

# Epidemiología y detección de biomarcadores en cáncer

**Daniel Redondo Sánchez**

Máster Universitario Oficial en Ciencia de Datos  
e Ingeniería de Computadores. Curso 2019/20.

**Tutores: Daniel Castillo Secilla, Luis Javier Herrera**



UNIVERSIDAD  
DE GRANADA

# Índice

1. Introducción
2. Metodología
3. Resultados
4. Conclusiones
5. Líneas abiertas de trabajo

# Índice

1. Introducción
2. Metodología
3. Resultados
4. Conclusiones
5. Líneas abiertas de trabajo

# Introducción

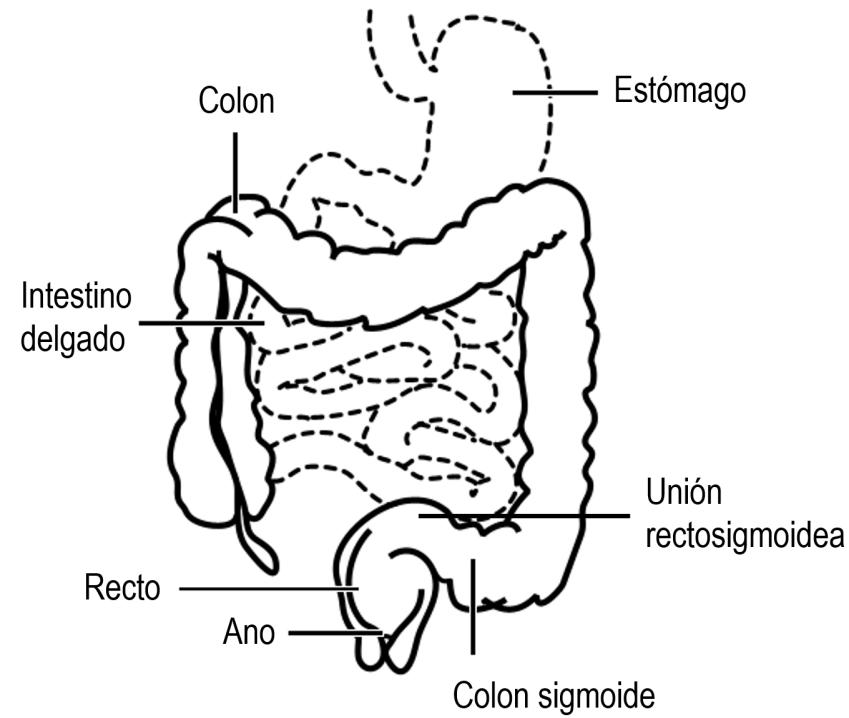
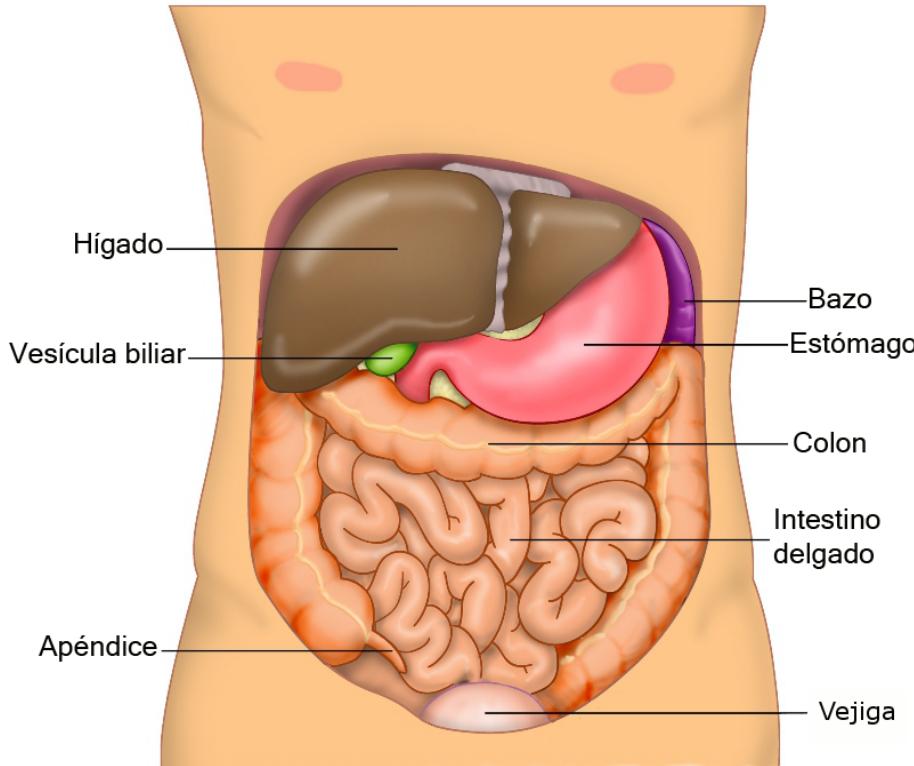
El cáncer se caracteriza por una división incontrolada de las células.

Es uno de los mayores problemas de salud pública:

- En el mundo: 17 millones de casos y 9,4 millones de defunciones al año.
- En España: 249.000 casos, 108.000 defunciones anuales.

# Introducción

Este trabajo se centra en dos localizaciones anatómicas:  
hígado y colon-recto.

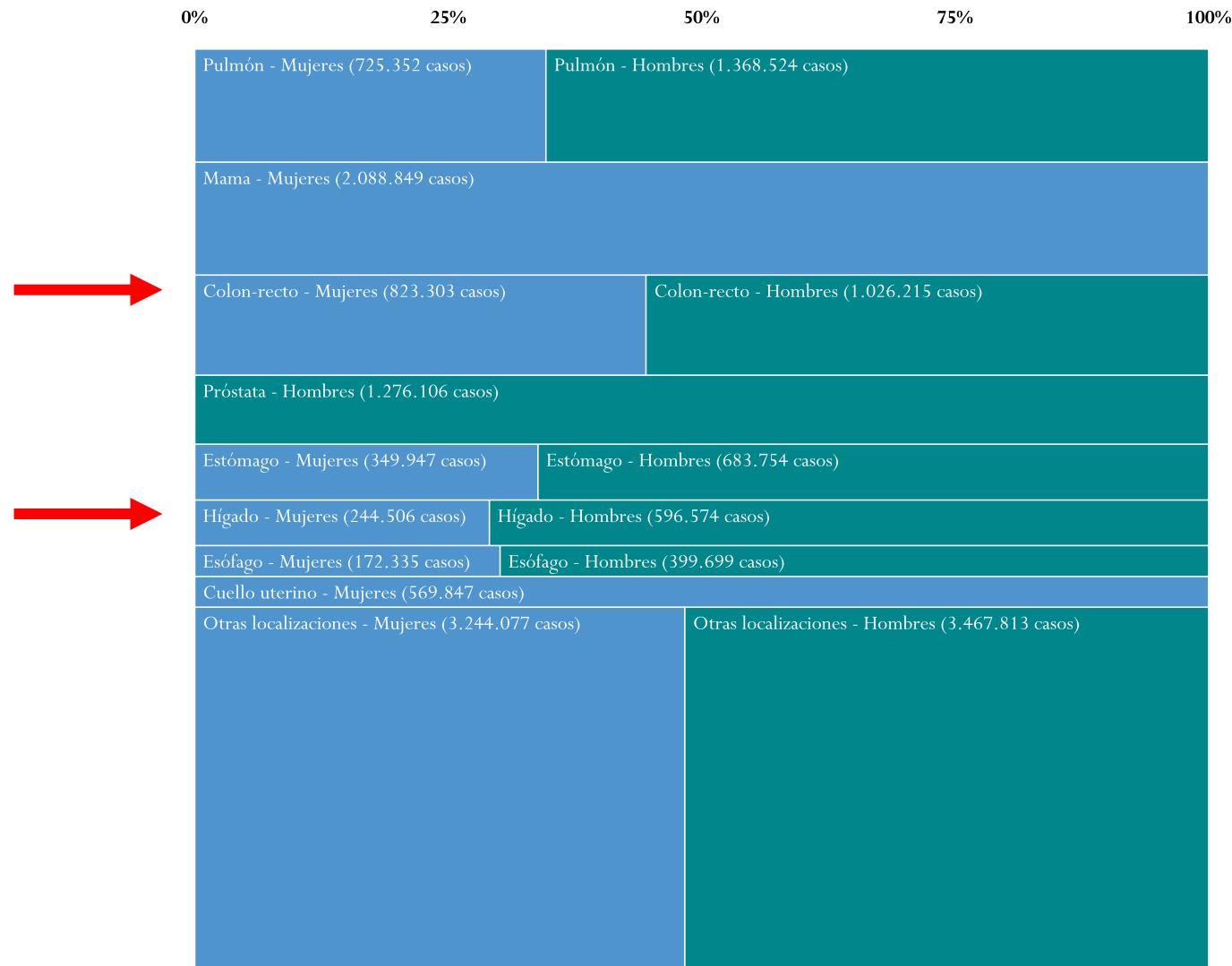


# Introducción

## Incidencia de cáncer en el mundo, 2018

Distribución de casos por sexo y localización anatómica.

Fuente: Global Cancer Observatory, Organización Mundial de la Salud.



# Introducción

## Mortalidad por cáncer en el mundo, 2018

Distribución de defunciones por sexo y localización anatómica.

Fuente: Global Cancer Observatory, Organización Mundial de la Salud.



# Índice

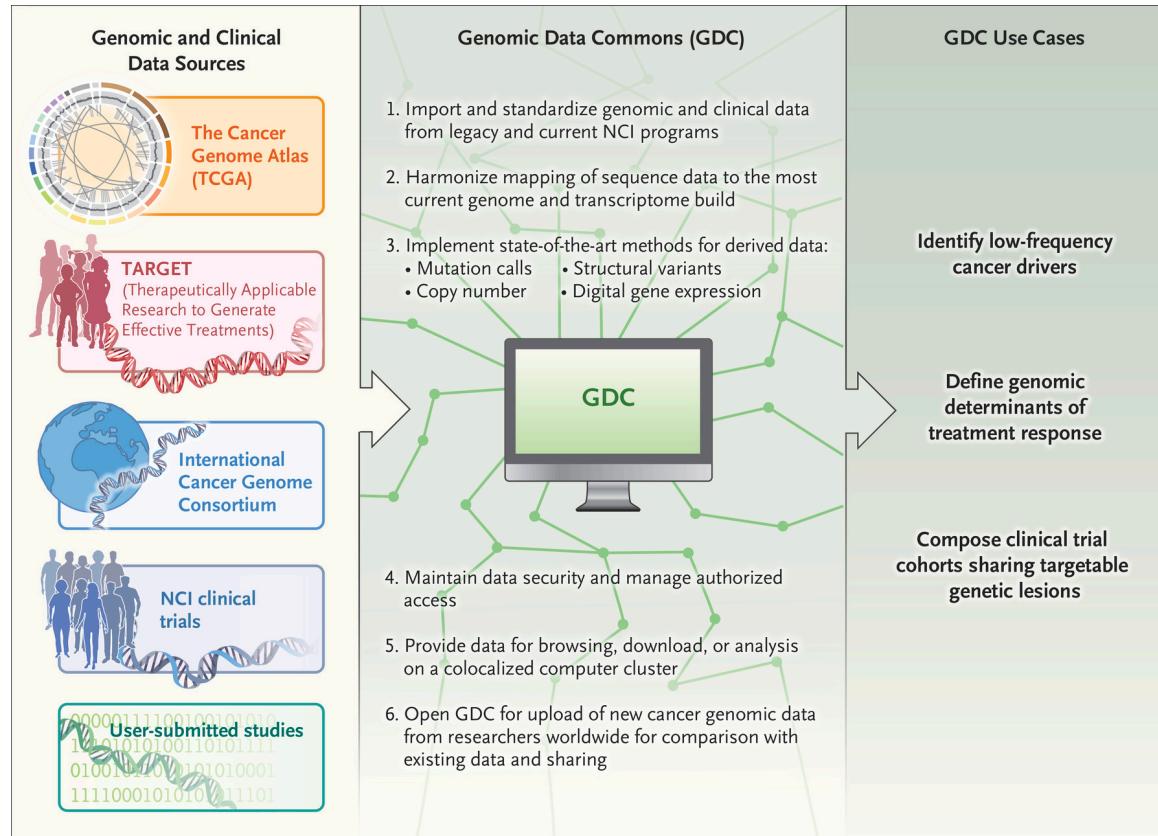
1. Introducción
2. Metodología
3. Resultados
4. Conclusiones
5. Líneas abiertas de trabajo

# Metodología

## Fuente de información: GDC Portal

Plataforma web que integra fuentes heterogéneas de datos.

Se descargan datos transcriptómicos de acceso abierto (RNA-Seq).



# Metodología

## Partición entrenamiento-test:

75% entrenamiento - 25% test, con balanceo de clases

## Métodos de selección de características:

- mRMR (mínima redundancia, máxima relevancia)
- RF (*random forest*)
- DA (asociación de enfermedades)

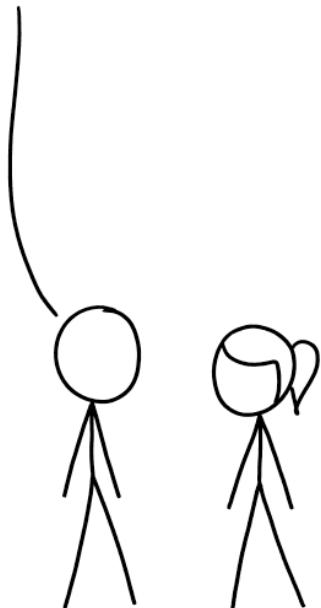
## Métodos de clasificación:

- SVM de kernel radial con optimización de coste y gamma
- RF (*random forest*)
- kNN (k vecinos más cercanos) con optimización de k

**Validación cruzada:** 5-fold

# Metodología

WANNA SEE THE CODE?



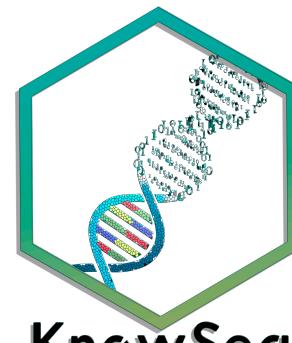
# Metodología



+



+

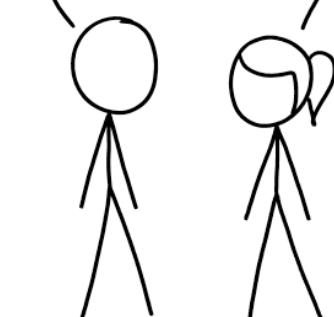


+



WANNA SEE THE CODE?

I WOULD, IF YOU HADN'T  
SAID THAT IN THE TONE  
OF VOICE OF "WANNA  
SEE A DEAD BODY?"



xkcd.com/2138/

[https://github.com/danielredondo/TFM\\_ciencia\\_de\\_datos](https://github.com/danielredondo/TFM_ciencia_de_datos)

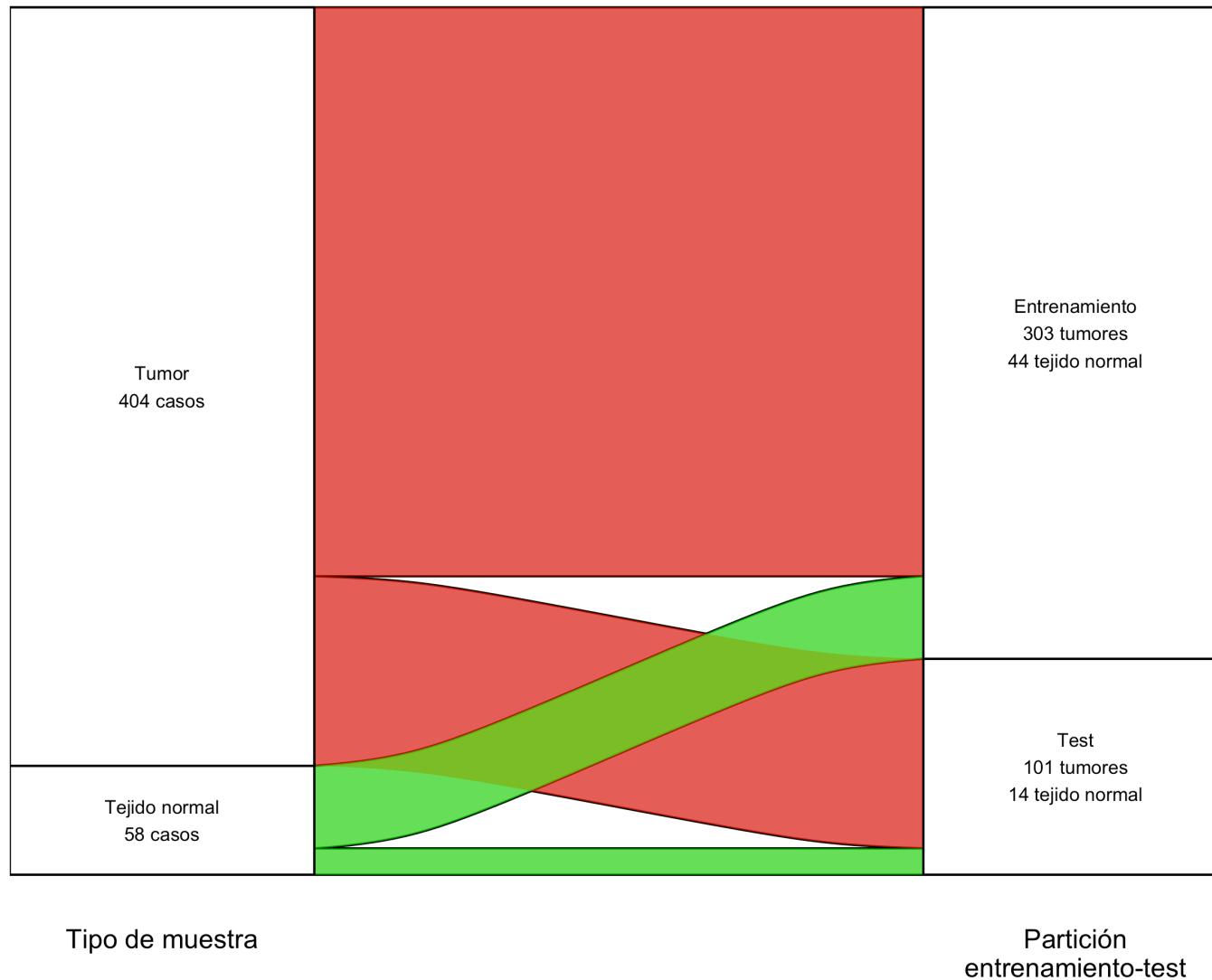
danielredondo actualizar presentación		e51992c 11 hours ago	161 commits
	analisis_cr	mejorar documento con comentarios de Luis Javier	last month
	analisis_higado	añadir script para exportar datos Shiny	last month
	documento	realizar maquetación final	3 days ago
	epidemiologia	renombrar carpetas	2 months ago
	presentacion	actualizar presentación	11 hours ago
	shiny	añadir script para crear datos de ejemplo	6 days ago
	.gitignore	ignorar figuras auxiliares de la presentación	2 days ago
	LICENSE	Initial commit	4 months ago
	README.md	actualizar README	2 days ago
	session_info.txt	actualizar README y session_info	2 months ago

Repositorio con licencia MIT

# Índice

1. Introducción
2. Metodología
- 3. Resultados**
4. Conclusiones
5. Líneas abiertas de trabajo

# Resultados - Hígado biclase



F1-Score como medida de evaluación, al ser un problema con desequilibrio de clases

# Resultados - Hígado biclase

Diez genes más relevantes según método de selección de características

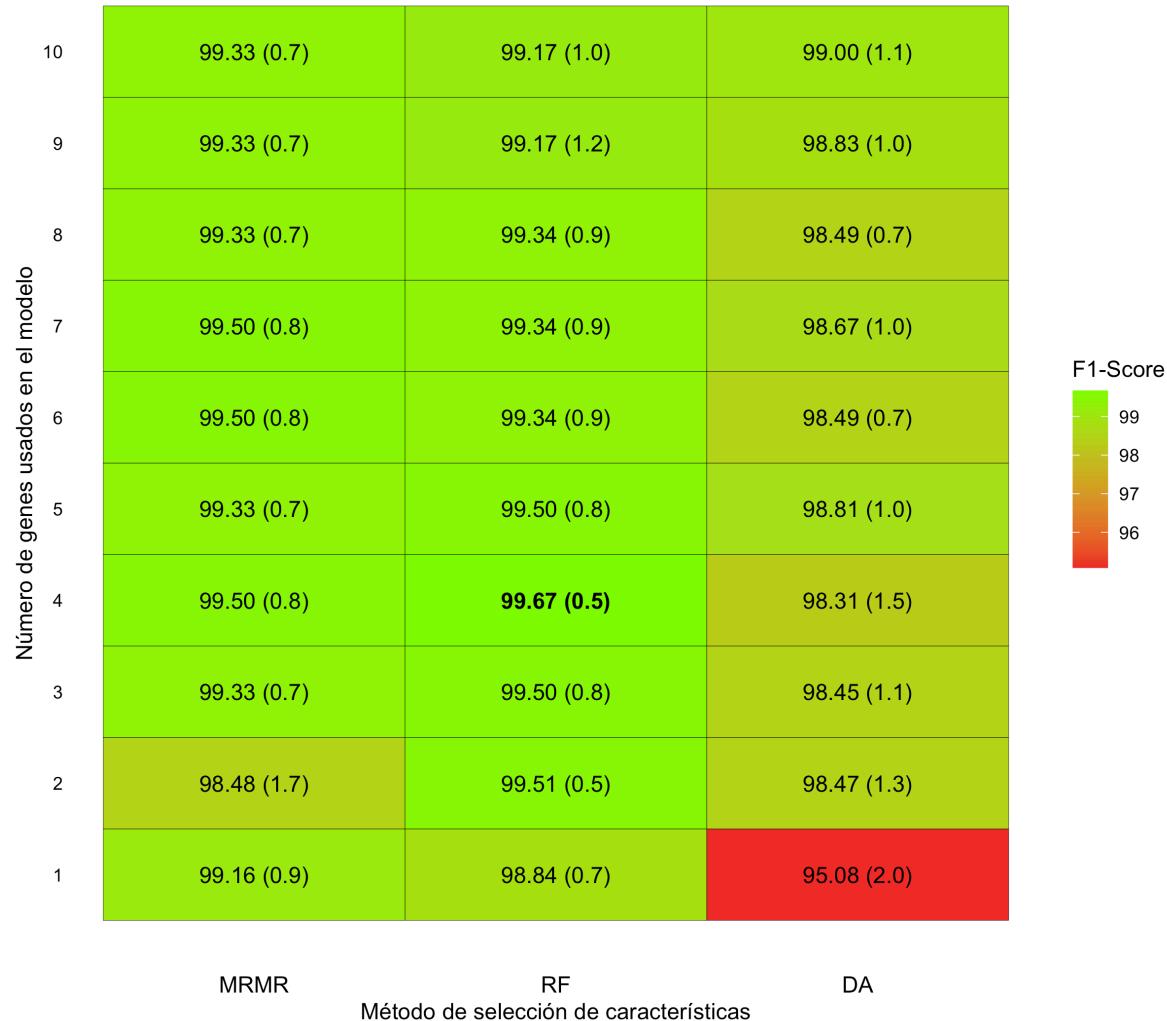
Ranking	mRMR	RF	DA
1	ANGPTL6	ANGPTL6	TERT
2	THY1	PTH1R	RSPO3
3	ADAMTS13	ADAMTS13	HOXA13
4	CELSR3	BMPER	SIX1
5	CCNE1	PRC1	TOP2A
6	CDH13	CLEC4G	GPC3
7	C14orf180	VIPR1	SSX1
8	GABRD	CLEC4M	BUB1B
9	AP000439.2	OIT3	RET
10	CEP152	GABRD	ESR1

# Resultados - Hígado biclase

## F1-Score para SVM con 5-fold

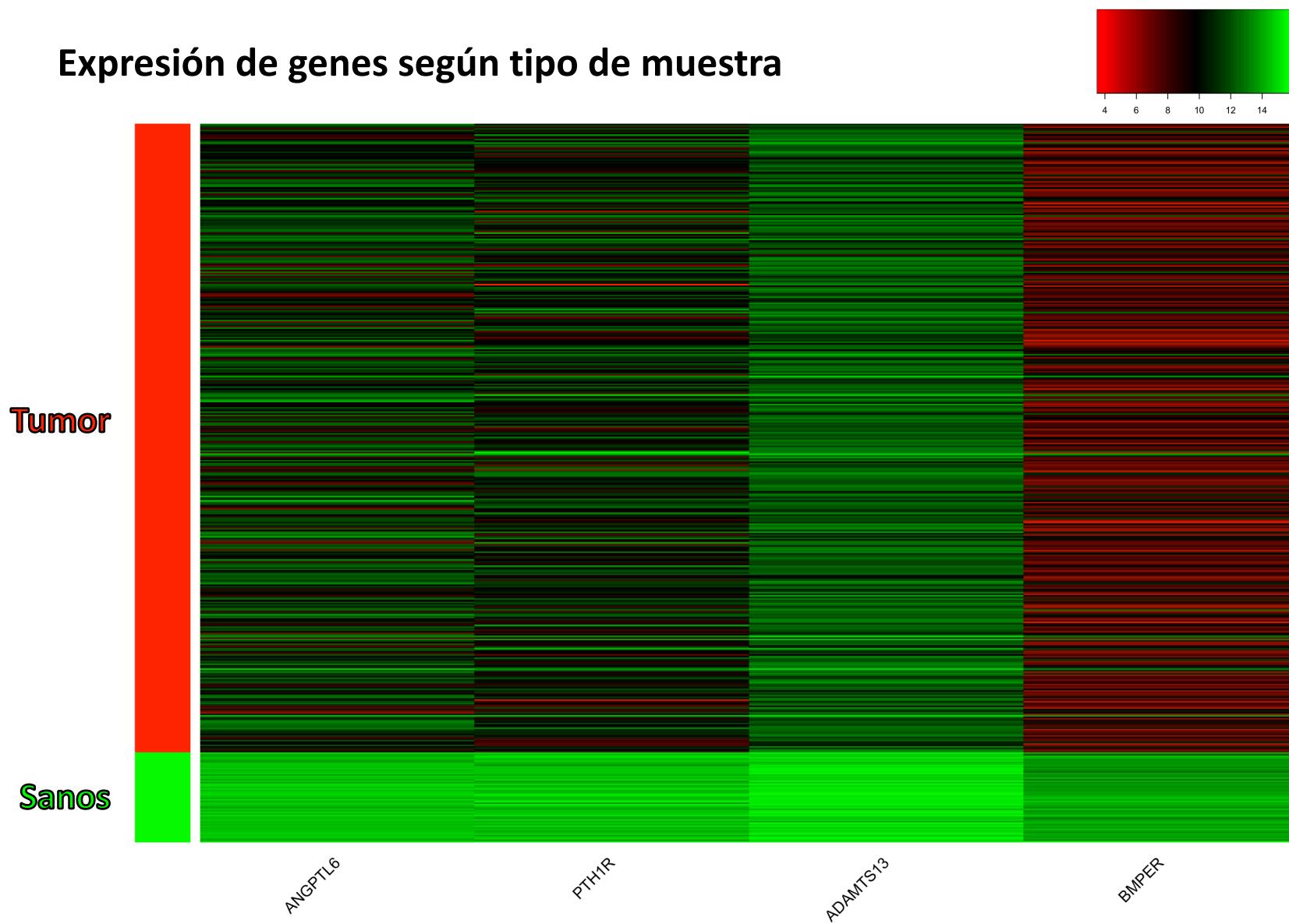
El mejor modelo se obtiene con 4 genes elegidos con RF.

F1-Score de SVM según método de selección de características  
y número de genes usados en el modelo  
F1-Score medio en los 5 fold (desviación típica)



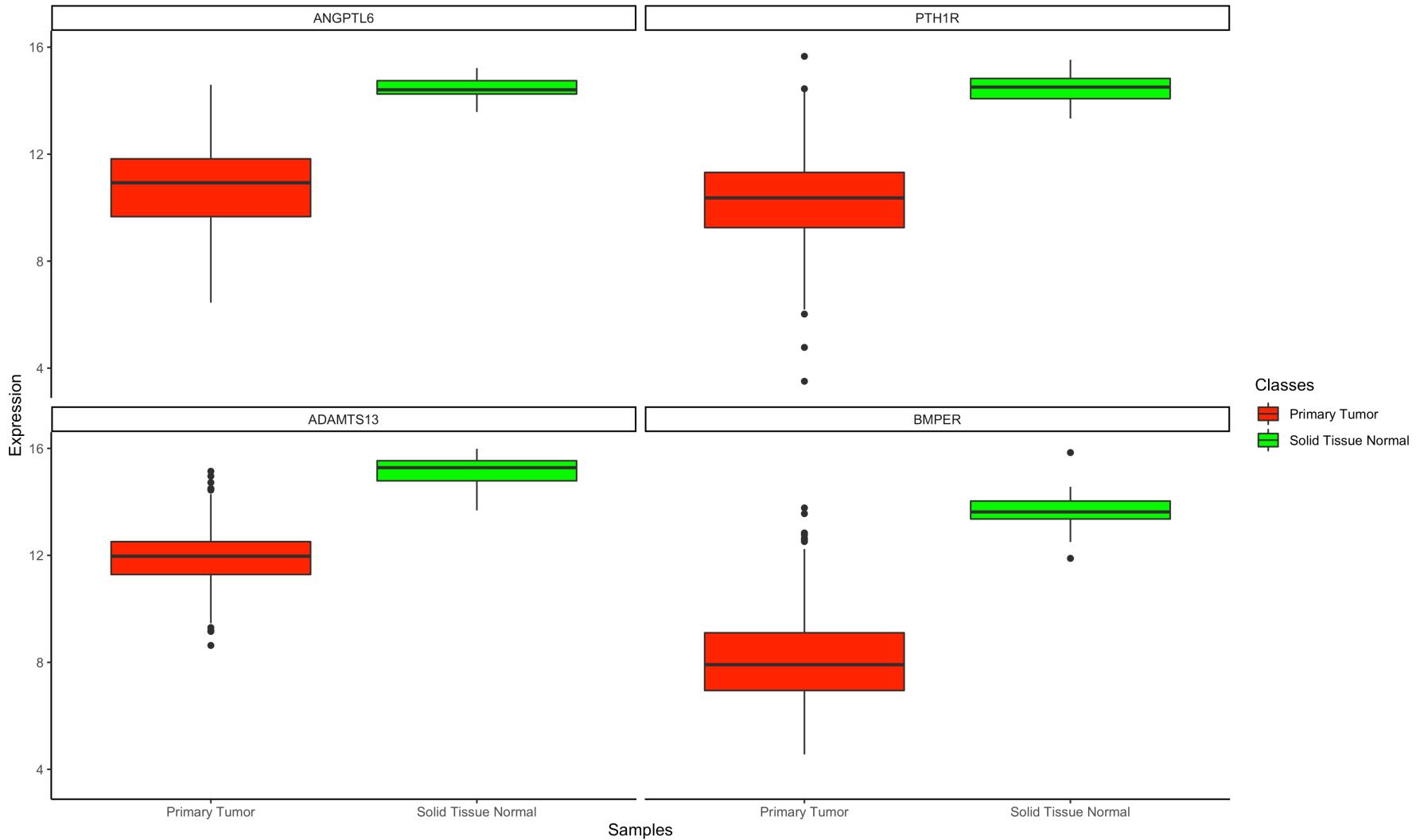
# Resultados - Hígado biclase

Expresión de genes según tipo de muestra



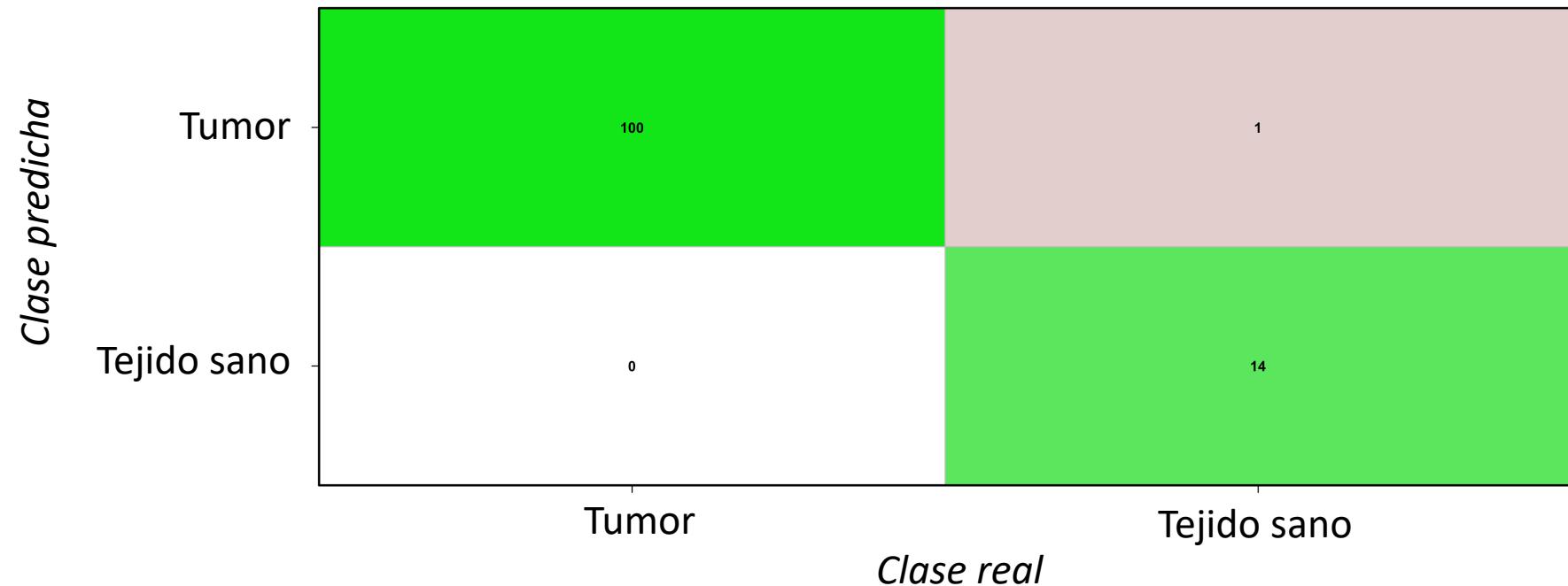
# Resultados - Hígado biclase

## Expresión de genes según tipo de muestra



# Resultados - Hígado biclase

## Validación en test



**Clasificación perfecta salvo 1 falso positivo.**

# Resultados - Hígado biclase

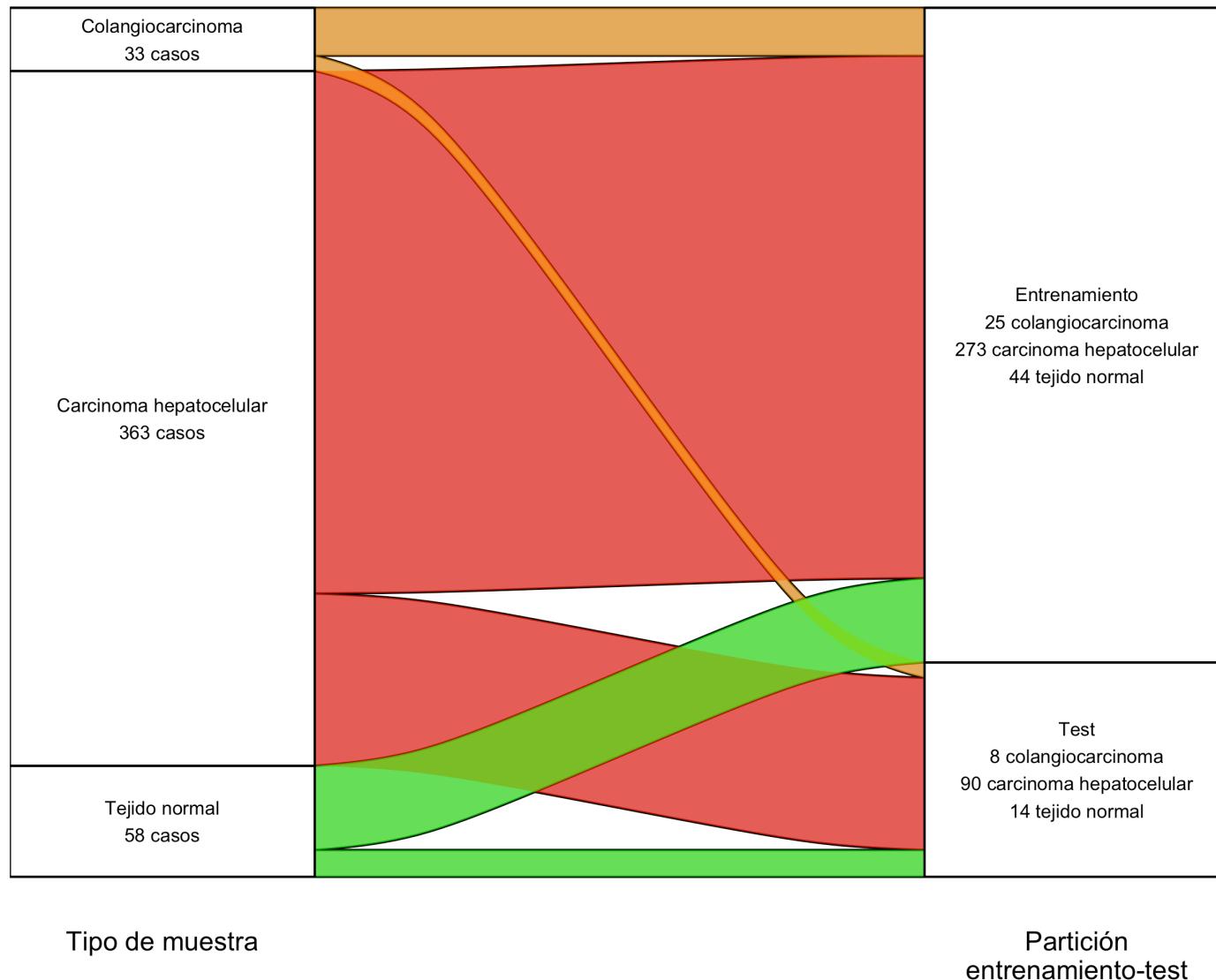
Tabla resumen de los mejores modelos según método de clasificación

	Biomarcadores	Parámetros	F1 train	Acc train	F1 test	Acc test
SVM	RF 4 genes	coste = 0.75 gamma = 0.1	99,67	99,42	99,5	99,13
RF	mRMR 7 genes	--	99,68	99,43	99,5	99,13
kNN	RF 2 genes	k = 5	99,67	99,42	99,5	99,13

Modelos muy similares, alto poder discriminatorio.

Misma matriz de confusión en test.

# Resultados - Hígado multiclase



# Resultados - Hígado multiclase

Diez genes más relevantes según método de selección de características

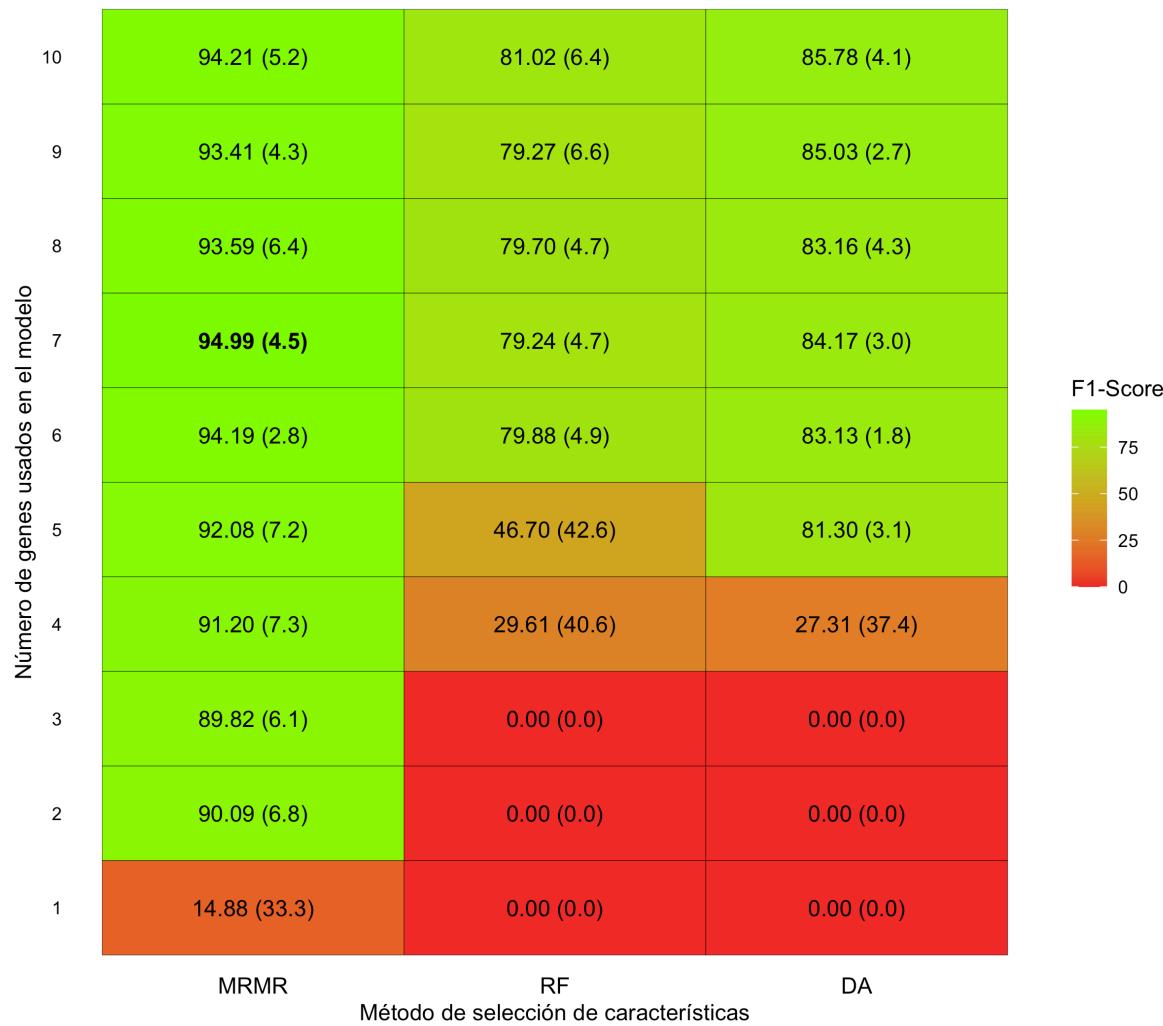
Ranking	mRMR	RF	DA
1	ANGPTL6	ANGPTL6	WWTR1
2	FTLP3	GABRD	BIRC3
3	PLXDC1	CDH13	CDH1
4	RAB25	STAB2	ROS1
5	WDR66	BMPER	POLQ
6	AP2B1	ECM1	FGFR2
7	CDH13	ADAMTS13	KLF6
8	PTPN13	GDF2	CBFB
9	SLC31A1	CLEC4G	FGFR3
10	ADAMTS13	SPDL1	CLTCL1

# Resultados - Hígado multiclase

## F1-Score para SVM con 5-fold

El mejor modelo se obtiene con 7 genes elegidos con mRMR.

F1-Score de SVM según método de selección de características  
y número de genes usados en el modelo  
F1-Score medio en los 5 fold (desviación típica)



# Resultados - Hígado multiclase

Tabla resumen de los mejores modelos según método de clasificación

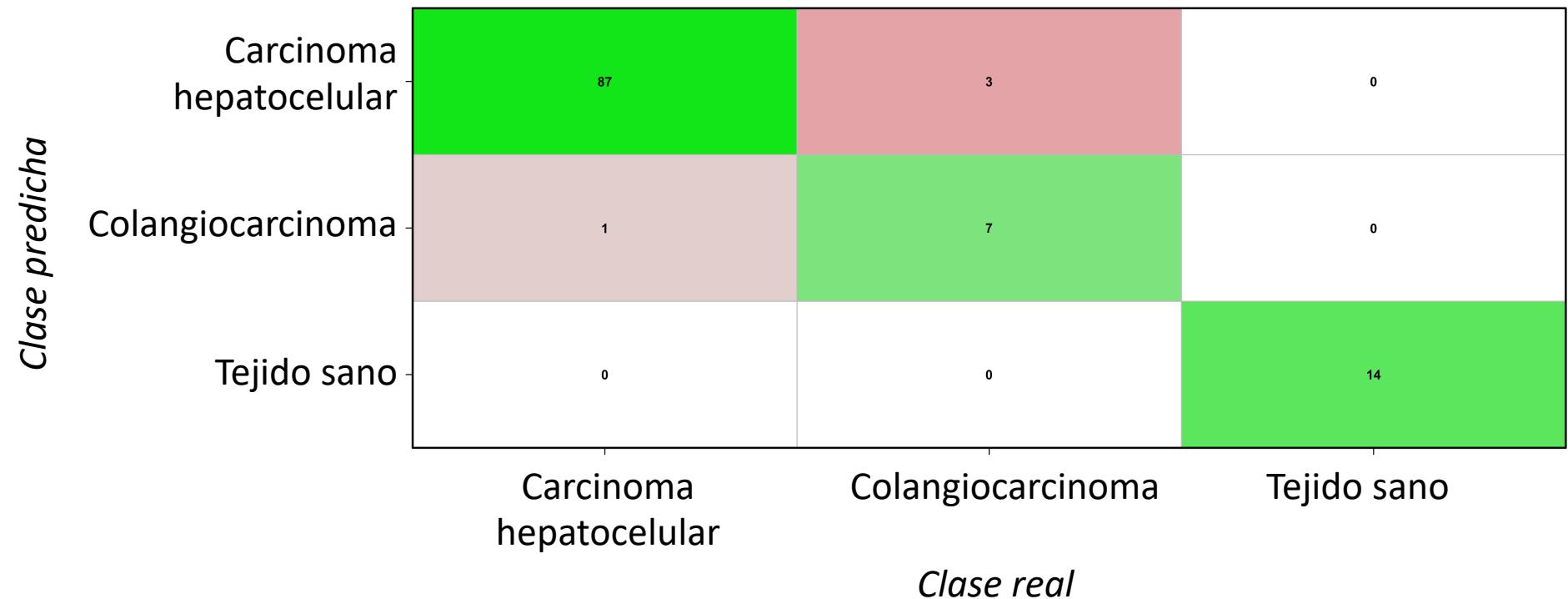
	Biomarcadores	Parámetros	F1 train	Acc train	F1 test	Acc test
SVM	mRMR 7 genes	coste = 1 gamma = 0.025	94,99	97,66	91,84	96,43
RF	mRMR 6 genes	--	94,08	97,08	90,63	95,54
kNN	mRmR 7 genes	k = 7	92,98	96,79	91,84	96,43

**mRMR mejor modelo para detectar biomarcadores.**

**SVM mejor modelo de clasificación.**

# Resultados - Hígado multiclase

## Validación en test



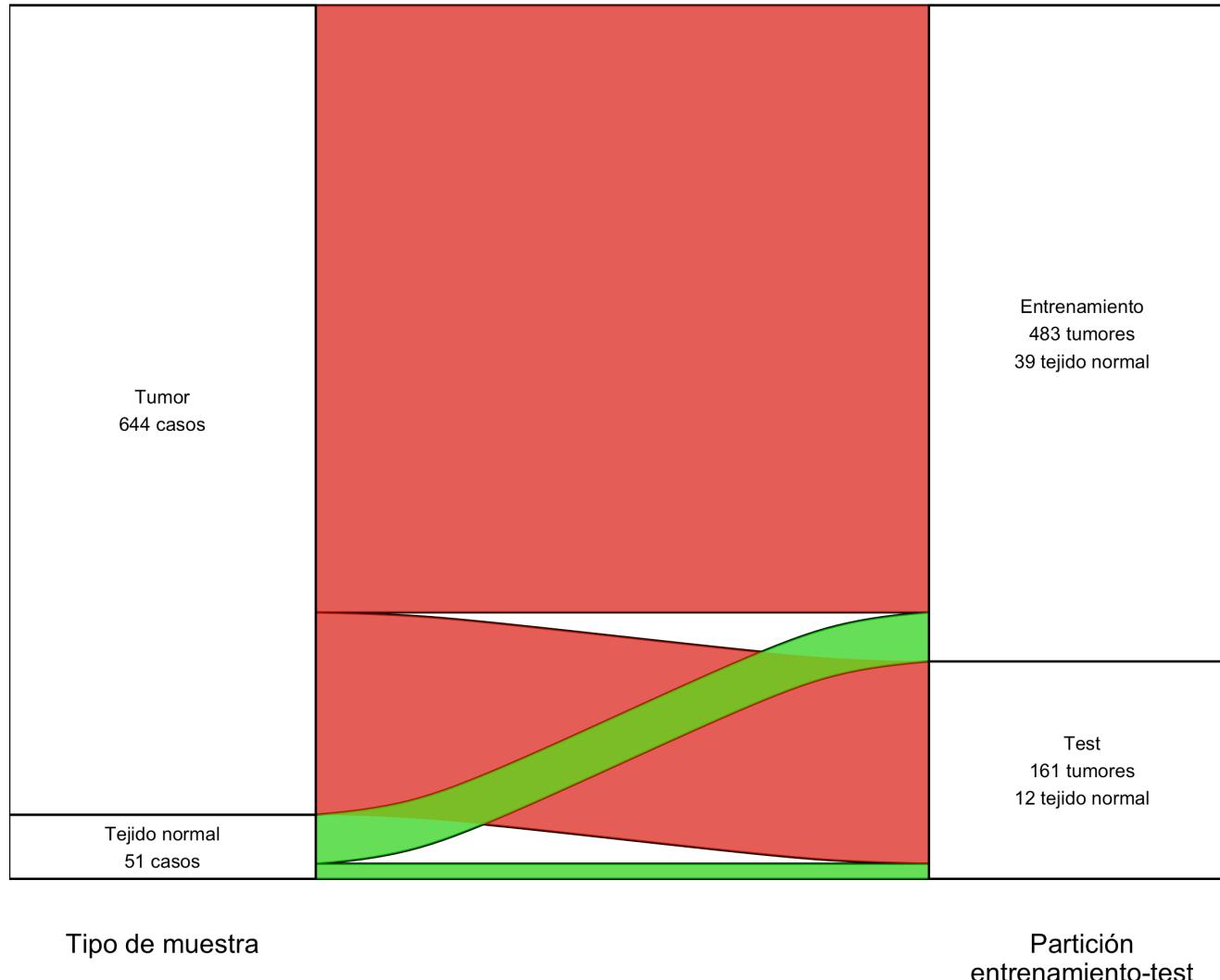
Perfecta discriminación entre tumores y tejidos sanos.

Dificultad en distinguir los dos tipos de tumores.

# Resultados - Colon-recto biclase

Partición en conjuntos de entrenamiento y test

Reparto 75% - 25% con balanceo de clases



# Resultados - Colon-recto biclase

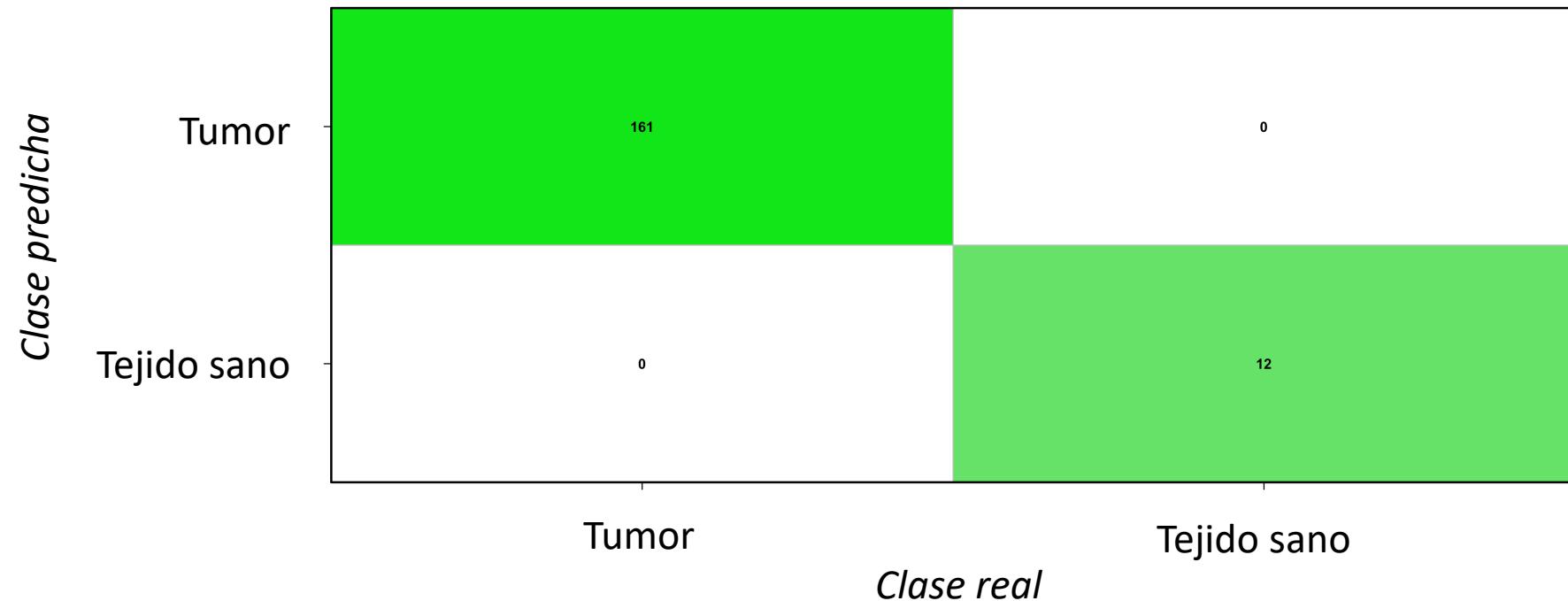
Tabla resumen de los mejores modelos según método de clasificación

	Biomarcadores	Parámetros	F1 train	Acc train	F1 test	Acc test
SVM	mRMR 3 genes	coste = 0,05 gamma = 0,06	100	100	100	100
	RF 3 genes	coste = 0,05 gamma = 0,07	100	100	100	100
RF	mRMR 4 genes	--	100	100	100	100
kNN	RF 3 genes	k = 23	100	100	100	100

Clasificación perfecta con 3 ó 4 genes  
en todos los métodos de clasificación.

# Resultados - Colon-recto biclase

## Validación en test

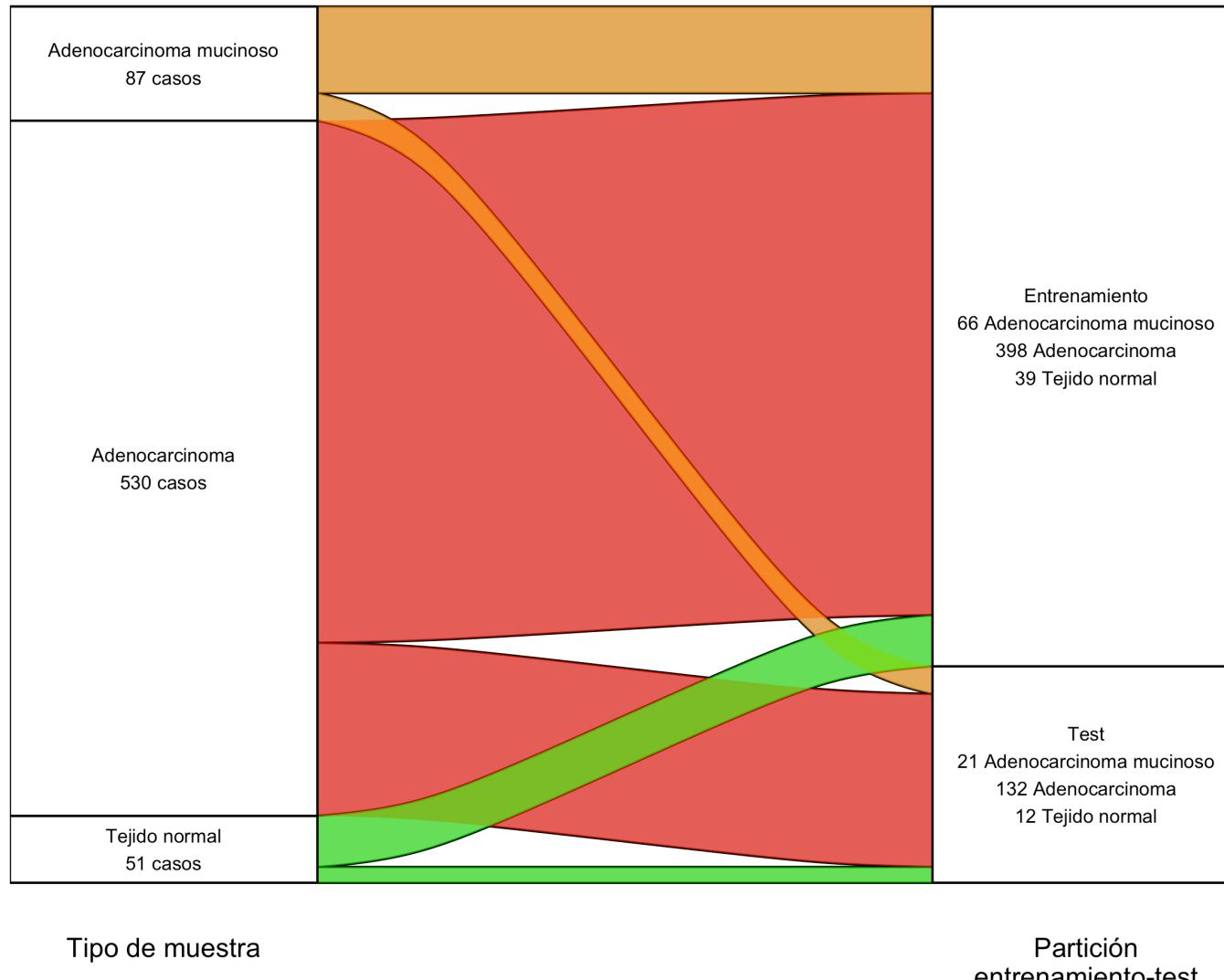


Clasificación perfecta.

# Resultados - Colon-recto multiclas

Partición en conjuntos de entrenamiento y test

Reparto 75% - 25% con balanceo de clases



# Resultados - Colon-recto multiclas

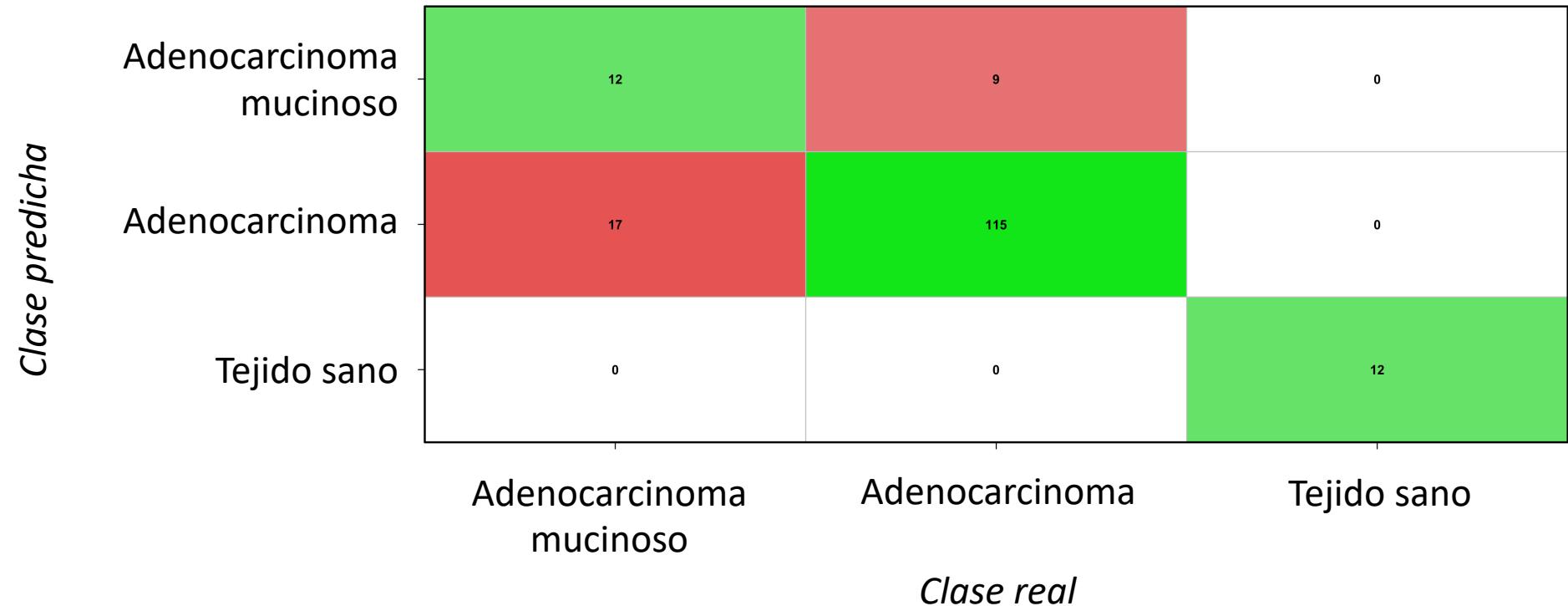
Tabla resumen de los mejores modelos según método de clasificación

	Biomarcadores	Parámetros	F1 train	Acc train	F1 test	Acc test
SVM	RF 3 genes	c = 5 gamma = 0,07	81,61	90,25	80,29	87,27
RF	mRMR 9 genes	--	83,38	90,66	79,28	84,24
kNN	RF 3 genes	k = 7	82,46	90,86	78,06	84,85

RF mejor modelo de clasificación en base a resultados en entrenamiento.  
Poco poder discriminatorio.

# Resultados - Colon-recto multiclase

## Validación en test



Perfecta discriminación entre tumores y tejidos sanos.

Dificultad en distinguir los dos tipos de tumores.

# Resultados

## biomarkerS <https://dredondo.shinyapps.io/biomarkerS/>

- Es una **aplicación web** que realiza análisis transcriptómicos basados en machine learning.
- Útil para **usuarios sin conocimientos previos de programación**. En inglés para favorecer su uso.

### Aspectos técnicos:

- Basada en {shiny}. Versión web y local.
- Mejorada con CSS.
- Contiene tablas interactivas con {DT}.
- Pantallas de carga con {waiter}.



# Resultados

## biomarkerS

<https://dredondo.shinyapps.io/biomarkerS/>

biomarkerS

- Introduction
- Data loading
- Genes selection
- Model training
- Model validation**
- Related diseases
- Authors
- Code

### Model validation

Feature selection algorithm:

mRMR

Classification algorithm (for SVM and kNN it must be trained first to obtain optimal parameters):

SVM

Select the number of genes to use (must be equal or less than the number of genes selected at 'Genes selection'):

10

Validate model in test

Actual	Primary Tumor	Solid Tissue Normal
Primary Tumor	99	2
Solid Tissue Normal	0	14

biomarkerS

### Data loading

Select CSV file with labels (see [here](#) an example)

Browse... higado\_200genes\_labels.csv

Upload complete

Select CSV file with DEGsMatrix (see [here](#) an example)

Browse... higado\_200genes\_degsmatrix.csv

Upload complete

Import file

### Distribution of classes

Label	Samples
Primary Tumor	404
Solid Tissue Normal	58

< > ↗ ↘

# Índice

1. Introducción
2. Metodología
3. Resultados
- 4. Conclusiones**
5. Líneas abiertas de trabajo

# Conclusiones

## Sobre los métodos de selección de características

En general se han obtenido **buenos resultados de clasificación con pocos genes**, lo que tiene muchas ventajas:

- poco **coste computacional**
- abaratamiento de **costes de recolección** de información
- gran **interpretabilidad**, buena **visualización de datos**

Es necesaria una **validación externa** e **interpretaciones clínicas** para establecer de forma clara una asociación gen-enfermedad.

# Conclusiones

## Sobre los algoritmos de clasificación

SVM, *random forest* y kNN obtienen **resultados muy similares**, y consiguen distinguir correctamente entre tejidos tumorales y sanos, con algunos problemas para distinguir entre diferentes tipos de cáncer.

La efectividad de los modelos puede favorecer un **diagnóstico temprano**, y por tanto una **mejora significativa en el pronóstico** (mayor efectividad del tratamiento, más supervivencia, mejora de la calidad de vida).

# Índice

1. Introducción
2. Metodología
3. Resultados
4. Conclusiones
5. Líneas abiertas de trabajo

# Líneas abiertas de trabajo

- Otros **métodos de selección de características** (p. ej. DARED, DA que añade control de redundancia) y de **clasificación** (p. ej. ensemble).
- Combinar análisis de expresión de RNA con análisis de **microRNA** o **alteraciones somáticas** para mejorar eficiencia.
- Mejoras en **{KnowSeq}** y **biomarkeRs**: nuevos gráficos, mejora del tuning, redacción de manual de uso, personalización de entrenamiento y test... El **código abierto** favorece la colaboración.
- **Artículo científico** que sintetiza este trabajo.

# ¡Gracias por la atención!

## Daniel Redondo Sánchez



daniel.redondo.easp@juntadeandalucia.es



@dredondosanchez



danielredondo



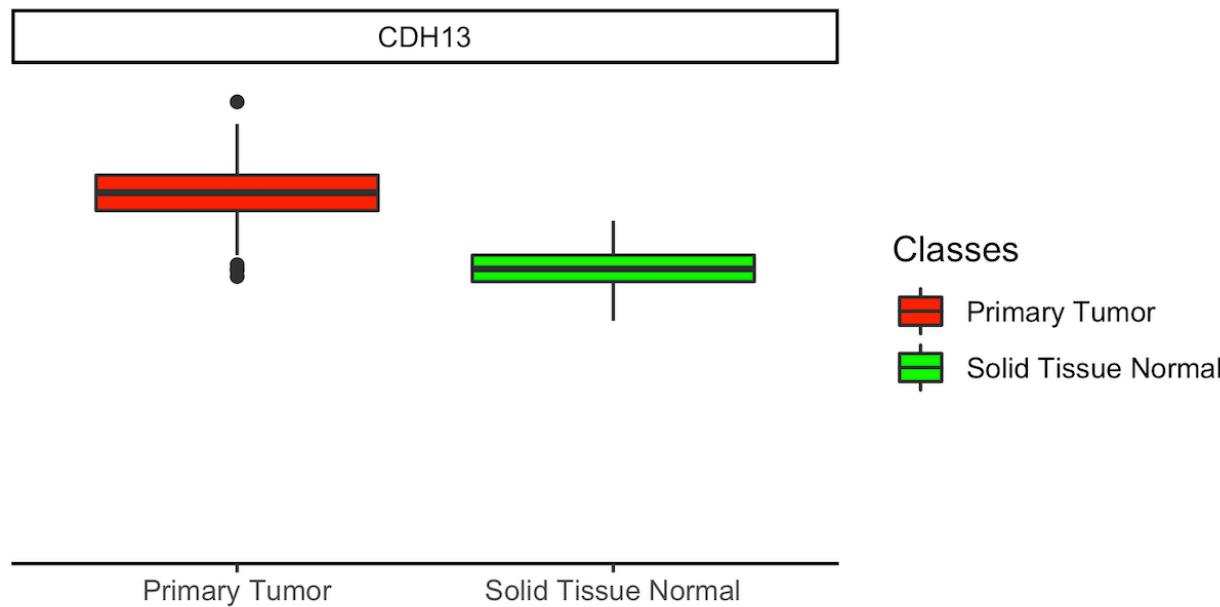
UNIVERSIDAD  
DE GRANADA

# **Diapositivas adicionales**

# Metodología

## Extracción de DEGs (Genes Diferencialmente Expresados):

- p-valor = 0,001
- Controlando por efecto batch con SVA
- Log Fold Change = 1



# Metodología

## Algoritmo mRMR (mínima redundancia, máxima relevancia)

*Etapa inicial:*

Variables seleccionadas =  $S = \text{Conjunto vacío}$

*Paso i:*

Se añade a  $S$  la variable  $X$  que maximiza:

$$I(X, Y) - \frac{1}{|S|} \sum_{W \in S} I(X, W)$$

Donde  $Y$  es la variable resultado, e  $I$  es la función de información mutua entre dos variables.

# Metodología

## Algoritmo RF como algoritmo de selección de características

La importancia de cada variable se mide como la reducción media en precisión del modelo al aleatorizar los valores de la variable manteniendo su distribución.

## Algoritmo DA

Usa una plataforma web (*targetValidation* de *Open Targets*) para conocer las evidencias científicas sobre asociación gen-enfermedad.