Análisis Supervivencia de Cáncer NSCLC

Pacientes con mutación de KRAS

Pere Pavón

4 de junio de 2018

Contents

1 Descripción del estudio original **2**
Abstract . . . 2
Título original de la Tesis doctoral . . . 2
Objetivos . . . 2
Materiales y métodos . . . 2

2 Analisis bioinformático y bioestadístico **3**
2.1 Estudio propuesto para el TFM y Pipeline para el proceso de análisis . . . 3
Objetivos . . . 3
Propuesta de pipeline para el proceso de análisis . . . 3
2.2 Proceso de análisis . . . 4
Obtención de los datos estadísticos . . . 4
Preparación y exploración de los datos . . . 4
Preparación de los datos. . . . 4
Exploración y descripción de los datos. . . . 4
Análisis de normalidad . . . 10
Test de correlación . . . 12
Análisis de supervivencia . . . 14
Supervivencia Libre de Progresión (SLP) . . . 16
Estimación de SLP sin agrupaciones . . . 16
Estimación de SLP con agrupación por sexo y por tipo de mutación KRAS . . . 18
Ajuste de un modelo de regresión de Cox para SLP . . . 21
Modelo de regresión de Cox Univariable (SLP) . . . 22
Modelo de regresión de Cox Univariable estratificado (SLP). . . . 23
Modelo de regresión de Cox Multivariable (SLP) . . . 31
Mejora del modelo de regresión de Cox Multivariable (SLP) . . . 32
Evaluación de validez del mejor modelo de regresión de Cox obtenido (SLP) . . . 34
Análisis de interacciones de covariables en el modelo ajustado (SLP) . . . 36
Resultados de análisis de SLP . . . 37
Supervivencia Global (SGL) . . . 38
Estimación de SGL sin agrupaciones . . . 38
Estimación de SGL con agrupación por sexo y por tipo de mutación KRAS . . . 40
Ajuste de un modelo de regresión de Cox para SGL . . . 42
Modelo de regresión de Cox Univariable (SGL). . . . 42
Modelo de regresión de Cox Univariable estratificado (SGL) . . . 42
Modelo de regresión de Cox Multivariable (SGL) . . . 49
Mejora del modelo de regresión de Cox Multivariable (SGL) . . . 50
Evaluación de validez del mejor modelo de regresión de Cox obtenido (SGL) . . . 52
Resultados de análisis de SGL . . . 54
Conclusiones . . . 54
Comparativa de resultados con estudio original . . . 57
Anexo de codigo R realizado para la generación de las tablas y gráficas . . . 58

Referencias **90**

1 Descripción del estudio original

Abstract

El estudio original parte de una Tesis doctoral de la licenciada en medicina María de los Llanos Gil Moreno (Gil Moreno (2018)) sobre el Carcinoma de Pulmón no célula pequeña (NSCLC). A diferencia de como se ha enfocado hasta el momento el tratamiento dirigido del adenocarcinoma de pulmón con mutación en KRAS como una enfermedad única con un mismo comportamiento, en la Tesis se plantea la hipótesis de que realmente no es así y que se trata de una enfermedad heterogénea con diferentes subtipos, que tienen comportamientos diferentes, tanto a nivel molecular, como a nivel de su evolución, indicando esto que será necesario aplicar tratamientos diferenciados a cada uno de los subtipos.

Título original de la Tesis doctoral

El título original es: *Carcinoma de pulmón no célula pequeña, con mutación de KRAS: diferenciación y caracterización de subtipos, así como los diferentes mecanismos de resistencia, para elaboración de tratamientos dirigidos.*

Objetivos

En la Tesis se plantean cuatro objetivos principales, que son:

1. Subdivisión en subtipos en función de las diferentes mutaciones que presenta KRAS.
2. Caracterizar por subgrupos dependiendo de la expresión génica de 7 genes analizados.
3. Diferencias entre subgrupos: característica, evolución y pronóstico.
4. Estudio de combinación de diferentes fármacos inhibidores de las vías moleculares relevantes que objetiva en el desarrollo previo de su trabajo, como posible dianas terapéuticas en 4 líneas celulares (H23, A549, H460, Calu-6).

Materiales y métodos

Se obtuvieron las muestras de pacientes diagnosticados de cáncer del pulmón célula no pequeña con mutación de KRAS, de diferentes centros previa aprobación del CEIC (Comité de ética e investigación clínica y consiguiendo los consentimientos informados de los pacientes según la ley 14/2007 de investigación biomédica. Se efectuó una valoración patológica de las muestras, tanto cuantitativa como cualitativa y de tipo de disección, para seleccionar finalmente las muestras a tratar en el estudio.

Se efectuó un análisis de las mutaciones del gen KRAS de los codones 12 y 13 mediante la técnica HRM (High Resolution Melting) de Applied Biosystems.

El análisis de la expresión génica de ARNm de los genes de interés, se realizó a partir de la obtención de RNA de las muestras primarias, mediante una lisis celular con una solución de EDTA 0.1M, SDS 2%, tris 0.5M y proteinasa K (20 mg/ml). El cDNA se sintetiza usando un Kit basado en la encima de retrotranscripción MMLV (200u/μL)/RNASE OUT (20u/μL), realizando controles de técnica: controles de extracción y calibradores. La amplificación se realizó mediante PCR a Tiempo Real introduciendo por triplicado 2.5 μLde ADNc obtenido de las muestras, con un sistema de detección ABI Quanti studio Flex de Applied Biosystems, utilizando sondas y cebadores específicos para cada uno de los genes a analizar. La cuantificación se realiza a partir de los valores de fluorescencia que se transforman mediante un software al valor CT, que es el número de ciclos necesarios para que haya un aumento de fluorescencia significativo respecto a la señal de base. Se realiza una validación de sondas y cebadores, mediante un control de calidad in silico.

El valor relativo de expresión se calculó utilizando el método de Ct comparativo usando $\beta - actin$ como endógeno y RNA comercial como calibrador (Liver and Lung; Stratagene, La jolla, CA USA). Los estadísticos

calculados para la expresión génica fueron: promedio de las tres réplicas, dCt de cada muestra para cada uno de los genes analizados, que se define como el promedio de la diferencia entre el gen problema con el gen endógeno, como δCt o dCt, la media de dCt de los calibradores y por último el $ddCt$ para las muestras y los calibradores, calculado según la fórmula: $ddCt = 2 - dCt_{gen} - dCt_{liver-lung}$. Adicionalmente, se realizan experimentos in Vitro para valorar el comportamiento de las líneas celulares tumorales H23, A549, H460 y Calu-6, frente a inhibidores antitumorales.

A nivel estadístico se realiza un estudio de SLP, supervivencia libre de progresión, que se definió como el tiempo desde el inicio del tratamiento de primera línea hasta la progresión o muerte y un estudio de la SG, supervivencia global, definida como el tiempo desde el inicio del tratamiento de la primera línea hasta la muerte. Se utilizaron modelos de regresión de riesgos proporcionales de Cox para los análisis multivariantes. Todo el análisis estadístico se realizó con el software SPSS v15 (Statistical Package for Social Sciencies).

2 Analisis bioinformático y bioestadístico

2.1 Estudio propuesto para el TFM y Pipeline para el proceso de análisis

Objetivos

Para consecución del tercer objetivo global del presente trabajo sobre el estudio de las implicaciones de KRAS en el cáncer de pulmón, se propone efectuar un análisis bioinformático y bioestadístico, basado en un estudio ya realizado (Tesis doctoral de Gil Moreno (2018)), para poder evidenciar a través del análisis estadístico de la expresión de varios genes implicados en líneas tumorales con mutación de KRAS, su implicación en la supervivencia de los pacientes libres de progresión tumoral (SLP) y en la supervivencia global (SG), para realizar posteriormente una comparativa de los resultados obtenidos respecto al estudio original. El trabajo original de la Tesis doctoral, abarca un ámbito de trabajo mayor, pero para el presente trabajo se ha seleccionado únicamente la parte del estudio realizado sobre muestras tumorales de adenocarcinoma de pulmón con mutación de KRAS, cuyo resultado se describe como parte A de los resultados obtenidos en el trabajo realizado para la Tesis, referente a la posible influencia de la expresión diferencial de los genes analizados en la supervivencia de los pacientes.

Los datos de las muestras del estudio han sido cedidos por la Dra. María de los Llanos Gil, para fines únicamente de realización del presente TFM, para compartir y contrastar posteriormente los resultados obtenidos. Todo el proceso se efectuará en el entorno de programación y plataforma estadística R (Open Access Statistics - R Consortium (2018)), generando un código reproducible de todas las funciones utilizadas, así como de un informe dinámico en RMarkdown, que será anexado al TFM como uno de los productos finales del trabajo. El informe dinámico será diseñado de forma que pueda ser reaprovechado para estudios similares posteriores, donde únicamente se deba modificar la fuente de datos origen del estudio. Tanto los textos, tablas y gráficas obtenidas, se incluirán en el capítulo 4 de la memoria del TFM.

Propuesta de pipeline para el proceso de análisis

Se propone el siguiente Pipeline de trabajo para la realización del análisis:

1. Obtención de los datos estadísticos del estudio
2. Preparación y exploración de los datos:

- 2.1 Preparación de los datos
- 2.2 Exploración de los datos
- 2.3 Análisis de normalidad y correlación

3. Proceso de análisis de Supervivencia
4. Conclusiones.
5. Comparativa de resultados con estudio original

2.2 Proceso de análisis

Obtención de los datos estadísticos

Los datos obtenidos para el estudio se encuentran en el fichero ***BaseFinalAnalisisEstadísticoKRAS_DraGil.csv*** que está ubicado en el directorio de trabajo *./Datos*. El fichero original del estudio contiene 32 registros de muestras de adenocarcinomas de pulmón, cada uno con 79 variables que corresponden a los datos clínicos recogidos para el estudio, los datos de seguimiento de los tratamientos administrados a los pacientes y los datos sobre los cálculos de expresión génica realizados a los genes objetivo del estudio original. Se han descartado 2 registros por falta de los datos básicos necesarios para la realización del análisis de supervivencia, por lo que finalmente la muestra contiene 30 pacientes. Se ha generado el data frame *dataKras* para el tratamiento posterior de los datos

Preparación y exploración de los datos

Preparación de los datos.

Para la realización del análisis se han seleccionado únicamente las variables requeridas para la obtención del análisis de supervivencia de los datos completos del estudio. Se ha preparado el data frame *dataKras*, que contiene las mismas 30 muestras seleccionadas, pero solo con las 48 variables que pueden ser de interés para el análisis. Se han factorizado las variables categóricas de mayor interés, para realizar posteriormente una mejor exploración de los datos.

Exploración y descripción de los datos.

A continuación, en la tabla 1, se muestra la relación de variables que contiene *dataKras*, algunas de las cuales serán objeto del análisis que se realiza en los siguientes apartados.

Tabla 1: Descripción variables del conjunto de datos *dataKras*

Id	Nombre Variable	Descripción	Clase
1	Nº.KRAS	Identificador de paciente	character
2	Date.of.Dx	Fecha de diagnóstico del cáncer	character
3	DOB	Fecha de nacimiento	character
4	Gender	Genero	factor
5	Smoking.history	Historial fumador	integer
6	PA.YEAR	Consumo de paquetes por año	integer
7	Histology	Histologia del tumor	factor
8	DateDxM1	Fecha de diagnóstico de primera metástasis	character
9	TtratStartL1	Fecha inicio tratamiento 1a línea	character
10	TratEndL1	Fecha fin tratamiento 1a línea	character
11	X2nd.line	Estado progresión en L1	character
12	TtratStartL2	Fecha inicio tratamiento 2a línea	character
13	TratEndL2	Fecha fin tratamiento 2a línea	character
14	X3rd.line	Estado progresión en L2	character
15	TtratStartL3	Fecha inicio tratamiento 3a línea	character
16	TratEndL3	Fecha fin tratamiento 3a línea	character
17	X4th.line	Estado progresión en L3	character
18	TtratStartL4	Fecha inicio tratamiento 4a línea	character
19	TratEndL4	Fecha fin tratamiento 4a línea	character
20	X5th.line	Estado progresión en L4	character
21	TtratStartL5	Fecha inicio tratamiento5a línea	character
22	TratEndL5	Fecha fin tratamiento 5a línea	character

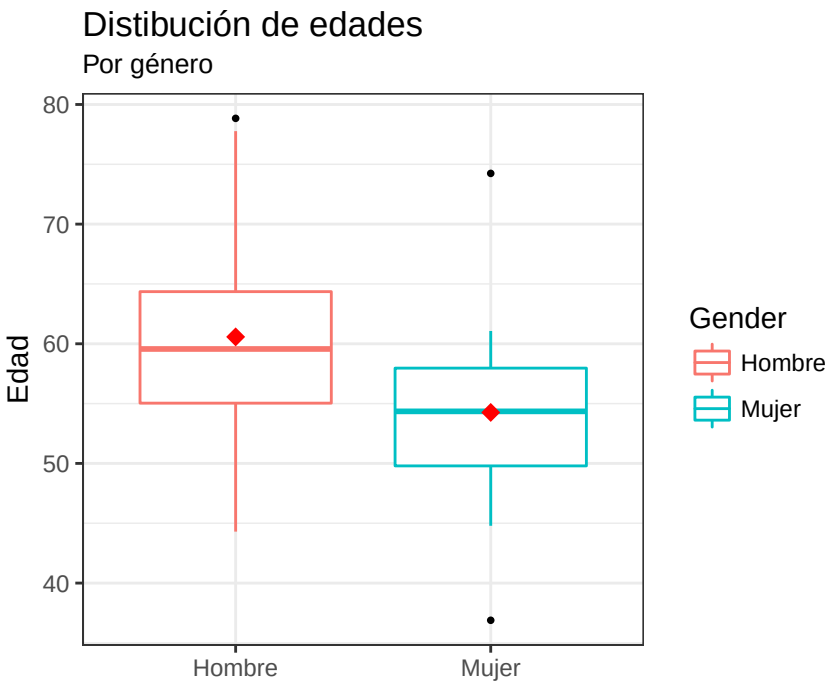
Id	Nombre Variable	Descripción	Clase
23	X6th.line	Estado progresión en L5	character
24	survival.status	Status de supervivencia	character
25	date.of.death	Fecha de la muerte	character
26	Death.Lung.Cancer	0 - No muerte por cáncer de pulmón 1- Muerte por cáncer de pulmón	character
27	DateLastFollowUpS	Fecha del último seguimiento	character
28	PFS	Tiempo desde el inicio del tratamiento hasta la progresión (original)	integer
29	OS	Tiempo desde metástasis hasta la muerte (original)	character
30	Mutacion	Tipo de mutación de KRAS	factor
31	ddCt.HES1	Valor ddCt expresión gen HES1	numeric
32	ddCt.CDCP1	Valor ddCt expresión gen CDCP1	numeric
33	ddCt.AXL	Valor ddCt expresión gen AXL	numeric
34	ddCt.YAP1	Valor ddCt expresión gen YAP1	numeric
35	ddCt.CREB	Valor ddCt expresión gen CREB	numeric
36	ddCt.LKB1	Valor ddCt expresión gen LKB1	numeric
37	ddCt.Src	Valor ddCt expresión gen Src	numeric
38	ddCt.STAT3	Valor ddCt expresión gen HES8	numeric
39	Comments	Comentarios	character
40	FecFinGl	Fecha final para supervivencia global	character
41	CensGl	Censura en supervivencia globa	integer
42	TimeGl	Tiempo en supervivencia global	integer
43	FecFinSLP	Fecha final para supervivencia libre progresión	character
44	CensSLP	Censura en supervivencia libre progresión	integer
45	TimeSLP	Tiempo en supervivencia libre progresión	integer
46	DateCalEdad	Fecha del momento para cálculo de edad (hasta fecha inicio tratamiento 1a línea o fecha de diagnóstico)	character
47	Smoking.historyFac	Historial fumador factorizado	factor
48	edad	Edad del paciente en el inicio de tratamiento 1a línea o diagnóstico	numeric

En primer lugar vamos a proceder a ver las características principales de los pacientes que componen la muestra de estudio en base a las variables disponibles. Hay un total de 30 de pacientes en la muestra de estudio, distribuidos por género según se puede ver en la tabla 2, donde la mayoría de los pacientes son hombres con un *73.33%*.

Table 2: Distribución de pacientes por género

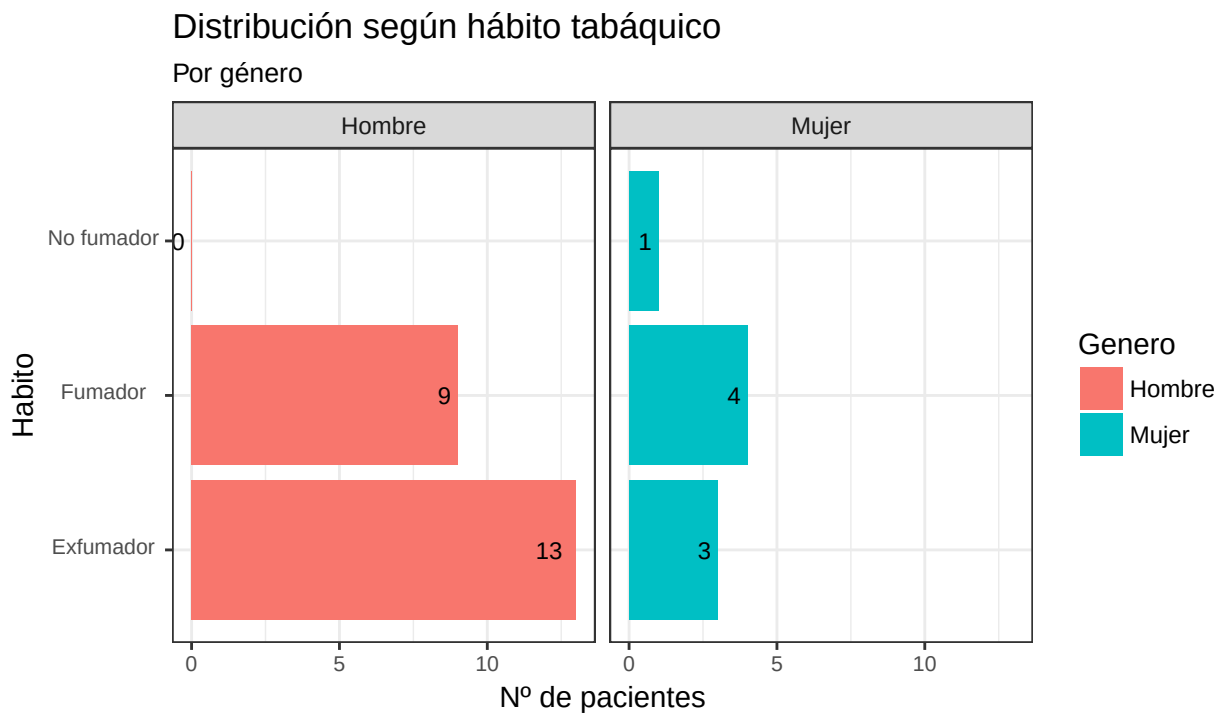
Género	Frecuencia	%
Hombre	22	73.33
Mujer	8	26.67

La edad media global de los pacientes es de 58.89 años. En la gráfica Gr.1 se muestra la distribución de las edades de los pacientes por género, donde se puede ver que la edad media de las mujeres es inferior a la de los hombres, de 54.26 y 60.58 años respectivamente. La mediana global de edad de la muestra es de 58.61 años.

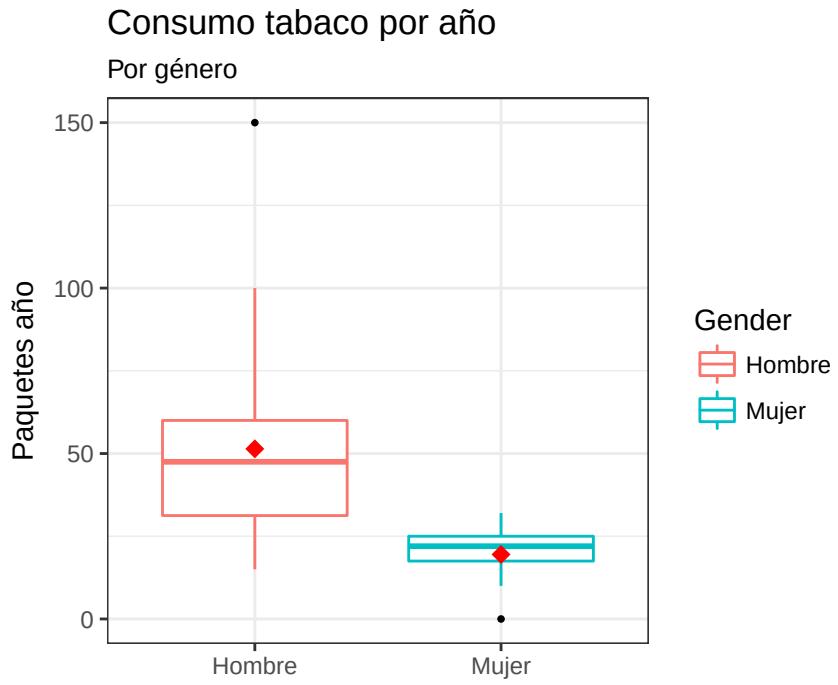


Gráfica Gr.1

En la gráfica Gr.2 se presenta la distribución de pacientes de la muestra respecto su hábito de consumo de tabaco, clasificados en función del sexo. Se puede observar que prácticamente todos los pacientes eran fumadores o exfumadores, solo hay dos casos que nunca habían fumado. Aunque la muestra ya determina su mayor porcentaje de hombres que mujeres, se puede ver que la proporción de fumadores hombres es el doble de mujeres, aunque el número de exfumadores es mucho mayor en hombres. En referencia al consumo, si se cuenta el consumo de los fumadores y el consumo que tenían los exfumadores, la media del consumo está en 44 paquetes por año. Si se analiza por género, en la gráfica Gr.3 se puede ver que a nivel de consumo de paquetes de tabaco por año, el hombre consume mucho más que la mujer, con un consumo medio de unos 51 paquetes por año en hombres y unos 22 paquetes por año en mujeres.

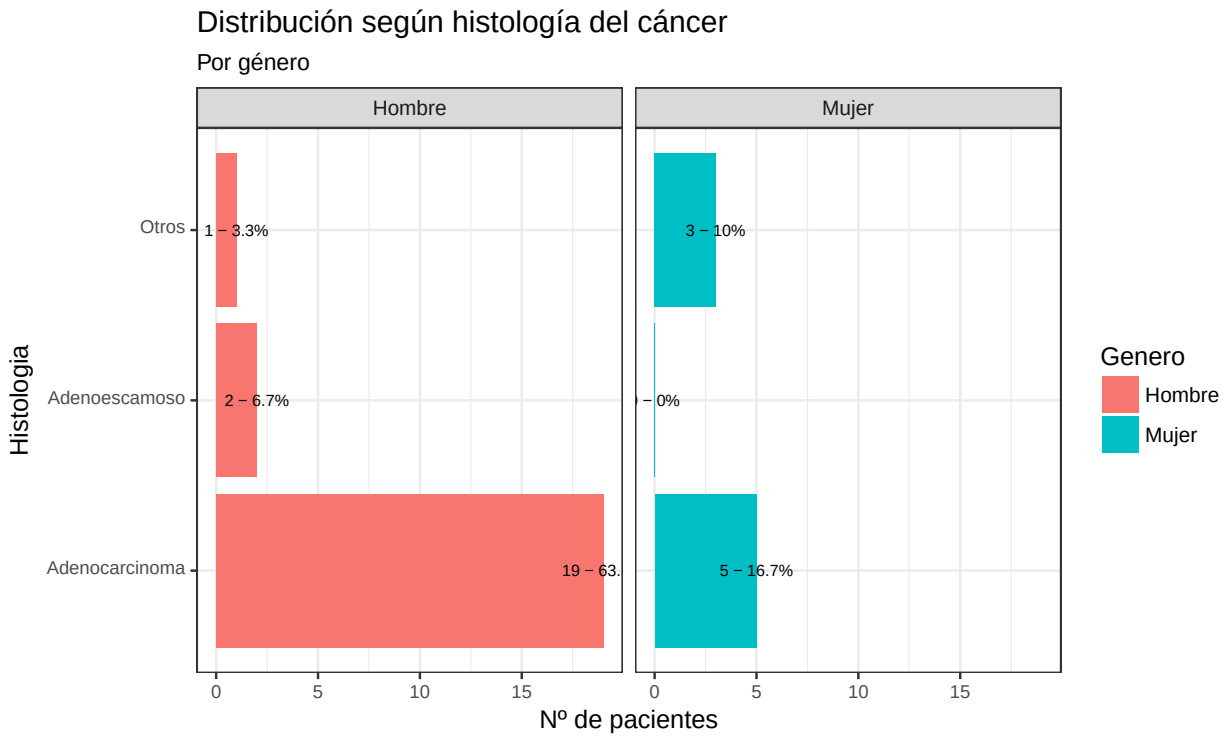


Gráfica Gr.2



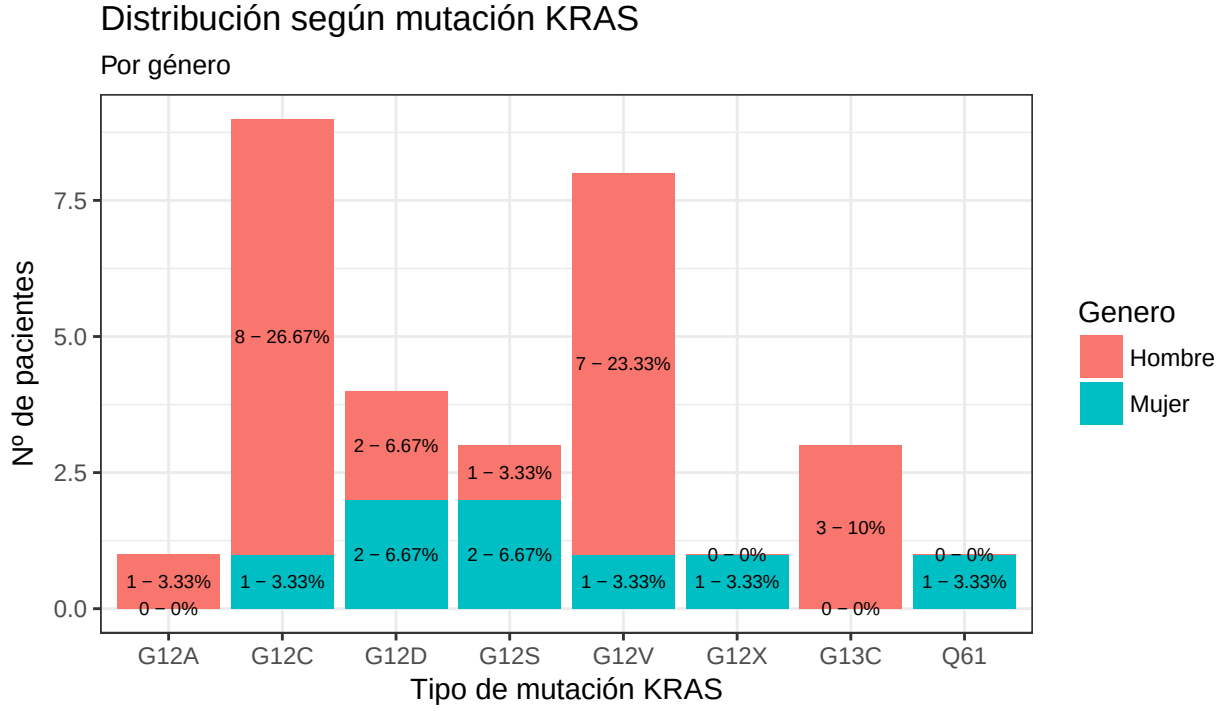
Gráfica Gr.3

Respecto a la histología de los pacientes de la muestra (gráfica Gr.4), el *80%* eran *Adenocarcinomas* y el *6.7%* son *Adenoescamosos* En ambos sexos el *Adenocarcinoma* es el más frecuente, aunque en proporción ha sido más habitual en los hombres que en las mujeres.



Gráfica Gr.4

Respecto a las características moleculares de las muestras, el análisis de la mutación de KRAS que se ha realizado muestra que los tres tipos de mutación más frecuentes son G12C y G12V con un 28.12%, seguido de G12C y G12S con un 9.38 cada una de ellas (ver gráfica Gr5).



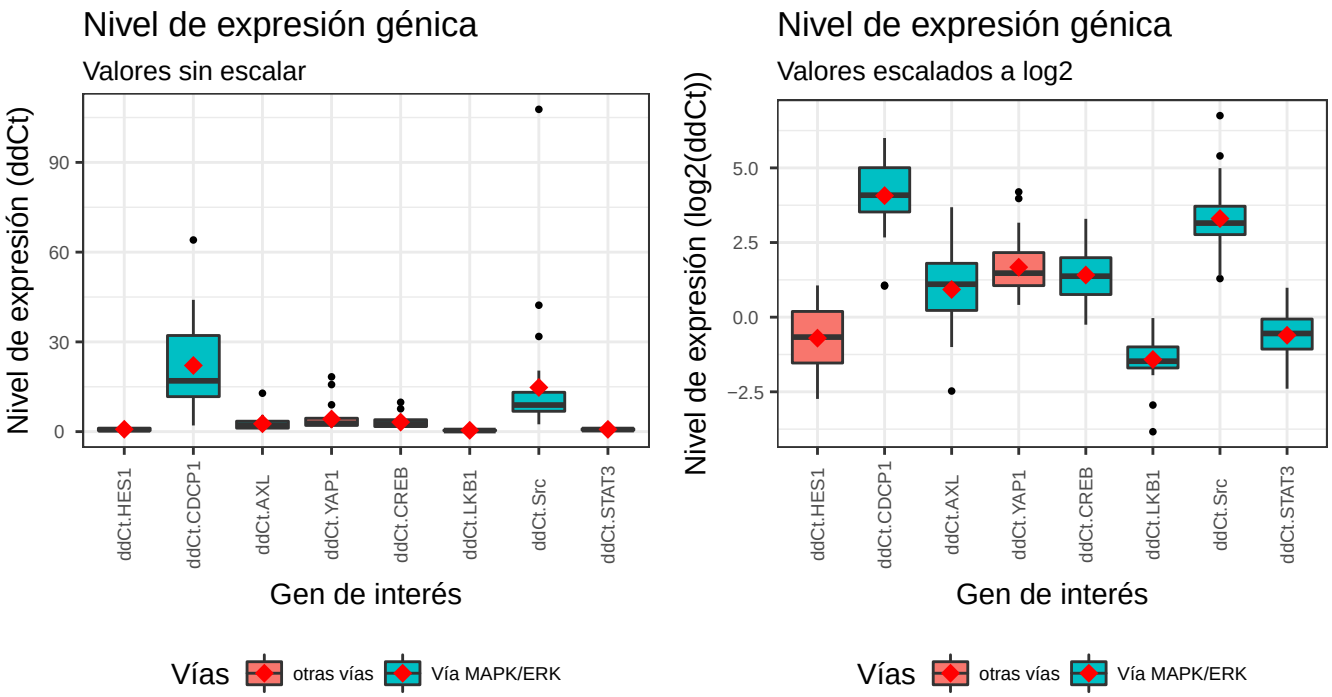
Gráfica Gr.5

Por otra parte, en el proyecto original se identificaron 8 genes que se suponen están implicados en los mecanismos de activación de las vías de señalización celular donde KRAS aparece mutado y que pueden generar que las terapias aplicadas a los pacientes no tengan el resultado esperado. Estos 8 genes parecen estar implicados como activadores o inhibidores de las vías MAP quinasas con KRAS mutado. Hay 6 de ellos que están relacionados más directamente con las vías de señalización MAPK/ERK y otros 2 que aunque no directamente y dada la redundancia de las vías acaban influyendo también en esta vía. Como una primera exploración de las expresiones de estos genes, se muestra a continuación un resumen estadístico de los mismos, para ver numéricamente sus características básicas, media, mediana, máximo, mínimo, NAs, etc.

```
   ddCt.HES1        ddCt.CDCP1       ddCt.AXL        ddCt.YAP1
 Min.   :0.1500   Min.   : 2.06   Min.   : 0.180   Min.   : 1.330
 1st Qu.:0.3450   1st Qu.:11.70   1st Qu.: 1.170   1st Qu.: 2.080
 Median :0.6300   Median :16.98   Median : 2.150   Median : 2.780
 Mean   :0.7953   Mean   :22.12   Mean   : 2.625   Mean   : 4.202
 3rd Qu.:1.1450   3rd Qu.:32.14   3rd Qu.: 3.490   3rd Qu.: 4.480
 Max.   :2.0900   Max.   :64.06   Max.   :12.840   Max.   :18.320
 NA's   :11       NA's   :2       NA's   :5        NA's   :5
   ddCt.CREB        ddCt.LKB1         ddCt.Src        ddCt.STAT3
 Min.   :0.840   Min.   :0.0700   Min.   :  2.45   Min.   :0.1900
 1st Qu.:1.695   1st Qu.:0.3075   1st Qu.:  6.80   1st Qu.:0.4775
 Median :2.590   Median :0.3600   Median :  8.89   Median :0.6850
 Mean   :3.174   Mean   :0.4400   Mean   : 14.77   Mean   :0.7508
 3rd Qu.:3.985   3rd Qu.:0.5050   3rd Qu.: 13.15   3rd Qu.:0.9575
 Max.   :9.810   Max.   :0.9800   Max.   :107.73   Max.   :1.9800
 NA's   :7       NA's   :10       NA's   :2        NA's   :4
```

Los datos muestran que hay una gran variabilidad de valores entre la expresión de los diferentes genes y además en cada uno de ellos existen valores omitidos, es decir, que no de todos ellos disponemos de valores de expresión para todos los pacientes, posiblemente debido a la calidad de las muestras analizadas en cada caso. Los genes con mayor número de valores omitidos son el HES1 y el LKB1 y los que menos valores omitidos presentan han sido el gen CDCP1 y Scr.

En el siguiente gráfico de cajas, Gr.6, se muestra la distribución de los valores de expresión de los 8 genes para ver si guardan cierta homogeneidad. Como se puede comprobar los datos de expresión no son homogéneos entre los diferentes genes y por tanto la media parece no ser una buena medida de resumen entre las diferentes expresiones. Tal como se había podido comprobar en el resumen numérico, debido a las diferencias de escalas de los resultados de expresión de los diferentes genes, no se pueden apreciar correctamente las diferencias, por lo que en el gráfico Gr.7 se muestran los datos de expresión donde se ha efectuado un re-escalado de los valores aplicando una transformación logarítmica de los mismos, para que de esta forma que se aproximen a una escala comparable. En ambos gráficos se han agrupado los genes en base a su implicación o no en la vía de señalización de interés. En cualquier caso, se confirma que hay divergencia en la distribución de los valores de expresión en los diferentes genes. En los siguientes apartados verificaremos la normalidad o no de los mismos para la aplicación de funciones estadísticas paramétricas o no paramétricas.

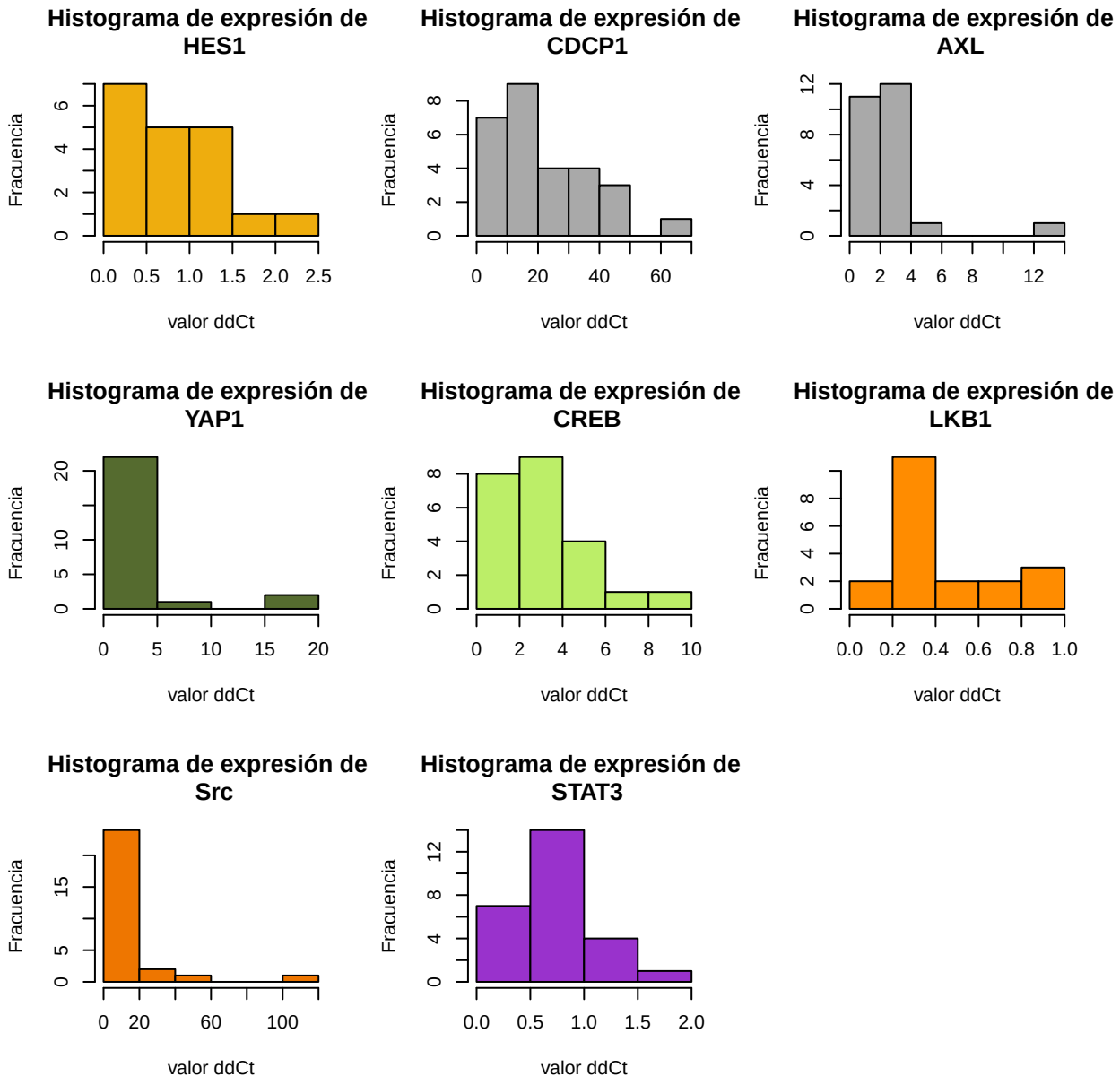


Gráficas Gr.6 y Gr.7

Análisis de normalidad

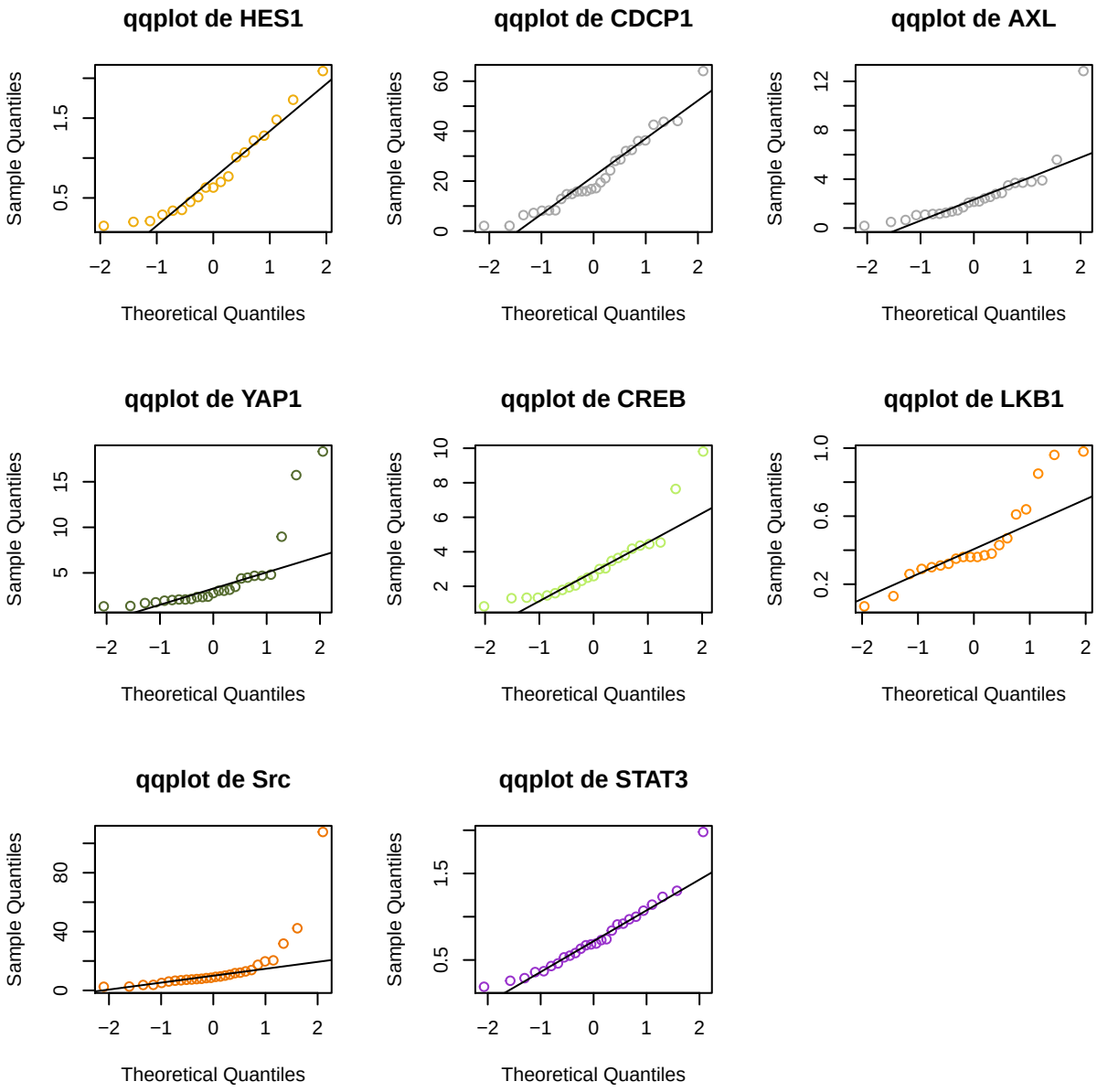
Un dato estadístico de interés a analizar para el estudio sobre las variables de expresión génica es su correlación. Puede ser importante ver si existe correlación entre la expresión de los diferentes genes objetivos del estudio. Con esta finalidad y para poder determinar cuál es el mejor test de correlación a aplicar, se debe efectuar un análisis de normalidad de estas variables de expresión para escoger entre un test de correlación paramétrico, en caso de evidencia de normalidad en los datos, o un test no paramétrico para datos de distribución no normal.

En primer lugar veamos gráficamente si los datos de expresión de los genes presentan a simple vista una distribución normal o no. Para ello se presentan los histogramas (gráfico Gr.8) de cada una de las variables de expresión de los diferentes genes.



Gráfica Gr.8

Como se puede apreciar, ninguno de los genes parece presentar una distribución simétrica de los datos de expresión, por lo que a simple vista no parece no presentar una distribución normal. Aun así, para ver de forma más clara este aspecto, en las gráficas QQ plot Gr.9 se presenta la desviación entre los valores de expresión respecto la recta normal.



Gráfica Gr.9

Según se puede ver en las gráficas QQ de normalidad, muchas de las expresiones se aproximan a la lineas de normalidad, pero en todos los casos hay algunos valores extremos que se desvían de la recta por las colas, por lo que no se puede confirmar la distribución normal de los datos. Para verificar estadísticamente si es cierta esta falta de normalidad en los valores de expresión de los genes, a continuación se va a aplicar el test de Shapiro-Wilk a cada una de las variables, para determinar si hay significación en el test para confirmar o no su distribución normal.

Viendo los resultados del test de normalidad de *Shapiro-Wilks*, se confirma que todas las distribuciones de la expresión de los genes tienen un p-valor < 0.05 por lo que se descarta la distribución normal, a excepción del gen STAT3 que tiene un p-valor > 0.05 por lo que en este caso no se puede descartar la normalidad.

Test de correlación

Table 3: Test de normalidad Shapiro-Wilk de expresiones génicas

gen	W	p.value
Src	0.51155	0.00000
YAP1	0.61403	0.00000
AXL	0.68917	0.00001
CREB	0.82433	0.00096
LKB1	0.86663	0.01026
CDCP1	0.92757	0.05355
STAT3	0.92661	0.06439
HES1	0.91484	0.09086

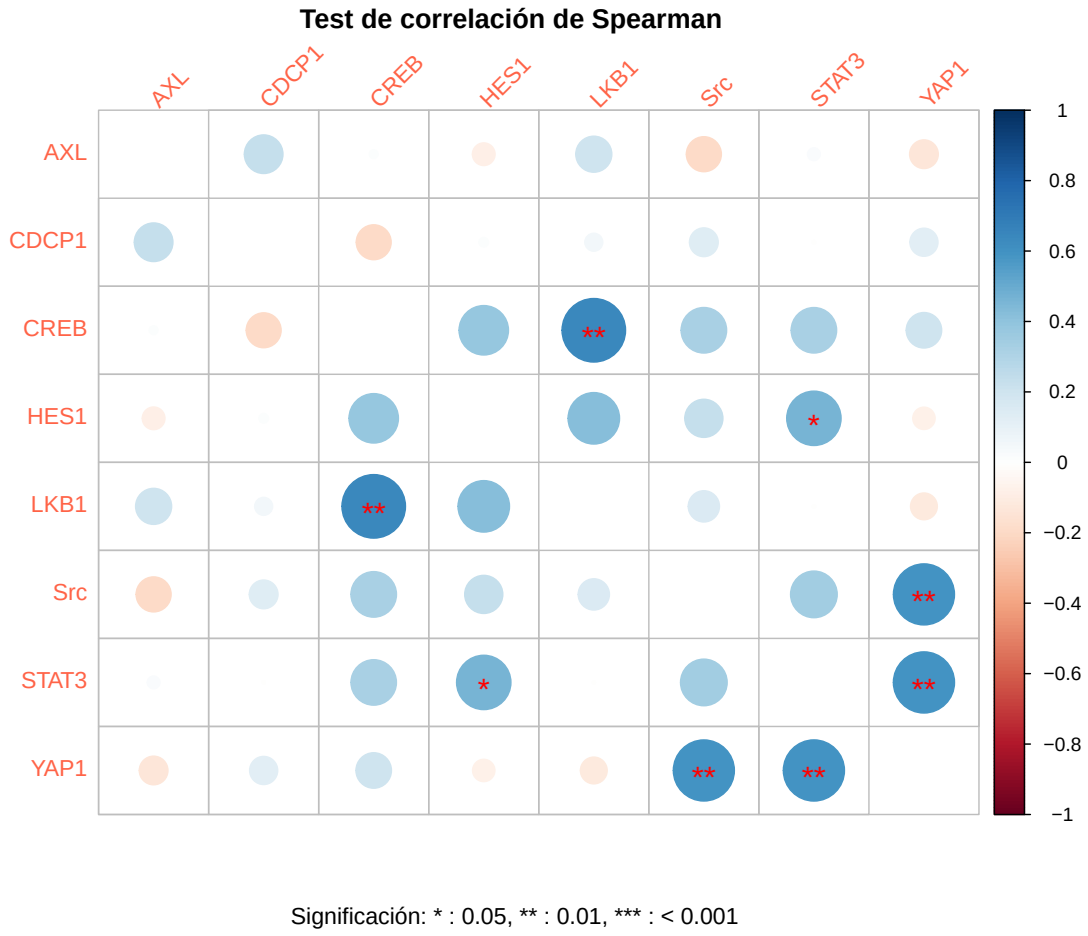
Para ahora determinar la existencia de correlación entre las expresiones de los diferentes genes, se debe aplicar un test de correlación. En este sentido, como se indicaba anteriormente, existen dos tipos de test, los paramétricos en los que se asume la normalidad de los datos y los no paramétricos en la que no es necesaria la asunción de normalidad de los datos. En base al resultado obtenido de no normalidad al aplicar el test de *Shaphiro-Wilks* para verificar la normalidad de los datos de la expresión, se aplicará el *test Spearman* no paramétrico de correlación. Dado que la implementación del test de correlación de R *cor.test()* es para realizarlas entre parejas, se ha definido la función rs.test() en R para obtener los resultados de correlación entre todas las parejas posibles de los 8 genes estudiados. En el anexo del código R se puede consultar la función.

Table 4: Correlación de Spearman entre expresiones génicas

gen	r	AXL	CDCP1	CREB	HES1	LKB1	Src	STAT3	YAP1
AXL	rho	1.0000	0.2369	0.0119	-0.0816	0.2063	-0.1948	0.0257	-0.1306
	p-value	0.0000	0.2530	0.9582	0.7476	0.3828	0.3601	0.9083	0.5526
CDCP1	rho	0.2369	1.0000	-0.1937	0.0140	0.0512	0.1313	-0.0017	0.1255
	p-value	0.2530	0.0000	0.3758	0.9545	0.8302	0.5124	0.9947	0.5501
CREB	rho	0.0119	-0.1937	1.0000	0.3882	0.6415	0.3296	0.3286	0.2000
	p-value	0.9582	0.3758	0.0000	0.1114	0.0031	0.1245	0.1257	0.3601
HES1	rho	-0.0816	0.0140	0.3882	1.0000	0.4211	0.2308	0.4695	-0.0791
	p-value	0.7476	0.9545	0.1114	0.0000	0.0818	0.3418	0.0425	0.7476
LKB1	rho	0.2063	0.0512	0.6415	0.4211	1.0000	0.1551	-0.0015	-0.1138
	p-value	0.3828	0.8302	0.0031	0.0818	0.0000	0.5137	0.9950	0.6328
Src	rho	-0.1948	0.1313	0.3296	0.2308	0.1551	1.0000	0.3456	0.5928
	p-value	0.3601	0.5124	0.1245	0.3418	0.5137	0.0000	0.0843	0.0018
STAT3	rho	0.0257	-0.0017	0.3286	0.4695	-0.0015	0.3456	1.0000	0.5943
	p-value	0.9083	0.9947	0.1257	0.0425	0.9950	0.0843	0.0000	0.0017
YAP1	rho	-0.1306	0.1255	0.2000	-0.0791	-0.1138	0.5928	0.5943	1.0000
	p-value	0.5526	0.5501	0.3601	0.7476	0.6328	0.0018	0.0017	0.0000

En la tabla 4 se muestran los resultados del test de correlación de *Spearman* para los diferentes valores de expresión de los genes analizados, en forma de matriz. Se muestran los valores de correlación ρ *(rho)* de Spearman y el *p-valor* obtenido entre cada pareja. Se puede apreciar que se aprecian diferentes correlaciones significativas (p-valor < 0.05) entre algunos de los genes. Para ver estas correlaciones de forma más clara, en el gráfico Gr.10 se pueden identificar tanto los valores de correlación, como su grado de significación (p-valor).

El grado de correlación vienen identificado por el área e intensidad del color de los círculos mostrados, cuanto mayor es su área y mayor intensidad de color, mayor es el valor de correlación *rho*, el color azul indica correlación positiva y el rojo correlación negativa. El grado de significación viene marcado por los asteriscos (*,**,***).



Gráfica Gr.10

Los genes LKB1 y CREB son los que presentan una mayor correlación de expresión, seguida de los genes YAP1 y SRC, de YAP1 y STAT3, y con menor fuerza de correlación entre STAT3 y HES1.

Análisis de supervivencia

Respecto al proceso de análisis de supervivencia, es necesario indicar que debido a las características especiales del tipo de estudio realizado sobre pacientes de cáncer de pulmón con mutación en KRAS a los que se les han aplicado diferentes terapias de tratamiento en momentos diferentes en temporalidad, no existe un periodo de análisis concreto definido para realizar el estudio de supervivencia, como sería la forma más convencional de realizar este tipo de estudios. Los datos recogidos para el estudio fueron obtenidos de diferentes fuentes, como se describió anteriormente, por lo que cada paciente pertenece a un periodo de tratamiento determinado diferente, que varía desde el año 2008 al año 2016. Por tanto, el estudio de supervivencia se efectúa relativo al inicio del tratamiento de primera línea una vez diagnosticada la primera metástasis de cada paciente o la fecha del diagnóstico del cáncer o la primera metástasis, en función del dato disponible.

De todos los pacientes que inicialmente formaban parte del estudio, se descartaron aquellos sobre los cuales

el análisis molecular del tejido tumoral no obtuvo los mínimos exigidos de calidad para obtener resultados correctos. Tal como se indicó en el apartado de descripción de los datos, el conjunto final de muestras está compuesto por 30 pacientes, de los cuales se descartan adicionalmente 2 pacientes por no disponer datos clínicos suficientes para poder establecer el periodo de tratamiento para calcular el estadístico de supervivencia. Se verá a continuación que en los diferente test y modelos de regresión se descartan siempre estos dos registros de pacientes, quedando la muestra final de análisis con 28 pacientes.

En la tabla 5 se muestra el conjunto de datos final para el análisis de supervivencia. La variable *ObjSLP* contiene los valores *time* par el análisis, donde el signo + indica que ese paciente está *censurado*.

Table 5: Datos de la muestra para el Análisis de Supervivencia

Nº.KRAS	Gender	Smoking.historyFac	PA.YEAR	Mutacion	ddCt.HES1	ddCt.CDCP1	ddCt.AXL	ddCt.YAP1	ddCt.CREB	ddCt.LKB1	ddCt.Src	ddCt.STAT3	ObjSLP	ObjSGL
kras 3	Hombre	Exfumador	60	G12V	0.35	16.81	1.06	1.37	1.34	0.36	3.68	0.46	36+	36+
kras 7	Hombre	Fumador	15	G12V	0.29	43.75	0.66	3.47	0.84	0.07	10.71	0.55	1	1
kras 10	Hombre	Exfumador	60	G12C	0.51	6.37	1.36	1.95	1.94	0.31	6.87	0.43	4	4
kras 11	Hombre	Exfumador	150	G12C	NA	28.67	3.49	1.69	1.31	0.26	2.62	0.19	1	2+
kras 13	Hombre	Exfumador	35	G12C	0.63	15.79	3.69	3.15	2.50	0.35	8.58	1.98	14	18
kras 14	Hombre	Exfumador	60	G12D	1.01	32.03	1.17	4.68	2.59	0.47	9.46	0.84	13+	13+
kras 15	Mujer	No fumador	0	G12D	1.22	36.03	2.87	3.04	3.64	0.61	11.72	0.97	3	23+
kras 16	Mujer	Exfumador	24	G12V	2.09	8.12	2.07	2.14	4.18	0.98	20.41	0.67	7	7
kras 17	Hombre	Exfumador	30	G12C	NA	32.47	1.10	18.32	7.64	NA	19.72	1.23	7	26+
kras 19	Hombre	Exfumador	36	G12V	1.28	7.22	1.15	4.39	3.00	0.32	7.22	1.14	4	30
kras 20	Hombre	Fumador	100	Otras	0.34	12.83	1.25	3.04	3.47	0.43	12.92	0.26	5	4+
kras 22	Hombre	Fumador	35	Otras	NA	21.13	5.59	NA	NA	NA	42.26	NA	14	16
kras 23	Hombre	Exfumador	20	G12C	0.45	15.90	1.70	4.81	4.45	0.64	17.40	0.68	2	37+
kras 24	Mujer	Fumador	20	G12C	0.77	2.06	2.15	2.78	2.05	0.38	7.68	0.63	6	19
kras 28	Mujer	Fumador	10	G12D	0.20	8.23	2.55	8.97	2.32	0.30	5.07	1.30	4	8
kras 31	Hombre	Exfumador	30	G12C	1.73	28.08	NA	4.48	3.04	NA	7.90	0.73	8	26
kras 32	Hombre	Exfumador	50	G12V	0.15	16.01	3.79	1.77	3.79	0.36	2.45	0.37	0	8
kras 33	Hombre	Exfumador	80	G12C	NA	64.06	NA	15.73	NA	NA	107.73	0.91	6	9
kras 34	Hombre	Fumador	45	G12D	NA	44.06	12.84	NA	NA	NA	NA	NA	0	0+
kras 36	Mujer	Fumador	25	Otras	NA	17.16	2.18	1.33	4.54	0.96	3.76	0.29	44+	44+
kras37	Mujer	Exfumador	32	Otras	NA	19.44	NA	NA	NA	NA	10.06	1.00	1	2
kras39	Mujer	Fumador	20	Otras	NA	42.55	3.71	2.39	1.60	NA	6.03	0.53	2	3
kras 42	Hombre	Fumador	60	G12V	0.70	36.28	3.90	2.35	1.79	0.36	9.20	0.74	9	15
kras43	Hombre	Exfumador	45	G12V	NA	NA	NA	NA	NA	NA	31.81	NA	2	2
kras 44	Mujer	Exfumador	25	Otras	NA	NA	NA	NA	NA	NA	NA	NA	1	1
kras 47	Hombre	Fumador	60	G12V	0.63	8.29	2.43	2.08	1.47	0.37	8.35	0.69	5	5
kras 48	Hombre	Fumador	15	G12C	1.07	14.74	2.79	2.35	9.81	0.85	6.59	0.92	7	18+
kras 49	Hombre	Fumador	25	Otras	0.21	2.10	1.45	4.68	NA	0.13	13.84	0.58	0	0
kras 50	Hombre	Fumador	50	Otras	NA	24.27	0.50	2.08	1.34	NA	7.42	0.36	18	38
kras 52	Hombre	Exfumador	70	Otras	1.48	14.84	0.18	2.02	4.36	0.29	12.05	1.07	0	0

Supervivencia Libre de Progresión (SLP)

Para calcular la variable *time* para el análisis de Supervivencia Libre de Progresión (SLP), se ha tenido en cuenta como fecha de inicio del estudio, la fecha de inicio del tratamiento que se administró a los pacientes en primera línea. El tiempo se ha calculado en meses. De todos los pacientes con datos clínicos disponibles, se identificaron 3 pacientes de los cuales no se disponía de la fecha de inicio del tratamiento, ni la fecha de diagnóstico de la primera metástasis, por lo que se tomó la fecha del diagnóstico de cáncer como fecha de inicio (según se adoptó en el estudio original). Como fecha de producción del evento, se tomó la fecha de éxitus del paciente o la fecha en que se determina que después de finalizar la primera línea del tratamiento, se constata que la progresión del cáncer continúa. Si se revisan los datos iniciales de la muestra, al inicio de cada línea de tratamiento (*X2nd.line .. X5th.line*) se codifica la progresión del cáncer. Todos los códigos diferentes de *A22* indican una progresión en la enfermedad. En caso de progresión se toma la fecha de inicio del tratamiento de la siguiente línea, como fecha de ocurrencia del evento. Hay tres casos en que se desconoce la fecha de progresión o éxitus, por lo que se toma la última fecha de seguimiento disponible, como fin de periodo y marcando estos pacientes como censurados (en la tabla 5, la variable *ObjSLP* de los pacientes censurados tiene un símbolo *+*).

Estimación de SLP sin agrupaciones

A continuación se muestra el resumen global del análisis SLP sin estratificación y en la tabla 6 se muestra el resultado detallado de los valores obtenidos del análisis SLP, con los valores estimados para cada *time* de pacientes en riesgo, número de eventos producidos, número de censuras, probabilidad de supervivencia, el error estándar y el intervalo de confianza de la probabilidad. Adicionalmente se ha estimado por máxima verosimilitud el riesgo acumulado, que se ha añadido como la columna *CumHaz* en la tabla.

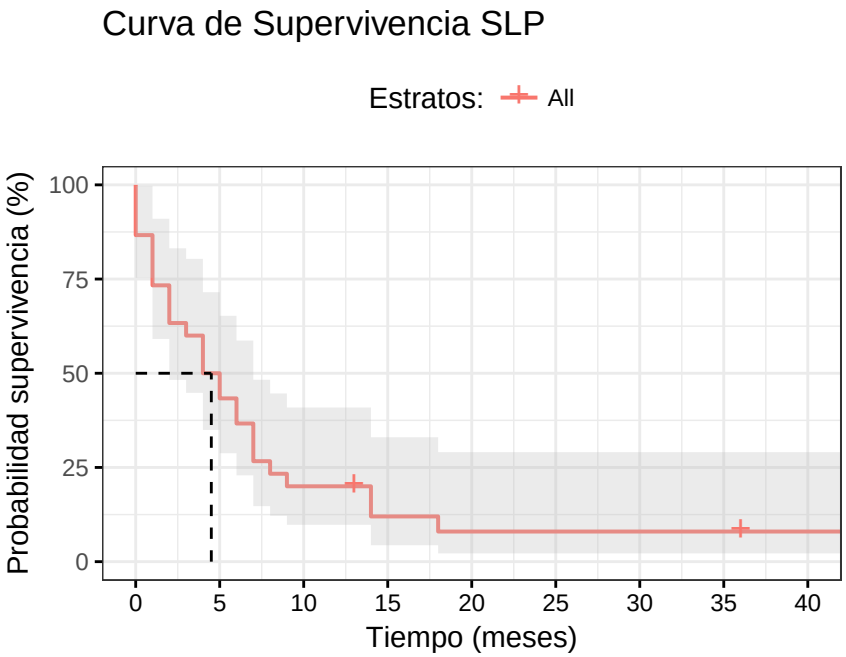
Como puede observar, la mediana de Supervivencia Libre de Progresión (50% probabilidad de supervivencia) es de *4.5* meses y la media de SLP de *8.2* meses.

```
Call: survfit(formula = ObjSLP ~ 1, data = Kras.mostra, type = "kaplan-meier")

        n     events     *rmean *se(rmean)     median    0.95LCL
    30.00      27.00       8.19       2.16       4.50       2.00
  0.95UCL
     7.00
   * restricted mean with upper limit =  44
```

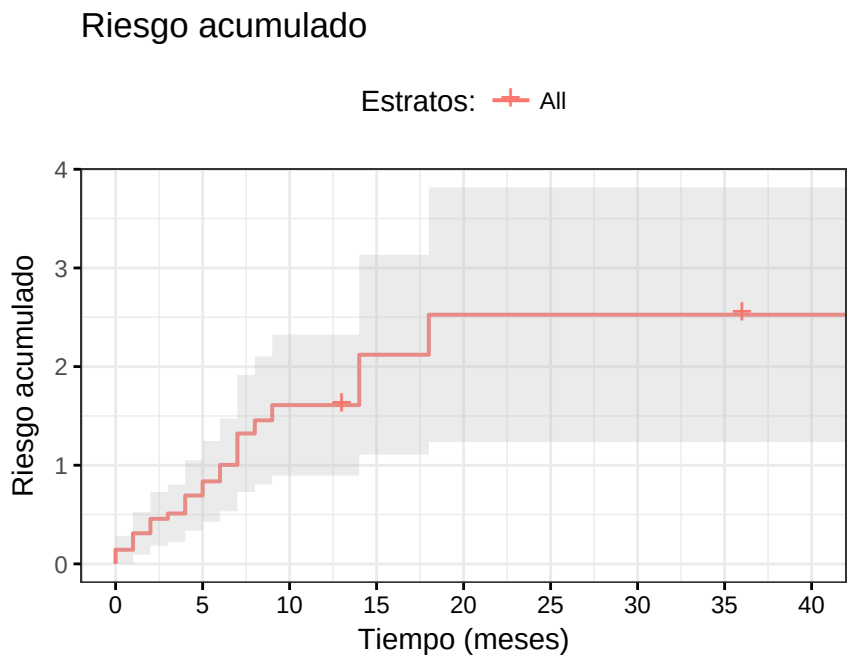
En las gráficas Gr.11a y 11b se muestran los datos de la tabla 6 mediante la curva de SLP y la curva de riesgos acumulados respectivamente, sin ninguna estratificación.

Table 6: Modelo de estimación SLP Kaplan-Meier

time	n.risk	n.event	n.censor	surv	std.err	upper	lower	CumHaz
0	30	4	0	0.8667	0.0716	0.9973	0.7532	0.1333
1	26	4	0	0.7333	0.1101	0.9099	0.5910	0.2872
2	22	3	0	0.6333	0.1389	0.8315	0.4824	0.4235
3	19	1	0	0.6000	0.1491	0.8036	0.4480	0.4762
4	18	3	0	0.5000	0.1826	0.7151	0.3496	0.6428
5	15	2	0	0.4333	0.2088	0.6524	0.2878	0.7762
6	13	2	0	0.3667	0.2399	0.5868	0.2291	0.9300
7	11	3	0	0.2667	0.3028	0.4827	0.1473	1.2027
8	8	1	0	0.2333	0.3309	0.4464	0.1220	1.3277
9	7	1	0	0.2000	0.3651	0.4091	0.0978	1.4706
13	6	0	1	0.2000	0.3651	0.4091	0.0978	1.4706
14	5	2	0	0.1200	0.5164	0.3302	0.0436	1.8706
18	3	1	0	0.0800	0.6583	0.2907	0.0220	2.2039
36	2	0	1	0.0800	0.6583	0.2907	0.0220	2.2039
44	1	0	1	0.0800	0.6583	0.2907	0.0220	2.2039



Gráfica Gr.11a



Gráfica Gr.11b

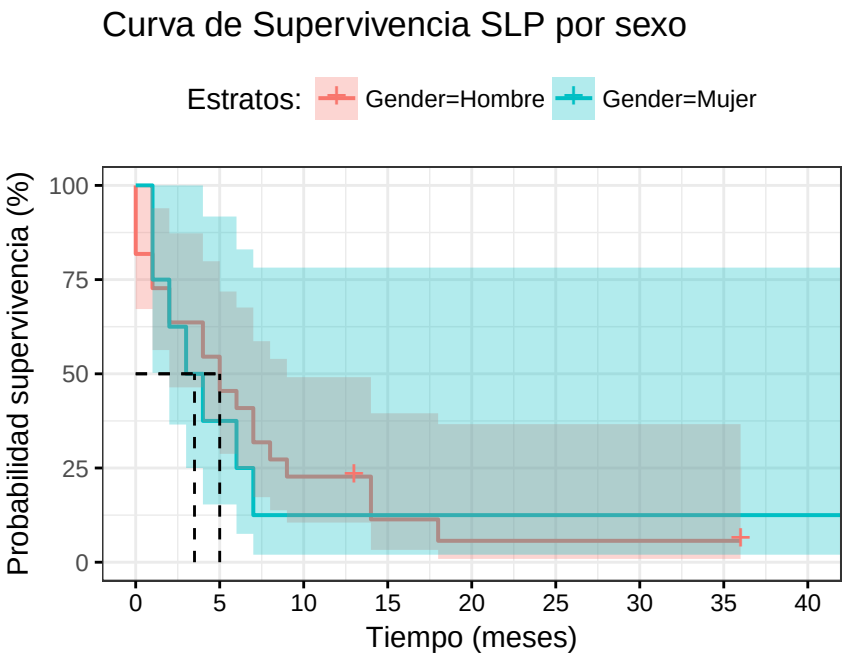
Estimación de SLP con agrupación por sexo y por tipo de mutación KRAS

Para verificar si el sexo de los pacientes o el tipo de mutación del gen KRAS tienen alguna influencia sobre el SLP, adicionalmente se ha efectuado un análisis SLP clasificado por sexo y por tipo de mutación en KRAS, esta última según las mutaciones más frecuentes y resto (*G12C, G12D, G12V y otras*). A continuación se muestran los resultados numéricos (tablas 7 y 8) y gráficos (gráficos Gr.12 y Gr. 13).

```
Call: survfit(formula = ObjSLP ~ Gender, data = Kras.mostra, type = "kaplan-meier")

               n events *rmean *se(rmean) median 0.95LCL 0.95UCL
Gender=Hombre 22     20   7.66       2.09    5.0       2       9
Gender=Mujer   8      7   8.00       4.34    3.5       2      NA
    * restricted mean with upper limit =  40
```

Table 7: Modelo de estimación SLP Kaplan-Meier por género

strata	time	n.risk	n.event	n.censor	surv	std.err	upper	lower
Hombre	0	22	4	0	0.8182	0.1005	0.9963	0.6719
	1	18	2	0	0.7273	0.1306	0.9394	0.5631
	2	16	2	0	0.6364	0.1612	0.8727	0.4640
	4	14	2	0	0.5455	0.1946	0.7988	0.3725
	5	12	2	0	0.4545	0.2335	0.7184	0.2876
	6	10	1	0	0.4091	0.2562	0.6760	0.2476
	7	9	2	0	0.3182	0.3121	0.5866	0.1726
	8	7	1	0	0.2727	0.3482	0.5396	0.1378
	9	6	1	0	0.2273	0.3931	0.4911	0.1052
	13	5	0	1	0.2273	0.3931	0.4911	0.1052
	14	4	2	0	0.1136	0.6360	0.3953	0.0327
	18	2	1	0	0.0568	0.9511	0.3665	0.0088
	36	1	0	1	0.0568	0.9511	0.3665	0.0088
Mujer	1	8	2	0	0.7500	0.2041	1.0000	0.5027
	2	6	1	0	0.6250	0.2739	1.0000	0.3654
	3	5	1	0	0.5000	0.3536	0.9998	0.2500
	4	4	1	0	0.3750	0.4564	0.9174	0.1533
	6	3	1	0	0.2500	0.6124	0.8302	0.0753
	7	2	1	0	0.1250	0.9354	0.7819	0.0200
	44	1	0	1	0.1250	0.9354	0.7819	0.0200



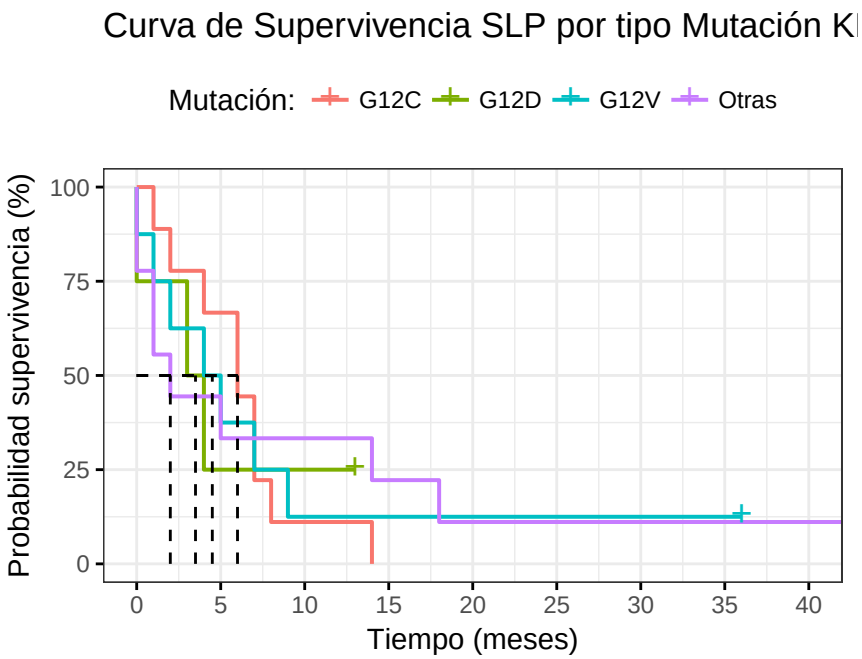
Gráfica Gr.12

```
Call: survfit(formula = ObjSLP ~ Mutacion, data = Kras.mostra, type = "kaplan-meier")
```

```
                n events *rmean *se(rmean) median 0.95LCL 0.95UCL
Mutacion=G12C   9      9   6.11       1.19    6.0       4      NA
Mutacion=G12D   4      3   8.00       4.96    3.5       0      NA
Mutacion=G12V   8      7   6.62       2.65    4.5       2      NA
Mutacion=Otras  9      8   7.33       2.92    2.0       1      NA
    * restricted mean with upper limit =  25
```

Table 8: Modelo de estimación SLP Kaplan-Meier por mutación KRAS

strata	time	n.risk	n.event	n.censor	surv	std.err	upper	lower
G12C	1	9	1	0	0.8889	0.1179	1.0000	0.7056
	2	8	1	0	0.7778	0.1782	1.0000	0.5485
	4	7	1	0	0.6667	0.2357	1.0000	0.4200
	6	6	2	0	0.4444	0.3727	0.9227	0.2141
	7	4	2	0	0.2222	0.6236	0.7544	0.0655
	8	2	1	0	0.1111	0.9428	0.7051	0.0175
	14	1	1	0	0.0000	Inf	NA	NA
G12D	0	4	1	0	0.7500	0.2887	1.0000	0.4259
	3	3	1	0	0.5000	0.5000	1.0000	0.1877
	4	2	1	0	0.2500	0.8660	1.0000	0.0458
	13	1	0	1	0.2500	0.8660	1.0000	0.0458
G12V	0	8	1	0	0.8750	0.1336	1.0000	0.6734
	1	7	1	0	0.7500	0.2041	1.0000	0.5027
	2	6	1	0	0.6250	0.2739	1.0000	0.3654
	4	5	1	0	0.5000	0.3536	0.9998	0.2500
	5	4	1	0	0.3750	0.4564	0.9174	0.1533
	7	3	1	0	0.2500	0.6124	0.8302	0.0753
	9	2	1	0	0.1250	0.9354	0.7819	0.0200
	36	1	0	1	0.1250	0.9354	0.7819	0.0200
Otras	0	9	2	0	0.7778	0.1782	1.0000	0.5485
	1	7	2	0	0.5556	0.2981	0.9966	0.3097
	2	5	1	0	0.4444	0.3727	0.9227	0.2141
	5	4	1	0	0.3333	0.4714	0.8397	0.1323
	14	3	1	0	0.2222	0.6236	0.7544	0.0655
	18	2	1	0	0.1111	0.9428	0.7051	0.0175
	44	1	0	1	0.1111	0.9428	0.7051	0.0175



Gráfica Gr.13

En ambas gráficas, por grupo de sexo y mutación, no parecen mostrar diferencias significativas de la SLP para los grupos comparados. Aun así, se efectúa un test de diferencias entre los resultados de las estimaciones de supervivencia de ambos grupos, por género y por mutación KRAS, para determinar si las diferencias son significativas estadísticamente. Vemos el resultado a continuación.

```
Call:
survdiff(formula = ObjSLP ~ Gender, data = Kras.mostra)

               N Observed Expected (O-E)^2/E (O-E)^2/V
Gender=Hombre 22       20    20.55    0.0145    0.0692
Gender=Mujer   8        7     6.45    0.0463    0.0692

 Chisq= 0.1  on 1 degrees of freedom, p= 0.8
```

```
Call:
survdiff(formula = ObjSLP ~ Mutacion, data = Kras.mostra)

                N Observed Expected (O-E)^2/E (O-E)^2/V
Mutacion=G12C   9        9     8.82   0.00378   0.00660
Mutacion=G12D   4        3     2.72   0.02819   0.03547
Mutacion=G12V   8        7     7.14   0.00276   0.00421
Mutacion=Otras  9        8     8.32   0.01225   0.02135

 Chisq= 0.1  on 3 degrees of freedom, p= 1
```

En ambos casos, se puede ver que el p-valor obtenido es > 0.05, por lo que se descarta la hipótesis de diferencias entre estimaciones por los diferentes grupos comparados. Se podría efectuar un modelo de regresión de Cox, para verificar si hay significación, pero viendo los valores de p-value en ambos casos, no se estima necesario y se determina que no tiene interés incluir estas variables en los modelos de análisis posteriores.

Ajuste de un modelo de regresión de Cox para SLP

Uno de los objetivos finales del trabajo era determinar si la expresión diferencial de los genes analizados molecularmente en el proyecto, tenían influencia sobre la supervivencia de los pacientes. Esto significa que es necesario efectuar un modelo de regresión para ver si la expresión de los diferentes genes puede ser un riesgo o un factor influyente para la supervivencia de los pacientes. Para ello, se va a ajustar un modelo de regresión de Cox, que nos proporcionará la significancia del riesgo proporcional para cada una de las variables de expresión de los genes en la supervivencia de los pacientes. En primer lugar, se efectúa un modelo de regresión univariable, es decir, un modelo de regresión con cada una de las covariables (variables de expresión génica) de forma independiente y sin estratificación o agrupación. Posteriormente, en base al resultado obtenido se realizará un modelo de regresión multivariable, buscando el modelo más ajustado utilizando una estrategia de descarte progresivo de variables, hasta encontrar el mejor modelo en base a la significación global del mismo, el índice de concordancia C y la significación de cada covariable. Por último y también en función de los resultados obtenidos, se efectuará un modelo de regresión estratificado en base a diferentes criterios, con la finalidad de buscar un modelo ajustado, que apoye o descarte la hipótesis de influencia de la expresión de algún gen sobre la supervivencia de los pacientes.

El modelo de riesgos proporcionales de Cox

El modelo de Cox es un modelo *semiparamétrico*, que permite examinar cómo las covariables o variables explicativas influyen en la tasa de un evento (en el caso de este trabajo, sobre la muerte) que ocurre en un momento particular. Esta tasa se conoce como la tasa de riesgo (HR hazard rate, en inglés). El modelo de Cox viene definida por la *función de riesgo* denotada por $h(t)$, que indica el riesgo de muerte h en un momento t con la siguiente formula:

$$h(t) = h_0(t) * exp(\beta_{1x1} + \beta_{2x2} + ... + \beta_{pxp})$$

Donde t es el tiempo de supervivencia, $h(t)$ es la función de riesgo determinado por el conjunto p de covariables $(x_1, x_2, ., x_p)$, las $(\beta_1, \beta_2, ., \beta_p)$ miden el impacto de las covariables (parte paramétrica del modelo) y h_0 es riesgo de referencia que puede variar en el tiempo t (parte no paramétrica del modelo). El valor de $HR = exp(\beta_i)$ es el ratio de riesgo que indica:

- HR = 1: No tiene efecto
- HR < 1: Reducción del riesgo
- HR > 1: Incremento del riesgo

Para poder asegurar que el modelo definido es adecuado, entre otras se debe verificar la suposición de que el peligro del evento en cualquier grupo es un múltiplo constante del peligro en cualquier otro, es decir la *suposición de riesgos proporcionales.*

Evaluación de validez del modelo de regresión de Cox

Para confirmar que el modelo estadístico de Cox es apropiado para extraer conclusiones válidas en referencia a la influencia de la expresión de los genes analizados en la supervivencia o progresión del cáncer en los pacientes del estudio, se debe verificar que el modelo sea una representación adecuada de los datos. Para ello, en este apartado se van analizar las siguientes suposiciones:

- Verificación de suposición de riesgos proporcionales
- Prueba de observaciones influyentes (outliers)

Dado en número bajo de muestras disponibles, será difícil validar la suposición de linealidad. Más adelante, después del ajuste de cada modelo planteado, se incorpora un apartado de *Evaluación de validez del modelo de Cox*, donde se realiza un análisis exhaustivo de verificación de las suposiciones del modelo de Cox planteado.

Modelo de regresión de Cox Univariable (SLP)

Para efectuar el modelo de ajuste de regresión de Cox de cada covariable, se define la función *fit.cox()* que implementa en R la ejecución de la función *coxph()* del paquete *survival* de R, para las variables de conjunto de datos a analizar. La implementación de esta función se puede ver en el apartado de anexo de código R de

este informe dinámico. En la misma función se han calculado los puntos de corte óptimos obtenidos mediante la función *cutp()* del paquete *survMisc* de R.

En la tabla 9, se muestra el resultado de la ejecución de la función para el modelo de ajuste de regresión de Cox univariable para la variable ObjSLP con cada covariable de expresión de cada gen. en la última columna se ha añadido el punto de corte óptimo para cada covariable.

Table 9: Resultados del modelo de regresión de Cox Univariable

	n	n.eventos	beta	HR (95% CI for HR)	wald.test	p.value	CutP
ddCt.HES1	19	17	-0.2439	0.7836 (0.2902-2.1155)	0.23	0.6303	0.35
ddCt.CDCP1	28	25	0.0016	1.0016 (0.9711-1.033)	0.01	0.9192	16.81
ddCt.AXL	25	22	0.1494	1.1611 (0.9444-1.4274)	2.01	0.1563	1.25
ddCt.YAP1	25	22	0.0191	1.0193 (0.9315-1.1153)	0.17	0.6779	1.69
ddCt.CREB	23	20	0.0105	1.0106 (0.8259-1.2365)	0.01	0.9186	4.54
ddCt.LKB1	20	17	-2.2456	0.1059 (0.0076-1.4838)	2.78	0.0955	0.35
ddCt.Src	28	25	0.0024	1.0024 (0.9845-1.0207)	0.07	0.7922	10.06
ddCt.STAT3	26	23	0.0656	1.0678 (0.432-2.6394)	0.02	0.8870	0.53

Como se puede comprobar, de forma individual no parece que la expresión de ninguno los genes influya de forma significativa en la supervivencia libre de progresión.

Viendo estos resultados, seguidamente se van a efectuar dos procesos de análisis adicionales, para ver si se puede identificar alguna evidencia de significación al 5% en la SLP según la expresión de los genes estudiados, en base a diferentes niveles de expresión de cada gen. En primer lugar, se va a proceder a efectuar la estratificación de los datos de expresión según diferentes quantiles, para ajustar con ellos modelos de Cox univariables. En segundo lugar, se va a realizar el modelo de regresión multivariable con todas las covariables, para ver determinamos un modelo lineal semiparamétrico que pueda explicar las variaciones en la SLP.

Modelo de regresión de Cox Univariable estratificado (SLP).

Dado que el número de muestras disponibles en el estudio es limitado y cualquier muestra con valores fuera de la media (outliers) puede afectar significativamente el resultado, a continuación se va a realizar un proceso de estratificación de los valores de expresión de los genes en diferentes cuantiles, de forma que se minimice el impacto de estos valores sobresalientes en los posibles resultados.

Se van a efectuar cuatro tipos de estratificación de los valores de expresión, que son los siguientes:

- Por tres cuantiles de expresión, etiquetados como *Alta, Media* y *Baja*, donde se agruparan en el percentil 33 la expresión Baja, entre el percentil 33 y 66 la expresión media y mayores que el percentil 66, la expresión alta.
- Por dos cuantiles de expresión, etiquetados como *Alta* y *Baja-Media*, donde se agruparan como Media-Baja hasta el percentil 66 y Alta los valores superiores la percentil 66.
- Por dos cuantiles de expresión, cuantil inferior al 25 etiquetados como $< Q25$ y superiores o iguales a 25, etiquetados como $>= Q25$.
- Por estratificación según los mejores puntos de corte basados en la función *cutp()* del paquete *survMisc* de R.

A continuación se detalla un resumen de las variables estratificadas y factorizadas según los tres criterios indicados, para su uso en un modelo de Cox Univariable estratificado. Para el cálculo de los percentiles indicados, se ha utilizado la función *quantile()* de R.

Modelos de regresión de Cox con estratificación según expresión Alta, Media y Baja

En la tabla 10 se muestra el resultado de la estratificación de cada covariable de expresión según los percentiles 33 y 66, como expresión *Alta*, *Media* y *Baja.* A continuación, en la tabla 11, se muestra el resultado del

modelo de regresión de Cox estratificado.

Table 10: Estratificación: Alta, Media y Baja

HES1.st	CDCP1.st	AXL.st	YAP1.st	CREB.st	LKB1.st	Src.st	STAT3.st
Alta : 7	Alta :10	Alta :9	Alta :9	Alta :8	Alta : 7	Alta :10	Alta :9
Baja : 6	Baja : 9	Baja :8	Baja :8	Baja :8	Baja : 7	Baja : 9	Baja :9
Media: 6	Media: 9	Media:8	Media:8	Media:7	Media: 6	Media: 9	Media:8
NA's :11	NA's : 2	NA's :5	NA's :5	NA's :7	NA's :10	NA's : 2	NA's :4

Table 11: Modelo de Cox Univariable Estratificado (Alta, Media y Baja)

Variable	n	n.eventos	wald.test	p.value	strat	beta	HR (95% CI for HR)	Pr(>|z|)
HES1.st	19	17	0.36	0.834	HES1.stBaja	0.3374	1.4013 (0.4118-4.7683)	0.5892
					HES1.stMedia	0.0199	1.0201 (0.3234-3.2172)	0.973
CDCP1.st	28	25	3.66	0.1602	CDCP1.stBaja	0.2756	1.3173 (0.5035-3.4464)	0.5743
					CDCP1.stMedia	-0.9321	0.3937 (0.1196-1.2963)	0.1252
AXL.st	25	22	1.27	0.53	AXL.stBaja	-0.6	0.5488 (0.1912-1.5749)	0.2646
					AXL.stMedia	-0.1735	0.8408 (0.3067-2.3044)	0.736
YAP1.st			0.96	0.6192	YAP1.stBaja	-0.5789	0.5605 (0.1756-1.789)	0.3282
					YAP1.stMedia	-0.1952	0.8227 (0.3068-2.2062)	0.6982
CREB.st	23	20	0.35	0.838	CREB.stBaja	-0.2198	0.8027 (0.2778-2.3197)	0.6848
					CREB.stMedia	-0.3235	0.7236 (0.2401-2.1808)	0.5655
LKB1.st	20	17	8.56	0.0139	LKB1.stBaja	2.0587	7.8361 (1.7509-35.0695)	0.0071
					LKB1.stMedia	-0.032	0.9685 (0.277-3.3864)	0.96
Src.st	28	25	2.12	0.347	Src.stBaja	-0.5302	0.5885 (0.2166-1.5986)	0.2984
					Src.stMedia	-0.6651	0.5142 (0.1988-1.3299)	0.1701
STAT3.st	26	23	1.13	0.5693	STAT3.stBaja	-0.5313	0.5878 (0.2051-1.6844)	0.3226
					STAT3.stMedia	-0.3853	0.6802 (0.2505-1.8474)	0.4497

Una vez realizado el modelo de ajuste de regresión de Cox con la expresión estratificada de los diferentes genes, se puede comprobar (datos en rojo en la tabla 11) que el gen *LKB1* muestra evidencia de significación al 5% en el estrato de expresión *Baja* respecto a la expresión *Alta*, con un $p - valor = 0.0071$ y con un test de Wald también significativo con un $p - valor = 0.0139$.

Modelos de regresión de Cox con estratificación según expresión Alta, Baja-Media

En la tabla 12 se muestra el resultado de la estratificación de cada covariable de expresión según los percentiles inferiores a 66 y superiores a 66, como expresión *Alta* o *Media-Baja*. A continuación, en la tabla 13, se muestra el resultado del modelo de regresión de Cox estratificado.

Table 12: Estratificación: Alta y Baja-Media

HES1.st	CDCP1.st	AXL.st	YAP1.st	CREB.st	LKB1.st	Src.st	STAT3.st
Alta : 7	Alta :10	Alta : 9	Alta : 9	Alta : 8	Alta : 7	Alta :10	Alta : 9
Baja-Media:12	Baja-Media:18	Baja-Media:16	Baja-Media:16	Baja-Media:15	Baja-Media:13	Baja-Media:18	Baja-Media:17
NA's :11	NA's : 2	NA's : 5	NA's : 5	NA's : 7	NA's :10	NA's : 2	NA's : 4

En esta ocasión, tal como reflejan los resultados de los diferentes modelos de Cox univariables, con esta estratificación no existe evidencia de significación de ninguno de los genes analizados.

Table 13: Modelo de Cox Univariable Estratificado (Alta y Baja-Media)

Variable	n	n.eventos	wald.test	p.value	strat	beta	HR (95% CI for HR)
HES1.st	19	17	0.08	0.7785	HES1.stBaja-Media	0.1467	1.158 (0.4167-3.2181)
CDCP1.st	28	25	0.39	0.5335	CDCP1.stBaja-Media	-0.2706	0.7629 (0.3255-1.7882)
AXL.st	25	22	0.78	0.3772	AXL.stBaja-Media	-0.392	0.6757 (0.2831-1.6129)
YAP1.st	25	22	0.58	0.447	YAP1.stBaja-Media	-0.3489	0.7055 (0.287-1.7341)
CREB.st	23	20	0.32	0.5698	CREB.stBaja-Media	-0.2691	0.7641 (0.3021-1.9326)
LKB1.st	20	17	1.37	0.2417	LKB1.stBaja-Media	0.6275	1.8728 (0.6551-5.3542)
Src.st	28	25	2.06	0.1507	Src.stBaja-Media	-0.6059	0.5456 (0.2387-1.2467)
STAT3.st	26	23	1.07	0.3005	STAT3.stBaja-Media	-0.4543	0.6349 (0.2687-1.5004)

Modelos de regresión de Cox con estratificación según expresión $< Q25$ ***y*** $>= Q25$

En la tabla 14 se muestra el resultado de la estratificación de cada covariable de expresión según los percentiles inferiores a cuantil 25 y superiores o iguales a cuantil 25, como expresión $< Q25$ o $>= Q25$. A continuación, en la tabla 15, se muestra el resultado del modelo de regresión de Cox estratificado.

Table 14: Estratificación: <Q25 y >=Q25

HES1.st	CDCP1.st	AXL.st	YAP1.st	CREB.st	LKB1.st	Src.st	STAT3.st
<Q25 : 5	<Q25 : 7	<Q25 : 6	<Q25 : 6	<Q25 : 6	<Q25 : 5	<Q25 : 7	<Q25 : 7
>=Q25:14	>=Q25:21	>=Q25:19	>=Q25:19	>=Q25:17	>=Q25:15	>=Q25:21	>=Q25:19
NA's :11	NA's : 2	NA's : 5	NA's : 5	NA's : 7	NA's :10	NA's : 2	NA's : 4

Table 15: Modelo de Cox Univariable Estratificado (<Q25 y >=Q25)

Variable	n	n.eventos	wald.test	p.value	strat	beta	HR (95% CI for HR)
HES1.st	19	17	6.15	0.0131	HES1.st>=Q25	-1.5425	0.2138 (0.0632-0.7234)
CDCP1.st	28	25	1.95	0.1622	CDCP1.st>=Q25	-0.6765	0.5084 (0.1969-1.3126)
AXL.st	25	22	0.4	0.5258	AXL.st>=Q25	0.3318	1.3935 (0.5-3.884)
YAP1.st	25	22	0.33	0.5659	YAP1.st>=Q25	0.3316	1.3932 (0.4492-4.321)
CREB.st	23	20	0.03	0.8622	CREB.st>=Q25	0.0929	1.0974 (0.3844-3.1324)
LKB1.st	20	17	9.78	0.0018	LKB1.st>=Q25	-2.1108	0.1211 (0.0323-0.4549)
Src.st	28	25	0.66	0.4164	Src.st>=Q25	0.4176	1.5182 (0.5546-4.1565)
STAT3.st	26	23	1.92	0.1659	STAT3.st>=Q25	0.7909	2.2053 (0.7204-6.7509)

Como resultado de los modelos de Cox analizados, se puede ver (en rojo) que hay dos covariables que tienen evidencia de significación al 5%, *HES1* y *LKB1*, en su estrato de $>= Q25$ respecto a $< Q25$, con *p-valores* de *0.0131* y *0.0018*, respectivamente.

Modelos de regresión de Cox con estratificación según mejores puntos de corte

Se ha efectuado un último modelo de ajuste de Cox Univariable estratificando la expresión de los genes según los mejores puntos de corte que se han encontrado mediante la función propuesta por Contal C y O'Quigley, implementada en la función *cutp()* del paquete *survMisc* de R. Los puntos de corte aplicados para hacer la estratificación de la expresión de cada gen es la que se ha indicado en la tabla 9.

En la tabla 16 se muestra el resultado de la estratificación, donde se pueden ver el número de muestras que han caído en cada lado del punto de corte. En la tabla 17 se muestran los resultados de los diferentes modelos de Cox ajustados para cada gen, aplicando la estratificación del mejor punto de corte. En rojo se indican los modelos que han resultado significativos al 5% y podemos ver que son los genes *HES1*, *LKB1* y *Src*, con un p-valor de *0.0131, 0.0034, 0.0346*, respectivamente.

Validación significación de genes

Table 16: Estratificación: Mejores puntos de corte

HES1.st	CDCP1.st	AXL.st	YAP1.st	CREB.st	LKB1.st	Src.st	STAT3.st
< 0.35 : 5	< 16.81 :13	< 1.25 : 7	< 1.69 : 2	< 4.54 :20	< 0.35 : 7	< 10.06 :16	< 0.53 : 7
>= 0.35 :14	>= 16.81 :15	>= 1.25 :18	>= 1.69 :23	>= 4.54 : 3	>= 0.35 :13	>= 10.06 :12	>= 0.53 :19
NA's :11	NA's : 2	NA's : 5	NA's : 5	NA's : 7	NA's :10	NA's : 2	NA's : 4

Table 17: Modelo de Cox Univariable Estratificado mejores puntos de corte

Variable	n	n.eventos	wald.test	p.value	strat	beta	HR (95% CI for HR)
HES1.st	19	17	6.15	0.0131	HES1.st>= 0.35	-1.5425	0.2138 (0.0632-0.7234)
CDCP1.st	28	25	3.65	0.056	CDCP1.st>= 16.81	-0.8201	0.4404 (0.1899-1.0211)
AXL.st	25	22	1.45	0.2281	AXL.st>= 1.25	0.6246	1.8675 (0.6762-5.1574)
YAP1.st	25	22	0	0.9978	YAP1.st>= 1.69	19.681	352637377.3775 (0-Inf)
CREB.st	23	20	1.28	0.258	CREB.st>= 4.54	-0.8479	0.4283 (0.0985-1.8616)
LKB1.st	20	17	8.56	0.0034	LKB1.st>= 0.35	-2.0732	0.1258 (0.0314-0.5046)
Src.st	28	25	4.46	0.0346	Src.st>= 10.06	0.8862	2.4258 (1.0663-5.5186)
STAT3.st	26	23	1.92	0.1659	STAT3.st>= 0.53	0.7909	2.2053 (0.7204-6.7509)

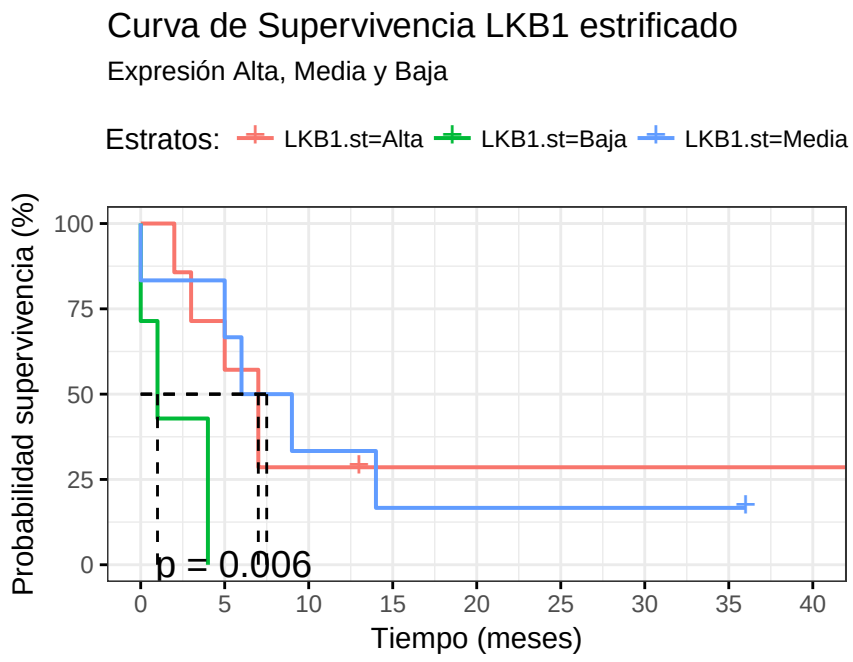
Para validar la significación de los modelos anteriormente descritos, se ha efectuado un test de Log-Rank de comparación por pares, cuyo resultado confirma la significación de todos ellos, según el resultado que se muestra a continuación.

```
    Pairwise comparisons using Log-Rank test

data:  Kras.strat33 and LKB1.st

      Alta  Baja
Baja  0.018 -
Media 0.977 0.018

P value adjustment method: BH


    Pairwise comparisons using Log-Rank test

data:  Kras.stratQ25 and HES1.st

      <Q25
>=Q25 0.0089

P value adjustment method: BH


    Pairwise comparisons using Log-Rank test

data:  Kras.stratQ25 and LKB1.st

      <Q25
>=Q25 0.00069

P value adjustment method: BH
```

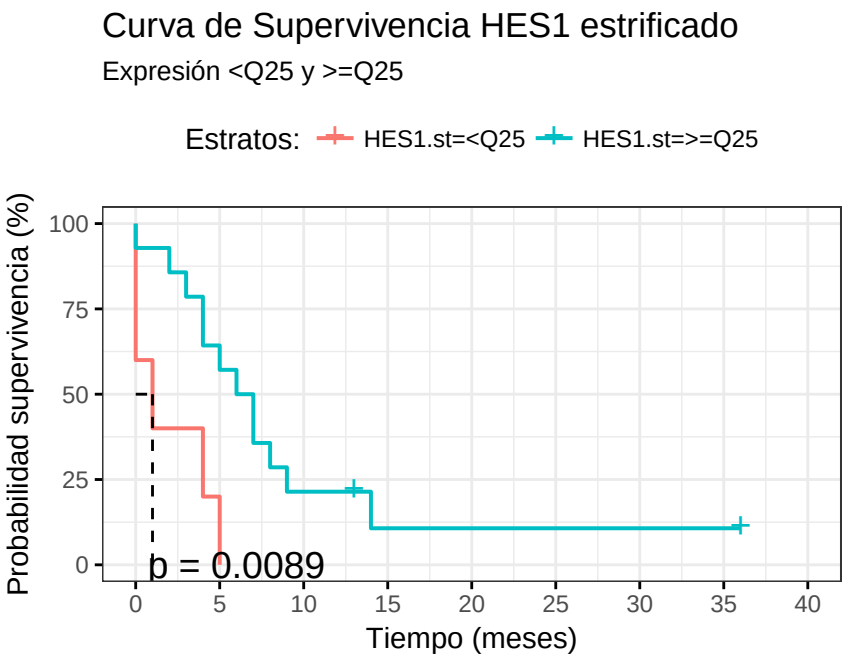
```
    Pairwise comparisons using Log-Rank test

data:  Kras.stratBCP and HES1.st

         < 0.35
>= 0.35  0.0089

P value adjustment method: BH


    Pairwise comparisons using Log-Rank test

data:  Kras.stratBCP and LKB1.st

         < 0.35
>= 0.35  0.0014

P value adjustment method: BH


    Pairwise comparisons using Log-Rank test

data:  Kras.stratBCP and Src.st

          < 10.06
>= 10.06  0.033

P value adjustment method: BH
```

Curvas de supervivencia de los genes con significación según estratificación

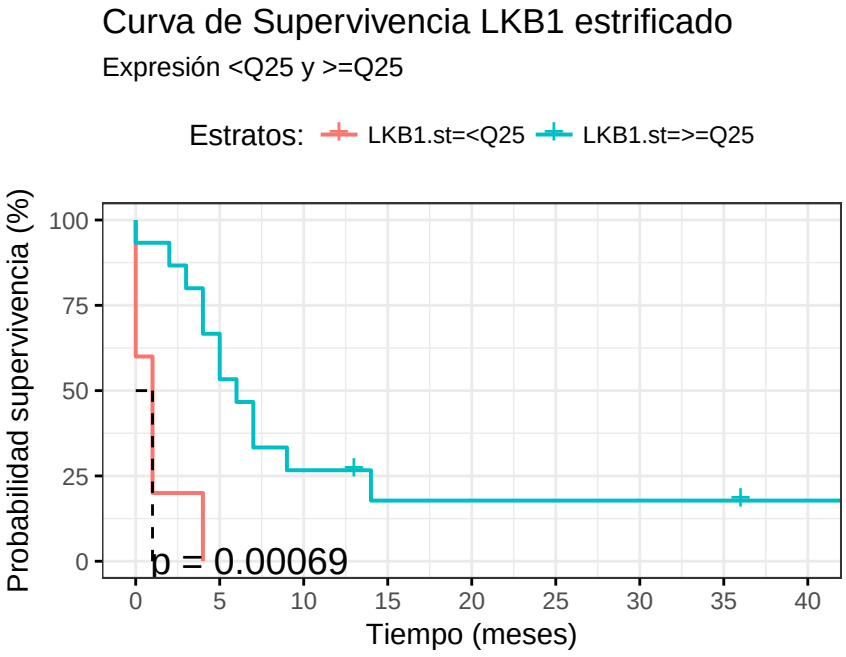
A continuación se muestra la gráfica de Supervivencia libre de progresión (SLP) de los genes *HES1*, *LKB1* y *Src*, que han evidenciado significancia en alguno de los modelos de Cox para las diferentes estratificaciones (gráficas Gr.14, Gr.15a-b, Gr.16a-c)



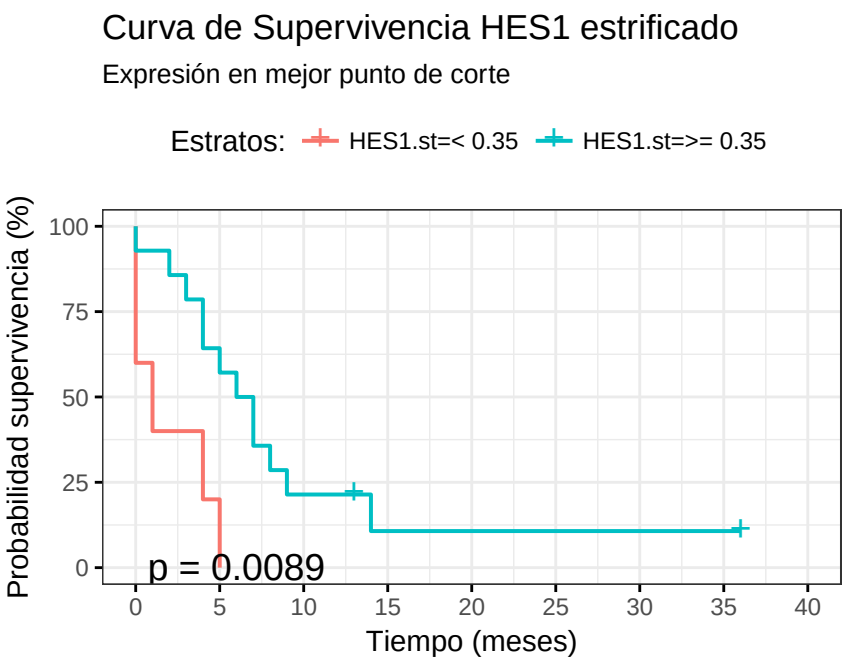
Gráfica Gr.14



Gráfica Gr.15a



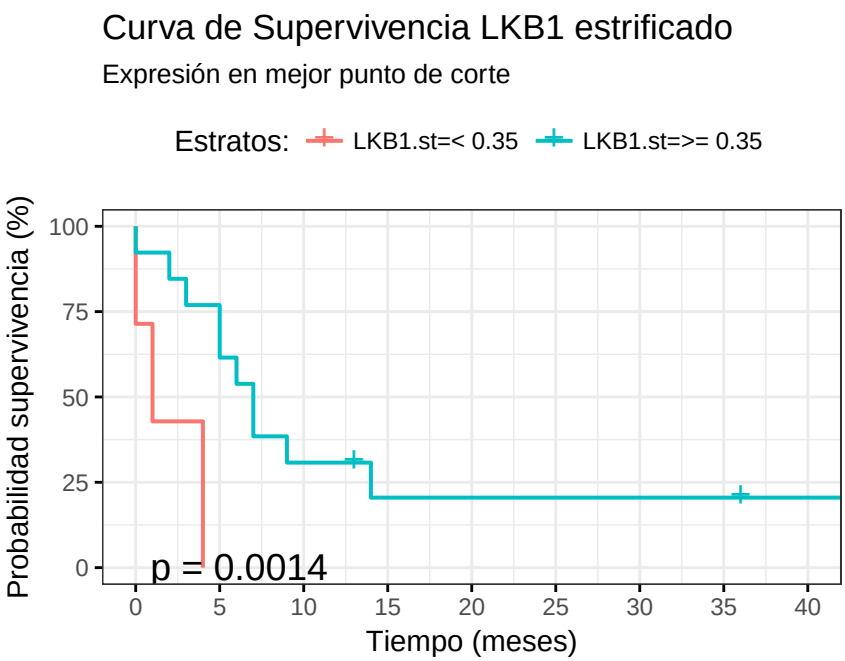
Gráfica Gr.15b



Gráfica Gr.16a

Estos son los valores de la media y mediana de la SLP estratificada con el mejor punto de corte para la expresión del gen *LKB1*:

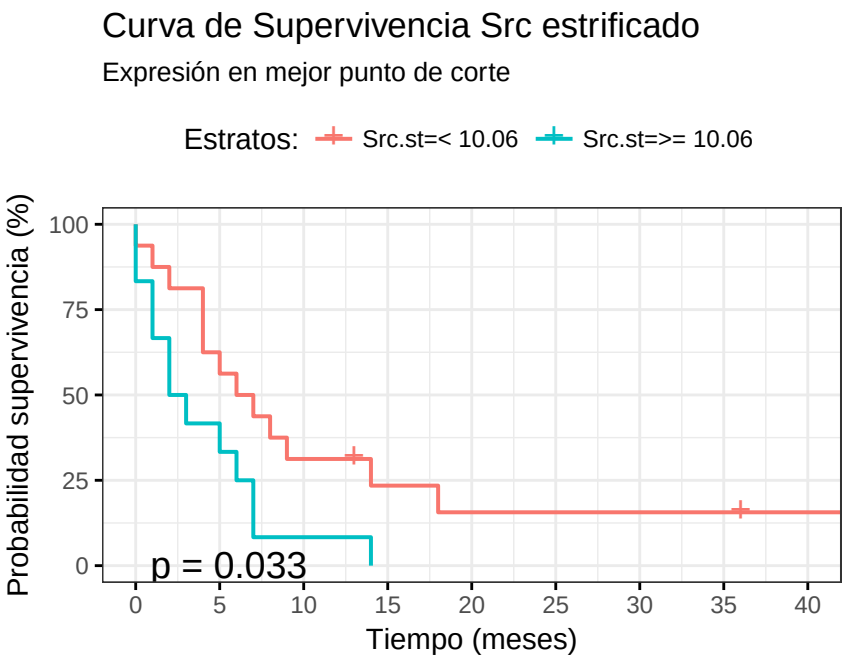
```
Call: survfit(formula = ObjSLP ~ HES1.st, data = Kras.stratBCP, type = "kaplan-meier")

   11 observations deleted due to missingness
                  n events *rmean *se(rmean) median 0.95LCL 0.95UCL
HES1.st=< 0.35    5      5   2.00      0.938    1.0       0      NA
HES1.st=>= 0.35  14     12   7.62      1.557    6.5       4      NA
    * restricted mean with upper limit =  20.5
```



Gráfica Gr.16b

Estos son los valores de la media y mediana de la SLP estratificada con el mejor punto de corte para la expresión del gen *LKB1*:

```
Call: survfit(formula = ObjSLP ~ LKB1.st, data = Kras.stratBCP, type = "kaplan-meier")

   10 observations deleted due to missingness
                  n events *rmean *se(rmean) median 0.95LCL 0.95UCL
LKB1.st=< 0.35    7      7   2.00       0.67      1       0      NA
LKB1.st=>= 0.35  13     10   9.74       2.27      7       5      NA
    * restricted mean with upper limit =  24
```



Gráfica Gr.16c

Estos son los valores de la media y mediana de la SLP estratificada con el mejor punto de corte para la expresión del gen *Src*:

```
Call: survfit(formula = ObjSLP ~ Src.st, data = Kras.stratBCP, type = "kaplan-meier")

   2 observations deleted due to missingness
                 n events *rmean *se(rmean) median 0.95LCL 0.95UCL
Src.st=< 10.06   16     13   10.2       2.39    6.5       4      NA
Src.st=>= 10.06  12     12    4.0       1.12    2.5       1      NA
    * restricted mean with upper limit =  29
```

Modelo de regresión de Cox Multivariable (SLP)

En este modelo se deberá considerar que se excluirán las muestras en las cuales la expresión de alguno de los genes no esté informada (valor NA), por lo que el modelo se ajustará a la n mínima de las muestras que tengan valor de expresión para todos los genes, por lo que posiblemente su resultado no sea demasiado ajustado. De la misma manera, el principio de linealidad de las covariables, será difícil de validar, dado que las muestras son escasas y un solo valor outlier, puede influir negativamente en este aspecto. Aun así se considera interesante para el trabajo hacer este análisis y se validará la suposición de riesgos proporcionales, básica para la aplicación de modelo de Cox.

```
Call:
coxph(formula = fmla, data = Kras.mostra)

  n= 17, number of events= 15
   (13 observations deleted due to missingness)

                coef  exp(coef)   se(coef)      z Pr(>|z|)
ddCt.HES1   1.879e+00  6.544e+00  1.145e+00  1.641  0.10083
ddCt.CDCP1 -4.661e-02  9.545e-01  3.258e-02 -1.431  0.15244
ddCt.AXL    6.852e-01  1.984e+00  4.745e-01  1.444  0.14871
ddCt.YAP1   3.930e-01  1.481e+00  2.664e-01  1.475  0.14015
```

```
ddCt.CREB   8.482e-01  2.335e+00  3.015e-01  2.813  0.00491 **
ddCt.LKB1  -1.368e+01  1.147e-06  4.553e+00 -3.004  0.00266 **
ddCt.Src    2.725e-01  1.313e+00  1.170e-01  2.329  0.01986 *
ddCt.STAT3 -2.515e+00  8.083e-02  1.358e+00 -1.853  0.06394 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

           exp(coef) exp(-coef) lower .95 upper .95
ddCt.HES1  6.544e+00  1.528e-01 6.939e-01  61.71631
ddCt.CDCP1 9.545e-01  1.048e+00 8.954e-01   1.01738
ddCt.AXL   1.984e+00  5.040e-01 7.829e-01   5.02831
ddCt.YAP1  1.481e+00  6.751e-01 8.789e-01   2.49685
ddCt.CREB  2.335e+00  4.282e-01 1.293e+00   4.21726
ddCt.LKB1  1.147e-06  8.715e+05 1.529e-10   0.00861
ddCt.Src   1.313e+00  7.615e-01 1.044e+00   1.65176
ddCt.STAT3 8.083e-02  1.237e+01 5.647e-03   1.15690

Concordance= 0.766  (se = 0.098 )
Rsquare= 0.555   (max possible= 0.978 )
Likelihood ratio test= 13.77  on 8 df,   p=0.09
Wald test            = 10.66  on 8 df,   p=0.2
Score (logrank) test = 13.08  on 8 df,   p=0.1
```

Los resultados del modelo multivariable confirman que se han descartado 13 muestras del total de 30, por lo que $n = 17$. en el modelo analizado. Hay tres genes significativos con un *p-valor* < 0.05, *CREB, LKB1 y Src*, pero la significancia global del modelo es > 0.05 por lo que el ajuste del modelo no es óptimo para confirmar que alguna de las expresiones de los genes sea un factor influyente en el tiempo libre de progresión, al menos de forma global.

Mejora del modelo de regresión de Cox Multivariable (SLP)

Para intentar buscar un mejor modelo de ajuste se va a proceder a ir eliminando las covariables no significativas del modelo una a una hasta dejar las tres significativas del modelo completo (*CREB, LKB1 y Src*), para ver qué modelo obtiene un ajuste mejor. Adicionalmente, se efectuará un modelo de ajuste con la combinación de las otras covariables *CREB, LKB1, Src*. Para ello se implementa la función *fit.cox2()* en R que admite como parámetro una lista de las covariables a introducir al modelo de Cox. La implementación de esta función se puede ver en el apartado de anexo de código R de este informe dinámico.

En las tablas 18 y 19 que se muestran a continuación, se puede comprobar que el modelo regresión que utiliza las covariables de la expresión de los genes *CREB, LKB1, Src y STAT3* es significativo en dos de los tres test de ajuste, en el de *razón de verosimilitud (Likekihook)* y en el *test de Wald*. De estas cuatro variables explicativas, el modelo indica que las variables **CREB, LKB1 y Src son significativas**, con un **p-valor** $<$ **0.05**, y la variable *STAT3* tiene un *p-valor* > 0.05, por lo que *no es significativa* al 5%. El indice de Concordancia del modelo $C = 0.7562$, que no está mal y como se puede ver, solo se han podido utilizar *19* muestras de las *30* de conjunto de datos inicial.

Por otra parte, también se puede ver que hay otros modelos que dan valores de significación de las variables explicativas *CREB, LKB1 y Src*. Los modelos en azul en la tabla también son significativos aunque en esos casos varían los resultados en los test de ajuste del modelo.

Viendo estos resultados se puede concluir que con más certeza, por obtener significancia en dos de los tres test del modelo, el modelo más ajustado es el marcado en color rojo en la tablas (*CREB, LKB1 y Src significativas*) con un índice de concordancia de $C = 0.7562$. Los dos modelos en azul en las tablas también serían un modelos ajustados con las mismas tres covariables significativas (*CREB, LKB1 y Src*), con un índice de concordancia de $C = 0.6688$ y $C = 0.7688$ valores bastante próximos en los tres modelos, pero donde solo el test del ratio de verosimilitud es significativo. Por tanto, parece ser que las variables *CREB, LKB1 y*

Table 18: Variables significativas de cada modelo de regresión de Cox

Modelo	Var.Significativas	p.valores
HES1+CDCP1+AXL+YAP1+CREB+LKB1+Src+STAT3	CREB, LKB1, Src	0.0049, 0.0027, 0.0199
CDCP1+AXL+YAP1+CREB+LKB1+Src+STAT3	CREB, LKB1, Src	0.0152, 0.0079, 0.0295
AXL+YAP1+CREB+LKB1+Src+STAT3	CREB, LKB1, Src	0.0155, 0.0081, 0.0474
YAP1+CREB+LKB1+Src+STAT3	CREB, LKB1	0.0161, 0.0074
CREB+LKB1+Src+STAT3	CREB, LKB1, Src	0.0146, 0.0059, 0.0464
CREB+LKB1+Src	CREB, LKB1	0.0294, 0.0114
CDCP1+YAP1+CREB+LKB1+Src+STAT3	CREB, LKB1, Src	0.016, 0.0067, 0.0422
CDCP1+CREB+LKB1+Src+STAT3	CREB, LKB1, Src	0.0144, 0.0054, 0.0354
CDCP1+CREB+LKB1+Src	CREB, LKB1, Src	0.03, 0.0103, 0.0466
AXL+CREB+LKB1+Src+STAT3	CREB, LKB1, Src	0.0138, 0.0063, 0.0406
AXL+YAP1+CREB+LKB1+Src	CREB, LKB1	0.0285, 0.0118

Table 19: Significación global de cada modelo Multivariable

Modelo	n	n.eventos	Conc.	likelih.pval	wald.pval	score.paval
HES1+CDCP1+AXL+YAP1+CREB+LKB1+Src+STAT3	17	15	0.7656	0.0879	0.2217	0.1091
CDCP1+AXL+YAP1+CREB+LKB1+Src+STAT3	19	16	0.7938	0.0569	0.192	0.1591
AXL+YAP1+CREB+LKB1+Src+STAT3	19	16	0.775	0.0522	0.1417	0.1263
YAP1+CREB+LKB1+Src+STAT3	19	16	0.7688	0.04	0.0893	0.0943
CREB+LKB1+Src+STAT3	19	16	0.7562	0.0216	0.0486	0.0546
CREB+LKB1+Src	19	16	0.6688	0.0214	0.0705	0.0862
CDCP1+YAP1+CREB+LKB1+Src+STAT3	19	16	0.775	0.0572	0.1388	0.1403
CDCP1+CREB+LKB1+Src+STAT3	19	16	0.7625	0.0342	0.0835	0.0903
CDCP1+CREB+LKB1+Src	19	16	0.7062	0.0403	0.1248	0.1514
AXL+CREB+LKB1+Src+STAT3	19	16	0.7625	0.0341	0.0874	0.0856
AXL+YAP1+CREB+LKB1+Src	19	16	0.6688	0.0838	0.2047	0.2392

Src pueden influir en el tiempo libre de progresión. Si revisamos el gráfico Gr.10 de los resultados del test de correlación de Spearman, vemos que algunos de los genes en los que se indicaban existía correlación, son los que se visualizan como significativos en el modelo de Cox.

A continuación se muestra el resumen completo del modelo (en rojo en la tabla), donde se pueden ver todos los estadísticos, entre otros los coeficientes β, los ratios $HR = exp(coef)$ (estimación de riesgos relativos), la desviación estándar y los intervalos de confianza de cada variable.

Si se analiza la estimación de riesgo $HR = exp(coef)$ en el resumen mostrado de cada covariable, se puede concluir con significancia de 0.05 que el modelo indica que a igualdad de valores en el resto de expresiones de los otros genes, el aumento de un punto de expresión del gen *CREB* incrementa un 68% *(HR=1.68)* el riesgo de muerte. En el caso del gen *Scr*, a igualdad de valores de expresión de los otros genes, el aumento de un punto en su expresión, significa un aumento del 18% *(HR=1.18)* en el riesgo de muerte, mientras que en el caso de *LKB1* el aumento de su nivel de expresión, manteniendo la expresión de los otros genes, tendría un efecto de disminución del riesgo de muerte ya que vemos que el valor de su coeficiente $\beta = -7.28$ es negativo, pero dado que su valor de $HR = 0.0007$ es muy cercano a cero, este no tendría una afectación significante en el pronóstico. Por otra parte, se puede comprobar que el intervalo de confianza de la covariable *HEST1* contiene el 1, es decir contiene el valor nulo del modelo, no tiene una contribución significativa en el modelo.

```
[1] "Formula =  ObjSLP ~ ddCt.CREB + ddCt.LKB1 + ddCt.Src + ddCt.STAT3"
```

```
Call:
coxph(formula = fmla, data = Kras.mostra)

  n= 19, number of events= 16
   (11 observations deleted due to missingness)
```

```
                coef  exp(coef)   se(coef)      z Pr(>|z|)
ddCt.CREB   0.5208156  1.6834001  0.2133533  2.441  0.01464 *
ddCt.LKB1  -7.2841556  0.0006863  2.6475223 -2.751  0.00594 **
ddCt.Src    0.1629121  1.1769333  0.0817963  1.992  0.04641 *
ddCt.STAT3 -0.7285865  0.4825906  0.5718622 -1.274  0.20264
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

           exp(coef) exp(-coef) lower .95 upper .95
ddCt.CREB  1.6834001     0.5940 1.108e+00    2.5574
ddCt.LKB1  0.0006863  1457.0303 3.828e-06    0.1231
ddCt.Src   1.1769333     0.8497 1.003e+00    1.3816
ddCt.STAT3 0.4825906     2.0721 1.573e-01    1.4803

Concordance= 0.756  (se = 0.092 )
Rsquare= 0.454   (max possible= 0.98 )
Likelihood ratio test= 11.49  on 4 df,   p=0.02
Wald test            = 9.56  on 4 df,   p=0.05
Score (logrank) test = 9.27  on 4 df,   p=0.05
```

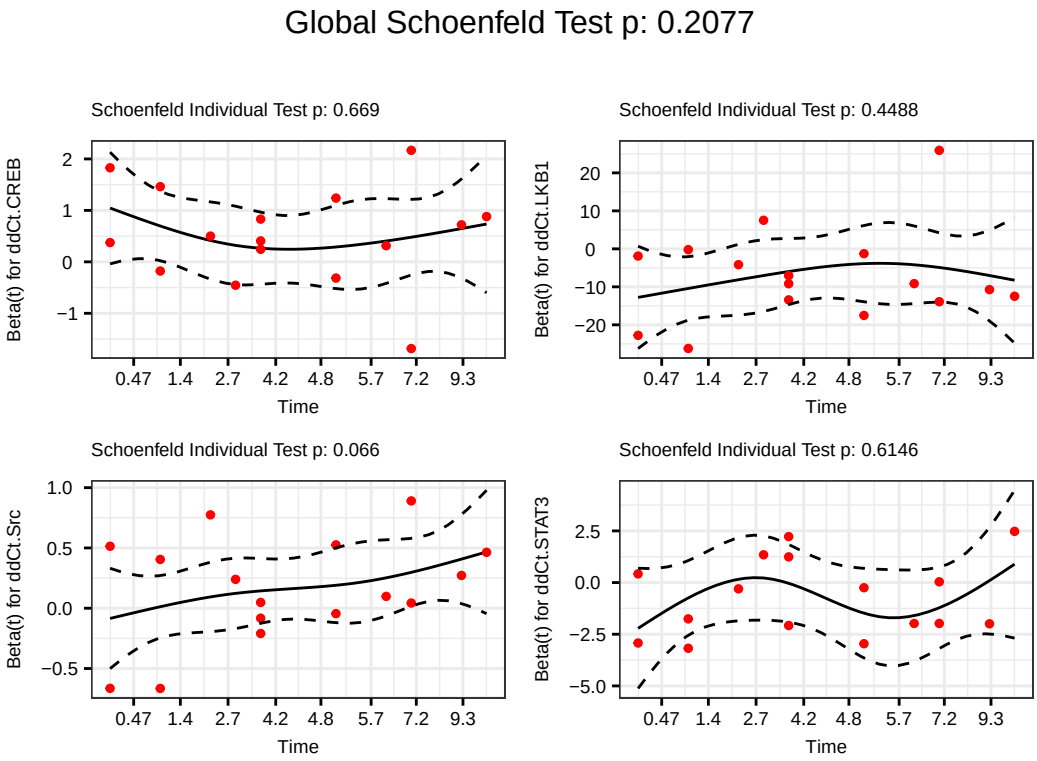
Evaluación de validez del mejor modelo de regresión de Cox obtenido (SLP)

Verificación de suposición de riesgos proporcionales

Tal como se indicaba anteriormente, para aplicar el modelo se debe efectuar la comprobación de la suposición de riesgos proporcionales y para ellos se utilizará la función *cox.zph()* del paquete *survival* de R, aplicándolo al modelo lineal de regresión obtenido. A continuación se muestra el resultado del test en formato de resumen y en la gráfica Gr.17.

```
              rho chisq     p
ddCt.CREB  -0.100 0.183 0.669
ddCt.LKB1   0.168 0.574 0.449
ddCt.Src    0.349 3.381 0.066
ddCt.STAT3  0.158 0.253 0.615
GLOBAL         NA 5.887 0.208
```

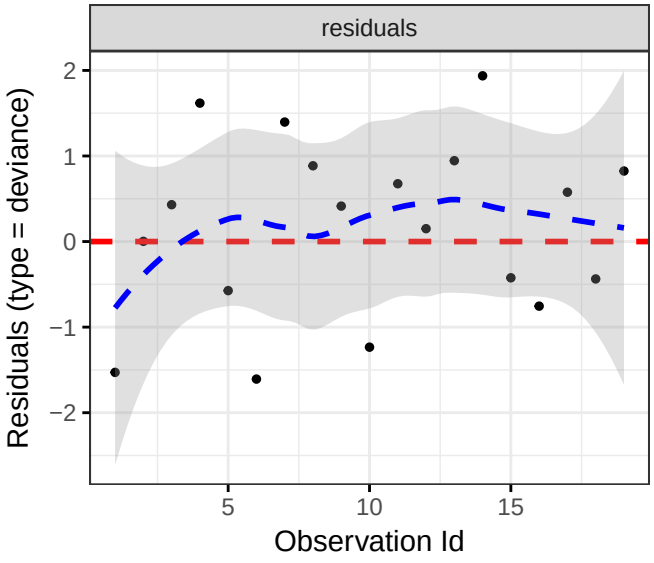
Según se puede ver, el resultado del test de riesgos proporcionales no es significativo al 5% ni globalmente, ni para las diferentes covariables del modelo, por lo que se puede concluir que no se viola el supuesto de riesgos proporcionales En la gráfica Gr.17 se muestra el test individual de Schoenfeld para las betas resultantes del modelo.



Gráfica Gr.17

Prueba de observaciones influyentes (outliers)

A partir de la desviación de los residuos, calculada como una transformada normalizada de la martingala residual, en la gráfica Gr.18, se pueden ver los outliers (valores alejados de la línea de ajuste) del modelo, que aunque hay algún valor que puede destacar, no parecen ser muy excesivos.

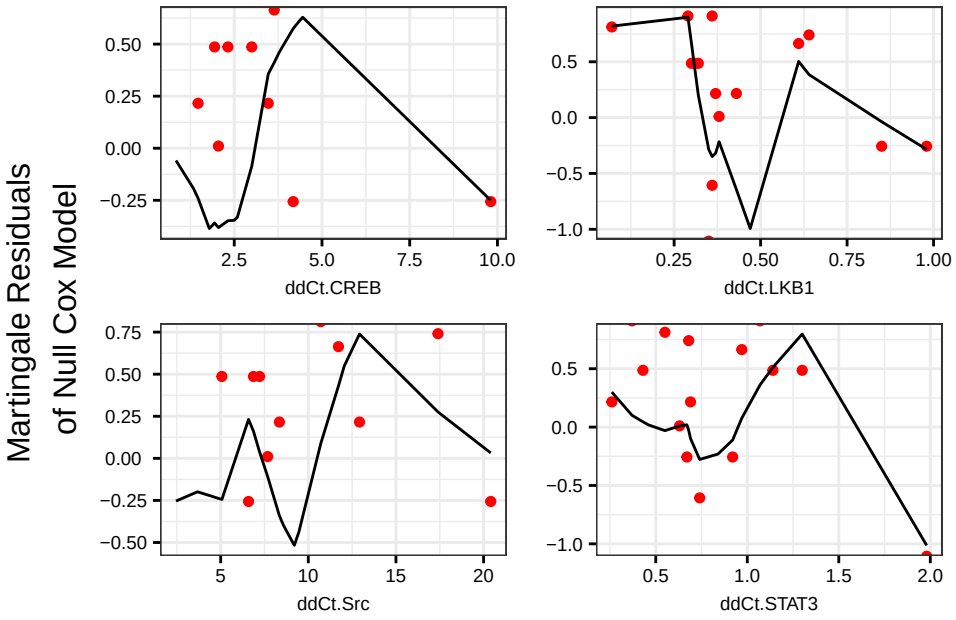


Gráfica Gr.18

Ver linealidad

Para ver la linealidad del modelo o para evaluar la forma funcional de las covariables, se muestra a continuación

la gráfica Gr.19, donde se muestran los residuos de martingala del modelo de Cox nulo contra las covariables continuas. Los residuos martingala con valor cercano a 1, indican individuos que murieron demasiado pronto respecto a lo esperado y los valores negativos grandes indican individuos que sobrepasaron la expectativa de vida según el modelo. Para efectuar la gráficas se utiliza la función *ggcoxfunctional()* del paquete *survminer* de R.



Gráfica Gr.19

Análisis de interacciones de covariables en el modelo ajustado (SLP)

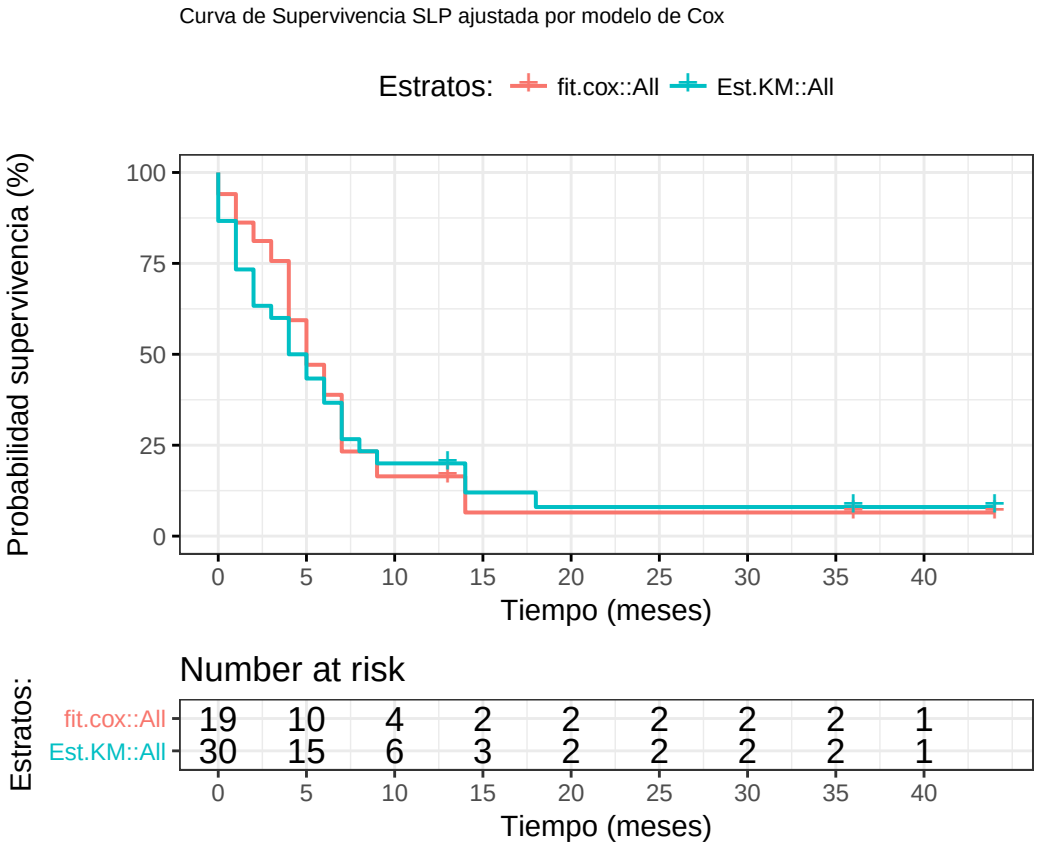
A continuación se analizan las interacciones entre los tres genes significativos en el modelo anterior (*CREB, LKB1 y Src*), para ver si alguna de las interacciones también es significante al 0.05.

```
[1] "Formula = ObjSLP ~ ddCt.CREB * ddCt.LKB1 * ddCt.Src + ddCt.STAT3"

Call:
coxph(formula = fmla, data = Kras.mostra)

                                coef exp(coef)  se(coef)     z     p
ddCt.CREB                   2.79e+00  1.62e+01  2.19e+00  1.27 0.203
ddCt.LKB1                  -2.93e+01  1.84e-13  2.95e+01 -1.00 0.319
ddCt.Src                   -3.68e-01  6.92e-01  9.42e-01 -0.39 0.696
ddCt.STAT3                 -1.49e+00  2.26e-01  6.76e-01 -2.20 0.028
ddCt.CREB:ddCt.LKB1        -1.91e+00  1.48e-01  5.48e+00 -0.35 0.727
ddCt.CREB:ddCt.Src         -7.49e-02  9.28e-01  2.21e-01 -0.34 0.735
ddCt.LKB1:ddCt.Src          9.07e-01  2.48e+00  3.05e+00  0.30 0.766
ddCt.CREB:ddCt.LKB1:ddCt.Src  1.67e-01  1.18e+00  6.65e-01  0.25 0.801

Likelihood ratio test=22.8  on 8 df, p=0.004
n= 19, number of events= 16
   (11 observations deleted due to missingness)
```

Vemos que no hay significación al 0.05 en ninguna de las interacciones entre las expresiones génicas.

En la gráfica Gr.20 se muestra la comparativa entre la estimación según el modelo de Kaplan - Meier y el modelo de regresión de Cox seleccionado como el mejor ajustado (en rojo en las tablas 16 y 17). En el

resumen de estratos debajo de la gráfica se puede ver la diferencia de la n de muestras en cada modelo, dado que en el modelo de Cox se han utilizado todas las covariables, muchas de las cuales tienen valores omitidos (NA).



Gráfica Gr.20

Se puede observar que el ajuste del modelo de Cox es casi paralelo a la función de estimación de supervivencia de Kaplan - Meier, aunque como indicaba anteriormente, el número pequeño de muestras disponibles en el estudio y la diferencia de muestras analizadas en cada uno de los dos modelos hace que el resultado sea difícil de evaluar.

Resultados de análisis de SLP

Tal como se ha podido comprobar en el proceso de análisis anterior, los modelos de regresión de Cox no estratificados no han mostrado evidencia significativa de que ninguno de los genes tenga influencia sobre el tiempo de supervivencia. En el caso de la estratificación por niveles Alto, Medio y Bajo, ha mostrado evidencia significativa al 5% de que la expresión del gen *LKB1* puede tener influencia en el tiempo de supervivencia, al igual que en el caso del modelo estratificado en cuantiles con percentil superior o igual a *25* (*>=Q25*), que han mostrados significancia al 5% también la expresión del gen *LKB1* y también el gen *HES1*.

En el análisis de los mejores puntos de corte, se han obtenido significancia en los genes *HES1*, *LKB1* y *Src*, obteniendo las siguientes particiones:

En el modelo de regresión de Cox Multivariable se ha ajustado un modelo que indica que las expresiones de los genes *CREB*, *LKB1* y *Src* presentan evidencia significativa de que tienen influencia en el SLP. El modelo es el indicado en rojo en la tabla 19.

Table 20: Mejores puntos de corte de HES1, LKB1 y Src

HES1.st	LKB1.st	Src.st
< 0.35 : 5	< 0.35 : 7	< 10.06 :16
>= 0.35 :14	>= 0.35 :13	>= 10.06 :12
NA's :11	NA's :10	NA's : 2

Superviviencia Global (SGL)

Supervivencia Global (SGL)

Para calcular la variable *time* para el análisis de Supervivencia Global (SGL), al igual que para el proceso de análisis SLP, como fecha de inicio del estudio se han tenido en cuenta las mismas fechas con los condicionantes de disponibilidad descritos en el apartado de SLP. El tiempo se ha calculado también en meses. Como fecha de producción del evento, se tomó la fecha de éxitus del paciente y en caso de desconocerse ésta se tomó la última fecha de seguimiento disponible como fin de periodo, marcando estos pacientes como censurados (en la tabla 5, la variable *ObjSGL* de los pacientes censurados tiene un símbolo +).

Estimación de SGL sin agrupaciones

A continuación se muestra el resumen global del análisis SGL sin estratificación y en la tabla 21 se muestra el resultado detallado de los valores obtenidos del análisis SGL, con los valores estimados para cada *time* de pacientes en riesgo, número de eventos producidos, número de censuras, probabilidad de supervivencia, el error estándar y el intervalo de confianza de la probabilidad. Adicionalmente se ha estimado por máxima verosimilitud el riesgo acumulado, que se ha añadido como la columna *CumHaz* en la tabla.

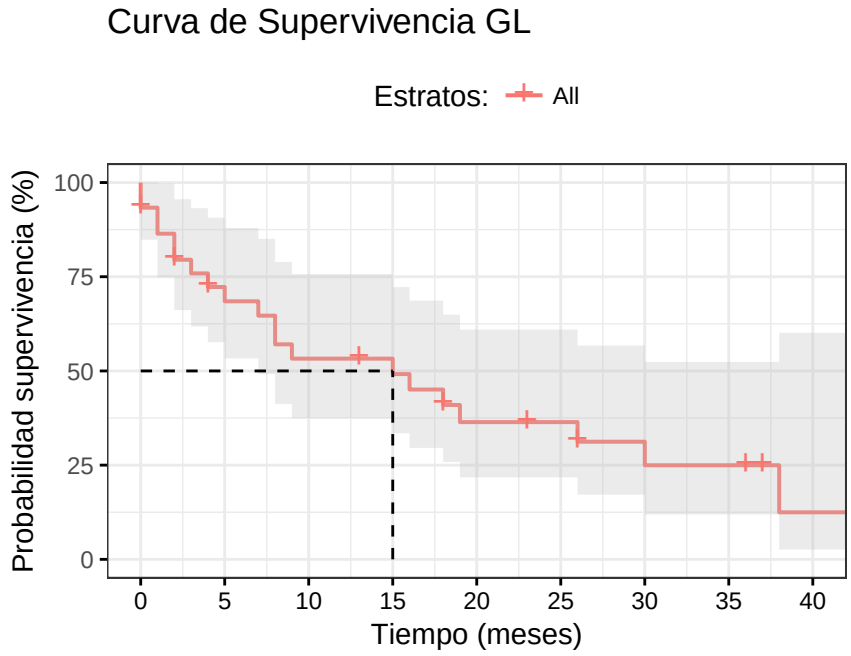
Como puede observar, la mediana de Supervivencia Global (50% probabilidad de supervivencia) es de *15* meses y la media de SGL de *18.20* meses.

```
Call: survfit(formula = ObjSGL ~ 1, data = Kras.mostra, type = "kaplan-meier")

        n     events     *rmean *se(rmean)     median    0.95LCL
    30.00      20.00      18.20       3.09      15.00       7.00
  0.95UCL
       NA
   * restricted mean with upper limit =  44
```

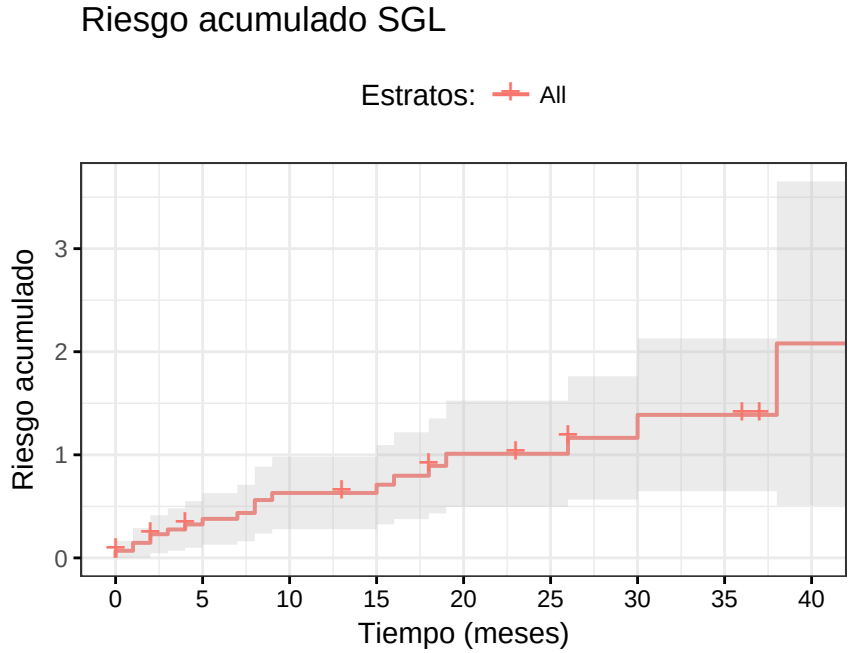
En las gráficas Gr.21a y 21b se muestran los datos de la tabla 21 mediante la curva de SGL y la curva de riesgos acumulados respectivamente, sin ninguna estratificación.

Table 21: Modelo de estimación SGL Kaplan-Meier

time	n.risk	n.event	n.censor	surv	std.err	upper	lower	CumHaz
0	30	2	1	0.9333	0.0488	1.0000	0.8482	0.0667
1	27	2	0	0.8642	0.0731	0.9973	0.7488	0.1407
2	25	2	1	0.7951	0.0939	0.9558	0.6614	0.2207
3	22	1	0	0.7589	0.1048	0.9320	0.6180	0.2662
4	21	1	1	0.7228	0.1156	0.9066	0.5762	0.3138
5	19	1	0	0.6847	0.1276	0.8794	0.5332	0.3664
7	18	1	0	0.6467	0.1399	0.8506	0.4917	0.4220
8	17	2	0	0.5706	0.1655	0.7893	0.4125	0.5396
9	15	1	0	0.5326	0.1793	0.7569	0.3747	0.6063
13	14	0	1	0.5326	0.1793	0.7569	0.3747	0.6063
15	13	1	0	0.4916	0.1964	0.7224	0.3345	0.6832
16	12	1	0	0.4506	0.2148	0.6866	0.2958	0.7666
18	11	1	1	0.4097	0.2350	0.6494	0.2585	0.8575
19	9	1	0	0.3642	0.2629	0.6097	0.2175	0.9686
23	8	0	1	0.3642	0.2629	0.6097	0.2175	0.9686
26	7	1	1	0.3121	0.3049	0.5673	0.1717	1.1114
30	5	1	0	0.2497	0.3781	0.5239	0.1190	1.3114
36	4	0	1	0.2497	0.3781	0.5239	0.1190	1.3114
37	3	0	1	0.2497	0.3781	0.5239	0.1190	1.3114
38	2	1	0	0.1249	0.8018	0.6011	0.0259	1.8114
44	1	0	1	0.1249	0.8018	0.6011	0.0259	1.8114



Gráfica Gr.21a



Gráfica Gr.21b

Estimación de SGL con agrupación por sexo y por tipo de mutación KRAS

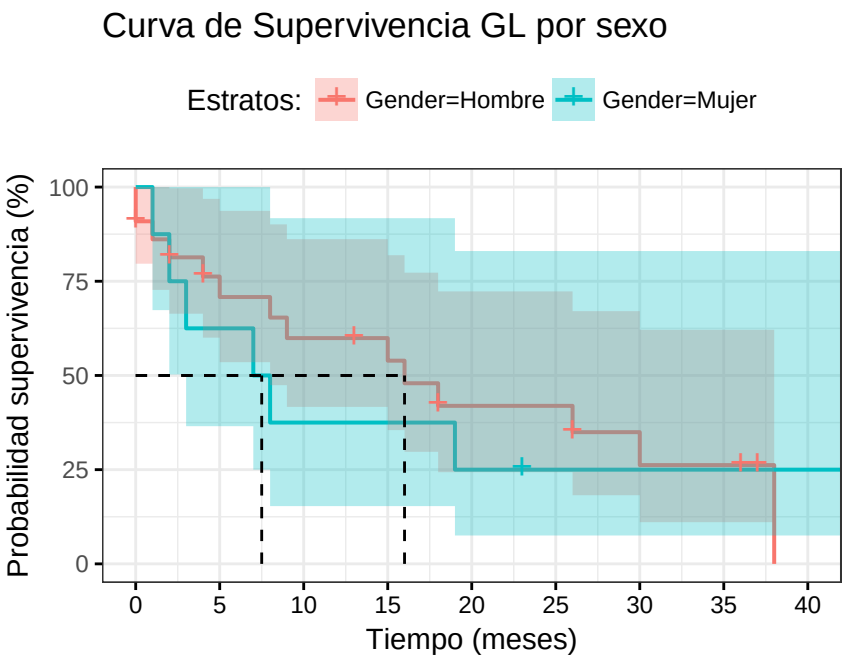
A continuación se muestra el resumen de estadísticos del modelo y los gráficos de supervivencia GL por estrato, Gr.22 y Gr. 23.

Resumen modelo estratificado por sexo:

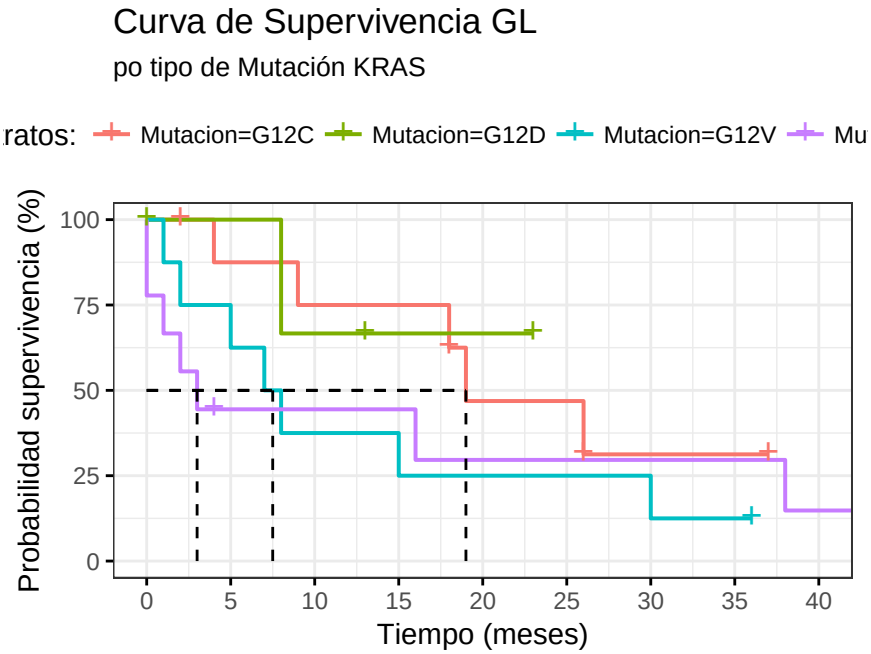
```
Call: survfit(formula = ObjSGL ~ Gender, data = Kras.mostra, type = "kaplan-meier")

               n events *rmean *se(rmean) median 0.95LCL 0.95UCL
Gender=Hombre 22     14   18.9       3.32   16.0       8      NA
Gender=Mujer   8      6   15.2       5.58    7.5       3      NA
    * restricted mean with upper limit =  41
```

Resumen del modelo estratificado por tipo de mutación KRAS, según las mutaciones más frecuentes y resto (*G12C, G12D, G12V y otras*):

```
Call: survfit(formula = ObjSGL ~ Mutacion, data = Kras.mostra, type = "kaplan-meier")

               n events *rmean *se(rmean) median 0.95LCL 0.95UCL
Mutacion=G12C  9      5   22.3       4.21   19.0      18      NA
Mutacion=G12D  4      1   27.0       7.76     NA       8      NA
Mutacion=G12V  8      7   13.1       4.39    7.5       5      NA
Mutacion=Otras 9      7   13.9       5.34    3.0       1      NA
    * restricted mean with upper limit =  36.5
```



Gráfica Gr.22



Gráfica Gr.23

Para verificar si las diferencias que se observan, tanto por sexo como por tipo de mutación, son significativas efectuaremos un test de diferencias con la función *survdiff()*, para determinar el grado de significancia.

```
Call:
survdiff(formula = ObjSGL ~ Gender, data = Kras.mostra)

                N Observed Expected (O-E)^2/E (O-E)^2/V
Gender=Hombre 22       14    14.66    0.0299     0.117
```

```
Gender=Mujer   8        6     5.34    0.0821     0.117

 Chisq= 0.1  on 1 degrees of freedom, p= 0.7
```

```
Call:
survdiff(formula = ObjSGL ~ Mutacion, data = Kras.mostra)

               N Observed Expected (O-E)^2/E (O-E)^2/V
Mutacion=G12C  9        5     7.36     0.756     1.290
Mutacion=G12D  4        1     2.18     0.640     0.756
Mutacion=G12V  8        7     5.00     0.804     1.127
Mutacion=Otras 9        7     5.46     0.432     0.717

 Chisq= 2.9  on 3 degrees of freedom, p= 0.4
```

En ambos casos, se puede ver que el p-valor obtenido es > 0.05, por lo que se descarta la hipótesis de diferencias entre estimaciones por los diferentes grupos comparados.

Ajuste de un modelo de regresión de Cox para SGL

Modelo de regresión de Cox Univariable (SGL).

Al igual que para el análisis de SLP, para efectuar el modelo de ajuste de regresión de Cox de SGL de cada covariable, se utiliza la función *fit.cox()*, para las variables de conjunto de datos a analizar. La implementación de esta función se puede ver en el apartado de anexo de código R de este informe dinámico.

En la tabla 22, se muestra el resultado de la ejecución de la función para el modelo de ajuste de regresión de Cox univariable para la variable *ObjSGL* con cada covariable de expresión de cada gen.

Table 22: Resultados del modelo de regresión de Cox Univariable (GL)

	n	n.eventos	beta	HR (95% CI for HR)	wald.test	p.value	CutP
ddCt.HES1	19	13	-0.1284	0.8795 (0.2912-2.6563)	0.05	0.82	0.34
ddCt.CDCP1	28	18	-0.0043	0.9957 (0.9581-1.0347)	0.05	0.8256	15.9
ddCt.AXL	25	15	0.0735	1.0763 (0.7881-1.4698)	0.21	0.6439	3.69
ddCt.YAP1	25	16	-0.0187	0.9815 (0.8682-1.1094)	0.09	0.7648	1.77
ddCt.CREB	23	14	-0.3565	0.7001 (0.4662-1.0514)	2.95	0.0857	4.45
ddCt.LKB1	20	12	-5.4628	0.0042 (0-0.5831)	4.73	0.0297	0.43
ddCt.Src	28	19	0.0082	1.0082 (0.9898-1.0269)	0.75	0.3857	5.07
ddCt.STAT3	26	17	0.3364	1.3999 (0.4401-4.4528)	0.32	0.5689	0.37

Se puede apreciar que la expresión del gen *LKB1* es significativa al 5% con $p - valor = 0.0297$ en la supervivencia global de los pacientes. A continuación se realiza un modelo de regresión de Cox multivarible para ver al igual que con la SLP si puede haber una relación lineal significativa entre la SGL y algunas de las expresiones génicas.

Modelo de regresión de Cox Univariable estratificado (SGL)

Se utilizará la misma estratificación que la realizada para SLP.

Modelos de regresión de Cox con estratificación según expresión Alta, Media y Baja

A continuación, en la tabla 23, se muestra el resultado del modelo de regresión de Cox estratificado.

Table 23: Resultados del modelo de regresión de Cox Univariable Estratificado (SGL)

Variable	n	n.eventos	wald.test	p.value	strat	beta	HR (95% CI for HR)	Pr(>|z|)
HES1.st	19	13	0.78	0.6774	HES1.stBaja	0.6102	1.8408 (0.452-7.4963)	0.3944
					HES1.stMedia	0.4326	1.5413 (0.4121-5.7643)	0.5204
CDCP1.st	28	18	1.9	0.3863	CDCP1.stBaja	0.4717	1.6027 (0.5027-5.1102)	0.4253
					CDCP1.stMedia	-0.3374	0.7136 (0.1996-2.5514)	0.6038
AXL.st	25	15	0.72	0.6974	AXL.stBaja	-0.4593	0.6317 (0.1608-2.4825)	0.5107
					AXL.stMedia	0.0678	1.0702 (0.3107-3.6858)	0.9144
YAP1.st		16	0.17	0.9188	YAP1.stBaja	-0.1574	0.8543 (0.2386-3.0595)	0.8089
					YAP1.stMedia	0.1307	1.1396 (0.3354-3.8723)	0.8342
CREB.st	23	14	2.26	0.3224	CREB.stBaja	1.0689	2.9121 (0.7198-11.7817)	0.1339
					CREB.stMedia	0.7669	2.1531 (0.5074-9.1362)	0.2984
LKB1.st	20	12	6.45	0.0397	LKB1.stBaja	2.6439	14.0681 (1.6694-118.5546)	0.015
					LKB1.stMedia	1.7932	6.0084 (0.7002-51.5612)	0.1021
Src.st	28	19	0.88	0.6439	Src.stBaja	-0.398	0.6717 (0.2014-2.2398)	0.5172
					Src.stMedia	0.1358	1.1455 (0.3922-3.3452)	0.8038
STAT3.st	26	17	0.43	0.8075	STAT3.stBaja	-0.2844	0.7525 (0.2075-2.7286)	0.6652
					STAT3.stMedia	0.143	1.1537 (0.3713-3.585)	0.8048

Una vez realizado el modelo de ajuste de regresión de Cox con la expresión estratificada de los diferentes genes, se puede comprobar (datos en rojo en la tabla 23) que el gen *LKB1* muestra evidencia de significación al 5% en el estrato de expresión Baja respecto a la expresión alta, con un $p - valor = 0.015$ y con un test de Wald también significativo con un $p - valor = 0.0397$.

Modelos de regresión de Cox con estratificación según expresión Alta, Baja-Media

A continuación, en la tabla 24, se muestra el resultado del modelo de regresión de Cox estratificado.

Table 24: Resultados del modelo de regresión de Cox Univariable Estratificado (SGL)

Variable	n	n.eventos	wald.test	p.value	strat	beta	HR (95% CI for HR)
HES1.st	19	13	0.7	0.4024	HES1.stBaja-Media	0.5061	1.6589 (0.5073-5.4244)
CDCP1.st	28	18	0.02	0.8932	CDCP1.stBaja-Media	0.0727	1.0754 (0.3723-3.1064)
AXL.st	25	15	0.09	0.7641	AXL.stBaja-Media	-0.1732	0.841 (0.2713-2.6065)
YAP1.st	25	16	0	0.9823	YAP1.stBaja-Media	-0.0118	0.9883 (0.349-2.7984)
CREB.st	23	14	1.99	0.1582	CREB.stBaja-Media	0.922	2.5143 (0.6986-9.0486)
LKB1.st	20	12	4.27	0.0389	LKB1.stBaja-Media	2.1602	8.6731 (1.1165-67.3727)
Src.st	28	19	0.04	0.8399	Src.stBaja-Media	-0.1013	0.9037 (0.3383-2.4143)
STAT3.st	26	17	0.01	0.9336	STAT3.stBaja-Media	-0.0434	0.9575 (0.3448-2.6589)

En esta ocasión, tal como reflejan los resultados de los diferentes modelos de Cox univariables, con esta estratificación también existe evidencia significativa al 5% en el gen *LKB1*, con un $p - valor = 0.0389$ entre expresión Alta y Baja-Media.

Modelos de regresión de Cox con estratificación según expresión $< Q25$ ***y*** $>= Q25$

A continuación, en la tabla 25, se muestra el resultado del modelo de regresión de Cox estratificado.

Como resultado de los modelos de Cox analizados, se puede ver (en rojo) que hay tres covariables que tienen evidencia de significación al 5%, *HES1*, *CDCP11* y *LKB1*, en su estrato de $>= Q25$ respecto a $< Q25$, con

Table 25: Resultados del modelo de regresión de Cox Univariable Estratificado (SGL)

Variable	n	n.eventos	wald.test	p.value	strat	beta	HR (95% CI for HR)
HES1.st	19	13	4.45	0.0349	HES1.st>=Q25	-1.5201	0.2187 (0.0533-0.8979)
CDCP1.st	28	18	3.95	0.0469	CDCP1.st>=Q25	-0.9896	0.3717 (0.1401-0.9864)
AXL.st	25	15	0.23	0.6315	AXL.st>=Q25	0.2892	1.3354 (0.4095-4.3544)
YAP1.st	25	16	0.17	0.6824	YAP1.st>=Q25	0.2681	1.3074 (0.3621-4.7211)
CREB.st	23	14	0.26	0.6132	CREB.st>=Q25	-0.3093	0.7339 (0.2212-2.4353)
LKB1.st	20	12	8.44	0.0037	LKB1.st>=Q25	-2.1632	0.115 (0.0267-0.4949)
Src.st	28	19	1.36	0.2432	Src.st>=Q25	0.7404	2.0968 (0.6047-7.2712)
STAT3.st	26	17	2.17	0.1404	STAT3.st>=Q25	1.1321	3.1021 (0.6887-13.9731)

p-valores de *0.0349*, *0.0469* y *0.0037*, respectivamente.

Modelos de regresión de Cox con estratificación según mejores puntos de corte

Al igual que en el análisis de SLP, para SGL se ha efectuado un último modelo de ajuste de Cox Univariable estratificando la expresión de los genes según los mejores puntos de corte. Los puntos de corte aplicados para hacer la estratificación de la expresión de cada gen es la que se ha indicado en la tabla 22.

Table 26: Estratificación: Mejores puntos de corte

HES1.st	CDCP1.st	AXL.st	YAP1.st	CREB.st	LKB1.st	Src.st	STAT3.st
< 0.34 : 5	< 15.9 :13	< 3.69 : 7	< 1.77 : 2	< 4.45 :20	< 0.43 : 7	< 5.07 :16	< 0.37 : 7
>= 0.34 :14	>= 15.9 :15	>= 3.69 :18	>= 1.77 :23	>= 4.45 : 3	>= 0.43 :13	>= 5.07 :12	>= 0.37 :19
NA's :11	NA's : 2	NA's : 5	NA's : 5	NA's : 7	NA's :10	NA's : 2	NA's : 4

Table 27: Modelo de Cox Univariable Estratificado mejores puntos de corte

Variable	n	n.eventos	wald.test	p.value	strat	beta	HR (95% CI for HR)
HES1.st	19	13	4.45	0.0349	HES1.st>= 0.34	-1.5201	0.2187 (0.0533-0.8979)
CDCP1.st	28	18	1.93	0.1652	CDCP1.st>= 15.9	-0.6894	0.5019 (0.1896-1.3288)
AXL.st	25	15	0.53	0.466	AXL.st>= 3.69	0.4345	1.5442 (0.4801-4.9666)
YAP1.st	25	16	0	0.9982	YAP1.st>= 1.77	19.5172	299361866.8478 (0-Inf)
CREB.st	23	14	0	0.9981	CREB.st>= 4.45	-19.5095	0 (0-Inf)
LKB1.st	20	12	6.1	0.0135	LKB1.st>= 0.43	-1.4624	0.2317 (0.0726-0.7396)
Src.st	28	19	1.12	0.2907	Src.st>= 5.07	0.5044	1.656 (0.6497-4.221)
STAT3.st	26	17	2.17	0.1404	STAT3.st>= 0.37	1.1321	3.1021 (0.6887-13.9731)

En la tabla 26 se muestra el resultado de la estratificación, donde se pueden ver el número de muestras que han caído en cada lado del punto de corte. En la tabla 27 se muestran los resultados de los diferentes modelos de Cox ajustados para cada gen, aplicando la estratificación del mejor punto de corte. En rojo se indican los modelos que han resultado significativos al 5% y podemos ver que son los genes *HES1* y *LKB1*, con un p-valor de *0.0349, 0.0135*, respectivamente.

Validación significación de genes

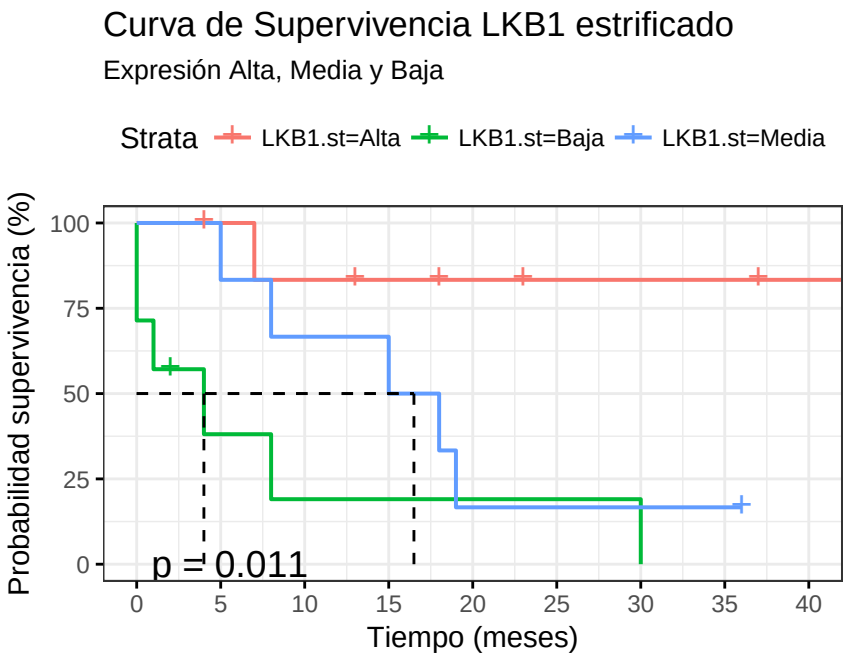
Para validar la significación de los modelos anteriormente descritos, se ha efectuado un test de Log-Rank de comparación por pares, cuyo resultado confirma la significación de todos ellos, según el resultado que se muestra a continuación.

```
    Pairwise comparisons using Log-Rank test

data:  Kras.strat33 and LKB1.st
```

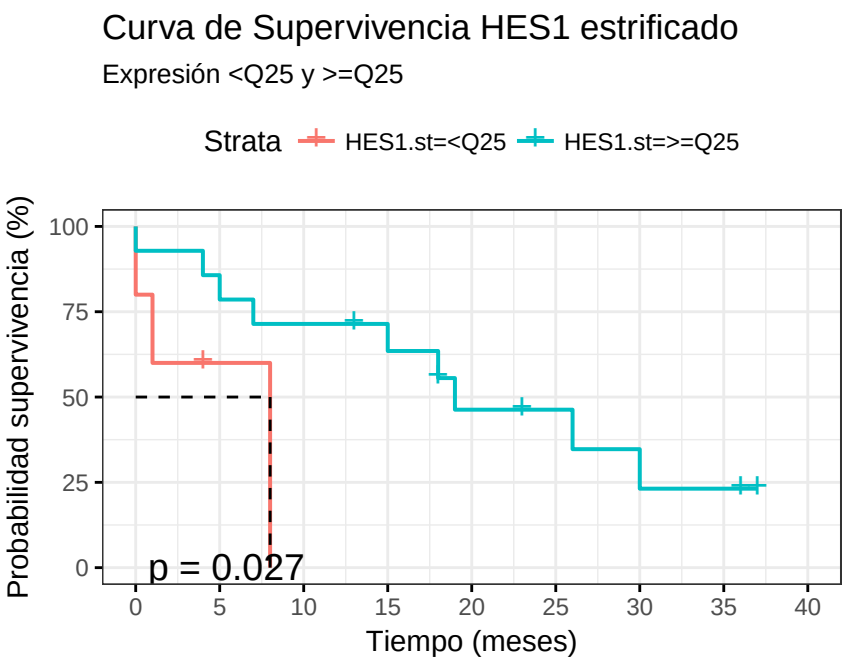
```
      Alta  Baja
Baja  0.017 -
Media 0.103 0.178

P value adjustment method: BH


    Pairwise comparisons using Log-Rank test

data:  Kras.strat3366 and LKB1.st

           Alta
Baja-Media 0.013

P value adjustment method: BH


    Pairwise comparisons using Log-Rank test

data:  Kras.stratQ25 and HES1.st

      <Q25
>=Q25 0.027

P value adjustment method: BH


    Pairwise comparisons using Log-Rank test

data:  Kras.stratQ25 and CDCP1.st

      <Q25
>=Q25 0.04

P value adjustment method: BH


    Pairwise comparisons using Log-Rank test

data:  Kras.stratQ25 and LKB1.st

      <Q25
>=Q25 0.0012

P value adjustment method: BH


    Pairwise comparisons using Log-Rank test

data:  Kras.stratBCP.GL and HES1.st

         < 0.34
>= 0.34  0.027

P value adjustment method: BH
```

```
    Pairwise comparisons using Log-Rank test

data:  Kras.stratBCP.GL and LKB1.st

         < 0.43
>= 0.43  0.0085

P value adjustment method: BH
```

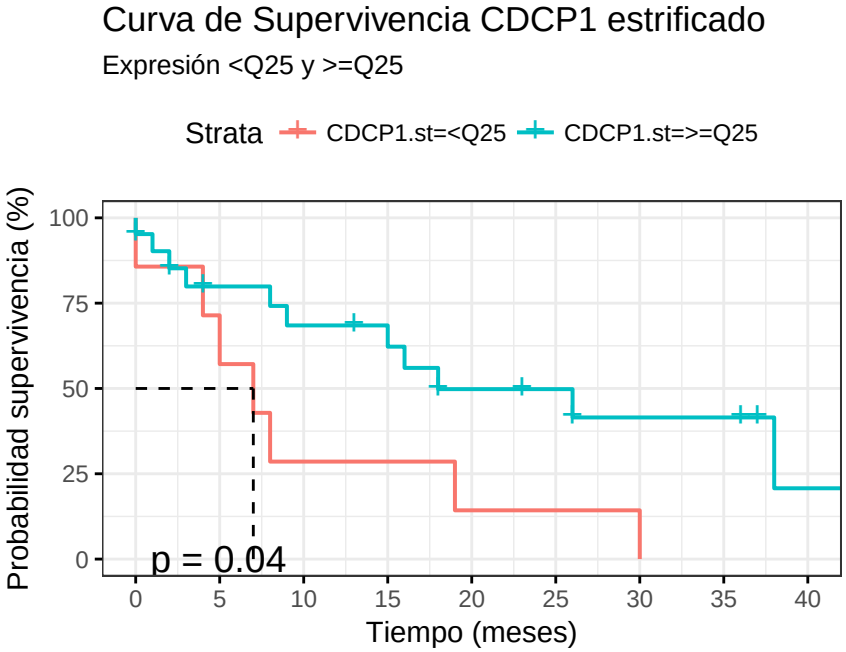
A continuación se muestra la gráfica de Supervivencia libre de progresión (SLP) que han evidenciado significancia en alguno de los modelos de Cox (gráficas Gr.24, Gr.25a-c, Gr.26a-b)



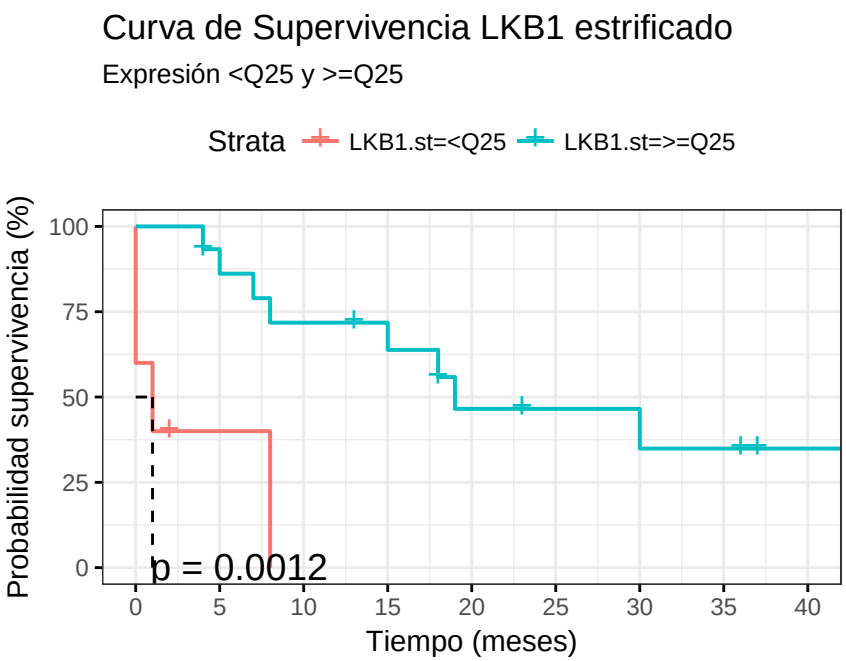
Gráfica Gr.24



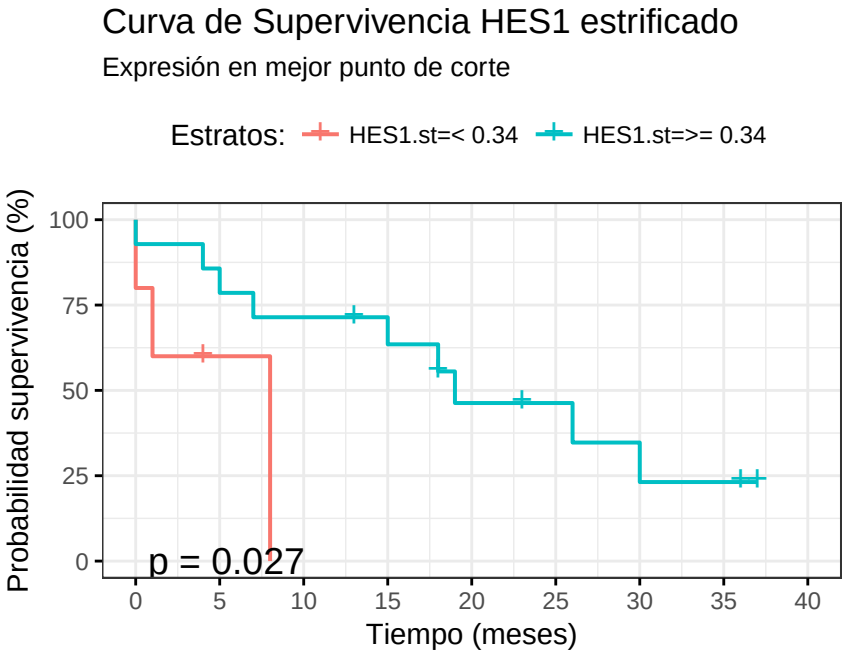
Gráfica Gr.25a



Gráfica Gr.25b



Gráfica Gr.25c

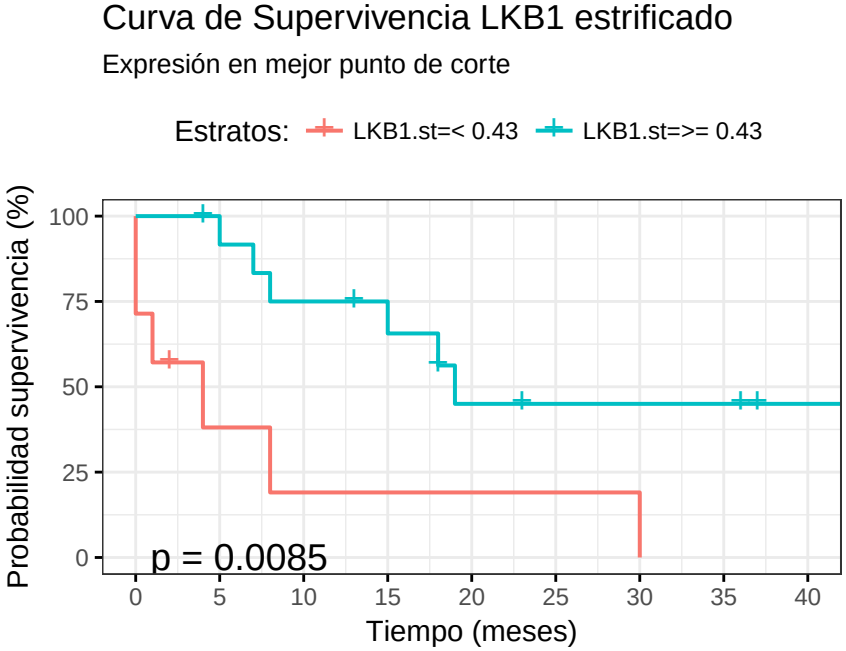


Gráfica Gr.26a

Estos son los valores de la media y mediana de la SLP estratificada con el mejor punto de corte para la expresión del gen *HES1*:

```
Call: survfit(formula = ObjSLP ~ HES1.st, data = Kras.stratBCP.GL,
    type = "kaplan-meier")

   11 observations deleted due to missingness
                n events *rmean *se(rmean) median 0.95LCL 0.95UCL
```

```
HES1.st=< 0.34    5       5   2.00       0.938    1.0        0       NA
HES1.st=>= 0.34  14      12   7.62       1.557    6.5        4       NA
    * restricted mean with upper limit =  20.5
```



Gráfica Gr.26b

Estos son los valores de la media y mediana de la SLP estratificada con el mejor punto de corte para la expresión del gen *HES1*:

```
Call: survfit(formula = ObjSLP ~ LKB1.st, data = Kras.stratBCP.GL,
    type = "kaplan-meier")

   10 observations deleted due to missingness
                 n events *rmean *se(rmean) median 0.95LCL 0.95UCL
LKB1.st=< 0.43   7      7   2.00       0.67      1       0      NA
LKB1.st=>= 0.43 13     10   9.74       2.27      7       5      NA
    * restricted mean with upper limit =  24
```

Modelo de regresión de Cox Multivariable (SGL)

En este modelo se deberá considerar que se excluirán las muestras en las cuales la expresión de alguno de los genes no esté informada (valor NA), por lo que el modelo se ajustará a la n mínima de las muestras que tengan valor de expresión para todos los genes.

```
Call:
coxph(formula = fmla, data = Kras.mostra)

  n= 17, number of events= 11
   (13 observations deleted due to missingness)

                coef  exp(coef)   se(coef)     z Pr(>|z|)
ddCt.HES1   1.208e+00  3.348e+00  1.100e+00  1.098   0.2722
```

```
ddCt.CDCP1 -4.824e-02  9.529e-01  3.787e-02 -1.274   0.2028
ddCt.AXL    6.588e-01  1.932e+00  5.729e-01  1.150   0.2502
ddCt.YAP1  -2.075e-01  8.126e-01  4.016e-01 -0.517   0.6054
ddCt.CREB   2.250e-01  1.252e+00  4.880e-01  0.461   0.6448
ddCt.LKB1  -1.337e+01  1.555e-06  6.070e+00 -2.203   0.0276 *
ddCt.Src    3.437e-01  1.410e+00  1.852e-01  1.855   0.0635 .
ddCt.STAT3 -1.160e+00  3.133e-01  1.098e+00 -1.057   0.2904
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

           exp(coef) exp(-coef) lower .95 upper .95
ddCt.HES1  3.348e+00  2.987e-01 3.873e-01   28.9358
ddCt.CDCP1 9.529e-01  1.049e+00 8.847e-01    1.0263
ddCt.AXL   1.932e+00  5.175e-01 6.287e-01    5.9400
ddCt.YAP1  8.126e-01  1.231e+00 3.699e-01    1.7854
ddCt.CREB  1.252e+00  7.986e-01 4.812e-01    3.2589
ddCt.LKB1  1.555e-06  6.432e+05 1.059e-11    0.2283
ddCt.Src   1.410e+00  7.091e-01 9.808e-01    2.0275
ddCt.STAT3 3.133e-01  3.191e+00 3.645e-02    2.6937

Concordance= 0.781  (se = 0.105 )
Rsquare= 0.465   (max possible= 0.946 )
Likelihood ratio test= 10.62  on 8 df,   p=0.2
Wald test            = 6.54  on 8 df,   p=0.6
Score (logrank) test = 9.15  on 8 df,   p=0.3
```

Los resultados del modelo multivariable confirman que se han descartado 13 muestras del total de 30, por lo que $n = 17$. en el modelo analizado. Solo aparece un gen significativo con un* p-valor < 0.05*, el* LKB1*, pero la significancia global del modelo es > 0.05 por lo que el ajuste del modelo no confirma que alguna de las expresiones de los genes sea un factor influyente en el tiempo libre de progresión, al menos de forma global.

Mejora del modelo de regresión de Cox Multivariable (SGL)

Para intentar buscar un mejor modelo de ajuste se va a proceder a ir eliminando las covariables no significativas del modelo una a una hasta dejar solo la variable significativa que se ha visto en el modelo completo (*LKB1*), para ver qué modelo obtiene un ajuste mejor. Para ello se utilizará al igual que en el modelo para SLP, la función *fit.cox2()*. La implementación de esta función se puede ver en el apartado de anexo de código R de este informe dinámico.

En las tablas 28 y 29 que se muestran a continuación, se puede comprobar que el modelo regresión que es significativo según los tres test de ajuste, el de *razón de verosimilitud (Likekihook)*, el *test de Wald* y el *test de los puntajes o score*, es el que utiliza las dos covariables de la expresión del gen *LKB1 y Scr*, las cuales tienen significancia al 5%. El índice de concordancia del modelo $C = 0.79$, que es bastante correcto. Solo se han podido utilizar *20* muestras de las 30 de conjunto de datos inicial y se han producido *12* eventos.

También se puede ver que en todos los modelos ajustados, la covariable *LKB1* tiene significación al 5% y hay un modelo que adicionalmente presenta también significación en la covariable *Scr*. Solo hay tres modelos que tienen significación al 5% al menos en la prueba de razón de verosimilitud (Likekihook).

Viendo estos resultados se puede concluir que por obtener significancia en los tres test del modelo, el modelo más ajustado es el marcado en color rojo en la tablas, con las covariables *LKB1 y Src*, con un índice de concordancia de $C = 0.79$. El modelo en azul en las tablas también sería un modelo ajustado con tres variables (*LKB1, Src y STAT3*), aunque solo con significancia *LKB1* y *Src*, con un índice de concordancia de $C = 0.80$, un poco más alto, pero donde solo el test del ratio de verosimilitud es significativo. Por tanto, parece ser que las variables *LKB1* y *Src* pueden influir en el tiempo de supervivencia global de los pacientes.

En conclusión, se puede ver que en el caso de SGL el resultado obtenido con el modelo de Cox Univariable solo presenta significancia en el gen *LKB1* y el modelo Cox Multivariable, presenta significancia en las covariables *LKB1* y *Src*.

Table 28: Variables significativas de cada modelo de regresión de Cox SGL

Modelo	Var.Significativas	p.valores
HES1+CDCP1+AXL+YAP1+CREB+LKB1+Src+STAT3	LKB1	0.0276
CDCP1+AXL+YAP1+CREB+LKB1+Src+STAT3	LKB1, Src	0.0313, 0.0451
CDCP1+YAP1+CREB+LKB1+Src+STAT3	LKB1	0.0277
AXL+YAP1+CREB+LKB1+Src+STAT3	LKB1	0.0312
YAP1+CREB+LKB1+Src+STAT3	LKB1	0.0328
CREB+LKB1+Src+STAT3	LKB1	0.0467
CREB+LKB1+Src	LKB1	0.0478
LKB1+Src+STAT3	LKB1, Src	0.0059, 0.0456
LKB1+Src	LKB1, Src	0.0058, 0.0494

Table 29: Significación global de cada modelo

Modelo	n	n.eventos	Conc.	likelih.pval	wald.pval	score.paval
HES1+CDCP1+AXL+YAP1+CREB+LKB1+Src+STAT3	17	11	0.781	0.2243	0.5867	0.3302
CDCP1+AXL+YAP1+CREB+LKB1+Src+STAT3	19	11	0.7881	0.103	0.5591	0.4758
CDCP1+YAP1+CREB+LKB1+Src+STAT3	19	11	0.7627	0.0859	0.4194	0.3759
AXL+YAP1+CREB+LKB1+Src+STAT3	19	11	0.7712	0.1224	0.3664	0.4212
YAP1+CREB+LKB1+Src+STAT3	19	11	0.7712	0.0779	0.2779	0.306
CREB+LKB1+Src+STAT3	19	11	0.7881	0.063	0.2174	0.2625
CREB+LKB1+Src	19	11	0.7203	0.0329	0.1335	0.1647
LKB1+Src+STAT3	20	12	0.8015	0.0061	0.0509	0.0624
LKB1+Src	20	12	0.7941	0.0022	0.0206	0.03

Vemos el modelo completo de ajuste indicado en rojo en la tabla, donde se pueden ver todos los estadísticos, entre otros los coeficientes β, los ratios $HR = exp(coef)$ (estimación de riesgos relativos), la desviación estándar y los intervalos de confianza de cada variable.

Si se analiza la estimación de riesgo $HR = exp(coef)$ en el resumen mostrado de cada covariable, se puede concluir con significancia de 0.05 que el modelo indica que a igualdad de valores en el resto de expresiones de los otros genes, el aumento de un punto de expresión del gen *Scr*, significa un aumento del 21.4% *(HR=1.214)* en el riesgo de muerte, mientras que en el caso de *LKB1* el aumento de su nivel de expresión, manteniendo la expresión de los otros genes, tendría un efecto de disminución del riesgo de muerte ya que vemos que el valor de su coeficiente $\beta = -7.80$ es negativo, pero dado que su valor de $HR = 0.0004$ es muy cercano a cero, este no tendría una afectación significante en el pronóstico.

```
[1] "Formula = ~"                    "Formula = ObjSGL"
[3] "Formula = ddCt.LKB1 + ddCt.Src"

Call:
coxph(formula = fmla, data = Kras.mostra)

  n= 20, number of events= 12
   (10 observations deleted due to missingness)

                coef  exp(coef)   se(coef)      z Pr(>|z|)
ddCt.LKB1 -7.8008989  0.0004094  2.8284986 -2.758  0.00582 **
ddCt.Src   0.1941104  1.2142303  0.0987992  1.965  0.04945 *
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

          exp(coef) exp(-coef) lower .95 upper .95
ddCt.LKB1 0.0004094  2442.7968 1.601e-06    0.1046
ddCt.Src  1.2142303     0.8236 1.000e+00    1.4737

Concordance= 0.794  (se = 0.1 )
Rsquare= 0.458   (max possible= 0.946 )
Likelihood ratio test= 12.24  on 2 df,   p=0.002
Wald test            = 7.76  on 2 df,   p=0.02
Score (logrank) test = 7.01  on 2 df,   p=0.03
```

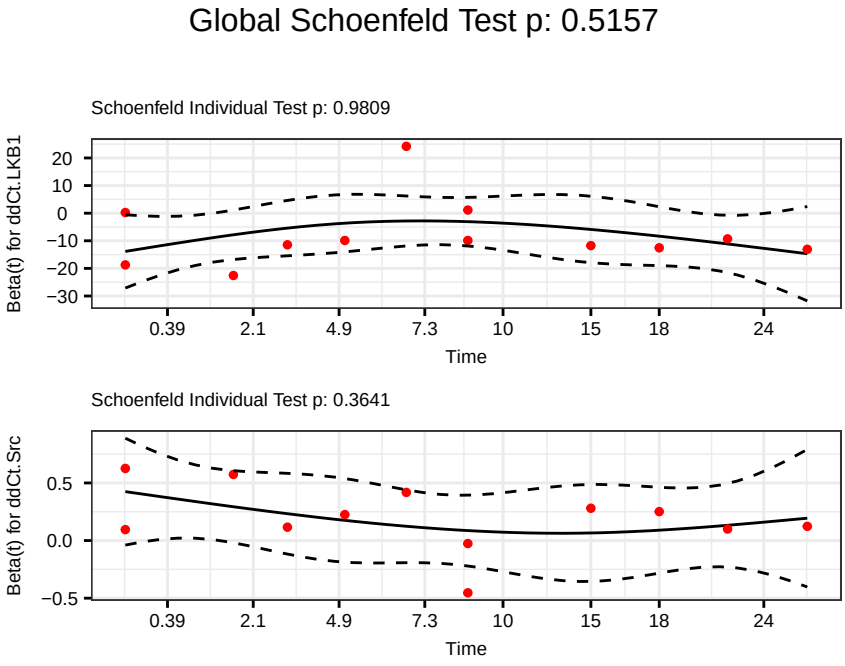
Evaluación de validez del mejor modelo de regresión de Cox obtenido (SGL)

Verificación de suposición de riesgos proporcionales

Tal como se indicaba en el modelo ajustado para SLP, para verificar la viabilidad de la aplicación del modelo de regresión de Cox, se debe efectuar la comprobación de la suposición de riesgos proporcionales. Para ello se utilizará la función *cox.zph()* del paquete *survival* de R, aplicándolo al modelo lineal de regresión obtenido. A continuación se muestra el resultado del test en formato de resumen y en la gráfica Gr.27.

```
               rho    chisq     p
ddCt.LKB1 -0.00588 0.000575 0.981
ddCt.Src  -0.32936 0.823583 0.364
GLOBAL           NA 1.324532 0.516
```

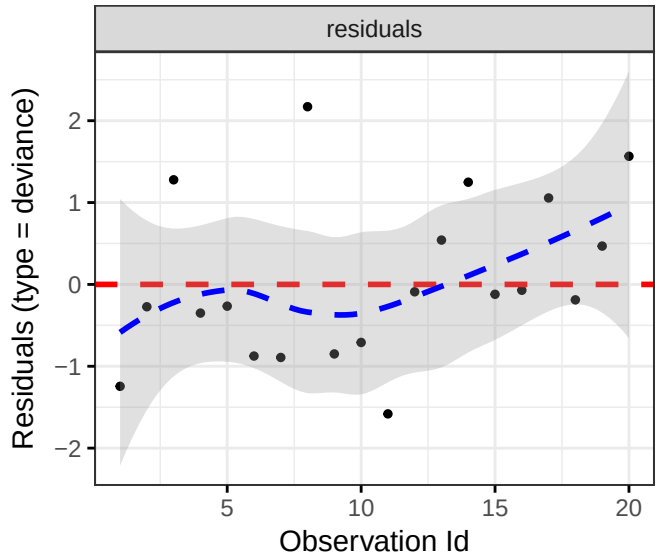
Según se puede ver, el resultado del test de riesgos proporcionales no es significativo al 5% ni globalmente, ni para las diferentes covariables del modelo, por lo que se puede concluir que no se viola el supuesto de riesgos proporcionales. En la gráfica Gr.27 se muestra el test individual de Schoenfeld para las betas resultantes del modelo.



Gráfica Gr.27

Prueba de observaciones influyentes (outliers)

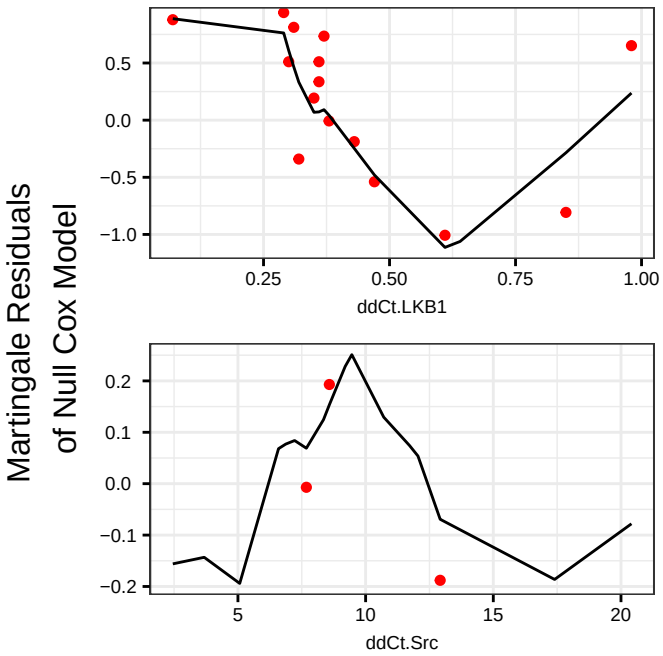
A partir de la desviación de los residuos, en la gráfica Gr.28, se pueden ver los outliers (valores alejados de la línea de ajuste) del modelo.



Gráfica Gr.28

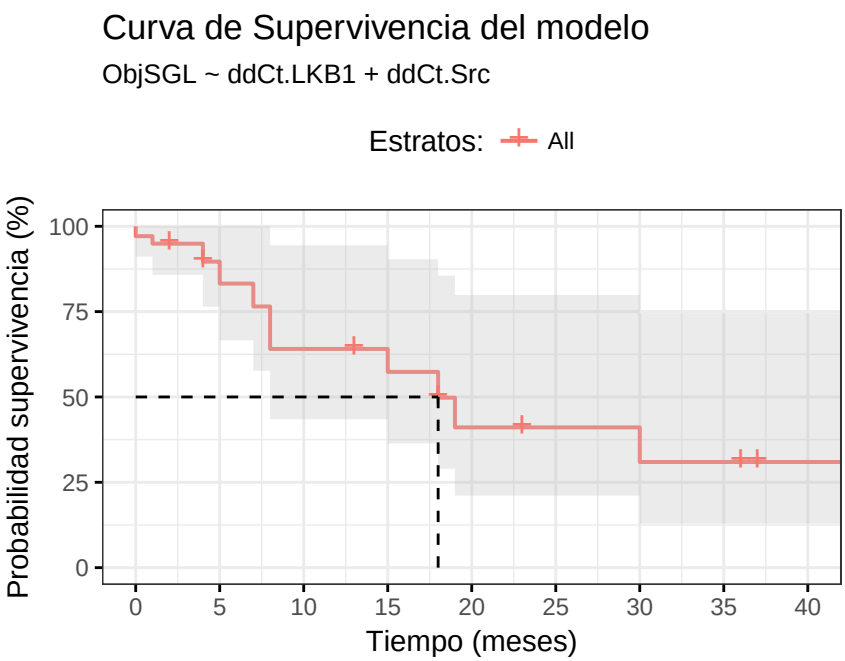
Ver linealidad

Se muestra a continuación la gráfica Gr.29, donde se pueden ver los residuos de martingala del modelo de Cox nulo contra las covariables continuas.



Gráfica Gr.29

En el gráfico Gr.30 se puede visualizar la proporción de supervivencia prevista en base a la expresión de los genes del modelo ajustado, según los valores medios de las covariables.



Gráfica Gr.30

Resultados de análisis de SGL

Los modelos de regresión de Cox Univariable no estratificados han mostrado evidencia significativa al 5% de que el gen *LKB1* tiene evidencia de influencia sobre el tiempo de supervivencia. En el caso de la estratificación por niveles Alto, Medio y Bajo, ha mostrado también evidencia significativa al 5% de que la expresión del gen *LKB1* puede tener influencia en el tiempo de supervivencia, al igual que en el caso del modelo estratificado en el contraste entre expresión Baja-Media vs Alta. En el modelo estratificado por quantiles del percentil >= 25, han mostrado significancia al 5% en los genes *HES1*, *CDCP1* y *LKB1*.

En el análisis de los mejores puntos de corte, se han obtenido significancia en los genes *HES1* y *LKB1*, obteniendo las siguientes particiones en las muestras de pacientes:

Table 30: Mejores puntos de corte de HES1, LKB1 y Src

HES1.st	LKB1.st
< 0.34 : 5	< 0.43 : 7
>= 0.34 :14	>= 0.43 :13
NA's :11	NA's :10

En el modelo de regresión de Cox Multivariable se ha ajustado un modelo que indica que las expresiones de los genes *LKB1* y *Scr* presentan evidencia significativa de que tienen influencia en el SLP. El modelo es el indicado en rojo en la tabla 29.

Conclusiones

Como primer aspecto a destacar del proceso de análisis bioestadístico realizado, que surge tras la preparación y exploración de los datos obtenidos del estudio real, es el número reducido de muestras finales disponibles (30) para el estudio, que además no son de forma homogénea completas para cada muestra, con lo que ello comporta para los diferentes análisis bioestadísticos realizados.

El análisis de supervivencia está enfocado a dilucidar si existe evidencia significativa sobre si la expresión de los genes analizados molecularmente son influyentes en la Supervivencia Libre de progresión (*SLP*) o en la Supervivencia General (*SGL*) de los pacientes, teniendo en cuenta el factor de que no siempre se dispone de la información de expresión de todos los genes. Aunque en los procesos de análisis reales siempre este factor es complejo, hay que considerar que los resultados obtenidos serían más precisos si el número de muestras fuera algo mayor y/o más homogéneo en cuanto a los datos obtenidos de cada paciente.

Por otra parte, también hay que mencionar que a diferencia de lo que sería un estudio teórico convencional de Supervivencia, en el casos del análisis de supervivencia relacionados con la progresión del cáncer, los periodos de análisis siempre estarán basados entre fechas en los que los pacientes han sido tratados, que evidentemente pueden variar entre las muestras, no siendo entre fechas fijas determinadas en el tiempo. Por ello, el periodo de análisis establecido para este estudio real de pacientes, no se ha fijado en un espacio temporal de inicio concreto, sino que se marca como inicio del tiempo de análisis, la fecha de inicio del tratamiento de primera línea que se administra al paciente, por lo que abarca diferentes años naturales. Este aspecto ha generado una dificultad añadida en el proceso del cálculo de los tiempos (time) de supervivencia, dado que la información sobre las características de las diferentes variables (fechas clave, progresión de la enfermedad, etc.) presentaba algunas dudas que se han tenido que resolver directamente del responsable del estudio original.

En la muestra analizada, el *73.33%* de los pacientes eran hombres y el *26.67%* eran mujeres, dato que como se ha podido comprobar difiere de las estadísticas poblacionales analizadas sobre este tipo de cáncer en el capítulo 3. Posiblemente esta diferencia está relacionada con la *"n"* reducida de la muestra tratada y solo referente a pacientes con la mutación en KRAS. Los hombres muestran una media de consumo de tabaco el doble superior al de las mujeres (Gr.3). No existe mucha diferencia de edades entre sexos, la media de edad de los pacientes está en el rango de entre 54 y 60 años, mujeres y hombres, respectivamente (Gr.2).

En referencia a la histología de los cánceres muestreados, el porcentaje mayor es de *Adenocarcinomas* con un *80%* y las mutaciones más habituales de KRAS han sido las *G12C, la G12V* y la *G12D* (Gr.4 y Gr.5).

Se ha detectado la existencia de correlación significativa (del *0.05* o inferior) entre la expresión de varios genes. Los genes *LKB1* y *CREB* son los que presentan una mayor correlación en su expresión, seguida de los genes *YAP1* y *SRC*, de *YAP1* y *STAT3* y con menor fuerza de correlación *STAT3* y *HES1* (tabla 4 y Gr.10). Esta correlación está relacionada con la significación estadística encontrada en estos genes como factor de influencia en la *SLP* y *SLG*.

En referencia a la estimación de supervivencia (*SLP*) mediante el modelo de Kaplan-Meier, se han tratado *30* muestras con *27* eventos y *3* censuras. La mediana de *SLP* general ha sido de *4.5* meses y la media de *8.2* meses (Gr.11a). En la estimación de *SLP* general estratificada por sexo y tipo de mutación de KRAS, no se han encontrado evidencias de diferencias significativas entre estos estratos.

Se han realizado diferentes análisis de supervivencia mediante el modelo de regresión de Cox. En el modelo univariable sin estratificar, no se han encontrado evidencia significativa de influencia de ninguno de los genes en la supervivencia de los pacientes. En el modelo univariable estratificado, se han encontrado evidencias significativas estadísticas al 5% de que la expresión de los genes *LKB1* y *HES1* son factor influyente en el tiempo de supervivencia, estratificando su expresión en percentil de *25* y en el caso de *LKB1* entre la expresión *Alta* vs *Baja*. Como se puede apreciar en las gráficas Gr.14, y Gr.15a-b, la baja expresión de los genes *LKB1* y *HES1* indican una supervivencia inferior de los pacientes. Adicionalmente, se han analizado los mejores puntos de corte para todos los genes (tabla 16) mediante la función *cutp()* del paquetes *survMisc*, que implementa el modelo de obtención de Contal C y O'Quigley J (1999). A partir de los mejores puntos de corte se ha realizado una nueva estratificación de la expresión de los genes y se ha efectuado nuevos modelos de regresión de Cox con esta estratificación, que han denotado evidencia significativa al 5% en los genes *HES1*, *LKB1* y *Src* (tabla 17). En las gráficas Gr.16a-c se muestran las curvas de supervivencia de los genes significativos y con los mejores puntos de corte, donde se puede apreciar los límites en la expresión de estos genes y su influencia en la supervivencia de los pacientes. Se puede ver que en el caso de la expresión gen *Src*, a diferencia de la de los genes *HES1* y *LKB1*, los niveles altos de expresión tienen un pronostico peor para la supervivencia de los pacientes.

Asumiendo la linealidad de las variables (difícil de validar dado en bajo número de muestras) y verificando la

suposición de riesgos proporcionales, se han realizado varios procesos de ajuste con diferentes modelos de regresión de Cox multivariable. Varios de estos modelos han evidenciado significancia al 5% de varios genes, siendo el más destacado el modelo con los genes *CREB + LKB1 + Scr + STAT3*, donde *CREB*, *LKB1* y *Src* indicaban evidencia estadística significativa al 5% y con un indice de Concordancia del modelo *C =0.7562*, indicando con ello la influencia de estos genes en el tiempo de SLP. Interpretando el modelo, nos indica que con significancia del 5% y a igualdad de valores en el resto de expresiones de los otros genes, el aumento de un punto de expresión del gen *CREB* incrementa un 68% *(HR=1.68)* el riesgo de muerte. En el caso del gen *Scr*, a igualdad de valores de expresión de los otros genes, el aumento de un punto en su expresión, significa un aumento del 18% *(HR=1.18)* en el riesgo de muerte, mientras que en el caso de *LKB1* el aumento de su nivel de expresión, manteniendo la expresión de los otros genes, tendría un efecto de disminución del riesgo de muerte ya que vemos que el valor de su coeficiente $\beta = -7.28$ es negativo, pero dado que su valor de $HR = 0.0007$ es muy cercano a cero, este no tendría una afectación significante en el pronóstico. Por otra parte, se puede comprobar que el intervalo de confianza de la covariable *HEST1* contiene el 1, es decir contiene el valor nulo del modelo, no tiene una contribución significativa en el modelo.

En el apartado de análisis de SLP se pueden ver en los diferentes gráficos y tablas que aportan los datos numéricos y gráficos de los resultados obtenidos. Adicionalmente, se han analizado las posibles interacciones entre las covariables, determinado que no hay evidencia significativa de ninguna interacción entre ellas. Se han efectuado las pruebas de riesgos proporcionales, dando esta que no son significativas.

Si se analiza la estimación de riesgo $HR = exp(coef)$ en el resumen mostrado de cada covariable, se puede concluir con significancia de 0.05 que el modelo indica que a igualdad de valores en el resto de expresiones de los otros genes,

En referencia a la estimación de supervivencia global (*SGL*) mediante el modelo de Kaplan-Meier, se han tratado *30* muestras con *20* eventos y *10* censuras. La mediana de SGL general ha sido de *15 meses* y la media de *18.2 meses* (Gr.21a). En la estimación de SGL general estratificado por sexo y tipo de mutación de KRAS, no se han encontrado evidencias de diferencias significativas para estos estratos.

Al igual que en la SLP, para el análisis de la SGL se han realizado diferentes análisis de ajuste de las covariables, mediante diferentes modelos de regresión de Cox Univariable. Al efectuar el modelo de ajuste Univariable no estratificado, se aprecia evidencia significativa al 5% solo del gen *LKB1* de influencia en el tiempo de supervivencia global de los pacientes. En el modelo Univariable estratificado, hay evidencia significativa al 5% tanto en la estratificación de *Alta* vs *Baja*, como en la estratificación *Baja-Media* vs *Alta*, también en el gen *LKB1*. En la estratificación por cuantiles en el percentil *>=25* o *<25*, aparecen tres genes significativos al 5%, que han sido el *HES1*, *CDCP1* y *LKB1* con con *p-valores* de *0.0349*, *0.0469* y *0.0037*, respectivamente (ver gráficas Gr.24 y Gr.25a-c). Analizando los mejores puntos de corte para el nivel de expresión con la función *cutp()* (tabla 26), se ha efectuado nuevos modelos de regresión de Cox con esta estratificación, que han denotado evidencia significativa al 5% en los genes *HES1* y *LKB1* (tabla 27). En las gráficas Gr.26a-b se muestran las curvas de supervivencia de los genes significativos y con los mejores puntos de corte, donde se puede apreciar los límites en la expresión de estos genes y su influencia en la supervivencia de los pacientes.

Asumiendo igualmente la linealidad de las covariables, se han efectuado varios modelos de regresión de Cox con las diferentes covariables de entrada, donde solo se ha encontrado significancia al 5% a nivel de test de *razón de verosimilitud (Likekihook)* en dos de los modelos ajustados: en el modelo con las variables de entrada *LKB1 + Src + STAT3*, con significancia solo las covariables *LKB1* y *sRC*; en el modelo con las covariable de entrada *LKB1 + Src*, además de dar significancia el test global de *razón de verosimilitud (Likekihook)*, también han sido significativos al 5% los test de Wald y de score. Interpretando el modelo lineal, podemos entender que a igualdad de valores en el resto de expresiones de los otros genes, el aumento de un punto de expresión del gen *Scr*, significa un aumento del 21.4% *(HR=1.214)* en el riesgo de muerte, mientras que en el caso de *LKB1* el aumento de su nivel de expresión, manteniendo la expresión de los otros genes, tendría un efecto de disminución del riesgo de muerte ya que vemos que el valor de su coeficiente $\beta = -7.80$ es negativo, pero dado que su valor de $HR = 0.0004$ es muy cercano a cero, este no tendría una afectación significante en el pronóstico.

Se han efectuado las pruebas de riesgos proporcionales, dando estas que no son significativas.

Después de los diferentes análisis de supervivencia realizados, se puede concluir que la expresión de los genes que presentan evidencia significativa al 5% que tienen influencia en el tiempo de supervivencia de los pacientes con cáncer de pulmón (NSCLC) con mutación en KRAS han sido la relativa a los genes *LKB1*, *Src*, *HES1*, *CREB* y *CDCP1.*

Comparativa de resultados con estudio original

A nivel global se han encontrado algunas diferencias entre los resultados obtenidos en el estudio original y los presentados en este análisis. Como se indicaba en la explicación del estudio original, la reducida cantidad de muestras y el no disponer de forma homogéneamente completa los diferentes datos para todas las muestras, puede haber sido un factor fundamental para que puedan existir algunas discrepancias en los cálculos realizados y por tanto en los resultados obtenidos. El cálculo del tiempo de supervivencia de cada muestra, presentó dificultades por la falta de información clara respecto a los diferentes límites de cálculo tanto para la fecha de inicio de análisis, como para la fecha de progresión (este último debido a la falta de conocimientos médicos sobre este aspecto), que finalmente se pudieron solventar directamente con el responsable del trabajo original.

A continuación se efectúa una revisión de los diferentes resultados obtenidos en ambos trabajos, indicando las divergencias y concordancias encontradas entre ambos análisis.

- En ambos análisis se ha trabajado con 30 muestras, de las cuales se han identificado 27 eventos y 3 censuras para la SLP y 20 eventos y 10 censuras para la SGL. La exploración y revisión de los datos disponibles para el estudio han mostrado una total convergencia en los resultados mostrados en ambos trabajos.
- En referencia al análisis de correlación entre las diferentes expresiones de los genes, han mostrado alguna divergencia de parejas de correlación, aun aplicando el test e Spearman en ambos casos:
- En el estudio original se ha detectado correlación significativa al 5% en 7 parejas, mientras en el presente estudio solo han mostrado correlación significativa en 4 de las 7 parejas del estudio original. Las parejas de *CREB1* y *Src*, *CREB1* y *STAT3*, así como de *Src* y *STAT3*, en el presente análisis (tabla 4 y gráfica Gr.10) no presentaban correlación significante al 5%. La discrepancia en este aspecto, podría venir causada por que en el estudio original se hubieran empleado 2 muestras que se han descartado del estudio, dada la inexistencia de las fechas de análisis necesarias para realizar el estudio de supervivencia, aunque estas dos muestras tenían los valores de expresión de los diferentes genes. En este estudio se han descartado estas dos muestras para todos los análisis realizados. Se desconoce si en el estudio original se han tenido en cuenta para algunos de los análisis realizados.
- En el análisis de Supervivencia Libre de Progresión (SLP) global existe discrepancia en la mediana del tiempo de progresión.
- En el estudio original se indica que la mediana de SLP es de 2.5 meses y no se describe la media.
- En el presente estudio se confirma que la mediana de SLP es de 4.5 meses y la media es de 8.19 meses.
- Del análisis de supervivencia SLP estratificado por género y por tipo de mutación de KRAS no se puede hacer comparativa, dado que es un análisis no realizado en el estudio original. En el presente trabajo se ha podido constatar que no hay evidencia significativa de diferencias entre la supervivencia entre hombres y mujeres, ni en base a las diferentes mutaciones del gen KRAS.
- En el análisis mediante un modelo de regresión de Cox Univariable sin estratos, se obtiene el mismo resultado en ambos trabajos, es decir no hay evidencia significativa de la influencia de la expresión de los genes en el tiempo de supervivencia.
- En el análisis mediante modelos de regresión de Cox Univariables estratificados, se han realizado las mismas agrupaciones de expresión en los genes y la comparativa de los resultados obtenidos es la siguiente:
  - En la estratificación entre niveles de expresión Alta, Media y Baja, en el estudio original no se ha encontrado evidencia significativa de la influencia de la expresión de ninguno de los genes analizados en el tiempo de supervivencia, en cambio, en el presente trabajo se han encontrados evidencia significativa al 5% en la expresión Alta vs Baja del gen *LKB1.*
  - En la estratificación entre el nivel de expresión Alta y Baja-Media, es al contrario, en el estudio

original se encuentra evidencia significativa al 5% en el gen *Src*, mientras que en el presente trabajo no se encuentra en ninguno de los genes analizados con esta agrupación.
  - En la estratificación según percentiles >= 25, en el estudio original se encuentra evidencia significativa al 5% entre la expresión $>= Q25$ vs $< Q25$ en el gen *LKB1*. En el presente trabajo se encuentra adicionalmente al gen *LKB1*, también significancia en el gen *HES1*.
  - En ambos estudios se ha realizado el cálculo de los mejores puntos de corte para el modelo de Cox Univariable de cada expresión génica y ha habido coincidencia en dos de los genes que presentaban evidencia significativa al 5% en la estratificación con el mejor punto de corte, que ha sido en los genes *LKB1* y *Src*, pudiendo verificar en estos casos que los puntos de corte en ambos casos eran los mismos, pero se ha mostrado discrepancia en el resto de resultados. En el estudio original se han identificado los genes CDCP1 y CREB1 con significación estadística al 5% y en cambio en el presente estudio, solo en el gen *HES1*.
- En referencia al análisis con modelo de Cox Multivariable, no se puede realizar comparativa, ya que este análisis no se realizó en el estudio original.
- En el análisis de Supervivencia Global, también existe algo de discrepancia entre los valores de mediana obtenidos en ambos trabajos, aunque menor que la existente en la SLP.
  - En el trabajo original se obtuvo una mediana de 13.4 meses de SLP. La media no se indicaba.
  - En el presente trabajo se obtuvo una mediana de 15 meses y una media de 18.2 meses.
- Del análisis de supervivencia SGL estratificado por género y por tipo de mutación de KRAS no se puede hacer comparativa, dado que es un análisis no realizado en el estudio original. En el presente trabajo se ha podido constatar que tampoco hay evidencia significativa de diferencias entre la supervivencia entre hombres y mujeres, ni en base a las diferentes mutaciones del gen KRAS a nivel global.
- En el análisis mediante un modelo de regresión de Cox Univariable sin estratos, se obtiene el mismo resultado en ambos trabajos, ambos muestran evidencia significativa al 5% de la influencia de la expresión de los genes en el tiempo de supervivencia para el gen *LKB1*.
- En el análisis mediante modelos de regresión de Cox Univariables estratificados, se han realizado solo algunas agrupaciones de expresión equivalentes en los genes y la comparativa de los resultados obtenidos en las estratificaciones equivalentes es la siguiente:
  - En las estratificaciones entre niveles de expresión Alta, Media y Baja y entre Alta, Baja-Media, son equivalentes en ambos trabajos, se han encontrado evidencia significativa al 5% en la expresión Alta vs Baja del gen *LKB1*.
  - Al igual que en la SLP, se ha realizado el cálculo de los mejores puntos de corte para el modelo de Cox Univariable de cada expresión génica y ha habido coincidencia solo en el gen *LKB1*, que ambos estudios han mostrado significancia, pero en el estudio original se encontró significancia, también en la expresión del gen *CDCP1* y en cambio en el presente trabajo se encontró en la expresión del gen *HES1*.
  - En el estudio original no se realizó la estratificación según percentiles >= 25, pero en el presente trabajo se realizó obteniendo evidencia significativa al 5% de influencia de los genes *HES1*, *CDCP* y *LKB1* en el tiempo de supervivencia global.

Como puede observarse, si revisamos las coincidencias en la expresión de los genes que en algún punto han sido significativos al 5% como factores influyentes en la supervivencia de los pacientes, aparecen los genes *LKB1*, *Src* y *CDCP*.

Anexo de codigo R realizado para la generación de las tablas y gráficas

2.2 Proceso de análisis

Obtención de los datos estadísticos

Preparación y exploración de los datos

Preparación de los datos.

```
stopifnot(require(car))
# Preparación de los datos
# Selección de variables de interés
# Se eliminan dos muestras por no disponer de datos suficientes Kras21 y Kras27
dataRaw <- dataRaw[-c(12,16),]
Vinter <- c(1,2,3,4,5,6,7,18,21,22,27,28,29,34,35,36,41,42,43,48,
            49,50,52,56:70,77,71:76,78)
dataKras <- dataRaw[,Vinter]

# Variables de nivel de expresión de genes
ExpGen <- c(31:38)
ExpGen2 <- c(32,33,35,36,37)
# Factorización de variables categóricas
dataKras$Mutacion <- factor(dataKras$Mutacion)
dataKras$Gender <- recode(dataKras$Gender,"0='Mujer';1='Hombre'",
                          as.factor=TRUE)
dataKras$Smoking.historyFac <-
  recode(dataKras$Smoking.history,"0= 'No fumador';1= 'Exfumador';
         2= 'Fumador';3='Sin datos'",as.factor=TRUE)
dataKras$Histology <-
  recode(dataKras$Histology,"0='Otros';1= 'Adenocarcinoma'; 2= 'Escamoso';
         3= 'Adenoescamoso';4= 'Otros';5= 'Otros';6= 'Otros';7= 'Otros';
         8= 'Otros';9= 'Otros';10= 'Otros'",as.factor=TRUE)
#
# Calculo de la edad en la fecha de inicio de tratamiento si no se conoce,
# se calcula a partir de la fecha de diagnóstico
#
d1 <- as.numeric(as.Date(dataKras$DOB, format="%d/%m/%Y"))
d2 <- as.numeric(as.Date(dataRaw$DateCalEdad, format="%d/%m/%Y"))
dataKras$edad <- round((d2-d1)/365.25,2)

# Grabar Excel con la tabla tratada de interés
# write.xlsx2(dataKras,file.path(folder.data,"DatosEstudioKras.xlsx"),
#             sheetName="dataKras", col.names=TRUE, row.names=TRUE, append=FALSE)
```

Exploración y descripción de los datos.

```
# Creamos un data frame con las columnas para que se genere la tabla posteriormente
FD2 <- ""
for (i in 1:length(DescVar[,1])) {
  FD2 <-paste(FD2,paste("|",DescVar[i,"Id"],"|",DescVar[i,"Variable"],"|",
                        DescVar[i,"Descripcion"],"|",
                        class(dataKras[,DescVar[i,"Id"]]),"|",sep=""),"\n",sep="")
}
```

```
kable(data.frame(table(dataKras$Gender),
                 round(table(dataKras$Gender)/
                         sum(table(dataKras$Gender))*100,2))[,c(1,2,4)],
      col.names= c("Género", "Frecuencia","%"),
      align= "rcc", caption = "Distribución de pacientes por género")
```

```
# Gráficos de cajas
# Distribución de edades de los pacientes por género
#
file <- "graf1_c4.png"
```

```
graf1 <- ggplot(dataKras, aes(x=Gender, y=edad, color=Gender))+
  theme_bw() +
  geom_boxplot(outlier.colour="black",
               outlier.shape=16,outlier.size=1, notch=FALSE) +
  stat_summary(fun.y=mean, geom="point", shape=18,size=3, color="red") +
  ylab("Edad") +
  xlab("") +
  ggtitle("Distibución de edades ",
          subtitle = "Por género")

graf1
ggsave(filename = file, path = folder.graf, width = 6, height = 4.5)
```

```
# Gráfico de barras
# Distribución del hábito tabáquico por género
#
file <- "graf2_c4.png"
aux <- data.frame(table(dataKras$Smoking.historyFac,dataKras$Gender))
colnames(aux) <- c("Habito","Genero","Frec")
graf2 <- ggplot(aux, aes(x = Habito, y= Frec, fill=Genero)) +
  geom_bar(stat="identity") + facet_wrap(~ Genero, nrow = 1) +
  geom_text(aes(label = Frec),
          position = position_dodge(width = 0.5), size=3,
          col="black", vjust=0.5, hjust=1.5 ) +
  theme_bw() +
  coord_flip()+
  ylab("Nº de pacientes") +
  theme(axis.text.x = element_text(hjust=0.5, vjust = 0.5, size=7.5))+
  theme(axis.text.y = element_text(hjust=0.5, vjust = 0.25, size=8))+
  ggtitle("Distribución según hábito tabáquico",
          subtitle = "Por género")
graf2
ggsave(filename = file, path = folder.graf, width = 6, height = 4.5)
```

Gr.3

```
# Distribución de consumo de tabaco por género
#
file <- "graf3_c4.png"
graf3 <- ggplot(dataKras, aes(x=Gender, y=PA.YEAR, color=Gender))+
  theme_bw() +
  geom_boxplot(outlier.colour="black",
               outlier.shape=16,outlier.size=1, notch=FALSE) +
  stat_summary(fun.y=mean, geom="point", shape=18,size=3, color="red") +
  ylab("Paquetes año") +
  xlab("") +
  ggtitle("Consumo tabaco por año ",
          subtitle = "Por género")
graf3
ggsave(filename = file, path = folder.graf, width = 6, height = 4.5)
```

```
# Gráficos de barras
# Distribución histología del cáncer por género

file <- "graf4_c4.png"
```

```
graf4 <- ggplot(auxH, aes(x = Histologia, y= Frec, fill=Genero)) +
  geom_bar(stat="identity") + facet_wrap(~ Genero, nrow = 1) +
  geom_text(aes(label = paste(Frec," - ",Por,"%",sep="")),check_overlap = TRUE,
            position = position_stack(vjust = 1),size =2.5) +
        # position = position_dodge(width = 0.5), size=3,
        #   col="black", vjust=0.5, hjust=1.5 ) +
  coord_flip()+
  ylab("Nº de pacientes") +
  theme_bw() +
  theme(axis.text.x = element_text(hjust=0.5, vjust = 0.5, size=7.5))+
  theme(axis.text.y = element_text(hjust=1, vjust = 0.25, size=8))+
  ggtitle("Distribución según histología del cáncer ",
          subtitle = "Por género")
graf4
ggsave(filename = file, path = folder.graf, width = 6, height = 4.5)
```

```
# Gráficos de cajas
# Distribución de consumo de tabaco por género
#

file <- "graf5_c4.png"
graf5 <- ggplot(aux, aes(x = Mutacion, y= Frec, fill=Genero)) +
  geom_bar(stat="identity",position=position_stack()) +
  geom_text(aes(label = paste(Frec," - ",Por,"%",sep="")),check_overlap = TRUE,
          position = position_stack(vjust = 0.5),size =2.5) +
  ylab("Nº de pacientes") +
  xlab("Tipo de mutación KRAS") +
  theme_bw() +
  ggtitle("Distribución según mutación KRAS ",
          subtitle = "Por género")
graf5
ggsave(filename = file, path = folder.graf, width = 6, height = 4.5)
```

```
summary(dataKras[,31:38])
```

```
# Distribución de consumo de tabaco por género
#
# Preparación de los datos para gráfica Boxplot
dataGraf <- melt(dataKras[,c(31:38)])
colnames(dataGraf) <- c("Gen","value")
dataGraf <- data.frame(dataGraf,
                       Vias = recode(dataGraf$Gen,"'ddCt.HES1'='otras vías';
                                     'ddCt.CDCP1'='Vía MAPK/ERK';
                                     'ddCt.AXL'='Vía MAPK/ERK';'ddCt.YAP1'='otras vías';
                                     'ddCt.CREB'='Vía MAPK/ERK';'ddCt.LKB1'='Vía MAPK/ERK';
                                     'ddCt.Src'='Vía MAPK/ERK';'ddCt.STAT3'='Vía MAPK/ERK'",
                                     as.factor=TRUE))
file <- "graf6_c4.png"
graf6 <- ggplot(dataGraf, aes(x=Gen, y=value, fill= Vias))+
  theme_bw() +
  geom_boxplot(outlier.colour="black",
               outlier.shape=16,outlier.size=1, notch=FALSE) +
  stat_summary(fun.y=mean, geom="point", shape=18,size=3, color="red") +
  theme(axis.text.x = element_text(size = 7, angle=90, vjust=0.5, hjust=0.5),
```

```
                    axis.text.y = element_text(size = 7, hjust=0.5),
                    legend.text = element_text(size = 7),
                    legend.position = "bottom") +
  ylab("Nivel de expresión (ddCt)") +
  xlab("Gen de interés") +
  guides(fill = guide_legend(title = "Vías"))+
  ggtitle("Nivel de expresión génica ",
          subtitle = "Valores sin escalar")

graf6
ggsave(filename = file, path = folder.graf, width = 6, height = 4.5)

file <- "graf7_c4.png"
graf7 <- ggplot(dataGraf, aes(x=Gen, y=log2(value), fill= Vias))+
  theme_bw() +
  geom_boxplot(outlier.colour="black",
               outlier.shape=16,outlier.size=1, notch=FALSE) +
  stat_summary(fun.y=mean, geom="point", shape=18,size=3, color="red") +
  theme(axis.text.x = element_text(size = 7, angle=90, vjust=0.5, hjust=0.5),
                    axis.text.y = element_text(size = 7, hjust=0.5),
                    legend.text = element_text(size = 7),
                    legend.position = "bottom") +
  ylab("Nivel de expresión (log2(ddCt))") +
  xlab("Gen de interés") +
  guides(fill = guide_legend(title = "Vías"))+
  ggtitle("Nivel de expresión génica ",
          subtitle = "Valores escalados a log2")

graf7
ggsave(filename = file, path = folder.graf, width = 6, height = 4.5)
```

Análisis de normalidad

```
# Histograma de variables de expresión génica
# Verificación de normalidad
#
color <- colors()
# png(filename = paste(folder.graf,"/","graf8_c4.png",sep=""),width = 960, height = 960)

par(mfrow=c(3,3))
for (i in ExpGen) {
  hist(dataKras[,i], col=color[i*5/2],
       main = paste("Histograma de expresión de \n", substr(names(dataKras)[i],6,10)),
       xlab = "valor ddCt", ylab = "Fracuencia" )
}

# dev.off()
```

```
# QQ Plot de normalidad de variables de expresión génica
# Verificación de normalidad
#
# png(filename = paste(folder.graf,"/",
#                      "graf9_c4.png",sep=""),width = 960, height = 960)
par(mfrow=c(3,3))
```

```
for (i in ExpGen) {
  qqnorm(dataKras[,i], col=color[i*5/2],
       main = paste("qqplot de", substr(names(dataKras)[i],6,10)))
  qqline(dataKras[,i])
}

# dev.off()
```

```
# Test de Normalidad Shapiro
# Verificación de normalidad
#
gen <- substring(names(dataKras)[c(ExpGen)],6,15)
Test.shp <- data.frame(gen, W=NA, p.value=NA)
x <-0
for (i in ExpGen) {
  x <- x+1
  rs <- shapiro.test(dataKras[,i])
  Test.shp[x,2:3] <- c(rs$statistic,rs$p.value)
}
kable(Test.shp[order(Test.shp$p.value),], row.names=FALSE,
      format = "latex", booktabs = T, digits = 5,
      col.names= colnames(Test.shp),
      align= "r",
      caption = "Test de normalidad Shapiro-Wilk de expresiones génicas") %>%
      kable_styling(latex_options = c("striped", "hold_position"))
```

Test de correlación

```
# Función para efectuar el test de correlacón de Spearman
# o Pearson entre varias variables.

rs.test <-function(m, method = "pearson", use = "pairwise") {
  n<-0

  # Preparacion de la matriz que contendrá los p-value
  pval<-matrix(rep(0,ncol(m)^2), nc=ncol(m),nr=ncol(m))
  # Recortamos el nombre de la variable dejando solo el nombre del gen
  colnames(pval) <- rownames(pval) <-substr(colnames(m),6,99)
  # Preparacion de la matriz que contendrá las correlaciones
  r<-matrix(rep(1,ncol(m)^2), nc=ncol(m),nr=ncol(m))
  colnames(r) <- rownames(r) <- substr(colnames(m),6,99)

  for (i in 1:(ncol(m)-1)) {
    for (j in (i+1):ncol(m)) {
      n <- n+1
      res.test <- cor.test(m[,i],m[,j],method=method,use=use)
      # Guardamos resultado del test en las matrices correspondientes
      pval[i,j] <- pval[j,i] <- res.test$p.value
      r[i,j] <- r[j,i] <- res.test$estimate
    }
  }

  return(list("Method"=res.test$method,"r"=r,"pValue"=pval))
}
```

```
# Realización del tests de Spearman entre las expresiones de
# los genes para detectar las relaciones significativas al 0.05
#
test <- rs.test(dataKras[,ExpGen], method="spearman")
# print(test)
```

```
# Preparación de los datos para la tabla
#
aux <- rbind(data.frame(gen=rownames(test$r),r="rho",test$r),
             data.frame(gen=rownames(test$r),
                        r="p-value",test$pValue))[,c(1,2,5,4,7,3,8,9,10,6)]
aux$r <- factor(aux$r)
```

```
# Tabla 4 Datos de test de correlación de Spearman entre expresiones génicas

kable(aux[order(rownames(aux)),], row.names=FALSE, format = "latex",
      booktabs = T, digits = 4,
      col.names= colnames(aux),
      align= "r", caption = "Correlación de Spearman entre expresiones génicas") %>%
      collapse_rows(1) %>%
      kable_styling(font_size = 8) %>%
      kable_styling(latex_options = c("striped", "hold_position"))
      # kable_as_image(paste(folder.graf,sep="","Tabla1"))
```

```
# Gráfica del test de correlación de Spearman
#
# png(filename = paste(folder.graf,"/",
#                      "graf10_c4.png",sep=""),width = 960, height = 960)

corrplot(test$r, method = "circle", p.mat = test$pValue, order="alphabet",
         insig= "label_sig", pch=c("*"),pch.cex=1.3,pch.col = c("red"),
         sig.level=c(0.001,0.01,0.05),
         diag=FALSE,tl.cex=1,tl.col="tomato",tl.srt=45,  win.asp = 0.8)
          title("Test de correlación de Spearman ",
            sub="Significación: * : 0.05, ** : 0.01, *** : < 0.001",outer = FALSE)
# dev.off()
```

Análisis de superivencia

```
# Análisis de supervivencia
# El tiempo de progresión está calculado en meses
#
# Preparación datos para el análsis de supervivencia
Kras.mostra <- dataKras[,c(1,4,47,6,30,31:38,44,45)]
# Recodificación de las mutaciones KRAS, para diferenciar las tres más relevantes
Kras.mostra$Mutacion <-
  recode(Kras.mostra$Mutacion,"'G12A'='Otras';'G12C'= 'G12C'; 'G12D'= 'G12D';
         'G12S'= 'Otras';'G12V'= 'G12V';'G12X'= 'Otras';'G13C'= 'Otras';
         'Q61'= 'Otras'",as.factor=TRUE)

#Creando objeto tipo Surv para SLP y GL
Kras.mostra$ObjSLP <- Surv(dataKras$TimeSLP, dataKras$CensSLP == 1)
Kras.mostra$ObjSGL <- Surv(dataKras$TimeGl, dataKras$CensGl == 1)

# Tabla 5 Datos de la muestra para el Análisis de Supervivencia
```

```
kable(Kras.mostra[,-c(14,15)], row.names=FALSE, format = "latex",
      booktabs = T, digits = 4,
      col.names= colnames(Kras.mostra[,-c(14,15)]),
      align= "r", caption = "Datos de la muestra para el
      Análisis de Supervivencia") %>%
      collapse_rows(1) %>%
      kable_styling(latex_options = c("striped", "scale_down", "hold_position"))
      # kable_as_image(paste(folder.graf,sep="","Tabla1"))
```

Superviviencia Libre de Prograsión (SLP)

Supervivencia Libre de Progresión (SLP)

Estimación de SLP sin agrupaciones

```
# Análisis de supervivencia
# SLP --> supervivencia libre de progresión

# Modelos de ajuste de estimación de la función de supervivencia según metodo Kaplan Meier
# Modelo sin agrupación (como 1)
 Kras.km1 <- survfit(ObjSLP ~ 1, data = Kras.mostra, type = "kaplan-meier")
```

```
# Tabla 6 Datos de Modelo de SLP sin agrupaciones
#
# Resumen del modelop
print(Kras.km1, print.rmean=TRUE)

# Estimación máxima verosimilitud del riesgo acumulado
Kras.km1RA <- Kras.km1 %>% fortify %>%
  mutate(CumHaz = cumsum(n.event/n.risk))

kable(fortify(Kras.km1RA), row.names=FALSE,
      format = "latex", booktabs = T, digits = 4,
      # col.names= colnames(Kras.km1),
      align= "r", caption = "Modelo de estimación SLP Kaplan-Meier") %>%
      kable_styling(font_size = 8) %>%
      kable_styling(latex_options = c("striped", "hold_position"))
      # kable_as_image(paste(folder.graf,sep="","Tabla1"))
```

```
# Gráficas de supervivencial Libre de Progresión SLP
# Gráfica SLP como 1
file <- "graf11a_c4.png"
graf11a <- ggsurvplot(Kras.km1,conf.int=TRUE,surv.median.line="hv",
                      size = 0.7, break.time.by = 5,
                      title= "Curva de Supervivencia SLP", legend.title = "Estratos:",
                      fun = "pct", xlab = "Tiempo (meses)",
                      ylab ="Probabilidad supervivencia (%)",
                      ggtheme=(theme_bw() + theme(axis.text.x =
                                        element_text(hjust=0.5,
                                       vjust = 0.5, colour = "black"))))
graf11a
ggsave(filename = file, path = folder.graf, width = 6, height = 4.5)
```

```
# Gráficas de supervivencial Libre de Progresión SLP
# Gráfica SLP como 1 Riesgos acumulados
file <- "graf11b_c4.png"
```

```
graf11b <- ggsurvplot(Kras.km1,conf.int=TRUE,surv.median.line="hv",
                      size = 0.7, break.time.by = 5,
                      title= "Riesgo acumulado",
                      legend.title = "Estratos:",
                      fun = "cumhaz", xlab = "Tiempo (meses)",
                      ylab ="Riesgo acumulado",
                      ggtheme=(theme_bw() +
                                 theme(axis.text.x = element_text(hjust=0.5,
                                                                  vjust = 0.5, colour = "black"))))
graf11b
ggsave(filename = file, path = folder.graf, width = 6, height = 4.5)
```

Estimación de SLP con agrupación por sexo y por tipo de mutación KRAS

```
# Tabla 7 Datos de Modelo de SLP agrupado por sexo
#
# Modelo estratificado por Genero
Kras.kmSex <- survfit(ObjSLP ~ Gender,
                      data = Kras.mostra, type = "kaplan-meier")
#
# Resumen del modelo
print(Kras.kmSex, print.rmean=TRUE)

kable(fortify(Kras.kmSex)[,c(9,1:8)], row.names=FALSE,
      format = "latex", booktabs = T, digits = 4,
      # col.names= colnames(Kras.km1),
      align= "r", caption = "Modelo de estimación SLP Kaplan-Meier por género") %>%
      collapse_rows(columns = 1) %>%
      kable_styling(font_size = 8) %>%
      kable_styling(latex_options = c("striped", "hold_position"))
      # kable_as_image(paste(folder.graf,sep="","Tabla1"))
```

```
# Gráficas de supervivencial Libre de Progresión SLP
# Gráfica SLP estratificada por sexo

file <- "graf12_c4.png"
graf12 <- ggsurvplot(Kras.kmSex,conf.int=TRUE,surv.median.line="hv",
                      size = 0.7, break.time.by = 5,
                      title= "Curva de Supervivencia SLP por sexo",
                      legend.title = "Estratos:",
                      fun = "pct", xlab = "Tiempo (meses)",
                      ylab ="Probabilidad supervivencia (%)",
                      ggtheme=(theme_bw() + theme(axis.text.x =
                                                  element_text(hjust=0.5,
                                                      vjust = 0.5, colour = "black"))))
graf12
ggsave(filename = file, path = folder.graf, width = 6, height = 4.5)
```

```
# Tabla 7 Datos de Modelo de SLP agrupado por mutación de KRAS
#
# Modelo estratificado por tipo de mutación en KRAS (3 principales y otras)
Kras.kmMut <- survfit(ObjSLP ~ Mutacion, data = Kras.mostra,
                      type = "kaplan-meier")  #Estimación Kaplan Meier
# Resumen del modelo
print(Kras.kmMut, print.rmean=TRUE)
```

```r
kable(fortify(Kras.kmMut)[,c(9,1:8)], row.names=FALSE,
      format = "latex", booktabs = T, digits = 4,
      # col.names= colnames(Kras.km1),
      align= "r", caption = "Modelo de estimación SLP
      Kaplan-Meier por mutación KRAS") %>%
      collapse_rows(columns = 1) %>%
      kable_styling(font_size = 8) %>%
      kable_styling(latex_options = c("striped", "hold_position"))
      # kable_as_image(paste(folder.graf,sep="","Tabla1"))
```

```r
# Gráficas de supervivencial Libre de Progresión SLP
# Gráfica SLP estratificada por tipo de mutación (3 mutaciones principales y otras)

file <- "graf13_c4.png"
graf13 <- ggsurvplot(Kras.kmMut,conf.int=FALSE,surv.median.line="hv",
                     size = 0.7, break.time.by = 5,
                     title= "Curva de Supervivencia SLP por tipo Mutación KRAS",
                     legend.title = "Mutación:",
                     fun = "pct", xlab = "Tiempo (meses)",
                     ylab ="Probabilidad supervivencia (%)",
                     legend.labs=c("G12C","G12D","G12V", "Otras"),
                     ggtheme=(theme_bw() ))
graf13
ggsave(filename = file, path = folder.graf, width = 6, height = 4.5)
```

```r
# Análisis de supervivencia
# SLP --> supervivencia libre de progresión
survdiff(ObjSLP ~ Gender, data = Kras.mostra)
survdiff(ObjSLP ~ Mutacion, data = Kras.mostra)
```

Ajuste de un modelo de regresión de Cox para SLP

Modelo de regresión de Cox Univariable (SLP)

```r
# Función fit.cox() para efectuar modelo de regresión de Cox univariable para cada covariable
#
fit.cox <-function(m, objSur, covars) {
  # Creeación de la formula del modelo
  All_frml <- sapply(covars,
                          function(x) as.formula(paste(objSur,'~', x)))
  # Ejecución del modelo de ajuste
  All_models <- lapply( All_frml, function(x){coxph(x, data = m)})
  # Obtención de datos de interés de los modelos
  All_results <- lapply(All_models,
                          function(x){
                             cp <- cutp(x)
                             cp <- as.numeric(cp[[1]][1,1])
                             x <- summary(x)
                             p.value<-round(x$wald["pvalue"], digits=4)
                             wald.test<-round(x$wald["test"], digits=4)
                             beta<-round(x$coef[1], digits=4);#coeficient beta
                             HR <-round(x$coef[2], digits=4);#exp(beta)
                             HR.confint.lower <- NA
                             HR.confint.upper <- NA
                             if (nrow(x$conf.int)==1) {
```

```
                                HR.confint.lower <- round(x$conf.int[,"lower .95"], 4)
                                HR.confint.upper <- round(x$conf.int[,"upper .95"], 4)
                            }
                            HR <- paste0(HR, " (",
                                         HR.confint.lower, "-", HR.confint.upper, ")")
                            n <- x$n
                            n.event <- x$nevent
                            res<-c(n, n.event,beta, HR, wald.test,p.value,cp)
                            names(res)<-c("n", "n.eventos",
                                          "beta", "HR (95% CI for HR)", "wald.test",
                                          "p.value", "CutP")
                            return(res)
                            #return(exp(cbind(coef(x),confint(x))))
                          })
  # Retorno de un data frame con los resultados del modelo para cada covariable
  res <- t(as.data.frame(All_results, check.names = FALSE))
    return(as.data.frame(res,stringsAsFactors=FALSE))
}
```

```
# Ejecución del modelo de ajuste para las covariables mediante fit.cox()
# Creación de tabla de resultados.
#
# Covariables para los diferentes modelos de ajuste
covariates <- colnames(Kras.mostra[6:13])
# Llamada a la función de regresión de Cox
Kras.CoxUni <- fit.cox(Kras.mostra,"ObjSLP",covariates)
for (i in c(1,2,3,5,6,7)) {
    Kras.CoxUni[i] <- as.numeric(Kras.CoxUni[[i]])
}
# Creación de la tabla de resultados
kable(Kras.CoxUni, row.names=TRUE, format = "latex", booktabs = T, digits = 4,
      # col.names= colnames(Kras.km1),
      align= "r", caption = "Resultados del modelo de regresión de Cox Univariable") %>%
      kable_styling(latex_options = c("striped", "hold_position"))
      # kable_as_image(paste(folder.graf,sep="","Tabla1"))
```

```
# Mejores puntos de corte para cada covariable
newnames <- paste(substring(colnames(Kras.mostra[6:13]),6,15),"st",sep=".")
cps.UV <- data.frame()
for (i in 1:8) {
  cps.UV[1,i] <- Kras.CoxUni[i,"CutP"]
}
colnames(cps.UV) <- newnames

# PARA ELIMINAR *******
cpHES1 <- Kras.CoxUni[1,"CutP"]
cpCDCP1 <- Kras.CoxUni[2,"CutP"]
cpAXL <- Kras.CoxUni[3,"CutP"]
cpYAP1 <- Kras.CoxUni[4,"CutP"]
cpCREB <- Kras.CoxUni[5,"CutP"]
cpLKB1 <- Kras.CoxUni[6,"CutP"]
cpSrc <-  Kras.CoxUni[7,"CutP"]
cpSTAT3 <- Kras.CoxUni[8,"CutP"]
```

Modelo de regresión de Cox Univariable estratificado (SLP).

```
# Creación de estratos
# Calculo de proporciones en expresiones de genes
newnames <- paste(substring(colnames(Kras.mostra[6:13]),6,15),"st",sep=".")
quant33.66 <-
  as.data.frame(lapply(as.matrix(6:13),
                       function(x){quantile(na.omit(Kras.mostra[,x]),c(.33,.66))}))
colnames(quant33.66) <-newnames
quant25 <- as.data.frame(lapply(as.matrix(6:13),
                                function(x){quantile(na.omit(Kras.mostra[,x]),c(.25))}))
colnames(quant25) <- newnames

# Estratificación quantiles <=33, 33:66, >66
Kras.strat33 <- Kras.mostra[,c(1:5,6:13,16)]
Kras.strat33 <- data.frame(Kras.strat33,as.data.frame(apply(matrix(6:13),1,
                           function(x){
                               ifelse (Kras.mostra[,x]<=quant33.66[1,x-5],"Baja",
                                       ifelse((Kras.mostra[,x]>quant33.66[1,x-5]
                                               &Kras.mostra[,x]<=quant33.66[2,x-5]),
                                              "Media","Alta"))
                             })))
colnames(Kras.strat33) <- c(colnames(Kras.strat33[1:14]),newnames)
for (i in 7:14) {Kras.strat33[,x] <- factor(Kras.strat33[,x])}
Kras.strat33 <- data.frame(Kras.strat33,ObjSGL=Kras.mostra[,17])

# Estratificación quantiles <=66, >66
Kras.strat3366 <- Kras.mostra[,c(1:5,6:13,16)]
Kras.strat3366 <- data.frame(Kras.strat3366,as.data.frame(apply(matrix(6:13),1,
                           function(x){
                               ifelse (Kras.mostra[,x]<=quant33.66[2,x-5],"Baja-Media","Alta")
                             })))
colnames(Kras.strat3366) <- c(colnames(Kras.strat3366[1:14]),newnames)
for (i in 7:14) {Kras.strat3366[,x] <- factor(Kras.strat3366[,x])}
Kras.strat3366 <- data.frame(Kras.strat3366,ObjSGL=Kras.mostra[,17])

# Estratificación quantiles <25, >=25
Kras.stratQ25 <- Kras.mostra[,c(1:5,6:13,16)]
Kras.stratQ25 <- data.frame(Kras.stratQ25,as.data.frame(apply(matrix(6:13),1,
                           function(x){
                               ifelse (Kras.mostra[,x]<quant25[1,x-5],"<Q25",">=Q25")
                             })))
colnames(Kras.stratQ25) <- c(colnames(Kras.stratQ25[1:14]),newnames)
for (i in 7:14) {Kras.stratQ25[,x] <- factor(Kras.stratQ25[,x])}
Kras.stratQ25 <- data.frame(Kras.stratQ25,ObjSGL=Kras.mostra[,17])
```

```
# Función fit.cox() para efectuar modelo de regresión de Cox univariable para cada covariable
#
fit.coxSt <-function(m, objSur, covars) {
  # Creeación de la formula del modelo
  All_frml <- sapply(covars,
                           function(x) as.formula(paste(objSur,'~', x)))
  # Ejecución del modelo de ajuste
  All_models <- lapply( All_frml, function(x){coxph(x, data = m)})
  # Obtención de datos de interés de los modelos
```

```
  All_results <- lapply(All_models,
                        function(x){
                           x <- summary(x)
                           p.valuew<-round(x$wald["pvalue"], digits=4)
                           wald.test<-round(x$wald["test"], digits=4)
                           n <- x$n
                           n.event <- x$nevent
                           iter <- length(x$coef[,1])
                           res <- matrix(NA,iter,8)
                           # names(res)<-c("n", "n.eventos", "beta", "HR (95% CI for HR)"
                           #  "wald.test", "p.value")
                           rownames(res) <- rownames(x$coefficients)
                           for (i in 1:iter){
                                 beta<-round(x$coef[i,1], digits=4);#coeficient beta
                                 p.value <- round(x$coef[i,5], digits=4);#p-value
                                 HR <-round(x$coef[i,2], digits=4);#exp(beta)
                                 HR.confint.lower <- round(x$conf.int[i,"lower .95"], 4)
                                 HR.confint.upper <- round(x$conf.int[i,"upper .95"], 4)
                                 HR <- paste0(HR, " (",
                                          HR.confint.lower, "-", HR.confint.upper, ")")
                                 strat <- rownames(x$coef)[i]
                                 res[i,1:8] <- c(strat,n, n.event,beta, HR,
                                                 p.value, wald.test, p.valuew)

                           }
                           return(res)

                          })
  # Retorno de un data frame con los resultados del modelo para cada covariable
  res <- rbind(All_results[[1]])
  for (i in 2:length(All_results)) {res <- rbind(res,All_results[[i]])}
  colnames(res) <- c("strat","n", "n.eventos", "beta",
                     "HR (95% CI for HR)","Pr(>|z|)", "wald.test","p.value")
  return(res)
}
```

Modelos de regresión de Cox con estratificación según expresión Alta, Media y Baja

```
# Covariable estratificada según percientil 33 y 66

kable(summary(Kras.strat33[15:22]), row.names=FALSE,
      format = "latex", booktabs = T, digits = 4,
      align= c("r","r","r","r","r","r","r","r"),
      caption = " Estratificación: Alta, Media y Baja") %>%
      kable_styling(font_size = 8,latex_options = c( "hold_position","striped"))
```

```
# Ejecución del modelo de ajuste  estratificdo para las covariables mediante fit.cox()
# Estratificación percentiles 33 y 66
#
# Covariables para los diferentes modelos de ajuste
covarstrat <- colnames(Kras.strat33[15:22])
# Llamada a la función de regresión de Cox
Kras.CoxUniSt <- fit.coxSt(Kras.strat33,"ObjSLP",covarstrat)
# Añadir columna con covariables
```

```
Kras.CoxUniSt <- cbind(Variable=as.character(apply(matrix(covarstrat),1,rep,2)),Kras.CoxUniSt)

# Creación de la tabla de resultados
kable(Kras.CoxUniSt[,c(1,3,4,8,9,2,5,6,7)], row.names=FALSE,
      format = "latex", booktabs = T, digits = 4,
      # col.names= colnames(Kras.km1),
      align= c("l","c","c","r","r","l","r","l","r"),
      caption = "Modelo de Cox Univariable Estratificado (Alta, Media y Baja)") %>%
      collapse_rows(columns = c(1:5)) %>%
      row_spec(11, bold = F, color = "red") %>%
      kable_styling(font_size = 8,latex_options = c("striped", "hold_position"))
      # kable_as_image(paste(folder.graf,sep="","Tabla1"))

# Ajuste del modelo significativo
vexpl <- "LKB1.st"
fmla <- as.formula(paste("ObjSLP ~", paste(vexpl, collapse= "+")))
Kras.CoxBestSt <- coxph(fmla, data = Kras.strat33)
```

Modelos de regresión de Cox con estratificación según expresión Alta, Baja-Media

```
# Covariable estratificada según percientil 66

kable(summary(Kras.strat3366[15:22]), row.names=FALSE,
      format = "latex", booktabs = T, digits = 4,
      align= c("r","r","r","r","r","r","r","r"),
      caption = " Estratificación: Alta y Baja-Media") %>%
      kable_styling(font_size = 7,latex_options = c("striped", "hold_position"))
```

```
# Ejecución del modelo de ajuste  estratificdo para las covariables mediante fit.cox()
# Estratificación percentiles 66
#
# Covariables para los diferentes modelos de ajuste
covarstrat <- colnames(Kras.strat3366[15:22])
# Llamada a la función de regresión de Cox
Kras.CoxUniSt66 <- fit.coxSt(Kras.strat3366,"ObjSLP",covarstrat)
# Añadir columna con covariables
Kras.CoxUniSt66 <- cbind(Variable=as.character(apply(matrix(covarstrat),1,rep,1)),Kras.CoxUniSt66)

# Creación de la tabla de resultados
kable(Kras.CoxUniSt66[,c(1,3,4,8,9,2,5,6)], row.names=FALSE,
      format = "latex", booktabs = T, digits = 4,
      align= c("l","c","c","r","r","r","r","r"),
      caption = "Modelo de Cox Univariable Estratificado (Alta y Baja-Media)") %>%
      kable_styling(font_size = 8,latex_options = c("striped", "hold_position"))
      # kable_as_image(paste(folder.graf,sep="","Tabla1"))
```

Modelos de regresión de Cox con estratificación según expresión $< Q25$ ***y*** $>= Q25$

```
# Covariable estratificada según percientil Q25

kable(summary(Kras.stratQ25[15:22]), row.names=FALSE,
      format = "latex", booktabs = T, digits = 4,
      align= c("r","r","r","r","r","r","r","r"),
      caption = " Estratificación: <Q25 y >=Q25") %>%
      kable_styling(font_size = 8,latex_options = c("striped", "hold_position"))
```

```
      # kable_as_image(paste(folder.graf,sep="","Tabla1"))
```

```
# Ejecución del modelo de ajuste  estratificdo para las covariables mediante fit.cox()
# Estratificación percentiles cuantil 25
#
# Covariables para los diferentes modelos de ajuste
covarstrat <- colnames(Kras.stratQ25[15:22])
# Llamada a la función de regresión de Cox
Kras.CoxUniStQ25 <- fit.coxSt(Kras.stratQ25,"ObjSLP",covarstrat)
# Añadir columna con covariables
Kras.CoxUniStQ25 <- cbind(Variable=as.character(
  apply(matrix(covarstrat),1,rep,1)),Kras.CoxUniStQ25)

# Creación de la tabla de resultados
kable(Kras.CoxUniStQ25[,c(1,3,4,8,9,2,5,6)], row.names=FALSE,
      format = "latex", booktabs = T, digits = 4,
      align= c("l","c","c","r","r","r","r","r"),
      caption = "Modelo de Cox Univariable Estratificado (<Q25 y >=Q25)") %>%
      row_spec(c(1,6), bold = F, color = "red") %>%
      kable_styling(font_size = 8,latex_options = c("striped", "hold_position"))
      # kable_as_image(paste(folder.graf,sep="","Tabla1"))

# Ajuste de los modelo significativos
vexpl <- "HES1.st"
fmla <- as.formula(paste("ObjSLP ~", paste(vexpl, collapse= "+")))
Kras.CoxBestQ25_1 <- coxph(fmla, data = Kras.stratQ25)
vexpl <- "LKB1.st"
fmla <- as.formula(paste("ObjSLP ~", paste(vexpl, collapse= "+")))
Kras.CoxBestQ25_2 <- coxph(fmla, data = Kras.stratQ25)
```

Modelos de regresión de Cox con estratificación según mejores puntos de corte

```
# Estratificación según Best cut points
#
Kras.stratBCP <- Kras.mostra[,c(1:5,6:13,16)]
Kras.stratBCP <- data.frame(Kras.stratBCP,as.data.frame(apply(matrix(6:13),1,
                          function(x){
                              ifelse (Kras.mostra[,x]<cps.UV[1,x-5],
                                      paste("<",cps.UV[1,x-5],SEP=""),
                                      paste(">=",cps.UV[1,x-5],SEP=""))
                            })))
colnames(Kras.stratBCP) <- c(colnames(Kras.stratBCP[1:14]),newnames)
for (i in 7:14) {Kras.stratBCP[,x] <- factor(Kras.stratBCP[,x])}
Kras.stratBCP <- data.frame(Kras.stratBCP,ObjSGL=Kras.mostra[,17])
```

```
# Covariable estratificada según percientil Q25

kable(summary(Kras.stratBCP[15:22]), row.names=FALSE,
      format = "latex", booktabs = T, digits = 4,
      align= c("r","r","r","r","r","r","r","r"),
      caption = " Estratificación: Mejores puntos de corte") %>%
      kable_styling(font_size = 7.5,latex_options = c("striped", "hold_position"))
      # kable_as_image(paste(folder.graf,sep="","Tabla1"))
```

```
# Ejecución del modelo de ajuste  estratificdo para las covariables mediante fit.cox()
# Estratificación segun mejores puntos de corte
#
# Covariables para los diferentes modelos de ajuste
covarstrat <- colnames(Kras.stratBCP[15:22])
# Llamada a la función de regresión de Cox
Kras.CoxUniStBCP <- fit.coxSt(Kras.stratBCP,"ObjSLP",covarstrat)
# Añadir columna con covariables
Kras.CoxUniStBCP <- cbind(Variable=as.character(
  apply(matrix(covarstrat),1,rep,1)),Kras.CoxUniStBCP)

# Creación de la tabla de resultados
kable(Kras.CoxUniStBCP[,c(1,3,4,8,9,2,5,6)], row.names=FALSE,
      format = "latex", booktabs = T, digits = 4,
      align= c("l","c","c","r","r","r","r","r"),
      caption = "Modelo de Cox Univariable Estratificado mejores puntos de corte") %>%
      row_spec(c(1,6,7), bold = F, color = "red") %>%
      kable_styling(font_size = 8,latex_options = c("striped", "hold_position"))
      # kable_as_image(paste(folder.graf,sep="","Tabla1"))
```

Validación significación de genes

```
# Pairwise survdiff

res <- pairwise_survdiff(ObjSLP ~ LKB1.st,
     data = Kras.strat33)
res
res <- pairwise_survdiff(ObjSLP ~ HES1.st,
     data = Kras.stratQ25)
res
res <- pairwise_survdiff(ObjSLP ~ LKB1.st,
     data = Kras.stratQ25)
res
res <- pairwise_survdiff(ObjSLP ~ HES1.st,
     data = Kras.stratBCP)
res
res <- pairwise_survdiff(ObjSLP ~ LKB1.st,
     data = Kras.stratBCP)
res
res <- pairwise_survdiff(ObjSLP ~ Src.st,
     data = Kras.stratBCP)
res
```

Curvas de supervivencia de los genes con significación según estratificación

```
# Gráficas de supervivencial Libre de Progresión SLP
# Gráfica SLP estratificada según expresión Alta, Baja-Media

Kras.aux <-survfit(ObjSLP ~ LKB1.st, data= Kras.strat33)
file <- "graf14_c4.png"
graf14 <- ggsurvplot(Kras.aux,conf.int=FALSE,surv.median.line="hv",
                     pval = TRUE, size = 0.7, break.time.by = 5,
                     title= "Curva de Supervivencia LKB1 estrificado",
                     subtitle= "Expresión Alta, Media y Baja",
                     legend.title = "Estratos:",
```

```
                      fun = "pct", xlab = "Tiempo (meses)",
                      ylab ="Probabilidad supervivencia (%)",
                      ggtheme=(theme_bw() ))
graf14
ggsave(filename = file, path = folder.graf, width = 6, height = 4.5)
```

```
# Gráficas de supervivencial Libre de Progresión SLP
# Gráfica SLP estratificada según expresión Alta, Baja-Media

Kras.aux <-survfit(ObjSLP ~ HES1.st, data= Kras.stratQ25)
file <- "graf15a_c4.png"
graf15a <- ggsurvplot(Kras.aux,conf.int=FALSE,surv.median.line="hv",
                      pval = TRUE, size = 0.7, break.time.by = 5,
                      title= "Curva de Supervivencia HES1 estrificado",
                      subtitle= "Expresión <Q25 y >=Q25",
                      legend.title = "Estratos:",
                      fun = "pct", xlab = "Tiempo (meses)",
                      ylab ="Probabilidad supervivencia (%)",
                      ggtheme=(theme_bw() ))
graf15a
ggsave(filename = file, path = folder.graf, width = 6, height = 4.5)
```

```
# Gráficas de supervivencial Libre de Progresión SLP
# Gráfica SLP estratificada según expresión <Q25 y >=Q25

Kras.aux <-survfit(ObjSLP ~ LKB1.st, data= Kras.stratQ25)
file <- "graf15b_c4.png"
graf15b <- ggsurvplot(Kras.aux,conf.int=FALSE,surv.median.line="hv",
                      pval = TRUE, size = 0.7, break.time.by = 5,
                      title= "Curva de Supervivencia LKB1 estrificado",
                      subtitle= "Expresión <Q25 y >=Q25",
                      legend.title = "Estratos:",
                      fun = "pct", xlab = "Tiempo (meses)",
                      ylab ="Probabilidad supervivencia (%)",
                      ggtheme=(theme_bw() ))
graf15b
ggsave(filename = file, path = folder.graf, width = 6, height = 4.5)
```

```
# Gráficas de supervivencial Libre de Progresión SLP
# Gráfica SLP estratificada según mejores puntos de corte

Kras.aux <-survfit(ObjSLP ~ HES1.st, data= Kras.stratBCP)
file <- "graf16a_c4.png"
graf16a <- ggsurvplot(Kras.aux,conf.int=FALSE,surv.median.line="hv",
                      pval = TRUE, size = 0.7, break.time.by = 5,
                      title= "Curva de Supervivencia HES1 estrificado",
                      subtitle= "Expresión en mejor punto de corte",
                      legend.title = "Estratos:",
                      fun = "pct", xlab = "Tiempo (meses)",
                      ylab ="Probabilidad supervivencia (%)",
                      ggtheme=(theme_bw() ))
graf16a
ggsave(filename = file, path = folder.graf, width = 6, height = 4.5)
```

```r
# Análisis de supervivencia
# Media y mediana de la curva de supervivencia
# Modelos de ajuste de estimación de la función de supervivencia
# según metodo Kaplan Meier
print(survfit(ObjSLP ~ HES1.st, data = Kras.stratBCP,
              type = "kaplan-meier"), print.rmean=TRUE)
```

```r
# Gráficas de supervivencial Libre de Progresión SLP
# Gráfica SLP estratificada según mejores puntos de corte

Kras.aux <-survfit(ObjSLP ~ LKB1.st, data= Kras.stratBCP)
file <- "graf16b_c4.png"
graf16b <- ggsurvplot(Kras.aux,conf.int=FALSE,surv.median.line="hv",
                      pval = TRUE, size = 0.7, break.time.by = 5,
                      title= "Curva de Supervivencia LKB1 estrificado",
                      subtitle= "Expresión en mejor punto de corte",
                      legend.title = "Estratos:",
                      fun = "pct", xlab = "Tiempo (meses)",
                      ylab ="Probabilidad supervivencia (%)",
                      ggtheme=(theme_bw() ))
graf16b
ggsave(filename = file, path = folder.graf, width = 6, height = 4.5)
```

```r
# Análisis de supervivencia
# Media y mediana de la curva de supervivencia
# Modelos de ajuste de estimación de la función de supervivencia
# según metodo Kaplan Meier
print(survfit(ObjSLP ~ LKB1.st, data = Kras.stratBCP,
              type = "kaplan-meier"), print.rmean=TRUE)
```

```r
# Gráficas de supervivencial Libre de Progresión SLP
# Gráfica SLP estratificada según mejores puntos de corte

Kras.aux <-survfit(ObjSLP ~ Src.st, data= Kras.stratBCP)
file <- "graf16c_c4.png"
graf16c <- ggsurvplot(Kras.aux, conf.int=FALSE,
                      pval = TRUE, size = 0.7, break.time.by = 5,
                      title= "Curva de Supervivencia Src estrificado",
                      subtitle= "Expresión en mejor punto de corte",
                      surv.median.line="hv",
                      legend.title = "Estratos:",
                      fun = "pct", xlab = "Tiempo (meses)",
                      ylab ="Probabilidad supervivencia (%)",
                      ggtheme=(theme_bw() ))
graf16c
ggsave(filename = file, path = folder.graf, width = 6, height = 4.5)
```

```r
# Análisis de supervivencia
# Media y mediana de la curva de supervivencia
# Modelos de ajuste de estimación de la función de supervivencia según metodo Kaplan Meier
print(survfit(ObjSLP ~ Src.st, data = Kras.stratBCP, type = "kaplan-meier"), print.rmean=TRUE)
```

Modelo de regresión de Cox Multivariable (SLP)

```r
# Modelo de regresión de Cox Multivariable
# contraste con todos los valores de expresión de genes objetivo
#
# Generación de la formaula con todas la covariables de expresión
sel <- c(7,9,10,11,12,13)
vexpl <- colnames(Kras.mostra[6:13])
# vexpl <- colnames(Kras.mostra[sel])
# Modelo de ajuste
fmla <- as.formula(paste("ObjSLP ~", paste(vexpl, collapse= "+")))

Kras.CoxMul <- coxph(fmla, data = Kras.mostra)
# Kras.CoxMul <- coxph(ObjSLP ~ ddCt.Src, data = Kras.mostra)
summary(Kras.CoxMul)
# cps <- cutp(Kras.CoxMul)
# cpCREB.mu <- as.numeric(cps[[5]][1,1])
# cpLKB1.mu <- as.numeric(cps[[6]][1,1])
# cpSrc.mu <- as.numeric(cps[[7]][1,1])
# cpSTAT3.mu <- as.numeric(cps[[8]][1,1])
```

Mejora del modelo de regresión de Cox Multivariable (SLP)

```r
# Modelo de regresión de Cox Multivariable
# contraste con los valores de expresión de los genes CREB, LKB1, Src y STAT3
#
fit.cox2 <-function(m, objSur, covars) {
  # Creeación de la formula del modelo
  All_frml <- lapply(covars,
                     function(x) as.formula(paste(objSur, "~", paste(x, collapse= "+"))))

  # Ejecución del modelo de ajuste
  All_models <- lapply( All_frml, function(x){coxph(x, data = m)})
  # Obtención de datos de interés de los modelos
  All_results <- lapply(All_models,
                        function(x){
                           x <- summary(x)
                           # print(rownames(x$coef))
                           fmla <- paste(substring(rownames(x$coef),6,15),
                                         collapse = "+")
                           # res <- c(fmla,"","","","","","","","","","")
                           sig <- which(x$coef[,5]<0.05)
                           if (!is.null(sig)) {
                               VarSig <- paste(substring(names(sig),6,15),
                                               collapse = ", ")
                               p.vals <- paste(round(x$coef[sig,5],digits=4),
                                               collapse = ", ")
                               # Solo p-value
                               lt <- round(x$logtest["pvalue"], digits=4)
                               wt <- round(x$wald["pvalue"], digits=4)
                               st <- round(x$sctest["pvalue"], digits=4)
                               beta <- paste(round(x$coef[sig,1], digits=4),
                                             collapse = ", ");#coeficient beta
                               HR <-paste(round(x$coef[sig,2], digits=4),
                                          collapse = ", ");#exp(beta)
                               C <- round(x$concordance["C"], digits = 4);
```

```
                                n <- x$n;
                                n.event <- x$nevent
                                res<-c(fmla, n, n.event, VarSig, beta, HR,
                                       p.vals, C, lt, wt, st)
                            }
                             names(res)<-c("fmla", "n", "n.eventos", "var.sign",
                                           "beta", "HR",
                                           "p.valores", "Conc.", "likelihood",
                                           "wald.test", "score")
                             return(res)
                            })
  # Retorno de un data frame con los resultados del modelo para cada covariable
  res <- t(as.data.frame(All_results, check.names = TRUE))
  rownames(res) <- NULL
  return((res))
}
```

```
# Creación de lista de formulas a generar modelo de Cox
#
selesp <- c(7,9:13) # selección especial de genes
selesp2 <- c(7,10:13) # selección especial de genes
selesp3 <- c(7,10:12) # selección especial de genes
selesp4 <- c(8,10:13) # selección especial de genes
selesp5 <- c(8,9:12) # selección especial de genes
#
covariates <- list(colnames(Kras.mostra[6:13]),colnames(Kras.mostra[7:13]),
              colnames(Kras.mostra[8:13]),colnames(Kras.mostra[9:13]),
              colnames(Kras.mostra[10:13]),colnames(Kras.mostra[10:12]),
              colnames(Kras.mostra[selesp]),colnames(Kras.mostra[selesp2]),
              colnames(Kras.mostra[selesp3]),colnames(Kras.mostra[selesp4])
              ,colnames(Kras.mostra[selesp5]))
#
# Llamada a la función de regresión de Cox para multivariable
#
Kras.CoxMul2 <- fit.cox2(Kras.mostra,"ObjSLP",covariates)
# Creación de la tabla de resultados
#
```

```
# Tablas de resultados
#
kable(Kras.CoxMul2[,c(1,4,7)], row.names=FALSE,
      format = "latex", booktabs = T, digits = 4,
      # col.names= c("Modelo", "Var.Significativas", betas","HR", "p.valores"),
      align=c("l","l","r"),col.names= c("Modelo", "Var.Significativas", "p.valores"),
      caption = "Variables significativas de cada modelo de regresión de Cox") %>%
      kable_styling(font_size = 8) %>%
      row_spec(c(4,6), bold = F, color = "blue") %>%
      row_spec(5, bold = F, color = "red") %>%
      kable_styling(latex_options = c("striped", "hold_position"),
                    full_width = FALSE)
      # kable_as_image(paste(folder.graf,sep="","Tabla1"))
kable(Kras.CoxMul2[,c(1,2,3,8:11)], row.names=FALSE, format = "latex",
      booktabs = T, digits = 4,
      col.names= c("Modelo", "n","n.eventos","Conc.","likelih.pval",
```

```
                    "wald.pval", "score.paval"),
      align= c("l","c","c","r","r","r","r"),
      caption = "Significación global de cada modelo Multivariable") %>%
      kable_styling(font_size = 8) %>%
      row_spec(c(4,6), bold = F, color = "blue") %>%
      row_spec(5, bold = F, color = "red") %>%
      kable_styling(latex_options = c("striped", "hold_position"),full_width = FALSE)
```

```
#
# Summary del mejor modelo

# # Modelo de ajuste
fmla <- as.formula(paste("ObjSLP ~", paste(colnames(Kras.mostra[10:13]), collapse= "+")))
# Generación del modelo
Kras.CoxBest <- coxph(fmla, data = Kras.mostra)
# Resumen
print(paste("Formula = ",c(fmla)))
summary(Kras.CoxBest)
```

Evaluación de validez del mejor modelo de regresión de Cox obtenido (SLP)

Verificación de suposición de riesgos proporcionales

```
# Verificación de suposición de riesgos proporcionales
#
# cox.zph() --> Probar el supuesto de riesgos proporcionales de la regresión
Kras.zphbest <- cox.zph(Kras.CoxBest)
Kras.zphbest
```

```
# Verificación de suposición de riesgos proporcionales
#
# Gráfica de riesgos proporcionales
file = "graf17_c4.png"
graf17 <- ggcoxzph(Kras.zphbest,
                   font.main = 7, ggtheme=(theme_bw() +
                             theme(axis.text.x =
                                     element_text(hjust=0.5,  vjust = 0.5,
                                                  size= 6.5, colour = "black"),
                                   axis.text.y = element_text(hjust=0.5, vjust = 0.5,
                                                  size= 6.5, colour = "black"),
                                   axis.title.y = element_text(size = 7),
                                   axis.title.x = element_text(size = 7),
                                   title = element_text(size=6))))
graf17
ggsave(filename = file, path = folder.graf, width = 6, height = 4.5)
```

Prueba de observaciones influyentes (outliers)

```
# Prueba de observciones influyentes - outliers
#
# ggcoxdiagnostics(Kras.CoxMul, type = "martingale",
#                  linear.predictions = FALSE, ggtheme=(theme_bw()))
# ggcoxdiagnostics(Kras.CoxMul, type = "dfbeta",
#                  linear.predictions = FALSE, ggtheme=(theme_bw()))

file = "graf18_c4.png"
```

```
graf18<- ggcoxdiagnostics(Kras.CoxBest, type = "deviance",
                          linear.predictions = FALSE, ggtheme=(theme_bw()))
graf18
ggsave(filename = file, path = folder.graf, width = 6, height = 4.5)
```

Ver linealidad

```
# No linealidad
#
# Generación de la formaula con todas la covariables de expresión
 # sel <- c(8,10,11,12,13)
vexpl <- colnames(Kras.mostra[10:13])
# vexpl <- colnames(Kras.mostra[sel])
fmla <- as.formula(paste("ObjSLP ~", paste(vexpl, collapse= "+")))
#
# Gráfica de ajuste linealidad
#
file = "graf19_c4.png"
graf19 <- ggcoxfunctional(fmla, data = na.omit(Kras.mostra),
                          font.main = 8, point.size=1.3,
                          ggtheme=(theme_bw() +
                                     theme(axis.text.x = element_text(hjust=0.5,
                                                 vjust = 0.5, size= 6.5, colour = "black"),
                                           axis.text.y = element_text(hjust=0.5, vjust = 0.5,
                                                 size= 6.5, colour = "black"),
                                           axis.title.y = element_text(size = 7),
                                           axis.title.x = element_text(size = 7),
                                           title = element_text(size=7))))
graf19
ggsave(filename = file, path = folder.graf, width = 6, height = 4.5)
```

Análisis de interacciones de covariables en el modelo ajustado (SLP)

```
# Modelo de regresión de Cox Multivariable
# contraste con los valores de expresión de los genes CREB, LKB1, Src y STAT3
#
# # Modelo de ajuste con interaciones
fmla <- as.formula(paste("ObjSLP ~", paste(c(paste(colnames(Kras.mostra[10:12]), collapse= " * "),
                                             colnames(Kras.mostra[13])),collapse = " + ")))
# Generación del modelo
Kras.CoxBestI <- coxph(fmla, data = Kras.mostra)
# Resumen
print(paste("Formula = ",c(fmla), sep=""))
Kras.CoxBestI
# summary(Kras.CoxBestI)
```

```
# survfit(Kras.Cox)
Kras.surcox <-survfit(Kras.CoxBest, data= Kras.mostra)
fits <- list(fit.cox= Kras.surcox, Est.KM = Kras.km1 )

file= "graf20_c4.png"
graf20 <- ggsurvplot_combine(fits,data= Kras.mostra, font.main = 8,
                             size = 0.7, break.time.by = 5,
                             conf.int=FALSE,surv.median.line="hv",
                             risk.table = "abs_pct",
```

```
                              title= "Curva de Supervivencia SLP ajustada por modelo de Cox",
                              legend.title = "Estratos:",
                              fun = "pct", xlab = "Tiempo (meses)",
                              ylab ="Probabilidad supervivencia (%)",
                              ggtheme=(theme_bw() ))
graf20
ggsave(filename = file, path = folder.graf, width = 6, height = 4.5)
```

Resultados de análisis de SLP

```
# Covariable estratificada según percientil Q25

kable(summary(Kras.stratBCP[c(15,20,21)]), row.names=FALSE,
      format = "latex", booktabs = T, digits = 4,
      align= c("r","r","r","r","r","r","r","r"),
      caption = "Mejores puntos de corte de HES1, LKB1 y Src") %>%
      kable_styling(font_size = 7.5,latex_options = c("striped", "hold_position"))
      # kable_as_image(paste(folder.graf,sep="","Tabla1"))
```

Superviviencia Global (SGL)

Supervivencia Global (SGL)

Estimación de SGL sin agrupaciones

```
# Análisis de supervivencia
# GL --> supervivencia Global

# Modelos de ajuste de estimación de la función de supervivencia según metodo Kaplan Meier
# Modelo sin agrupación (como 1)
 Kras.km1GL <- survfit(ObjSGL ~ 1, data = Kras.mostra, type = "kaplan-meier")

# Modelo estratificado por Genero
Kras.kmSexGL <- survfit(ObjSGL ~ Gender, data = Kras.mostra, type = "kaplan-meier")

# Modelo estratificado por tipo de mutación en KRAS (3 principales y otras)
Kras.kmMutGL <- survfit(ObjSGL ~ Mutacion, data = Kras.mostra, type = "kaplan-meier")  #Estimación Kapl
```

```
# Tabla 6 Datos de Modelo de SLP sin agrupaciones
#
# Resumen del modelop
print(Kras.km1GL, print.rmean=TRUE)

# Estimación máxima verosimilitud del riesgo acumulado

Kras.km1RAGL <- Kras.km1GL %>% fortify %>% mutate(CumHaz = cumsum(n.event/n.risk))

kable(fortify(Kras.km1RAGL), row.names=FALSE, format = "latex", booktabs = T, digits = 4,
      align= "r", caption = "Modelo de estimación SGL Kaplan-Meier") %>%
      kable_styling(font_size = 8) %>%
      kable_styling(latex_options = c("striped", "hold_position"))
      # kable_as_image(paste(folder.graf,sep="","Tabla1"))
```

```
# Gráficas de supervivencial Libre de Progresión SLP
# Gráfica GL como 1
file <- "graf21a_c4.png"
graf21a <- ggsurvplot(Kras.km1GL,conf.int=TRUE,surv.median.line="hv",
```

```
                      size = 0.7, break.time.by = 5,
                      title= "Curva de Supervivencia GL",
                      legend.title = "Estratos:",
                      fun = "pct", xlab = "Tiempo (meses)",
                      ylab ="Probabilidad supervivencia (%)",
                      ggtheme=(theme_bw() +
                        theme(axis.text.x =
                                element_text(hjust=0.5, vjust = 0.5,
                                             colour = "black"))))
graf21a
ggsave(filename = file, path = folder.graf, width = 6, height = 4.5)
```

```
# Gráficas de supervivencial Libre de Progresión SLP
# Gráfica GL como 1 Riesgo acumulado

file <- "graf21b_c4.png"
graf21b <- ggsurvplot(Kras.km1GL,conf.int=TRUE,surv.median.line="hv",
                      size = 0.7, break.time.by = 5,
                      title= "Riesgo acumulado SGL",
                      legend.title = "Estratos:",
                      fun = "cumhaz", xlab = "Tiempo (meses)",
                      ylab ="Riesgo acumulado",
                      ggtheme=(theme_bw() +
                               theme(axis.text.x =
                                       element_text(hjust=0.5,
                                                    vjust = 0.5, colour = "black"))))
graf21b
ggsave(filename = file, path = folder.graf, width = 6, height = 4.5)
```

Estimación de SGL con agrupación por sexo y por tipo de mutación KRAS

```
# Datos de Modelo de GL agrupado por sexo
#
# Resumen del modelo por sexo
print(Kras.kmSexGL, print.rmean=TRUE)
```

```
# Datos de Modelo de GL agrupado por mutación de KRAS
#
# Resumen del modelo por mutación de KRAS
print(Kras.kmMutGL, print.rmean=TRUE)
```

```
# Gráficas de supervivencial Libre de Progresión SLP
# Gráfica SLP estratificada por sexo

file <- "graf22_c4.png"
graf22 <- ggsurvplot(Kras.kmSexGL,conf.int=TRUE,surv.median.line="hv",
                     size = 0.7, break.time.by = 5,
                     title= "Curva de Supervivencia GL por sexo",
                     legend.title = "Estratos:",
                     fun = "pct", xlab = "Tiempo (meses)",
                     ylab ="Probabilidad supervivencia (%)",
                     ggtheme=(theme_bw() +
                              theme(axis.text.x = element_text(hjust=0.5,
                                                               vjust = 0.5, colour = "black"))))
graf22
```

```
ggsave(filename = file, path = folder.graf, width = 6, height = 4.5)
```

```
# Gráficas de supervivencial Libre de Progresión SLP
# Gráfica SLP estratificada por tipo de mutación (3 mutaciones principales y otras)

file <- "graf23_c4.png"
graf23 <- ggsurvplot(Kras.kmMutGL,conf.int=FALSE,surv.median.line="hv",
                     size = 0.7, break.time.by = 5,
                     title= "Curva de Supervivencia GL",
                     subtitle= "po tipo de Mutación KRAS",
                     legend.title = "Estratos:",
                     fun = "pct", xlab = "Tiempo (meses)",
                     ylab ="Probabilidad supervivencia (%)",
                     ggtheme=(theme_bw() ))
graf23
ggsave(filename = file, path = folder.graf, width = 6, height = 4.5)
```

```
# Análisis de supervivencia
# SLP --> supervivencia libre de progresión
survdiff(ObjSGL ~ Gender, data = Kras.mostra)
survdiff(ObjSGL ~ Mutacion, data = Kras.mostra)
```

Ajuste de un modelo de regresión de Cox para SGL

Modelo de regresión de Cox Univariable (SGL).

```
# Ejecución del modelo de ajuste para las covariables mediante fit.cox()
# Creación de tabla de resultados.
#
# Covariables para los diferentes modelos de ajuste
covariates <- colnames(Kras.mostra[6:13])
# Llamada a la función de regresión de Cox
Kras.CoxUniGL <- fit.cox(Kras.mostra,"ObjSGL",covariates)
# Creación de la tabla de resultados
kable(Kras.CoxUniGL, row.names=TRUE, format = "latex", booktabs = T, digits = 4,
      # col.names= colnames(Kras.km1),
      align= "r", caption = "Resultados del modelo de regresión de Cox Univariable (GL)") %>%
      row_spec(6, bold = F, color = "red") %>%
      kable_styling(font_size = 8) %>%
      kable_styling(latex_options = c("striped", "hold_position"))
      # kable_as_image(paste(folder.graf,sep="","Tabla1"))
```

```
# Mejores puntos de corte para cada covariable
newnames <- paste(substring(colnames(Kras.mostra[6:13]),6,15),"st",sep=".")
cps.UVGL <- data.frame()
for (i in 1:8) {
  cps.UVGL[1,i] <- Kras.CoxUniGL[i,"CutP"]
}
colnames(cps.UVGL) <- newnames
```

Modelo de regresión de Cox Univariable estratificado (SGL)

Modelos de regresión de Cox con estratificación según expresión Alta, Media y Baja

```
# Ejecución del modelo de ajuste  estratificdo para las covariables mediante fit.cox()
# Estratificación percentiles 33 y 66
#
```

```
# Covariables para los diferentes modelos de ajuste
covarstrat <- colnames(Kras.strat33[15:22])
# Llamada a la función de regresión de Cox
Kras.CoxUniStGL <- fit.coxSt(Kras.strat33,"ObjSGL",covarstrat)
# Añadir columna con covariables
Kras.CoxUniStGL <- cbind(Variable=as.character(
  apply(matrix(covarstrat),1,rep,2)),Kras.CoxUniStGL)

# Creación de la tabla de resultados
kable(Kras.CoxUniStGL[,c(1,3,4,8,9,2,5,6,7)], row.names=FALSE,
      format = "latex", booktabs = T, digits = 4,
      # col.names= colnames(Kras.km1),
      align= c("l","c","c","r","r","l","r","l","r"),
      caption = "Resultados del modelo de regresión de
      Cox Univariable Estratificado (SGL)") %>%
      collapse_rows(columns = c(1:5)) %>%
      row_spec(11, bold = F, color = "red") %>%
      kable_styling(font_size = 8,latex_options = c("striped", "hold_position"))
      # kable_as_image(paste(folder.graf,sep="","Tabla1"))

# Ajuste del modelo significativo
vexpl <- "LKB1.st"
fmla <- as.formula(paste("ObjSGL ~", paste(vexpl, collapse= "+")))
Kras.CoxBestSt <- coxph(fmla, data = Kras.strat33)
```

Modelos de regresión de Cox con estratificación según expresión Alta, Baja-Media

```
# Ejecución del modelo de ajuste  estratificdo para las covariables mediante fit.cox()
# Estratificación percentiles 66
#
# Covariables para los diferentes modelos de ajuste
covarstrat <- colnames(Kras.strat3366[15:22])
# Llamada a la función de regresión de Cox
Kras.CoxUniSt66 <- fit.coxSt(Kras.strat3366,"ObjSGL",covarstrat)
# Añadir columna con covariables
Kras.CoxUniSt66 <- cbind(Variable=as.character(
  apply(matrix(covarstrat),1,rep,1)),Kras.CoxUniSt66)

# Creación de la tabla de resultados
kable(Kras.CoxUniSt66[,c(1,3,4,8,9,2,5,6)], row.names=FALSE,
      format = "latex", booktabs = T, digits = 4,
      align= c("l","c","c","r","r","r","r","r"),
      caption = "Resultados del modelo de regresión de Cox
      Univariable Estratificado (SGL)") %>%
      row_spec(c(6), bold = F, color = "red") %>%
      kable_styling(font_size = 8,latex_options = c("striped", "hold_position"))
      # kable_as_image(paste(folder.graf,sep="","Tabla1"))
```

Modelos de regresión de Cox con estratificación según expresión =Q25

```
# Ejecución del modelo de ajuste  estratificdo para las covariables mediante fit.cox()
# Estratificación percentiles cuantil 25
#
# Covariables para los diferentes modelos de ajuste
covarstrat <- colnames(Kras.stratQ25[15:22])
```

```
# Llamada a la función de regresión de Cox
Kras.CoxUniStQ25 <- fit.coxSt(Kras.stratQ25,"ObjSGL",covarstrat)
# Añadir columna con covariables
Kras.CoxUniStQ25 <- cbind(Variable=as.character(
  apply(matrix(covarstrat),1,rep,1)),Kras.CoxUniStQ25)

# Creación de la tabla de resultados
kable(Kras.CoxUniStQ25[,c(1,3,4,8,9,2,5,6)], row.names=FALSE,
      format = "latex", booktabs = T, digits = 4,
      align= c("l","c","c","r","r","r","r","r"),
      caption = "Resultados del modelo de regresión
      de Cox Univariable Estratificado (SGL)") %>%
      row_spec(c(1,2,6), bold = F, color = "red") %>%
      kable_styling(font_size = 8,latex_options = c("striped", "hold_position"))
      # kable_as_image(paste(folder.graf,sep="","Tabla1"))

# Ajuste de los modelo significativos
vexpl <- "HES1.st"
fmla <- as.formula(paste("ObjSGL ~", paste(vexpl, collapse= "+")))
Kras.CoxBestQ25_1 <- coxph(fmla, data = Kras.stratQ25)
vexpl <- "LKB1.st"
fmla <- as.formula(paste("ObjSGL ~", paste(vexpl, collapse= "+")))
Kras.CoxBestQ25_2 <- coxph(fmla, data = Kras.stratQ25)
```

Modelos de regresión de Cox con estratificación según mejores puntos de corte

```
# Estratificación según Best cut points
#
Kras.stratBCP.GL <- Kras.mostra[,c(1:5,6:13,16)]
Kras.stratBCP.GL <- data.frame(Kras.stratBCP.GL,as.data.frame(apply(matrix(6:13),1,
                          function(x){
                              ifelse (Kras.mostra[,x]<cps.UV[1,x-5],
                                      paste("<",cps.UVGL[1,x-5],SEP=""),
                                      paste(">=",cps.UVGL[1,x-5],SEP=""))
                            })))
colnames(Kras.stratBCP.GL) <- c(colnames(Kras.stratBCP.GL[1:14]),newnames)
for (i in 7:14) {Kras.stratBCP.GL[,x] <- factor(Kras.stratBCP.GL[,x])}
Kras.stratBCP.GL <- data.frame(Kras.stratBCP.GL,ObjSGL=Kras.mostra[,17])
```

```
# Covariable estratificada según percientil Q25

kable(summary(Kras.stratBCP.GL[15:22]), row.names=FALSE,
      format = "latex", booktabs = T, digits = 4,
      align= c("r","r","r","r","r","r","r","r"),
      caption = " Estratificación: Mejores puntos de corte") %>%
      kable_styling(font_size = 7.5,latex_options = c("striped", "hold_position"))
      # kable_as_image(paste(folder.graf,sep="","Tabla1"))
```

```
# Ejecución del modelo de ajuste  estratificdo para las covariables mediante fit.cox()
# Estratificación segun mejores puntos de corte
#
# Covariables para los diferentes modelos de ajuste
covarstrat <- colnames(Kras.stratBCP.GL[15:22])
# Llamada a la función de regresión de Cox
Kras.CoxUniStBCP.GL <- fit.coxSt(Kras.stratBCP.GL,"ObjSGL",covarstrat)
```

```
# Añadir columna con covariables
Kras.CoxUniStBCP.GL <- cbind(Variable=as.character(
  apply(matrix(covarstrat),1,rep,1)),Kras.CoxUniStBCP.GL)

# Creación de la tabla de resultados
kable(Kras.CoxUniStBCP.GL[,c(1,3,4,8,9,2,5,6)], row.names=FALSE,
      format = "latex", booktabs = T, digits = 4,
      align= c("l","c","c","r","r","r","r","r"),
      caption = "Modelo de Cox Univariable Estratificado mejores puntos de corte") %>%
      row_spec(c(1,6), bold = F, color = "red") %>%
      kable_styling(font_size = 8,latex_options = c("striped", "hold_position"))
      # kable_as_image(paste(folder.graf,sep="","Tabla1"))
```

Validación significación de genes

```
# Pairwise survdiff

res <- pairwise_survdiff(ObjSGL ~ LKB1.st,
     data = Kras.strat33)
res
res <- pairwise_survdiff(ObjSGL ~ LKB1.st,
     data = Kras.strat3366)
res
res <- pairwise_survdiff(ObjSGL ~ HES1.st,
     data = Kras.stratQ25)
res
res <- pairwise_survdiff(ObjSGL ~ CDCP1.st,
     data = Kras.stratQ25)
res
res <- pairwise_survdiff(ObjSGL ~ LKB1.st,
     data = Kras.stratQ25)
res
res <- pairwise_survdiff(ObjSGL ~ HES1.st,
     data = Kras.stratBCP.GL)
res
res <- pairwise_survdiff(ObjSGL ~ LKB1.st,
     data = Kras.stratBCP.GL)
res
```

```
# Gráficas de supervivencial Libre de Progresión SLP
# Gráfica SLP estratificada según expresión Alta, Baja-Media

Kras.aux <-survfit(ObjSGL ~ LKB1.st, data= Kras.strat33)
file <- "graf24_c4.png"
graf24 <- ggsurvplot(Kras.aux,conf.int=FALSE,surv.median.line="hv",
                     pval = TRUE, size = 0.7, break.time.by = 5,
                     title= "Curva de Supervivencia LKB1 estrificado",
                     subtitle= "Expresión Alta, Media y Baja",
                     fun = "pct", xlab = "Tiempo (meses)",
                     ylab ="Probabilidad supervivencia (%)",
                     ggtheme=(theme_bw() ))
graf24
ggsave(filename = file, path = folder.graf, width = 6, height = 4.5)
```

```r
# Gráficas de supervivencial Libre de Progresión SLP
# Gráfica SLP estratificada según expresión Alta, Baja-Media

Kras.aux <-survfit(ObjSGL ~ HES1.st, data= Kras.stratQ25)
file <- "graf25a_c4.png"
graf25a <- ggsurvplot(Kras.aux,conf.int=FALSE,surv.median.line="hv",
                      pval = TRUE, size = 0.7, break.time.by = 5,
                      title= "Curva de Supervivencia HES1 estrificado",
                      subtitle= "Expresión <Q25 y >=Q25",
                      fun = "pct", xlab = "Tiempo (meses)",
                      ylab ="Probabilidad supervivencia (%)",
                      ggtheme=(theme_bw() ))
graf25a
ggsave(filename = file, path = folder.graf, width = 6, height = 4.5)
```

```r
# Gráficas de supervivencial Libre de Progresión SLP
# Gráfica SLP estratificada según expresión <Q25 y >=Q25

Kras.aux <-survfit(ObjSGL ~ CDCP1.st, data= Kras.stratQ25)
file <- "graf25b_c4.png"
graf25b <- ggsurvplot(Kras.aux,conf.int=FALSE,surv.median.line="hv",
                      pval = TRUE, size = 0.7, break.time.by = 5,
                      title= "Curva de Supervivencia CDCP1 estrificado",
                      subtitle= "Expresión <Q25 y >=Q25",
                      fun = "pct", xlab = "Tiempo (meses)",
                      ylab ="Probabilidad supervivencia (%)",
                      ggtheme=(theme_bw() ))
graf25b
ggsave(filename = file, path = folder.graf, width = 6, height = 4.5)
```

```r
# Gráficas de supervivencial Libre de Progresión SLP
# Gráfica SLP estratificada según expresión <Q25 y >=Q25

Kras.aux <-survfit(ObjSGL ~ LKB1.st, data= Kras.stratQ25)
file <- "graf25c_c4.png"
graf25c <- ggsurvplot(Kras.aux,conf.int=FALSE,surv.median.line="hv",
                      pval = TRUE, size = 0.7, break.time.by = 5,
                      title= "Curva de Supervivencia LKB1 estrificado",
                      subtitle= "Expresión <Q25 y >=Q25",
                      fun = "pct", xlab = "Tiempo (meses)",
                      ylab ="Probabilidad supervivencia (%)",
                      ggtheme=(theme_bw() ))
graf25c
ggsave(filename = file, path = folder.graf, width = 6, height = 4.5)
```

```r
# Gráficas de supervivencial Libre de Progresión SLP
# Gráfica SLP estratificada según mejores puntos de corte

Kras.aux <-survfit(ObjSGL ~ HES1.st, data= Kras.stratBCP.GL)
file <- "graf26a_c4.png"
graf26a <- ggsurvplot(Kras.aux,conf.int=FALSE,surv.median.line="hv",
                      pval = TRUE, size = 0.7, break.time.by = 5,
                      title= "Curva de Supervivencia HES1 estrificado",
                      subtitle= "Expresión en mejor punto de corte",
```

```
                      legend.title = "Estratos:",
                      fun = "pct", xlab = "Tiempo (meses)",
                      ylab ="Probabilidad supervivencia (%)",
                      ggtheme=(theme_bw() ))
graf26a
ggsave(filename = file, path = folder.graf, width = 6, height = 4.5)
```

```
# Análisis de supervivencia
# Media y mediana de la curva de supervivencia
# Modelos de ajuste de estimación de la función de supervivencia
# según metodo Kaplan Meier
print(survfit(ObjSLP ~ HES1.st, data = Kras.stratBCP.GL,
              type = "kaplan-meier"), print.rmean=TRUE)
```

```
# Gráficas de supervivencial Libre de Progresión SLP
# Gráfica SLP estratificada según mejores puntos de corte

Kras.aux <-survfit(ObjSGL ~ LKB1.st, data= Kras.stratBCP.GL)
file <- "graf26b_c4.png"
graf26b <- ggsurvplot(Kras.aux,conf.int=FALSE,surv.median.line="hv",
                      pval = TRUE, size = 0.7, break.time.by = 5,
                      title= "Curva de Supervivencia LKB1 estrificado",
                      subtitle= "Expresión en mejor punto de corte",
                      legend.title = "Estratos:",
                      fun = "pct", xlab = "Tiempo (meses)",
                      ylab ="Probabilidad supervivencia (%)",
                      ggtheme=(theme_bw() ))
graf26b
ggsave(filename = file, path = folder.graf, width = 6, height = 4.5)
```

```
# Análisis de supervivencia
# Media y mediana de la curva de supervivencia
# Modelos de ajuste de estimación de la función de supervivencia según metodo Kaplan Meier
print(survfit(ObjSLP ~ LKB1.st, data = Kras.stratBCP.GL,
              type = "kaplan-meier"), print.rmean=TRUE)
```

Modelo de regresión de Cox Multivariable (SGL)

```
# Modelo de regresión de Cox Multivariable
# contraste con todos los valores de expresión de genes objetivo
#
# Generación de la formaula con todas la covariables de expresión
vexpl <- colnames(Kras.mostra[6:13])
# Modelo de ajuste
fmla <- as.formula(paste("ObjSGL ~", paste(vexpl, collapse= "+")))

Kras.CoxMulGL <- coxph(fmla, data = Kras.mostra)
summary(Kras.CoxMulGL)
# cps <- cutp(Kras.CoxMulGL)
# cpHES1 <- as.numeric(cps[[1]][1,1])
# cpCDCP1 <- as.numeric(cps[[2]][1,1])
# cpLKB1 <- as.numeric(cps[[6]][1,1])
# cpSrc <- as.numeric(cps[[7]][1,1])
```

Mejora del modelo de regresión de Cox Multivariable (SGL)

```
# Creación de lista de formulas a generar modelo de Cox
#
selesp <- c(7,9,10,11,12,13) # selección especial de genes
#
covariates <- list(colnames(Kras.mostra[6:13]),colnames(Kras.mostra[7:13]),
                   colnames(Kras.mostra[selesp]),colnames(Kras.mostra[8:13]),
                   colnames(Kras.mostra[9:13]),colnames(Kras.mostra[10:13]),
                   colnames(Kras.mostra[10:12]),colnames(Kras.mostra[11:13]),
                   colnames(Kras.mostra[11:12]))
#
# Llamada a la función de regresión de Cox para multivariable
#
Kras.CoxMulGL2 <- fit.cox2(Kras.mostra,"ObjSGL",covariates)
# Creación de la tabla de resultados
#
kable(Kras.CoxMulGL2[,c(1,4,7)], row.names=FALSE, format = "latex", booktabs = T, digits = 4,
      # col.names= c("Modelo", "Var.Significativas", "betas","HR", "p.valores"),
      align=c("l","l","r"),col.names= c("Modelo", "Var.Significativas", "p.valores"),
      caption = "Variables significativas de cada modelo de regresión de Cox SGL") %>%
      kable_styling(font_size = 8) %>%
      row_spec(8, bold = F, color = "blue") %>%
      row_spec(9, bold = F, color = "red") %>%
      kable_styling(latex_options = c("striped", "hold_position"),full_width = FALSE)
      # kable_as_image(paste(folder.graf,sep="","Tabla1"))
kable(Kras.CoxMulGL2[,c(1,2,3,8:11)], row.names=FALSE, format = "latex", booktabs = T, digits = 4,
      col.names= c("Modelo", "n","n.eventos","Conc.","likelih.pval","wald.pval", "score.paval"),
      align= c("l","c","c","r","r","r","r"), caption = "Significación global de cada modelo") %>%
      kable_styling(font_size = 8) %>%
      row_spec(8, bold = F, color = "blue") %>%
      row_spec(9, bold = F, color = "red") %>%
      kable_styling(latex_options = c("striped", "hold_position"),full_width = FALSE)
```

```
#
# Summary del mejor modelo

# # Modelo de ajuste
fmla <- as.formula(paste("ObjSGL ~", paste(colnames(Kras.mostra[11:12]), collapse= "+")))
# Generación del modelo
Kras.CoxBestGL <- coxph(fmla, data = Kras.mostra)
# Resumen
print(paste("Formula = ",fmla, sep=""))
summary(Kras.CoxBestGL)
```

Evaluación de validez del mejor modelo de regresión de Cox obtenido (SGL)

Verificación de suposición de riesgos proporcionales

```
# Verificación de suposición de riesgos proporcionales
#
# cox.zph() --> Probar el supuesto de riesgos proporcionales de la regresión
Kras.zphbestGL <- cox.zph(Kras.CoxBestGL)
Kras.zphbestGL
```

```
# Verificación de suposición de riesgos proporcionales
#
```

```
# Gráfica de riesgos proporcionales
file = "graf27_c4.png"
graf27 <- ggcoxzph(Kras.zphbestGL,
                   font.main = 7, ggtheme=(theme_bw() +
                             theme(axis.text.x =
                                     element_text(hjust=0.5,  vjust = 0.5,
                                                  size= 6.5, colour = "black"),
                                   axis.text.y = element_text(hjust=0.5, vjust = 0.5,
                                                  size= 6.5, colour = "black"),
                                   axis.title.y = element_text(size = 7),
                                   axis.title.x = element_text(size = 7),
                                   title = element_text(size=6))))
graf27
ggsave(filename = file, path = folder.graf, width = 6, height = 4.5)
```

Prueba de observaciones influyentes (outliers)

```
# Prueba de observciones influyentes - outliers
#
# ggcoxdiagnostics(Kras.CoxMul, type = "martingale",
#                  linear.predictions = FALSE, ggtheme=(theme_bw()))
# ggcoxdiagnostics(Kras.CoxMul, type = "dfbeta",
#                  linear.predictions = FALSE, ggtheme=(theme_bw()))

file = "graf28_c4.png"
graf28<- ggcoxdiagnostics(Kras.CoxBestGL, type = "deviance",
                          linear.predictions = FALSE, ggtheme=(theme_bw()))
graf28
ggsave(filename = file, path = folder.graf, width = 6, height = 4.5)
```

Ver linealidad

```
# No linealidad
#
# Generación de la formaula con todas la covariables de expresión
 # sel <- c(8,10,11,12,13)
vexpl <- colnames(Kras.mostra[11:12])
# vexpl <- colnames(Kras.mostra[sel])
fmla <- as.formula(paste("ObjSGL ~", paste(vexpl, collapse= "+")))
#
# Gráfica de ajuste linealidad
#
file = "graf29_c4.png"
graf29 <- ggcoxfunctional(fmla, data = na.omit(Kras.mostra),
                          font.main = 8, point.size=1.3,
                          ggtheme=(theme_bw() +
                                     theme(axis.text.x = element_text(hjust=0.5,
                                                  vjust = 0.5, size= 6.5, colour = "black"),
                                           axis.text.y = element_text(hjust=0.5, vjust = 0.5,
                                                  size= 6.5, colour = "black"),
                                           axis.title.y = element_text(size = 7),
                                           axis.title.x = element_text(size = 7),
                                           title = element_text(size=7))))
graf29
ggsave(filename = file, path = folder.graf, width = 6, height = 4.5)
```

```
# Gráficas de supervivencial Libre de Progresión SGL
# Gráfica mejor modelo SGL

file <- "graf30_c4.png"
graf30 <- ggsurvplot(survfit(Kras.CoxBestGL,data = Kras.mostra),conf.int=TRUE,
                     surv.median.line="hv", size = 0.7, break.time.by = 5,
                     title= "Curva de Supervivencia del modelo",
                     subtitle= paste("ObjSGL ~",
                                     paste(colnames(Kras.mostra[11:12]), collapse= " + ")),
                     legend.title = "Estratos:",
                     fun = "pct", xlab = "Tiempo (meses)",
                     ylab ="Probabilidad supervivencia (%)",
                     ggtheme=(theme_bw() ))
graf30
ggsave(filename = file, path = folder.graf, width = 6, height = 4.5)
```

Resultados de análisis de SGL

```
# Covariable estratificada según percientil Q25

kable(summary(Kras.stratBCP.GL[c(15,20)]), row.names=FALSE,
      format = "latex", booktabs = T, digits = 4,
      align= c("r","r","r","r","r","r","r","r","r"),
      caption = "Mejores puntos de corte de HES1, LKB1 y Src") %>%
      kable_styling(font_size = 7.5,latex_options = c("striped", "hold_position"))
      # kable_as_image(paste(folder.graf,sep="","Tabla1"))
```

Referencias

Contal, O'Quigley, C. 1999. "An Application of Changepoint Methods in Studying the Effect of Age on Survival in Breast Cancer." Vol. 30.

Dardis, Chris. 2016. *Miscellaneous Functions for Survival Data'*. https://cran.r-project.org/package=survMiscl.

Dragulescu, Adrian A. 2014. *Advanced Graphics and Image-Processing in R*. https://cran.r-project.org/package=xlsxl.

Fox, John, and Sanford Weisberg. 2011. *An R Companion to Applied Regression*. Second. Thousand Oaks CA: Sage. http://socserv.socsci.mcmaster.ca/jfox/Books/Companion.

Gil Moreno, María de los Llanos. 2018. "Carcinoma de Pulmón No Celula Pequeña, Con Mutación de Kras: Diferenciación Y Caracterización de Subtipos, Así Como Los Diferentes Mecanismos de Resistencia, Para Elaboración de Trataminetos Dirigidos." PhD thesis, uab.

Horikoshi, Masaaki, and Yuan Tang. 2018. *Ggfortify: Data Visualization Tools for Statistical Analysis Results*. https://CRAN.R-project.org/package=ggfortify.

Kassambara, Alboukadel. 2018. *Drawing Survival Curves Using 'Ggplot2'*. https://cran.r-project.org/web/packages/survminer/index.html.

Kuhn, Max. 2018. *Misc Functions for Training and Plotting Classification and Regression Models*. https://cran.r-project.org/package=caretl.

Ooms, Jeroen. 2018. *Read, Write, Format Excel 2007 and Excel 97/2000/Xp/2003 Files*. https://cran.

r-project.org/package=magickl.

Open Access Statistics - R Consortium, FOAS: Foundation for. 2018. "RStudio - Open Source and Enterprise-Ready Professional Software for Rr." https://www.rstudio.com/.

Therneau, Terry M. 2015. *A Package for Survival Analysis in S*. https://CRAN.R-project.org/package=survival.

Warnes, Gregory R. et al. 2015. *Various R Programming Tools for Model Fitting*. https://cran.r-project.org/package=gmodelsl.

Wei, Taiyun, and Viliam Simko. 2017. *R Package "Corrplot": Visualization of a Correlation Matrix*. https://github.com/taiyun/corrplot.

Wickham, Hadley. 2007. "Reshaping Data with the reshape Package." *Journal of Statistical Software* 21 (12): 1–20. http://www.jstatsoft.org/v21/i12/.

———. 2009. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. http://ggplot2.org.

———. 2011. "The Split-Apply-Combine Strategy for Data Analysis." *Journal of Statistical Software* 40 (1): 1–29. http://www.jstatsoft.org/v40/i01/.

Xie, Yihui. 2018. *Knitr: A General-Purpose Package for Dynamic Report Generation in R*. https://yihui.name/knitr/.

Zeileis, Achim, and Torsten Hothorn. 2002. "Diagnostic Checking in Regression Relationships." *R News* 2 (3): 7–10. https://CRAN.R-project.org/doc/Rnews/.

Zhu, Hao. 2018. *Construct Complex Table with 'Kable' and Pipe Syntax*. https://cran.r-project.org/package=kableExtral.