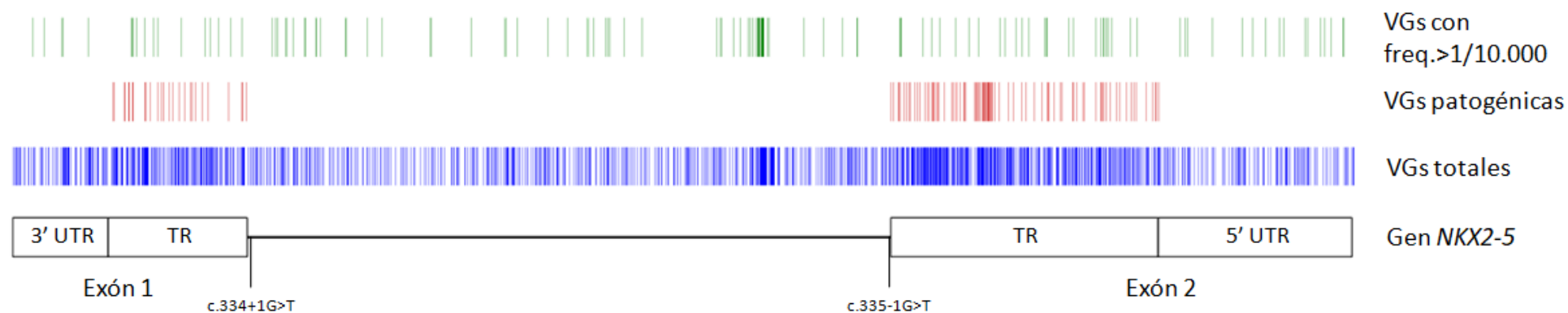
Centrándonos en las variantes clasificadas como patogénicas por publicaciones y/o bases de datos, del total de 146, 144 se ubican en la región traducida, y las otras 2 afectan sitios de *splicing*. Su distribución se observa en rojo en la Figura 23, y su distribución por tipo de variante se muestra en la Tabla 5, en números entre paréntesis.

**Tabla 5. Tipos de variantes génicas en las diferentes regiones conservadas del gen *NKX2-5*.**

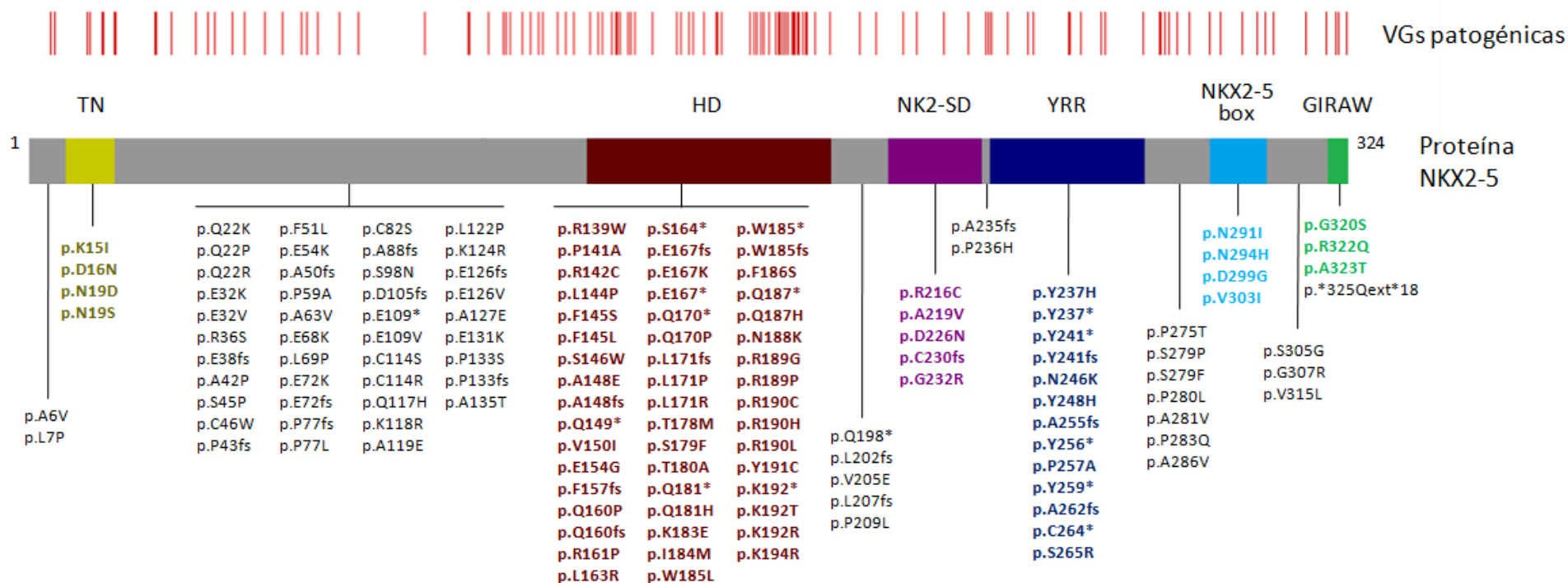
|                     |               | Cambio de sentido | Sinónimas | Sin sentido | Indel sin cambio de marco de lectura | Indel con cambio de marco de lectura | Codón stop | Total             |
|---------------------|---------------|-------------------|-----------|-------------|--------------------------------------|--------------------------------------|------------|-------------------|
| Región traducida    | Tinman        | 10 (4)            | 6 (0)     | 0 (0)       | 0 (0)                                | 1 (0)                                | 0 (0)      | 17 (4)            |
|                     | Hélice 1      | 18 (4)            | 11 (0)    | 1 (1)       | 0 (0)                                | 2 (2)                                | 0 (0)      | 32 (7)            |
|                     | Hélice 2      | 13 (4)            | 6 (0)     | 2 (2)       | 1 (0)                                | 2 (2)                                | 0 (0)      | 24 (8)            |
|                     | Hélice 3      | 28 (18)           | 11 (0)    | 4 (4)       | 0 (0)                                | 1 (1)                                | 0 (0)      | 44 (23)           |
|                     | Entre hélices | 26 (10)           | 13 (0)    | 2 (1)       | 0 (0)                                | 1 (1)                                | 0 (0)      | 42 (12)           |
|                     | Total         | 85 (36)           | 41 (0)    | 9 (8)       | 1 (0)                                | 6 (6)                                | 0 (0)      | 142 (50)          |
|                     | NK2-SD        | 27 (4)            | 9 (0)     | 1 (0)       | 4 (0)                                | 2 (1)                                | 0 (0)      | 43 (5)            |
|                     | YRR           | 44 (5)            | 10 (0)    | 7 (7)       | 0 (0)                                | 5 (4)                                | 0 (0)      | 66 (16)           |
|                     | NKX2-5 box    | 20 (4)            | 9 (0)     | 0 (0)       | 1 (0)                                | 1 (0)                                | 0 (0)      | 31 (4)            |
|                     | GIRAW         | 6 (3)             | 2 (0)     | 0 (0)       | 0 (0)                                | 0 (0)                                | 0 (0)      | 8 (3)             |
| Región no traducida | Otros         | 199 (47)          | 93 (0)    | 3 (2)       | 8 (0)                                | 18 (12)                              | 1 (1)      | 322 (62)          |
|                     | Total         | 391 (103)         | 170 (0)   | 20 (17)     | 14 (0)                               | 33 (23)                              | 1 (1)      | 629 (144)         |
|                     |               |                   |           |             |                                      |                                      |            | 760 (2)           |
|                     |               |                   |           |             |                                      |                                      |            | Total: 1381 (144) |

Entre paréntesis, variantes clasificadas como patogénicas. HD: Homeodominio; NK2-SD: Dominio NK2 específico; YRR: Dominio rico en tirosinas. GIRAW: Dominio compuesto por aminoácidos Gly-Ile-Arg-Ala-Trp.

A



B



**Figura 23. Distribución de variantes compiladas sobre el gen *NKX2-5* y la proteína que codifica.** A. Ubicación de las variantes en la secuencia del gen *NKX2-5*. Las cajas blancas representan los dos exones del gen, mientras que la línea que los une representa el intrón. El análisis fue limitado al transcripto canónico de *NKX2-5* (chr5:173.232.109-173.235.311; RefSeq: NC\_000005.10). Cada línea sobre el gen *NKX2-5* representa la posición de cada variante. En azul, la posición inicial de las 1389 variantes compiladas. En rojo, variantes clasificadas como patogénicas. En verde, variantes con una frecuencia mayor a 1/10.000 en bases de datos poblacionales. Las dos variantes patogénicas que afectan sitios de *splicing* son detalladas explícitamente bajo el esquema del gen (RefSeq: NM\_004378.1). B. Posiciones de las variantes compiladas clasificadas como patogénicas, sobre una representación de la proteína codificada por *NKX2-5*, resaltando dominios conservados conocidos. Por encima de la proteína, cada línea roja representa la posición de una variante clasificada como patogénica. Por debajo de la proteína, se enumeran las variantes por región con sus descriptores HGVS (RefSeq: NP\_004378.1). No se esquematizan variantes distintas que causan el mismo cambio aminoacídico. VG, variante génica. HD, homeodominio. NK2-SD, dominio NK2-específico. TN, dominio Tinman. TR, región traducida. UTR, región no traducida. YRR, región rica en tirosinas.

#### 4.2.1.1. Distribución de variantes que afectan motivos lineales

Se registraron aquellas VGs que se ubicaran sobre alguno de los 97 motivos lineales o el sitio de acetilación encontrados en la sección 4.1.3. Para esto, sólo se tuvieron en cuenta aquellas VGs que sean de cambio de sentido o inserciones/deleciones sin cambio del marco de lectura. Estas fueron un total de 405 (391 VGs de cambio de sentido y 14 inserciones/deleciones), de las cuales 276 (265 missense y 11 inserciones/deleciones) ocurren sobre motivos lineales predichos o confirmados experimentalmente.

En el caso de las variantes patogénicas, se hallaron 69 VGs: 57 sobre motivos lineales predichos y 12 sobre motivos lineales confirmados experimentalmente.

En la Tabla 6 se presentan los 5 motivos lineales confirmados experimentalmente con las variantes patogénicas que ocurren en sus residuos. De estas 12 variantes patogénicas, 2 se encuentran en el motivo de SUMOylación (aminoácidos 51-54), 3 en la señal de localización nuclear N-terminal (aminoácidos 136-143), 3 en el sitio reconocido por caseína-kinasa 2 (aminoácidos 161-167), 1 en una

lisina acetilada (aminoácido 183) y 3 en la señal de localización nuclear C-terminal (aminoácidos 192-198).

**Tabla 6. Variantes patogénicas en motivos lineales confirmados.**

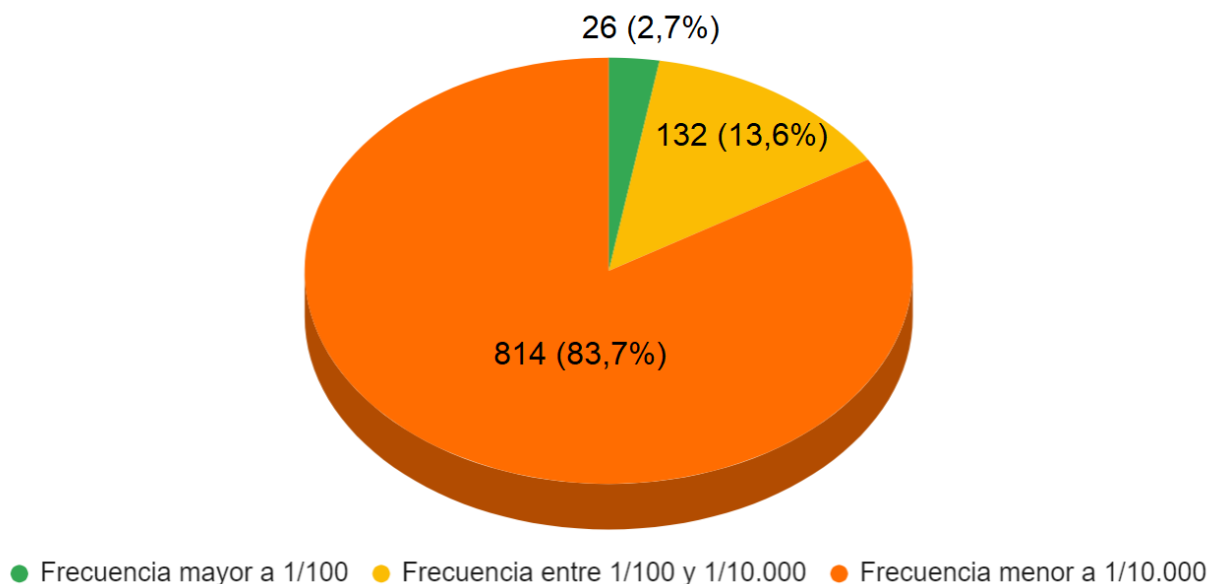
| Descripción                                    | Residuos | Número de variantes patogénicas | Variantes patogénicas en la región | Fuentes y referencias                    |
|------------------------------------------------|----------|---------------------------------|------------------------------------|------------------------------------------|
| <b>Motivo de SUMOylación</b>                   | 51-54    | 2                               | p.(F51L), p.(E54K)                 | ELM; Wang y col., 2008; Kim y col., 2011 |
| <b>Sitio de acetilación</b>                    | 183      | 1                               | p.(K183E)                          | Li y col., 2007; Tang y col., 2016       |
| <b>Motivo de fosforilación Caseín-kinasa 2</b> | 161-167  | 3                               | p.R161P, p.(L163R), p.(E167K)      | ELM; Kasahara e Izumo, 1999              |
| <b>Señal de localización nuclear</b>           | 136-143  | 3                               | p.R139W, p.(P141A), p.R142C        | ELM; Kasahara e Izumo, 1999              |
| <b>Señal de localización nuclear</b>           | 192-198  | 3                               | p.(K192T), p.(K192R), p.(K194R)    | ELM; Ouyang y col., 2016                 |

Para cada motivo lineal, se registran sus posiciones en la proteína NKX2-5, las variantes patogénicas que lo afectan y la bibliografía que respalda el motivo lineal y las variantes patogénicas. ELM: Predicho por Eukaryotic Linear Motif (<http://elm.eu.org/>) [154].

#### 4.2.1.2. Distribución de variantes con información de frecuencia poblacional

En cuanto a la frecuencia poblacional de las variantes, se recopiló información para 972 variantes de las 3 bases de datos estudiadas. Esta información se presentó como frecuencia cada 1.000 individuos.

Se observó que 26 de las variantes poseen una frecuencia mayor a 1/100, que 132 poseen una frecuencia entre 1/100 y 1/10.000 y que 814 poseen una frecuencia menor a 1/10.000 (Figura 24). La distribución de VGs con frecuencia poblacional mayor a 1/10.000 se puede ver en la Figura 23A, en líneas en verde.



**Figura 24. Proporciones de variantes según su frecuencia en bases de datos poblacionales.**

En cuanto a las variantes patogénicas, se observó que 32 de las 146 variantes patogénicas poseen datos de frecuencias en las bases de datos poblacionales. En particular, 7 variantes tienen una frecuencia entre 1/100 y 1/10.000, y 1 variante tiene una frecuencia de 1,07/100. Esta información se puede ver específicamente en la tabla de VGs patogénicas en la sección Anexo I.2.

#### 4.2.1.3. Distribución de variantes patogénicas en NKX2-5

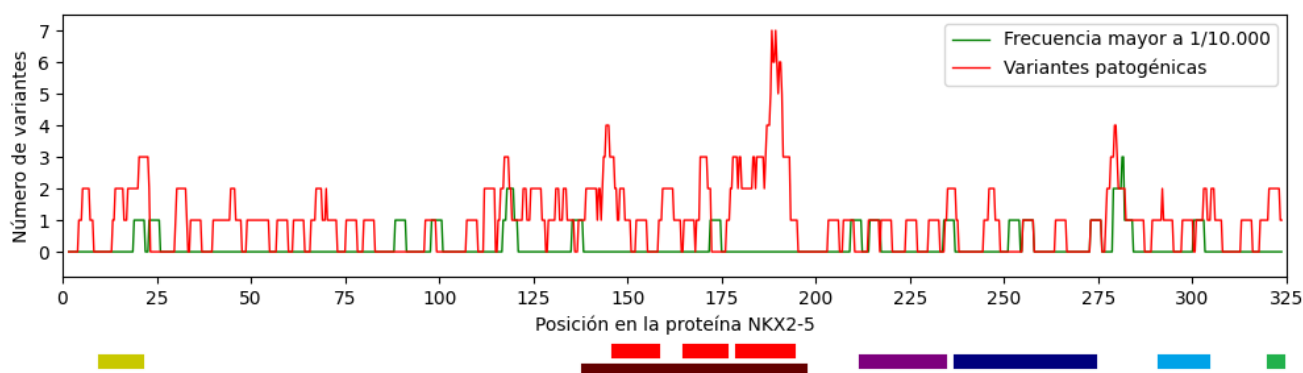
Para el estudio de la distribución de variantes patogénicas se analizaron las 103 variantes patogénicas de cambio de sentido (Tabla 5), ya que son las únicas que afectan la proteína localmente, cambiando aminoácidos puntuales. Por otro lado, las 17 variantes patogénicas que agregan un codón *stop*, 23 que producen una inserción con cambio de sentido y 2 que afectan sitios de *splicing* afectan a la totalidad de la proteína. La variante patogénica de pérdida de codón *stop* no se incluyó en este grupo por agregar varios aminoácidos en vez de cambiar uno solo.

Las 103 variantes patogénicas de cambio de sentido se distribuyeron en las distintas regiones de NKX2-5 de la siguiente manera (Figura 23B): 4 en el dominio Tinman, 36 en el Homeodominio, 4 en el dominio NK2 específico, 4 en la región rica en tirosinas, 4 en el dominio NKX2-5 box, 3 en el dominio GIRAW y 48 en el resto de la

proteína. De las 36 variantes en el Homeodominio, 19 afectaron residuos altamente conservados, 14 residuos medianamente conservados y 3 residuos levemente conservados.

Para visualizar de forma más intuitiva la concentración de variantes a lo largo de la proteína, se utilizó el programa desarrollado en la sección 3.2.6. para graficar la cantidad de variantes en una ventana de 3 aminoácidos (9 nucleótidos) alrededor de cada posición del gen (Figura 25).

Como se observa, una gran cantidad de variantes patogénicas se ubican en la tercera hélice alfa del homeodominio. Además, esta región no posee ninguna variante con frecuencia poblacional mayor a 1/10.000, lo que sugiere que es poco tolerante a variantes de secuencia.



**Figura 25. Distribución de variantes génicas en la proteína NKX2-5.** Se observa el número de variantes de cambio de sentido a lo largo de la proteína NKX2-5 (residuos 1-324), contadas en ventanas de 9 residuos. En verde, variantes de bases de datos poblacionales con una frecuencia de más de 1/10.000. En rojo, variantes clasificadas como patogénicas. Las líneas gruesas debajo del gráfico denotan distintas regiones y dominios de la proteína: En amarillo, el dominio Tinman. En bordó, el homeodominio. En rojo, las tres hélices alfa del homeodominio. En violeta, el dominio NK2-específico. En azul, la región rica en tirosinas. En celeste, el motivo Nkx2-5 box. En verde, el motivo GIRAW.

#### 4.2.2. Evaluación de patogenicidad de variantes génicas

Con el objetivo de tener una definición estandarizada de patogenicidad que pueda compararse con otras fuentes, se clasificaron 165 variantes de acuerdo a los criterios del ACMG. El enlace para la tabla con la clasificación de variantes según los criterios de ACMG se encuentra disponible en la sección Anexo I.3. Su clasificación

según este criterio se observa en la columna “ACMG”, y su clasificación en publicaciones y/o bases de datos se observa en la columna “Clasificación”.

De las 165 variantes clasificadas:

- 144 corresponden a variantes clasificadas como patogénicas en publicaciones o bases de datos, de las cuales 71 fueron clasificadas según ACMG como Patogénicas (“*Pathogenic*”), 35 como Probablemente patogénicas (“*Likely pathogenic*”), 35 como Variantes de significado incierto (“*Variant of uncertain significance*”) y 3 como Probablemente benignas (“*Likely benign*”).
- 13 corresponden a variantes clasificadas como benignas en publicaciones o bases de datos, de las cuales 2 fueron clasificadas según ACMG como Variantes de significado incierto (“*Variant of uncertain significance*”), 6 como Probablemente benignas (“*Likely benign*”) y 5 como Benignas (“*Benign*”).
- 6 corresponden a variantes con evidencia conflictiva en publicaciones o bases de datos, de las cuales 4 fueron clasificadas según ACMG como Variantes de significado incierto (“*Variant of uncertain significance*”), 1 como Probablemente patogénica (“*Likely pathogenic*”) y 1 como Probablemente benigna (“*Likely benign*”).
- 2 corresponden a variantes encontradas en pacientes del CNGM, de las cuales ambas fueron clasificadas como Variantes de significado incierto (“*Variant of uncertain significance*”) según ACMG

Cabe destacar que todas las variantes clasificadas como Patogénicas o como Benignas por el criterio de ACMG fueron consistentes con la clasificación encontrada en publicaciones y bases de datos.

#### 4.2.3. Distribución de subtipos de CC

Para el estudio de la distribución de fenotipos, se analizaron variantes clasificadas como patogénicas por publicaciones y/o bases de datos (Tabla 5, números entre paréntesis), y se hizo foco en aquellas con subtipos de cardiopatías asociados. No se analizaron variantes con evidencia conflictiva por tener una posibilidad mayor de que los fenotipos no estén asociados a la variante.

Por otro lado, se analizaron por separado aquellas variantes que involucran un aminoácido y aquellas que conllevan grandes cambios sobre la secuencia de la proteína. Mientras que las variantes que afectan puntualmente un aminoácido pueden estudiarse en relación a su posición específica en la proteína, las que conllevan grandes cambios en la proteína probablemente reflejan un efecto global en la función de NKX2-5.

De las 146 variantes patogénicas (Anexo I.2.), 131 tuvieron subtipos específicos de cardiopatías asociados, 5 estuvieron asociadas a cardiopatías congénitas en general, 5 estuvieron asociadas únicamente a fenotipos no cardíacos y 5 no tuvieron un fenotipo específico asociado. De las 131 variantes en las que se concentró este análisis, 34 predicen la introducción de un codón *stop* prematuro, 1 elimina el codón *stop* agregando 18 aminoácidos en el extremo C-terminal de la proteína y 96 cambian un aminoácido puntual. La distribución de subtipos de cardiopatías para estas variantes se visualiza en la Tabla 7.

**Tabla 7. Distribución de subtipos de cardiopatías acorde al tipo de variante patogénica.**

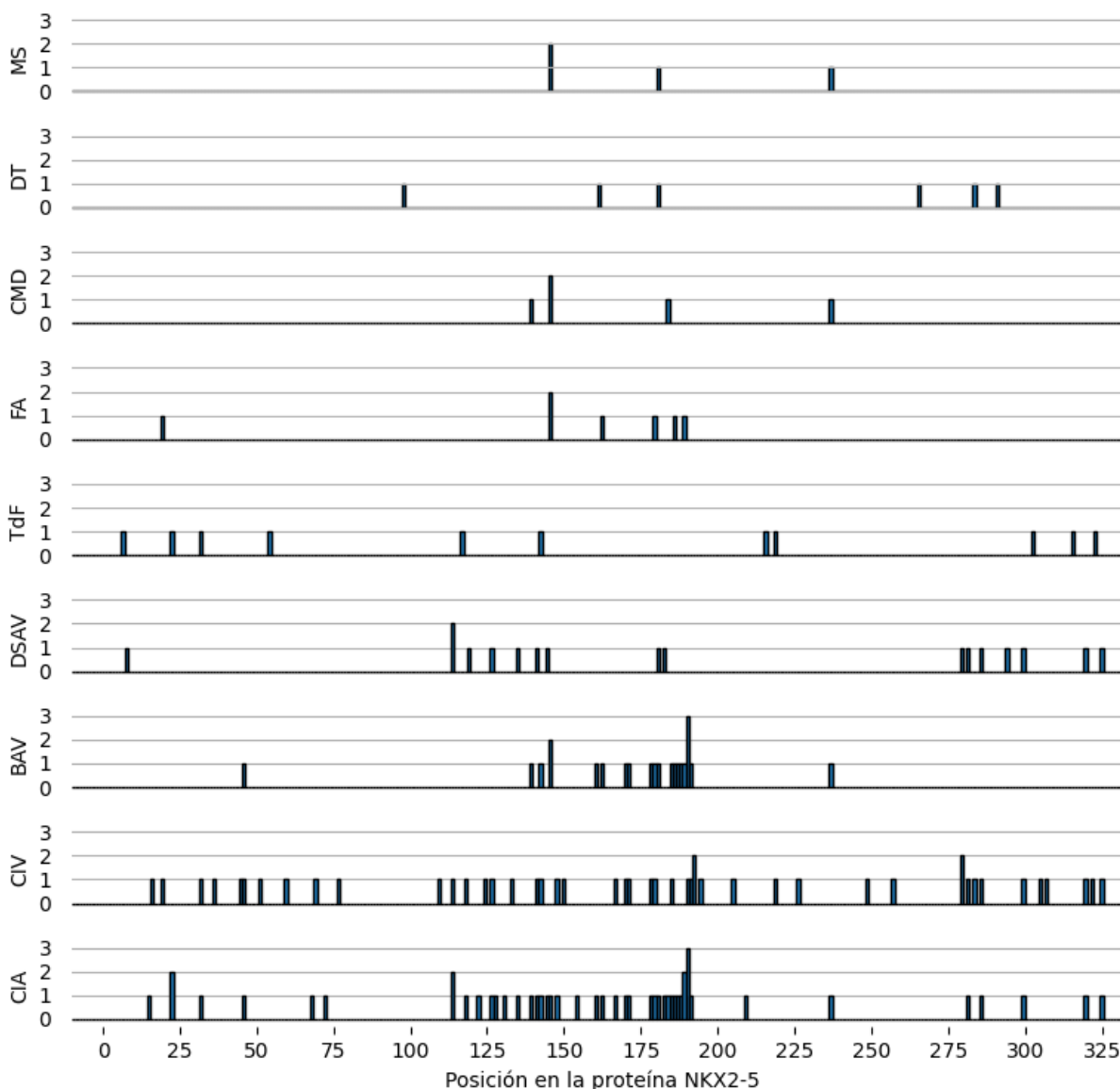
| Variantes con grandes cambios en la proteína (n=35) |            | Variantes de efecto en un aminoácido (n=96) |            |
|-----------------------------------------------------|------------|---------------------------------------------|------------|
| Comunicación interauricular                         | 31 (88,6%) | Comunicación interauricular                 | 48 (50,0%) |
| Bloqueo atrioventricular                            | 26 (74,3%) | Comunicación interventricular               | 46 (47,9%) |
| Comunicación interventricular                       | 9 (25,7%)  | Bloqueo atrioventricular                    | 22 (22,9%) |
| Muerte súbita                                       | 4 (11,4%)  | Defecto en septo atrioventricular           | 16 (16,7%) |
| Cardiomiopatía dilatada                             | 3 (8,6%)   | Tetralogía de Fallot                        | 11 (11,5%) |
| Defecto en septo atrioventricular                   | 2 (5,7%)   | Fibrilación atrial                          | 7 (7,3%)   |
| Fibrilación atrial                                  | 2 (5,7%)   | Disgénesis de tiroides                      | 6 (6,2%)   |
| Heterotaxia                                         | 2 (5,7%)   | Cardiomiopatía dilatada                     | 5 (5,2%)   |

Los 3 subtipos más comunes de cardiopatías observadas están coloreados para una mejor visualización.

Como se observa en la Tabla 7, las variantes que tendrían como efecto la introducción de un codón *stop* prematuro o la eliminación del codón *stop* estarían principalmente asociadas a comunicación interauricular (CIA) y bloqueo atrioventricular (bloqueo AV), los cuales se observan asociados a la mayoría de las variantes con grandes cambios en la proteína (88,6% y 74,3%, respectivamente). El tercer subtipo más común en asociación a variantes que introducen grandes cambios en NKX2-5 es la comunicación interventricular (CIV), viéndose asociada a una proporción marcadamente menor (25,7%) de las variantes encontradas.

En las variantes puntuales, por su parte, se observan los mismos 3 subtipos más comunes, aunque en órdenes diferentes. Los dos subtipos más comunes son CIA y CIV, los cuales se observan asociados a aproximadamente la mitad de las variantes de efecto en un aminoácido (50% y 47,9%, respectivamente). Por su parte, el bloqueo AV se encuentra en un tercer lugar con una proporción marcadamente menor (22,9%) a la vista en variantes que introducen grandes cambios.

La distribución de las variantes de efecto en un aminoácido asociadas a distintas patologías se puede visualizar en la Figura 26.



**Figura 26. Distribución de variantes patogénicas de cambio de sentido en la proteína NKX2-5, separadas por patologías específicas a las que se encuentran asociadas.** Cada barra representa una posición en la proteína NKX2-5, y su altura, la cantidad de variantes patogénicas asociadas a una patología ubicadas en esa posición. CIA: Comunicación interauricular. CIV: Comunicación interventricular. BAV: Bloqueo atrioventricular. DSAV: Defecto en septo atrioventricular. TdF: Tetralogía de Fallot. FA: Fibrilación atrial. CMD: Cardiomiopatía dilatada. DT: Disgénesis de tiroides. MS: Muerte súbita.

Como se puede observar en la Figura 26, tanto para CIA como para CIV, las variantes asociadas están distribuidas a lo largo de la proteína, sin una agrupación específica. Sin embargo para el bloqueo AV, 20 de las 22 variantes asociadas se encuentran en el homeodominio y 11 de ellas se encuentran específicamente en la tercera hélice alfa, lo cual sugiere su posible relación con una disfunción de la

interacción con el ADN. Por otro lado, para Tetralogía de Fallot (TdF), se observó que sólo 1 de las 11 variantes puntuales asociadas se encuentran en el homeodominio, lo cual sugiere que para este subtipo de CC el efecto deletéreo no se relacionaría con la unión al ADN. De destacar, es que sólo 1 de las 35 variantes que generan grandes cambios en la secuencia de la proteína estaría asociada a TdF.

Aunque las patologías no cardíacas no fueron foco de estudio, se encontraron en 9 de las 146 variantes patogénicas y se remarcán dos observaciones:

- Los fenotipos tiroideos fueron los más comunes entre los fenotipos no cardíacos, con 6 variantes asociadas a distintas disgénesis de la tiroides. Todas las variantes patogénicas asociadas a disgénesis de tiroides fueron puntuales (de cambio de sentido), aunque no se las vió concentradas en una región en particular.
- Se encontraron 2 variantes asociadas a heterotaxia, ambas de cambio de marco de lectura antes del homeodominio, por lo que se puede inferir que provocarían la pérdida completa de la funcionalidad de NKX2-5.

### **4.3. Análisis de todas las variantes de cambio de sentido en el homeodominio de NKX2-5**

Con el objetivo de hacer un análisis específico y en profundidad de variantes en regiones de NKX2-5 con estructura tridimensional, se seleccionó el homeodominio de NKX2-5 para realizar un análisis de los posibles efectos funcionales de cualquier variante que pudiera originar un cambio aminoacídico en esta región.

#### **4.3.1. Generación de todas las VGs de cambio de sentido posibles en la región del homeodominio**

Utilizando el programa desarrollado, se generaron 337 variantes de cambio de sentido en la región correspondiente al homeodominio de NKX2-5 (residuos 137-194). Entre ellas, 250 están compiladas en la lista de variantes de la sección 4.2. Esto deja

un total de 87 variantes aún no reportadas en la lista de 337 variantes de cambio de sentido posibles en el homeodominio de NKX2-5.

#### 4.3.2. Estudios *in silico* para cada VG generada

Estas 337 variantes fueron analizadas para predecir su potencial efecto en la interacción con el ADN, en SLiMs, en la estabilidad de la proteína (predicha por  $\Delta\Delta G$  o por conservación evolutiva) y para registrar cambios que introduzcan cisteínas. Los resultados se presentan resumidos en la Tabla 8 y completos en un enlace externo de la sección Anexo I.4.

**Tabla 8. Conteo de los resultados significativos de análisis *in silico*.**

|                                            |     |
|--------------------------------------------|-----|
| Efecto en la interacción con el ADN        | 32  |
| Efecto en la interacción con otra proteína | 13  |
| Motivo lineal confirmado                   | 52  |
| Motivo lineal predicho                     | 58  |
| Variante que introduce Cisteína            | 13  |
| Efecto en estabilidad según FoldX          | 122 |
| Probablemente dañinas según PolyPhen2      | 248 |

Específicamente, se observó que:

- De las 32 variantes que afectan la unión al ADN, 26 se deben a un cambio de energía de interacción en un residuo en contacto con el ADN, 5 eliminarían un contacto con el ADN y 1 se encuentra en un residuo que no está en contacto con el ADN pero la variante introduce un aminoácido en contacto con una base nitrogenada.

- De las 13 variantes que introducen cisteínas, 1 se encuentra en un residuo inaccesible al agua, pero introduce un cambio significativo en la energía de desplegado. Las otras 12 ocurren en residuos accesibles al agua.
- De las 13 variantes que afectan la interacción con otras proteínas, 11 afectan la interacción con TBX5 y 2, la interacción con MEF2B.
- Entre las variantes que afectan motivos lineales, 52 afectan motivos lineales confirmados experimentalmente y 58 afectan motivos lineales predichos.
- Entre las variantes con efectos predichos, 122 afectarían a la estabilidad de la proteína según lo calculado con FoldX y 248 fueron clasificadas como probablemente dañinas por PolyPhen2.

#### 4.3.3. Clasificación de potenciales efectos de VGs de cambio de sentido en la funcionalidad de NKX2-5

Con los resultados obtenidos de los estudios *in silico*, se procedió a agruparlos y asignar diferentes mecanismos potenciales de patogenicidad. A saber, se agruparon los resultados de efecto en la interacción con otras proteínas y el efecto en motivos lineales por consistir de interacciones con otras macromoléculas. De esta agrupación se diferenció la interacción con el ADN por las particularidades del reconocimiento de una secuencia de ADN y la importancia de la interacción de un factor de transcripción con el ADN.

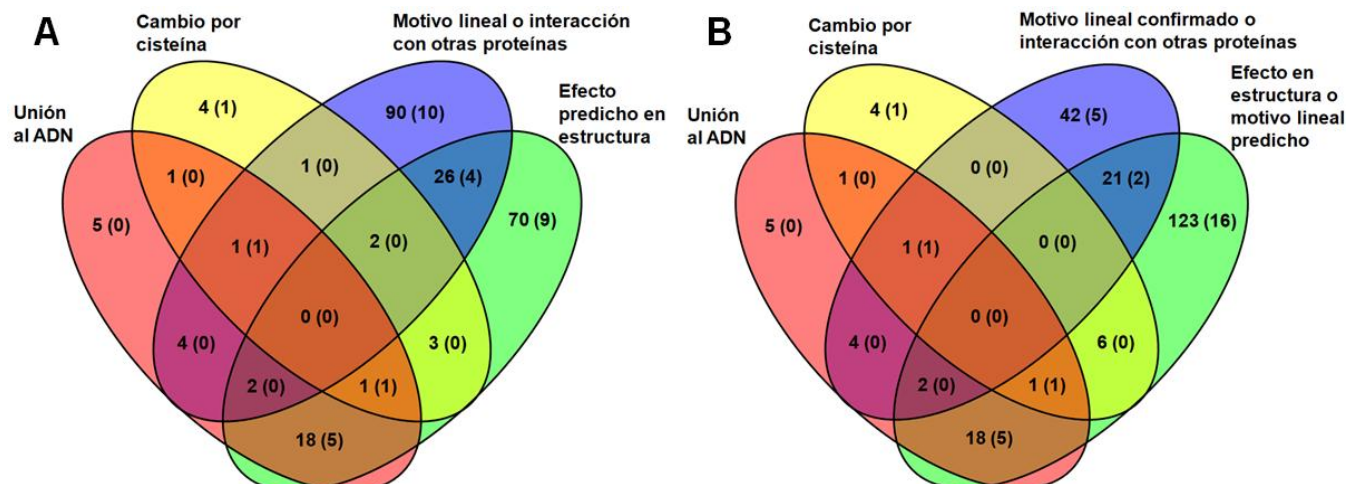
Esto agrupó los resultados en 4 mecanismos potenciales de patogenicidad: desestabilización de la estructura terciaria, efecto en la interacción con el ADN, efecto en la interacción con otras macromoléculas (incluyendo el efecto en interacción con proteínas y el efecto en un motivos lineales) e introducción de una cisteína.

En la Figura 27A se observa la distribución de los 4 efectos de patogénesis como un diagrama de Venn. Como se observa en la figura, 59 de las variantes afectarían la función de la proteína por más de un mecanismo posible. Por esta razón, se implementó una jerarquía de potenciales efectos en la función de NKX2-5, de acuerdo a lo establecido en la sección 3.3.3. De esta forma, si una variante tiene

asociados dos o más posibles efectos, sólo se considera el efecto más alto en la jerarquía, y se simplifica el análisis a realizar.

El orden de la jerarquía fue: en primer lugar, variantes que afectan la interacción con el ADN, en segundo lugar, variantes que introducen cisteínas, en tercer lugar, variantes que afectan motivos lineales confirmados o interacciones con proteínas y, por último, variantes con efecto predicho en estabilidad o en motivos lineales no confirmados.

Antes de aplicar la jerarquización, fue necesario separar las variantes con efecto en motivos lineales predichos por ELM sin confirmación experimental de las variantes con efecto en motivos lineales confirmados y las variantes con efecto en la interacción con otras proteínas. Las variantes con efecto en motivos lineales predichos fueron agrupadas con las variantes con predicción de cambio de estabilidad de desplegado. Ambos tipos de efectos serían potencialmente leves y con mayor posibilidad de representar falsos positivos y por ende estarían en el último grupo de la jerarquía. Los cambios de valores se pueden ver en la Figura 27B.



**Figura 27. Diagramas de Venn representando la jerarquización de variantes génicas de acuerdo a los distintos criterios seleccionados.** Los números representan la cantidad de variantes correspondientes a cada clasificación. Los números entre paréntesis corresponden a números de variantes patogénicas de la sección 4.2.

**A:** Clasificación por los distintos criterios de patogenicidad definidos.

**B:** Similar al de A, con reacomodación de los grupos "Motivo lineal" y "Efecto predicho en estructura". Se reacomodaron motivos lineales predichos y efectos predichos en estructura en un grupo y motivos lineales confirmados en otro.

Con este análisis se observó que de las 337 GVs generadas, 32 afectarían la unión al ADN (9,5%), 10 se corresponden con cambios por cisteína (3,0%), 63 afectan la interacción con otras proteínas o a un motivo lineal confirmado experimentalmente (18,7%), en 123 se predice que afectarían la estabilidad de la proteína o a un motivo lineal no confirmado (36,5%) y para 109 no se obtuvo un efecto en la función asociado (32,3%) (Tabla 9). La clasificación completa con criterio jerarquizado se encuentra disponible en un enlace externo de la sección Anexo I.4.

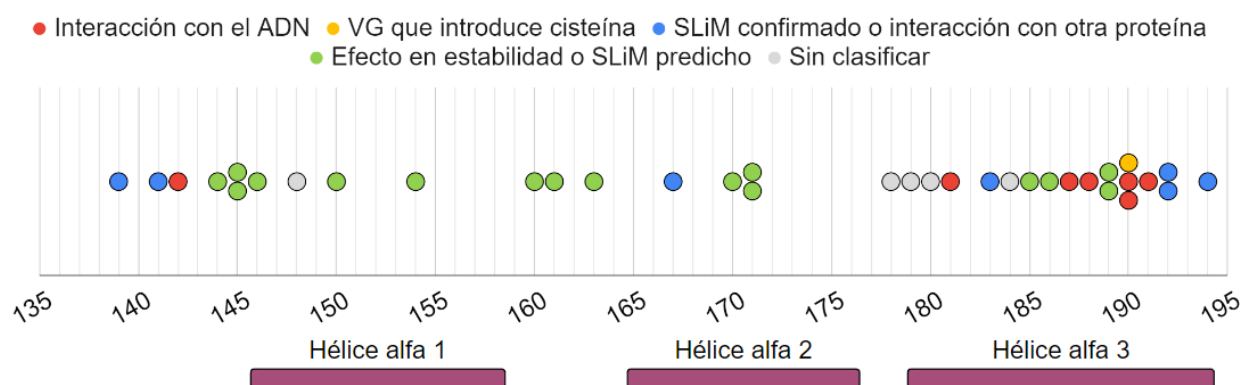
**Tabla 9. Conteo de los resultados de análisis *in silico* tras aplicar la jerarquización.**

|                                                                         | Todas las VGs | VGs patogénicas |
|-------------------------------------------------------------------------|---------------|-----------------|
| Efecto en interacción con el ADN                                        | 32 (9,5%)     | 7 (19,4%)       |
| Variante que introduce Cisteína                                         | 10 (3,0%)     | 1 (2,8%)        |
| Efecto en la interacción con otras proteínas o motivo lineal confirmado | 63 (18,7%)    | 7 (19,4%)       |
| Efecto en estabilidad o motivo lineal predicho                          | 123 (36,5%)   | 16 (44,5%)      |
| Sin mecanismo de patogenicidad asociado                                 | 109 (32,3%)   | 5 (13,9%)       |

VGs: Variantes génicas.

Para complementar lo visto en la distribución de variantes patogénicas de la sección 4.2., se hizo un conteo específico para la tercera hélice alfa del homeodominio (residuos 179 a 194), donde se generaron 96 variantes posibles, de las cuales 68 (70,8%) tuvieron un efecto asociado en la función y 28 (29,2%) no lo tuvieron. Cabe resaltar que esta distribución es notablemente distinta entre los primeros residuos de la hélice (residuos 179 a 184) y los últimos (185 a 194). En los residuos 179 a 184, de las 37 VGs posibles que se generaron, 16 (43,2%) tuvieron asociados mecanismos de efecto en la función, mientras que de las 59 VGs posibles que se generaron en los residuos 185 a 195, 52 (88,1%) tuvieron asociados mecanismos de efecto en la función.

Por último, se estudiaron las 36 variantes clasificadas como patogénicas en la lista de la sección 4.2. (Anexo II. Tabla S2). Se observó que 7 variantes afectarían la unión al ADN (19,4%), 1 se corresponde con un cambio por cisteína (2,8%), 7 afectan a un motivo lineal confirmado experimentalmente (19,4%), 16 tienen estudios que predicen un efecto sobre la estabilidad o un motivo lineal predicho (44,5%) y 5 no tuvieron un efecto funcional asociado (13,9%) (Tabla 9 y Figura 28). Ninguna de las variantes patogénicas tuvo un efecto predicho en la interacción con otras proteínas.



**Figura 28. Resultados del potencial efecto funcional de variantes génicas clasificadas como patogénicas de acuerdo a su posición en el homeodominio.** Con los círculos de colores se representa la clasificación según nuestra jerarquía de cada variante patogénica y su posición en la secuencia de NKX2-5 en el eje X. Abajo, en cuadrados violetas, se esquematiza la posición de las 3 hélices alfa del homeodominio.

Como se esquematiza en la Figura 28, las variantes patogénicas y sus potenciales mecanismos de patogenicidad se corresponden con regiones relevantes del homeodominio. Las variantes antes de la primera hélice alfa se encuentran sobre una señal de localización nuclear y posiblemente entren en contacto con el ADN. Las variantes entre la primera y segunda alfa hélice consisten principalmente de variantes patogénicas que desestabilizan la estructura proteica o afectan al motivo lineal entre ellas. En la tercera alfa hélice se encuentran, por un lado, variantes que afectan la interacción con el ADN, pero también 4 de las 5 variantes patogénicas sin una función asociada, las cuales se agrupan en un rango de 7 residuos (entre la posición 178 y 184).

Tal como se describe en la sección 4.2., la tercera hélice alfa del homeodominio tiene una gran cantidad de variantes patogénicas, lo cual coincide con las variantes que

afectan la unión con el ADN así como con aquellas sin un mecanismo de patogénesis asociado.

#### **4.4. Recopilación y estudio de todos los sitios de unión de NKX2-5 en el genoma humano y de ratón**

Esta parte del trabajo complementa los estudios de variantes en *NKX2-5* recopilando información de sitios de unión de *NKX2-5* en el genoma humano y de ratón y de genes que podría estar transactivando *NKX2-5* al unirse en estos sitios.

Para tal fin se buscaron en la bibliografía trabajos reportando experimentos de ChIP-seq de tejidos o células relacionadas al corazón que tuvieran también asociados experimentos de RNA-seq en su condición *wild type* como luego del *knock out* o *knock down* de *NKX2-5*.

A partir de estos trabajos, se caracterizaron los picos de ChIP-seq con la información de los experimentos de RNA-seq con expresión diferencial de *NKX2-5*. De esta forma, se obtuvo una lista de sitios de unión de *NKX2-5* y genes cercanos que aumentan o disminuyen su expresión según presencia o ausencia de *NKX2-5* (transactivación aumentada o disminuida).

##### **4.4.1. Selección de resultados**

Teniendo en cuenta lo explicado precedentemente, se hallaron 5 publicaciones:

- He y col., 2011 [206] realizaron un trabajo con cardiomiocitos de ratón *in vitro*.
- van den Boogaard y col., 2012 [207], trabajaron con cardiomiocitos de ratón *in vivo* (extractos de corazón adulto).
- Ye y col., 2015 [208], trabajaron con cardiomiocitos de ratón *in vivo* a partir de extractos de corazón embrionario a día 12,5.
- Dupays y col., 2015 [209], trabajaron con cardiomiocitos de ratón *in vivo* a partir de extractos de corazón embrionario a día 11,5. El estudio incluye dos

repeticiones del experimento y tiene seleccionados sitios de unión que hayan aparecido en ambas réplicas.

- Anderson y col., 2018 [210], es un trabajo a partir de cardiomiocitos de humano *in vitro*.

Para nuestro estudio se seleccionaron y analizaron los datos de Anderson y col., 2018 [210], por provenir de un experimento en células humanas. También se utilizaron los datos presentados por Dupays y col., 2015 [209], por ser el trabajo más completo y más reciente en ratones *in vivo*.

En el trabajo de Anderson y col., 2018 [210] se utilizaron cardiomiocitos derivados de células madre HES3 10 días después de su diferenciación. Para el análisis de RNA-seq, se comparó una línea hemicigota para *NKX2-5* con otra con *knock out* del gen.

En el trabajo de Dupays y col., 2015 [209], utilizaron corazones de embriones de ratón al día 11,5 de una cepa diseñada para modelar cardiopatías [211]. En el análisis de RNA-seq realizado, los autores compararon una línea *wild type* para *Nkx2-5* con otra línea con expresión reducida (hipomórfica), ya que el *knock out* de *Nkx2-5* es letal en fase embrionaria. Se estudiaron los corazones de embriones al día 11,5 por ser el estado más avanzado sin diferencias morfológicas con la línea *wild type*.

Los resultados del trabajo de Anderson y col., 2018 [210], incluyen un archivo .bed con 6885 picos de ChIP-seq y un archivo con información de RNA-seq con 1174 genes con una expresión diferencial 2 veces mayor o menor que en presencia de *NKX2-5*.

Los resultados del trabajo de Dupays y col., 2015 [209], incluyen un archivo .bed con 2610 picos de ChIP-seq y un archivo con información de RNA-seq con 1492 genes con una expresión diferencial 1,5 veces mayor o menor que en presencia del silenciamiento de *NKX2-5*.

#### 4.4.2. Recopilación de sitios de unión de NKX2-5

Con los resultados de los experimentos seleccionados en humano y ratón, se recopilaron los picos de ChIP-seq de los archivos .bed y se les agregó información de sitios de unión confirmados para NKX2-5 y de genes cercanos a distintas distancias. Los enlaces para las tablas con la clasificación agregada se presentan en la sección Anexo I.5.

Los sitios de unión de NKX2-5 con los que se trabajaron fueron, por un lado, los 10 sitios utilizados experimentalmente listados en la sección 4.1.4.: GCAAGTG, GGAAGTG, GAAAGTG, ATAAGTG, GTAAGTG, CTAAGTG, TCAAGTG, TGAAGTG, TAAAGTG y TTAAGTG. Por otro lado, se utilizó la matriz de puntajes “NKX25\_HUMAN.H11MO.0.B” de HOCOMOCO ([https://hocomoco11.autosome.org/motif/NKX25\\_HUMAN.H11MO.0.B](https://hocomoco11.autosome.org/motif/NKX25_HUMAN.H11MO.0.B)) con un valor de corte para la selección del 90% del puntaje máximo posible de la matriz. Este valor de corte arrojó 10 secuencias distintas: TGGAGAGC, TGAAGAGC, TTGAGAGG, GGGAGTGG, TGAAGAGG, GGAAGAGG, TGAAGTGG, TGGAGTGG, GGGAGAGG y TGGAGAGG.

Del total de 6885 y 2610 picos de ChIP-seq en humanos y en ratón, respectivamente, 5692 contenían sitios de unión de NKX2-5 en humano y 2064 en ratón (Tabla 10). En el caso del experimento en humanos, 4887 corresponden a picos con sitios confirmados experimentalmente y 2895 a sitios con un *score* alto en la matriz de puntaje. En el caso del ratón, 2022 se corresponden a picos con sitios confirmados experimentalmente y 398 a sitios con *score* alto en la matriz de puntaje.

Por su parte, a partir de los 6885 picos que se obtuvieron en el experimento en células humanas, se definieron 4210 picos que poseían al menos un gen a una distancia menor a 100.000 pb. Entre esos picos, 728 se corresponden con picos en los que se observan genes comprendidos a la distancia de estudio con expresión diferencial en el experimento de RNA-seq. Para los 2610 picos del experimento en ratón, se definieron 1926 en los que se observa por lo menos un gen. Entre esos picos,

en 406 se observan genes con expresión diferencial en el experimento de RNA-seq (Tabla 10).

Los 728 picos a menos de 100.000 pb de genes con expresión diferencial en el experimento de RNA-seq en humano y los 406 en ratón son los picos que denominamos **confirmados**. Estos picos fueron seleccionados para el estudio de regiones con mayor evidencia experimental de ser funcionales.

**Tabla 10. Cantidad de picos de ChIP-seq con sitios de unión de NKX2-5 y/o genes cercanos.**

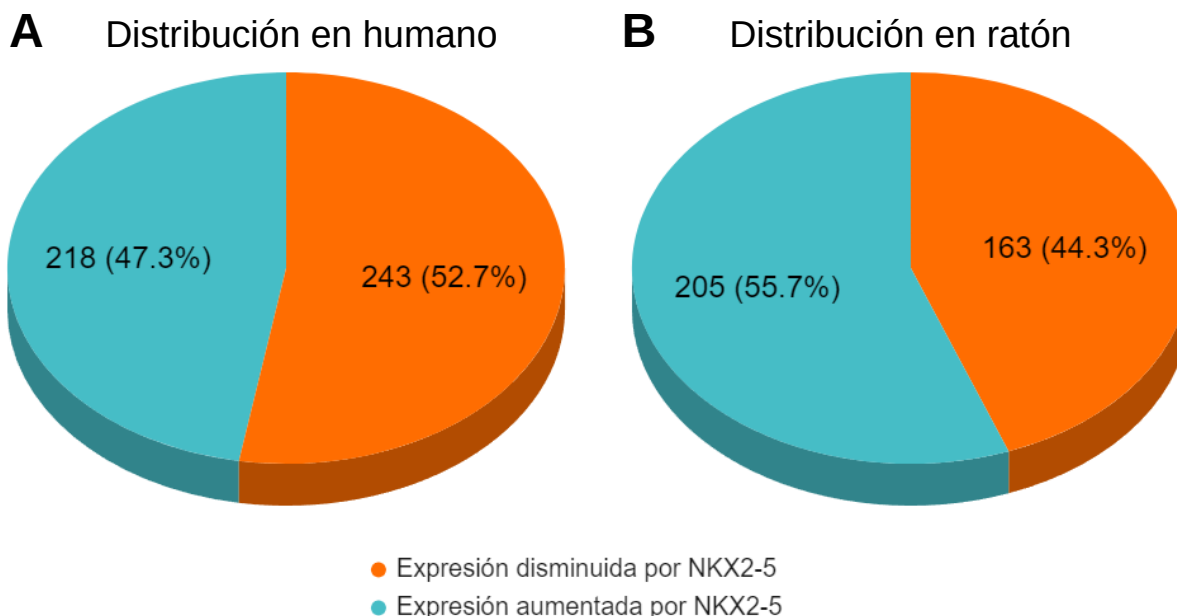
| Experimento                                       | Anderson y col., 2018 |        |        |        | Dupays y col., 2015 |        |        |        |
|---------------------------------------------------|-----------------------|--------|--------|--------|---------------------|--------|--------|--------|
| Total de picos ChIP-seq                           | 6885                  |        |        |        | 2610                |        |        |        |
| Con sitios de unión de NKX2-5                     | 5692                  |        |        |        | 2064                |        |        |        |
| Distancia para búsqueda de genes cercanos         | 1Mpb                  | 500kpb | 100kpb | 1500pb | 1Mpb                | 500kpb | 100kpb | 1500pb |
| Picos con genes cercanos                          | 6780                  | 6466   | 4210   | 429    | 2604                | 2555   | 1926   | 354    |
| Gen confirmado (todos los genes cerca)            | 3876                  | 2536   | 728    | 29     | 1721                | 1208   | 406    | 25     |
| Sitio de unión y gen confirmado (todos los genes) | 3233                  | 2125   | 615    | 17     | 1356                | 944    | 318    | 19     |
| Gen confirmado (gen más cercano)                  | 1441                  | 1248   | 574    | 27     | 404                 | 384    | 251    | 25     |
| Sitio de unión y gen confirmado (gen más cercano) | 1246                  | 1082   | 499    | 15     | 318                 | 303    | 196    | 19     |

Cuando se consigna “gen confirmado”, se hace referencia a genes con expresión diferencial en experimentos de RNA-seq con *knock out* de *NKX2-5* (en humano) o *knock down* de *Nkx2-5* (en ratón).

En el experimento en humanos, los 728 picos confirmados poseen 461 genes a una distancia menor a 100.000 pb, de los cuales 243 (53%) aumentan su expresión en ausencia de NKX2-5, por lo que serían genes que normalmente se encuentran regulados negativamente (*downregulados*) por este factor de transcripción, mientras que 218 (47%) disminuyen su expresión, en ausencia de NKX2-5, o sea que son genes para los cuales NKX2-5 activa su transcripción (*upregulados*) (Figura 29).

En ratón, se observaron 368 genes para los 406 picos confirmados fue de 368, de los cuales 163 (44%) aumentan su transcripción en ausencia de NKX2-5 (serían

genes normalmente *downregulados*), mientras que 205 (56%) disminuyen su expresión en ausencia de NKX2-5 (normalmente *upregulados*) (Figura 29).



**Figura 29. Proporción de genes que aumentan o disminuyen su expresión por NKX2-5 cerca de sitios de unión en picos de ChIP-seq.**

A: Valores en humano [210]. B: Valores en ratón [209].

A partir de estos primeros resultados, se decidió analizar en profundidad los datos provenientes de los experimentos en células humanas [210].

#### 4.4.3. Estudio de asociación de sitios de unión de NKX2-5 con regulación río abajo

Con los resultados obtenidos en células humanas, se buscó analizar la posible asociación entre la condición de regulación de genes a menos de 100.000 pb de sitios de unión de NKX2-5 y los sitios de unión en particular.

En los 728 picos analizados, se hallaron 1.444 sitios de unión de NKX2-5, los cuales se presentan en la sección Anexo I.6. Los sitios de unión que se consideraron para el análisis fueron las 20 secuencias nucleotídicas nombradas precedentemente, las 10 confirmadas experimentalmente y las 10 de la matriz de puntajes.

En la Figura 30 se observa un histograma representativo de la cantidad de genes en los que se observó expresión diferencial significativa en los ensayos de RNA-

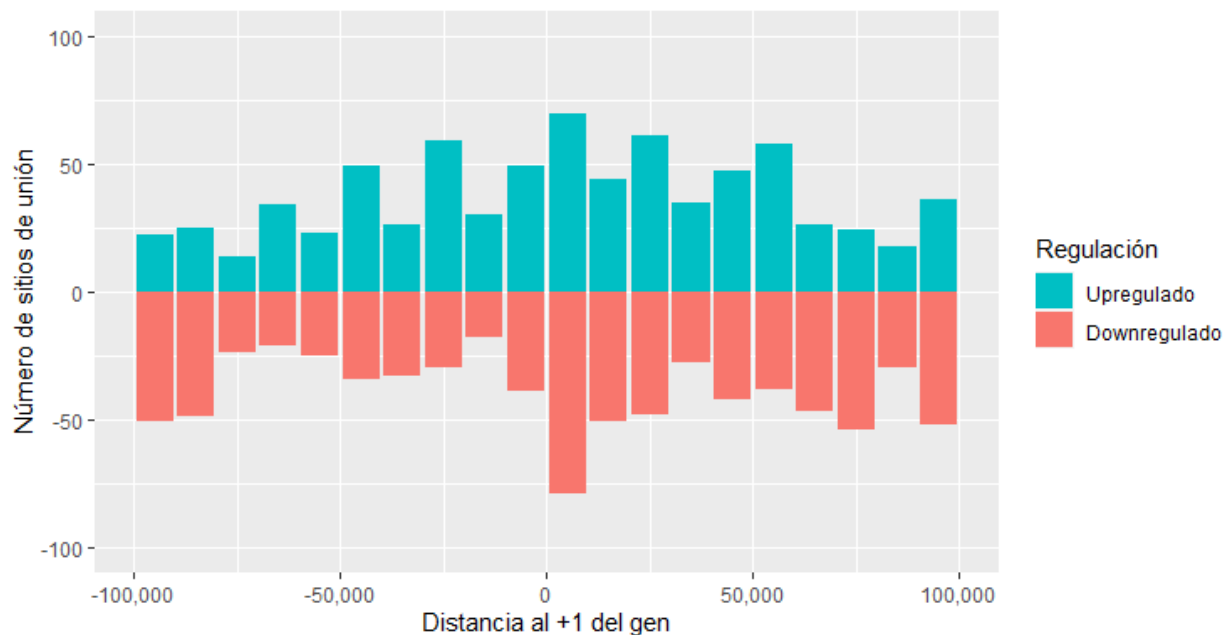
seq, ya sea con expresión aumentada o disminuida por NKX2-5, para cada una de las 20 secuencias nucleotídicas de sitios de unión de este factor de transcripción. Las secuencias más largas, correspondientes con las encontradas por la matriz de pesos, son las menos frecuentes. También se observa que, dentro de las secuencias confirmadas experimentalmente, las secuencias “TCAAGTG”, “GGAAGTG” y “GCAAGTG” tienen la mayor proporción de genes cercanos con expresión aumentada por NKX2-5 (54,3%, 54,1% y 53,2%, respectivamente), mientras que la secuencia “TTAAGTG” tiene la mayor proporción de genes cercanos con expresión disminuida por NKX2-5 (61,1%).



**Figura 30. Cantidad de sitios de unión a menos de 100.000 pb de genes por secuencia de unión.** En celeste, número de sitios con genes cercanos que aumentan su expresión en presencia de NKX2-5 en los ensayos de RNA-seq (*upregulados*). En naranja, número de sitios con genes cercanos que disminuyen su expresión en presencia de NKX2-5 en los ensayos de RNA-seq (*downregulados*).

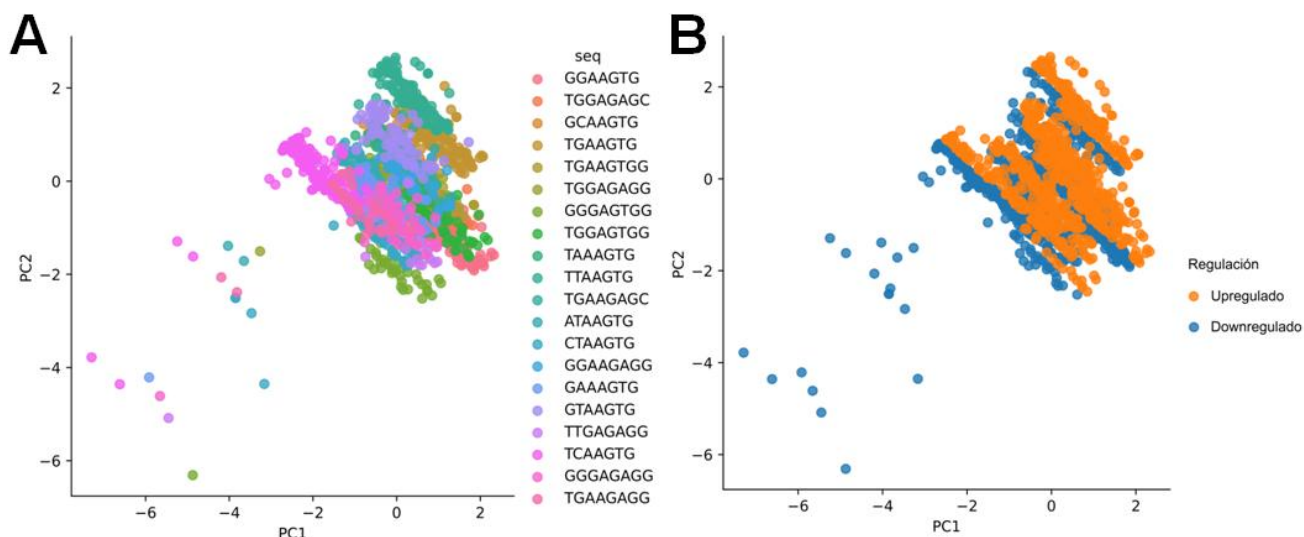
En la Figura 31 se observa un histograma representativo de la cantidad de genes con expresión diferencial significativa en los ensayos de RNA-seq, ya sea con expresión aumentada o disminuida por NKX2-5, a distintas distancias de los 1.444 sitios de unión de NKX2-5 seleccionados para esta sección. Se observa que hay una concentración de genes en las 10.000 pb río arriba o río abajo del inicio de

transcripción del gen, pero la cantidad de genes a distancias mayores sigue siendo relevante (Anexo II. Figura S1).



**Figura 31. Cantidad de sitios de unión por rangos de distancia al +1 del gen más cercano hasta una distancia de 100.000 pb.** En celeste, número de sitios con genes cercanos que aumentan su expresión en presencia de NKX2-5 (*upregulados*). En naranja, número de sitios con genes cercanos que disminuyen su expresión en presencia de NKX2-5 (*downregulados*).

En la Figura 32 se observa un gráfico representativo de la distribución de los 1.444 sitios de unión seleccionados para el análisis de componentes principales de acuerdo a sus dos primeros componentes. El conjunto de datos utilizados para el análisis fue la distancia del sitio de unión al gen cercano sin preprocesar, el valor de *fold change* del gen cercano sin modificar y las secuencias de los sitios de unión con codificación *one-hot*. Los sitios de unión de la figura se encuentran coloreados por secuencia (Figura 32A) o por regulación (Figura 32B). No se observó una separación clara de los componentes analizados.



**Figura 32. Distribución de los 1.444 sitios de unión seleccionados para el análisis de componentes principales de acuerdo a sus dos primeros componentes principales.**

**A:** Análisis por diferentes sitios de unión. Cada una de las secuencias encontradas se presentan de un color diferente. **B:** Análisis por tipo de regulación. En naranja, número de sitios con genes cercanos que aumentan su expresión en presencia de NKX2-5 (*upregulados*). En azul, número de sitios con genes cercanos que disminuyen su expresión en presencia de NKX2-5 (*downregulados*). PC1: Primer componente principal. PC2: Segundo componente principal.

Estos resultados son similares para los 8 conjuntos diferentes generados para el análisis de componentes principales. En todos los casos, las varianzas representadas por los primeros componentes principales no superó el 5,5% y de 20 componentes principales, 19 de ellos poseen proporciones de varianza entre 5,2% y 4,3% (Anexo II. Tabla S3). No fue posible seleccionar primeros componentes principales de utilidad para nuestro estudio.

En resumen, con el análisis de componentes principales no se encontró una asociación entre ninguna de las secuencias de unión, sus distancias a genes cercanos y la regulación de dichos genes por NKX2-5.

## 4.5. Estudio de sitios de unión cercanos de otros factores de transcripción y genes transactivados

En esta última parte del trabajo se estudiaron otros genes y su potencial relación a los sitios de unión confirmados para NKX2-5. Para ello, se estudió la función de los genes cercanos, así como los sitios de unión de factores de transcripción que podrían estar interactuando con NKX2-5 en la transactivación de genes río abajo.

En total, se registraron sitios de unión de 6 factores de transcripción que podrían potencialmente interactuar con NKX2-5 para la transactivación y se estudiaron los genes que podrían estar siendo transactivados o inhibidos gracias a esta interacción.

### 4.5.1. Sitios de unión de otros factores de transcripción cercanos a sitios de NKX2-5

En primer lugar, se definió una lista de factores de transcripción que probablemente interactúen con NKX2-5 y se buscaron sus sitios de unión cerca de los sitios confirmados de la sección 4.4.3. (Anexo I.6.).

En la búsqueda de factores de transcripción que interactúan con NKX2-5, se encontraron 12: *TBX5*, *GATA4*, *SMAD4*, *SRF*, *MEF2A*, *MEF2B*, *MEF2C*, *CAMTA2*, *HIPK1*, *JARID2*, *MYOCD* y *PIAS1* (Tabla 11). De estos, *TBX5* y *GATA4* tienen evidencia de interacción en células humanas en la bibliografía, *TBX5* y *MEF2A/B/C* tienen cristales heterodimerizando con NKX2-5 y el ADN, *TBX5*, *GATA4*, *SMAD4* y *MEF2A/B/C* se encontraron en STRING y *TBX5*, *GATA4*, *SMAD4*, *SRF*, *CAMTA2*, *HIPK1*, *JARID2*, *MYOCD* y *PIAS1* se encontraron en GeneMANIA (Tabla 11).

**Tabla 11. Factores de transcripción que interactúan con NKX2-5.**

| Factor de transcripción | STRING | GeneMANIA | Bibliografía específica |
|-------------------------|--------|-----------|-------------------------|
| TBX5                    | X      | X         | [146,212]               |
| GATA4                   | X      | X         | [144,213,214]           |
| SMAD4                   | X      | X         |                         |
| MEF2A                   | X      |           | [178] (*)               |
| MEF2B                   | X      |           | [178]                   |
| MEF2C                   | X      |           | [178] (*)               |
| SRF                     |        | X         |                         |
| CAMTA2                  |        | X         |                         |
| HIPK1                   |        | X         |                         |
| JARID2                  |        | X         |                         |
| MYOCD                   |        | X         |                         |
| PIAS1                   |        | X         |                         |

Bibliografía de estudios masivos para STRING y GeneMANIA: [215-219]. (\*) Uno de los cristales correspondientes a la estructura en contacto con NKX2-5 es una proteína quimérica representando a MEF2A, MEF2B y MEF2C, y el otro cristal es sólo de NKX2-5 en contacto con MEF2B.

Al analizar las secuencias de los picos de ChIP-seq con MEME-ChIP, se encontraron 553 secuencias sobrerrepresentadas que se corresponden con sitios de unión de factores de transcripción. Se utilizó un E-valor de corte de  $4,80 \times 10^{-116}$ , correspondiente al sitio de unión de NKX2-5, y se seleccionaron factores de transcripción entre los sitios de unión con un E-valor menor. La tabla con la lista completa con todos los motivos hallados se presenta en el Anexo I.7.

Con el puntaje de corte definido, se seleccionaron 7 genes que codifican para factores de transcripción con matrices de puntaje en HOCOMOCO: *TBX20*, *MEIS1*, *TGIF1*, *GATA4*, *TAL1*, *MAF* y *NKX2-5* (Tabla 12) y se procedió a realizar la búsqueda de sitios de unión de estos factores de transcripción a menos de 250 pb de distancia de los 1.444 sitios confirmados de la sección 4.4.3. (Anexo I.6.).

**Tabla 12. Factores de transcripción seleccionados en orden de mayor a menor puntaje (columna E-value) de acuerdo a MEME-ChIP.**

| Identificador del motivo de unión | Secuencia consenso  | Largo del motivo | Número de sitios | E-Valor   |
|-----------------------------------|---------------------|------------------|------------------|-----------|
| TBX20_HUMAN.H11MO.0.D             | DRGTGHTGACAGSH      | 14               | 6860             | 3.1e-718  |
| MEIS1_HUMAN.H11MO.1.B             | GMTGACAGV           | 9                | 6871             | 9.1e-366  |
| TGIF1_HUMAN.H11MO.0.A             | TGACAGS             | 7                | 6871             | 1.20E-299 |
| GATA4_HUMAN.H11MO.0.A             | DSAGATAAGA          | 10               | 6848             | 2.60E-268 |
| TAL1_HUMAN.H11MO.0.A              | CTGSNVVSBRSWGATAAGV | 19               | 6580             | 1.00E-239 |
| MAF_HUMAN.H11MO.1.B               | TGCTGAST            | 8                | 6871             | 1.10E-194 |
| NKX25_HUMAN.H11MO.0.B             | KKRAGWGS            | 8                | 6871             | 4.80E-116 |

Los recuadros marcan los motivos correspondientes a los genes con evidencia experimental de interacción con NKX2-5. También se recuadra el motivo correspondiente a NKX2-5, que es el factor de transcripción con el cual se seleccionaron las secuencias de ChIP-seq, por lo que era de esperar que se encontrara en esta lista. El identificador del motivo de unión corresponde al nombre de la matriz en la base de datos HOCOMOCO v11. El número de sitios corresponde a la cantidad de secuencias analizadas que contienen el sitio de unión correspondiente.

En el análisis se observó que a 250 pb de 636 sitios de unión de NKX2-5 confirmados experimentalmente existirían sitios de unión para al menos 1 de los 6 factores de transcripción mencionados más arriba (excluyendo a NKX2-5) y que 861 sitios tendrían al menos otro sitio de unión de NKX2-5. Los datos de los sitios de unión de NKX2-5 confirmados experimentalmente que poseen otros sitios de unión a menos de 250 pb se presentan en una tabla del Anexo I.8.

#### 4.5.2. Análisis de genes cercanos a sitios de unión seleccionados

En esta sección, se estudiaron los GO-terms de genes a menos de 100.000 pb de sitios de unión de NKX2-5. Estos fueron analizados teniendo en cuenta el total de genes regulados por NKX2-5 o bien agrupados entre aquellos que mostraron regulación positiva o negativa por NKX2-5 en los ensayos de expresión diferencial. También se realizó el análisis de GO-terms teniendo en cuenta la presencia de sitios de unión de otros factores de transcripción a menos de 250 pb de distancia. Para este análisis, se trabajó con los 1.444 sitios de unión de NKX2-5 referenciados en las tablas del Anexo I.6., y los resultados se presentan en el Anexo I.8.

Con los sitios de unión seleccionados, se compilaron 24 listas de genes, agrupadas de a 3, considerando genes en los que se observó expresión aumentada

por NKX2-5, genes con expresión disminuida, o el total de genes con expresión diferencial (Tabla 13).

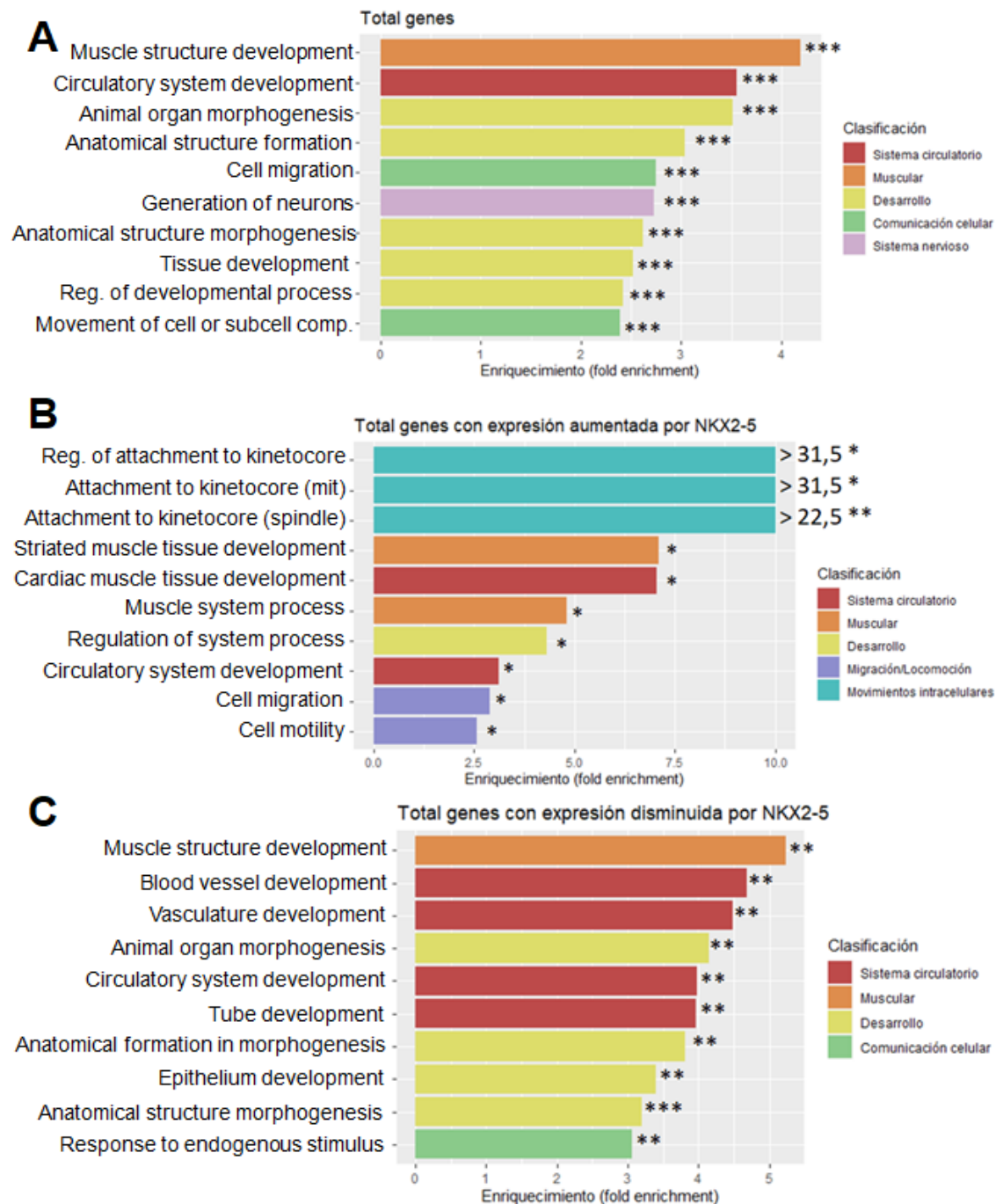
**Tabla 13. Número de genes recopilados para los factores de transcripción seleccionados por poseer sitios de unión cercanos a NKX2-5 incluidos en el análisis con ShinyGO.**

|                                          | Todos los sitios de unión | Sitio NKX2-5 cercano | Sitio TGIF1 cercano | Sitio MAF cercano | Sitio MEIS1 cercano | Sitio GATA4 cercano | Sitio TBX20 cercano | Sitio TAL1 cercano |
|------------------------------------------|---------------------------|----------------------|---------------------|-------------------|---------------------|---------------------|---------------------|--------------------|
| Total de genes regulados por NKX2-5      | 356                       | 303                  | 178                 | 110               | 83                  | 46                  | 12                  | 0                  |
| Genes regulados positivamente por NKX2-5 | 174                       | 146                  | 91                  | 58                | 50                  | 26                  | 6                   | 0                  |
| Genes regulados negativamente por NKX2-5 | 182                       | 157                  | 87                  | 52                | 33                  | 20                  | 6                   | 0                  |

Los GO-terms hallados se agruparon -a grandes rasgos- en 7 categorías, 3 de ellas específicas: sistema circulatorio, función y desarrollo muscular y sistema nervioso, y 4 más generales: desarrollo y diferenciación de tejidos/órganos, comunicación celular, migración/locomoción y movimientos intracelulares.

Los GO-terms asociados al sistema circulatorio y a la función y desarrollo muscular fueron los de mayor representación, especialmente entre los genes que se regulan en forma negativa por *NKX2-5* (con expresión disminuida o *downregulados*). Entre estos últimos, se encuentran los asociados al desarrollo de vasos sanguíneos. Asimismo, cabe destacar que se observó menor FDR entre los GO-terms de genes regulados en forma positiva (con expresión aumentada o *upregulados*).

En la Figura 33 se esquematizan los primeros 10 GO-terms seleccionados por ShinyGO para la lista de todos los genes confirmados experimentalmente a menos de 100.000 pb de un sitio de unión de NKX2-5. Los GO-terms se encuentran coloreados de acuerdo a sus clasificaciones en código de color y ordenados en función de su enriquecimiento relativo en la lista de genes.



**Figura 33. GO-terms sobrerrepresentados de genes cerca de sitios de unión de NKX2-5 confirmados experimentalmente.** El eje X representa el enriquecimiento relativo de cada GO-term con respecto al esperado por azar. Los valores extremos fueron recortados y anotados al costado de las columnas correspondientes. **A**, GO-terms sobrerrepresentados en la lista de todos los genes a menos de 100.000 pb de sitios de unión de NKX2-5 confirmados experimentalmente. **B**, GO-terms sobrerrepresentados en la lista de sólo los genes con expresión disminuida por NKX2-5 (*downregulados*). **C**, GO-terms sobrerrepresentados en la lista de sólo los genes con expresión aumentada por NKX2-5 (*upregulados*). Los asteriscos denotan la probabilidad de falso positivo (FDR).

\*: FDR<0,001; \*\*: FDR<0,00001; \*\*\*: FDR<10<sup>-10</sup>.

Se puede ver, en los GO-terms sobrerrepresentados en genes con expresión aumentada o disminuida por NKX2-5 (Figura 33A), que los primeros 2 GO-terms se corresponden con desarrollo muscular y del sistema circulatorio, mientras que 7 de los otros 8 GO-terms se corresponden con funciones generales de desarrollo o comunicación celular. Analizando sólo los genes con expresión aumentada por NKX2-5 (Figura 33B), en cambio, 9 de los 10 GO-terms dejan de estar significativamente sobrerrepresentados, y los primeros 3 GO-terms se encuentran asociados a microtúbulos del huso cromosómico. Por otro lado, en los genes con expresión disminuida por NKX2-5 (Figura 33C) se ven 4 GO-terms asociados al sistema circulatorio, 3 de los cuales corresponden al desarrollo vascular.

A continuación se resumen los resultados obtenidos para los primeros 10 GO-terms seleccionados por ShinyGO teniendo en cuenta los otros sitios de unión cercanos para los 6 factores de transcripción seleccionados, los cuales se encuentran esquematizados en las Figuras S2-S7 del Anexo II.

- TGIF1 (Anexo II. Figura S2)

Se observó, a grandes rasgos, que hay una mayor cantidad de GO-terms asociados al sistema circulatorio y a músculos que en el total de genes sin filtrar por sitios de unión de TGIF1 (Anexo II. Figura S2 y Figura 33). El total de genes con expresión aumentada o disminuida por NKX2-5 (Anexo II. Figura S2A) y los genes con disminución de expresión en presencia de NKX2-5 (Anexo II. Figura S2C) tienen sobrerrepresentación de GO-terms de sistema circulatorio/muscular, y los genes con aumento de expresión en presencia de NKX2-5 (Anexo II. Figura S2B) contienen GO-terms asociados a la separación de cromosomas durante la división celular. También se observó que ninguno de los GO-terms asociados a los genes con aumento de expresión en presencia de NKX2-5 fueron significativos.

En particular, entre los genes con aumento de expresión en presencia de NKX2-5, el GO-term con mayor enriquecimiento fue “*neuronal action potential*”, el cual hace referencia a la generación del impulso nervioso en las neuronas y tiene 4 genes asociados: *CACNA1G*, *DRD1*, *P2RX1*, *SCN5A*.

- MAF (Anexo II. Figura S3)

Se observó una sobrerrepresentación de un GO-term muy específico asociado al sistema nervioso, que se corresponde con 3 genes: *ZIC3*, *MAFB* y *EGR2*, 2 de ellos (*MAFB* y *EGR2*) con expresión diferencial disminuida y el otro (*ZIC3*) aumentada por NKX2-5.

Tanto en el análisis teniendo en cuenta todos los genes regulados por NKX2-5 como en el caso de tomar en cuenta sólo los genes con expresión diferencial disminuida, se observó una gran cantidad de GO-terms asociados al desarrollo muscular. El único GO-term cardíaco fue específico del desarrollo de músculo cardíaco.

Aunque sólo 3 de los GO-terms encontrados tuvieron FDR significativos, en el análisis tomando en cuenta sólo aquellos genes con expresión diferencial aumentada, se encontraron GO-terms con FDR muy bajos, correspondientes a un máximo de 4 genes, ninguno de los cuales fue significativo. Cabe destacar en esta lista que también están presentes *NPPA* y *SCN5A*, asociados al GO-term con mayor enriquecimiento, correspondiente a la repolarización de la membrana muscular auricular.

- MEIS1 (Anexo II. Figura S4)

Se encontraron GO-terms representados por pocos genes y sin significancia estadística. En el análisis realizado teniendo en cuenta el total de genes que estarían regulados por NKX2-5 (Anexo II. Figura S4A), se encontró que 8 de los 10 GO-terms están relacionados al desarrollo del sistema circulatorio o muscular.

Entre los GO-terms hallados para aquellos genes con expresión diferencial disminuida por NKX2-5 (Anexo II. Figura S4C), cabe destacar 2 genes relacionados al desarrollo del nodo sinoatrial: *ISL1* y *TBX18*. Estos 2 genes están listados en los 4 primeros GO-terms.

Entre las vías halladas para aquellos genes con expresión diferencial aumentada (Anexo II. Figura S4B) se encontró una situación similar a la observada para MAF, con GO-terms representados por un máximo de 6 genes. También comparten el GO-term

con mayor enriquecimiento, correspondiente a la repolarización de la membrana muscular auricular, representado por los mismos 2 genes: *NPPA* y *SCN5A*.

- GATA4 (Anexo II. Figura S5)

El análisis de ShinyGO no arrojó resultados de GO-terms para aquellos genes con expresión diferencial aumentada o disminuida por *NKX2-5*. Cuando el análisis se realizó teniendo en cuenta el total de genes que estarían regulados por *NKX2-5*, ShinyGO obtiene 5 GO-terms para desarrollo muscular y 1 GO-term asociado a funciones del sistema nervioso, si bien no poseen significancia estadística.

- TBX20 (Anexo II. Figura S6)

Ninguno de los GO-terms encontrados tuvo significancia estadística, pero la revisión de genes individuales asociados a los GO-terms encontrados fue relevante para este trabajo.

En particular, los 10 GO-terms encontrados entre sus 6 genes con expresión diferencial aumentada por *NKX2-5* provienen de 4 de los genes: *FHL2*, *NOS3*, *NPPA* y *ZIC3*, y todos están asociados al desarrollo cardíaco. Analizando los genes con expresión diferencial disminuida por *NKX2-5*, se suma un gen de interés: *TGFB3*, que está asociado al desarrollo muscular.

- Múltiples sitios de *NKX2-5* (Anexo II. Figura S7)

Los agrupamientos de GO-terms para los genes cerca de múltiples sitios de unión de *NKX2-5* fueron similares a los observados cuando se analizó el total de genes regulados por este factor de transcripción (Figura 33).

- Genes específicos identificados

Se identificaron 11 genes de interés, 2 de ellos seleccionados por estar asociados a GO-terms específicos de interés y 9 por encontrarse representados en múltiples factores de transcripción:

- Entre los genes cerca de sitios de unión de MEIS1 se observaron 2 genes: *ISL1* y *TBX18*, que son regulados negativamente por NKX2-5 y que se relacionan específicamente al desarrollo del nodo sinoatrial.
- Entre los diferentes análisis realizados, se observaron 9 genes comúnmente representados: *FHL2*, *LIMCH1*, *NOS3*, *NPPA*, *SCN5A*, *SERPINE1*, *TGFB3*, *UCHL1* y *ZIC3*. De éstos, hay 5 genes que son regulados en forma positiva por NKX2-5 (*FHL2*, *NOS3*, *NPPA*, *SCN5A* y *ZIC3*) y aparecen representados principalmente en GO-terms cardíacos y musculares. Por otro lado, hay 4 genes que son regulados en forma negativa por NKX2-5 (*LIMCH1*, *SERPINE1*, *TGFB3* y *UCHL1*) y que aparecen representados en distintos grupos de GO-terms asociados al desarrollo y diferenciación de tejidos y/u órganos y señalización.

#### 4.5.3. Análisis de genes seleccionados en relación a subtipos de CC

A lo largo de este trabajo, se seleccionaron 21 genes con posibles interacciones con NKX2-5: 4 genes que codifican para proteínas encontradas en contacto con NKX2-5 en cristales de PDB (*TBX5* y *MEF2A/B/C*), 2 genes que codifican para factores de transcripción con interacción confirmada específicamente con NKX2-5 (*TBX5* y *GATA4*), 5 genes que codifican para factores de transcripción predichos con el análisis por MEME-ChIP (*MAF*, *MEIS1*, *TAL1*, *TBX20* y *TGIF1*) y 11 genes de interés transactivados o inhibidos por NKX2-5 (*FHL2*, *ISL1*, *LIMCH1*, *NOS3*, *NPPA*, *SCN5A*, *SERPINE1*, *TBX18*, *TGFB3*, *UCHL1* y *ZIC3*).

En la Tabla 14 se presentan los fenotipos más comúnmente asociados a VGs de NKX2-5 y aquellos genes de los seleccionados con VGs asociadas a estos fenotipos. Esta tabla expande la información presentada en la Tabla 7, cuya información se repite para facilitar su visualización.

**Tabla 14. Fenotipos asociados a cada uno de los genes seleccionados con relación funcional a *NKX2-5*.**

| Fenotipo asociado                 | N° de VGs <i>missense</i> en <i>NKX2-5</i> | N° de VGs globales en <i>NKX2-5</i> | Genes con interacción confirmada con <i>NKX2-5</i> | Factores de transcripción seleccionados con MEME-ChIP | Genes con expresión aumentada por <i>NKX2-5</i> | Genes con expresión disminuida por <i>NKX2-5</i> |
|-----------------------------------|--------------------------------------------|-------------------------------------|----------------------------------------------------|-------------------------------------------------------|-------------------------------------------------|--------------------------------------------------|
| Comunicación interauricular       | 48                                         | 31                                  | <i>GATA4, TBX5</i>                                 | <i>TBX20</i>                                          | <i>NPPA, ZIC3</i>                               | <i>ISL1</i>                                      |
| Comunicación interventricular     | 46                                         | 9                                   | <i>GATA4, TBX5</i>                                 | <i>TBX20</i>                                          | <i>ZIC3</i>                                     | <i>ISL1, TBX18</i>                               |
| Bloqueo atrioventricular          | 22                                         | 26                                  | <i>TBX5</i>                                        |                                                       |                                                 |                                                  |
| Defecto en septo atrioventricular | 16                                         | 2                                   | <i>GATA4, TBX5</i>                                 |                                                       | <i>ZIC3</i>                                     |                                                  |
| Tetralogía de Fallot              | 11                                         | 1                                   | <i>GATA4, TBX5</i>                                 |                                                       |                                                 |                                                  |
| Fibrilación atrial                | 7                                          | 2                                   | <i>GATA4, TBX5</i>                                 | <i>TBX20</i>                                          | <i>NPPA, SCN5A</i>                              | <i>ISL1</i>                                      |
| Disgénesis de tiroides            | 6                                          | 0                                   |                                                    |                                                       |                                                 |                                                  |
| Cardiomiopatía dilatada           | 5                                          | 3                                   | <i>GATA4, TBX5, MEF2C</i>                          | <i>TBX20</i>                                          | <i>FHL2, NOS3, SCN5A</i>                        |                                                  |
| Muerte súbita                     | 4                                          | 4                                   | <i>TBX5</i>                                        |                                                       | <i>SCN5A</i>                                    |                                                  |
| Heterotaxia                       | 0                                          | 2                                   |                                                    |                                                       | <i>ZIC3</i>                                     |                                                  |

VGs: Variantes génicas. CIA: Comunicación interauricular. CIV: Comunicación interventricular. CM: Cardiomiopatía. FA: Fibrilación auricular. TGV: Transposición de grandes vasos. TdF: Tetralogía de Fallot. DSAV: Defecto en septo atrioventricular. <sup>1</sup>Con asociación al desarrollo del corazón y/o sistema circulatorio. <sup>2</sup>Con asociación al desarrollo del corazón y/o cardiopatías en organismos modelo. <sup>3</sup>Con efectos observados en genes con asociación a CC. <sup>4</sup>Con función de protección y/o respuesta a daños en el corazón y/o sistema circulatorio.

Fuentes: [32] (*GATA4* y *TBX5*); [220-223] (*GATA4*); [224-231] (*TBX5*); [232] (*MEF2A/B/C*); [233] (*MEF2C*); [234] (*MAF*); [235,236] (*MEIS1*); [237-240] (*TAL1*); [241-244] (*TBX20*); [245,246] (*TGIF1*); [247] (*FHL2*); [248-251] (*NOS3*); [252-254] (*NPPA*); [255-259] (*SCN5A*); [260-262] (*ZIC3*); [263-266] (*ISL1*); [267,268] (*LIMCH1*); [269-271] (*SERPINE1*); [272] (*TBX18*); [273-275] (*TGFB3*); [276-279] (*UCHL1*).

Se observó, entre los genes con interacción confirmada con *NKX2-5*, que las variantes en *GATA4* y *TBX5* se encuentran asociadas a muchos de los mismos fenotipos que las variantes de *NKX2-5*. Las variantes en *MEF2C* sólo parecen tener asociación con cardiomiopatía dilatada y no se encontraron variantes en *MEF2A/B* asociadas con ninguno de los fenotipos seleccionados.

En cuanto a los factores de transcripción seleccionados con MEME-ChIP, sólo *TBX20* tuvo variantes asociadas a los fenotipos más comunes entre variantes de *NKX2-5*. De los demás factores de transcripción seleccionados, se encontraron

potenciales asociaciones al desarrollo cardíaco para *MEIS1* [235,236], *TGIF1* [28,245,246,280] y *TAL1* [237-240]. No se encontró una asociación de *MAF* con el desarrollo cardíaco [234].

Entre los genes seleccionados con expresión aumentada por *NKX2-5*, todos tuvieron alguna variante asociada a alguno de los fenotipos más comunes de variantes en *NKX2-5*.

Por otro lado, entre los genes con expresión disminuida por *NKX2-5*, *ISL1* y *TBX18* fueron los únicos con variantes asociadas a los fenotipos seleccionados. Se pudo encontrar potenciales asociaciones con fenotipos cardíacos para *SERPINE1* [269-271], *TGFB3* [273-275] y *UCHL1* [276-279], pero no para *LIMCH1*.

A grandes rasgos, se pudo ver que la mayoría de los genes seleccionados se asociaron a 4 fenotipos (CMD, FA, CIA y CIV), mientras que ninguno se encontró asociado a fenotipos tiroideos.

Como no se encontraron variantes patogénicas que afecten específicamente la interacción entre *NKX2-5* y otras proteínas, no se pudieron comparar sus fenotipos asociados a los fenotipos asociados a otros genes. Aún así, se destacan dos observaciones:

- *TBX5* es el único gen seleccionado con variantes asociadas a bloqueo atrioventricular, el cual es un fenotipo observado principalmente en variantes del homeodominio de *NKX2-5* o de efecto truncante.
- *ZIC3* es el único gen de todos los seleccionados con variantes asociadas a heterotaxia, y el resto de sus variantes asociadas son de efecto estructural.

## 4.6. Flujo de trabajo replicable

Teniendo en cuenta todo lo visto en este trabajo, se creó un flujo de trabajo lineal y simplificado comenzando con la obtención de VGs y finalizando con las

conclusiones sobre potencial patogenicidad correspondientes. El flujo de trabajo consiste de los siguientes pasos:

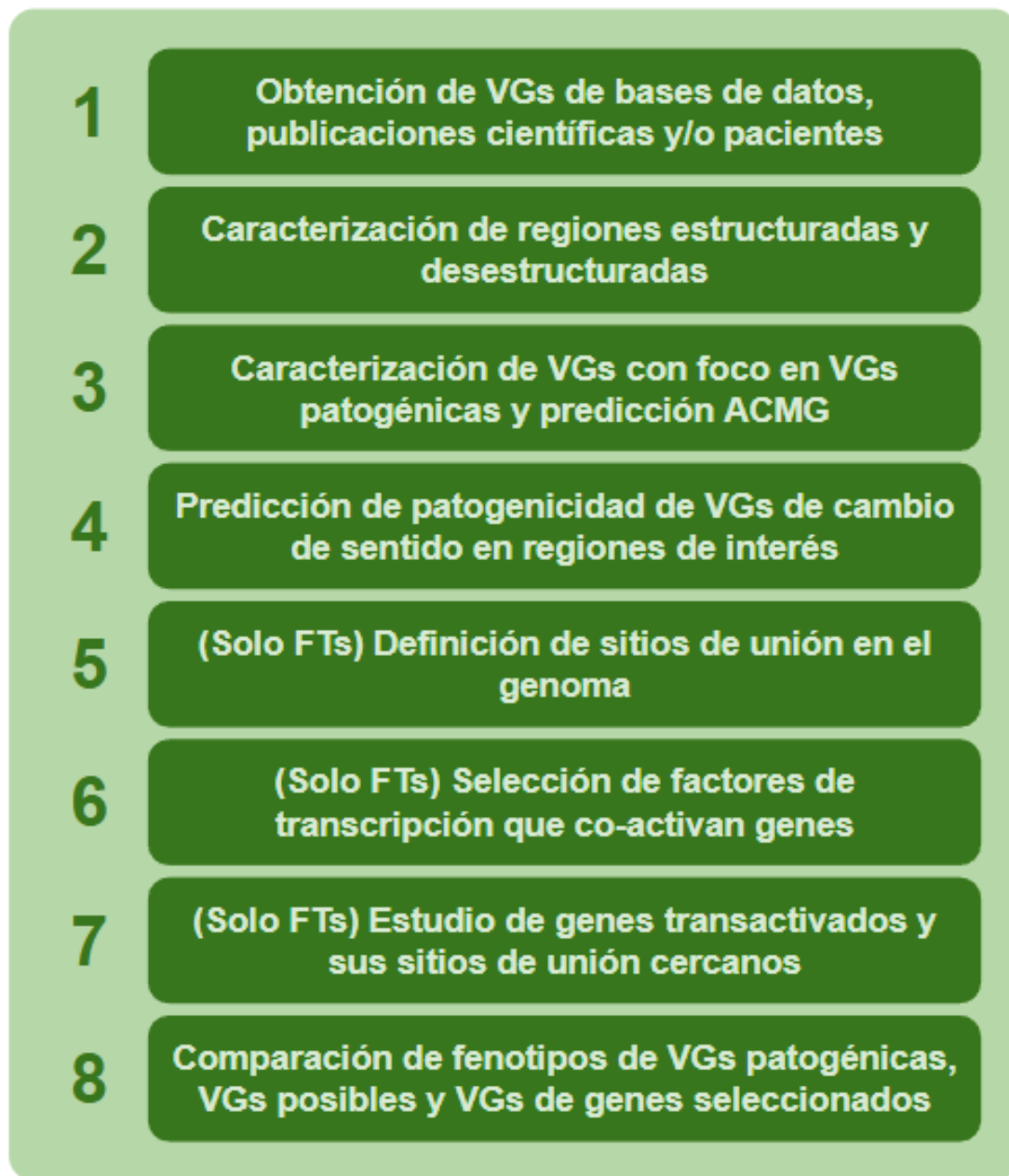
1. **Obtención de VGs para el gen de interés.** Estas variantes pueden ser obtenidas de bases de datos, publicaciones científicas y/o pacientes. Se obtiene una lista de VGs de distintas fuentes.
2. **Caracterización de regiones estructuradas y desestructuradas.** Esto implica la búsqueda de estructuras tridimensionales -o su obtención *in silico*-, la caracterización de todos los residuos en esta estructura y la predicción de SLiMs y regiones conservadas. Se obtiene un modelo estructural del gen completo.
3. **Caracterización de VGs.** Incluye el estudio de su posición respecto del modelo estructural obtenido en el paso 2, con foco en VGs patogénicas, y una predicción de su patogenicidad por criterios estandarizados como ACMG. Se obtiene una lista procesada de VGs y patrones de distribución de las distintas características estudiadas (concentración de VGs por región del gen, distribución de VGs patogénicas por subtipo, etc.).
4. **Predicción de patogenicidad de VGs en regiones de interés.** Pensado específicamente para el estudio de variantes de cambio de sentido, aunque se puede aplicar a otros tipos de VGs si se poseen criterios *in silico* para determinar su potencial patogenicidad. Las regiones de interés pueden ser regiones estructuradas u otras regiones con relevancia funcional, aunque esta selección depende de la posibilidad de hacer los ensayos *in silico* correspondientes sin una estructura tridimensional. Se obtiene una lista de VGs posibles y sus potenciales efectos en la funcionalidad del gen.
5. **Definición de sitios de unión en el genoma.** Sólo relevante para factores de transcripción y, potencialmente, otros genes que se unen al ADN. Su caracterización es necesaria para el paso 7 y aporta información utilizable en el paso 6. Se obtiene una lista de sitios de unión en el genoma, preferiblemente confirmada por ensayos de transactivación de genes cercanos.
6. **Selección de proteínas que entren en contacto con la proteína codificada por el gen de interés.** Para el caso de factores de transcripción, implica la selección de potenciales co-activadores. Puede realizarse por estructuras

tridimensionales conocidas en contacto, por sobrerepresentación de sitios de unión en picos de ChIP-seq o por otros ensayos no realizados en este trabajo. Se obtiene una lista de proteínas potencialmente en contacto con el ADN.

7. **Estudio de genes transactivados y sus sitios de unión cercanos.** Sólo relevante para factores de transcripción. Se caracterizan genes transactivados y los sitios de unión en su región regulatoria, a grandes rasgos por sus GO-terms y por estudio de referencias bibliográficas en caso de seleccionarse unos pocos genes de interés. Se obtiene una lista de genes transactivados y una lista de sitios de unión incluyendo sitios de unión del paso 5 y sitios de unión de proteínas seleccionadas en el paso 6.
8. **Comparación de fenotipos de las distintas fuentes.** Esto implica comparar los fenotipos asociados a VGs patogénicas, VGs posibles y VGs de genes seleccionados. Se priorizan VGs patogénicas en residuos con contacto confirmado con otros genes y VGs de genes seleccionados con fenotipos específicos similares. Se obtienen genes y VGs patogénicas asociadas a fenotipos específicos.

Los genes y VGs patogénicas asociadas a fenotipos específicos permiten ver patrones de distribución comunes. De esta forma, si un fenotipo poco frecuente se ve asociado a una VG en un residuo en contacto con otro gen y este gen, a su vez, se encuentra asociado a este fenotipo, se puede estimar una potencial vía de desarrollo de patogénesis. En casos de fenotipos más frecuentes, se puede estudiar su proporción en las VGs patogénicas y su asociación (o no asociación) a los distintos genes seleccionados para direccionar futuros estudios, aunque no sea probable que se pueda determinar una vía única para su desarrollo.

Un esquema del flujo de trabajo resumido se presenta en la Figura 34.



**Figura 34. Flujo de trabajo para el estudio de la patogenicidad de VGs en un gen (preferiblemente un factor de transcripción) determinado.** VGs: Variantes génicas. FTs: Factores de transcripción.

Particularmente, para el caso de *NKX2-5* y el estudio de CC, y considerando lo visto en este trabajo (Tablas 11 y 14), determinamos 3 grupos de genes de interés para estudios futuros. Estos genes serían los que, de acuerdo a la información disponible, tendrían una mayor cantidad de información para aportar al estudio del desarrollo de CC y, a su vez, cuyos resultados puedan complementar los obtenidos en este trabajo. De más relevante a menos relevante, estos grupos serían:

1. ***GATA4* y *TBX5***. Son los dos genes con más evidencia de interacción con NKX2-5, así como cruciales en el desarrollo del corazón. *TBX5*, además, tiene una estructura tridimensional en contacto con NKX2-5 para estudiar.
2. ***MEF2A/B/C***. Son tres genes con una estructura tridimensional de sus proteínas codificadas en contacto con NKX2-5, y el único otro gen con esta característica es *TBX5*.
3. ***ISL1*, *TBX20*, *TBX18* y *ZIC3***. Son genes seleccionados por distintas razones entre factores de transcripción que potencialmente interactúen con NKX2-5 (*TBX20*), genes con expresión aumentada por NKX2-5 (*ZIC3*) y genes con expresión disminuida por NKX2-5 (*ISL1* y *TBX18*). *TBX20* es el único entre los factores de transcripción que potencialmente interactúen con NKX2-5 con asociación específica a CC. *ZIC3* tiene una asociación específica con heterotaxia que es potencialmente interesante para entender el desarrollo de algunos subtipos estructurales de CC. *ISL1* y *TBX18* tienen una asociación específica con el desarrollo del nodo sinoatrial que es potencialmente relevante para el desarrollo de cardiopatías de conducción (específicamente bloqueo atrioventricular).

## **5. Discusión**

El trabajo presentado tuvo como eje central el estudio de uno de los principales genes asociados a CC, *NKX2-5*.

Este gen codifica para un factor de transcripción crucial en la formación del corazón, la proteína *NKX2-5* [40,47] y fue el primer gen donde una variante génica se asoció al desarrollo de CC [38]. Desde entonces se han reportado más de 100 VGs patogénicas, la mayoría de ellas asociadas a CC, si bien se puede encontrar asociada a otras patologías [32,40,42]. Asimismo, y como se observa en general en CC, se describió que un mismo subtipo de CC puede estar asociado a distintas VGs y una misma variación en el gen puede estar asociada a distintas CC [29].

Como es común en factores de transcripción de mamíferos, *NKX2-5* actúa en interacción con numerosos factores de transcripción. Entre éstos, se encuentran *TBX5* [146,209], *GATA4* [144,210,211], *MEF2B* [212], entre otros (Tabla 11), así como factores de transcripción que no habían sido descritos al inicio de este trabajo. Estas interacciones pueden ser tanto de cooperación, para actuar sobre genes río abajo [281], como de competencia, por sitios superpuestos [282]. Esta complejidad de interacción entre factores de transcripción podría aportar información crucial para entender la relación entre los genotipos y sus fenotipos asociados en pacientes con CC.

En este trabajo nos propusimos desarrollar un flujo de trabajo para VGs de *NKX2-5* que permita el estudio descriptivo-predictivo de su patogenicidad. También estudiamos la interacción de *NKX2-5* con otros factores de transcripción y las implicancias en el desarrollo del corazón que podrían tener a través de sus sitios de unión en el genoma.

Para ello, recopilamos todas las VGs de *NKX2-5* de publicaciones y bases de datos, desarrollamos un modelo estructural de la proteína *NKX2-5* de largo completo, estudiando todas las variantes de cambio de sentido posibles en su homeodominio, la ocurrencia de SLiMs y su interacción potencial con sitios de unión de genes cercanos y de otros factores de transcripción. Con esta información, se colaboraría a disminuir la brecha entre la información de genotipos y sus fenotipos asociados al proponer

variantes de interés en *NKX2-5*, factores de transcripción con los que interactúa y genes transactivados que puedan afectar vías metabólicas relevantes para el desarrollo de patologías específicas.

## 5.1. Estudio de VGs de *NKX2-5*

Las primeras observaciones de las VGs recopiladas dieron cuenta de que su distribución es más densa en la región traducida de la isoforma canónica de la proteína (Figura 23A). Esta observación es de esperar si se tiene en cuenta que la mayoría de las variantes provienen de estudios de la región codificante del gen y, por lo tanto, sesga la proporción de variantes patogénicas encontradas. Considerando este sesgo, el hecho de que todas las variantes patogénicas encontradas tengan un efecto en la secuencia de la proteína *NKX2-5* puede representar una limitación de la información disponible y, de hecho, tanto en ratón como en *Drosophila* se han encontrado asociaciones funcionales entre determinadas secuencias en la región no codificante de homólogos de *NKX2-5* y el desarrollo del corazón [283,284]. Por lo tanto, no se puede descartar esta posibilidad también para humanos, más teniendo en cuenta la existencia de algunos estudios de variantes sinónimas polimórficas con resultados contradictorios entre sí [285,286]. De cualquier manera, la ausencia de VGs patogénicas no codificantes conocidas en humano hace que su estudio esté fuera del alcance de este trabajo.

Por todo lo expuesto, se hizo foco en las variantes con efecto en la secuencia aminoacídica. Esto realza la importancia del modelo estructural de *NKX2-5* para el estudio de la patogénesis, ya que provee información que permite realizar estudios más complejos en las distintas regiones de la proteína y se le dió importancia al estudio de las variantes de cambio de sentido, por ser consideradas las más interesantes. Esto es así ya que, por un lado, ninguna variante sinónima fue clasificada como patogénica y, por el otro, las deleciones e inserciones de cambio de sentido y variantes que introducen señales de *stop* prematuros tienen un efecto sobre toda la secuencia que es

muy difícil de estudiar más allá de predecir que la proteína probablemente pierda gran parte de su funcionalidad normal.

Para realizar el análisis de variantes patogénicas, y especialmente aquellas de cambio de sentido, se dividió a la proteína -a grandes rasgos- en dos partes: por un lado la región fuera del homeodominio y por otro el homeodominio propiamente dicho.

Para la región fuera del homeodominio de NKX2-5 no se halló ninguna estructura tridimensional asociada y la concentración de variantes patogénicas fue menor en comparación al homeodominio. En particular, la caracterización de motivos lineales confirmados experimentalmente permitió asignar una funcionalidad específica potencial para 9 de estos residuos. Estos motivos lineales con residuos fuera del homeodominio se encuentran asociados a dos funcionalidades específicas: la localización nuclear y la SUMOylación.

La señal de localización nuclear dirige a NKX2-5 al núcleo [176,177], por lo que es de esperar que su modificación afecte la ubicación de NKX2-5 dentro de la célula y esto explicaría el efecto patogénico de las VGs halladas (Tabla 6). Por otro lado, se describió que la SUMOylación de Nkx2-5 en ratón es crucial para su función [178,179]. Esta modificación postraducciona se condice con un motivo lineal conservado entre ratón y humano (residuos 51-54 en humano y 50-53 en ratón), por lo que podría tener una función similar en NKX2-5 humano. Si se confirmara que el efecto es también importante en células humanas, podría ser el mecanismo por el cual la presencia de VGs en la región se asocian a una disfuncionalidad de la proteína y a la aparición de CC. Esto podría explicar, por ejemplo, la patogenicidad de dos variantes de cambio de sentido encontradas en este trabajo (p.Phe51Leu y p.Glu54Lys; Tabla 6), así como variantes que pudieran encontrarse en el futuro.

Por último, es posible que parte de la región fuera del homeodominio mantenga una estructura tridimensional relativamente estable en presencia de otras proteínas, como se sabe que sucede con proteínas y regiones intrínsecamente desordenadas [95,287,288]. De hallarse evidencia del plegado de regiones desestructuradas que adoptaran estructuras secundarias o terciarias cuando entran en contacto con otras

proteínas, se podría estudiar la funcionalidad de residuos en la región desestructurada para los cuales actualmente no hay información. Esto permitiría encarar estudios como los realizados en la sección 4.3. y entender mejor la patogenicidad de algunas de las 67 VGs patogénicas de cambio de sentido fuera del homeodominio.

Por su parte, la región del homeodominio propiamente dicho es la única región de NKX2-5 con estructura tridimensional conocida, es crucial en su funcionalidad [61] y es la que posee una mayor cantidad de variantes patogénicas que otras regiones (Figura 25). De hecho en estudios previos ya se había observado que el homeodominio es la región con mayor concentración de variantes patogénicas de cambio de sentido [32,65]. En este sentido, este trabajo de tesis refuerza esta observación y aumenta su precisión, al describir una concentración de variantes patogénicas en la tercera hélice alfa del homeodominio y destacar la ausencia de variantes con frecuencia mayor a 1/10.000 en bases de datos poblacionales en esta región. De destacar además, es que los motivos lineales confirmados experimentalmente para NKX2-5 y las modificaciones post-traduccionales conocidas se concentraron también en el homeodominio, donde se observaron 3 de 4 de estos motivos lineales, así como una lisina probablemente acetilada con implicancias funcionales. Esto suma evidencia de la importancia funcional del homeodominio que complementa lo visto para VGs.

Dentro del homeodominio, se encontraron 23 variantes clasificadas como patogénicas en 15 de los 16 residuos de su tercera hélice alfa (residuos 179 a 194). Esta particularidad no se observa en la primera y segunda alfa hélice y es consistente con lo que se observa en las estructuras tridimensionales, ya que el homeodominio es la región de NKX2-5 que está en contacto con el ADN y la tercera hélice alfa se inserta en el surco mayor del ADN confiriéndole la especificidad de secuencia [146,176-178]. Cuando se analizaron posibles variantes de cambio de sentido del homeodominio se identificó una diferencia entre variantes que se ubican en los primeros 6 residuos (179 a 184) de la tercera alfa hélice de aquellas en los últimos 10 (aminoácidos 185 a 194). Mientras que los últimos 10 residuos se encuentran en contacto con el ADN, los primeros 6 residuos contienen una lisina (Lys183) que se acetila con implicancias funcionales en Nkx2-5 de ratón [180,181]. Esta lisina no se asoció a ningún motivo

lineal predicho por ELM y la proteína que acetila a Nkx2-5 de ratón, p300, no tiene un motivo lineal conocido. Ya sea que se encuentre un motivo lineal o que el reconocimiento sea por la estructura tridimensional alrededor de la lisina, es muy probable que los residuos a su alrededor tengan relevancia funcional. De esta forma, se podría explicar la patogenicidad de 4 de las 5 variantes de cambio de sentido patogénicas en el homeodominio sin explicación funcional encontrada en la sección 4.3. Además, sabiendo que se conoce la interacción de p300 con MEF2 [183], proponemos que es posible que la interacción de MEF2A/B/C con NKX2-5 ayude a determinar la acetilación de la lisina 183.

Más generalmente, si se comparan los mecanismos probables de patogenicidad de todas las VGs de cambio de sentido posibles con aquellos asociados a las VGs patogénicas ya descritas (Tabla 9), se observa una disminución de 32,3% a 13,9% de VGs sin un mecanismo de patogenicidad asociado. Esto es esperable, considerando que las variantes patogénicas deben tener un mecanismo asociado para el desarrollo de la patogénesis. También se observó que las variantes que introducen cisteína, que afectan la interacción con otras proteínas o motivos lineales confirmados o que afectan la estabilidad o motivos lineales predichos, tienen proporciones similares en ambos casos. Estas observaciones estarían sugiriendo que muy probablemente estos mecanismos poseen una relevancia relativa menor en el desarrollo de la patogénesis en el homeodominio, aunque no descarta que siguen siendo mecanismos relevantes para estudiar.

Con el objetivo de tener una definición estandarizada de patogenicidad que pueda compararse con otras fuentes, se utilizó la clasificación de patogenicidad según los criterios del ACMG. Desde su publicación en 2015, los criterios de ACMG han tenido revisiones y cambios en recomendaciones de distinto alcance. El grupo SVI (por sus siglas en inglés para *Sequence Variant Interpretation*) (<https://www.clinicalgenome.org/working-groups/sequence-variant-interpretation/>) de ClinGen se encargó de refinar muchos de estos criterios y compiló distintos documentos con actualizaciones y recomendaciones (<https://clinicalgenome.org/curation-activities/variant-pathogenicity/documents/>). En el

contexto de este trabajo, es importante recordar que parte de la importancia de la clasificación según el criterio de ACMG es que el registro de los criterios de clasificación permite la reinterpretación de los datos en caso de que se actualizara la importancia relativa de las distintas fuentes si esta cambiara en el criterio estándar. Por otro lado, una de las limitaciones del criterio ACMG es su generalidad al aplicarse de la misma forma a distintas patologías. En este sentido, se desarrollaron criterios y guías específicas para algunas patologías (<https://www.clinicalgenome.org/working-groups/sequence-variant-interpretation/>). Si bien no existen guías específicas para la interpretación de variantes asociadas a CC, sí existen para cardiomiopatías [289]. Estas guías y criterios, sin embargo, no aplican masivamente para *NKX2-5*, ya que tiene una asociación limitada a estas patologías. De hecho, sólo 8 VGs se asocian a cardiomiopatías en nuestra compilación. SVI también tiene paneles de expertos que hacen curación de clasificaciones para variantes específicas y potencialmente podrían proponer criterios específicos. Entre estos paneles se encuentra uno de CC (<https://clinicalgenome.org/affiliation/40130>), por lo que no se debe descartar que en un futuro se publique un criterio específico para CC.

Para las 165 VGs clasificadas según el criterio general ACMG, se halló un gran nivel de concordancia con la clasificación dada en la literatura ya sea como benignas o como patogénicas. Todas las variantes clasificadas como patogénicas y como benignas, fueron concordantes. Las variantes en las que se observaron diferencias se tornan relevantes como variantes a estudiar, ya que un ensayo funcional podría determinar cambios de clasificación.

De este trabajo se desprenden 3 grupos de variantes que resultan de interés para realizar futuros estudios:

- En primer lugar, 35 VGs clasificadas como de significado incierto según los criterios de ACMG pero clasificadas como patogénicas en publicaciones y/o bases de datos, que potencialmente podrían ser clasificadas como probablemente patogénicas o probablemente benignas.
- Una variante, p.Lys183Asn, que tiene evidencia conflictiva en la bibliografía pero que según el criterio de ACMG sería probablemente patogénica.

- Por último, 3 VGs (g.173232721G>T, que codifica para p.Pro275Thr; g.173234791C>T, que codifica para p.Ser98Asn; y g.173234960C>G, que codifica para p.Ala42Pro) clasificadas como patogénicas en publicaciones y/o bases de datos y que se clasificarían como probablemente benignas según los criterios de ACMG.

Más allá de lo estudiado en las variantes encontradas y los potenciales mecanismos de patogenicidad considerados, el homeodominio de NKX2-5 es una región altamente conservada con muy poca variabilidad en los vertebrados [47]. En este trabajo, se complementó el estudio de variantes patogénicas con información de variantes benignas y posiblemente benignas, de las cuales no se encontró ninguna que afectara la secuencia de aminoácidos del homeodominio. También se analizaron las frecuencias poblacionales de VGs, para las cuales se observó que ninguna de las variantes que afectan la secuencia del homeodominio tienen una frecuencia mayor a 1/10.000. Esto da un contexto que está ausente en otros trabajos concentrados en analizar solamente las variantes patogénicas [32,40,42,110].

Una de las postulaciones que se puede hacer a partir de estas observaciones es que es de esperar que la mayoría de las variantes génicas que afectan la secuencia de NKX2-5 en la región del homeodominio tengan un efecto en la funcionalidad de la proteína. Esta fue la motivación para realizar el estudio del potencial efecto biológico de todas las variantes de cambio de sentido posibles en el homeodominio. Efectivamente, PolyPhen2 predijo que 248 de las 337 variantes de cambio de sentido serían potencialmente deletéreas, y definimos un mecanismo de patogenicidad asociado para 228 variantes. Para las variantes remanentes, es muy probable que si se incluye el análisis de otros mecanismos de patogenicidad como ser, que se afecte la interacción con otras proteínas, se logre sumar un número de variantes que predigan un mecanismo de patogenicidad asociado. Por otro lado, si bien no se puede descartar la posibilidad de que se generen variantes de secuencia en el homeodominio de NKX2-5 que sean polimórficas, por lo visto en los estudios recopilados en este trabajo es poco probable que una variante nóvel que afecte esta región no tenga un efecto en su funcionalidad. Esta aseveración puede hacerse con mayor certeza cuando se trata

específicamente de variantes en la tercera alfa hélice del homeodominio. Considerando lo observado en la jerarquización de variantes patogénicas (Figura 28 y Tabla 9), se podrían acotar criterios a manera de guía sobre cuál sería el mecanismo más probable de desarrollo de patogenicidad:

- VGs entre los residuos 185 y 194 de la tercer alfa hélice probablemente sean patogénicas por afectar la interacción con el ADN
- VGs entre el final de la segunda alfa hélice (aproximadamente residuo 175) y el residuo 185 de la tercer alfa hélice probablemente sean patogénicas por afectar la acetilación de la lisina 183
- VGs entre el principio de la primer alfa hélice y el final de la segunda probablemente sean patogénicas por un efecto la estabilidad del homeodominio
- VGs antes de la primer alfa hélice probablemente sean patogénicas por afectar el motivo lineal de traslocación nuclear

Aunque estos mecanismos no son 100% seguros, pueden orientar la clasificación de una variante novel y/o direccionar futuros estudios funcionales sobre la patogenicidad de una variante dada de manera predictiva.

Más allá de estudiar la patogénesis de VGs en general, el estudio de variantes patogénicas resulta de especial interés en relación a sus subtipos de CC asociados. Considerando que CIA y CIV son los dos fenotipos más frecuentes de CC [290,291], es de esperar una alta frecuencia para estos subtipos que se asocien a VGs patogénicas, tal como se observa en la Tabla 7. El bloqueo atrioventricular, por otro lado, tiene una incidencia menor, aunque es un fenotipo que suele estar asociado a algunos subtipos de CC como ser la transposición de grandes vasos, y *NKX2-5* es uno de los principales genes asociados a su desarrollo [292]. Por su mayor especificidad, es un fenotipo de interés para centrar estudios dirigidos, ya que no es tan común como CIA o CIV.

Para las variantes patogénicas que afecten específicamente a las regiones desestructuradas, se desprenden principalmente 2 observaciones:

1. Por un lado, las variantes que generan codones *stop* carboxi-terminales al homeodominio (truncando la proteína pero dejando el homeodominio intacto)

parecen estar asociadas a los mismos subtipos de CC que las variantes que truncan la proteína eliminando el homeodominio. Esta observación sugiere que es probable que el extremo carboxi-terminal al homeodominio sea tan crucial para la función de NKX2-5 como el homeodominio en sí mismo.

2. Por otro lado, 10 de las 11 variantes de cambio de sentido asociadas a Tetralogía de Fallot se encuentran fuera del homeodominio, lo que sugiere que este fenotipo específico se encuentra asociado a una funcionalidad de NKX2-5 que no es su unión al ADN. Estas 10 variantes no están agrupadas en ninguna parte específica de la región desestructurada, por lo que no fue posible proponer una funcionalidad en particular que esté asociada a Tetralogía de Fallot.

En relación al homeodominio, se observó que 20 de las 22 variantes de cambio de sentido que se relacionan a bloqueo atrioventricular se ubican en esta región. Este mismo subtipo de CC se encuentra asociado a 26 de las 34 variantes que truncan la proteína. Estas dos observaciones sugieren dos hipótesis:

- A. La pérdida total de función de NKX2-5 se encuentra asociada al desarrollo de bloqueo atrioventricular.
- B. La pérdida de interacción con el ADN es fenotípicamente similar a la pérdida completa de funcionalidad de NKX2-5.

Para el resto de los subtipos analizados (Figura 26: DSAV, FA, CMD, DS, MS) no se observaron patrones claros de distribución. Esto se podría deber a 3 razones:

1. Se requieren más variantes para observar más claramente un patrón de distribución. Para tal fin, es necesario contar con una mayor cantidad de fuentes de información.
2. Su patrón de distribución no está agrupado linealmente a lo largo de la proteína. Esto puede ser porque existen residuos con funciones no conocidas o porque existen patrones complejos que no fueron tenidos en cuenta. En ambos casos, sería posible, en teoría, definir estos patrones con estudios *in silico* más específicos.

3. Estos subtipos de CC no tienen un patrón de distribución específico. Esto puede ser porque el desarrollo de estas CC depende de factores que existen más allá de la proteína NKX2-5, como ser el contexto génico o una variabilidad estocástica en el fenotipo.

De cualquier manera, el análisis realizado a partir de este trabajo cobra singular importancia a los efectos de plantear experimentos en modelos biológicos que pudieran corroborar o no las hipótesis planteadas.

## **5.2. Interacción de NKX2-5 con el ADN y otras proteínas**

Lo observado en el homeodominio, y particularmente en su tercera alfa hélice, remarca la importancia de la interacción de NKX2-5 con el ADN. En este trabajo, esta interacción fue caracterizada mediante 2 abordajes complementarios: análisis de los aminoácidos en contacto con el ADN en las distintas estructuras tridimensionales con las que se trabajó por un lado, y el análisis de los sitios de unión de NKX2-5 en el genoma, por el otro.

Con el modelo estructural obtenido en la sección 4.1., se identificaron 17 residuos cercanos al ADN, 13 de los cuales están en contacto con fosfatos o bases nitrogenadas en las estructuras seleccionadas. Los otros 4 no están en contacto pero son "cercaños" y variantes que introduzcan un aminoácido con una cadena lateral más larga podrían agregar un contacto. Esta identificación permitió, en la sección 4.3., determinar 32 variantes de cambio de sentido que posiblemente afecten la interacción con el ADN. De estas 32 variantes, 7 se clasificaron como patogénicas en la sección 4.2., y 6 de estas 7 se ubican en la tercera hélice alfa del homeodominio. Estos resultados nos permiten proponer que los contactos con el ADN fuera de la tercera alfa hélice no serían esenciales para la función de NKX2-5

Recientemente, los modelos de aprendizaje automático como AlphaFold [77] han permitido predecir la estructura tridimensional de proteínas con precisión muy superior a predictores anteriores, en parte gracias a la gran cantidad de estructuras

confirmadas experimentalmente disponibles en la actualidad. Esta nueva herramienta podría detectar regiones con estructura tridimensional no cristalizada, o incluso ayudar a predecir superficies de interacción entre NKX2-5 y otras proteínas [293,294]. En este trabajo no se utilizó AlphaFold ya que se publicó cuando la sección de análisis estructural estaba completada, pero es una herramienta que potencialmente sería útil para otras proteínas a las cuales se aplique este flujo de trabajo y no tengan estructuras tridimensionales disponibles.

Por su parte, el estudio de los distintos sitios de unión de NKX2-5 al ADN se realizó de dos formas distintas: con secuencias exactas utilizadas en experimentos de transactivación por un lado y con matrices de la base de datos HOCOMOCO, por el otro. La lista de secuencias exactas permite asegurar unas pocas secuencias simples, pero ignora secuencias no consideradas explícitamente. Las matrices, por su parte, permiten buscar patrones más complejos en secuencias, aunque es más difícil asegurar que todas las secuencias halladas sean correctas.

En este trabajo, se hizo una selección orientada para minimizar falsos positivos, aplicando un valor de corte estricto para la matriz de pesos y aprovechando la abundancia de secuencias confirmadas experimentalmente. Una vez definidos los sitios de unión al ADN, se estudiaron genes que podrían ser activados o inhibidos por NKX2-5, para lo cual fue necesario definir genes confirmados experimentalmente y definir que estuvieran ubicados a menos de una cierta distancia de un sitio de unión confirmado de este factor de transcripción.

Para la búsqueda de sitios de unión en el genoma se seleccionaron dos estudios que proveyeron dos ejes de análisis complementarios. En ambos casos se trata de estudios de ChIP-seq en cardiomiocitos recientemente diferenciados en los que se busca confirmar genes transactivados por NKX2-5 con ensayos de RNA-seq. El trabajo en ratón [206] estudia cardiomiocitos *in vivo* y el trabajo en humano [207] utiliza líneas celulares *in vitro*. Ambos estudios tenían sus ventajas, por lo que en parte del trabajo utilizamos ambos estudios en paralelo para la búsqueda de sitios de unión de NKX2-5.

Los beneficios y limitaciones del uso de ratones o líneas celulares de humanos son conocidos con relativa profundidad [295]. Por un lado, el trabajo en cardiomiocitos de ratón *in vivo* aporta información acerca de la función del gen en un organismo completo, con sus efectos a nivel de órganos. Estos resultados también son de utilidad para estudiar efectos en la función de *NKX2-5* asociados a su estructura tridimensional, ya que el homeodominio tiene 100% de identidad de secuencia proteica entre humano y ratón. Por otro lado, el trabajo en cardiomiocitos humanos *in vitro* provee un mejor contexto génico para el estudio de CC, por lo que representa mejor las interacciones entre los distintos genes. En particular, para los estudios de genes cercanos a sitios de unión de *NKX2-5*, se tomaron en cuenta sólo los resultados en células humanas por considerar que aportan información más relevante al estudio de CC.

Para los análisis y la selección de distancias de genes cercanos se utilizaron 2 métodos de selección de genes y 4 distancias. Los métodos de selección fueron el método de “dos genes más cercanos” de GREAT y un criterio personalizado de selección de todos los genes hasta una distancia dada. El criterio de GREAT busca los genes más cercanos (1 río arriba y 1 río abajo) hasta cierta distancia límite. Esto prioriza genes más cercanos, por lo que es una buena herramienta para buscar a distancias límites altas (mayores a 100.000 pb) y permite seleccionar los dos genes con mayor probabilidad de ser transactivados por *NKX2-5* al unirse a un sitio dado. La selección de todos los genes, por el contrario, puede incluir varios genes que muy probablemente no sean transactivados por *NKX2-5*, y puede cubrir una gran porción del genoma si la distancia máxima es muy alta y los sitios de unión están muy dispersos entre sí. Aún así, la confirmación experimental por ensayos de RNA-seq con expresión diferencial permitió que se trabajara con todos los genes hasta la distancia máxima. Esta confirmación permite minimizar los falsos positivos al seleccionar muchos genes, mientras que el criterio de GREAT podría ignorar genes transactivados por *NKX2-5* que tuvieran otro gen entre su promotor y el sitio de unión de *NKX2-5* y que en rigor no sea el gen blanco. Por otro lado, las distintas distancias de búsqueda se seleccionaron de tal manera de analizar distintos genes de acuerdo a la forma en la que probablemente sean transactivados por *NKX2-5*. La búsqueda de hasta 1500 pb implicaría sitios de unión en el promotor de los genes, mientras que las búsquedas de

hasta 100.000 pb, 500.000 pb y 1.000.000 pb permiten ver transactivación de largo alcance (regulación como *enhancer* o *silencer*). Se trabajó específicamente con una distancia de 100.000 pb porque a distancias cortas (1500 pb) se encontraron muy pocos genes y a distancias largas (500.000 y 1.000.000 pb) se incluiría una proporción muy grande del genoma. Al analizar la distribución de distancias entre sitios de unión de NKX2-5 y genes confirmados experimentalmente, se observó que éstos se encuentran en mayor abundancia en el rango entre -50.000 pb y 50.000 pb (Figuras 31 y Anexo II. Figura S1). Esto sugiere que la búsqueda hasta 100.000 pb de distancia es del orden de magnitud correcto y es consistente con las distancias de transactivación a largo alcance de NKX2-5 [170-172].

Al realizar el análisis, se observó que algunas de las secuencias de unión parecen tener más genes con expresión aumentada por NKX2-5 mientras que otras, más genes que disminuyen su expresión en presencia de NKX2-5 (Figura 30). Sin embargo, para ninguna se observó una proporción mayor a 65%. Esto es consistente con la proporción del total de genes con expresión aumentada o disminuida por NKX2-5 obtenidas de los estudios de RNA-seq en células humanas y de ratón (Figura 29) [206,207]. Observando otros factores de transcripción importantes en el desarrollo del corazón con ensayos similares en ratón, *Gata4* aumenta la expresión en 63% de genes transactivados y la disminuye en el 37% [145], mientras que para *Tbx5* se observa un 51% de genes con expresión aumentada y 49% disminuida [225]. En resumen, las proporciones de genes río abajo activados o inhibidos por NKX2-5 parecen ser consistentes con estos factores de transcripción, y en todos los casos no parecen tener una función principal de activación o de inhibición.

Teniendo en cuenta la ausencia de patrones claros en los datos analizados en la sección 4.4., podemos postular que la activación o inhibición de genes río abajo por NKX2-5 no estaría determinada de manera simple por la secuencia de unión ni su distancia a los genes transactivados. La disponibilidad reciente de ensayos comparables en *Tbx5* y *Gata4* en ratón [145,225] abren la posibilidad de aplicar el flujo de trabajo de este estudio en esos genes y comparar sus resultados para confirmar si

la ausencia de patrones de secuencia de unión y su distancia a genes transactivados es extensible a estos genes.

Para entender la forma en la que *NKX2-5* activa o inhibe genes, y extender el análisis de posibles mecanismos de patogénesis estudiados en las secciones 4.2. y 4.3., es necesario estudiar otros factores que pudieran afectar esta función. En particular, el enfoque de nuestro trabajo fue estudiar la interacción de *NKX2-5* con otros factores de transcripción.

Por un lado, con el estudio de estructuras tridimensionales de *NKX2-5* se hizo un análisis de potenciales interacciones y de VGs patogénicas que pudieran afectar estas interacciones. Este estudio arrojó 5 residuos potencialmente involucrados en la interacción entre *NKX2-5* y *TBX5* o *MEF2B*. Entre todas las VGs de cambio de sentido posibles en estos residuos, 11 afectan la interacción con *TBX5* y 2, la interacción con *MEF2B*, aunque ninguna de estas variantes posee una clasificación como patogénica. Esto nos impidió estudiar fenotipos asociados a la interacción entre proteínas. Sin embargo, nuestro análisis de variantes posibles permite hacer un análisis más rápido si alguna de estas variantes apareciera en futuros pacientes. Las estructuras tridimensionales seleccionadas para el análisis de estas interacciones consisten en heterodímeros con poco contacto entre cada cadena aminoacídica en *cis*, por lo que es posible que la superficie de interacción esté siendo subestimada, y que un mejor conocimiento de la interacción de *NKX2-5* con otros factores de transcripción sea la mejor forma de profundizar en el entendimiento de estas interacciones. También es posible que las interacciones estén mediadas por las regiones desestructuradas de alguna de las proteínas [146,183], por otros factores de transcripción participando de la interacción, como *GATA4* [296,297], o que actúen en *trans*. El trabajo que determina la estructura tridimensional de *MEF2B* con *NKX2-5* propone la interacción en *trans*, aunque no logra una confirmación *in vitro* de la interacción [178].

La interacción de *NKX2-5* con otras macromoléculas fue evaluada indirectamente a través del análisis de sus motivos lineales y de residuos con modificaciones post-traduccionales. Esto es así porque tanto los motivos lineales como las modificaciones post-traduccionales conllevan el reconocimiento de *NKX2-5* por

otros elementos (principalmente otras proteínas) para llevar a cabo funciones específicas. En este trabajo, se encontraron 4 motivos lineales con confirmación funcional, además de un residuo con una modificación post-traducciona, para los cuales se hallaron 12 VGs patogénicas de cambio de sentido (Tabla 6). Se estima, por lo tanto, que su efecto deletéreo estaría relacionado con la función biológica de estas regiones. En particular, resultó de interés el residuo Lys183, que se discutió con anterioridad, en relación a la posibilidad de explicar el mecanismo de 4 VGs patogénicas.

En la sección 4.5.3. se definieron 21 genes con posibles interacciones con *NKX2-5*: 2 genes que codifican para factores de transcripción con interacción confirmada específicamente con *NKX2-5* (*TBX5* y *GATA4*), 4 genes que codifican para proteínas en contacto con *NKX2-5* provenientes de cristales de PDB (*TBX5* y *MEF2A/B/C*), 5 genes que codifican para factores de transcripción predichos con el análisis por MEME-ChIP (*MAF*, *MEIS1*, *TAL1*, *TBX20* y *TGIF1*) y 11 genes que resultaron de interés a partir de los análisis de GO-terms (sección 4.5.2.) por ser activados o inhibidos por *NKX2-5* (*FHL2*, *ISL1*, *LIMCH1*, *NOS3*, *NPPA*, *SCN5A*, *SERPINE1*, *TBX18*, *TGFB3*, *UCHL1* y *ZIC3*)

Es sabido que *NKX2-5*, *TBX5* y *GATA4* son tres genes centrales en el desarrollo y diferenciación del corazón humano [32,298,299] y sus interacciones están también muy documentadas [144,146,177,213,300]. Se conoce la asociación con cardiopatías de cada uno de estos factores de transcripción y tanto *TBX5* como *GATA4* tienen variantes patogénicas asociadas a distintos tipos de cardiopatías [32,220,241-243]. Por lo tanto, ambos serían candidatos ideales para analizar las implicancias biológicas de todas las posibles VGs utilizando este mismo flujo de trabajo.

Por su parte, *MEF2A/B/C* son 3 genes emparentados con función en numerosas vías de desarrollo, incluyendo al corazón, [232,233,301] y se conocen cristales de sus proteínas con *NKX2-5* [178]. Esto los hace potencialmente interesantes para estudiar en conjunto con *NKX2-5*, aunque su información de asociación a CC y de interacción con *NKX2-5* no es tan extensiva como la de *GATA4* y *TBX5*.

A pesar de que *GATA4*, *TBX5* y *MEF2* serían los genes cuya interacción con *NKX2-5* sería la más fácil de estudiar, se incluyó el análisis de potenciales factores relacionados (sección 4.5.1.) mediante MEME-ChIP. La racionalidad de este abordaje se basó en que con la selección mediante MEME-ChIP incluiríamos aquellos factores de transcripción que con mayor probabilidad estén interactuando con *NKX2-5* en el contexto de los ensayos de ChIP-seq y RNA-seq cuyos resultados analizamos.

De esta forma, se seleccionó a *GATA4* y a otros 5 factores de transcripción, de los cuales 4 tienen asociaciones al desarrollo tanto de cardiopatías como del corazón en general. Específicamente:

- ***TBX20***. Codifica para un factor de transcripción clave en el desarrollo y reparación del corazón [243]. Es un gen evolutiva- y funcionalmente cercano a *TBX5*. Su homólogo en ratón, *Tbx20*, tiene interacción confirmada con *Nkx2-5* [302].
- ***TGIF1***. Codifica para un factor de transcripción importante en la diferenciación de la cresta cardíaca [245].
- ***MAF***. Codifica para un factor de transcripción de función muy extendida, que suele heterodimerizar para definir activación/inhibición de otros TF y está asociado con expresión en ojos/retina y músculo esquelético, entre otros [234,303]. No se encontró una asociación clara de *MAF* con el desarrollo del corazón ni con otras funciones conocidas de *NKX2-5* en el desarrollo (e.g. tiroides, bazo). Es posible que esto se deba a que *NKX2-5* inhiba la función y/o unión de *MAF* o, en caso de tratarse de un falso positivo, que exista otro factor con sitios de unión similares y se haya identificado *MAF* incorrectamente.
- ***MEIS1***. Codifica para un factor de transcripción con función en la diferenciación celular, en particular hematopoyesis y limita la proliferación de cardiomiocitos (para su regeneración) [235,236]. Fue estudiado en ratón en el paper de ChIP-seq Dupays y col., 2015 [206], donde parece tener superposición de sitios de unión con *NKX2-5*.

- **TAL1.** Codifica para un factor de transcripción esencial en la hematopoyesis, particularmente en la diferenciación de células del mesodermo y precursoras del endotelio [237]. En organismos modelo, se observó que interviene en la diferenciación de células del endotelio, suprimiendo la cardiomiogénesis y participando en la migración celular [238-240].

Estos factores de transcripción fueron identificados buscando secuencias sobrerrepresentadas en **todos** los picos de ChIP-seq y asignándolos a sitios de unión de factores de transcripción conocidos. Aún así, para algunos se hallaron muy pocos sitios de unión y, en algunos casos, ninguno (Tabla 13). Las razones de estas observaciones podrían ser tres:

1. La selección de factores de transcripción con MEME-ChIP se hizo sobre todos los picos de ChIP-seq, a diferencia de la búsqueda que se realizó cerca de sitios de unión de NKX2-5 (a menos de 100.000 pb de genes con expresión diferencial en ensayos de RNA-seq). Por lo tanto, es posible que algunos de los factores de transcripción seleccionados se unan a sitios en el ADN cuyos genes cercanos no sean transactivados por NKX2-5. Esto puede ser por dos motivos:
  - a. El factor de transcripción seleccionado inhibe la función de transactivación de NKX2-5.
  - b. NKX2-5 se une a estos factores de transcripción en vez de a un sitio de unión propio.
2. MEME-ChIP busca secuencias sobrerrepresentadas y luego asigna sitios de unión de factores de transcripción a estas secuencias. Esta estrategia es diferente de la búsqueda hecha en este trabajo.
3. El puntaje estricto de corte puede estar ignorando sitios de unión de bajo puntaje, y estos sitios de bajo puntaje pueden ser los que se encuentran cerca de NKX2-5.

De cualquier manera, este flujo de trabajo se diseñó con el objetivo de estudiar los sitios de unión de cualquier factor de transcripción, por lo que se podría utilizar para confirmar sitios de unión de alguno de los genes seleccionados y tener un análisis más robusto de su potencial interacción. Esta profundidad de análisis escapa al alcance de

este trabajo ya que implicaría encontrar o realizar estudios de ChIP-seq/RNA-seq con expresión diferencial para distintos genes en las mismas condiciones experimentales. En efecto, este análisis se estudia entre *Nkx2-5* y *Meis1* en Dupays y col., 2015 [206], pero no se profundizó para este trabajo de tesis, porque nos centramos en los resultados en humano. Sin embargo, y a partir del flujo de trabajo desarrollado en este trabajo, el análisis de los resultados puede ser realizado con mayor facilidad y celeridad en estudios futuros.

Tras la identificación y selección de sitios de unión, el análisis de genes cercanos se hizo principalmente por medio del estudio de sus GO-terms asociados, ya que permite estimar patrones de vías de acción en grandes listas de genes. De este análisis, se identificaron 11 genes de interés potencialmente relacionados al desarrollo de patogenicidad mediada por *NKX2-5*, los cuales permiten definir estudios futuros más específicos.

Los GO-terms mencionados se clasificaron en 7 categorías diferentes, todas ellas relacionadas directa o indirectamente con el desarrollo del sistema circulatorio:

- Sistema circulatorio. *NKX2-5* está asociado con su desarrollo embrionario y nuestro trabajo se concentró en su relación con cardiopatías.
- Función y desarrollo muscular. Estrechamente relacionado con la función del corazón.
- Desarrollo y diferenciación de tejidos y/u órganos. Función general relevante para un factor de transcripción de expresión temprana en el desarrollo.
- Comunicación celular. Función general consistente con interacciones entre distintas células en desarrollo y diferenciación, así como funciones de músculos y neuronas del sistema circulatorio.
- Migración y locomoción celular. Función general correspondiente con la migración celular de tejido que forma parte del corazón, por ejemplo las células de la cresta neural, y con motilidad celular, asociada con heterotaxia.
- Movimientos intracelulares. Función general, que en nuestros resultados fue principalmente asociada a la separación de cromosomas, lo cual es consistente con el desarrollo y diferenciación de tejidos y/u órganos.

- Sistema nervioso. Función específica tangencialmente relacionada con el sistema circulatorio. En este trabajo en particular se observó una relación con el nodo sinoatrial.

Más específicamente, los GO-terms sobrerrepresentados en la lista de 356 genes con expresión diferencial en ensayos de RNA-seq (Figura 33) fueron consistentes con la función de *NKX2-5*. A notar:

1. Los GO-terms asociados a genes para los cuales *NKX2-5* aumenta su expresión, se relacionan a la migración y motilidad celular. Estos hallazgos serían consistentes con la migración de células del mesodermo a tiroides y corazón (George y col., 2020).
2. Los GO-terms asociados a genes cuya expresión se ve disminuida por *NKX2-5*, se relacionan con el desarrollo vascular. Estas observaciones indicarían el silenciamiento de genes de diferenciación vascular para fomentar específicamente la diferenciación cardíaca.

Al realizar un análisis de los genes con expresión diferencial cerca de sitios de unión de *NKX2-5* y de otros factores de transcripción, se obtuvieron observaciones relevantes.

- Genes cercanos de múltiples sitios de unión de *NKX2-5*: Estos genes tuvieron sobrerrepresentación de GO-terms similares a los de todos los genes transactivados por *NKX2-5* (Anexo II. Figura S7). Esto es consistente considerando que 303 de los 356 poseen otro sitio de unión de *NKX2-5* cercano (Tabla 13), por lo que es posible que no haya diferencia funcional asociada a la presencia de múltiples sitios de unión de *NKX2-5*. Esta observación sugiere que *NKX2-5* suele actuar por unión a varios sitios (al mismo tiempo o alternativamente), lo cual potencialmente aporta relevancia a las estructuras tridimensionales del homodímero de *NKX2-5* unido al ADN [146,176].
- Genes cerca de sitios de unión a *GATA4*: Se identificaron pocos genes, a pesar de que se sabe que *GATA4* es un cofactor de *NKX2-5*. Específicamente, hay evidencia experimental de que *GATA4* coregula con *NKX2-5* para transactivar

el gen del factor natriurético atrial (*NPPA*) [213], pero no se halló un sitio de unión de GATA4 en la región regulatoria de *NPPA*. Es posible que se deba a que el criterio de búsqueda del sitio de unión fue muy restrictivo, ya sea porque el sitio de unión de GATA4 se encuentra a mayor distancia de sitios de NKX2-5 que la utilizada en este trabajo, o que el corte utilizado para la matriz de puntajes haya sido muy alto para GATA4. Alternativamente, es posible que GATA4 actúe como co-factor a través de una unión a NKX2-5 sin mediar unión al ADN.

- Genes cerca de sitios de unión a TBX20: Se identificaron 12 genes, de los cuales se destacan 5: *FHL2*, *NOS3*, *NPPA*, *TGFB3* y *ZIC3*. Estos genes poseen sitios de unión para múltiples de los factores de transcripción seleccionados, y pueden representar una transactivación mediada por un super-enhancer.
- Genes cerca de sitios de unión a TGIF1: Representan la mitad de los genes transactivados por NKX2-5. Esto se refleja en una gran cantidad de GO-terms significativos, asociados principalmente al desarrollo cardíaco, en particular entre los genes con expresión disminuida por NKX2-5. De destacar es que esta relación actualmente no está caracterizada en mucha profundidad.
- Genes cerca de sitios de unión a MAF: A pesar de que se observaron 100 genes con sitios cercanos de unión a este factor de transcripción, la cantidad de GO-terms con sobrerrepresentación significativa son escasos. Si bien podría tratarse de un falso positivo, no se puede descartar que tenga una interacción compleja para la función de transactivación con NKX2-5.
- Genes cerca de sitios de unión a MEIS1: No se halló que hubiera algún GO-term con sobrerrepresentación significativa, pero se identificaron los genes *ISL1* y *TBX18* como genes de interés por su asociación al desarrollo del nodo sinoatrial.
- Genes cerca de sitios de unión a TAL1: No se hallaron genes transactivados por NKX2-5 cerca de sitios de unión a *TAL1* y *NKX2-5*. Sin embargo, y considerando su función de supresión de diferenciación de cardiomiocitos, una explicación posible es que *TAL1* inhiba la función de transactivación de NKX2-5.

Asimismo, los 11 genes de interés seleccionados al final de la sección 4.5.2. se pueden agrupar de acuerdo a si son regulados en forma positiva (*FHL2*, *NOS3*, *NPPA*, *SCN5A* y *ZIC3*) o negativa (*ISL1*, *LIMCH1*, *SERPINE1*, *TBX18*, *TGFB3* y *UCHL1*) por *NKX2-5*. Su funcionalidad se resume a continuación.

#### Genes regulados en forma positiva por *NKX2-5*

- *NPPA* -también conocido como *ANF*- codifica para el factor natriurético atrial, el cual se expresa en el corazón (principalmente las aurículas) y su principal función es regular la presión sanguínea en las aurículas [304]. Se hallaron variantes de secuencia en este gen que se asocian a fibrilación auricular [252,253] e hipertensión [305]. Como se aludió anteriormente, hay evidencia experimental de que *NKX2-5* transactiva este gen [306], particularmente en conjunto con *GATA4* [213], y su promotor es ampliamente utilizado en ensayos de transactivación por *NKX2-5* [70,186-205]. Dada esta observación y que no hallamos sitios de unión de *GATA4* en este trabajo, es muy probable que el criterio de búsqueda del sitio de unión de *GATA4* haya sido muy restrictivo. Idealmente, este conflicto debería solucionarse con un análisis específico de *GATA4* utilizando el flujo de trabajo propuesto en la sección 4.6.
- *ZIC3* codifica para un factor de transcripción que determina lateralidad y asimetría en estadios muy tempranos [307]. Se hallaron variantes de secuencia en este gen asociadas principalmente a heterotaxia, pero también a defectos en órganos definidos por su asimetría, como el corazón [260,262]. Los subtipos asociados son variados, pero entre ellos se encuentran la transposición de grandes vasos, arco aórtico interrumpido, comunicación interauricular, comunicación interventricular y conexiones venosas anormales [260].
- *SCN5A* codifica para parte de un canal de sodio con un rol en la función normal del corazón, y variantes de secuencia en este gen están asociadas a muerte súbita infantil y fibrilación atrial [259,308-311].
- *FHL2* codifica para un factor de transcripción involucrado en varias vías de señalización involucradas en regulación tejido-específica [247]. Tiene un rol en diferenciación celular, apoptosis, ciclo celular y ensamblado de la matriz

extracelular, entre otros [312]. Variantes de secuencia en este gen están asociadas a cardiomiopatías dilatadas [313]. Es sabido que su expresión reduce los efectos de hipertrofia cardíaca inducida por endotelina-1 o fenilefrina [247,314], lo cual sugiere una función de protección relevante en el contexto del desarrollo de CC.

- *NOS3* codifica para una enzima de síntesis de monóxido de nitrógeno (NO), que actúa como vasodilatador, entre otras funciones [315-317]. Se observaron variantes de secuencia en este gen asociadas con hipertensión en el embarazo [318], problemas en la arteria coronaria [249-251] y cardiomiopatías dilatadas [248].

#### Genes regulados en forma negativa por *NKX2-5*

- *TGFB3* codifica para una de las tres isoformas de “*Transforming growth factor-beta*”, y tiene un rol importante en la regulación de crecimiento, diferenciación y reparación celular en una gran variedad de tejidos [319]. Su activación inadecuada se encuentra asociada con el desarrollo de fibrosis en los tejidos en los que sucede [273,319], lo cual es relevante para un gen que es inhibido por *NKX2-5*, ya que fallas en la función de *NKX2-5* podrían causar su sobreexpresión en tejido cardíaco y producir fibrosis.
- *LIMCH1* codifica para una proteína que regula la actividad de una miosina no muscular involucrada en la supresión de la migración celular [268]. No se encontró un estudio que confirme o refute su asociación con cardiopatías y/o el desarrollo del corazón. Como es un gen silenciado por *NKX2-5*, es probable que su expresión produzca problemas cardíacos, si bien no hay registros de pacientes con alteraciones activantes (o duplicaciones) de este gen. Es posible que aún no se haya registrado uno de estos pacientes y/o que sea un fenotipo letal embrionario.
- *UCHL1* codifica para una enzima deubiquitinadora con expresión en neuronas. Si bien se la encuentra asociada con defectos del sistema nervioso central [320,321], se observó que se expresa en cardiomiocitos luego de infartos de miocardio [276]. En particular, en ratones, se observó que su inhibición

disminuye síntomas de fibrilación auricular [279]. Es consistente con el hecho que es un gen inhibido por NKX2-5.

- *SERPINE1* codifica para una proteína que participa en la disolución de coágulos [322,323], y su patogénesis está relacionada a sangrado y falta de coagulación [324,325]. En particular, se ha visto que interviene en migración celular, particularmente neuronal [326,327]. Esto es consistente con una relevancia en el desarrollo del corazón, ya que parte de sus células provienen de migración de la cresta neural [328,329] y el correcto desarrollo de neuronas es crucial para evitar fenotipos de conducción. Es posible que NKX2-5 inhiba este gen para delimitar la migración celular y/o el desarrollo neuronal.
- *ISL1* codifica para un factor de transcripción clave en el desarrollo cardíaco, particularmente del nodo sinoatrial [330,331]. Se observaron variantes de secuencia en este gen asociadas a cardiopatías, particularmente transposición de grandes arterias, ductus arterioso persistente y comunicación intraventricular [263,266].
- *TBX18* codifica para un factor de transcripción asociado con la diferenciación del nodo sinoatrial [332] y está involucrado en el desarrollo del corazón en ratón [333,334]. Se sugirió que variantes de secuencia en este gen podrían estar potencialmente asociadas a comunicación interventricular [272].

Es interesante resaltar que tanto *TBX5* como *GATA4* se asocian con varios de los fenotipos cardíacos con los que también se asocia *NKX2-5*, lo que refleja la relación de estos factores de transcripción con el desarrollo del corazón. Sin embargo, una observación importante es el hecho de que sólo *TBX5* se asocia a bloqueo atrioventricular, fenotipo que presentan varios de los pacientes que poseen VGs patogénicas en *NKX2-5* (Tabla 7). Esta observación sugiere una implicancia de *TBX5* y *NKX2-5* en el desarrollo de los sistemas de conducción del corazón que no ocurre con *GATA4*.

Por su parte, entre los genes seleccionados en este trabajo, se encontraron dos, *ISL1* y *TBX18*, asociados con el desarrollo del nodo sinoatrial, y que podrían tener relación con patologías de conducción. Ambos disminuyen su expresión en presencia

de *NKX2-5*, pero para ninguno se reportaron aún variantes asociadas a bloqueo atrioventricular.

La CIA y CIV son dos de los fenotipos más comunes de CC [290,291]. Por su parte, la fibrilación auricular y las cardiomiopatías dilatadas (CMD) también son fenotipos que se presentan con alta frecuencia (Tabla 14). En este contexto, no llama la atención la asociación de la mayoría de los genes seleccionados a este tipo de cardiopatías. De manera similar, la fibrilación atrial afecta a 1-4% de la población general [335-337], y su incidencia es mayor si se acompaña con cardiomiopatía dilatada [338,339]. Es posible, por lo tanto, que estos fenotipos no sean representativos de vías funcionales específicas de los genes seleccionados, sino subtipos comunes debidos a fallas en la organogénesis general durante el desarrollo del corazón.

Otro fenotipo específico de interés es el bloqueo atrioventricular. Este fenotipo es uno de los más comúnmente asociados a VGs en *NKX2-5*, específicamente para variantes con efecto en toda la proteína (Tabla 7). Teniendo en cuenta que no es un fenotipo tan común, es posible que alguna de las funciones de *NKX2-5* sea específicamente relevante para este fenotipo. Esto se extiende a *TBX5*, que es el único de los genes seleccionados en los que se hallaron VGs asociadas al fenotipo, por lo que su interacción podría ser el factor que determine su desarrollo. Estudios específicos de la interacción entre *TBX5* y *NKX2-5* podrían contribuir al desarrollo de este fenotipo particularmente más que otros.

Por último, para *ZIC3* observamos una fuerte asociación a heterotaxia. Una desregulación de *ZIC3* en regiones específicas del corazón podría explicar el desarrollo de CC que se asemejan a efectos locales de heterotaxia. Este efecto sería interesante de estudiar mediante un *knock-out* condicional dirigido a regiones específicas del corazón, y de esta forma entender mejor la distribución de subtipos de CC que puede conllevar.

Del análisis realizado, llama la atención la ausencia de genes asociados al desarrollo de fenotipos tiroideos. Sin embargo se debe tener en cuenta, que el criterio de selección aplicado estuvo centrado en resultados de experimentos diseñados para

estudiar la función de *NKX2-5* en el desarrollo del corazón. Aunque *NKX2-5* se expresa en el desarrollo de tiroides en períodos similares a los de corazón [340,341], los datos de ChIP-seq analizados en este trabajo se basan en estudios de cardiomiocitos. También es posible que algunos de los genes que interactúen con *NKX2-5* en el desarrollo del corazón se asocien al desarrollo de disgénesis de tiroides de formas no documentadas. Estudios específicos de esta interacción aportarían información clave para entender las funciones menos estudiadas de *NKX2-5* en el desarrollo embrionario.

Aunque se encontraron muchas asociaciones entre genes y fenotipos particulares, es importante recordar las limitaciones inherentes al estudio de pacientes. Es posible que algunos fenotipos sean registrados de formas no específicas (e.g. el bloqueo atrioventricular puede ser registrado como “problemas de conducción”), que algunos fenotipos no se registren por ser opacados por otros (e.g. un diagnóstico de muerte súbita se puede hacer sin investigar si fue causada por una malformación no detectada) o que algunos de los genes no hayan sido estudiados en relación a paneles de CC, entre otras posibilidades, y que por eso no se encuentren algunas asociaciones. En estos casos, es crucial remarcar la importancia de poder elaborar y mantener estudios propios que permitan realizar análisis más robustos.

Como ya se mencionara, con el análisis realizado no se hallaron VGs patogénicas que afecten específicamente regiones de interacción con otras proteínas. Esta limitación puede superarse, a futuro, de dos formas. Por un lado, que se encuentren pacientes con CC y alguna de las 13 VGs sinónimas que afectan la interacción con *TBX5* o *MEF2*. Por otro lado, que se aplique el flujo de trabajo de la sección 4.6. a *TBX5* o *MEF2* y se analicen VGs patogénicas en residuos de estos genes que interactúen con *NKX2-5*. Esta segunda opción es la más accesible, y la que más probablemente encuentre información de utilidad, aunque aún así excede al alcance de este trabajo.

A futuro, esperamos que nuestros resultados faciliten el estudio de variantes de interés actuales y futuras, colaboren a interpretar nueva información estructural de *NKX2-5* y complementen estudios en otros genes asociados a *NKX2-5*. Este último punto es de particular interés, ya que este mismo flujo de trabajo aplicado a otro gen

asociado a *NKX2-5* permitiría la generación de resultados complementarios a los obtenidos en este trabajo.

## **6. Conclusiones**

Esta tesis se realizó con el objetivo de obtener un flujo de trabajo con una función específica y aplicarlo a *NKX2-5*. Su aplicación arrojó una serie de resultados que permitieron evaluar la potencial patogenicidad de VGs en *NKX2-5*, así como proponer mecanismos por los cuales este gen estaría involucrado con distintos subtipos de CC.

Tanto el modelo estructural de *NKX2-5*, como la recopilación de VGs conocidas y la evaluación de todas las VGs posibles en el homeodominio que se analizaron fueron de utilidad para profundizar en la comprensión de las distintas regiones del gen. Los motivos lineales, las regiones desestructuradas, el homeodominio y, específicamente, su tercera alfa hélice, se analizaron con una profundidad mayor a la existente en la bibliografía. Esta profundidad permitió, en particular:

- Describir la concentración de VGs patogénicas y ausencia de VGs benignas en el homeodominio de *NKX2-5*, observando que las VGs patogénicas están especialmente concentradas en su tercera alfa hélice.
- Describir una región (entre los residuos 175 y 185) que probablemente tenga relevancia funcional, proponiendo un mecanismo para su patogenicidad (acetilación de Lys183).
- Definir potenciales efectos funcionales para 228 de las 337 VGs de cambio de sentido posibles en el homeodominio de *NKX2-5*.

También se allanó el camino para el estudio de variantes futuras, las cuales podrían, por un lado ser estudiadas con mayor facilidad con la información obtenida en este trabajo y, por el otro, agregar información que contribuya con nuestras predicciones.

La recopilación de sitios de unión de *NKX2-5* y descripción de genes que potencialmente interactúen y que fuesen transactivados, por su parte, permiten sugerir vías funcionales para el estudio de subtipos específicos de CC, como ser *ZIC3* para fenotipos relacionados a heterotaxia y *TBX5* para fenotipos relacionados a bloqueo atrioventricular. Estos resultados se encuentran más alejados de interpretaciones

funcionales de relevancia clínica, pero representan un primer paso en la comprensión de la asociación de *NKX2-5* con subtipos específicos de CC.

Por último, se obtuvo un flujo de trabajo con la potencialidad de ser aplicado a otros genes, y se propuso una lista de genes prioritarios para ser estudiados de esta manera. Este flujo de trabajo redundaría en poder confirmar o refutar las hipótesis propuestas, que no pueden ser confirmadas con absoluta certeza sólo con el estudio de *NKX2-5*. De esta forma, sentamos las bases para un modelo computacional predictivo de la patogenicidad potencial de VGs en *NKX2-5* no caracterizadas funcionalmente y direccionamos futuros estudios hacia un uso de las redes de interacciones de *NKX2-5* con otros genes en la comprensión del desarrollo de CC.

## **7. Bibliografía**

1. Pierpont, M. E., Brueckner, M., Chung, W. K., Garg, V., Lacro, R. V., McGuire, A. L., Mital, S., Priest, J. R., Pu, W. T., Roberts, A., Ware, S. M., Gelb, B. D., Russell, M. W., & American Heart Association Council on Cardiovascular Disease in the Young; Council on Cardiovascular and Stroke Nursing; and Council on Genomic and Precision Medicine (2018). Genetic Basis for Congenital Heart Disease: Revisited: A Scientific Statement From the American Heart Association. *Circulation*, 138(21), e653–e711. <https://doi.org/10.1161/CIR.0000000000000606>
2. Liu, Y., Chen, S., Zühlke, L., Black, G. C., Choy, M. K., Li, N., & Keavney, B. D. (2019). Global birth prevalence of congenital heart defects 1970-2017: updated systematic review and meta-analysis of 260 studies. *International journal of epidemiology*, 48(2), 455–463. <https://doi.org/10.1093/ije/dyz009>
3. Singh, Y., & Lakshminrusimha, S. (2021). Perinatal Cardiovascular Physiology and Recognition of Critical Congenital Heart Defects. *Clinics in perinatology*, 48(3), 573–594. <https://doi.org/10.1016/j.clp.2021.05.008>
4. Sun H. Y. (2021). Prenatal diagnosis of congenital heart defects: echocardiography. *Translational pediatrics*, 10(8), 2210–2224. <https://doi.org/10.21037/tp-20-164>
5. Bolundut, A. C., Lazea, C., & Mihu, C. M. (2023). Genetic Alterations of Transcription Factors and Signaling Molecules Involved in the Development of Congenital Heart Defects-A Narrative Review. *Children (Basel, Switzerland)*, 10(5), 812. <https://doi.org/10.3390/children10050812>
6. Azamian, M., & Lalani, S. R. (2016). Cytogenomic Aberrations in Congenital Cardiovascular Malformations. *Molecular syndromology*, 7(2), 51–61. <https://doi.org/10.1159/000445788>
7. Suluba, E., Shuwei, L., Xia, Q. & Mwanga, A. (2020). Congenital heart diseases: genetics, non-inherited risk factors, and signaling pathways. *Egyptian Journal of Medical Human Genetics*. 21. 11. <https://doi.org/10.1186/s43042-020-0050-1>
8. Shabana, N. A., Shahid, S. U., & Irfan, U. (2020). Genetic Contribution to Congenital Heart Disease (CHD). *Pediatric cardiology*, 41(1), 12–23. <https://doi.org/10.1007/s00246-019-02271-4>

9. Rohit, M., & Shrivastava, S. (2018). Acyanotic and Cyanotic Congenital Heart Diseases. Indian journal of pediatrics, 85(6), 454–460.  
<https://doi.org/10.1007/s12098-017-2454-6>
10. Shanley, T.P.; Wheeler, D.S. & Wong, H.R (2007). Pediatric critical care medicine: basic science and clinical evidence. Berlin: Springer. p. 665-667.
11. Pierpont, M.E.M. & Moller, J.H. (1987). The genetics of cardiovascular disease. Martinus Nijhoff Publishing, Boston.
12. Garg V. (2006). Insights into the genetic basis of congenital heart disease. Cellular and molecular life sciences : CMLS, 63(10), 1141–1148.  
<https://doi.org/10.1007/s00018-005-5532-2>
13. Martinez, J.M., Puerto, B. & Gomez, O. (2002). Fetal congenital cardiac disease: conotruncal anomalies. Ciencia Ginecologica 6:54-61.
14. Zaidi S, Brueckner M. Genetics and Genomics of Congenital Heart Disease. Circ Res. 2017 Mar 17;120(6):923-940.  
<https://doi.org/10.1161/CIRCRESAHA.116.309140>
15. Diab, N. S., Barish, S., Dong, W., Zhao, S., Allington, G., Yu, X., Kahle, K. T., Brueckner, M., & Jin, S. C. (2021). Molecular Genetics and Complex Inheritance of Congenital Heart Disease. Genes, 12(7), 1020.  
<https://doi.org/10.3390/genes12071020>
16. Kopf, P. G., & Walker, M. K. (2009). Overview of developmental heart defects by dioxins, PCBs, and pesticides. Journal of environmental science and health. Part C, Environmental carcinogenesis & ecotoxicology reviews, 27(4), 276–285.  
<https://doi.org/10.1080/10590500903310195>
17. Zhu, H., Kartiko, S., & Finnell, R. H. (2009). Importance of gene-environment interactions in the etiology of selected birth defects. Clinical genetics, 75(5), 409–423. <https://doi.org/10.1111/j.1399-0004.2009.01174.x>
18. Watts, D. H., Huang, S., Culnane, M., Kaiser, K. A., Scheuerle, A., Mofenson, L., Stanley, K., Newell, M. L., Mandelbrot, L., Delfraissy, J. F., & Cunningham, C. K. (2011). Birth defects among a cohort of infants born to HIV-infected women on antiretroviral medication. Journal of perinatal medicine, 39(2), 163–170.  
<https://doi.org/10.1515/jpm.2010.139>

19. Smedts, H. P., van Uitert, E. M., Valkenburg, O., Laven, J. S., Eijkemans, M. J., Lindemans, J., Steegers, E. A., & Steegers-Theunissen, R. P. (2012). A derangement of the maternal lipid profile is associated with an elevated risk of congenital heart disease in the offspring. *Nutrition, metabolism, and cardiovascular diseases : NMCD*, 22(6), 477–485. <https://doi.org/10.1016/j.numecd.2010.07.016>
20. Madsen, N. L., Schwartz, S. M., Lewin, M. B., & Mueller, B. A. (2013). Prepregnancy body mass index and congenital heart defects among offspring: a population-based study. *Congenital heart disease*, 8(2), 131–141. <https://doi.org/10.1111/j.1747-0803.2012.00714.x>
21. Sun, R., Liu, M., Lu, L., Zheng, Y., & Zhang, P. (2015). Congenital Heart Disease: Causes, Diagnosis, Symptoms, and Treatments. *Cell biochemistry and biophysics*, 72(3), 857–860. <https://doi.org/10.1007/s12013-015-0551-6>
22. Williams, K., Carson, J., & Lo, C. (2019). Genetics of Congenital Heart Disease. *Biomolecules*, 9(12), 879. <https://doi.org/10.3390/biom9120879>
23. van der Bom, T., Zomer, A. C., Zwinderman, A. H., Meijboom, F. J., Bouma, B. J., & Mulder, B. J. (2011). The changing epidemiology of congenital heart disease. *Nature reviews. Cardiology*, 8(1), 50–60. <https://doi.org/10.1038/nrcardio.2010.166>
24. Baker, K., Sanchez-de-Toledo, J., Munoz, R., Orr, R., Kiray, S., Shiderly, D., Clemens, M., Wearden, P., Morell, V. O., & Chrysostomou, C. (2012). Critical congenital heart disease--utility of routine screening for chromosomal and other extracardiac malformations. *Congenital heart disease*, 7(2), 145–150. <https://doi.org/10.1111/j.1747-0803.2011.00585.x>
25. Glessner, J. T., Bick, A. G., Ito, K., Homsy, J., Rodriguez-Murillo, L., Fromer, M., Mazaika, E., Vardarajan, B., Italia, M., Leipzig, J., DePalma, S. R., Golhar, R., Sanders, S. J., Yamrom, B., Ronemus, M., Iossifov, I., Willsey, A. J., State, M. W., Kaltman, J. R., White, P. S., ... Chung, W. K. (2014). Increased frequency of de novo copy number variants in congenital heart disease by integrative analysis of single nucleotide polymorphism array and exome sequence data. *Circulation research*, 115(10), 884–896. <https://doi.org/10.1161/CIRCRESAHA.115.304458>

26. Kim, D. S., Kim, J. H., Burt, A. A., Crosslin, D. R., Burnham, N., Kim, C. E., McDonald-McGinn, D. M., Zackai, E. H., Nicolson, S. C., Spray, T. L., Stanaway, I. B., Nickerson, D. A., Heagerty, P. J., Hakonarson, H., Gaynor, J. W., & Jarvik, G. P. (2016). Burden of potentially pathologic copy number variants is higher in children with isolated congenital heart disease and significantly impairs covariate-adjusted transplant-free survival. *The Journal of thoracic and cardiovascular surgery*, 151(4), 1147–51.e4. <https://doi.org/10.1016/j.jtcvs.2015.09.136>
27. Delea, M., Espeche, L. D., Bruque, C. D., Bidondo, M. P., Massara, L. S., Oliveri, J., Brun, P., Cosentino, V. R., Martinoli, C., Tolaba, N., Picon, C., Ponce Zaldua, M. E., Ávila, S., Gutnisky, V., Perez, M., Furforo, L., Buzzalino, N. D., Liascovich, R., Groisman, B., Rittler, M., ... Dain, L. (2018). Genetic Imbalances in Argentinean Patients with Congenital Conotruncal Heart Defects. *Genes*, 9(9), 454. <https://doi.org/10.3390/genes9090454>
28. Zaidi, S., Choi, M., Wakimoto, H., Ma, L., Jiang, J., Overton, J. D., Romano-Adesman, A., Bjornson, R. D., Breitbart, R. E., Brown, K. K., Carriero, N. J., Cheung, Y. H., Deanfield, J., DePalma, S., Fakhro, K. A., Glessner, J., Hakonarson, H., Italia, M. J., Kaltman, J. R., Kaski, J., ... Lifton, R. P. (2013). De novo mutations in histone-modifying genes in congenital heart disease. *Nature*, 498(7453), 220–223. <https://doi.org/10.1038/nature12141>
29. Fahed, A. C., Gelb, B. D., Seidman, J. G., & Seidman, C. E. (2013). Genetics of congenital heart disease: the glass half empty. *Circulation research*, 112(4), 707–720. <https://doi.org/10.1161/CIRCRESAHA.112.300853>
30. Blue, G. M., Kirk, E. P., Giannoulatou, E., Dunwoodie, S. L., Ho, J. W., Hilton, D. C., White, S. M., Sholler, G. F., Harvey, R. P., & Winlaw, D. S. (2014). Targeted next-generation sequencing identifies pathogenic variants in familial congenital heart disease. *Journal of the American College of Cardiology*, 64(23), 2498–2506. <https://doi.org/10.1016/j.jacc.2014.09.048>
31. Jia, Y., Louw, J. J., Breckpot, J., Callewaert, B., Barrea, C., Sznajder, Y., Gewillig, M., Souche, E., Dehaspe, L., Vermeesch, J. R., Lambrechts, D., Devriendt, K., & Corveleyn, A. (2015). The diagnostic value of next generation sequencing in

- familial nonsyndromic congenital heart defects. American journal of medical genetics. Part A, 167A(8), 1822–1829. <https://doi.org/10.1002/ajmg.a.37108>
32. Su, W., Zhu, P., Wang, R., Wu, Q., Wang, M., Zhang, X., Mei, L., Tang, J., Kumar, M., Wang, X., Su, L., & Dong, N. (2017). Congenital heart diseases and their association with the variant distribution features on susceptibility genes. Clinical genetics, 91(3), 349–354. <https://doi.org/10.1111/cge.12835>
  33. Delea, M., Massara, L. S., Espeche, L. D., Bidondo, M. P., Barbero, P., Oliveri, J., Brun, P., Fabro, M., Galain, M., Fernández, C. S., Taboas, M., Bruque, C. D., Kolomenski, J. E., Izquierdo, A., Berenstein, A., Cosentino, V., Martinoli, C., Vilas, M., Rittler, M., Mendez, R., ... On Behalf Of The Pid Acn-Cc Group (2022). Genetic Analysis Algorithm for the Study of Patients with Multiple Congenital Anomalies and Isolated Congenital Heart Disease. Genes, 13(7), 1172. <https://doi.org/10.3390/genes13071172>
  34. Zhang, Y., Ai, F., Zheng, J., & Peng, B. (2017). Associations of GATA4 genetic mutations with the risk of congenital heart disease: A meta-analysis. Medicine, 96(18), e6857. <https://doi.org/10.1097/MD.0000000000006857>
  35. Steimle, J. D., & Moskowitz, I. P. (2017). TBX5: A Key Regulator of Heart Development. Current topics in developmental biology, 122, 195–221. <https://doi.org/10.1016/bs.ctdb.2016.08.008>
  36. Zhu H. (2016). Forkhead box transcription factors in embryonic heart development and congenital heart disease. Life sciences, 144, 194–201. <https://doi.org/10.1016/j.lfs.2015.12.001>
  37. MacGrogan, D., Luxán, G., & de la Pompa, J. L. (2014). Genetic and functional genomics approaches targeting the Notch pathway in cardiac development and congenital heart disease. Briefings in functional genomics, 13(1), 15–27. <https://doi.org/10.1093/bfpg/elt036>
  38. Schott, J. J., Benson, D. W., Basson, C. T., Pease, W., Silberbach, G. M., Moak, J. P., Maron, B. J., Seidman, C. E., & Seidman, J. G. (1998). Congenital heart disease caused by mutations in the transcription factor NKX2-5. Science (New York, N.Y.), 281(5373), 108–111. <https://doi.org/10.1126/science.281.5373.108>

39. Muntean, I., Togănel, R., & Benedek, T. (2017). Genetics of Congenital Heart Disease: Past and Present. *Biochemical genetics*, 55(2), 105–123. <https://doi.org/10.1007/s10528-016-9780-7>
40. Cao, C., Li, L., Zhang, Q., Li, H., Wang, Z., Wang, A., & Liu, J. (2023). Nkx2.5: a crucial regulator of cardiac development, regeneration and diseases. *Frontiers in cardiovascular medicine*, 10, 1270951. <https://doi.org/10.3389/fcvm.2023.1270951>
41. Shiojima, I., Komuro, I., Inazawa, J., Nakahori, Y., Matsushita, I., Abe, T., Nagai, R., & Yazaki, Y. (1995). Assignment of cardiac homeobox gene CSX to human chromosome 5q34. *Genomics*, 27(1), 204–206. <https://doi.org/10.1006/geno.1995.1027>
42. Ellesøe, S. G., Johansen, M. M., Bjerre, J. V., Hjortdal, V. E., Brunak, S., & Larsen, L. A. (2016). Familial Atrial Septal Defect and Sudden Cardiac Death: Identification of a Novel NKX2-5 Mutation and a Review of the Literature. *Congenital heart disease*, 11(3), 283–290. <https://doi.org/10.1111/chd.12317>
43. Fagman, H., & Nilsson, M. (2011). Morphogenetics of early thyroid development. *Journal of molecular endocrinology*, 46(1), R33–R42. <https://doi.org/10.1677/jme-10-0084>
44. Brendolan, A., Ferretti, E., Salsi, V., Moses, K., Quaggin, S., Blasi, F., Cleary, M. L., & Selleri, L. (2005). A Pbx1-dependent genetic and transcriptional network regulates spleen ontogeny. *Development (Cambridge, England)*, 132(13), 3113–3126. <https://doi.org/10.1242/dev.01884>
45. Burn, S. F., Boot, M. J., de Angelis, C., Doohan, R., Arques, C. G., Torres, M., & Hill, R. E. (2008). The dynamics of spleen morphogenesis. *Developmental biology*, 318(2), 303–311. <https://doi.org/10.1016/j.ydbio.2008.03.031>
46. Bodmer R. (1993). The gene tinman is required for specification of the heart and visceral muscles in *Drosophila*. *Development (Cambridge, England)*, 118(3), 719–729. <https://doi.org/10.1242/dev.118.3.719>
47. Harvey R. P. (1996). NK-2 homeobox genes and heart development. *Developmental biology*, 178(2), 203–216. <https://doi.org/10.1006/dbio.1996.0212>

48. Komuro, I., & Izumo, S. (1993). Csx: a murine homeobox-containing gene specifically expressed in the developing heart. *Proceedings of the National Academy of Sciences of the United States of America*, 90(17), 8145–8149. <https://doi.org/10.1073/pnas.90.17.8145>
49. Holland P. W. (2013). Evolution of homeobox genes. *Wiley interdisciplinary reviews. Developmental biology*, 2(1), 31–45. <https://doi.org/10.1002/wdev.78>
50. Garcia-Fernández J. (2005). The genesis and evolution of homeobox gene clusters. *Nature reviews. Genetics*, 6(12), 881–892. <https://doi.org/10.1038/nrg1723>
51. Shiojima, I., Komuro, I., Mizuno, T., Aikawa, R., Akazawa, H., Oka, T., Yamazaki, T., & Yazaki, Y. (1996). Molecular cloning and characterization of human cardiac homeobox gene CSX1. *Circulation research*, 79(5), 920–929. <https://doi.org/10.1161/01.res.79.5.920>
52. Ota, T., Suzuki, Y., Nishikawa, T., Otsuki, T., Sugiyama, T., Irie, R., Wakamatsu, A., Hayashi, K., Sato, H., Nagai, K., Kimura, K., Makita, H., Sekine, M., Obayashi, M., Nishi, T., Shibahara, T., Tanaka, T., Ishii, S., Yamamoto, J., Saito, K., ... Sugano, S. (2004). Complete sequencing and characterization of 21,243 full-length human cDNAs. *Nature genetics*, 36(1), 40–45. <https://doi.org/10.1038/ng1285>
53. Liberatore, C. M., Searcy-Schrick, R. D., Vincent, E. B., & Yutzey, K. E. (2002). Nkx-2.5 gene induction in mice is mediated by a Smad consensus regulatory region. *Developmental biology*, 244(2), 243–256. <https://doi.org/10.1006/dbio.2002.0604>
54. Lien, C. L., McAnally, J., Richardson, J. A., & Olson, E. N. (2002). Cardiac-specific activity of an Nkx2-5 enhancer requires an evolutionarily conserved Smad binding site. *Developmental biology*, 244(2), 257–266. <https://doi.org/10.1006/dbio.2002.0603>
55. Lien, C. L., Wu, C., Mercer, B., Webb, R., Richardson, J. A., & Olson, E. N. (1999). Control of early cardiac-specific transcription of Nkx2-5 by a GATA-dependent enhancer. *Development (Cambridge, England)*, 126(1), 75–84. <https://doi.org/10.1242/dev.126.1.75>

56. Schwartz, R. J., & Olson, E. N. (1999). Building the heart piece by piece: modularity of cis-elements regulating Nkx2-5 transcription. *Development* (Cambridge, England), 126(19), 4187–4192. <https://doi.org/10.1242/dev.126.19.4187>
57. Tanaka, M., Wechsler, S. B., Lee, I. W., Yamasaki, N., Lawitts, J. A., & Izumo, S. (1999). Complex modular cis-acting elements regulate expression of the cardiac specifying homeobox gene Csx/Nkx2.5. *Development* (Cambridge, England), 126(7), 1439–1450. <https://doi.org/10.1242/dev.126.7.1439>
58. Reecy, J. M., Li, X., Yamada, M., DeMayo, F. J., Newman, C. S., Harvey, R. P., & Schwartz, R. J. (1999). Identification of upstream regulatory regions in the heart-expressed homeobox gene Nkx2-5. *Development* (Cambridge, England), 126(4), 839–849. <https://doi.org/10.1242/dev.126.4.839>
59. Searcy, R. D., Vincent, E. B., Liberatore, C. M., & Yutzey, K. E. (1998). A GATA-dependent nkx-2.5 regulatory element activates early cardiac gene expression in transgenic mice. *Development* (Cambridge, England), 125(22), 4461–4470. <https://doi.org/10.1242/dev.125.22.4461>
60. Chen, C. Y., & Schwartz, R. J. (1995). Identification of novel DNA binding targets and regulatory domains of a murine tinman homeodomain factor, nkx-2.5. *The Journal of biological chemistry*, 270(26), 15628–15633. <https://doi.org/10.1074/jbc.270.26.15628>
61. Akazawa, H., & Komuro, I. (2005). Cardiac transcription factor Csx/Nkx2-5: Its role in cardiac development and diseases. *Pharmacology & therapeutics*, 107(2), 252–268. <https://doi.org/10.1016/j.pharmthera.2005.03.005>
62. Elliott, D. A., Solloway, M. J., Wise, N., Biben, C., Costa, M. W., Furtado, M. B., Lange, M., Dunwoodie, S., & Harvey, R. P. (2006). A tyrosine-rich domain within homeodomain transcription factor Nkx2-5 is an essential element in the early cardiac transcriptional regulatory machinery. *Development* (Cambridge, England), 133(7), 1311–1322. <https://doi.org/10.1242/dev.02305>
63. Bouveret, R., Waardenberg, A. J., Schonrock, N., Ramialison, M., Doan, T., de Jong, D., Bondue, A., Kaur, G., Mohamed, S., Fonoudi, H., Chen, C. M., Wouters, M. A., Bhattacharya, S., Plachta, N., Dunwoodie, S. L., Chapman, G.,

- Blanpain, C., & Harvey, R. P. (2015). NKX2-5 mutations causative for congenital heart disease retain functionality and are directed to hundreds of targets. *eLife*, 4, e06942. <https://doi.org/10.7554/eLife.06942>
64. Evans S. M. (1999). Vertebrate tinman homologues and cardiac differentiation. *Seminars in cell & developmental biology*, 10(1), 73–83. <https://doi.org/10.1006/scdb.1999.0282>
65. Elliott, D. A., Kirk, E. P., Schaft, D., & Harvey, R. P. (2010). Chapter 9.1 - NK- 2 class homeodomain proteins: Conserved regulators of cardiogenesis. In N. Rosenthal & R. P. Harvey (Eds.), *Heart development and regeneration* (pp. 569–597). Boston: Academic Press.
66. Kasahara, H., Bartunkova, S., Schinke, M., Tanaka, M., & Izumo, S. (1998). Cardiac and extracardiac expression of Csx/Nkx2.5 homeodomain protein. *Circulation research*, 82(9), 936–946. <https://doi.org/10.1161/01.res.82.9.936>
67. Costa, M. W., Lee, S., Furtado, M. B., Xin, L., Sparrow, D. B., Martinez, C. G., Dunwoodie, S. L., Kurtenbach, E., Mohun, T., Rosenthal, N., & Harvey, R. P. (2011). Complex SUMO-1 regulation of cardiac transcription factor Nkx2-5. *PLoS one*, 6(9), e24812. <https://doi.org/10.1371/journal.pone.0024812>
68. Kolomenski, J. E. (2018, Marzo 28). Implicancias estructurales de variantes génicas en el gen NKX2.5 en cardiopatías congénitas [Tesis de licenciatura]. Universidad de Buenos Aires, Buenos Aires.
69. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic acids research*, 28(1), 235–242. <https://doi.org/10.1093/nar/28.1.235>
70. Koss, M., Bolze, A., Brendolan, A., Saggese, M., Capellini, T. D., Bojilova, E., Boisson, B., Prall, O. W., Elliott, D. A., Solloway, M., Lenti, E., Hidaka, C., Chang, C. P., Mahlaoui, N., Harvey, R. P., Casanova, J. L., & Selleri, L. (2012). Congenital asplenia in mice and humans with mutations in a Pbx/Nkx2-5/p15 module. *Developmental cell*, 22(5), 913–926. <https://doi.org/10.1016/j.devcel.2012.02.009>

71. Kostopoulou, E., Miliordos, K., & Spiliotis, B. (2021). Genetics of primary congenital hypothyroidism-a review. *Hormones (Athens, Greece)*, 20(2), 225–236. <https://doi.org/10.1007/s42000-020-00267-x>
72. Szczepanek-Parulska, E., Budny, B., Borowczyk, M., Zhukov, I., Szutkowski, K., Zawadzka, K., Tahir, R., Minczykowski, A., Niedziela, M., & Ruchała, M. (2022). NKX2-5 Variant in Two Siblings with Thyroid Hemiagenesis. *International journal of molecular sciences*, 23(6), 3414. <https://doi.org/10.3390/ijms23063414>
73. Izumi, K., Noon, S., Wilkens, A., & Krantz, I. D. (2014). NKX2.5 mutation identification on exome sequencing in a patient with heterotaxy. *European journal of medical genetics*, 57(10), 558–561. <https://doi.org/10.1016/j.ejmg.2014.08.003>
74. Zhu, L., Belmont, J. W., & Ware, S. M. (2006). Genetics of human heterotaxias. *European journal of human genetics : EJHG*, 14(1), 17–25. <https://doi.org/10.1038/sj.ejhg.5201506>
75. Berman, H. M., Kleywegt, G. J., Nakamura, H., & Markley, J. L. (2014). The Protein Data Bank archive as an open data resource. *Journal of computer-aided molecular design*, 28(10), 1009–1014. <https://doi.org/10.1007/s10822-014-9770-y>
76. Jaroszewski, L., Li, Z., Krishna, S. S., Bakolitsa, C., Wooley, J., Deacon, A. M., Wilson, I. A., & Godzik, A. (2009). Exploration of uncharted regions of the protein universe. *PLoS biology*, 7(9), e1000205. <https://doi.org/10.1371/journal.pbio.1000205>
77. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., ... Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873), 583–589. <https://doi.org/10.1038/s41586-021-03819-2>
78. Magliery T. J. (2015). Protein stability: computation, sequence statistics, and new experimental methods. *Current opinion in structural biology*, 33, 161–168. <https://doi.org/10.1016/j.sbi.2015.09.002>

79. Dill K. A. (1990). Dominant forces in protein folding. *Biochemistry*, 29(31), 7133–7155. <https://doi.org/10.1021/bi00483a001>
80. Pace, C. N., Scholtz, J. M., & Grimsley, G. R. (2014). Forces stabilizing proteins. *FEBS letters*, 588(14), 2177–2184. <https://doi.org/10.1016/j.febslet.2014.05.006>
81. Chandler D. (2005). Interfaces and the driving force of hydrophobic assembly. *Nature*, 437(7059), 640–647. <https://doi.org/10.1038/nature04162>
82. Pace, C. N., Fu, H., Fryar, K. L., Landua, J., Trevino, S. R., Shirley, B. A., Hendricks, M. M., Iimura, S., Gajiwala, K., Scholtz, J. M., & Grimsley, G. R. (2011). Contribution of hydrophobic interactions to protein stability. *Journal of molecular biology*, 408(3), 514–528. <https://doi.org/10.1016/j.jmb.2011.02.053>
83. Pace, C. N., Fu, H., Lee Fryar, K., Landua, J., Trevino, S. R., Schell, D., Thurlkill, R. L., Imura, S., Scholtz, J. M., Gajiwala, K., Sevcik, J., Urbanikova, L., Myers, J. K., Takano, K., Hebert, E. J., Shirley, B. A., & Grimsley, G. R. (2014). Contribution of hydrogen bonds to protein stability. *Protein science : a publication of the Protein Society*, 23(5), 652–661. <https://doi.org/10.1002/pro.2449>
84. Zhou, H. X., & Pang, X. (2018). Electrostatic Interactions in Protein Structure, Folding, Binding, and Condensation. *Chemical reviews*, 118(4), 1691–1741. <https://doi.org/10.1021/acs.chemrev.7b00305>
85. Wada, A., & Nakamura, H. (1981). Nature of the charge distribution in proteins. *Nature*, 293(5835), 757–758. <https://doi.org/10.1038/293757a0>
86. Pace, C. N., Grimsley, G. R., & Scholtz, J. M. (2009). Protein ionizable groups: pK values and their contribution to protein stability and solubility. *The Journal of biological chemistry*, 284(20), 13285–13289. <https://doi.org/10.1074/jbc.R800080200>
87. Pace, C. N., Grimsley, G. R., Thomson, J. A., & Barnett, B. J. (1988). Conformational stability and activity of ribonuclease T1 with zero, one, and two intact disulfide bonds. *The Journal of biological chemistry*, 263(24), 11820–11825.
88. Dombkowski, A. A., Sultana, K. Z., & Craig, D. B. (2014). Protein disulfide engineering. *FEBS letters*, 588(2), 206–212. <https://doi.org/10.1016/j.febslet.2013.11.024>

89. Anfinsen C. B. (1972). The formation and stabilization of protein structure. The Biochemical journal, 128(4), 737–749. <https://doi.org/10.1042/bj1280737>
90. Babu, M. M., Kriwacki, R. W., & Pappu, R. V. (2012). Structural biology. Versatility from protein disorder. Science (New York, N.Y.), 337(6101), 1460–1461. <https://doi.org/10.1126/science.1228775>
91. Dyson, H. J., & Wright, P. E. (2005). Intrinsically unstructured proteins and their functions. Nature reviews. Molecular cell biology, 6(3), 197–208. <https://doi.org/10.1038/nrm1589>
92. Dunker, A. K., Silman, I., Uversky, V. N., & Sussman, J. L. (2008). Function and structure of inherently disordered proteins. Current opinion in structural biology, 18(6), 756–764. <https://doi.org/10.1016/j.sbi.2008.10.002>
93. Oates, M. E., Romero, P., Ishida, T., Ghalwash, M., Mizianty, M. J., Xue, B., Dosztányi, Z., Uversky, V. N., Obradovic, Z., Kurgan, L., Dunker, A. K., & Gough, J. (2013). D<sup>2</sup>P<sup>2</sup>: database of disordered protein predictions. Nucleic acids research, 41(Database issue), D508–D516. <https://doi.org/10.1093/nar/gks1226>
94. Wright, P. E., & Dyson, H. J. (2015). Intrinsically disordered proteins in cellular signalling and regulation. Nature reviews. Molecular cell biology, 16(1), 18–29. <https://doi.org/10.1038/nrm3920>
95. Már, M., Nitsenko, K., & Heidarsson, P. O. (2023). Multifunctional Intrinsically Disordered Regions in Transcription Factors. Chemistry (Weinheim an der Bergstrasse, Germany), 29(21), e202203369. <https://doi.org/10.1002/chem.202203369>
96. Tompa, P., Davey, N. E., Gibson, T. J., & Babu, M. M. (2014). A million peptide motifs for the molecular biologist. Molecular cell, 55(2), 161–169. <https://doi.org/10.1016/j.molcel.2014.05.032>
97. Gould, C. M., Diella, F., Via, A., Puntervoll, P., Gemünd, C., Chabanis-Davidson, S., Michael, S., Sayadi, A., Bryne, J. C., Chica, C., Seiler, M., Davey, N. E., Haslam, N., Weatheritt, R. J., Budd, A., Hughes, T., Pas, J., Rychlewski, L., Travé, G., Aasland, R., ... Gibson, T. J. (2010). ELM: the status of the 2010 eukaryotic linear motif resource. Nucleic acids research, 38(Database issue), D167–D180. <https://doi.org/10.1093/nar/gkp1016>

98. Davey, N. E., Van Roey, K., Weatheritt, R. J., Toedt, G., Uyar, B., Altenberg, B., Budd, A., Diella, F., Dinkel, H., & Gibson, T. J. (2012). Attributes of short linear motifs. *Molecular bioSystems*, 8(1), 268–281. <https://doi.org/10.1039/c1mb05231d>
99. Cheng, T. M., Blundell, T. L., & Fernandez-Recio, J. (2007). pyDock: electrostatics and desolvation for effective scoring of rigid-body protein-protein docking. *Proteins*, 68(2), 503–515. <https://doi.org/10.1002/prot.21419>
100. Jiménez-García, B., Pons, C., & Fernández-Recio, J. (2013). pyDockWEB: a web server for rigid-body protein-protein docking using electrostatics and desolvation scoring. *Bioinformatics (Oxford, England)*, 29(13), 1698–1699. <https://doi.org/10.1093/bioinformatics/btt262>
101. Guerois, R., Nielsen, J. E., & Serrano, L. (2002). Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *Journal of molecular biology*, 320(2), 369–387. [https://doi.org/10.1016/S0022-2836\(02\)00442-4](https://doi.org/10.1016/S0022-2836(02)00442-4)
102. Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., & Serrano, L. (2005). The FoldX web server: an online force field. *Nucleic acids research*, 33(Web Server issue), W382–W388. <https://doi.org/10.1093/nar/gki387>
103. Pearce, R., Huang, X., Setiawan, D., & Zhang, Y. (2019). EvoDesign: Designing Protein-Protein Binding Interactions Using Evolutionary Interface Profiles in Conjunction with an Optimized Physical Energy Function. *Journal of molecular biology*, 431(13), 2467–2476. <https://doi.org/10.1016/j.jmb.2019.02.028>
104. Huang, X., Pearce, R., & Zhang, Y. (2020). EvoEF2: accurate and fast energy function for computational protein design. *Bioinformatics (Oxford, England)*, 36(4), 1135–1142. <https://doi.org/10.1093/bioinformatics/btz740>
105. Vangone, A., & Bonvin, A. M. (2015). Contacts-based prediction of binding affinity in protein-protein complexes. *eLife*, 4, e07454. <https://doi.org/10.7554/eLife.07454>
106. Xue, L. C., Rodrigues, J. P., Kastiris, P. L., Bonvin, A. M., & Vangone, A. (2016). PRODIGY: a web server for predicting the binding affinity of protein-

- protein complexes. *Bioinformatics* (Oxford, England), 32(23), 3676–3678. <https://doi.org/10.1093/bioinformatics/btw514>
107. Dehouck, Y., Kwasigroch, J. M., Rooman, M., & Gilis, D. (2013). BeAtMuSiC: Prediction of changes in protein-protein binding affinity on mutations. *Nucleic acids research*, 41(Web Server issue), W333–W339. <https://doi.org/10.1093/nar/gkt450>
  108. Pires, D. E., Ascher, D. B., & Blundell, T. L. (2014). mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics* (Oxford, England), 30(3), 335–342. <https://doi.org/10.1093/bioinformatics/btt691>
  109. Blaabjerg, L. M., Kassem, M. M., Good, L. L., Jonsson, N., Cagiada, M., Johansson, K. E., Boomsma, W., Stein, A., & Lindorff-Larsen, K. (2023). Rapid protein stability prediction using deep learning representations. *eLife*, 12, e82593. <https://doi.org/10.7554/eLife.82593>
  110. Reamon-Buettner, S. M., & Borlak, J. (2010). NKX2-5: an update on this hypermutable homeodomain protein and its role in human congenital heart disease (CHD). *Human mutation*, 31(11), 1185–1194. <https://doi.org/10.1002/humu.21345>
  111. den Dunnen, J. T., Dalgleish, R., Maglott, D. R., Hart, R. K., Greenblatt, M. S., McGowan-Jordan, J., Roux, A. F., Smith, T., Antonarakis, S. E., & Taschner, P. E. (2016). HGVS Recommendations for the Description of Sequence Variants: 2016 Update. *Human mutation*, 37(6), 564–569. <https://doi.org/10.1002/humu.22981>
  112. Gulley, M. L., Brazier, R. M., Halling, K. C., Hsi, E. D., Kant, J. A., Nikiforova, M. N., Nowak, J. A., Ogino, S., Oliveira, A., Polesky, H. F., Silverman, L., Tubbs, R. R., Van Deerlin, V. M., Vance, G. H., Versalovic, J., & Molecular Pathology Resource Committee, College of American Pathologists (2007). Clinical laboratory reports in molecular pathology. *Archives of pathology & laboratory medicine*, 131(6), 852–863. <https://doi.org/10.5858/2007-131-852-CLRIMP>
  113. Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W. W., Hegde, M., Lyon, E., Spector, E., Voelkerding, K., Rehm, H. L., &

- ACMG Laboratory Quality Assurance Committee (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in medicine : official journal of the American College of Medical Genetics*, 17(5), 405–424. <https://doi.org/10.1038/gim.2015.30>
114. O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C. M., ... Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic acids research*, 44(D1), D733–D745. <https://doi.org/10.1093/nar/gkv1189>
  115. van Weerd, J. H., & Christoffels, V. M. (2016). The formation and function of the cardiac conduction system. *Development (Cambridge, England)*, 143(2), 197–210. <https://doi.org/10.1242/dev.124883>
  116. Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T. R., & Weirauch, M. T. (2018). The Human Transcription Factors. *Cell*, 172(4), 650–665. <https://doi.org/10.1016/j.cell.2018.01.029>
  117. Gonzalez-Teran, B., Pittman, M., Felix, F., Thomas, R., Richmond-Buccola, D., Hüttenhain, R., Choudhary, K., Moroni, E., Costa, M. W., Huang, Y., Padmanabhan, A., Alexanian, M., Lee, C. Y., Maven, B. E. J., Samse-Knapp, K., Morton, S. U., McGregor, M., Gifford, C. A., Seidman, J. G., Seidman, C. E., ... Srivastava, D. (2022). Transcription factor protein interactomes reveal genetic determinants in heart disease. *Cell*, 185(5), 794–814.e30. <https://doi.org/10.1016/j.cell.2022.01.021>
  118. DeCaprio, J., & Kohl, T. O. (2020). Chromatin Immunoprecipitation. *Cold Spring Harbor protocols*, 2020(8), 098665. <https://doi.org/10.1101/pdb.prot098665>

119. Park P. J. (2009). ChIP-seq: advantages and challenges of a maturing technology. *Nature reviews. Genetics*, 10(10), 669–680. <https://doi.org/10.1038/nrg2641>
120. Furey T. S. (2012). ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. *Nature reviews. Genetics*, 13(12), 840–852. <https://doi.org/10.1038/nrg3306>
121. Nakato, R., & Sakata, T. (2021). Methods for ChIP-seq analysis: A practical workflow and advanced applications. *Methods (San Diego, Calif.)*, 187, 44–53. <https://doi.org/10.1016/j.ymeth.2020.03.005>
122. Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C. L., Serova, N., Davis, S., & Soboleva, A. (2013). NCBI GEO: archive for functional genomics data sets--update. *Nucleic acids research*, 41(Database issue), D991–D995. <https://doi.org/10.1093/nar/gks1193>
123. Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics*, 10(1), 57–63. <https://doi.org/10.1038/nrg2484>
124. Zoabi, Y., & Shomron, N. (2021). Processing and Analysis of RNA-seq Data from Public Resources. *Methods in molecular biology (Clifton, N.J.)*, 2243, 81–94. [https://doi.org/10.1007/978-1-0716-1103-6\\_4](https://doi.org/10.1007/978-1-0716-1103-6_4)
125. Costa-Silva, J., Domingues, D., & Lopes, F. M. (2017). RNA-Seq differential expression analysis: An extended review and a software tool. *PloS one*, 12(12), e0190152. <https://doi.org/10.1371/journal.pone.0190152>
126. Stark, R., Grzelak, M., & Hadfield, J. (2019). RNA sequencing: the teenage years. *Nature reviews. Genetics*, 20(11), 631–656. <https://doi.org/10.1038/s41576-019-0150-2>
127. Stormo G. D. (2000). DNA binding sites: representation and discovery. *Bioinformatics (Oxford, England)*, 16(1), 16–23. <https://doi.org/10.1093/bioinformatics/16.1.16>

128. Stormo G. D. (2015). DNA Motif Databases and Their Uses. Current protocols in bioinformatics, 51, 2.15.1–2.15.6. <https://doi.org/10.1002/0471250953.bi0215s51>
129. Kulakovskiy, I. V., Vorontsov, I. E., Yevshin, I. S., Sharipov, R. N., Fedorova, A. D., Rumynskiy, E. I., Medvedeva, Y. A., Magana-Mora, A., Bajic, V. B., Papatsenko, D. A., Kolpakov, F. A., & Makeev, V. J. (2018). HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. Nucleic acids research, 46(D1), D252–D259. <https://doi.org/10.1093/nar/gkx1106>
130. Schneider, T. D., & Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. Nucleic acids research, 18(20), 6097–6100. <https://doi.org/10.1093/nar/18.20.6097>
131. Collins, F. S., & Fink, L. (1995). The Human Genome Project. Alcohol health and research world, 19(3), 190–195.
132. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., ... International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. Nature, 409(6822), 860–921. <https://doi.org/10.1038/35057062>
133. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature genetics, 25(1), 25–29. <https://doi.org/10.1038/75556>
134. Gene Ontology Consortium, Aleksander, S. A., Balhoff, J., Carbon, S., Cherry, J. M., Drabkin, H. J., Ebert, D., Feuermann, M., Gaudet, P., Harris, N. L., Hill, D. P., Lee, R., Mi, H., Moxon, S., Mungall, C. J., Muruganugan, A., Mushayahama, T., Sternberg, P. W., Thomas, P. D., Van Auken, K., ...

- Westerfield, M. (2023). The Gene Ontology knowledgebase in 2023. *Genetics*, 224(1), iyad031. <https://doi.org/10.1093/genetics/iyad031>
135. Duck, G., Nenadic, G., Filannino, M., Brass, A., Robertson, D. L., & Stevens, R. (2016). A Survey of Bioinformatics Database and Software Usage through Mining the Literature. *PloS one*, 11(6), e0157989. <https://doi.org/10.1371/journal.pone.0157989>
  136. Yoshino, S., & Suzuki, H. I. (2022). The molecular understanding of super-enhancer dysregulation in cancer. *Nagoya journal of medical science*, 84(2), 216–229. <https://doi.org/10.18999/nagjms.84.2.216>
  137. Wang, M., Chen, Q., Wang, S., Xie, H., Liu, J., Huang, R., Xiang, Y., Jiang, Y., Tian, D., & Bian, E. (2023). Super-enhancers complexes zoom in transcription in cancer. *Journal of experimental & clinical cancer research : CR*, 42(1), 183. <https://doi.org/10.1186/s13046-023-02763-5>
  138. Hnisz, D., Abraham, B. J., Lee, T. I., Lau, A., Saint-André, V., Sigova, A. A., Hoke, H. A., & Young, R. A. (2013). Super-enhancers in the control of cell identity and disease. *Cell*, 155(4), 934–947. <https://doi.org/10.1016/j.cell.2013.09.053>
  139. Saint-André, V., Federation, A. J., Lin, C. Y., Abraham, B. J., Reddy, J., Lee, T. I., Bradner, J. E., & Young, R. A. (2016). Models of human core transcriptional regulatory circuitries. *Genome research*, 26(3), 385–396. <https://doi.org/10.1101/gr.197590.115>
  140. Warde-Farley, D., Donaldson, S. L., Comes, O., Zuberi, K., Badrawi, R., Chao, P., Franz, M., Grouios, C., Kazi, F., Lopes, C. T., Maitland, A., Mostafavi, S., Montojo, J., Shao, Q., Wright, G., Bader, G. D., & Morris, Q. (2010). The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic acids research*, 38(Web Server issue), W214–W220. <https://doi.org/10.1093/nar/gkq537>
  141. Szklarczyk, D., Kirsch, R., Koutrouli, M., Nastou, K., Mehryary, F., Hachilif, R., Gable, A. L., Fang, T., Doncheva, N. T., Pyysalo, S., Bork, P., Jensen, L. J., & von Mering, C. (2023). The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced

- genome of interest. *Nucleic acids research*, 51(D1), D638–D646.  
<https://doi.org/10.1093/nar/gkac1000>
142. Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., Campbell, N. H., Chavali, G., Chen, C., del-Toro, N., Duesbury, M., Dumousseau, M., Galeota, E., Hinz, U., Iannuccelli, M., Jagannathan, S., Jimenez, R., Khadake, J., Lagreid, A., Licata, L., ... Hermjakob, H. (2014). The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic acids research*, 42(Database issue), D358–D363.  
<https://doi.org/10.1093/nar/gkt1115>
143. Oughtred, R., Rust, J., Chang, C., Breitkreutz, B. J., Stark, C., Willems, A., Boucher, L., Leung, G., Kolas, N., Zhang, F., Dolma, S., Coulombe-Huntington, J., Chatr-Aryamontri, A., Dolinski, K., & Tyers, M. (2021). The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein science : a publication of the Protein Society*, 30(1), 187–200. <https://doi.org/10.1002/pro.3978>
144. Jumppanen, M., Kinnunen, S. M., Välimäki, M. J., Talman, V., Auno, S., Bruun, T., Boije Af Gennäs, G., Xhaard, H., Aumüller, I. B., Ruskoaho, H., & Yli-Kauhaluoma, J. (2019). Synthesis, Identification, and Structure-Activity Relationship Analysis of GATA4 and NKX2-5 Protein-Protein Interaction Modulators. *Journal of medicinal chemistry*, 62(17), 8284–8310.  
<https://doi.org/10.1021/acs.jmedchem.9b01086>
145. Zhou, P., Zhang, Y., Sethi, I., Ye, L., Trembley, M. A., Cao, Y., Akerberg, B. N., Xiao, F., Zhang, X., Li, K., Jardin, B. D., Mazumdar, N., Ma, Q., He, A., Zhou, B., & Pu, W. T. (2022). GATA4 Regulates Developing Endocardium Through Interaction With ETS1. *Circulation research*, 131(11), e152–e168.  
<https://doi.org/10.1161/CIRCRESAHA.120.318102>
146. Pradhan, L., Gopal, S., Li, S., Ashur, S., Suryanarayanan, S., Kasahara, H., & Nam, H. J. (2016). Intermolecular Interactions of Cardiac Transcription Factors NKX2.5 and TBX5. *Biochemistry*, 55(12), 1702–1710.  
<https://doi.org/10.1021/acs.biochem.6b00171>

147. Delea, M (2023, Diciembre 19). Defectos genómicos en cardiopatías congénitas [Tesis de doctorado]. Universidad de Buenos Aires, Buenos Aires.
148. Gutmanas, A., Alhroub, Y., Battle, G. M., Berrisford, J. M., Bochet, E., Conroy, M. J., Dana, J. M., Fernandez Montecelo, M. A., van Ginkel, G., Gore, S. P., Haslam, P., Hatherley, R., Hendrickx, P. M., Hirshberg, M., Lagerstedt, I., Mir, S., Mukhopadhyay, A., Oldfield, T. J., Patwardhan, A., Rinaldi, L., ... Kleywegt, G. J. (2014). PDBe: Protein Data Bank in Europe. *Nucleic acids research*, 42(Database issue), D285–D291. <https://doi.org/10.1093/nar/gkt1180>
149. Velankar, S., van Ginkel, G., Alhroub, Y., Battle, G. M., Berrisford, J. M., Conroy, M. J., Dana, J. M., Gore, S. P., Gutmanas, A., Haslam, P., Hendrickx, P. M., Lagerstedt, I., Mir, S., Fernandez Montecelo, M. A., Mukhopadhyay, A., Oldfield, T. J., Patwardhan, A., Sanz-García, E., Sen, S., Slowley, R. A., ... Kleywegt, G. J. (2016). PDBe: improved accessibility of macromolecular structure data from PDB and EMDB. *Nucleic acids research*, 44(D1), D385–D395. <https://doi.org/10.1093/nar/gkv1047>
150. Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., & Ferrin, T. E. (2004). UCSF Chimera--a visualization system for exploratory research and analysis. *Journal of computational chemistry*, 25(13), 1605–1612. <https://doi.org/10.1002/jcc.20084>
151. Pauling, L. C. (1960). *The Nature of the Chemical Bond*, 3rd edn. Cornell University Press, Ithaca, New York.
152. Ashkenazy, H., Abadi, S., Martz, E., Chay, O., Mayrose, I., Pupko, T., & Ben-Tal, N. (2016). ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic acids research*, 44(W1), W344–W350. <https://doi.org/10.1093/nar/gkw408>
153. Ng, P. C., & Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. *Nucleic acids research*, 31(13), 3812–3814. <https://doi.org/10.1093/nar/gkg509>
154. Gouw, M., Michael, S., Sámano-Sánchez, H., Kumar, M., Zeke, A., Lang, B., Bely, B., Chemes, L. B., Davey, N. E., Deng, Z., Diella, F., Gürth, C. M., Huber, A. K., Kleinsorg, S., Schlegel, L. S., Palopoli, N., Roey, K. V., Altenberg,

- B., Reményi, A., Dinkel, H., ... Gibson, T. J. (2018). The eukaryotic linear motif resource - 2018 update. *Nucleic acids research*, 46(D1), D428–D434. <https://doi.org/10.1093/nar/gkx1077>
155. Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic acids research*, 29(1), 308–311. <https://doi.org/10.1093/nar/29.1.308>
  156. Mottaz, A., David, F. P., Veuthey, A. L., & Yip, Y. L. (2010). Easy retrieval of single amino-acid polymorphisms and phenotype information using SwissVar. *Bioinformatics* (Oxford, England), 26(6), 851–852. <https://doi.org/10.1093/bioinformatics/btq028>
  157. Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., O'Donnell-Luria, A. H., Ware, J. S., Hill, A. J., Cummings, B. B., Tukiainen, T., Birnbaum, D. P., Kosmicki, J. A., Duncan, L. E., Estrada, K., Zhao, F., Zou, J., Pierce-Hoffman, E., Berghout, J., Cooper, D. N., ... Exome Aggregation Consortium (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616), 285–291. <https://doi.org/10.1038/nature19057>
  158. Fokkema, I. F., Taschner, P. E., Schaafsma, G. C., Celli, J., Laros, J. F., & den Dunnen, J. T. (2011). LOVD v.2.0: the next generation in gene variant databases. *Human mutation*, 32(5), 557–563. <https://doi.org/10.1002/humu.21438>
  159. Landrum, M. J., Lee, J. M., Benson, M., Brown, G. R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., Karapetyan, K., Katz, K., Liu, C., Maddipatla, Z., Malheiro, A., McDaniel, K., Ovetsky, M., Riley, G., Zhou, G., Holmes, J. B., ... Maglott, D. R. (2018). ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic acids research*, 46(D1), D1062–D1067. <https://doi.org/10.1093/nar/gkx1153>
  160. J. D. Hunter, "Matplotlib: A 2D Graphics Environment", *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90-95, 2007
  161. Thoen, H., Moreel, L., Adriaenssens, B., Keuleers, S., & Voet, J. (2020). Undiagnosed adult congenital heart disease presenting with postpartum

- dyspnoea: a case report. *ESC heart failure*, 7(5), 3193–3197.  
<https://doi.org/10.1002/ehf2.12868>
162. Warnes, C. A., Liberthson, R., Danielson, G. K., Dore, A., Harris, L., Hoffman, J. I., Somerville, J., Williams, R. G., & Webb, G. D. (2001). Task force 1: the changing profile of congenital heart disease in adult life. *Journal of the American College of Cardiology*, 37(5), 1170–1175.  
[https://doi.org/10.1016/s0735-1097\(01\)01272-4](https://doi.org/10.1016/s0735-1097(01)01272-4)
  163. Winston, J. B., Erlich, J. M., Green, C. A., Aluko, A., Kaiser, K. A., Takematsu, M., Barlow, R. S., Sureka, A. O., LaPage, M. J., Janss, L. L., & Jay, P. Y. (2010). Heterogeneity of genetic modifiers ensures normal cardiac development. *Circulation*, 121(11), 1313–1321.  
<https://doi.org/10.1161/CIRCULATIONAHA.109.887687>
  164. Baglivo, M., Dassati, S., Krasi, G., Fanelli, F., Kurti, D., Bonelli, A., Arabia, G., Fabbicatore, D., Muneretto, C., & Bertelli, M. (2019). Atrial septal defects, supra-ventricular aortic stenosis and syndromes predisposing to aneurysm of large vessels. *Acta bio-medica : Atenei Parmensis*, 90(10-S), 53–57.  
<https://doi.org/10.23750/abm.v90i10-S.8760>
  165. Benson D. W. (2002). The genetics of congenital heart disease: a point in the revolution. *Cardiology clinics*, 20(3), 385–vi. [https://doi.org/10.1016/s0733-8651\(02\)00009-7](https://doi.org/10.1016/s0733-8651(02)00009-7)
  166. Benson, D. W., Sharkey, A., Fatkin, D., Lang, P., Basson, C. T., McDonough, B., Strauss, A. W., Seidman, J. G., & Seidman, C. E. (1998). Reduced penetrance, variable expressivity, and genetic heterogeneity of familial atrial septal defects. *Circulation*, 97(20), 2043–2048.  
<https://doi.org/10.1161/01.cir.97.20.2043>
  167. Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S., & Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nature methods*, 7(4), 248–249.  
<https://doi.org/10.1038/nmeth0410-248>
  168. McLean, C. Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B. T., Lowe, C. B., Wenger, A. M., & Bejerano, G. (2010). GREAT improves functional

- interpretation of cis-regulatory regions. *Nature biotechnology*, 28(5), 495–501.  
<https://doi.org/10.1038/nbt.1630>
169. Tanigawa, Y., Dyer, E. S., & Bejerano, G. (2022). WhichTF is functionally important in your open chromatin data?. *PLoS computational biology*, 18(8), e1010378. <https://doi.org/10.1371/journal.pcbi.1010378>
  170. Wang, F., Yang, Q., Wu, F., Zhang, Y., Sun, S., Wang, X., Gui, Y., & Li, Q. (2019). Identification of a 42-bp heart-specific enhancer of the notch1b gene in zebrafish embryos. *Developmental dynamics : an official publication of the American Association of Anatomists*, 248(6), 426–436.  
<https://doi.org/10.1002/dvdy.31>
  171. von Both, I., Silvestri, C., Erdemir, T., Lickert, H., Walls, J. R., Henkelman, R. M., Rossant, J., Harvey, R. P., Attisano, L., & Wrana, J. L. (2004). Foxh1 is essential for development of the anterior heart field. *Developmental cell*, 7(3), 331–345. <https://doi.org/10.1016/j.devcel.2004.07.023>
  172. Shiratori, H., Sakuma, R., Watanabe, M., Hashiguchi, H., Mochida, K., Sakai, Y., Nishino, J., Saijoh, Y., Whitman, M., & Hamada, H. (2001). Two-step regulation of left-right asymmetric expression of Pitx2: initiation by nodal signaling and maintenance by Nkx2. *Molecular cell*, 7(1), 137–149.  
[https://doi.org/10.1016/s1097-2765\(01\)00162-9](https://doi.org/10.1016/s1097-2765(01)00162-9)
  173. Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., & de Hoon, M. J. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics (Oxford, England)*, 25(11), 1422–1423.  
<https://doi.org/10.1093/bioinformatics/btp163>
  174. Machanick, P., & Bailey, T. L. (2011). MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics (Oxford, England)*, 27(12), 1696–1697.  
<https://doi.org/10.1093/bioinformatics/btr189>
  175. Ge, S. X., Jung, D., & Yao, R. (2020). ShinyGO: a graphical gene-set enrichment tool for animals and plants. *Bioinformatics (Oxford, England)*, 36(8), 2628–2629. <https://doi.org/10.1093/bioinformatics/btz931>

176. Pradhan, L., Genis, C., Scone, P., Weinberg, E. O., Kasahara, H., & Nam, H. J. (2012). Crystal structure of the human NKX2.5 homeodomain in complex with DNA target. *Biochemistry*, 51(32), 6312–6319. <https://doi.org/10.1021/bi300849c>
177. Luna-Zurita, L., Stirnimann, C. U., Glatt, S., Kaynak, B. L., Thomas, S., Baudin, F., Samee, M. A., He, D., Small, E. M., Mileikovsky, M., Nagy, A., Holloway, A. K., Pollard, K. S., Müller, C. W., & Bruneau, B. G. (2016). Complex Interdependence Regulates Heterotypic Transcription Factor Distribution and Coordinates Cardiogenesis. *Cell*, 164(5), 999–1014. <https://doi.org/10.1016/j.cell.2016.01.004>
178. Lei, X., Zhao, J., Sagendorf, J. M., Rajashekar, N., Xu, J., Dantas Machado, A. C., Sen, C., Rohs, R., Feng, P., & Chen, L. (2020). Crystal Structures of Ternary Complexes of MEF2 and NKX2-5 Bound to DNA Reveal a Disease Related Protein-Protein Interaction Interface. *Journal of molecular biology*, 432(19), 5499–5508. <https://doi.org/10.1016/j.jmb.2020.07.004>
179. Kasahara, H., & Izumo, S. (1999). Identification of the in vivo casein kinase II phosphorylation site within the homeodomain of the cardiac tissue-specifying homeobox gene product Csx/Nkx2.5. *Molecular and cellular biology*, 19(1), 526–536. <https://doi.org/10.1128/MCB.19.1.526>
180. Ouyang, P., Zhang, H., Fan, Z., Wei, P., Huang, Z., Wang, S., & Li, T. (2016). A R/K-rich motif in the C-terminal of the homeodomain is required for complete translocating of NKX2.5 protein into nucleus. *Gene*, 592(2), 276–280. <https://doi.org/10.1016/j.gene.2016.07.022>
181. Kim, E. Y., Chen, L., Ma, Y., Yu, W., Chang, J., Moskowitz, I. P., & Wang, J. (2011). Expression of sumoylation deficient Nkx2.5 mutant in Nkx2.5 haploinsufficient mice leads to congenital heart defects. *PloS one*, 6(6), e20803. <https://doi.org/10.1371/journal.pone.0020803>
182. Wang, J., Zhang, H., Iyer, D., Feng, X. H., & Schwartz, R. J. (2008). Regulation of cardiac specific nkx2.5 gene activity by small ubiquitin-like modifier. *The Journal of biological chemistry*, 283(34), 23235–23243. <https://doi.org/10.1074/jbc.M709748200>

183. Li, T., Li, Y. M., Jia, Z. Q., Chen, P., Ma, K. T., & Zhou, C. Y. (2007). Carboxyl terminus of Nkx2.5 impairs its interaction with p300. *Journal of molecular biology*, 370(5), 976–992. <https://doi.org/10.1016/j.jmb.2007.05.033>
184. Tang, X., Ma, H., Han, L., Zheng, W., Lu, Y. B., Chen, X. F., Liang, S. T., Wei, G. H., Zhang, Z. Q., Chen, H. Z., & Liu, D. P. (2016). SIRT1 deacetylates the cardiac transcription factor Nkx2.5 and inhibits its transcriptional activity. *Scientific reports*, 6, 36576. <https://doi.org/10.1038/srep36576>
185. Keleş, S., Warren, C. L., Carlson, C. D., & Ansari, A. Z. (2008). CSI-Tree: a regression tree approach for modeling binding properties of DNA-binding molecules based on cognate site identification (CSI) data. *Nucleic acids research*, 36(10), 3171–3184. <https://doi.org/10.1093/nar/gkn057>
186. Xu, J. H., Gu, J. Y., Guo, Y. H., Zhang, H., Qiu, X. B., Li, R. G., Shi, H. Y., Liu, H., Yang, X. X., Xu, Y. J., Qu, X. K., & Yang, Y. Q. (2017). Prevalence and Spectrum of NKX2-5 Mutations Associated With Sporadic Adult-Onset Dilated Cardiomyopathy. *International heart journal*, 58(4), 521–529. <https://doi.org/10.1536/ihj.16-440>
187. Xu, Y. J., Qiu, X. B., Yuan, F., Shi, H. Y., Xu, L., Hou, X. M., Qu, X. K., Liu, X., Huang, R. T., Xue, S., Yang, Y. Q., & Li, R. G. (2017). Prevalence and spectrum of NKX2.5 mutations in patients with congenital atrial septal defect and atrioventricular block. *Molecular medicine reports*, 15(4), 2247–2254. <https://doi.org/10.3892/mmr.2017.6249>
188. Yuan, F., Qiu, X. B., Li, R. G., Qu, X. K., Wang, J., Xu, Y. J., Liu, X., Fang, W. Y., Yang, Y. Q., & Liao, D. N. (2015). A novel NKX2-5 loss-of-function mutation predisposes to familial dilated cardiomyopathy and arrhythmias. *International journal of molecular medicine*, 35(2), 478–486. <https://doi.org/10.3892/ijmm.2014.2029>
189. Qu, X. K., Qiu, X. B., Yuan, F., Wang, J., Zhao, C. M., Liu, X. Y., Zhang, X. L., Li, R. G., Xu, Y. J., Hou, X. M., Fang, W. Y., Liu, X., & Yang, Y. Q. (2014). A novel NKX2.5 loss-of-function mutation associated with congenital bicuspid aortic valve. *The American journal of cardiology*, 114(12), 1891–1895. <https://doi.org/10.1016/j.amjcard.2014.09.028>

190. Yu, H., Xu, J. H., Song, H. M., Zhao, L., Xu, W. J., Wang, J., Li, R. G., Xu, L., Jiang, W. F., Qiu, X. B., Jiang, J. Q., Qu, X. K., Liu, X., Fang, W. Y., Jiang, J. F., & Yang, Y. Q. (2014). Mutational spectrum of the NKX2-5 gene in patients with lone atrial fibrillation. *International journal of medical sciences*, 11(6), 554–563. <https://doi.org/10.7150/ijms.8407>
191. Reamon-Buettner, S. M., Sattlegger, E., Ciribilli, Y., Inga, A., Wessel, A., & Borlak, J. (2013). Transcriptional defect of an inherited NKX2-5 haplotype comprising a SNP, a nonsynonymous and a synonymous mutation, associated with human congenital heart disease. *PloS one*, 8(12), e83295. <https://doi.org/10.1371/journal.pone.0083295>
192. Xie, W. H., Chang, C., Xu, Y. J., Li, R. G., Qu, X. K., Fang, W. Y., Liu, X., & Yang, Y. Q. (2013). Prevalence and spectrum of Nkx2.5 mutations associated with idiopathic atrial fibrillation. *Clinics (Sao Paulo, Brazil)*, 68(6), 777–784. [https://doi.org/10.6061/clinics/2013\(06\)09](https://doi.org/10.6061/clinics/2013(06)09)
193. Costa, M. W., Guo, G., Wolstein, O., Vale, M., Castro, M. L., Wang, L., Otway, R., Riek, P., Cochrane, N., Furtado, M., Semsarian, C., Weintraub, R. G., Yeoh, T., Hayward, C., Keogh, A., Macdonald, P., Feneley, M., Graham, R. M., Seidman, J. G., Seidman, C. E., ... Harvey, R. P. (2013). Functional characterization of a novel mutation in NKX2-5 associated with congenital heart disease and adult-onset cardiomyopathy. *Circulation. Cardiovascular genetics*, 6(3), 238–247. <https://doi.org/10.1161/CIRCGENETICS.113.000057>
194. Huang, R. T., Xue, S., Xu, Y. J., Zhou, M., & Yang, Y. Q. (2013). A novel NKX2.5 loss-of-function mutation responsible for familial atrial fibrillation. *International journal of molecular medicine*, 31(5), 1119–1126. <https://doi.org/10.3892/ijmm.2013.1316>
195. van Engelen, K., Mommersteeg, M. T., Baars, M. J., Lam, J., Ilgun, A., van Trotsenburg, A. S., Smets, A. M., Christoffels, V. M., Mulder, B. J., & Postma, A. V. (2012). The ambiguous role of NKX2-5 mutations in thyroid dysgenesis. *PloS one*, 7(12), e52685. <https://doi.org/10.1371/journal.pone.0052685>
196. Kodo, K., Nishizawa, T., Furutani, M., Arai, S., Ishihara, K., Oda, M., Makino, S., Fukuda, K., Takahashi, T., Matsuoka, R., Nakanishi, T., &

- Yamagishi, H. (2012). Genetic analysis of essential cardiac transcription factors in 256 patients with non-syndromic congenital heart defects. *Circulation journal : official journal of the Japanese Circulation Society*, 76(7), 1703–1711. <https://doi.org/10.1253/circj.cj-11-1389>
197. Hermanns, P., Grasberger, H., Refetoff, S., & Pohlenz, J. (2011). Mutations in the NKX2.5 gene and the PAX8 promoter in a girl with thyroid dysgenesis. *The Journal of clinical endocrinology and metabolism*, 96(6), E977–E981. <https://doi.org/10.1210/jc.2010-2341>
  198. Wang, J., Xin, Y. F., Liu, X. Y., Liu, Z. M., Wang, X. Z., & Yang, Y. Q. (2011). A novel NKX2-5 mutation in familial ventricular septal defect. *International journal of molecular medicine*, 27(3), 369–375. <https://doi.org/10.3892/ijmm.2010.585>
  199. Ouyang, P., Saarel, E., Bai, Y., Luo, C., Lv, Q., Xu, Y., Wang, F., Fan, C., Younoszai, A., Chen, Q., Tu, X., & Wang, Q. K. (2011). A de novo mutation in NKX2.5 associated with atrial septal defects, ventricular noncompaction, syncope and sudden death. *Clinica chimica acta; international journal of clinical chemistry*, 412(1-2), 170–175. <https://doi.org/10.1016/j.cca.2010.09.035>
  200. Rauch, R., Hofbeck, M., Zweier, C., Koch, A., Zink, S., Trautmann, U., Hoyer, J., Kaulitz, R., Singer, H., & Rauch, A. (2010). Comprehensive genotype-phenotype analysis in 230 patients with tetralogy of Fallot. *Journal of medical genetics*, 47(5), 321–331. <https://doi.org/10.1136/jmq.2009.070391>
  201. Esposito, G., Grutter, G., Drago, F., Costa, M. W., De Santis, A., Bosco, G., Marino, B., Bellacchio, E., Lepri, F., Harvey, R. P., Sarkozy, A., & Dallapiccola, B. (2009). Molecular analysis of PRKAG2, LAMP2, and NKX2-5 genes in a cohort of 125 patients with accessory atrioventricular connection. *American journal of medical genetics. Part A*, 149A(7), 1574–1577. <https://doi.org/10.1002/ajmg.a.32907>
  202. Dentice, M., Cordeddu, V., Rosica, A., Ferrara, A. M., Santarpia, L., Salvatore, D., Chiovato, L., Perri, A., Moschini, L., Fazzini, C., Olivieri, A., Costa, P., Stoppioni, V., Baserga, M., De Felice, M., Sorcini, M., Fenzi, G., Di Lauro, R., Tartaglia, M., & Macchia, P. E. (2006). Missense mutation in the transcription

- factor NKX2-5: a novel molecular event in the pathogenesis of thyroid dysgenesis. *The Journal of clinical endocrinology and metabolism*, 91(4), 1428–1433. <https://doi.org/10.1210/jc.2005-1350>
203. Kasahara, H., & Benson, D. W. (2004). Biochemical analyses of eight NKX2.5 homeodomain missense mutations causing atrioventricular block and cardiac anomalies. *Cardiovascular research*, 64(1), 40–51. <https://doi.org/10.1016/j.cardiores.2004.06.004>
  204. Zhu, W., Shiojima, I., Hiroi, Y., Zou, Y., Akazawa, H., Mizukami, M., Toko, H., Yazaki, Y., Nagai, R., & Komuro, I. (2000). Functional analyses of three Csx/Nkx-2.5 mutations that cause human congenital heart disease. *The Journal of biological chemistry*, 275(45), 35291–35296. <https://doi.org/10.1074/jbc.M000525200>
  205. Kasahara, H., Lee, B., Schott, J. J., Benson, D. W., Seidman, J. G., Seidman, C. E., & Izumo, S. (2000). Loss of function and inhibitory effects of human CSX/NKX2.5 homeoprotein mutations associated with congenital heart disease. *The Journal of clinical investigation*, 106(2), 299–308. <https://doi.org/10.1172/JCI9860>
  206. He, A., Kong, S. W., Ma, Q., & Pu, W. T. (2011). Co-occupancy by multiple cardiac transcription factors identifies transcriptional enhancers active in heart. *Proceedings of the National Academy of Sciences of the United States of America*, 108(14), 5632–5637. <https://doi.org/10.1073/pnas.1016959108>
  207. van den Boogaard, M., Wong, L. Y., Tessadori, F., Bakker, M. L., Dreizehnter, L. K., Wakker, V., Bezzina, C. R., 't Hoen, P. A., Bakkers, J., Barnett, P., & Christoffels, V. M. (2012). Genetic variation in T-box binding element functionally affects SCN5A/SCN10A enhancer. *The Journal of clinical investigation*, 122(7), 2519–2530. <https://doi.org/10.1172/JCI62613>
  208. Ye, W., Wang, J., Song, Y., Yu, D., Sun, C., Liu, C., Chen, F., Zhang, Y., Wang, F., Harvey, R. P., Schrader, L., Martin, J. F., & Chen, Y. (2015). A common Shox2-Nkx2-5 antagonistic mechanism primes the pacemaker cell fate in the pulmonary vein myocardium and sinoatrial node. *Development (Cambridge, England)*, 142(14), 2521–2532. <https://doi.org/10.1242/dev.120220>

209. Dupays, L., Shang, C., Wilson, R., Kotecha, S., Wood, S., Towers, N., & Mohun, T. (2015). Sequential Binding of MEIS1 and NKX2-5 on the Popdc2 Gene: A Mechanism for Spatiotemporal Regulation of Enhancers during Cardiogenesis. *Cell reports*, 13(1), 183–195. <https://doi.org/10.1016/j.celrep.2015.08.065>
210. Anderson, D. J., Kaplan, D. I., Bell, K. M., Koutsis, K., Haynes, J. M., Mills, R. J., Phelan, D. G., Qian, E. L., Leitoguinho, A. R., Arasaratnam, D., Labonne, T., Ng, E. S., Davis, R. P., Casini, S., Passier, R., Hudson, J. E., Porrello, E. R., Costa, M. W., Rafii, A., Curl, C. L., ... Elliott, D. A. (2018). NKX2-5 regulates human cardiomyogenesis via a HEY2 dependent transcriptional network. *Nature communications*, 9(1), 1373. <https://doi.org/10.1038/s41467-018-03714-x>
211. Prall, O. W., Menon, M. K., Solloway, M. J., Watanabe, Y., Zaffran, S., Bajolle, F., Biben, C., McBride, J. J., Robertson, B. R., Chaulet, H., Stennard, F. A., Wise, N., Schaft, D., Wolstein, O., Furtado, M. B., Shiratori, H., Chien, K. R., Hamada, H., Black, B. L., Saga, Y., ... Harvey, R. P. (2007). An Nkx2-5/Bmp2/Smad1 negative feedback loop controls heart progenitor specification and proliferation. *Cell*, 128(5), 947–959. <https://doi.org/10.1016/j.cell.2007.01.042>
212. Hiroi, Y., Kudoh, S., Monzen, K., Ikeda, Y., Yazaki, Y., Nagai, R., & Komuro, I. (2001). Tbx5 associates with Nkx2-5 and synergistically promotes cardiomyocyte differentiation. *Nature genetics*, 28(3), 276–280. <https://doi.org/10.1038/90123>
213. Durocher, D., Charron, F., Warren, R., Schwartz, R. J., & Nemer, M. (1997). The cardiac transcription factors Nkx2-5 and GATA-4 are mutual cofactors. *The EMBO journal*, 16(18), 5687–5696. <https://doi.org/10.1093/emboj/16.18.5687>
214. Shiojima, I., Komuro, I., Oka, T., Hiroi, Y., Mizuno, T., Takimoto, E., Monzen, K., Aikawa, R., Akazawa, H., Yamazaki, T., Kudoh, S., & Yazaki, Y. (1999). Context-dependent transcriptional cooperation mediated by cardiac transcription factors Csx/Nkx-2.5 and GATA-4. *The Journal of biological chemistry*, 274(12), 8231–8239. <https://doi.org/10.1074/jbc.274.12.8231>

215. Li, X., Wang, W., Wang, J., Malovannaya, A., Xi, Y., Li, W., Guerra, R., Hawke, D. H., Qin, J., & Chen, J. (2015). Proteomic analyses reveal distinct chromatin-associated and soluble transcription factor complexes. *Molecular systems biology*, 11(1), 775. <https://doi.org/10.15252/msb.20145504>
216. Mallon, B. S., Chenoweth, J. G., Johnson, K. R., Hamilton, R. S., Tesar, P. J., Yavatkar, A. S., Tyson, L. J., Park, K., Chen, K. G., Fann, Y. C., & McKay, R. D. (2013). StemCellDB: the human pluripotent stem cell database at the National Institutes of Health. *Stem cell research*, 10(1), 57–66. <https://doi.org/10.1016/j.scr.2012.09.002>
217. Wu, G., Feng, X., & Stein, L. (2010). A human functional protein interaction network and its application to cancer data analysis. *Genome biology*, 11(5), R53. <https://doi.org/10.1186/gb-2010-11-5-r53>
218. Lin, A., Wang, R. T., Ahn, S., Park, C. C., & Smith, D. J. (2010). A genome-wide map of human genetic interactions inferred from radiation hybrid genotypes. *Genome research*, 20(8), 1122–1132. <https://doi.org/10.1101/gr.104216.109>
219. Rual, J. F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G. F., Gibbons, F. D., Dreze, M., Ayivi-Guedehoussou, N., Klitgord, N., Simon, C., Boxem, M., Milstein, S., Rosenberg, J., Goldberg, D. S., Zhang, L. V., Wong, S. L., Franklin, G., Li, S., ... Vidal, M. (2005). Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437(7062), 1173–1178. <https://doi.org/10.1038/nature04209>
220. Afouda B. A. (2022). Towards Understanding the Gene-Specific Roles of GATA Factors in Heart Development: Does GATA4 Lead the Way?. *International journal of molecular sciences*, 23(9), 5255. <https://doi.org/10.3390/ijms23095255>
221. Jiang, J. Q., Shen, F. F., Fang, W. Y., Liu, X., & Yang, Y. Q. (2011). Novel GATA4 mutations in lone atrial fibrillation. *International journal of molecular medicine*, 28(6), 1025–1032. <https://doi.org/10.3892/ijmm.2011.783>
222. Yang, Y. Q., Wang, M. Y., Zhang, X. L., Tan, H. W., Shi, H. F., Jiang, W. F., Wang, X. H., Fang, W. Y., & Liu, X. (2011). GATA4 loss-of-function mutations

- in familial atrial fibrillation. *Clinica chimica acta; international journal of clinical chemistry*, 412(19-20), 1825–1830. <https://doi.org/10.1016/j.cca.2011.06.017>
223. Posch, M. G., Boldt, L. H., Polotzki, M., Richter, S., Rolf, S., Perrot, A., Dietz, R., Ozcelik, C., & Haverkamp, W. (2010). Mutations in the cardiac transcription factor GATA4 in patients with lone atrial fibrillation. *European journal of medical genetics*, 53(4), 201–203. <https://doi.org/10.1016/j.ejmg.2010.03.008>
  224. Møller Nielsen, A. K., Dehn, A. M., Hjortdal, V., & Larsen, L. A. (2024). TBX5 variants and cardiac phenotype: A systematic review of the literature and a novel variant. *European journal of medical genetics*, 68, 104920. <https://doi.org/10.1016/j.ejmg.2024.104920>
  225. Rathjens, F. S., Blenkle, A., Iyer, L. M., Renger, A., Syeda, F., Noack, C., Jungmann, A., Dewenter, M., Toischer, K., El-Armouche, A., Müller, O. J., Fabritz, L., Zimmermann, W. H., Zelarayan, L. C., & Zafeiriou, M. P. (2021). Preclinical evidence for the therapeutic value of TBX5 normalization in arrhythmia control. *Cardiovascular research*, 117(8), 1908–1922. <https://doi.org/10.1093/cvr/cvaa239>
  226. Hata, Y., Kinoshita, K., Mizumaki, K., Yamaguchi, Y., Hirono, K., Ichida, F., Takasaki, A., Mori, H., & Nishida, N. (2016). Postmortem genetic analysis of sudden unexplained death syndrome under 50 years of age: A next-generation sequencing study. *Heart rhythm*, 13(7), 1544–1551. <https://doi.org/10.1016/j.hrthm.2016.03.038>
  227. Zhou, W., Zhao, L., Jiang, J. Q., Jiang, W. F., Yang, Y. Q., & Qiu, X. B. (2015). A novel TBX5 loss-of-function mutation associated with sporadic dilated cardiomyopathy. *International journal of molecular medicine*, 36(1), 282–288. <https://doi.org/10.3892/ijmm.2015.2206>
  228. Zhang, X. L., Qiu, X. B., Yuan, F., Wang, J., Zhao, C. M., Li, R. G., Xu, L., Xu, Y. J., Shi, H. Y., Hou, X. M., Qu, X. K., Xu, Y. W., & Yang, Y. Q. (2015). TBX5 loss-of-function mutation contributes to familial dilated cardiomyopathy. *Biochemical and biophysical research communications*, 459(1), 166–171. <https://doi.org/10.1016/j.bbrc.2015.02.094>

229. Baban, A., Postma, A. V., Marini, M., Trocchio, G., Santilli, A., Pelegrini, M., Sirleto, P., Lerone, M., Albanese, S. B., Barnett, P., Boogerd, C. J., Dallapiccola, B., Digilio, M. C., Ravazzolo, R., & Pongiglione, G. (2014). Identification of TBX5 mutations in a series of 94 patients with Tetralogy of Fallot. *American journal of medical genetics. Part A*, 164A(12), 3100–3107. <https://doi.org/10.1002/ajmq.a.36783>
230. Patel, C., Silcock, L., McMullan, D., Brueton, L., & Cox, H. (2012). TBX5 intragenic duplication: a family with an atypical Holt-Oram syndrome phenotype. *European journal of human genetics : EJHG*, 20(8), 863–869. <https://doi.org/10.1038/ejhg.2012.16>
231. Muru, K., Kalev, I., Teek, R., Sõnajalg, M., Kuuse, K., Reimand, T., & Ounap, K. (2011). A Boy with Holt-Oram Syndrome Caused by Novel Mutation c.1304delT in the TBX5 Gene. *Molecular syndromology*, 1(6), 307–310. <https://doi.org/10.1159/000330109>
232. Pon, J. R., & Marra, M. A. (2016). MEF2 transcription factors: developmental regulators and emerging cancer genes. *Oncotarget*, 7(3), 2297–2312. <https://doi.org/10.18632/oncotarget.6223>
233. Friedrich, F. W., Dilanian, G., Khattar, P., Juhr, D., Gueneau, L., Charron, P., Fressart, V., Vilquin, J. T., Isnard, R., Gouya, L., Richard, P., Hammoudi, N., Komajda, M., Bonne, G., Eschenhagen, T., Dubourg, O., Villard, E., & Carrier, L. (2013). A novel genetic variant in the transcription factor Islet-1 exerts gain of function on myocyte enhancer factor 2C promoter activity. *European journal of heart failure*, 15(3), 267–276. <https://doi.org/10.1093/eurjhf/hfs178>
234. Fujino, M., Ojima, M., & Takahashi, S. (2023). Exploring Large MAF Transcription Factors: Functions, Pathology, and Mouse Models with Point Mutations. *Genes*, 14(10), 1883. <https://doi.org/10.3390/genes14101883>
235. Jiang, M., Xu, S., Bai, M., & Zhang, A. (2021). The emerging role of MEIS1 in cell proliferation and differentiation. *American journal of physiology. Cell physiology*, 320(3), C264–C269. <https://doi.org/10.1152/ajpcell.00422.2020>

236. Paul, S., Zhang, X., & He, J. Q. (2020). Homeobox gene Meis1 modulates cardiovascular regeneration. *Seminars in cell & developmental biology*, 100, 52–61. <https://doi.org/10.1016/j.semcdb.2019.10.003>
237. Porcher, C., Chagraoui, H., & Kristiansen, M. S. (2017). SCL/TAL1: a multifaceted regulator from blood development to disease. *Blood*, 129(15), 2051–2060. <https://doi.org/10.1182/blood-2016-12-754051>
238. Schumacher, J. A., Bloomekatz, J., Garavito-Aguilar, Z. V., & Yelon, D. (2013). tal1 Regulates the formation of intercellular junctions and the maintenance of identity in the endocardium. *Developmental biology*, 383(2), 214–226. <https://doi.org/10.1016/j.ydbio.2013.09.019>
239. Van Handel, B., Montel-Hagen, A., Sasidharan, R., Nakano, H., Ferrari, R., Boogerd, C. J., Schredelseker, J., Wang, Y., Hunter, S., Org, T., Zhou, J., Li, X., Pellegrini, M., Chen, J. N., Orkin, S. H., Kurdistani, S. K., Evans, S. M., Nakano, A., & Mikkola, H. K. (2012). Scl represses cardiomyogenesis in prospective hemogenic endothelium and endocardium. *Cell*, 150(3), 590–605. <https://doi.org/10.1016/j.cell.2012.06.026>
240. Bussmann, J., Bakkers, J., & Schulte-Merker, S. (2007). Early endocardial morphogenesis requires Scl/Tal1. *PLoS genetics*, 3(8), e140. <https://doi.org/10.1371/journal.pgen.0030140>
241. Li, N., Li, Y. J., Guo, X. J., Wu, S. H., Jiang, W. F., Zhang, D. L., Wang, K. W., Li, L., Sun, Y. M., Xu, Y. J., Yang, Y. Q., & Qiu, X. B. (2023). Discovery of TBX20 as a Novel Gene Underlying Atrial Fibrillation. *Biology*, 12(9), 1186. <https://doi.org/10.3390/biology12091186>
242. Chang, Y., Wacker, J., Ingles, J., Macciocca, I., King, I., Australian Genomics Cardiovascular Disorders Flagship, Semsarian, C., McGaughran, J., Weintraub, R. G., & Bagnall, R. D. (2024). TBX20 loss-of-function variants in families with left ventricular non-compaction cardiomyopathy. *Journal of medical genetics*, 61(2), 171–175. <https://doi.org/10.1136/jmg-2023-109455>
243. Chen, Y., Xiao, D., Zhang, L., Cai, C. L., Li, B. Y., & Liu, Y. (2021). The Role of Tbx20 in Cardiovascular Development and Function. *Frontiers in cell and developmental biology*, 9, 638542. <https://doi.org/10.3389/fcell.2021.638542>

244. Kirk, E. P., Sunde, M., Costa, M. W., Rankin, S. A., Wolstein, O., Castro, M. L., Butler, T. L., Hyun, C., Guo, G., Otway, R., Mackay, J. P., Waddell, L. B., Cole, A. D., Hayward, C., Keogh, A., Macdonald, P., Griffiths, L., Fatkin, D., Sholler, G. F., Zorn, A. M., ... Harvey, R. P. (2007). Mutations in cardiac T-box factor gene TBX20 are associated with diverse cardiac pathologies, including defects of septation and valvulogenesis and cardiomyopathy. *American journal of human genetics*, 81(2), 280–291. <https://doi.org/10.1086/519530>
245. Dinsmore, C. J., & Soriano, P. (2020). A Fateful Decision: Tgif1 and Cardiac Neural Crest Identity. *Developmental cell*, 53(3), 255–256. <https://doi.org/10.1016/j.devcel.2020.04.013>
246. Guca, E., Suñol, D., Ruiz, L., Konkol, A., Cordero, J., Torner, C., Aragon, E., Martin-Malpartida, P., Riera, A., & Macias, M. J. (2018). TGIF1 homeodomain interacts with Smad MH1 domain and represses TGF- $\beta$  signaling. *Nucleic acids research*, 46(17), 9220–9235. <https://doi.org/10.1093/nar/gky680>
247. Friedrich, F. W., Reischmann, S., Schwalm, A., Unger, A., Ramanujam, D., Münch, J., Müller, O. J., Hengstenberg, C., Galve, E., Charron, P., Linke, W. A., Engelhardt, S., Patten, M., Richard, P., van der Velden, J., Eschenhagen, T., Isnard, R., & Carrier, L. (2014). FHL2 expression and variants in hypertrophic cardiomyopathy. *Basic research in cardiology*, 109(6), 451. <https://doi.org/10.1007/s00395-014-0451-8>
248. Matsa, L. S., Rangaraju, A., Vengaldas, V., Latifi, M., Jahromi, H. M., Ananthapur, V., & Nallari, P. (2013). Haplotypes of NOS3 gene polymorphisms in dilated cardiomyopathy. *PloS one*, 8(7), e70523. <https://doi.org/10.1371/journal.pone.0070523>
249. Berger, K., Stögbauer, F., Stoll, M., Wellmann, J., Hüge, A., Cheng, S., Kessler, C., John, U., Assmann, G., Ringelstein, E. B., & Funke, H. (2007). The glu298asp polymorphism in the nitric oxide synthase 3 gene is associated with the risk of ischemic stroke in two large independent case-control studies. *Human genetics*, 121(2), 169–178. <https://doi.org/10.1007/s00439-006-0302-2>
250. Casas, J. P., Bautista, L. E., Humphries, S. E., & Hingorani, A. D. (2004). Endothelial nitric oxide synthase genotype and ischemic heart disease: meta-

- analysis of 26 studies involving 23028 subjects. *Circulation*, 109(11), 1359–1365.  
<https://doi.org/10.1161/01.CIR.0000121357.76910.A3>
251. Hingorani, A. D., Liang, C. F., Fatibene, J., Lyon, A., Monteith, S., Parsons, A., Haydock, S., Hopper, R. V., Stephens, N. G., O'Shaughnessy, K. M., & Brown, M. J. (1999). A common variant of the endothelial nitric oxide synthase (Glu298-->Asp) is a major risk factor for coronary artery disease in the UK. *Circulation*, 100(14), 1515–1520. <https://doi.org/10.1161/01.cir.100.14.1515>
  252. Huang, Y., Wang, L. L., Liu, Z. B., Chen, C., Ren, X., Luo, A. T., Ma, J. H., Antzelevitch, C., Barajas-Martínez, H., & Hu, D. (2024). Underlying mechanism of atrial fibrillation-associated Nppa-I137T mutation and cardiac effect of potential drug therapy. *Heart rhythm*, 21(2), 184–196.  
<https://doi.org/10.1016/j.hrthm.2023.10.025>
  253. Ren, X., Xu, C., Zhan, C., Yang, Y., Shi, L., Wang, F., Wang, C., Xia, Y., Yang, B., Wu, G., Wang, P., Li, X., Wang, D., Xiong, X., Liu, J., Liu, Y., Liu, M., Liu, J., Tu, X., & Wang, Q. K. (2010). Identification of NPPA variants associated with atrial fibrillation in a Chinese GeneID population. *Clinica chimica acta; international journal of clinical chemistry*, 411(7-8), 481–485.  
<https://doi.org/10.1016/j.cca.2009.12.019>
  254. Kuehl, K., Loffredo, C., Lammer, E. J., Iovannisci, D. M., & Shaw, G. M. (2010). Association of congenital cardiovascular malformations with 33 single nucleotide polymorphisms of selected cardiovascular disease-related genes. *Birth defects research. Part A, Clinical and molecular teratology*, 88(2), 101–110.  
<https://doi.org/10.1002/bdra.20630>
  255. Fu, Y., & Eisen, H. J. (2018). Genetics of Dilated Cardiomyopathy. *Current cardiology reports*, 20(11), 121. <https://doi.org/10.1007/s11886-018-1061-0>
  256. Denham, N. C., Pearman, C. M., Ding, W. Y., Waktare, J., Gupta, D., Snowdon, R., Hall, M., Cooper, R., Modi, S., Todd, D., & Mahida, S. (2019). Systematic re-evaluation of SCN5A variants associated with Brugada syndrome. *Journal of cardiovascular electrophysiology*, 30(1), 118–127.  
<https://doi.org/10.1111/jce.13740>

257. Kroncke, B. M., Glazer, A. M., Smith, D. K., Blume, J. D., & Roden, D. M. (2018). SCN5A (NaV1.5) Variant Functional Perturbation and Clinical Presentation: Variants of a Certain Significance. *Circulation. Genomic and precision medicine*, 11(5), e002095. <https://doi.org/10.1161/CIRCGEN.118.002095>
258. Veerman, C. C., Wilde, A. A., & Lodder, E. M. (2015). The cardiac sodium channel gene SCN5A and its gene product NaV1.5: Role in physiology and pathophysiology. *Gene*, 573(2), 177–187. <https://doi.org/10.1016/j.gene.2015.08.062>
259. Turillazzi, E., La Rocca, G., Anzalone, R., Corrao, S., Neri, M., Pomara, C., Riezzo, I., Karch, S. B., & Fineschi, V. (2008). Heterozygous nonsense SCN5A mutation W822X explains a simultaneous sudden infant death syndrome. *Virchows Archiv : an international journal of pathology*, 453(2), 209–216. <https://doi.org/10.1007/s00428-008-0632-7>
260. Bellchambers, H. M., & Ware, S. M. (2018). ZIC3 in Heterotaxy. *Advances in experimental medicine and biology*, 1046, 301–327. [https://doi.org/10.1007/978-981-10-7311-3\\_15](https://doi.org/10.1007/978-981-10-7311-3_15)
261. Cowan, J., Tariq, M., & Ware, S. M. (2014). Genetic and functional analyses of ZIC3 variants in congenital heart disease. *Human mutation*, 35(1), 66–75. <https://doi.org/10.1002/humu.22457>
262. Ware, S. M., Peng, J., Zhu, L., Fernbach, S., Colicos, S., Casey, B., Towbin, J., & Belmont, J. W. (2004). Identification and functional analysis of ZIC3 mutations in heterotaxy and related congenital heart defects. *American journal of human genetics*, 74(1), 93–105. <https://doi.org/10.1086/380998>
263. Ren, J., Miao, D., Li, Y., & Gao, R. (2021). Spotlight on Isl1: A Key Player in Cardiovascular Development and Diseases. *Frontiers in cell and developmental biology*, 9, 793605. <https://doi.org/10.3389/fcell.2021.793605>
264. Wu, S. H., Wang, X. H., Xu, Y. J., Gu, J. N., Yang, C. X., Qiao, Q., Guo, X. J., Guo, Y. H., Qiu, X. B., Jiang, W. F., & Yang, Y. Q. (2020). ISL1 loss-of-function variation causes familial atrial fibrillation. *European journal of medical genetics*, 63(11), 104029. <https://doi.org/10.1016/j.ejmg.2020.104029>

265. Wang, Z., Song, H. M., Wang, F., Zhao, C. M., Huang, R. T., Xue, S., Li, R. G., Qiu, X. B., Xu, Y. J., Liu, X. Y., & Yang, Y. Q. (2019). A New ISL1 Loss-of-Function Mutation Predisposes to Congenital Double Outlet Right Ventricle. *International heart journal*, 60(5), 1113–1122. <https://doi.org/10.1536/ihj.18-685>
266. Ma, L., Wang, J., Li, L., Qiao, Q., Di, R. M., Li, X. M., Xu, Y. J., Zhang, M., Li, R. G., Qiu, X. B., Li, X., & Yang, Y. Q. (2019). ISL1 loss-of-function mutation contributes to congenital heart defects. *Heart and vessels*, 34(4), 658–668. <https://doi.org/10.1007/s00380-018-1289-z>
267. Penna, M. S., Hu, R. C., Rodney, G. G., & Cooper, T. A. (2023). The role of Limch1 alternative splicing in skeletal muscle function. *Life science alliance*, 6(6), e202201868. <https://doi.org/10.26508/lsa.202201868>
268. Lin, Y. H., Zhen, Y. Y., Chien, K. Y., Lee, I. C., Lin, W. C., Chen, M. Y., & Pai, L. M. (2017). LIMCH1 regulates nonmuscle myosin-II activity and suppresses cell migration. *Molecular biology of the cell*, 28(8), 1054–1065. <https://doi.org/10.1091/mbc.E15-04-0218>
269. Chen, T. Y., Zhou, M., Lin, M. Q., Liang, S. T., Yan, Y., Wang, S. M., Fang, C. S., Li, D., & Ruan, Y. (2021). Research Progress on the SERPINE1 Protein and Chronic Inflammatory Diseases of the Upper Respiratory Tract: A Literature Review. *International archives of allergy and immunology*, 182(11), 1097–1102. <https://doi.org/10.1159/000516195>
270. Xu, Z., Castellino, F. J., & Ploplis, V. A. (2010). Plasminogen activator inhibitor-1 (PAI-1) is cardioprotective in mice by maintaining microvascular integrity and cardiac architecture. *Blood*, 115(10), 2038–2047. <https://doi.org/10.1182/blood-2009-09-244962>
271. Zaman, A. K., French, C. J., Schneider, D. J., & Sobel, B. E. (2009). A profibrotic effect of plasminogen activator inhibitor type-1 (PAI-1) in the heart. *Experimental biology and medicine* (Maywood, N.J.), 234(3), 246–254. <https://doi.org/10.3181/0811-RM-321>
272. Ma, L., Li, J., Liu, Y., Pang, S., Huang, W., & Yan, B. (2013). Novel and functional variants within the TBX18 gene promoter in ventricular septal defects.

Molecular and cellular biochemistry, 382(1-2), 121–126.  
<https://doi.org/10.1007/s11010-013-1725-4>

273. Sun, T., Huang, Z., Liang, W. C., Yin, J., Lin, W. Y., Wu, J., Vernes, J. M., Lutman, J., Caplazi, P., Jeet, S., Wong, T., Wong, M., DePianto, D. J., Morshead, K. B., Sun, K. H., Modrusan, Z., Vander Heiden, J. A., Abbas, A. R., Zhang, H., Xu, M., ... Arron, J. R. (2021). TGF $\beta$ 2 and TGF $\beta$ 3 isoforms drive fibrotic disease pathogenesis. *Science translational medicine*, 13(605), eabe0407. <https://doi.org/10.1126/scitranslmed.abe0407>
274. Wilson S. E. (2021). TGF beta -1, -2 and -3 in the modulation of fibrosis in the cornea and other organs. *Experimental eye research*, 207, 108594. <https://doi.org/10.1016/j.exer.2021.108594>
275. Hanna, A., & Frangogiannis, N. G. (2019). The Role of the TGF- $\beta$  Superfamily in Myocardial Infarction. *Frontiers in cardiovascular medicine*, 6, 140. <https://doi.org/10.3389/fcvm.2019.00140>
276. Geng, B., Wang, X., Park, K. H., Lee, K. E., Kim, J., Chen, P., Zhou, X., Tan, T., Yang, C., Zou, X., Janssen, P. M., Cao, L., Ye, L., Wang, X., Cai, C., & Zhu, H. (2022). UCHL1 protects against ischemic heart injury via activating HIF-1 $\alpha$  signal pathway. *Redox biology*, 52, 102295. <https://doi.org/10.1016/j.redox.2022.102295>
277. Bi, H. L., Zhang, X. L., Zhang, Y. L., Xie, X., Xia, Y. L., Du, J., & Li, H. H. (2020). The deubiquitinase UCHL1 regulates cardiac hypertrophy by stabilizing epidermal growth factor receptor. *Science advances*, 6(16), eaax4826. <https://doi.org/10.1126/sciadv.aax4826>
278. Matuszczak, E., Tylicka, M., Komarowska, M. D., Debek, W., & Hermanowicz, A. (2020). Ubiquitin carboxy-terminal hydrolase L1 - physiology and pathology. *Cell biochemistry and function*, 38(5), 533–540. <https://doi.org/10.1002/cbf.3527>
279. Bi, H. L., Zhang, Y. L., Yang, J., Shu, Q., Yang, X. L., Yan, X., Chen, C., Li, Z., & Li, H. H. (2020). Inhibition of UCHL1 by LDN-57444 attenuates Ang II-Induced atrial fibrillation in mice. *Hypertension research : official journal of the*

- Japanese Society of Hypertension, 43(3), 168–177.  
<https://doi.org/10.1038/s41440-019-0354-z>
280. Granadillo, J. L., Chung, W. K., Hecht, L., Corsten-Janssen, N., Wegner, D., Nij Bijvank, S. W. A., Toler, T. L., Pineda-Alvarez, D. E., Douglas, G., Murphy, J. J., Shimony, J., & Shinawi, M. (2018). Variable cardiovascular phenotypes associated with SMAD2 pathogenic variants. *Human mutation*, 39(12), 1875–1884. <https://doi.org/10.1002/humu.23627>
  281. Pikkarainen, S., Tokola, H., Kerkelä, R., & Ruskoaho, H. (2004). GATA transcription factors in the developing and adult heart. *Cardiovascular research*, 63(2), 196–207. <https://doi.org/10.1016/j.cardiores.2004.03.025>
  282. Watanabe, Y., Zaffran, S., Kuroiwa, A., Higuchi, H., Ogura, T., Harvey, R. P., Kelly, R. G., & Buckingham, M. (2012). Fibroblast growth factor 10 gene regulation in the second heart field by Tbx1, Nkx2-5, and Islet1 reveals a genetic switch for down-regulation in the myocardium. *Proceedings of the National Academy of Sciences of the United States of America*, 109(45), 18273–18280. <https://doi.org/10.1073/pnas.1215360109>
  283. Nie, J., Jiang, M., Zhang, X., Tang, H., Jin, H., Huang, X., Yuan, B., Zhang, C., Lai, J. C., Nagamine, Y., Pan, D., Wang, W., & Yang, Z. (2015). Post-transcriptional Regulation of Nkx2-5 by RHAU in Heart Development. *Cell reports*, 13(4), 723–732. <https://doi.org/10.1016/j.celrep.2015.09.043>
  284. Sloutskin, A., Itzhak, D., Vogler, G., Pozeilov, H., Ideses, D., Alter, H., Adato, O., Shachar, H., Doniger, T., Shohat-Ophir, G., Frasch, M., Bodmer, R., Duttke, S. H., & Juven-Gershon, T. (2024). From promoter motif to cardiac function: a single DPE motif affects transcription regulation and organ function in vivo. *Development (Cambridge, England)*, 151(14), dev202355. <https://doi.org/10.1242/dev.202355>
  285. González-Castro, T. B., Tovilla-Zárate, C. A., López-Narvaez, M. L., Juárez-Rojop, I. E., Calderón-Colmenero, J., Sandoval, J. P., García-Montes, J. A., Blachman-Braun, R., Castillo-Avila, R. G., García-Flores, E., Cazarín-Santos, B. G., Borgonio-Cuadra, V. M., Posadas-Sánchez, R., Vargas-Alarcón, G., Rodríguez-Pérez, J. M., & Pérez-Hernández, N. (2020). Association between

- congenital heart disease and NKX2.5 gene polymorphisms: systematic review and meta-analysis. *Biomarkers in medicine*, 14(18), 1747–1757. <https://doi.org/10.2217/bmm-2020-0190>
286. Chen, H., Li, T., Wu, Y., Wang, X., Wang, M., Wang, X., & Fang, X. (2022). Association between single-nucleotide polymorphisms of NKX2.5 and congenital heart disease in Chinese population: A meta-analysis. *Open life sciences*, 17(1), 473–482. <https://doi.org/10.1515/biol-2022-0058>
  287. Sanganna Gari, R. R., Tagiltsev, G., Pumroy, R. A., Jiang, Y., Blackledge, M., Moiseenkova-Bell, V. Y., & Scheuring, S. (2023). Intrinsically disordered regions in TRPV2 mediate protein-protein interactions. *Communications biology*, 6(1), 966. <https://doi.org/10.1038/s42003-023-05343-7>
  288. Uversky V. N. (2014). Introduction to intrinsically disordered proteins (IDPs). *Chemical reviews*, 114(13), 6557–6560. <https://doi.org/10.1021/cr500288y>
  289. Kelly, M. A., Caleshu, C., Morales, A., Buchan, J., Wolf, Z., Harrison, S. M., Cook, S., Dillon, M. W., Garcia, J., Haverfield, E., Jongbloed, J. D. H., Macaya, D., Manrai, A., Orland, K., Richard, G., Spoonamore, K., Thomas, M., Thomson, K., Vincent, L. M., Walsh, R., ... Funke, B. (2018). Adaptation and validation of the ACMG/AMP variant classification framework for MYH7-associated inherited cardiomyopathies: recommendations by ClinGen's Inherited Cardiomyopathy Expert Panel. *Genetics in medicine : official journal of the American College of Medical Genetics*, 20(3), 351–359. <https://doi.org/10.1038/gim.2017.218>
  290. Chung, I. M., & Rajakumar, G. (2016). Genetics of Congenital Heart Defects: The NKX2-5 Gene, a Key Player. *Genes*, 7(2), 6. <https://doi.org/10.3390/genes7020006>
  291. Moons, P., Sluysmans, T., De Wolf, D., Massin, M., Suys, B., Benatar, A., & Gewillig, M. (2009). Congenital heart disease in 111 225 births in Belgium: birth prevalence, treatment and survival in the 21st century. *Acta paediatrica (Oslo, Norway : 1992)*, 98(3), 472–477. <https://doi.org/10.1111/j.1651-2227.2008.01152.x>

292. Melim, C., Pimenta, J., & Areias, J. C. (2022). Congenital atrioventricular heart block: From diagnosis to treatment. *Revista portuguesa de cardiologia : orgao oficial da Sociedade Portuguesa de Cardiologia = Portuguese journal of cardiology : an official journal of the Portuguese Society of Cardiology*, 41(3), 231–240. <https://doi.org/10.1016/j.repc.2019.09.021>
293. Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., Bodenstein, S. W., Evans, D. A., Hung, C. C., O'Neill, M., Reiman, D., Tunyasuvunakool, K., Wu, Z., Žemgulytė, A., Arvaniti, E., Beattie, C., ... Jumper, J. M. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630(8016), 493–500. <https://doi.org/10.1038/s41586-024-07487-w>
294. Hu, W., & Ohue, M. (2024). SpatialPPI: Three-dimensional space protein-protein interaction prediction with AlphaFold Multimer. *Computational and structural biotechnology journal*, 23, 1214–1225. <https://doi.org/10.1016/j.csbj.2024.03.009>
295. Rao, K. S., Kameswaran, V., & Bruneau, B. G. (2022). Modeling congenital heart disease: lessons from mice, hPSC-based models, and organoids. *Genes & development*, 36(11-12), 652–663. <https://doi.org/10.1101/gad.349678.122>
296. Kinnunen, S., Välimäki, M., Tölli, M., Wohlfahrt, G., Darwich, R., Komati, H., Nemer, M., & Ruskoaho, H. (2015). Nuclear Receptor-Like Structure and Interaction of Congenital Heart Disease-Associated Factors GATA4 and NKX2-5. *PloS one*, 10(12), e0144145. <https://doi.org/10.1371/journal.pone.0144145>
297. Garg, V., Kathiriya, I. S., Barnes, R., Schluterman, M. K., King, I. N., Butler, C. A., Rothrock, C. R., Eapen, R. S., Hirayama-Yamada, K., Joo, K., Matsuoka, R., Cohen, J. C., & Srivastava, D. (2003). GATA4 mutations cause human congenital heart defects and reveal an interaction with TBX5. *Nature*, 424(6947), 443–447. <https://doi.org/10.1038/nature01827>
298. Canac, R., Cimarosti, B., Girardeau, A., Forest, V., Olchesqui, P., Poschmann, J., Redon, R., Lemarchand, P., Gaborit, N., & Lamirault, G. (2022).

- Deciphering Transcriptional Networks during Human Cardiac Development. *Cells*, 11(23), 3915. <https://doi.org/10.3390/cells11233915>
299. McCulley, D. J., & Black, B. L. (2012). Transcription factor pathways and congenital heart disease. *Current topics in developmental biology*, 100, 253–277. <https://doi.org/10.1016/B978-0-12-387786-4.00008-7>
  300. Ang, Y. S., Rivas, R. N., Ribeiro, A. J. S., Srivas, R., Rivera, J., Stone, N. R., Pratt, K., Mohamed, T. M. A., Fu, J. D., Spencer, C. I., Tippens, N. D., Li, M., Narasimha, A., Radzinsky, E., Moon-Grady, A. J., Yu, H., Pruitt, B. L., Snyder, M. P., & Srivastava, D. (2016). Disease Model of GATA4 Mutation Reveals Transcription Factor Cooperativity in Human Cardiogenesis. *Cell*, 167(7), 1734–1749.e22. <https://doi.org/10.1016/j.cell.2016.11.033>
  301. Qiao, X. H., Wang, F., Zhang, X. L., Huang, R. T., Xue, S., Wang, J., Qiu, X. B., Liu, X. Y., & Yang, Y. Q. (2017). MEF2C loss-of-function mutation contributes to congenital heart defects. *International journal of medical sciences*, 14(11), 1143–1153. <https://doi.org/10.7150/ijms.21353>
  302. Stennard, F. A., Costa, M. W., Elliott, D. A., Rankin, S., Haast, S. J., Lai, D., McDonald, L. P., Niederreither, K., Dolle, P., Bruneau, B. G., Zorn, A. M., & Harvey, R. P. (2003). Cardiac T-box factor Tbx20 directly interacts with Nkx2-5, GATA4, and GATA5 in regulation of gene expression in the developing heart. *Developmental biology*, 262(2), 206–224. [https://doi.org/10.1016/s0012-1606\(03\)00385-3](https://doi.org/10.1016/s0012-1606(03)00385-3)
  303. Sadaki, S., Fujita, R., Hayashi, T., Nakamura, A., Okamura, Y., Fuseya, S., Hamada, M., Warabi, E., Kuno, A., Ishii, A., Muratani, M., Okada, R., Shiba, D., Kudo, T., Takeda, S., & Takahashi, S. (2023). Large Maf transcription factor family is a major regulator of fast type IIb myofiber determination. *Cell reports*, 42(4), 112289. <https://doi.org/10.1016/j.celrep.2023.112289>
  304. Song, W., Wang, H., & Wu, Q. (2015). Atrial natriuretic peptide in cardiovascular biology and disease (NPPA). *Gene*, 569(1), 1–6. <https://doi.org/10.1016/j.gene.2015.06.029>
  305. Rubattu, S., Bigatti, G., Evangelista, A., Lanzani, C., Stanzione, R., Zagato, L., Manunta, P., Marchitti, S., Venturelli, V., Bianchi, G., Volpe, M., &

- Stella, P. (2006). Association of atrial natriuretic peptide and type a natriuretic peptide receptor gene polymorphisms with left ventricular mass in human essential hypertension. *Journal of the American College of Cardiology*, 48(3), 499–505. <https://doi.org/10.1016/j.jacc.2005.12.081>
306. Durocher, D., Chen, C. Y., Ardati, A., Schwartz, R. J., & Nemer, M. (1996). The atrial natriuretic factor promoter is a downstream target for Nkx-2.5 in the myocardium. *Molecular and cellular biology*, 16(9), 4648–4655. <https://doi.org/10.1128/MCB.16.9.4648>
307. Herman, G. E., & El-Hodiri, H. M. (2002). The role of ZIC3 in vertebrate development. *Cytogenetic and genome research*, 99(1-4), 229–235. <https://doi.org/10.1159/000071598>
308. Yamagata, K., Horie, M., Aiba, T., Ogawa, S., Aizawa, Y., Ohe, T., Yamagishi, M., Makita, N., Sakurada, H., Tanaka, T., Shimizu, A., Hagiwara, N., Kishi, R., Nakano, Y., Takagi, M., Makiyama, T., Ohno, S., Fukuda, K., Watanabe, H., Morita, H., ... Shimizu, W. (2017). Genotype-Phenotype Correlation of SCN5A Mutation for the Clinical and Electrocardiographic Characteristics of Proband With Brugada Syndrome: A Japanese Multicenter Registry. *Circulation*, 135(23), 2255–2270. <https://doi.org/10.1161/CIRCULATIONAHA.117.027983>
309. Kinoshita, K., Takahashi, H., Hata, Y., Nishide, K., Kato, M., Fujita, H., Yoshida, S., Murai, K., Mizumaki, K., Nishida, K., Yamaguchi, Y., Kano, M., Tabata, T., & Nishida, N. (2016). SCN5A(K817E), a novel Brugada syndrome-associated mutation that alters the activation gating of NaV1.5 channel. *Heart rhythm*, 13(5), 1113–1120. <https://doi.org/10.1016/j.hrthm.2016.01.008>
310. Darbar, D., Kannankeril, P. J., Donahue, B. S., Kucera, G., Stubblefield, T., Haines, J. L., George, A. L., Jr, & Roden, D. M. (2008). Cardiac sodium channel (SCN5A) variants associated with atrial fibrillation. *Circulation*, 117(15), 1927–1935. <https://doi.org/10.1161/CIRCULATIONAHA.107.757955>
311. Wang, D. W., Desai, R. R., Crotti, L., Arnestad, M., Insolia, R., Pedrazzini, M., Ferrandi, C., Vege, A., Rognum, T., Schwartz, P. J., & George, A. L., Jr (2007). Cardiac sodium channel dysfunction in sudden infant death syndrome.

- Circulation, 115(3), 368–376.  
<https://doi.org/10.1161/CIRCULATIONAHA.106.646513>
312. Johannessen, M., Møller, S., Hansen, T., Moens, U., & Van Ghelue, M. (2006). The multifunctional roles of the four-and-a-half-LIM only protein FHL2. Cellular and molecular life sciences : CMLS, 63(3), 268–284.  
<https://doi.org/10.1007/s00018-005-5438-z>
  313. Arimura, T., Hayashi, T., Matsumoto, Y., Shibata, H., Hiroi, S., Nakamura, T., Inagaki, N., Hinohara, K., Takahashi, M., Manatsu, S. I., Sasaoka, T., Izumi, T., Bonne, G., Schwartz, K., & Kimura, A. (2007). Structural analysis of four and half LIM protein-2 in dilated cardiomyopathy. Biochemical and biophysical research communications, 357(1), 162–167.  
<https://doi.org/10.1016/j.bbrc.2007.03.128>
  314. Purcell, N. H., Darwis, D., Bueno, O. F., Müller, J. M., Schüle, R., & Molkentin, J. D. (2004). Extracellular signal-regulated kinase 2 interacts with and is negatively regulated by the LIM-only protein FHL2 in cardiomyocytes. Molecular and cellular biology, 24(3), 1081–1095.  
<https://doi.org/10.1128/MCB.24.3.1081-1095.2004>
  315. Russell, T. M., & Richardson, D. R. (2022). Glutathione-S-Transferases as Potential Targets for Modulation of Nitric Oxide-Mediated Vasodilation. Biomolecules, 12(9), 1292. <https://doi.org/10.3390/biom12091292>
  316. Oliveira-Paula, G. H., Lacchini, R., & Tanus-Santos, J. E. (2016). Endothelial nitric oxide synthase: From biochemistry and gene structure to clinical implications of NOS3 polymorphisms. Gene, 575(2 Pt 3), 584–599.  
<https://doi.org/10.1016/j.gene.2015.09.061>
  317. Michel, T., & Vanhoutte, P. M. (2010). Cellular signaling and NO production. Pflugers Archiv : European journal of physiology, 459(6), 807–816.  
<https://doi.org/10.1007/s00424-009-0765-9>
  318. Kobashi, G., Yamada, H., Ohta, K., Kato, E., Ebina, Y., & Fujimoto, S. (2001). Endothelial nitric oxide synthase gene (NOS3) variant and hypertension in pregnancy. American journal of medical genetics, 103(3), 241–244.

319. Lim, H., & Zhu, Y. Z. (2006). Role of transforming growth factor-beta in the progression of heart failure. *Cellular and molecular life sciences : CMLS*, 63(22), 2584–2596. <https://doi.org/10.1007/s00018-006-6085-8>
320. Park, J., Tucci, A., Cipriani, V., Demidov, G., Rocca, C., Senderek, J., Butryn, M., Velic, A., Lam, T., Galanaki, E., Cali, E., Vestito, L., Maroofian, R., Deininger, N., Rautenberg, M., Admard, J., Hahn, G. A., Bartels, C., van Os, N. J. H., Horvath, R., ... Hengel, H. (2022). Heterozygous UCHL1 loss-of-function variants cause a neurodegenerative disorder with spasticity, ataxia, neuropathy, and optic atrophy. *Genetics in medicine : official journal of the American College of Medical Genetics*, 24(10), 2079–2090. <https://doi.org/10.1016/j.gim.2022.07.006>
321. Lowe, J., McDermott, H., Landon, M., Mayer, R. J., & Wilkinson, K. D. (1990). Ubiquitin carboxyl-terminal hydrolase (PGP 9.5) is selectively present in ubiquitinated inclusion bodies characteristic of human neurodegenerative diseases. *The Journal of pathology*, 161(2), 153–160. <https://doi.org/10.1002/path.1711610210>
322. Flevaris, P., & Vaughan, D. (2017). The Role of Plasminogen Activator Inhibitor Type-1 in Fibrosis. *Seminars in thrombosis and hemostasis*, 43(2), 169–177. <https://doi.org/10.1055/s-0036-1586228>
323. Asselbergs, F. W., Pattin, K., Snieder, H., Hillege, H. L., van Gilst, W. H., & Moore, J. H. (2008). Genetic architecture of tissue-type plasminogen activator and plasminogen activator inhibitor-1. *Seminars in thrombosis and hemostasis*, 34(6), 562–568. <https://doi.org/10.1055/s-0028-1103367>
324. Lin, M., Griessenauer, C. J., Starke, R. M., Tubbs, R. S., Shoja, M. M., Foreman, P. M., Vyas, N. A., Walters, B. C., Harrigan, M. R., Hendrix, P., Fisher, W. S., Pittet, J. F., Mathru, M., & Lipsky, R. H. (2019). Haplotype analysis of SERPINE1 gene: Risk for aneurysmal subarachnoid hemorrhage and clinical outcomes. *Molecular genetics & genomic medicine*, 7(8), e737. <https://doi.org/10.1002/mgg3.737>

325. Mehta, R., & Shapiro, A. D. (2008). Plasminogen activator inhibitor type 1 deficiency. *Haemophilia* : the official journal of the World Federation of Hemophilia, 14(6), 1255–1260. <https://doi.org/10.1111/j.1365-2516.2008.01834.x>
326. Genestine, M., Ambriz, D., Crabtree, G. W., Dummer, P., Molotkova, A., Quintero, M., Mela, A., Biswas, S., Feng, H., Zhang, C., Canoll, P., Hargus, G., Agalliu, D., Gogos, J. A., & Au, E. (2021). Vascular-derived SPARC and SerpinE1 regulate interneuron tangential migration and accelerate functional maturation of human stem cell-derived interneurons. *eLife*, 10, e56063. <https://doi.org/10.7554/eLife.56063>
327. Planus, E., Barlovatz-Meimon, G., Rogers, R. A., Bonavaud, S., Ingber, D. E., & Wang, N. (1997). Binding of urokinase to plasminogen activator inhibitor type-1 mediates cell adhesion and spreading. *Journal of cell science*, 110 ( Pt 9), 1091–1098. <https://doi.org/10.1242/jcs.110.9.1091>
328. Trainor P. A. (2014). *Neural Crest Cells: Evolution, Development, and Disease*, Amsterdam; Boston, Elsevier/Ap, Academic Press is an imprint of Elsevier.
329. George, R. M., Maldonado-Velez, G., & Firulli, A. B. (2020). The heart of the neural crest: cardiac neural crest cells in development and regeneration. *Development* (Cambridge, England), 147(20), dev188706. <https://doi.org/10.1242/dev.188706>
330. Galang, G., Mandla, R., Ruan, H., Jung, C., Sinha, T., Stone, N. R., Wu, R. S., Mannion, B. J., Allu, P. K. R., Chang, K., Rammohan, A., Shi, M. B., Pennacchio, L. A., Black, B. L., & Vedantham, V. (2020). ATAC-Seq Reveals an Isl1 Enhancer That Regulates Sinoatrial Node Development and Function. *Circulation research*, 127(12), 1502–1518. <https://doi.org/10.1161/CIRCRESAHA.120.317145>
331. Liang, X., Zhang, Q., Cattaneo, P., Zhuang, S., Gong, X., Spann, N. J., Jiang, C., Cao, X., Zhao, X., Zhang, X., Bu, L., Wang, G., Chen, H. S., Zhuang, T., Yan, J., Geng, P., Luo, L., Banerjee, I., Chen, Y., Glass, C. K., ... Evans, S. M. (2015). Transcription factor ISL1 is essential for pacemaker development and

- function. The Journal of clinical investigation, 125(8), 3256–3268.  
<https://doi.org/10.1172/JCI68257>
332. Zhang, W., Zhao, H., Quan, D., Tang, Y., Wang, X., & Huang, C. (2022). Tbx18 promoted the conversion of human-induced pluripotent stem cell-derived cardiomyocytes into sinoatrial node-like pacemaker cells. Cell biology international, 46(3), 403–414. <https://doi.org/10.1002/cbin.11738>
  333. Wu, S. P., Dong, X. R., Regan, J. N., Su, C., & Majesky, M. W. (2013). Tbx18 regulates development of the epicardium and coronary vessels. Developmental biology, 383(2), 307–320.  
<https://doi.org/10.1016/j.ydbio.2013.08.019>
  334. Wiese, C., Grieskamp, T., Airik, R., Mommersteeg, M. T., Gardiwal, A., de Gier-de Vries, C., Schuster-Gossler, K., Moorman, A. F., Kispert, A., & Christoffels, V. M. (2009). Formation of the sinus node head and differentiation of sinus node myocardium are independently regulated by Tbx18 and Tbx3. Circulation research, 104(3), 388–397.  
<https://doi.org/10.1161/CIRCRESAHA.108.187062>
  335. Rahman, F., Kwan, G. F., & Benjamin, E. J. (2014). Global epidemiology of atrial fibrillation. Nature reviews. Cardiology, 11(11), 639–654.  
<https://doi.org/10.1038/nrcardio.2014.118>
  336. Friberg, L., & Bergfeldt, L. (2013). Atrial fibrillation prevalence revisited. Journal of internal medicine, 274(5), 461–468. <https://doi.org/10.1111/joim.12114>
  337. Colilla, S., Crow, A., Petkun, W., Singer, D. E., Simon, T., & Liu, X. (2013). Estimates of current and future incidence and prevalence of atrial fibrillation in the U.S. adult population. The American journal of cardiology, 112(8), 1142–1147. <https://doi.org/10.1016/j.amjcard.2013.05.063>
  338. Reichart, D., Magnussen, C., Zeller, T., & Blankenberg, S. (2019). Dilated cardiomyopathy: from epidemiologic to genetic phenotypes: A translational review of current literature. Journal of internal medicine, 286(4), 362–372.  
<https://doi.org/10.1111/joim.12944>
  339. Yeung, C., Enriquez, A., Suarez-Fuster, L., & Baranchuk, A. (2019). Atrial fibrillation in patients with inherited cardiomyopathies. Europace : European

pacing, arrhythmias, and cardiac electrophysiology : journal of the working groups on cardiac pacing, arrhythmias, and cardiac cellular electrophysiology of the European Society of Cardiology, 21(1), 22–32.  
<https://doi.org/10.1093/europace/euy064>

340. Penha, R. C. C., Buexm, L. A., Rodrigues, F. R., de Castro, T. P., Santos, M. C. S., Fortunato, R. S., Carvalho, D. P., Cardoso-Weide, L. C., & Ferreira, A. C. F. (2018). NKX2.5 is expressed in papillary thyroid carcinomas and regulates differentiation in thyroid cells. BMC cancer, 18(1), 498.  
<https://doi.org/10.1186/s12885-018-4399-1>
341. Tanaka, M., Schinke, M., Liao, H. S., Yamasaki, N., & Izumo, S. (2001). Nkx2.5 and Nkx2.6, homologs of Drosophila tinman, are required for development of the pharynx. Molecular and cellular biology, 21(13), 4391–4398.  
<https://doi.org/10.1128/MCB.21.13.4391-4398.2001>

## **8. Anexo**

## ANEXO I. Enlaces externos

Anexo I.1. Variantes génicas en NKX2-5.

[https://docs.google.com/spreadsheets/d/1Dx\\_muV-QJayOwDYgFyYx5x4zkThtz0s4/](https://docs.google.com/spreadsheets/d/1Dx_muV-QJayOwDYgFyYx5x4zkThtz0s4/)

Anexo I.2. Variantes génicas clasificadas como patogénicas por publicaciones y/o bases de datos.

<https://docs.google.com/spreadsheets/d/1xfCj9MST1tfAZVTjtvZrd8u17VWKLKun/>

Anexo I.3. Clasificación según ACMG de variantes seleccionadas.

<https://docs.google.com/spreadsheets/d/1xso0pCMvARB4hMlfhb8noMYveLXp8IN9/>

Anexo I.4. Resultados de estudios *in silico* para cada variante génica generada en el homeodominio de NKX2-5.

[https://docs.google.com/spreadsheets/d/1uCtr4UYW0SAx7J7Tc9XUpRXKGGRxCSO\\_ /](https://docs.google.com/spreadsheets/d/1uCtr4UYW0SAx7J7Tc9XUpRXKGGRxCSO_/)

Anexo I.5. Resultados de caracterización de picos de ChIP-seq con confirmación de genes por RNA-seq con *knock out* de NKX2-5.

A) Tabla de resultados en humano [207]:

[https://docs.google.com/spreadsheets/d/1HIRxUpOUNKCkBMtOZKI41Bv2JL4cF\\_5JuO4KqOgWIB4/](https://docs.google.com/spreadsheets/d/1HIRxUpOUNKCkBMtOZKI41Bv2JL4cF_5JuO4KqOgWIB4/)

B) Tabla de resultados en ratón [206]:

<https://docs.google.com/spreadsheets/d/1M2AgfMM8VBjSH--fs8DbpZG6y5FLlwmMIPpwGbriSbl/>

Anexo I.6. Sitios de unión de NKX2-5 encontrados en picos de ChIP-seq de Anderson y col., 2018 [207] a menos de 100 kpb de genes con expresión diferencial en presencia de NKX2-5.

<https://docs.google.com/spreadsheets/d/1axe9O8P7qdxYQjo9jWXEeFNLunz54pMWk6omaJ1hWtl/>

Anexo I.7. Resultados completos del análisis con MEME-ChIP.

<https://docs.google.com/spreadsheets/d/1IGsFg4hZglY206bTWzBSJOGVWmLmPmx3c4AnzkK3SOY/>

Anexo I.8. Sitios de unión de NKX2-5 confirmados experimentalmente y sitios de otros factores de transcripción a menos de 250 pares de bases de distancia.

[https://docs.google.com/spreadsheets/d/1qbrPa8XJPLs4UPZT1xlw1W6-h\\_e3czvwNIH5GUZE9c4/](https://docs.google.com/spreadsheets/d/1qbrPa8XJPLs4UPZT1xlw1W6-h_e3czvwNIH5GUZE9c4/)

## ANEXO II. Tablas y figuras suplementarias

**Tabla S1. Motivos lineales predichos por Eukaryotic Linear Motif en NKX2-5.**

| Identificador de ELM | Posición en la proteína | Función del SLiM            | Proteína o dominio que lo reconoce |
|----------------------|-------------------------|-----------------------------|------------------------------------|
| CLV_C14_Caspase3-7   | 125-129                 | Corte                       | Caspasa                            |
| DOC_CYCLIN_RXL_1     | 136-146                 | Fosforilación               | Ciclina/CDK                        |
| DOC_CYCLIN_RXL_1     | 137-146                 | Fosforilación               | Ciclina/CDK                        |
| DOC_CYCLIN_RXL_1     | 138-146                 | Fosforilación               | Ciclina/CDK                        |
| DOC_CYCLIN_RXL_1     | 139-146                 | Fosforilación               | Ciclina/CDK                        |
| DOC_MAPK_gen_1       | 136-145                 | Fosforilación               | MAPK                               |
| DOC_MAPK_gen_1       | 137-145                 | Fosforilación               | MAPK                               |
| DOC_MAPK_gen_1       | 138-145                 | Fosforilación               | MAPK                               |
| DOC_MAPK_gen_1       | 139-145                 | Fosforilación               | MAPK                               |
| DOC_MAPK_gen_1       | 194-204                 | Fosforilación               | MAPK                               |
| DOC_MAPK_gen_1       | 195-204                 | Fosforilación               | MAPK                               |
| DOC_MAPK_gen_1       | 216-224                 | Fosforilación               | MAPK                               |
| DOC_MAPK_MEF2A_6     | 136-145                 | Fosforilación               | MAPK                               |
| DOC_MAPK_MEF2A_6     | 137-145                 | Fosforilación               | MAPK                               |
| DOC_MAPK_MEF2A_6     | 138-145                 | Fosforilación               | MAPK                               |
| DOC_MAPK_MEF2A_6     | 216-224                 | Fosforilación               | MAPK                               |
| DOC_MAPK_MEF2A_6     | 217-224                 | Fosforilación               | MAPK                               |
| DOC_PP1_RVXF_1       | 140-146                 | Desfosforilación            | PP1                                |
| DOC_PP2B_LxvP_1      | 163-166                 | Desfosforilación            | Calcineurinas                      |
| DOC_PP2B_LxvP_1      | 207-210                 | Desfosforilación            | Calcineurinas                      |
| DOC_PP4_FxxP_1       | 2-5                     | Fosforilación               | Dominio EVH1                       |
| DOC_WW_Pin1_4        | 1-6                     | Prolil-isomerización        | Pin1                               |
| DOC_WW_Pin1_4        | 5-10                    | Prolil-isomerización        | Pin1                               |
| DOC_WW_Pin1_4        | 7-12                    | Prolil-isomerización        | Pin1                               |
| DOC_WW_Pin1_4        | 75-80                   | Prolil-isomerización        | Pin1                               |
| DOC_WW_Pin1_4        | 262-267                 | Prolil-isomerización        | Pin1                               |
| DOC_WW_Pin1_4        | 276-281                 | Prolil-isomerización        | Pin1                               |
| DOC_WW_Pin1_4        | 302-307                 | Prolil-isomerización        | Pin1                               |
| LIG_14-3-3_CanoR_1   | 95-100                  | Retención de proteínas      | 14-3-3                             |
| LIG_14-3-3_CanoR_1   | 142-151                 | Retención de proteínas      | 14-3-3                             |
| LIG_14-3-3_CanoR_1   | 161-166                 | Retención de proteínas      | 14-3-3                             |
| LIG_ANK_PxLPxL_1     | 64-73                   | Interacción entre complejos | ANKRA2                             |
| LIG_EH1_1            | 11-19                   | Unión de correpressor       | Groucho/TLE                        |
| LIG_FHA_1            | 199-205                 | Diversas funciones          | FHA                                |
| LIG_FHA_2            | 123-129                 | Diversas funciones          | FHA                                |
| LIG_MYND_1           | 5-9                     | Modificación de histonas    | Dominio MYND                       |

|                     |         |                               |               |
|---------------------|---------|-------------------------------|---------------|
| LIG_SUMO_SIM_anti_2 | 201-206 | SUMOylación                   | SUMO          |
| LIG_SUMO_SIM_par_1  | 220-226 | SUMOylación                   | SUMO          |
| MOD_CDK_SPxK_1      | 75-81   | Fosforilación                 | CDK           |
| MOD_CK1_1           | 313-319 | Fosforilación                 | CK1           |
| MOD_CK2_1           | 161-167 | Fosforilación                 | CK2           |
| MOD_DYRK1A_RPxSP_1  | 75-79   | Fosforilación                 | DYRK1A        |
| MOD_GSK3_1          | 1-8     | Fosforilación                 | GSK3          |
| MOD_GSK3_1          | 37-44   | Fosforilación                 | GSK3          |
| MOD_GSK3_1          | 262-269 | Fosforilación                 | GSK3          |
| MOD_GSK3_1          | 310-317 | Fosforilación                 | GSK3          |
| MOD_NEK2_1          | 37-42   | Fosforilación                 | NEK2          |
| MOD_NEK2_1          | 175-180 | Fosforilación                 | NEK2          |
| MOD_NEK2_1          | 177-182 | Fosforilación                 | NEK2          |
| MOD_NEK2_1          | 231-236 | Fosforilación                 | NEK2          |
| MOD_NEK2_2          | 314-319 | Fosforilación                 | NEK2          |
| MOD_PIKK_1          | 143-149 | Fosforilación                 | PIKK          |
| MOD_Plk_1           | 37-43   | Fosforilación                 | Plk1          |
| MOD_Plk_1           | 198-204 | Fosforilación                 | Plk1          |
| MOD_ProDKin_1       | 1-7     | Fosforilación                 | MAPK          |
| MOD_ProDKin_1       | 5-11    | Fosforilación                 | MAPK          |
| MOD_ProDKin_1       | 7-13    | Fosforilación                 | MAPK          |
| MOD_ProDKin_1       | 75-81   | Fosforilación                 | MAPK          |
| MOD_ProDKin_1       | 262-268 | Fosforilación                 | MAPK          |
| MOD_ProDKin_1       | 276-282 | Fosforilación                 | MAPK          |
| MOD_ProDKin_1       | 302-308 | Fosforilación                 | MAPK          |
| MOD_SUMO_for_1      | 51-54   | Fosforilación                 | SUMO-1        |
| MOD_SUMO_rev_2      | 105-113 | Fosforilación                 | SUMO-1        |
| TRG_NES_CRM1_1      | 112-126 | Señal de exportación nuclear  |               |
| TRG_NLS_MonoCore_2  | 138-143 | Señal de localización nuclear |               |
| TRG_NLS_MonoExtC_3  | 193-199 | Señal de localización nuclear |               |
| TRG_NLS_MonoExtN_4  | 136-143 | Señal de localización nuclear |               |
| TRG_NLS_MonoExtN_4  | 137-143 | Señal de localización nuclear |               |
| TRG_NLS_MonoExtN_4  | 138-143 | Señal de localización nuclear |               |
| TRG_NLS_MonoExtN_4  | 192-198 | Señal de localización nuclear |               |
| LIG_EVH1_1          | 207-211 | Reconocimiento de actinas     | Dominio EVH1  |
| LIG_IBAR_NPY_1      | 246-248 | Polimerización de actinas     | Dominio I-BAR |
| LIG_PROFILIN_1      | 208-214 | Ensamblaje de filamentos      | Profilina     |
| LIG_SH2_NCK_1       | 97-101  | Transducción de señales       | Dominio SH2   |
| LIG_SH2_NCK_1       | 237-241 | Transducción de señales       | Dominio SH2   |
| LIG_SH2_SRC         | 237-240 | Transducción de señales       | Dominio SH2   |
| LIG_SH2_STAT5       | 162-165 | Transducción de señales       | Dominio SH2   |
| LIG_SH3_2           | 211-216 | Transducción de señales       | Dominio SH3   |

|                      |         |                                                  |             |
|----------------------|---------|--------------------------------------------------|-------------|
| LIG_SH3_3            | 61-67   | Transducción de señales                          | Dominio SH3 |
| LIG_SH3_3            | 84-90   | Transducción de señales                          | Dominio SH3 |
| LIG_SH3_3            | 202-208 | Transducción de señales                          | Dominio SH3 |
| LIG_SH3_3            | 205-211 | Transducción de señales                          | Dominio SH3 |
| LIG_SH3_3            | 206-212 | Transducción de señales                          | Dominio SH3 |
| LIG_SH3_3            | 207-213 | Transducción de señales                          | Dominio SH3 |
| LIG_SH3_3            | 208-214 | Transducción de señales                          | Dominio SH3 |
| LIG_SH3_3            | 233-239 | Transducción de señales                          | Dominio SH3 |
| LIG_SH3_3            | 251-257 | Transducción de señales                          | Dominio SH3 |
| LIG_SH3_3            | 272-278 | Transducción de señales                          | Dominio SH3 |
| LIG_SH3_3            | 277-283 | Transducción de señales                          | Dominio SH3 |
| LIG_SH3_3            | 300-306 | Transducción de señales                          | Dominio SH3 |
| LIG_SH3_3            | 303-309 | Transducción de señales                          | Dominio SH3 |
| TRG_ER_diArg_1       | 133-136 | Señal de localización en retículo endoplasmático |             |
| TRG_ER_diArg_1       | 134-137 | Señal de localización en retículo endoplasmático |             |
| TRG_ER_diArg_1       | 136-139 | Señal de localización en retículo endoplasmático |             |
| TRG_ER_diArg_1       | 137-139 | Señal de localización en retículo endoplasmático |             |
| TRG_ER_diArg_1       | 216-218 | Señal de localización en retículo endoplasmático |             |
| TRG_LysEnd_APsAcLL_1 | 200-205 | Señal de localización transmembrana              |             |

ELM: Eukaryotic Linear Motif.

**Tabla S2. Clasificación de efecto predicho jerarquizado de variantes génicas patogénicas de cambio de sentido en el homeodominio de NKX2-5.**

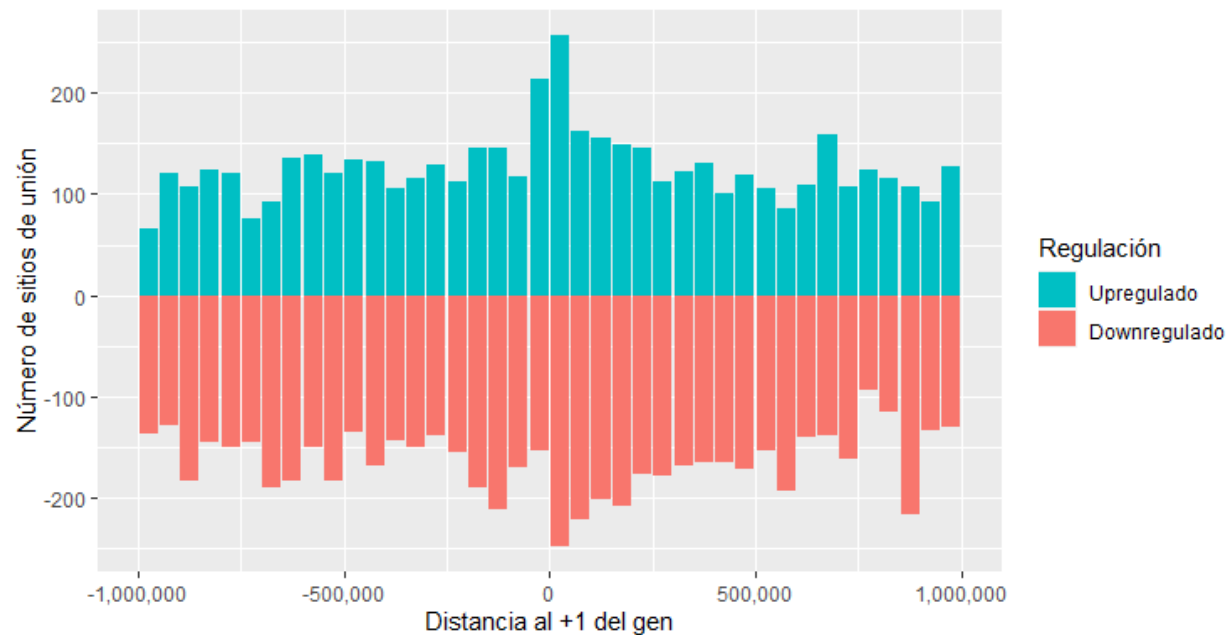
| Variante    | Fenotipos asociados                   | Frecuencia poblacional  | Efecto predicho                      |
|-------------|---------------------------------------|-------------------------|--------------------------------------|
| p.Arg139Trp | Cardiopatías                          | VG no encontrada en BDP | Motivo lineal confirmado             |
| p.Pro141Ala | Cardiopatías                          | Menor a 1/10.000        | Motivo lineal confirmado             |
| p.Arg142Cys | Cardiopatías                          | VG no encontrada en BDP | Interacción con el ADN               |
| p.Leu144Pro | Cardiopatías                          | 0,465%                  | Estabilidad o motivo lineal predicho |
| p.Phe145Leu | Cardiopatías                          | VG no encontrada en BDP | Estabilidad o motivo lineal predicho |
| p.Phe145Ser | Cardiopatías                          | Menor a 1/10.000        | Estabilidad o motivo lineal predicho |
| p.Ser146Trp | Cardiopatías                          | VG no encontrada en BDP | Estabilidad o motivo lineal predicho |
| p.Ala148Glu | Cardiopatías                          | VG no encontrada en BDP | Sin efecto predicho                  |
| p.Val150Ile | Cardiopatías                          | Menor a 1/10.000        | Estabilidad o motivo lineal predicho |
| p.Glu154Gly | Cardiopatías                          | VG no encontrada en BDP | Estabilidad o motivo lineal predicho |
| p.Gln160Pro | Cardiopatías                          | VG no encontrada en BDP | Estabilidad o motivo lineal predicho |
| p.Arg161Pro | Disgénesis de tiroides                | Menor a 1/10.000        | Estabilidad o motivo lineal predicho |
| p.Leu163Arg | Cardiopatías                          | 0,443%                  | Estabilidad o motivo lineal predicho |
| p.Glu167Lys | Cardiopatías                          | VG no encontrada en BDP | Motivo lineal confirmado             |
| p.Gln170Pro | Cardiopatías                          | VG no encontrada en BDP | Estabilidad o motivo lineal predicho |
| p.Leu171Arg | Cardiopatías                          | VG no encontrada en BDP | Estabilidad o motivo lineal predicho |
| p.Leu171Pro | Cardiopatías                          | VG no encontrada en BDP | Estabilidad o motivo lineal predicho |
| p.Thr178Met | Cardiopatías                          | VG no encontrada en BDP | Sin efecto predicho                  |
| p.Ser179Phe | Cardiopatías                          | VG no encontrada en BDP | Sin efecto predicho                  |
| p.Thr180Ala | Cardiopatías                          | VG no encontrada en BDP | Sin efecto predicho                  |
| p.Gln181His | Cardiopatías y disgénesis de tiroides | VG no encontrada en BDP | Interacción con el ADN               |
| p.Lys183Glu | Cardiopatías                          | VG no encontrada en BDP | Sin efecto predicho                  |
| p.Ile184Met | Cardiopatías                          | VG no encontrada en BDP | Sin efecto predicho                  |
| p.Trp185Leu | Cardiopatías                          | VG no encontrada en BDP | Estabilidad o motivo lineal predicho |
| p.Phe186Ser | Cardiopatías                          | VG no encontrada en BDP | Estabilidad o motivo lineal predicho |
| p.Gln187His | Cardiopatías                          | VG no encontrada en BDP | Interacción con el ADN               |
| p.Asn188Lys | Cardiopatías                          | VG no encontrada en BDP | Interacción con el ADN               |
| p.Arg189Gly | Cardiopatías                          | VG no encontrada en BDP | Estabilidad o motivo lineal predicho |
| p.Arg189Pro | Cardiopatías                          | VG no encontrada en BDP | Estabilidad o motivo lineal predicho |
| p.Arg190His | Cardiopatías                          | VG no encontrada en BDP | Interacción con el ADN               |
| p.Arg190Leu | Cardiopatías                          | VG no encontrada en BDP | Interacción con el ADN               |
| p.Arg190Cys | Cardiopatías                          | Menor a 1/10.000        | Introducción de cisteína             |
| p.Tyr191Cys | Cardiopatías                          | VG no encontrada en BDP | Interacción con el ADN               |
| p.Lys192Arg | Cardiopatías                          | VG no encontrada en BDP | Motivo lineal confirmado             |
| p.Lys192Thr | Cardiopatías                          | VG no encontrada en BDP | Motivo lineal confirmado             |
| p.Lys194Arg | Cardiopatías                          | VG no encontrada en BDP | Motivo lineal confirmado             |

VG: Variante génica. BDP: Base de datos poblacionales.

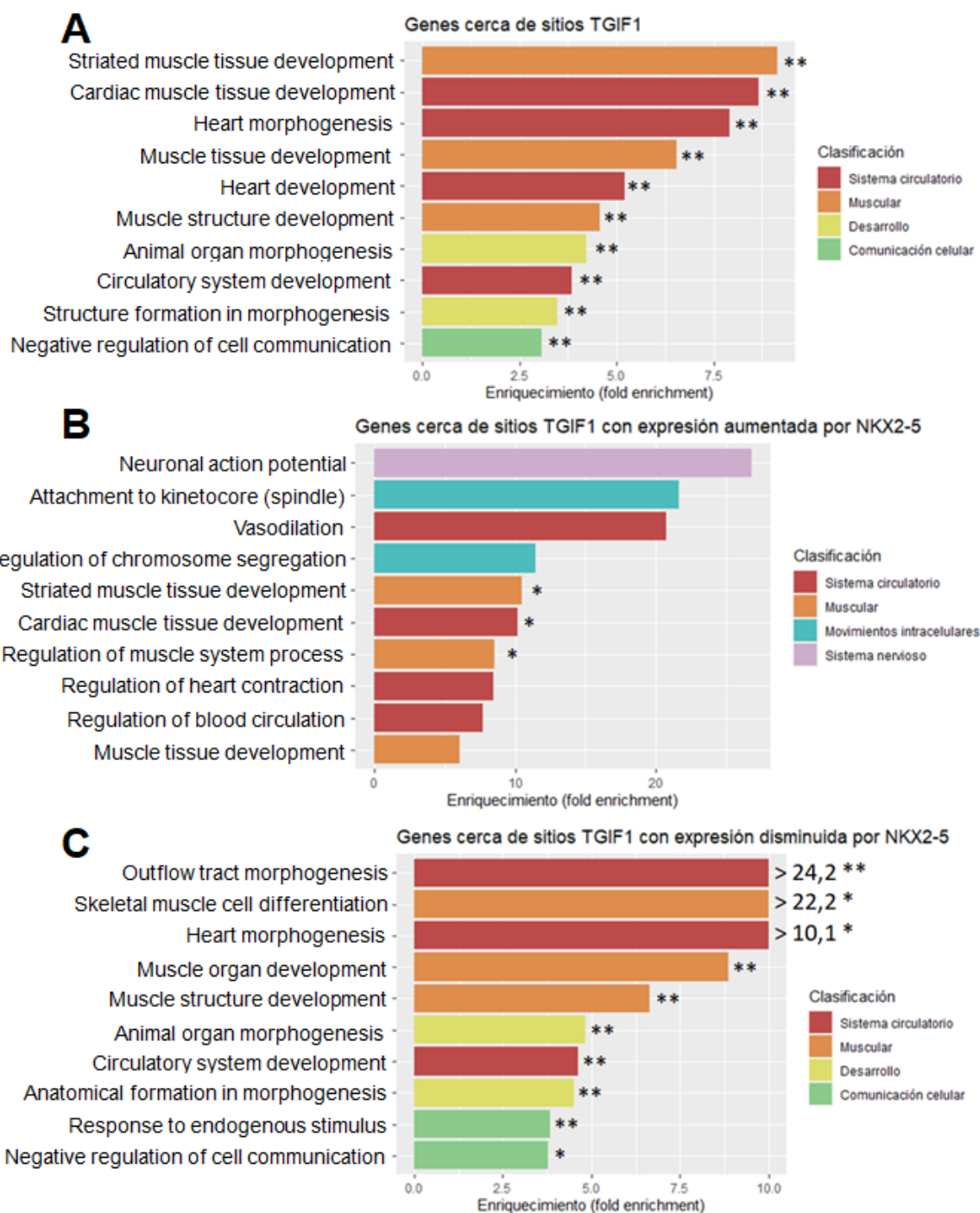
**Tabla S3. Primeros 10 componentes principales de distintas selecciones de datos para sitios de unión y sus distancias a genes con expresión diferencial en ensayo de RNA-seq con *knock out* de *NKX2-5*.**

| Distancia             | Fold change  | PC1   | PC2   | PC3   | PC4   | PC5   | PC6   | PC7   | PC8   | PC9   | PC10  |
|-----------------------|--------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| Inversa               | Normal       | 0.051 | 0.050 | 0.049 | 0.049 | 0.048 | 0.048 | 0.048 | 0.048 | 0.048 | 0.048 |
| Inversa               | Simplificada | 0.050 | 0.050 | 0.049 | 0.049 | 0.048 | 0.048 | 0.048 | 0.048 | 0.048 | 0.048 |
| Inversa del logaritmo | Normal       | 0.050 | 0.050 | 0.049 | 0.049 | 0.048 | 0.048 | 0.048 | 0.048 | 0.048 | 0.048 |
| Inversa del logaritmo | Simplificada | 0.051 | 0.050 | 0.049 | 0.048 | 0.048 | 0.048 | 0.048 | 0.048 | 0.048 | 0.048 |
| Logaritmo             | Normal       | 0.050 | 0.050 | 0.049 | 0.049 | 0.048 | 0.048 | 0.048 | 0.048 | 0.048 | 0.048 |
| Logaritmo             | Simplificada | 0.050 | 0.049 | 0.049 | 0.048 | 0.048 | 0.048 | 0.048 | 0.048 | 0.048 | 0.048 |
| Normal                | Normal       | 0.051 | 0.050 | 0.049 | 0.049 | 0.048 | 0.048 | 0.048 | 0.048 | 0.048 | 0.048 |
| Normal                | Simplificada | 0.050 | 0.050 | 0.049 | 0.048 | 0.048 | 0.048 | 0.048 | 0.048 | 0.048 | 0.048 |

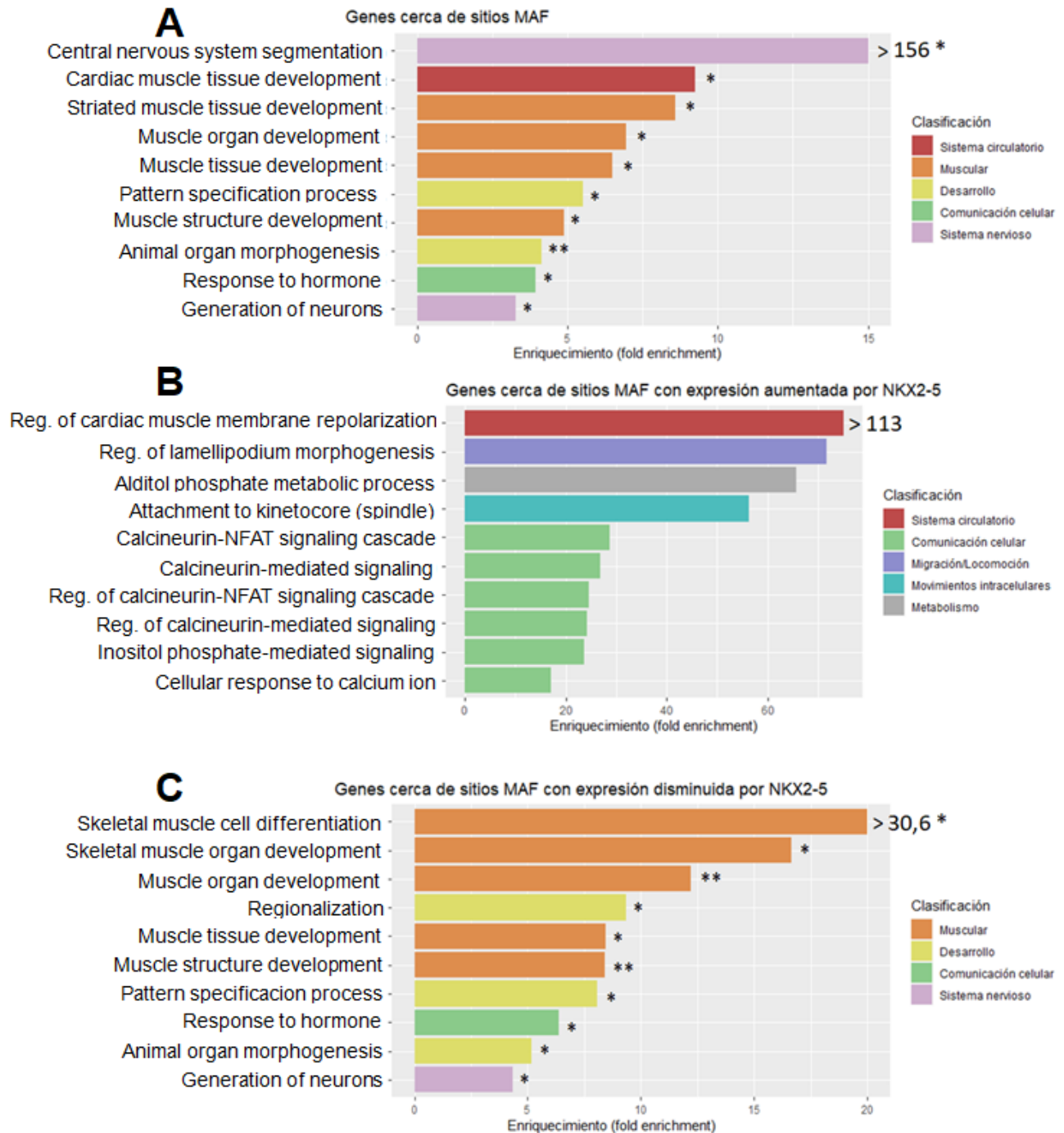
La columna “Distancia máxima” refiere la distancia máxima a la cual se buscó un gen cerca del sitio de unión. La columna “Distancia” refiere la transformación que se hizo al valor de distancia desde el sitio de unión. “Inversa”: 1/distancia. “Logaritmo”: signo(distancia)\*log(valor absoluto(distancia)). “Inversa del logaritmo”: 1/Logaritmo. “Normal”: Se utilizó el valor sin modificar. La columna “Fold change” refiere la transformación que se hizo al valor de fold change. “Simplificada”: Se utilizó -1 para valores menores a 0 y 1 para valores mayores a 0. “Normal”: Se utilizó el valor sin modificar.



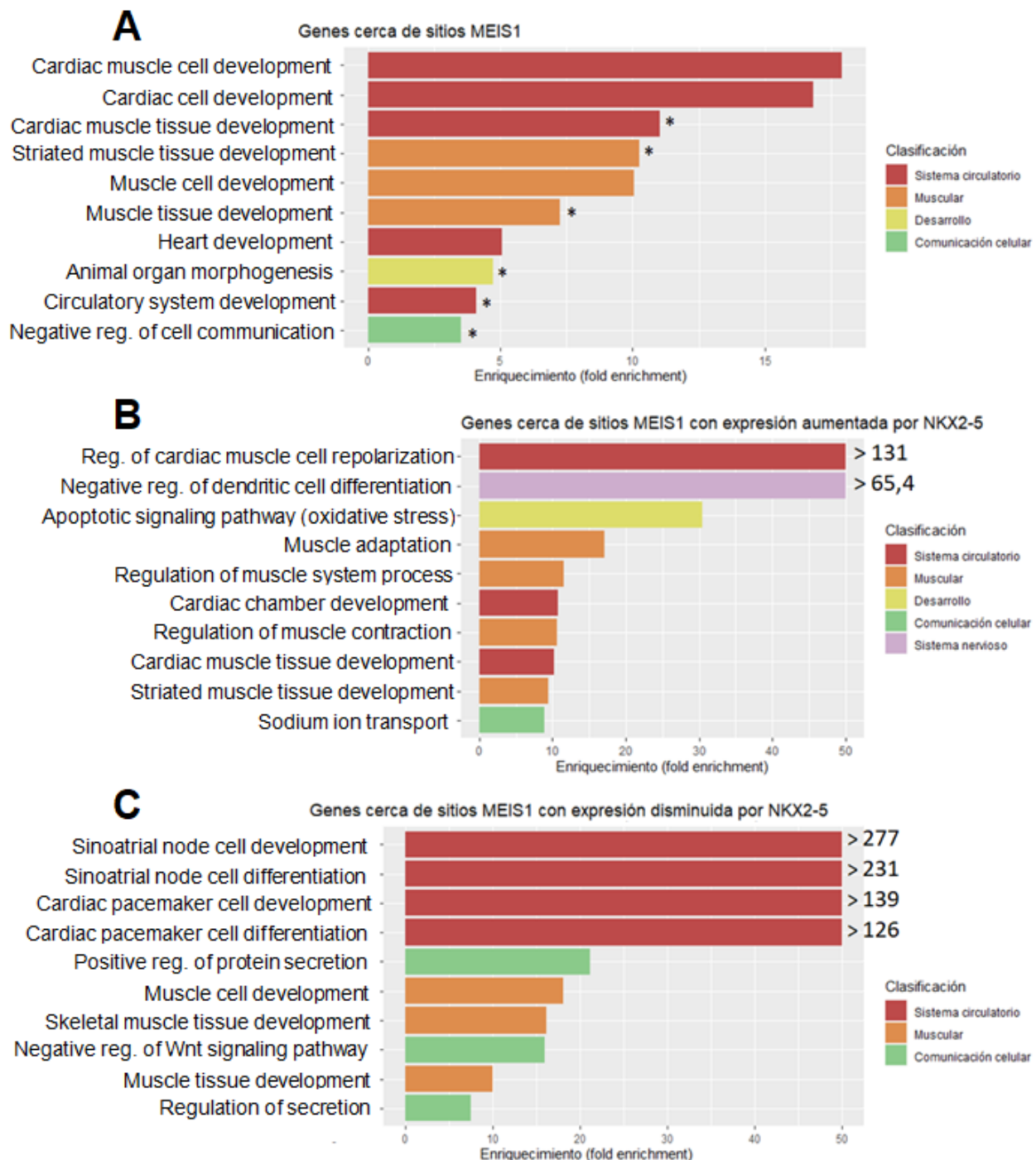
**Figura S1. Cantidad de sitios de unión por rangos de distancia al +1 del gen más cercano hasta una distancia de 1.000.000 de pares de bases.** En celeste, número de sitios con genes cercanos que aumentan su expresión en presencia de *NKX2-5* (upregulados). En naranja, número de sitios con genes cercanos que disminuyen su expresión en presencia de *NKX2-5* (downregulados).



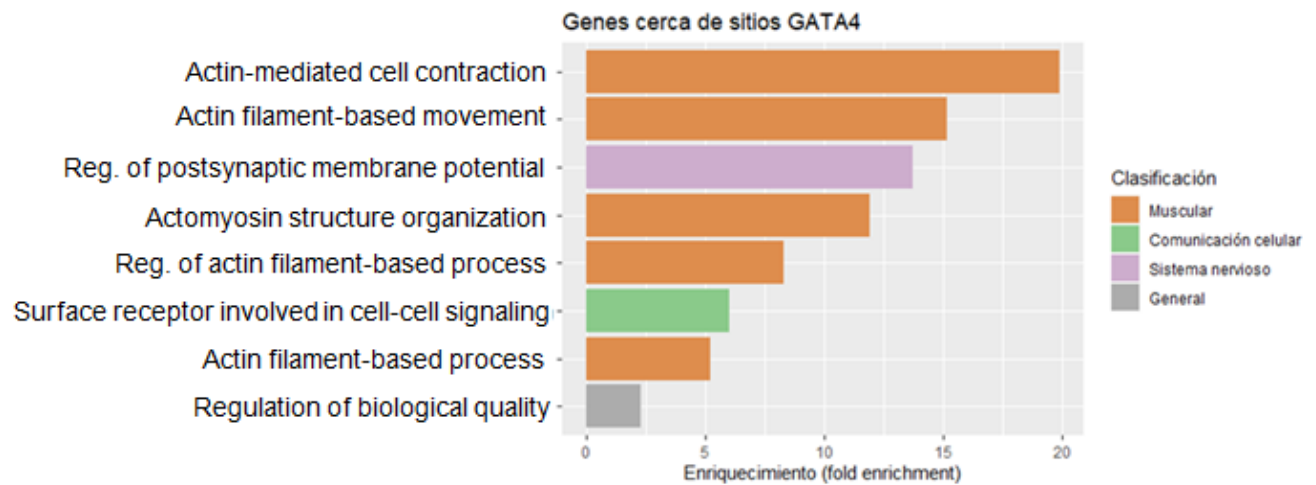
**Figura S2. GO-terms sobrerrepresentados de genes cerca de sitios de unión de NKX2-5 confirmados experimentalmente, seleccionados por presencia de sitios de unión de TGIF1.** El eje X representa el enriquecimiento relativo de cada GO-term con respecto al esperado por azar. Los valores extremos fueron recortados y anotados al costado de las columnas correspondientes. **A**, GO-terms sobrerrepresentados en la lista de todos los genes. **B**, GO-terms sobrerrepresentados en la lista de sólo los genes con expresión disminuida por NKX2-5 (downregulados). **C**, GO-terms sobrerrepresentados en la lista de sólo los genes con expresión aumentada por NKX2-5 (upregulados). Los asteriscos denotan la probabilidad de falso positivo (FDR). \*: FDR<0,001; \*\*: FDR<0,00001; \*\*\*: FDR<10<sup>-10</sup>.



**Figura S3. GO-terms sobrerrepresentados de genes cerca de sitios de unión de NKX2-5 confirmados experimentalmente, seleccionados por presencia de sitios de unión de MAF.** El eje X representa el enriquecimiento relativo de cada GO-term con respecto al esperado por azar. Los valores extremos fueron recortados y anotados al costado de las columnas correspondientes. **A**, GO-terms sobrerrepresentados en la lista de todos los genes. **B**, GO-terms sobrerrepresentados sólo de los genes con expresión disminuida por NKX2-5 (downregulados). **C**, GO-terms sobrerrepresentados sólo de los genes con expresión aumentada por NKX2-5 (upregulados). Los asteriscos denotan la probabilidad de falso positivo (FDR). \*: FDR<0,001; \*\*: FDR<0,00001; \*\*\*: FDR<10<sup>-10</sup>.



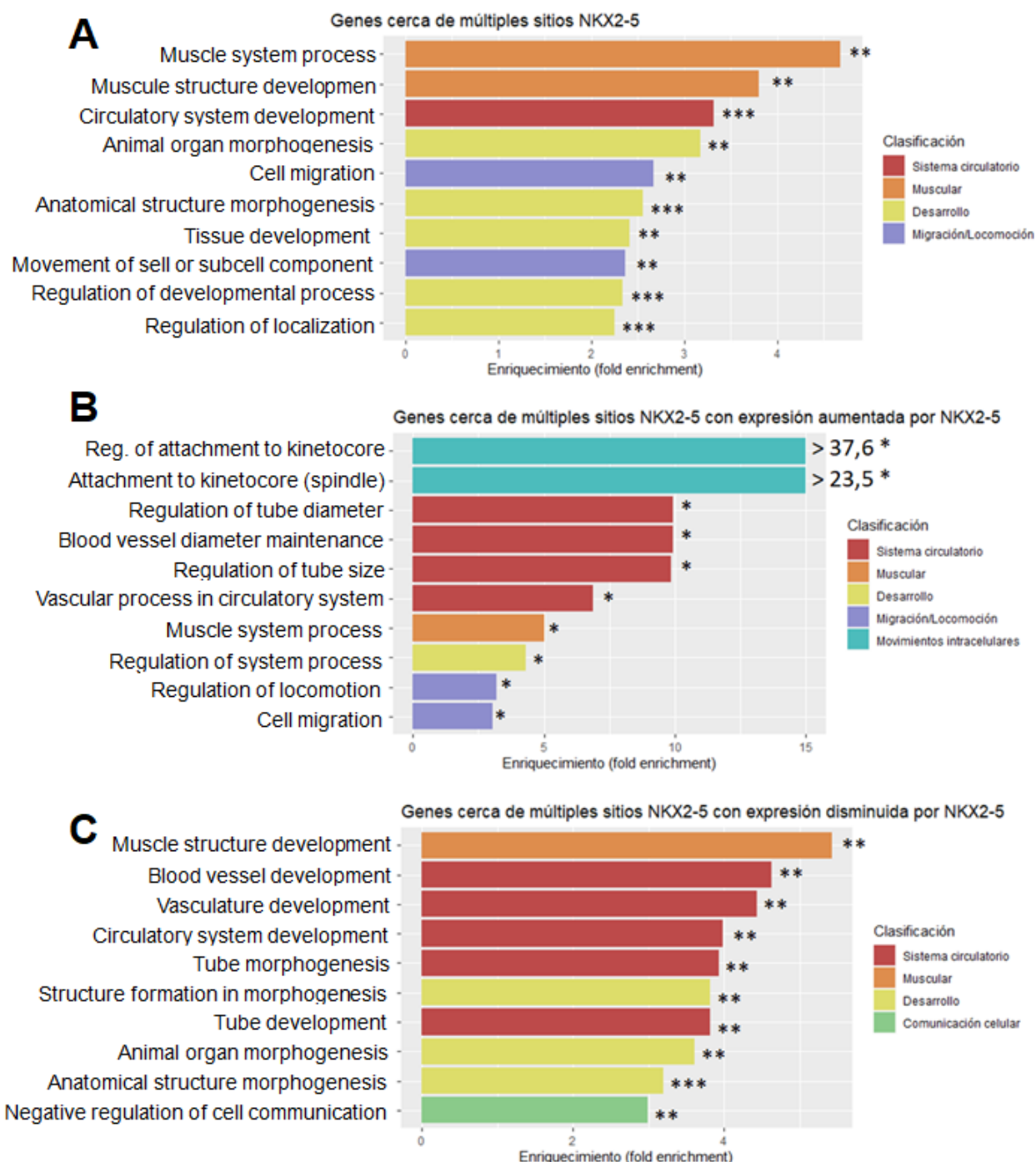
**Figura S4. GO-terms sobrerrepresentados de genes cerca de sitios de unión de NKX2-5 confirmados experimentalmente, seleccionados por presencia de sitios de unión de MEIS1.** El eje X representa el enriquecimiento relativo de cada GO-term con respecto al esperado por azar. Los valores extremos fueron recortados y anotados al costado de las columnas correspondientes. **A**, GO-terms sobrerrepresentados en la lista de todos los genes. **B**, GO-terms sobrerrepresentados en la lista de sólo los genes con expresión disminuida por NKX2-5 (downregulados). **C**, GO-terms sobrerrepresentados en la lista de sólo los genes con expresión aumentada por NKX2-5 (upregulados). Los asteriscos denotan la probabilidad de falso positivo (FDR). \*: FDR<0,001; \*\*: FDR<0,00001; \*\*\*: FDR<10<sup>-10</sup>.



**Figura S5. GO-terms sobrerrepresentados de genes cerca de sitios de unión de NKX2-5 confirmados experimentalmente, seleccionados por presencia de sitios de unión de GATA4.** En el eje X se representa el enriquecimiento relativo de cada GO-term respecto al esperado por azar. Sólo se encontraron GO-terms sobrerrepresentados cuando se analizaron todos los genes (sin filtrar por expresión diferencial aumentada o disminuida en presencia de NKX2-5). Ninguno de los GO-terms tuvo una probabilidad de falso positivo (FDR) menor a 0,001.



**Figura S6. GO-terms sobrerrepresentados de genes cerca de sitios de unión de NKX2-5 confirmados experimentalmente, seleccionados por presencia de sitios de unión de TBX20.** En el eje X se representa el enriquecimiento relativo de cada GO-term en cada lista con respecto al esperado por azar. Los valores extremos fueron recortados y anotados al costado de las columnas correspondientes. **A**, GO-terms sobrerrepresentados en la lista de todos los genes. **B**, GO-terms sobrerrepresentados en la lista de sólo los genes con expresión disminuida por NKX2-5 (downregulados). **C**, GO-terms sobrerrepresentados en la lista de sólo los genes con expresión aumentada por NKX2-5 (upregulados). Los asteriscos denotan la probabilidad de falso positivo (FDR). \*: FDR<0,001; \*\*: FDR<0,00001; \*\*\*: FDR<10<sup>-10</sup>.



**Figura S7. GO-terms sobrerrepresentados de genes cerca de sitios de unión de NKX2-5 confirmados experimentalmente, seleccionados por presencia de otros sitios de unión de NKX2-5 cercanos.** En el eje X se mide el enriquecimiento relativo de cada GO-term con respecto al esperado por azar. Los valores extremos fueron recortados y anotados al costado de las columnas correspondientes. **A**, GO-terms sobrerrepresentados en la lista de todos los genes. **B**, GO-terms sobrerrepresentados en la lista de sólo los genes con expresión disminuida por NKX2-5 (downregulados). **C**, GO-terms sobrerrepresentados en la lista de sólo los genes con expresión aumentada por NKX2-5 (upregulados). Los asteriscos denotan la probabilidad de falso positivo (FDR). \*: FDR<0,001; \*\*: FDR<0,00001; \*\*\*: FDR<10<sup>-10</sup>.