Detección de ARNs Circulares y Estudio de su Implicación en Adenocarcinoma Pulmonar

Alumno: Alfonso Sánchez-Macián Pérez
Máster universitario en Bioinformática y bioestadística UOC-UB
Genómica Computacional

Amadís Pagès Pinós
Carles Ventura Royo
TFM Bioinformática y Bioestadística

ÍNDICE

- Introducción.
- Objetivos.
- Planificación
- Desarrollo.
- Resultados.
- Conclusiones y trabajo futuro

Introducción

CircRNAs: RNA circular que:

- Se supuso inicialmente que era resultado de procesado erróneo.
- Alta expresión en gran número de genes humanos.
- Enlace covalente especial cabeza a cola (backsplice).
- Múltiples funciones, una de ella biomarcador tumoral.

Existen diversas herramientas de alineamiento basadas en tablas Hash o en el algoritmo Burrows-Wheeler.

También hay disponibles diferentes herramientas de identificación de circRNAs.

Introducción (II)

Los árboles de decisión son una técnica de Machine Learning que:

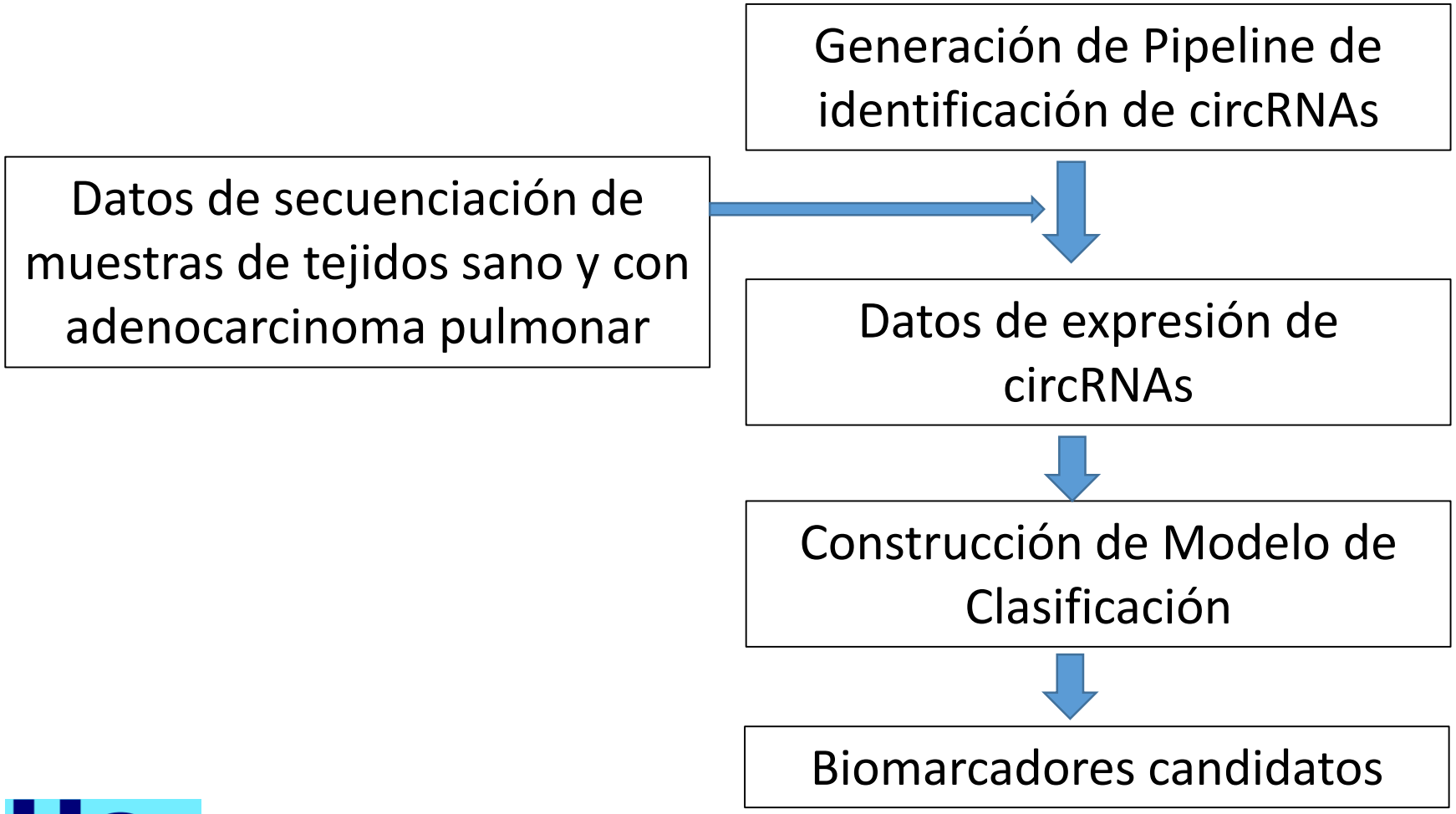
- Permite resolver problemas de clasificación.
- Destaca por su simplicidad y eficiencia.
- Están suficientemente probados.
- Se encuentra disponible en R y otros paquetes de software.

Se puede emplear la métrica Information Gain para reducir la entropía.

Objetivo

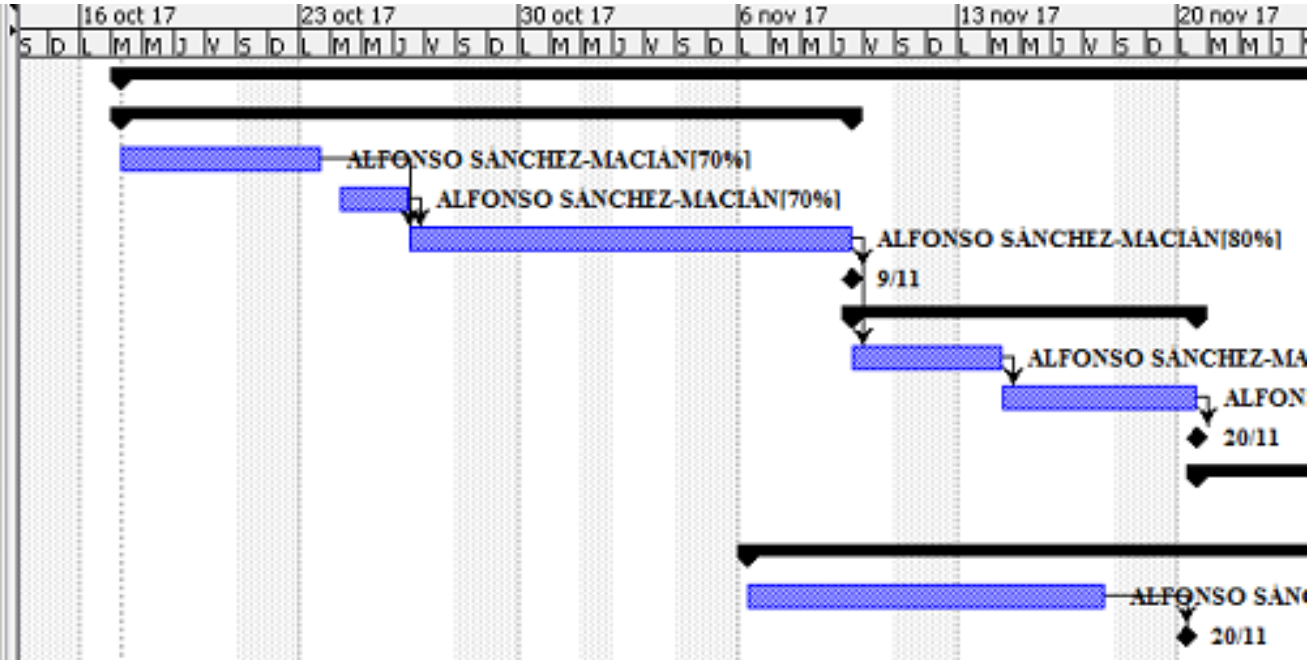
Generación de un modelo bioinformático que permita establecer un conjunto de circRNAs que actúen como biomarcadores moleculares para el adenocarcinoma pulmonar, capaz de clasificar muestras de tejido pulmonar como sanas o tumorales.

Objetivo. Proceso



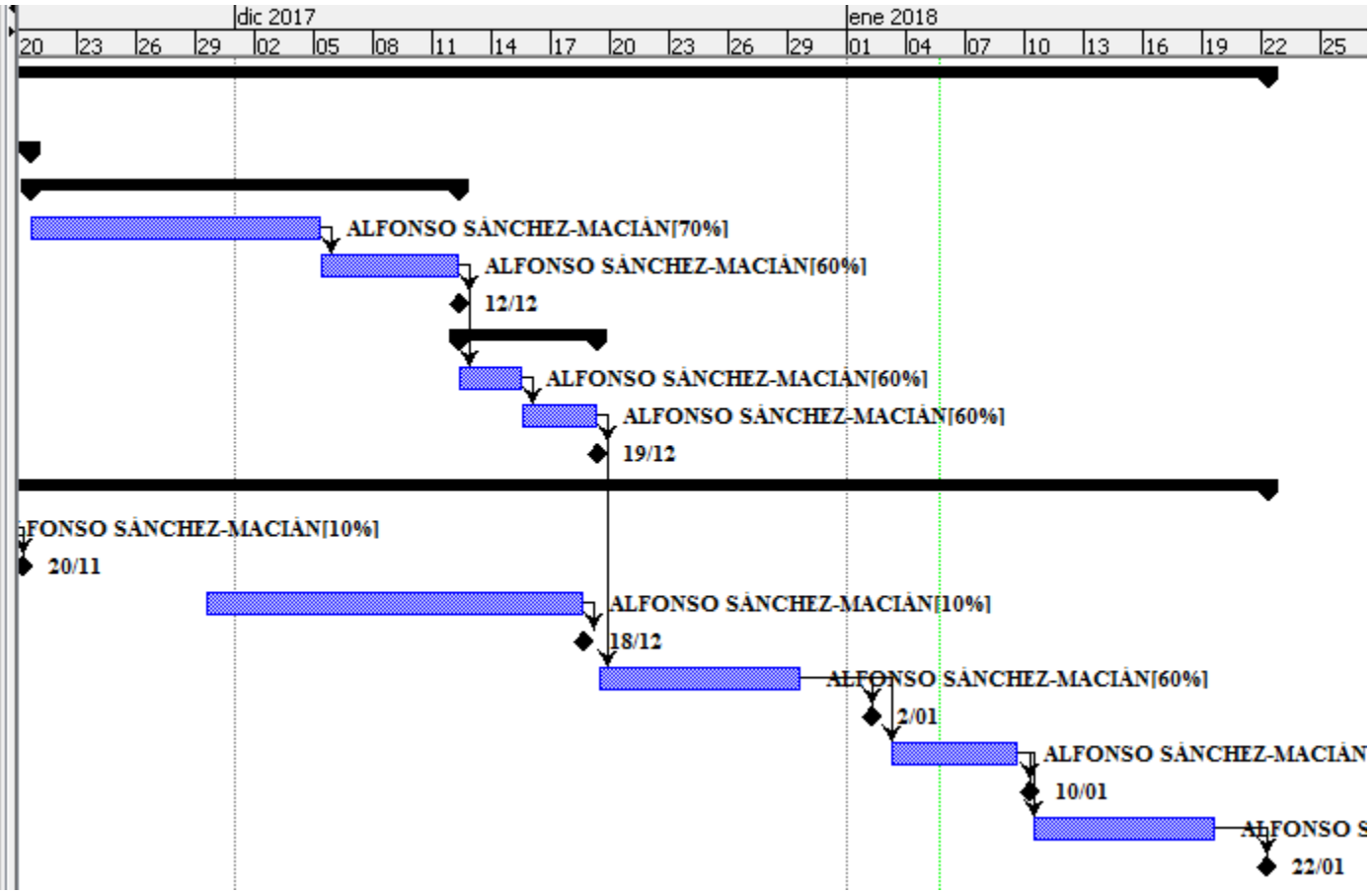
Planificación (I)

	Nombre	Trabajo
1	**1. Trabajo Fin de Máster**	300 horas
2	**1.1 Desarrollar un pipeline que permita identificar circRNAs.**	102 horas
3	1.1.1 Identificación de algoritmos y herramientas de identificación de circRNAs.	28 horas
4	1.1.2 Análisis de la información disponible en The Cancer Genome Atlas.	14 horas
5	1.1.3 Generación del pipeline a partir de las herramientas seleccionadas.	60 horas
6	M1.1 Pipeline de identificación de circRNAs generado	0 horas
7	**1.2 Aplicar el pipeline a muestras de adenocarcinoma pulmonar**	34 horas
8	1.2.1 División de la información en un grupo de entrenamiento y uno de test. Co	12 horas
9	1.2.2 Aplicación del pipeline definido al conjunto de datos de entrenamiento y a	22 horas
10	M1.2 Datos de circRNA obtenidos	0 horas
11	**1.3 Generar un modelo que facilite la clasificación de muestras como s**	72 horas
15	**1.4 Obtener un conjunto inicial de biomarcadores circRNAs**	24 horas
19	**1.5 Generación de la documentación y presentación**	68 horas
20	1.5.1 Generación de Informe de seguimiento: Desarrollo del proyecto Fase I.	8 horas
21	M1.5.1 Entrega del Primer Informe de Seguimiento	0 horas

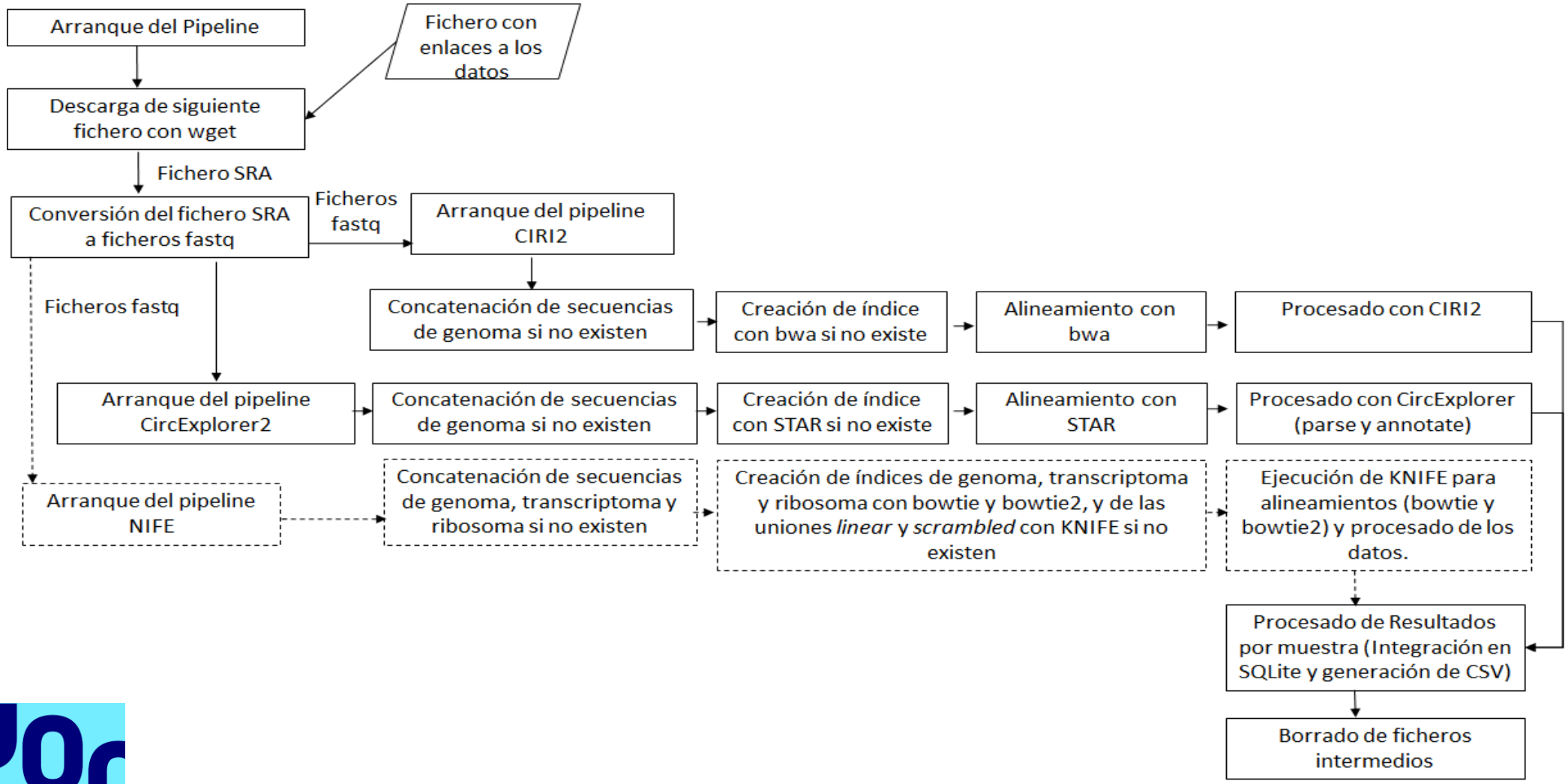


Planificación (II)

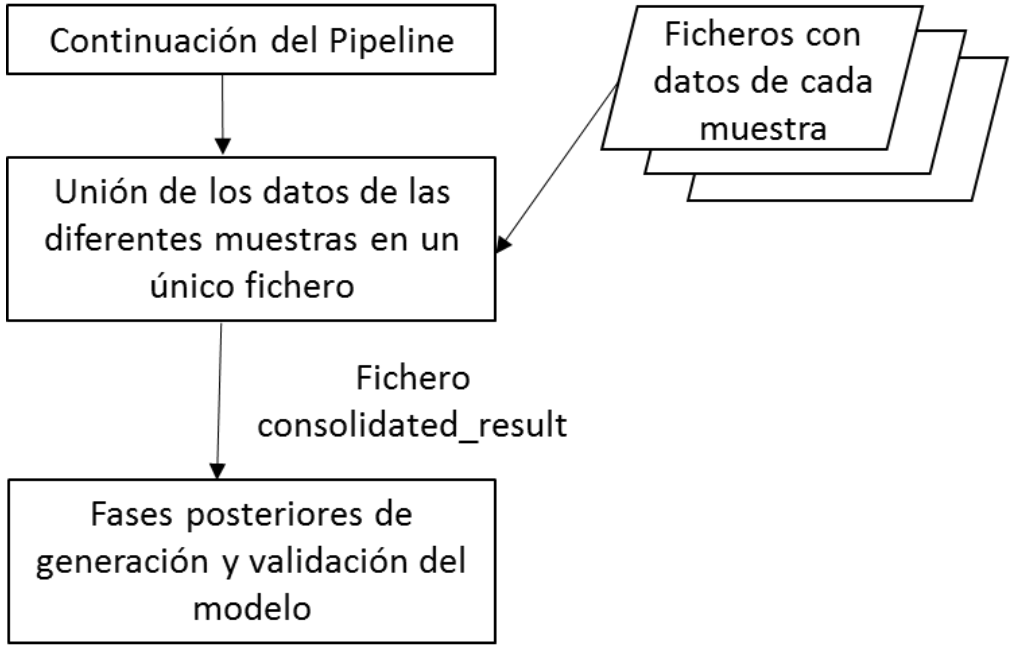
	Nombre	Trabajo
1	**1. Trabajo Fin de Máster**	300 horas
2	**1.1 Desarrollar un pipeline que permita identificar circRNAs.**	102 horas
7	**1.2 Aplicar el pipeline a muestras de adenocarcinoma pulmonar**	34 horas
11	**1.3 Generar un modelo que facilite la clasificación de muestras como s**	72 horas
12	1.3.1 Construcción del modelo de Árboles de decisión e Information Gain aplicár	58 horas
13	1.3.2 Verificación del modelo con el conjunto de test.	14 horas
14	M1.3 Modelo verificado	0 horas
15	**1.4 Obtener un conjunto inicial de biomarcadores circRNAs**	24 horas
16	1.4.1 Extracción de un conjunto de circRNAs con diferente expresión, a partir c	16 horas
17	1.4.2 Comprobación de existencia de información sobre los circRNAs obtenidos.	8 horas
18	M1.4 Biomarcadores de circRNA disponibles	0 horas
19	**1.5 Generación de la documentación y presentación**	68 horas
20	1.5.1 Generación de Informe de seguimiento: Desarrollo del proyecto Fase I.	8 horas
21	M1.5.1 Entrega del Primer Informe de Seguimiento	0 horas
22	1.5.2 Generación de Informe de seguimiento: Desarrollo del proyecto Fase II.	9 horas
23	M1.5.2 Entrega del Segundo Informe de Seguimiento	0 horas
24	1.5.3 Redacción de la memoria.	36 horas
25	M1.5.3 Entrega de la Memoria	0 horas
26	1.5.4 Elaboración de la presentación.	8 horas
27	M1.5.4 Entrega de la Presentación	0 horas
28	1.5.5 Ensayo de la Defensa pública.	6 horas
29	M1.5.5 Defensa pública	1 hora



Desarrollo. Identificación de circRNAs



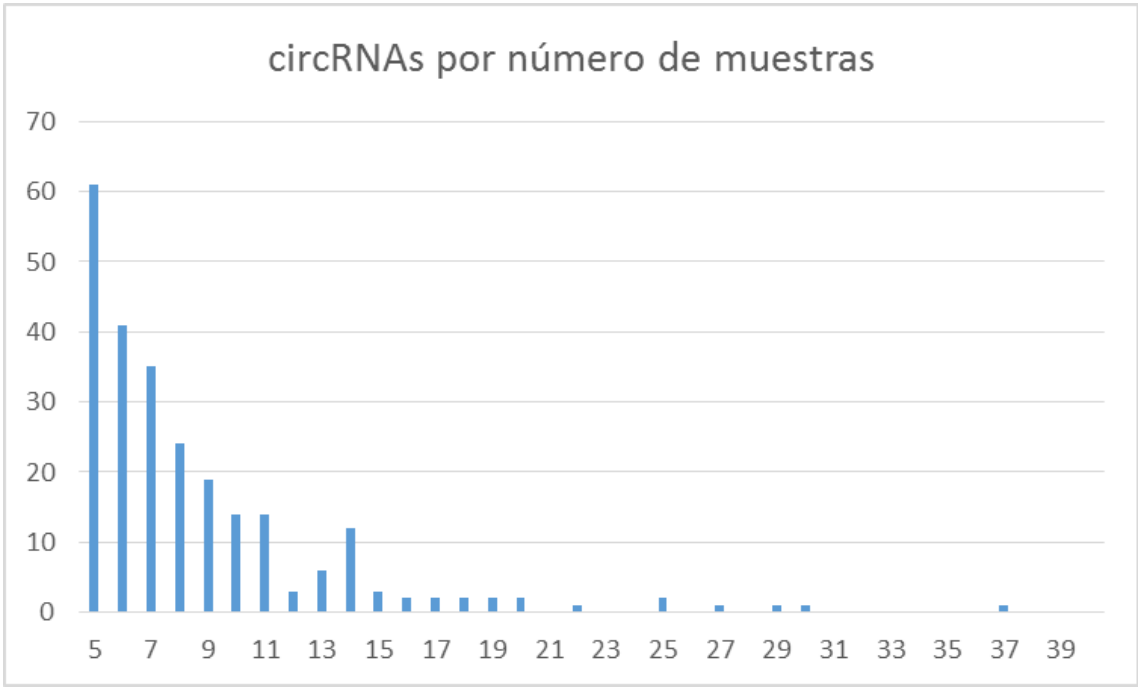
Desarrollo. Identificación de circRNAs (II)



Desarrollo. Datos de expresión de circRNAs

- 7577 circRNAs detectados, 118 por los 2 CIRI2 y CIRCExplorer2

- De ellos, 6217 en una sola muestra, 749 en dos muestras, 249 en 3 muestras, 112 en 4 muestras. El resto:



Desarrollo. Preparación de los datos

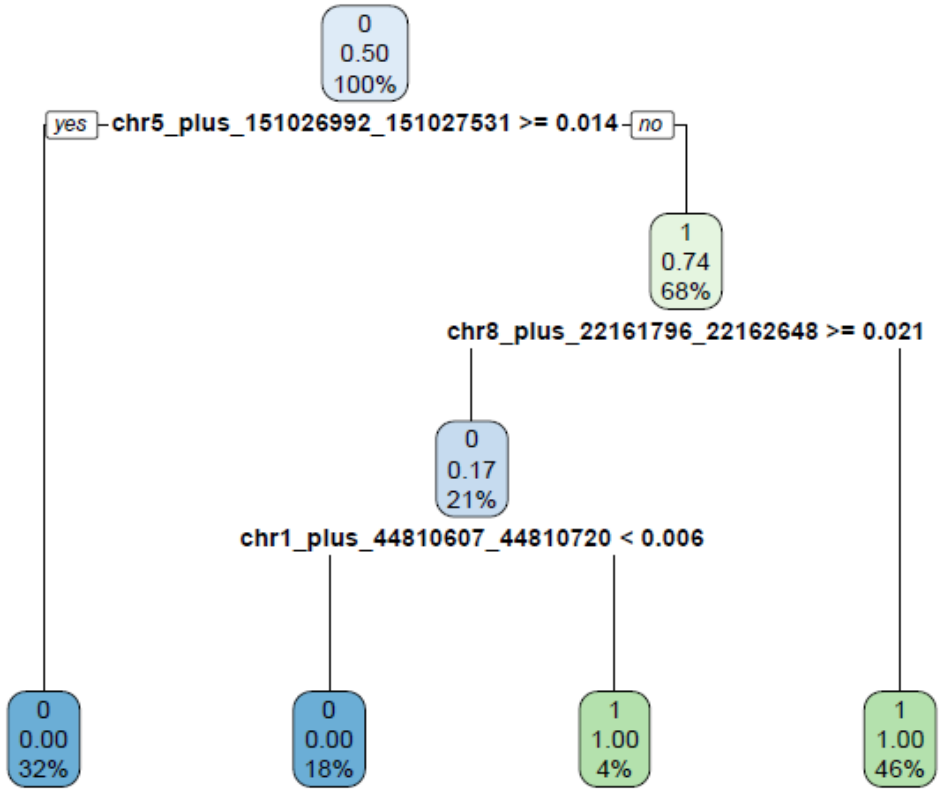
- Normalización usando métrica RPM (Read Per Million Mapped Reads).

$$RPM = (reads_{\_circRNA} * 10^6) / reads_{\_simple}$$

- Conjuntos de entrenamiento (28 muestras, 14 sanas y 14 tumorales) y test (12 muestras, 6 sanas y 6 tumorales).

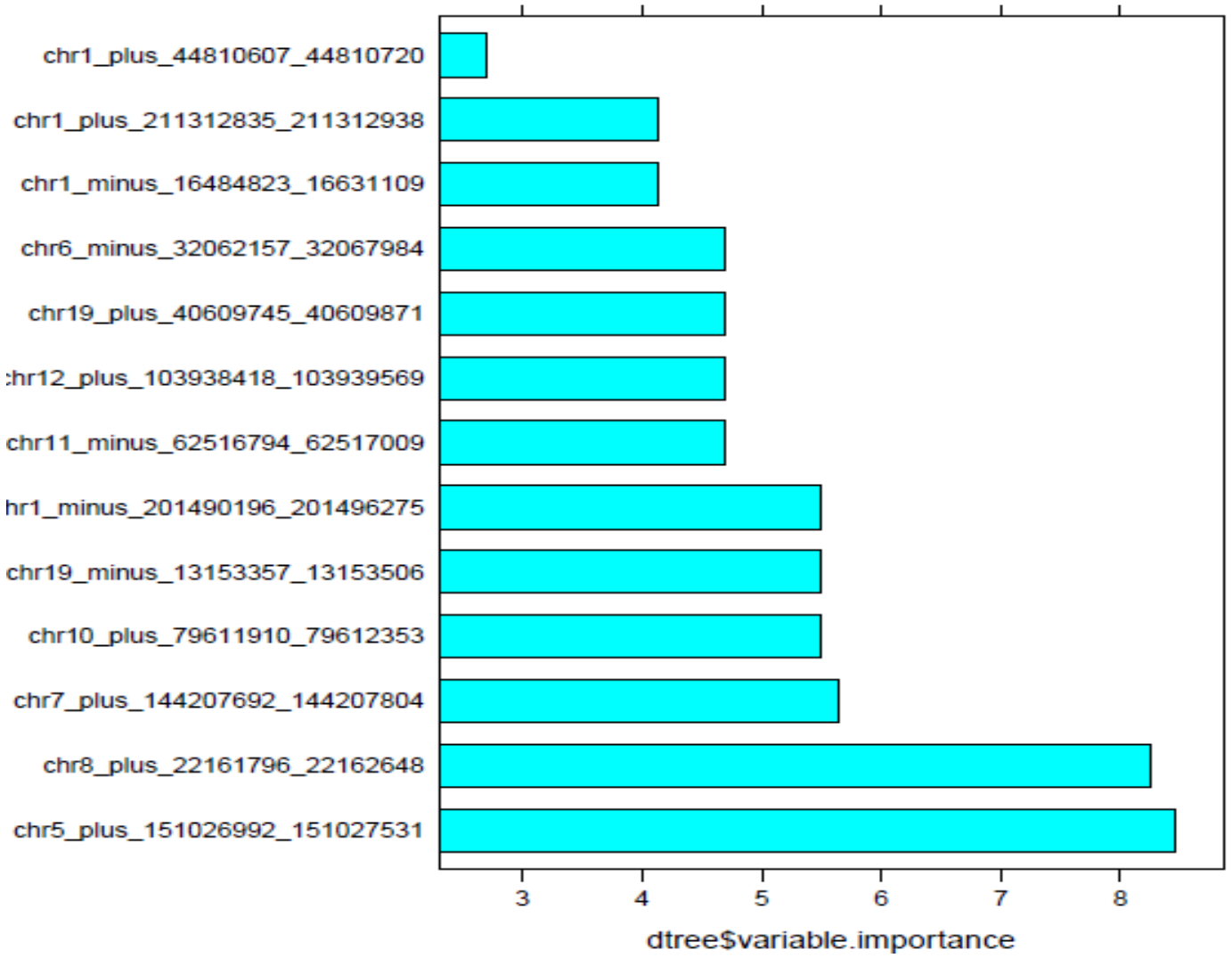
- Desarrollado en R con paquetes específicos (*rpart, caret, gmodel)*

Desarrollo. Modelo de árbol de decisión

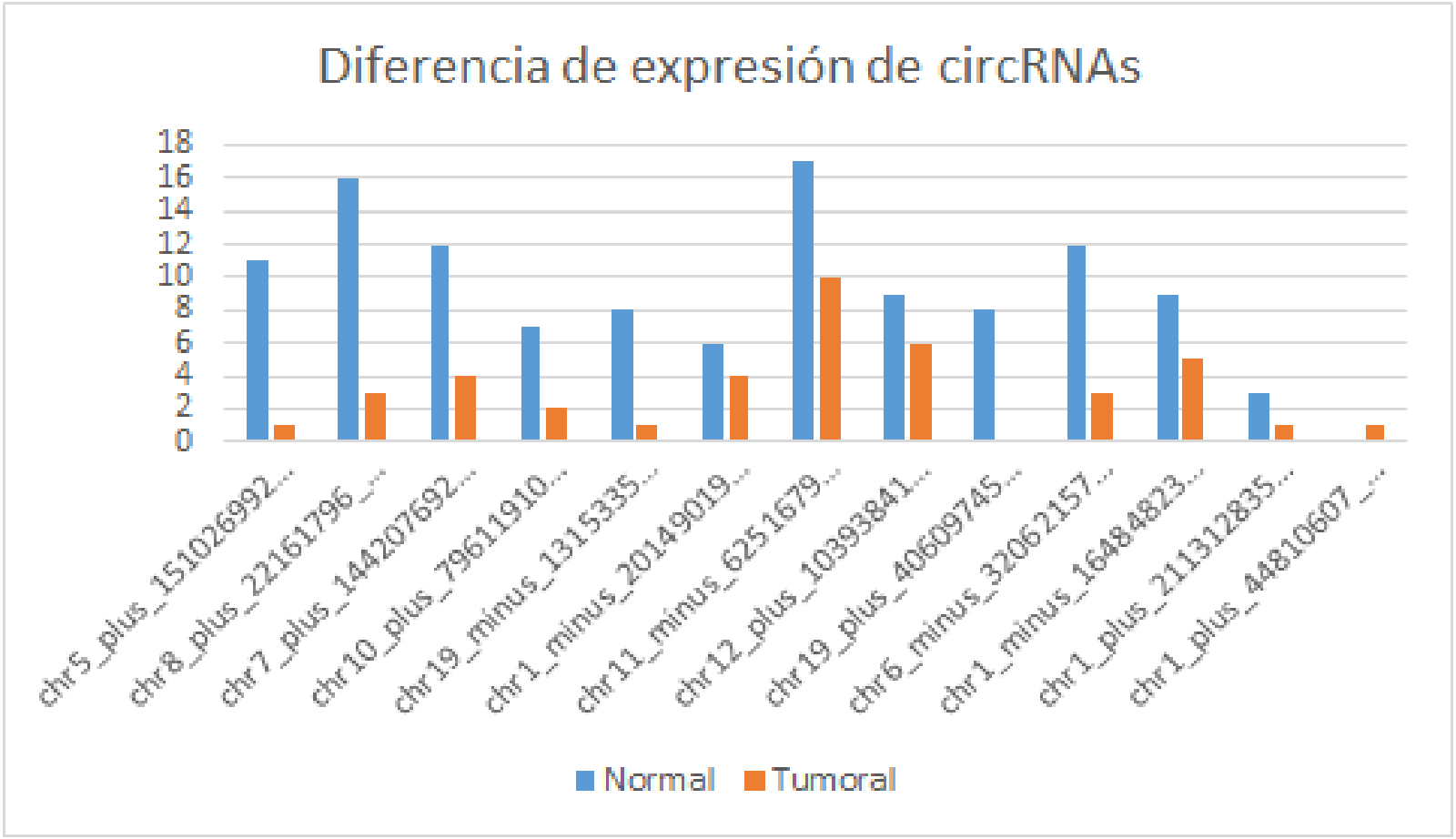


	test[, "pred_class"]		
test[, "condition"]	0	1	Row Total
0	6	0	6
	1.000	0.000	0.500
	0.857	0.000	
	0.500	0.000	
1	1	5	6
	0.167	0.833	0.500
	0.143	1.000	
	0.083	0.417	
Column Total	7	5	12
	0.583	0.417	

Resultados. Importancia de las variables



Resultados. Diferencia de expresión



Resultados. circRNAs

	circBase	circRNAdb	CSCD	Gen asociado	Programas
chr5:151026992-151027531 (+) [chr5:150406553-150407092]	hsa_circ_0074576	-	chr5:151025375|151028130	GPX3	CIRI2
chr8:22161796-22162648 (+) [chr8:22019309-22020161]		chr8:22019183-22020245		SFTPC	CIRI2, CIRCExplorer2 y KNIFE
chr7:144207692-144207804 (+) [chr7:143904785-143904897]	-	-	-	RP4-545C24.1	CIRCExplorer2 y KNIFE
chr10:79611910-79612353 (+) [chr10:81371666-81372109]	-	-	-	SFTPA1	CIRCExplorer2
chr19:13153357-13153506 (-) [chr19:13264171-13264320]	-	-	-	CTC-250I14.6	CIRCExplorer2
chr1:201490196-201496275 (-) [chr1:201459324-201465403]	-	-	chr1:201459327|201465386	CSRP1	CIRCExplorer2
chr11:62516794-62517009 (-) [chr11:62284266-62284481]	-	-	chr11:62284210|62284417	AHNAK	CIRI2 y CIRCExplorer2

Resultados. circRNAs

	circBase	circRNAdb	CSCD	Gen asociado	Programas
chr12:103938418-103939569 (+) [chr12:104332196-104333347]	-	-	2 solapados parcialmente	HSP90B1	CIRCExplorer2
chr19:40609745-40609871 (+) [chr19:41115651-41115777]	-	-	-	LTBP4	CIRCExplorer2
chr6:32062157-32067984 (-) [chr6:32029934-32035761]	-	-	-	TNXB e intergénico	CIRI2 y KNIFE
chr1:16484823-16631109 (-) [chr1:16811318-16957604]	-	-	-	Varios genes	CIRI2 y KNIFE
chr1:211312835-211312938 (+) [chr1:211486177-211486280]	-	-	chr1:211485784|211486360	RCOR3	CIRCExplorer2
chr1:44810607-44810720 (+) [chr1:45276279-45276392]	-	-	-	BTBD19	CIRCExplorer2

Conclusiones

Generado un modelo de árbol de decisión:

- Capaz de predecir con un 91% de precisión la condición normal o tumoral de las muestras.
- Sensibilidad de un 83,33% sobre el conjunto de test
- Especificidad de un 100%

Los circRNAs más relevantes resultado del trabajo han sido identificados y validados experimentalmente en estudios previos.

Trabajo futuro

- Aplicación del *pipeline* a un conjunto mayor de muestras de adenocarcinoma pulmonar.
- Aplicación del *pipeline* a otros tipos de muestras.
- Verificación experimental de los circRNAs identificados como posibles biomarcadores de adenocarcinoma pulmonar.
- Incorporación de otras herramientas de identificación del circRNA al análisis para poder obtener un mayor número de circRNAs.