Universidad de Chile

Instituto de Salud Poblacional

Programa de Magíster en Bioestadística

Estudio predictivo de sobrevida en base a una propuesta de genes en pacientes con Glioblastoma

JESSICA VALDEBENITO SILVA

TESIS PARA OPTAR AL GRADO DE MAGÍSTER

EN BIOESTADÍSTICA

Director de tesis:

Tutor: Felipe Medina, M.Sc. (Universidad de Chile, Santiago, Chile)

Co-directores:

Eliseo Eugenin, Ph.D. (UTMB, Galveston, Texas, US)

Krishna Bhat, Ph.D. (M.D. Anderson, Houston, Texas, US)

2019

TABLE OF CONTENTS

Acrónimos y abreviaturas más utilizadas

ADN (molécula) ácido desoxirribonucleico.

AIC (medida de bondad de ajuste) criterio de información akaike, *akaike information criterion.*

BIC (medida de bondad de ajuste) criterio de información bayesiano, *bayesian information criterion*.

CCC (análisis estadístico) coeficiente de correlación cofenética, *cophenetic correlation coefficient*.

CHZ (Función de riesgo), cociente de riesgo acumulado, *cummulative hazard ratio*.

CP (resultado análisis de componentes principales) componentes principales.

FPKM (método de normalización) Fragmentos de transcrito por millón de secuencias mapeadas, *fragments per kilobase of transcript per million*.

RPKM (método de normalización) lecturas por kilobase de transcrito por millón de lecturas mapeadas, *reads per kilo base per million mapped reads*.

G-CIMP (subtipo de tumor) glioma con fenotipo metilador de islas de CpG (citosina-fosfato-guanina), *glioma CpG (*cytosine-phosphate-guanine) *island methylator phenotype.*

GB (tipo de tumor) glioblastoma.

GBSC (subtipo celular) células madre de glioblastoma, *glioblastoma stem cells.*

HR (cociente) cociente de riesgo, *hazard ratio*.

OMS (organismo) Organización Mundial de la Salud.

OOB (muestra de observaciones) Fuera de la bolsa, *out-of-bag*.

RNA (molécula) ácido ribonucleico, *ribonucleic acid.*

ROC *(representación gráfica, curva)* Característica Operativa del Receptor, *Receiver Operating Characteristic.*

RSF (modelo estadístico) bosque aleatorio de sobrevida, *random survival forest.*

TMZ (alquilante) temozolamida, *temozolamide*.

MAR (mecanismo de pérdida de datos) pérdida al azar, *missing at random*.

MCAR (mecanismo de pérdida de datos) pérdida completamente al azar, *missing completely at random.*

MNAR (mecanismo de pérdida de datos) pérdida no al azar, *missing not at random.*

ML (métodos) Machine learning o aprendizaje automático.

mRNA (molécula) RNA mensajero, *messenger RNA.*

RNA-seq (método de biología molecular) secuenciación de RNA.

TCGA (proyecto) El Atlas del Genoma del Cáncer, *The Cancer Genome Atlas.*

TNT (estructura subcelular) nanotubos, *tunneling nanotubes.*

VIH (patógeno) virus de inmunodeficiencia humana.

VIMP (medida de importancia de variables obtenida de RSF) importancia de variables, *variable importance*.

Nombres de los genes de interés para este trabajo

FH Fumarasa o fumarato hidratasa, *fumarate hydratase*

GAP43 Proteína asociada al crecimiento 43*, growth Associated Protein 43.*

GJA1 Proteína alfa 1de la unión Gap (Conexina 43)*, gap junction alpha-1 protein.*

GLUL Glutamato-Amonio ligasa (Glutamina sintetasa)*, glutamate-Amonnia ligase.*

IDH1 Isocitrato deshidrogenasa 1, *isocitrate dehydrogenases 1*.

IDH2 Isocitrato deshidrogenasa 2, *isocitrate dehydrogenases 2*.

JMJD8 *Jumonji Domain-Containing 8.*

KEAP *Kelch-like ECH-associated protein 1.*

NANOG *Nanog Homebox.*

OGDH Oxoglutarato deshidrogenasa, *oxoglutarate Dehydrogenase.*

OLIG2 Factor de transcripción Olig2*, oligodendrocite transcription factor.*

SDHB Succinato deshidrogenasa*, succinate dehydrogenase complex iron sulfur subunit B.*

SLC1A5 *Solute Carrier Family 1 Member 5.*

TTYH1 *Tweety family member 1.*

YWHAG Proteína 14-3-3-gamma*, 14-3-3-gamma protein.*

WWTR1 Dominio WW contenedor del regulador transcripcional 1*, WW domain containing transcription regulator 1 or TAZ.*

RESUMEN

La presente tesis pretende estudiar un conjunto de genes de interés debido a la particular coincidencia de dichos genes en estudios experimentales previos realizados en células madre de glioblastoma en ratones, células madre de glioblastoma aisladas de muestras humanas de gliobastoma y nanotubos (*tunneling nanotubes*, TNT) aislados por microdisección por captura de láser (*laser capture microdisection*) de macrófagos infectados con VIH. Dichos genes son de interés científico debido al rol común que estos cumplen en vías metabólicas y que en este trabajo se postulan como una vía característica para el desarrollo y propagación de dichas enfermedades, cáncer y VIH.

Para su abordaje inicial, se realizó un análisis de sobrevida de dichos genes utilizando datos públicos de transcriptómica de muestras de glioblastoma con datos de acceso público. Conociendo las limitaciones de realizar dichos análisis en ensayos que priorizan la señal de genes sobre-regulados en el macerado del tumor, se pudo comprobar la escasa representación que tienen los genes seleccionados debido a la baja representación de su señal en el tumor completo.

Los genes seleccionados mostraron tener una estructura de agrupación mixta en cuanto a sus funciones biológicas, es decir, no se descarta la participación conjunta de dichos genes en múltiples mecanismos de resistencia al tratamiento a nivel de esta reducida, pero importante subpoblación celular. Su poder predictivo fue muy bajo, lo cual descarta mayor relevancia de estos en la sobrevida del paciente a nivel de información genética de larga-escala. Sin embargo, se espera poder entregar un contexto biológico

adicional en estudios posteriores, que crucen la información de estos pacientes a nivel de las subpoblaciones celulares de interés y no del tumor completo.

ABSTRACT

The present thesis aspires to contribute to the goal of understanding resistance to treatment mechanisms in Glioblastoma (GB). In particular, this study used transcriptomic data of 151 GB patients from the "TCGA-GBM" project in a multidimensional survival method based on Random Forest to analyze the predictive power of 16 genes identified from multiple previous experiments in isolated TNT and HIV infected macrophages. Most of these genes are represented in a small population of cells and cell to cell communication structures present in the tumor, thus we expected a diluted signal of our transcripts. Further biological and methodological specifications to understand and assess these genes have been well documented throughout the thesis. The results obtained through statistical analysis showed that our transcripts did not contribute to the prediction of survival in GB patients, thus we rescue the relevance that took in the model genes related to metabolic routes and TNT mechanisms took. We expect to continue assessing the predictive power and the effect these genes have in GB disease through an approach that allow us to deconvolute the diluted signal of them.

1. INTRODUCCIÓN

El glioblastoma (GB) es un cáncer altamente maligno y con un negativo pronóstico. De acuerdo con la Organización Mundial de la Salud (OMS), el GB corresponde al tumor cerebral más frecuente y tiene una incidencia anual estimada de 2-3 casos por cada 100.000 habitantes.[1,2] El GB es una enfermedad que progresa rápida y agresivamente sobre las estructuras cerebrales, por lo cual la media de sobrevida en estos pacientes alcanza los 3 meses cuando estos no reciben tratamiento y en el mejor de los casos, bajo tratamiento, no suele extenderse más allá de los 2 años.[3] En la actualidad, el único tratamiento aprobado para pacientes con GB corresponde a la extracción quirúrgica del tumor y, posterior radioterapia y temozolamida (TMZ) como coadyuvante. A pesar de la gran cantidad de datos recolectados del análisis de pacientes sanos y con GB, ya sea, de su genética, genómica, proteómica, y metabolismo (resumidas como "ómicas"), nuevos tratamientos o biomarcadores no han sido descritos o aprobados.

Actualmente, la mayoría de los enfoques genéticos, traslacionales, proteómicos y fosfo-proteómicos utilizados para detectar biomarcadores de enfermedad se basan en la utilización de grandes bases de datos que generalmente provienen del análisis molecular del tumor completo. Estos análisis han permitido identificar las asociaciones entre su fenotipo y determinados marcadores moleculares (e.g. mutaciones genéticas, expresión diferencial de genes basada en mRNA o proteínas). Sin embargo, la mayoría de estos análisis no estudian genes sobre o sub regulados, en pequeñas, pero importantes sub poblaciones celulares, como lo son las células madre de GB (*Glioblastoma Stem Cells,* GBSC).

Es decir, la identificación de estos marcadores en poblaciones celulares pequeñas se ve enmascarada debido a su abundancia en poblaciones celulares que existen en una baja proporción comparado con el tumor completo (e.g. poblaciones de células del sistema inmune, células madre).

Este estudio plantea analizar si la expresión de una propuesta única de genes, expresados en GBSC y en HIV, y que codifican transcritos o proteínas involucrados en el desarrollo de células madre, diferenciación, metabolismo, metilación y comunicación celular, basado en el análisis de formación, comunicación y transporte de *tunneling nanotubes* (TNT) en el tumor pueden predecir la sobrevida del paciente. Estos genes fueron seleccionados a partir de diversos estudios realizados por los laboratorios de Winkler, Bhat y Eugenin,[4–7] tanto en reservorios de VIH, GBSC y proteínas aisladas de TNT, tanto en modelos animales como en tejidos y células humanas. En dichos estudios se observó una coincidencia de 24 genes, de los cuales se seleccionarion 16 para nuestro estudio: *“OGDH", "IDH1", "IDH2", "WWTR1", "NANOG", "OLIG2", "TTYH1", "GAP43", "SDHB", "FH", "SLC1A5", "GLUL", "GJA1", "YWHAG”, “KEAP, “JMJD8”*. Estos genes participan en vías biológicas comunes para el desarrollo de GB. Esta notable y estable coincidencia de genes podría representar la población de GBSC aisladas de tumores primarios y sistemas de comunicación celular debido a la asociación biológica que se ha visto tienen con TNT, comunicación celular y dependencia metabólica de glutamato. Esta dependencia adaptativa del tumor les permite generar energía y contribuir a la sobrevida de células tumorales. Se piensa que, debido a esta particular coincidencia encontrada en múltiples análisis y

pacientes, el rol que estos genes cumplen en la resistencia al tratamiento del tumor puede resultar en una mejor comprensión de la sobrevida de los pacientes.

El estudio del poder predictivo de los genes resulta importante para su posterior validación como biomarcadores de sobrevida.[8] Algoritmos predictivos en Machine Learning (ML) poseen importantes ventajas sobre los modelos estadísticos clásicos, ya que generan predicciones más exactas para nuevas observaciones cuando el modelo de probabilidad, del mecanismo generador de los datos, es desconocido. Pese a que las relaciones entre los predictores y la respuesta no siempre son fáciles de interpretar a partir de los algoritmos, la flexibilidad de estos últimos permiten capturar complicadas asociaciones logrando buenas predicciones.

Para efectos de esta tesis sólo se considerarán datos transcriptómicos (ARN-seq) obtenidos de la base de datos pública *The Cancer Genome Atlas (TCGA)*. El poder predictivo del conjunto de genes propuestos será evaluado utilizando el modelo *Random Survival Forest* (RSF), un método "*ensemble*" (conjunto de modelos) que combina la aleatorización utilizada en la construcción de un conjunto de árboles de decisión e incorpora la censura en los datos de sobrevida. Debido a la escasez de trancritos de algunos genes en el tumor completo o limitaciones del instrumentos para detectarlos, los recuentos de lecturas perdidas de ARN-seq se modelarán mediante el método *Adaptative tree Imputation*, un método que permite imputar los datos perdidos en la medida que el bosque es construido por el modelo RSF.

Este trabajo busca responder la siguiente pregunta: ¿La propuesta de genes asociados a diferenciación, transmisión de señales, metilación y metabolismo predicen la sobrevida en pacientes adultos con glioblastoma?

2. MARCO TEÓRICO

2.1 GLIOBLASTOMA

De acuerdo con la clasificación de la Organización Mundial de la Salud el GB o tumor de grado IV corresponde a un tipo de neoplasmas citológicamente malignos, mitóticamente activos y con tendencia a la necrosis.[9] Dado el pobre pronóstico y complejidad de este tipo de tumor, en los últimos años se ha estudiado extensamente tanto en términos genéticos como epigenéticos e histológicos. Esto ha permitido observar una alta heterogeneidad celular inter- e intratumoral que le permite al GB tener un proceder agresivo y resistente al tratamiento.[10]

Considerando la alta heterogeneidad del tumor, estudios de expresión genética en muestras de tumores de pacientes con glioblastoma han contribuido a entregar mejores predicciones respecto a su evolución, vías moleculares y alternativas de tratamiento.[11–14] Estos estudios han permitido definir sub tipos de tumor, basados en información obtenida de la expresión de genes, perfiles de metilación de ADN y el micro-ambiente inmunológico. Estos sub grupos corresponden al proneural, neural, mesénquimal y clásico.[12] El subtipo mesénquimal usualmente ha sido asociado con un peor pronóstico que el resto y su

transición desde un subtipo proneural ha sido sugerida como un mecanismo de resistencia del tumor a la radiación y quimioterapia.[11] Brennan et al [15] logró clasificar 396 tumores de GB en 6 grupos o clusters usando datos de expresión genética y de metilación. Los subtipos genéticos de GB se encontraban enriquecidos en dichos clusters, siendo uno de ellos un nuevo fenotipo denominado G-CIMP (*glioma CpG island methylator phenotype*), el cual se encontró enriquecido principalmente por el subtipo proneural. La reciente identificación de subsets de tumores con este fenotipo han contribuido a la clasificación actual de gliomas de una manera independiente del grado e histología del tumor.[16]

La clasificación de tumores de acuerdo a particulares perfiles genéticos sigue siendo una promesa para estudiar la sobrevida de pacientes con glioblastoma,[17] sin embargo su uso aún debe ser cuidadosamente interpretado en los análisis. Recientemente, el subtipo neural se ha reportado podría corresponder a una contaminación de la muestra de tumor con células no-tumorales.[18] Los subtipos de GB parecieran demostrar plasticidad en su evolución no reteniendo muchas veces su clasificación original en el transcurso del desarrollo del tumor.

En relación a su abordaje terapéutico, el tratamiento de GB con el alquilante TMZ no ha cambiado por décadas y no ha sido posible prolongar la sobrevida de los pacientes debido a su adaptación y resistencia a tratamiento por mecanismos que hasta hoy son desconocidos. Un tipo de línea celular de interés para comprender los mecanismos de iniciación del tumor y resistencia terapéutica, son las células madre. Estas células permiten replicar la ontogenia del tumor, cuando ocurre apoptosis o resección del tumor, en tejidos altamente jerárquicos y homeostáticos, como lo es el cerebro.[19] En este lugar, dichas células

varían tanto en ubicación como en desarrollo, por lo tanto, en tumores cerebrales como el GB, ellas optan por reproducir diferentes jerarquías contribuyendo a la complejidad y proliferación del tumor.

La mayoría de las investigaciones [7,15,20–24] estudian aquellos genes sobre o sub regulados basados en la abundancia de sus transcritos o en productos de proteínas dentro del tumor, que corresponden a células de proliferación. Por lo tanto, en los análisis de expresión diferencial aquellas células asociadas a procesos de proliferación o diferenciación celular suelen verse mayormente representadas debido a su elevado número y esparcimiento del tumor. Por otro lado, otras poblaciones celulares como las GBSC, que se encuentran en una baja proporción dentro del tumor, no suelen verse representadas. Se ha visto bien documentado que dichas células son resistentes a la quimioterapia y radioterapia, lo cual resalta el rol que ellas cumplen en la progresión y recurrencia del cáncer.[19]

La literatura actual sugiere que aquellos genes involucrados en procesos de diferenciación, proliferación celular, comunicación o marcadores de células madre son los más representados en el tumor. En la **Tabla 1**, se pueden ver los marcadores de sobrevida y tratamiento más conocidos y propuestos por la literatura. [25] La mayoría de estos son genes mutados o expresados en la minoría de las células y ninguna de ellas corresponde a una población de células en particular, excepto CD15 y CD133 en células inmune. Si bien estos genes cumplen un rol importante en el desarrollo del tumor, muchos de ellos siguen siendo considerados controversiales como predictores de sobrevida y tratamiento. Por ende, se necesitan más estudios que verifiquen su poder predictivo.

Tabla 1. Marcadores de tipo de tumor para Glioblastoma (GB) y marcadores de Células Madre de Glioblastoma (GBSC).

Marcadores GB	Marcadores GBSC
MGMT	*CD133*
1p/19Q	*CD15*
IDH	*A2B5*
EGFR	*Nestin*
P53	*ALDH1*
PI3K	*ABC*
Rb	
RAF	

El cáncer ha sido estudiado como un proceso evolutivo donde las células o clones mejor adaptados sobreviven y son responsables del crecimiento del tumor. [26,27] La actividad de las GBSC y los mecanismos de comunicación celular se piensa que explican la heterogeneidad y la progresión del tumor.[27] Un nuevo sistema de comunicación observado en GB corresponde a TNTs,[6] los cuales son generados en condiciones patológicas. Esta conexión anatómica entre los astrocitos proporcionan la funcionalidad y resistencia del

tumor.[6] Winkler en dicho estudio reportó aquella información obtenida en GBSC de ratones. Estudios de Eugenin y Bhat han identificado estos procesos en células madre de GB y los han aislado a través de la técnica *laser capture microdisection*.[4,28] Asimismo han estudiado estas estructuras de comunicación celular en reservorios de VIH, en particular, en macrófagos infectados con VIH (**Tabla 2**). La mayoría de estos genes pueden ser clasificados en 4 familias asociados a: diferenciación, nanotubos, metilación y metabolismo.

Los genes que participan en diferenciación exhiben un rol importante en la proliferación celular e invasión a otros tejidos. Por ejemplo, cuando ocurre una transición de tejido epitelial a mesenquimal, la célula adquiere propiedades de células madre y, por ende, logra migrar e invadir otros tejidos. Esta transición se ha visto explicada por un tipo de coactivador transcripcional llamado TAZ con dominio PDZ.[28]

Otro rol importante en las familias antes mencionadas corresponde a los mecanismos involucrados en el crecimiento y resistencia del tumor a tratamiento, los cuales siguen siendo completamente desconocidos. Sin embargo, datos preliminares indican que los TNTs son esenciales para el crecimiento del tumor, invasión y resistencia al tratamiento.[4]

En relación con los genes propuestos asociados al metabolismo, varios de ellos cumplen un rol común en el ciclo de los ácidos tricarboxílicos (CAT), ya sea, como transportadores de glutamina (*ASCT2*), codificador de moléculas que participan en esta ruta (*OGHD*), entre otras funciones. Así también, se puede ver que existen genes que comparten funciones entre familias, como por ejemplo los genes *SDH* y *FH*, quienes

codifican la molécula alfa cetoglutarato que participa del CAT. Todos estos genes son esenciales para la adaptación del tumor al ambiente.

Tabla 2. Propuesta de genes.

Winkler[6]		Eugenin[4,21] & Bhat[22]
Genes	**Fold Change**	**Genes**
1.*GAP43*	27,59	1.*GAP43*
2.*VGF*	20,88	2.*VGF*
3.*TTHY1*	14,78	3.*TTHY1*
4.*KCNF1*	14,08	4.*OLIG2*
5.*OLIG2*	12,79	5.*OLIG1*
6.*CA9*	12,29	6.*CDC20*
7.*OLIG1*	11,53	7.*ASCL1*
8.*H19*	10,79	8.*OLIG1*
9.*FXYD6*	10,07	9.*CDC20*
10.*HESS*	9,51	10.*OLIG1*
11.*NDUFA4L2*	9,36	11.*OGDH*
12.*C18orf51*	9,13	12.*SDH*
13.*TEK*	9,05	13.*FH*
14.*Clorf61*	8,32	14.*IDH*
15.*Clorf115*	8,05	15.*JMJD8*
16.*GRIK1*	8,05	16.*NANOG*

17.*HEPACAM*	8,05	17.*MYH10*
18.*FAM181B*	7,39	18.*MYH5*
19.*CDC20*	7,06	19. *GS/GLUD1*
20.*ASCL1*	6,57	20. *ASCT2*

Los valores referidos en la tabla anterior (**Tabla 2**) corresponden valores proporcionados por los datos de Winkler en GBSC humanas.[6] Como se puede observar, algunos los genes encontrados en los datos de Eugenin y Bhat, usando técnica de *laser capture microdissection* en TNT y en macrófagos infectados en VIH, coincidieron descendentemente, de acuerdo a su *Fold Change*, con los resportados por Winkler. *Fold change* corresponde a una medida que refleja el cambio en el nivel de expresión de un gen en condiciones experimentales, con respecto a una condición control, y usualmente se utiliza en estudios de expresión genética.[29] Este es calculado como el cociente entre la diferencia de los valores de expresión y el valor de expresión de referencia.

2.2 ESTUDIO DE LA HETEROGENEIDAD DEL TUMOR

El GB representa uno de los tipos de cancer genómicamente más caracterizado. El reconocimiento de los sub-tipos de tumores para GB según su perfil de transcripción, genético o epigenético,[19] ha permitido predecir de mejor manera el pronóstico, sobrevida y respuesta a tratamiento de los pacientes.[25] Debido a esto, se han destinado más esfuerzos entorno al estudio de marcadores moleculares de GB. Existen distintos métodos de medición para marcadores moleculares (FISH, IHC, WB, PCR, entre otros), entre ellos está

el uso del secuenciamiento de ARN (RNA-seq). A continuación, se especificará el proceso de secuenciamiento de ARN, utilidad y limitaciones.

2.2.1 DATOS DE EXPRESIÓN GENÉTICA

La información genética de un organismo, como el ser humano, se guarda en el ADN de su genoma y es expresado mediante la transcripción. Este proceso es el primer paso en la expresión genética y consiste en la copia de información contenida en la secuencia de ADN de un gen para producir una molécula de ARN mensajero (mRNA), el cual posteriormente es utilizado en la producción de proteínas. La actividad celular está regulada por la cantidad de mRNA transcrito. El estudio del transcriptoma (i.e. conjunto de todos los transcritos de ARN en una o muchas células) permite cuantificar el cambio en los niveles de expresión de cada transcrito, en un momento determinado, durante el desarrollo de un estadío o condición fisiológica. Su estudio mediante el uso de técnologías de microarreglos (*microarray*) y secuenciamiento (RNA-seq), es esencial para la interpretación del funcionamiento del genoma, las estructuras moleculares de células o tejidos de interés, y por supuesto, para el entendimiento del desarrollo y pronóstico de una enfermedad.[30]

En este trabajo se utilizarán datos de expresión genética. Estos datos pueden ser obtenidos mediante el uso de tecnologías de *microarray* o *RNA-seq*. En RNA-seq, la muestra total o fraccionada del ARN es convertida en ADN complementario (cDNA). Luego, con el uso de adaptadores se obtienen secuencias de cDNA que son leídas por una plataforma en particular (e.g. Illumina). Para medir la abundancia de estos transcritos de ARN, la lectura

del cDNA obtenida por la plataforma luego es alineada con un genoma o transcriptoma de referencia y clasificada en tipos de lecturas. Estos tipos serán los que permitirán generar un perfil de expresión para cada gen.[30]

La proporción de lecturas que coincidan con un determinado transcrito se utiliza para cuantificar el nivel de expresión de un gen. Se sabe que, para un mismo nivel de expresión, transcritos más largos se asocian con una mayor proporción de lecturas, por lo que la normalización de las lecturas es requerida para comparar niveles de expresión. Luego del mapeo de los datos, la normalización se suele realizar utilizando lecturas por kilobase por millón de lecturas mapeadas (*reads per kilobase per million mapped reads,* RPKM) o (*fragments per kilobase per million,* FPKM). Ésta permite tomar en cuenta el efecto del largo del gen dividiendo la suma de los recuentos de los transcritos por el largo de ese gen.[31]

A continuación, y con el fin de cuantificar aquellos genes que fueron mayormente expresados en la muestra, se realiza un análisis estadístico de expresión diferencial. Para esto, se mapea cada valor de expresión para un gen dado de acuerdo con una distribución en particular. A diferencia del *microarray* donde la señal es continua, en el RNA-seq la naturaleza de los datos es discreta, por lo cual las distribuciones que han sido sugeridas por la literatura para su estudio corresponden a las distribuciones de Poisson y Binomial negativa.[32]

Genes con un bajo nivel de expresión suelen ser más difíciles de detectar en el secuenciamiento debido a una pobre amplificación. Este es el caso de muestras del tumor completo (*bulk*), donde existen células de interés y, por ende, genes que pueden estar

subrepresentados. El interés de este estudio consiste en considerar la heterogeneidad de la población de células del tumor completo en la sobrevida del paciente. Esto implica que aquellos genes que se cree cumplen un rol biológico relevante en la génesis del tumor, pueden resultar en una baja proporción de transcrito. Por ende, los transcritos obtenidos de muestras del tumor completo pueden contener valores perdidos para estos genes de interés y pueden ser confundidos como una ausencia de expresión.

2.2.1.1 Normalización por FPKM

Los datos de RNA-seq tienen una naturaleza de discreta y corresponden a datos de conteo de aquellas lecturas que coinciden con un determinado transcrito y permiten cuantificar la expresión de un gen. Sin embargo, es necesario su normalización considerando que tanto el largo como la profundidad de la lectura pueden intervenir en su interpretación. Aquellas secuencias que tengan una mayor profundidad (i.e. mayor número de lecturas únicas de nucléotidos para cada región de la secuencia), van a tener mayor cantidad de lecturas que coincidan con un gen. Mientras que aquellos genes más largos (i.e. mayor cantidad de pares de bases nucleotídicas), van a tener mayor cantidad de lecturas que los mapeen. Una forma de realizar esta normalización, es considerando el método RPKM. Este método se realiza cuando se considera un trozo de secuencia por fragmento, entonces se procede a normalizar de acuerdo a la profundidad y luego de acuerdo al largo del gen.[33]

Puede ocurrir que existan fragmentos con dos trozo secuenciados que mapeen al mismo gen, generando dos lecturas por fragmento. Con el fin de estipular estos casos, se

utiliza el método FPKM, el cual considera este secuenciamiento pareado de los fragmentos a secuenciar. Este método, es muy similar al método RPKM, la única diferencia es que éste considera aquellos fragmentos que tienen dos lecturas, con el fin de que no sean contabilizadas dos veces.

2.2.1.2 Análisis descriptivo para datos de expresión genética

Es importante realizar un estudio previo al modelamiento estadístico de nuestros datos de expresión genética que permitan cuantificar, identificar y describir en su totalidad a aquellos grupos de genes que presentan características similares según los datos. Considerando los objetivos particulares de cada estudio, siempre resulta interesante realizar un análisis de la distribución de los genes de interés utilizando diagramas de caja o histogramas. Asimismo, es apropiado pensar en el uso de otros análisis de tipo multivariado que permitan resumir y visualizar los datos, especialmente si estos son multivariantes. A continuación se mencionarán algunos de los métodos multivariados más utilizados bajo este contexto:

2.2.1.2.1 Análisis de componentes principales

El análisis de componentes principales corresponde a un método de tipo no supervisado, es decir, cuya respuesta no es conocida y su propósito es reducir, agrupar o explorar los patrones que existen en los datos. Esta técnica organiza amplios sets de datos

correlacionados proyectándolos geométricamente en un número representativo de variables denominadas componentes principales (CP). Los CP son ortogonales en el espacio euclídeo y, por ende, no se correlacionan entre sí. Su construcción se realiza considerando combinaciones lineales de los datos originales, tal que cada CP contenga la mayor variabilidad de los datos y sea ortogonal a los CPs anteriores. Su visualización puede incluir gráficos que estipulen dos CP, usualmente se consideran los dos primeros, pues ellos justamente contienen la mayor proporción de variabilidad.

2.2.1.2.2 ANÁLISIS JERÁRQUICO

Otra forma de explorar los datos, es estipulando el uso de otras técnicas multivariadas, que esta vez, buscan particionar el espacio de observaciones en subconjuntos cada vez más homogéneos, como lo hace el análisis jerárquico de grupos o conglomerados (*clusters*). Para ello este método utiliza una métrica o función que describe la distancia entre pares de observaciones o grupos (e.g. distancia euclidiana) y una regla que permite agrupar observaciones o grupos según aquella distancia. Estos algoritmos se dividen en dos paradigmas: aglomerativos (de abajo hacia arriba), los cuales construyen recursivamente pares de clusters hasta llegar a uno común; o de división (de arriba hacia abajo), los cuales comienzan con toda la data formando parte de un gran cluster, el cual comienza recursivamente a dividirse.[34] Los métodos jerárquicos aglomerativos han sido mayormente estudiados. En este tipo aglomerativo, el par de observaciones escogido para fusionarse consiste en aquellos grupos con la menor disimilitud entre ellos. Sean H y G dos clusters distintos, la medida de disimilitud d(H y G) se calculará a partir de las disimilitudes

$d_{ii'}$ calculadas entre los miembros i del cluster G y los miembros i' del clsuster H. Existen distintos criterios para su cálculo, entre ellos están:

- *Single linkage* (enlace simple) o *nearest-neighbor* (vecino más próximo): la disimilitud entre grupos corresponde a la disimilitud de las observaciones más cercanas (menos disímiles), donde cada observación corresponde a un grupo.

- *Complete linkage* (enlace completo) o *furthest-neighbor* (vecino más alejado): toma las disimilitudes entre grupos como los pares de observaciones más alejados (más disímiles), donde una observación corresponde a cada grupo.

- Average linkage: la disimilitud entre grupos corresponde al promedio de distancias entre todos los pares de observaciones, donde cada par está formado por una observación de cada grupo.

Un ejemplo de visualización del subtipo aglomerativo, son los dendrogramas. Es importante mencionar que estos métodos buscan describir la información y no corresponden a medidas resumen de los datos. Esto es, pues los métodos jerárquicos imponen una estructura ordenada, sin importar si ésta existe realmente en los datos y puede ser sensible a pequeños cambios en los datos. Si se quiere evaluar la estructura de los datos de manera fidedigna, es posible realizar otros análisis más elaborados como lo sería un análisis de correlación cofenética (*Cophenetic Correlation Coefficient*, CCC).[34] Este coeficiente corresponde a la correlación entre la disimilitud entre observaciones $d_{ii'}$ (input del algoritmo) y la disimilitud cofenética $d_{ii'}^{coph}$ (output del algoritmo) de los $\frac{N(N-1)}{2}$ pares

posibles de observaciones derivados del dendrograma. Las disimilitudes cofenéticas entre dos observaciones i e i' corresponden a la disimilitud entre grupos, en la cual dichas observaciones fueron agrupadas por primera vez en el mismo cluster.

Sea $\boldsymbol{D}$ la matriz de distancia de acuerdo a d y sea $\boldsymbol{Z}$ la matriz de distancia de acuerdo a d^{coph}. Luego, $\bar{\bar{\boldsymbol{D}}}$ y $\bar{\bar{\boldsymbol{Z}}}$ denotan las medias de $d_{ii'}$ y $d_{ii'}^{coph}$, respectivamente. Luego, CCC queda como:[35]

$$\mathrm{CCC}(\boldsymbol{D},\boldsymbol{Z}) = \mathrm{Cor}(\boldsymbol{D},\boldsymbol{Z}) = \frac{\sum_{i<i'}(D_{ii'}-\bar{\bar{D}})(Z_{ii'}-\bar{\bar{Z}})}{\sqrt{\sum_{i<i'}(D_{ii'}-\bar{\bar{D}})^2 \times \sum_{i<i'}(Z_{ii'}-\bar{\bar{Z}})^2}}$$

Usualmente para $\boldsymbol{D}$ se utiliza la distancia euclidiana como medida de distancia entre las observaciones. La distancia euclidiana corresponde a la distancia en línea recta entre dos puntos en el espacio euclidiano. El espacio euclidiano en $\mathbb{R}^p$, donde p corresponde al número de variables, corresponde a un vector finito-dimensional que contiene un producto interno definido y que induce a una métrica.[36] Luego, la matriz de distancia euclidiana corresponde a una tabla $d_{ii'}$ entre los N puntos en $\mathbb{R}^n$. Diremos que la distancia euclidiana entre los puntos i e i' es el largo del segmento lineal que los contecta, quedando como:[37]

$$d(i,i') = d(i',i) = \sqrt{\sum_{j=1}^{n}(i'_j - i_j)^2}$$

El uso de algoritmos jerárquicos debe justificar el uso del número de grupos o *clusters* a utilizar. El método más utilizado para ello corresponde al método del codo (*elbow method*). Este método se usa para calcular la cantidad óptima de grupos tal que la cantidad total de la varianza intra-grupo (*within-cluster sum of square*) sea minimizada. Su cálculo

consiste básicamente en computar el algoritmo de clustering para diferentes valores de k, siendo éste el número de clusters. Luego, para cada cluster se calcula la varianza intra-grupo. A continuación, se grafica la varianza intra-grupo versus el número de clusters y el lugar donde ocurra un quiebre en la curva, ese corresponderá al valor óptimo de clusters a utilizar.

2.2.1.2.3 ANÁLISIS Y VISUALIZACIÓN DE GRAFOS

Los grafos corresponden a interconexiones entre un conjunto de entidades V (nodos o vértices), donde sus conexiones corresponden a enlaces E (o bordes). Estos grafos pueden ser dirigidos o unidireccionales o no dirigidos o bidireccionales. Además, pueden considerar pesos o no. El tipo de grafo que se quiera utilizar dependerá de los objetivos que se pretendan alcanzar. Una forma de construir estas redes es usando las medidas de similitud o asociación, $\text{sim}(u, v)$ entre cada par de nodos $u, v \epsilon V$. Estas similitudes se pueden usar como pesos de enlaces para un gráfico con pesos wuv = $\text{sim}(u, v)$ o se pueden conectar dos nodos si su similitud es distinta de cero.[38,39]

Dentro de las medidas de conectividad están el diámetro y su densidad, y serán definidas a continuación.

El diámetro (*diameter*) del grafo G, es la longitud máxima del la longitud del camino más corto [38,39]

$$Diámetro(G) = max_{u,v}[\text{dist(u,v)}]$$

Donde dist corresponde a la distancia entre u y v.

Su densidad (*density*) es la proporción del número presente de enlaces dividido en el número posible de enlaces E [38,39]

$$Densidad(G) = \frac{|E|}{\frac{|V|(|V|-1)}{2}}$$

Otras medidas son las medidas de centralidad, entre ellas: grado, cercanía e intermediación, que serán definidas a continuación.

El grado (*degree*) de un nodo v, $C_D(v)$, nos indica qué tan bien está conectado un nodo y es el número de enlaces conectados al nodo v, el cual podemos computar como la suma de la columna v o fila v de **A**, siendo **A** la matriz de adyacencia cuadrada n x n. Esta matriz indica la relación entre dos variables (e.g. correlación) [38,39]

$$C_D(v) = \sum_{i=1}^{|V|} a_{iv} = \sum_{j=1}^{|V|} a_{vj}$$

Donde i corresponde a la fila v y j a la columna v.

La cercanía (*closeness*) nos dice qué tan fácil es alcanzar a otros nodos desde un nodo. Esta medida suele ser de interés cuando el grafo demuestra estar altamente conectado. Se dice de su cercanía $C_C(v)$ que un nodo será central si está cercano a otros nodos. Esta cercanía se puede calcular tomando la inversa de la suma de todas las longitudes de caminos que van de un nodo v a todos los otros nodos [38,39]

$$C_C(v) = \frac{1}{\sum_{i=1}^{|V|} dist(v,i)}$$

La intermediación (*betweenness*) de un nodo v nos dice qué tan bien un nodo conecta a otros nodos. Y se define como la suma de las proporciones del número de caminos más cortos entre todos los pares de nodos n que pasan por el nodo v [38,39]

$$C_B(i) = \sum_{i \neq j \neq k \epsilon v}^{n} \frac{\sigma(i,j|v)}{\sigma(i,j)}$$

Donde $\sigma(i,j)$ corresponde al número total de caminos cortos entre cualquiera de dos nodos y $\sigma(i,j|v)$ a la cantidad de esos caminos que pasa por v.

Los grafos pueden tener pesos (weighted) o no tener pesos (unweighted). Un gráfico con pesos corresponde a aquel en que la fuerza de sus uniones puede variar. Para un grafo de n nodos su estructura se definirá como los pesos |V| x |V| de la matriz de adyacencia **A** [38,39]

$$aij = \begin{cases} 1 & si\ hay\ un\ enlace\ desde\ el\ nodo\ i\ al\ j \\ 0 & en\ otro\ caso \end{cases}$$

Donde sus pesos |V| x |V| corresponderán a la matriz de pesos **W**, donde

$$w_{ij} = \begin{cases} 1 & si\ a_{ij} = 0 \\ w_{ij} \epsilon\ \mathbb{R}\ en\ otro\ caso \end{cases}$$

Las diagonales de **A** y **W** son cero. Aquí, cada fila y columna representa un nodo.

Por último, diremos que una matriz es unidireccional sólo si **A** y **W** son simétricas [38,39]

$$\mathbf{A}=\mathbf{A}^{T}$$

$$\mathbf{W}=\mathbf{W}^{T}$$

Y un grafo no tendrá pesos sólo si **A** es igual a **W** multiplicado por un escalar [38,39]

$$\mathbf{A}= c\mathbf{W}$$

Sin embargo, esto no implica que un gráfico no sea unidireccional o no tenga pesos.

2.2.2 Método de imputación adaptativo

En general perfiles de poblaciones celulares pequeñas, como las células madre cancerígenas, son enmascarados debido a su reducida abundancia en el tumor completo. Este fenómeno se puede observar, por ejemplo, cuando se buscan asociaciones entre la sobrevida y la expresión de genes en muestras de sangre,[10] en donde las señales de interés de la subpoblación minoritaria se pierden o diluyen entre las señales de los tipos celulares prominentes.

Algoritmos de predicción utilizados en análisis genéticos (e.g. Random Forest) han permitido incorporar grandes volúmenes de información, logrando una mejor representación de las dimensiones que intervienen en la predicción de un fenómeno. Sin embargo, poblaciones celulares poco representadas, pueden implicar una alta cantidad de valores perdidos. Omitir el tratamiento previo de estos datos, puede resultar en una pérdida de valiosa información y poder predictivo.[40]

La pérdida de los valores pueden tener diferentes orígenes, dichos mecanismos se pueden categorizar en: datos perdidos completamente al azar (MCAR = *missing completely*

at random), datos perdidos al azar (MAR = *missing at random*) y datos pérdidos no al azar (MNAR = *missing not at random*). MCAR ocurre cuando la probabilidad de obtener un valor perdido no se relaciona a los valores perdidos u observados de alguna variable; y MAR ocurre cuando la probabilidad de obtener un valor perdido no se relaciona al valor perdido en sí mismo, pero sí a los valores observados de alguna variable. NMAR, ocurre cuando existe una relación entre la pérdida de los datos y su valor, es decir, el mecanismo de perdida de estos debe ser modelado.

El método de imputación a utilizar en este estudio corresponde al método *Adaptative Tree Imputation* propuesto por Iswaran, 2008.[41] Este método adaptativo, permite imputar los datos perdidos durante el proceso de construcción del bosque aleatorio. A continuación, se presentará el algoritmo para el cual se considerará el caso donde sólo *x* predictores tiene datos perdidos:

1. Para cada nodo *h*, impute los datos perdidos previo a la división (*splitting*). Sea $\mathbf{X}^*_{k,h}$ el set de datos sin valores perdidos para la *k*-ésima coordenada de las *x*-variables in-bag en *h*. Sea $\mathbb{P}^*_{k,h}$ la función de distribución empírica para $\mathbf{X}^*_{k,h}$. Para cada caso in-bag en *h* con un valor perdido para la *k*-ésima coordenada, impute obteniendo un valor aleatorio de $\mathbb{P}^*_{k,h}$. Repita para cada *k*. La división se continúa realizando de acuerdo con la regla de decisión escogida, una vez que los datos hayan sido imputados. Nótese que sólo los datos in-bag son usados como base para la imputación y la división.
2. Los nodos hijas no contendrán datos perdidos porque el nodo parental fue imputado anteriormente. Reseteé los datos imputados en las hijas a valores

perdidos. Proceda como en el paso 1, continuando hasta que el árbol no pueda seguir dividiéndose.

3. Los valores perdidos en los nodos terminales luego son imputados utilizando datos OOB (out of bag) sin valores perdidos de los nodos terminales de todos los árboles. Estos son imputados extrayendo valores de $\mathbb{P}^{*}_{k,h}$. Para variables categóricas, se imputa usando el valor más frecuente y para variables continuas, se imputa usando el promedio.

Se debe enfatizar que los datos imputados en el paso 1 solo se obtienen con el fin de asignar casos a los nodos hijas. La data imputada no es utilizada para calcular la estadística de división, ya que la data imputada es temporal y se reestablece a sus valores perdidos luego de su asignación a los nodos hijas.

El valor resumen finalmente imputado, expresado en la matriz de datos, para una variable continua es el promedio de los valores in-bag imputados de ese caso. Para una variable categórica, es el valor imputado in-bag más frecuente.

Este algoritmo de imputación también puede ser usado para datos de testeo. Para ello, los datos se dejan caer sobre los árboles y los valores perdidos se van imputando de manera dinámica como en el paso 1. Una vez que se ha establecido la pertenencia en los nodos terminales, los datos perdidos son imputados mediante medidas resumen como se describió anterioremente.

El mecanismo de pérdida de los datos también juega un rol en la exactitud del modelo. La exactitud de las predicciones, disminuye desde MCAR, a MAR y NMAR. El

desempeño de este modelo en NMAR, en general, es pobre a menos que la correlación de las variables sea alta.[40]

2.3 Modelos de Sobrevida

Los métodos clásicos estadísticos usan los datos de una o más muestras para inferir acerca de los parámetros de un modelo estadístico. Los algoritmos de ML tratan el mecanismo de los datos como desconocido, aprendiendo directamente de los datos.[42] Si bien ambas culturas permiten construir modelos de predicción, el modelamiento algorítmico de ML es más apropiado para la predicción de nuevas observaciones, debido a que generan una buena exactitud predictiva en nuevas observaciones, asumen un modelo de probabilidad para los datos desconocido, no requieren directa interpretabilidad sobre las relaciones entre los predictores y la respuesta. Además, permiten capturar complicadas asociaciones en los datos, sin la necesidad de justificarlas en el modelo.

Los modelos de ML permiten predecir o clasificar una variable respuesta de acuerdo a un conjunto de predictores (i.e. análisis supervisado) o encontrar patrones en la estructura de los datos (i.e. análisis no supervisado). De acuerdo con las especificaciones que se le entreguen al modelo, ML puede ser visto como un continuo de técnicas que van desde ser totalmente guiadas por los usuarios a ser totalmente guiadas por las máquinas.[43]

Estos métodos se consideran flexibles en el análisis de datos debido a que requieren mínimos supuestos sobre su procedencia. Además, permiten trabajar con un amplio número de variables, incluso cuando estas superan el número de observaciones. Por otra

parte, debido al desarrollo de algoritmos de optimización eficientes, también permiten ajustar modelos en un contexto de relaciones complejas (e.g. no lineales) entre variables.

Entre los métodos de ML más utilizados para clasificación en GB, se encuentra *Random Forest* o RF (árboles aleatorios o bosque aleatorio).[44–47] Este modelo permite clasificar una variable respuesta de tipo categórica como predecir una variable respuesta de tipo continua. Una variable respuesta de interés para este estudio es la sobrevida. A continuación, se describirán los modelos de sobrevida que serán implementados.

2.3.1 Modelo de Cox

El modelo de Cox es un modelo estadístico semi-paramétrico ampliamente utilizado para análisis de sobrevida. Su popularidad se atribuye a que no requiere de especificar una función de riesgo en particular y entrega razonables estimaciones de los coeficientes de regresión, *Hazard Ratios* (HR) de interés y curvas de sobrevida. Por ende, ante el desconocimiento de un modelo particular, el modelo de Cox continuará entregando resultados confiables. Además, es considerado un modelo "robusto", pues sus resultados se aproximan a los resultados de un modelo paramétrico.[48] A continuación, se define el modelo como

$$h(t, \boldsymbol{X}) = h_0(t) e^{\sum_{i=1}^{p} \beta_i X_i}$$

Esta fórmula nos dice que el riesgo de morir a un determinado tiempo corresponde al producto de la función de riesgo $h_0(t)$ y de la exponencial de la sumatoria, sobre p variables, de $\beta_i X_i$. Una característica importante de este modelo es el supuesto de riesgos

proporcionales. Considerando que la función de riesgo es la misma para todos los individuos, es de interés estimar la relación entre los tiempos de muertes para individuos expuestos a factores distintos. Para esto, el modelo supone riesgos proporcionales, lo cual implica que para cualquier tiempo *t*, el cociente entre las funciones de riesgo de dos individuos diferentes es constante e independiente del tiempo *t*.

2.3.2 RANDOM SURVIVAL FOREST (RSF)

Los datos de sobrevida usualmente son analizados con métodos paramétricos o semi-paramétricos como el anterior, basados en el supuesto de riesgos proporcionales. RSF corresponde a otro modelo de sobrevida que permite trabajar con estructuras de datos complejas en el contexto de una variable respuesta censurada hacia la derecha, izquierda o en intervalos. RSF deriva de *random forest* o bosques aleatorios, un método ensamblado o combinado (*ensemble*) que basa sus resultados en decisiones obtenidas a partir de un conjunto de modelos de *decision trees* o árboles de decisión. Los árboles de decisión corresponden a un conjunto ordenado de reglas de decisión que particionan el espacio de predictores, respecto de la respuesta, en regiones más simples y homogéneas con cada segmentación. Estos árboles crecen desde un nodo inicial ubicado en la parte más alta del árbol y continúan creciendo de acuerdo con un determinado criterio (e.g. *information gain* o ganancia de información). Posteriormente dicho árbol es cortado con el fin de evitar el sobreajuste del modelo y consecuente pérdida de extrapolación de resultados a otras muestras. Finalmente, en cada nodo terminal se realiza una votación por la clase más frecuente, i.e. clase que acumuló mayor cantidad de observaciones.

Random forest se puede considerar como el paso siguiente a los árboles aleatorios, pues éste utiliza muestras *bootstrap* (i.e. muestreo con reposición de la muestra original), para la construcción de muchos árboles de decisión. Estos árboles eventualmente conformarán un bosque aleatorio. Luego, RSF corresponde a un modelo de bosques aleatorios que incorpora la censura al modelo cuando la variable de interés es la sobrevida.

Utilizando un criterio de sobrevida predeterminado (e.g. criterio de log-rank), los árboles de sobrevida en el modelo de RSF crecen desde un nodo inicial y continúan su división de manera recursiva en nodos hijas. Una buena división corresponderá a aquella que maximice las diferencias de sobrevida en los nodos hijas. Esto quiere decir que de todas las variables x y de todos los posibles valores de división c, se seleccionarán aquellos x^* y c^* que maximicen la diferencia de sobrevida en los nodos hijas. Cada nodo debe contener un mínimo de $d_0 > 0$ eventos o muertes únicas, de no ser así el árbol alcanzará un punto de saturación donde no podrá continuar creciendo. A aquellos nodos más extremos, se les denominará nodos terminales (ζ).

La división de los nodos puede seguir diferentes criterios de división, como lo son[41]:

- Log-rank, divide los nodos maximizando la estadística de log-rank;
- Regla de conservación de eventos, divide los nodos buscando hijas cercanas al principio de conservación de eventos;[41]
- Log-rank score, divide los nodos utilizando la estadística estandarizada de log-rank; y

- Random log-rank, realiza una división aleatoria para cada una de las variables *p* candidatas en un nodo, y la variable que tenga la estadística log-rank mayor será utilizada para dividir el nodo.

A continuación, se describe el algoritmo utilizado por RSF [40,41]:

Algoritmo de RSF

1. Extraer *B* muestras *bootstrap* de la muestra original. Notar que cada muestra *bootstrap* excluye aproximadamente, en promedio, un 37% de los datos originales. Los datos excluidos son denominados datos OOB.

2. Haga crecer un árbol de sobrevida en cada muestra *boostrap*. En cada nodo del árbol, seleccione aleatoriamente *p* variables candidatas. El nodo será dividido usando aquella variable candidata que maximice la diferencia de sobrevida entre los nodos hijas.

3. Haga crecer el árbol a tamaño completo bajo la restricción que un nodo terminal debería tener no menos que $d_0 > 0$ muertes únicas.

4. Calcule la función de riesgo acumulado (CHF, *Cummulative Hazard Function*) para cada árbol. Promedie para obtener el CHF del ensemblado.

5. Usando la data OOB, calcule el error de predicción y calcule la función CHF.

2.3.2.1 PREDICCIÓN NODOS TERMINALES

La predicción de los nodos terminales se basa en la estadística de Nelson-Aalen y en el principio de conservación de eventos, los cuales serán descritos a continuación. Se entenderá por tiempos de sobrevida ($T_{1,h}$; $\delta_{1,h}$), …, ($T_{n(h),h}$; $\delta_{n(h),h}$), donde $h \epsilon \zeta$ corresponde a un nodo terminal. Luego, los tiempos de eventos serán $t_{1,h} < t_{2,h} < \ldots < t_{N(h),h}$ donde $N(h)$ corresponde al número total de tiempos distintos de eventos en el nodo h.[41]

La función de riesgo acumulada (*Cummulative Hazard Function*, *CHF*) será la misma para los casos dentro de un nodo $h \epsilon \zeta$, y será estimada a partir del estimador no paramétrico de Nelson-Aalen

$$\widehat{H_h}(t) = \sum_{t_{l,h} \leq t} \frac{d_{l,h}}{y_{l,h}}$$

donde, $d_{l,h}$ y $y_{l,h}$ corresponden al número de muertes e individuos en riesgo al tiempo $t_{l,h}$.

El caso anterior corresponde a una coordenada, por lo tanto, cuando un individuo i tiene un vector de covariables x_i con d dimensiones, la CHF para i corresponderá al estimador Nelson-Aalen para el nodo terminal de x_i

$$\widehat{H_h}(t|x_i) = \widehat{H_h}(t), \text{ si } x_i \epsilon h.$$

El principio de conservación de eventos asegura que la suma de los CHF estimados sobre los tiempos observados (ambos censurados y no censurados) equivalen al total de muertes.[41] Considérese como tiempos de sobrevida e indicadores de censura de los datos originales (no *bootstrap*) (T_1,δ_1),…,(T_n, δ_n), entonces el principio de conservación de eventos implica que:

$$\sum_{i=1}^{n} H(T_i|x_i) = \sum_{h\epsilon\zeta} \sum_{i=1}^{n(h)} \widehat{H}_h(T_{i,h}) = \sum_{h\epsilon\zeta} \sum_{i=1}^{n(h)} \delta_{i,h} = \sum_{i=1}^{n} \delta_i$$

donde $\sum_{i=1}^{n} \delta_i$ equivale al número total de muertes.

Luego, la mortalidad para cada individuo equivale al valor esperado de los CHF sumados en el tiempo T_j, condicionados a un vector de covariables específico x_i. De esta forma, la mortalidad mide el número de muertes esperadas bajo la hipótesis nula de sobrevidas similares para sujetos con el mismo valor de x_i.[41] En términos estadísticos, Ishwaran et al definen la mortalidad como

$$M_i = E_i\left(\sum_{j=1}^{n} H(T_j|x_i)\right)$$

donde E_i es la esperanza bajo la hipótesis nula de que todos los j son similares a i, es decir, todos los tiempos de sobrevida observados son consecuencia de una misma función de riesgo acumulado.

Con el propósito de obtener estimaciones para un bosque aleatorio, se construirá un ensamblaje o *ensemble* de los CHF de todos los árboles del bosque. Para ello se promediarán *B* árboles de sobrevida, obteniéndose estimaciones *OOB* y *bootstrap* que serán definidas a continuación.

Recordando el algoritmo, diremos que una observación es OOB para una remuestra si es que tal observación no está presente en esa remuestra. Los datos OOB se dejarán caer

sobre los árboles construidos con los datos *in-bag* o bootstrap y se obtendrá un **OOB ensemble** o ensamblado de los CHF

$$H_e^{**}(t|x_i) = \frac{\sum_{b=1}^{B} I_{i,b} H_b^*(t|x_i)}{\sum_{b=1}^{B} I_{i,b}}$$

donde $I_{i,b}$ = 1 si i es un caso OOB para la remuestra b (de otra forma $I_{i,b}$ = 0), y $H_b^*(t|x_i)$ es el CHF para el árbol de sobrevida construido a partir de la b-ésima muestra boostrap.

Además, se obtendrá un **boostrap ensemble** de los CHF, el cual considerará todos los datos, sean OOB o in-bag. El bootstrap ensemble de CHF para i es

$$H_e^*(t|x_i) = \frac{1}{B} \sum_{b=1}^{B} H_b^*(t|x_i)$$

Considerando esta última expresión y el principio de conservación de eventos, el cual supone que los nodos terminales comparten una misma función de *hazard* estimada, es posible obtener una estimación ensamblada de la mortalidad para i

$$\widehat{M}_{e,i}^* = \sum_{j=1}^{n} H_e^*(T_j|x_i).$$

Asimismo, una estimación OOB ensemble de la mortalidad definida como

$$\widehat{M}_{e,i}^{**} = \sum_{j=1}^{n} H_e^{**}(T_j|x_i).$$

2.4 Estrategias de validación y evaluación del modelo

Los modelos usados en ML deben ser validados y evaluados. La validación para un modelo predictivo se enfoca en la generalización, que es la habilidad del modelo estimado de poder predecir correctamente la respuesta en nuevas observaciones.[49] La forma de validar un modelo predictivo es evaluando su grado de sobreajuste en los datos de entrenamiento. Para esto, se compara el rendimiento del modelo estimado en los datos de entrenamiento y de testeo. Un set de entrenamiento permite ajustar y entrenar el modelo, mientras que el set de testeo permite probar el modelo en un conjunto de datos que no han sido utilizados previamente. Si su rendimiento es mejor en el set de entrenamiento que en el de testeo, entonces hay sobreajuste. La validación del modelo es necesaria para evitar resultados que sobre estimen la exactitud del clasificador. Existen muchas técnicas de validación, dentro de ellas está la de testear el modelo en un set de datos que es independiente de la muestra de entrenamiento, o por ejemplo usar un método de submuestreo conocido como validación cruzada.

Por otro lado, la evaluación del modelo busca evaluar su poder predictivo.[49] El poder predictivo se refiere al desempeño, del modelo estimado, en nuevos datos. Usualmente se utilizan métricas que usan un set de datos independiente de la muestra de entrenamiento. Estas métricas actúan de manera similar a la evaluación de los test médicos de tamizaje. Es decir, pueden ser evaluados con respecto a su sensibilidad, especificidad, curva Característica Operativa del Receptor (ROC), y tasas de clasificación erróneas.

2.4.1 Error de predicción

Otra forma de evaluar la exactitud del modelo RSF es utilizando el error de predicción OOB. Para estimar este error se utiliza el índice de concordancia de Harrell, el cual estima la probabilidad de que, en un par de casos seleccionados aleatoriamente, aquel que falle primero coincida en tener el peor resultado predicho. Este valor se calcula mediante 1-C, donde C es el índice de concordancia de Harrell. Una ventaja de utilizar el error de predicción OOB en lugar de la validación cruzada, es que el primero utiliza los datos originales tanto para construir el RSF como para estimar el error. Por otro lado, el segundo deja de lado una proporción de la muestra para construir el RSF y otra para testearlo, lo cual genera árboles con menor rendimiento. Además, utilizar el error de predicción OOB implica una mayor eficiencia computacional que la validación cruzada, puesto que esta última implica la construcción de k RSF según k subconjuntos sean seleccionados.[50] El error de predicción OOB también puede ser utilizado para una selección de los parámetros número de variables candidatas elegidas aleatoriamente en cada división de los nodos (x^*).

2.4.2 Puntaje Brier

Otra forma de evaluar el un modelo, cuya respuesta es binaria, es mediante el puntaje Brier (*Brier score*) definido[51] como el cuadrado de la diferencia entre el estado de sobrevida observado (e.g., 1 = vivo al tiempo t o 0 = muerto al tiempo t) y un modelo basado en la predicción del tiempo de sobrevida t. Este puntaje permite evaluar el rendimiento global del modelo global, a diferencia del índice de concordancia, el cual

permite evaluar la habilidad discriminativa del modelo. Por ende, dado un tiempo particular t, el puntaje tiempo-dependiente Brier para un sujeto particular es:

$$BS(t, \hat{S}) = E\left[Y_i(t) - \hat{S}(t|X_i)\right]^2$$

Este puntaje se clasifica como 0=perfecto, 1= pobre y 0,25= no informativo. Además, es estratificado en 4 grupos según la mortalidad ensamblada (*ensemble mortality*) y su cálculo se basa en la ponderación en base al recíproco o inverso de la probabilidad de censura (*Inverse Probability of Censoring Weights,* IPCW). Este método,[51,52] se utiliza para evitar el sesgo en el promedio de la población. En particular, el sesgo se produce a nivel de las estimaciones de las probabilidades de sobrevida, pues los tiempos de eventos observados no son representativos de los tiempos de eventos de toda la población (e.g. pacientes jóvenes son más propensos a renunciar a un tratamiento), haciendo la inferencia acerca de la población, sesgada. Usualmente, los métodos de sobrevida consideran una censura no informativa, que asume independencia en la censura, generando un sesgo en la estimación cuando tal supuesto no se cumple en la población. El puntaje Brier asume dependencia en la censura, ya sea del evento y los tiempos de sobrevida o de las covariables. Su cálculo es posible debido a que el método IPCW pondera aquellas observaciones no censuradas y con carácterísticas similares a la de las observaciones censuradas.

Los 4 grupos de estratificación antes mencionados corresponden a los cuartiles 0-25, 25-50, 50-75 y 75-100 de los valores de la mortalidad.

2.4.3 BONDAD DE AJUSTE

Como vimos anteriormente, una forma de evitar el sobre ajuste de un modelo de ML es dejando un set de entrenamiento y otro de testeo, lo cual permite verificar el ajuste de nuestro modelo en nuevos datos. La validación cruzada es, en general, un algoritmo que intenta balancear los beneficios de dividir los datos en sets de entrenamiento y testeo para obtener una mejor estimación de la exactitud de la clasificación en el modelo entrenado, con los beneficios de ganar precisión en el ajuste del modelo usando toda la muestra para entrenarlo. Este método consiste en dividir los datos en *k* subconjuntos (*k-fold cross validation*), donde cada uno de ellos es utilizado como set de testeo de un modelo entrenado con el resto de la muestra. Otra forma alternativa a este método es usar métodos clásicos para evaluar el ajuste de modelos, tales como el criterio de información de Akaike (AIC), criterio de información Bayesiano (BIC), Mallow's Cp o R^2 ajustado.

En el modelo estadístico de riesgos proporcionales de Cox, también se utilizan medidas de bondad de ajuste para su evaluación, como lo es verificar el supuesto de riesgos proporcionales utilizando los residuos de Schoenfeld, definidos como:[53]

$$S_{ij}(\boldsymbol{\beta}) = Z_{ij}(t_i) - \bar{Z}_j(\boldsymbol{\beta}, t_i)$$

Donde i corresponde al i-ésimo individuo y j a la j-ésima variable Z, y t corresponde al tiempo de ocurrencia del evento. Mientras que $\boldsymbol{\beta}$corresponde al coeficiente de riesgo.

Estos residuos se definen para cada predictor del modelo y para cada sujeto que presenta el evento. Luego, los pasos a seguir para este análisis son:[48]

1. Se realiza el cálculo de los residuos para cada covariable. El residuo para el sujeto i-ésimo al tiempo t de la j-ésima variable corresponderá al valor observado de aquella variable menos el promedio ponderado de los valores para esa variable en el resto de los sujetos que siguen en riesgo al tiempo t. Los pesos corresponderan con el hazard de cada sujeto perteneciente a aquellos que aún están en riesgo.

2. Posterior a la obtención de los residuos es necesario crear una variable que haga un ranking de los sujetos en riesgo de acuerdo a los tiempos. El sujeto que presente el evento primero, tendrá el valor 1, y así sucesivamente.

3. Luego, se hará un test de correlación entre las variables creadas en los puntos 1 y 2. La hipótesis nula se referirá a que la correlación entre los residuos de Schoenfeld y los eventos en ranking es cero. De ser rechazada la hipótesis nula, diremos que el supuesto de riesgos proporcionales ha sido violado.

3. OBJETIVOS

3.1 PREGUNTA DE INVESTIGACIÓN

¿La propuesta de genes asociados a diferenciación, transmisión de señales, metilación y metabolismo predicen la sobrevida en pacientes adultos con glioblastoma?

3.2 OBJETIVO GENERAL

Predecir la sobrevida en pacientes adultos con glioblastoma, en base a la expresión de un conjunto de genes candidatos identificados de TNTs por un grupo de expertos.

3.3 OBJETIVOS ESPECÍFICOS

- Predecir la sobrevida de pacientes adultos con GB en base a los genes candidatos e información clínica.
- Evaluar la contribución de los genes candidatos en la predicción de la sobrevida de los pacientes adultos con GB.

4. METODOLOGÍA

4.1 DISEÑO DE ESTUDIO

Estudio analítico, observacional, longitudinal y retrospectivo.

4.2 UNIVERSO Y MUESTRA

Para efectos de este análisis, se hará uso de información recolectada a partir de la base de datos de acceso público, *TCGA*. Esta iniciativa del *National Cancer Institute* (NCI) de Estados Unidos y el *National Human Genome Research Institute* (NHGRI) han permitido recolectar, organizar y analizar datos de genética, epigenética, transcriptómica y proteómica, así como también de procesar la calidad y cantidad de muestras biológicas de 33 tipos de cáncer. Esta data se encuentra disponible de manera online a través del repositorio de NCI en *Genomic Data Commons* (GDC). Para este trabajo, se consideró información clínica, genómica, en particular, de transcriptómica e información asociada al bioespecimen, es decir, a la muestra física del tumor de cada paciente. La información reunida en este trabajo ha sido procesada por el *International Genomics Consortium* (IGC), el cual corresponde al *Bioespecimen Care Repository* (BCR) del NCI *Center of Cancer Genomics* (CCG).

4.3 VARIABLES

Variables respuesta

- Variables de sobrevida.
    - Tiempo de sobrevida en meses.
    - Evento o variable censura (1: muerte).

Variables predictoras

- Genes de expresión:

Expresión genética (datos de RNA-seq) de transcritos propuestos por los expertos:

1. *OGDH*
2. *IDH1 (+)*
3. *IDH2 (+)*
4. *WWTR1*
5. *NANOG*
6. *OLIG2*
7. *TTYH1*
8. *GAP43*
9. *SDHB*

10. *FH*

11. *SLC1A5*

12. *GLUL*

13. *YWHAG*

14. *GJA1(-)*

15. *KEAP1*

16. *JMJD8*

(+) control positivo

(-) control negativo

Variables clínicas del participante y tumor

1. Género (1: hombre, 0: mujer)
2. Raza (blanco, negro o afroamericano, asiático).
3. Edad, en años, al diagnóstico.
4. Recuento de mutaciones. Marcador de ADN.
5. Estado del gen IDH (mutado o *wild type*). Marcador de ADN.
6. Puntaje de stroma en el tumor (algoritmo ESTIMATE).[54]
7. Subtipo de tumor (Clásico, neural, proneural, mesenquimal, G-CIMP).

Variables auxiliares (para imputar datos perdidos en variables predictoras):

Expresión genética (datos de RNA-seq) de transcritos disponibles en la base de datos, pero no en el conjunto de variables predictoras. Estos corresponden a 51,690 transcritos.

4.4. ANÁLISIS DESCRIPTIVO

Previo al modelamiento estadístico, se realizará un análisis exploratorio y descriptivo de los datos de testeo, de entrenamiento y de los datos en su totalidad, con el fin de obtener una primera aproximación al estado de los datos, su comportamiento o patrones que sean de interés considerar en los análisis posteriores. Esto además permitirá describir las variables estipuladas en cada subconjunto de datos y verificar que la descripción entre ellas sea similar, corroborando una buena aleatorización de la muestra. Para esto, se utilizarán medidas de resumen, histogramas y métodos no supervisados multivariados, como lo son el análisis de componentes principales y métodos jerárquicos. El análisis de componentes principales permitirá conocer aquellos componentes que registren la mayor cantidad de variabilidad en nuestros datos. Con el fin de ilustrar estos análisis, se utilizarán mapas de calor que reflejen la correlación de la información.

Respecto al método de análisis jerárquico aglomerativo, se realizará un dendrograma y para ello se considerarán las distancias entre variables, particularmente, entre los genes de interés (espacio de variables). Las medidas de disimilitud entre variables se basarán en la distancia euclidiana y el método (*linkage method*) utilizado para su cálculo corresponderá al método *complete linkage*. Éste método ha demostrado ser invariante ante transformaciones monótonas, como lo es el método de normalización y, por ende, las medidas de disimilitud que tienen el mismo ranking relativo u orden resultan en la misma estructura de cluster.[55] Utilizando análisis jerárquico, se podrá conocer la estructura jerárquica que entrega el algoritmo y con el propósito de corroborar si ella presentada la realidad, se utilizará el coeficiente de correlación CCC.

En los datos de testeo, se aplicará además un algoritmo de dibujo de grafo que permita estudiar la matriz de proximidad obtenida con las distancias euclidianas utilizadas entre las observaciones. Estudiando las propiedades de la matriz de proximidad, se definirá si el algoritmo definirá un grafo uni o bidireccional, y con pesos o sin pesos.

4.5. MODELAMIENTO ESTADÍSTICO

Para el modelamiento estadístico, se analizarán 3 subconjuntos diferentes de predictores, una contendrá sólo los transcritos de los genes seleccionados; la siguiente, contendrá transcritos y variables clínicas; y la última, contendrá sólo datos clínicos. Esto se hará con el fin de estudiar en detalle la contribución de los genes estudiados. A continuación, se detallarán los modelos estadísticos a utilizar en este trabajo.

4.5.1 MODELO DE IMPUTACIÓN PARA DATOS ARN-SEQ

Aquellos transcritos que presenten niveleves de expresión 0, serán considerados como datos perdidos, y para ellos se utilizará el modelo adaptativo propuesto por Ishwaran, 2008, *Adaptative Tree Imputation (ATI)*. Para datos clínicos, no se realizará imputación. ATI permite imputar los datos en la medida que el bosque aleatorio es construido, en este caso el bosque para el modelo RSF. La imputación funciona extrayendo datos de manera aleatoria desde el set de datos no perdidos in-bag, en cada nodo del árbol construido.

Este modelo puede ser utilizado en sets de datos con baja a moderada cantidad de valores perdidos. En el caso de que exista un alta proporción de valores perdidos en algunos de los genes propuestos para este estudio, el algoritmo puede ser usado, pero es necesario iterarlo con el fin de mejorar la exactitud en los datos imputados. La iteración funciona de la siguiente manera. Posterior a la obtención de los valores imputados en la construcción del primer bosque, se construye un nuevo bosque usando los datos imputados. A continuación, en el nuevo bosque, para cada caso que originalmente era perdido, se extrae un valor aleatorio de los datos no perdidos in-bag del nodo terminal de esa observación, en todos los árboles del bosque. Entonces, la imputación se realiza considerando una estadística resumen completa (datos in-bag y OOB). Luego, se utilizan los datos reimputados y se vuelve a crecer otro bosque. Repita iterativamente.

Tang & Ishwaran,[40] estudiaron el desempeño de la iteración de este algoritmo de imputación con un máximo de 5 iteraciones. Esto permitió obtener valores imputados exactos y un error de predicción OOB que calzó con el error de predicción en los datos de testeo, bajo moderadas cantidades de datos perdidos (5-10%). A medida que aumentó la proporción de valores perdidos (25-50%), el error de predicción OOB sobre estimaba el error en los datos de testeo. Esto a su vez, puede afectar el desempeño de la importancia de las variables (VIMP), definidas más adelante.

4.5.2 MODELO PREDICTIVO DE SOBREVIDA BASADO EN RSF

En este trabajo se utilizará el modelo RSF, un método *ensemble* para el análisis de datos censurados hacia la derecha. Este método deriva del modelo de RF, el cual también corresponde a un modelo de tipo *ensemble*. Estos modelos se dicen son esemble, pues combinan múltiples modelos que producen predicciones precisas promediando las predicciones que provienen de diferentes modelos, y han probado ser útiles en numerosas aplicaciones.[49]

4.5.2.1 REGLAS DE DIVISIÓN

Para el análisis se utilizará el modelo de RSF implementando el paquete del software R randomSurvivalForest y el paquete survival. En cada instancia se harán crecer 1.000 árboles. Las divisiones en cada nodo se realizarán siguiendo la regla de división log-rank, la cual se basa en la estadística de log-rank. La división propuesta es de la forma $x \leq c$ y $x \geq c$, donde x denota uno de los genes y c el valor de corte. Luego la estadística de log-rank es:

$$LogRank\ (X, c) = \frac{\sum_{i=1}^{E} d_{t_i,hija_j} - E(D_i)}{\sqrt{[\sum_{i=1}^{E} var(D_i)]}}$$

donde E es el número de tiempos distintos en el nodo maternal, $d_{t_i,hija_j}$ es el número de eventos al tiempo t_i en el nodo hija j = 1,2, $R_{t_i,hija_j}$ es el número de individuos en riesgo al tiempo t_i en el nodo hija j=1,2, y $R_{t_i} = \sum_{j=1}^{2} R_{t_i,hija_j}$, $d_{t_i} = \sum_{j=1}^{2} d_{t_i,hija_j}$, D_i es la variable

aleatoria correspondiente al número de eventos en el nodo hija j=1 para el tiempo de evento i-ésimo distinto. $E(D_i) = R_{t_i,hija_1}(d_{t_i})$ es la esperanza de D_i, $Var(D_i) = \frac{d_{t_i}(R_{t_i} - d_{t_i})}{R_{t_i}{}^{-1}} \times \frac{R_{t_i,hija_1}}{R_{t_i}}\left(1 - \frac{R_{t_i,hija_1}}{R_{t_i}}\right)$ es la varianza de D_i. La mejor división será definida como aquella que maximiza el valor absoluto de $LogRank\ (X, c)$.

4.5.2.2 Error de predicción

El error de predicción se calculará utilizando el índice de concordancia de Harrell. Este índice permite distinguir aquellos individuos con o sin el evento de interés. Para calcular C, se requerirá del ensemble OOB de los CHF descrito en el marco teórico. Sean $t_1^o, \dots, t_m^o$ los puntos únicos preseleccionados (se hará uso de los tiempos de eventos $t_1, \dots, t_N$). Para rankear dos casos i y j, diremos que i tiene un peor resultado predicho que j si

$$\sum_{l=1}^{m} H_e^{**}(t_l^o | x_i) > \sum_{l=1}^{m} H_e^{**}(t_l^o | x_j).$$

Utilizando esta regla, se calculará C siguiendo los pasos descritos a continuación.[41]

1. Forme todos los pares posibles de casos de los datos.
2. Omita aquellos pares donde la sobrevida más corta esté censurada. Omita los pares i y j si $T_i = T_j$ a menos que uno tenga el evento (muerte). Se dirá Permisible al número total de pares permisibles.

3. Para aquellos pares permisibles donde $T_i \neq T_j$, cuente 1 si la sobrevida más corta tiene un peor resultado predicho; cuente 0,5 si los resultados predichos están empatados. Luego, para aquellos pares donde $T_i = T_j$ y ambos están muertos, cuente 1 si los resultados predichos están empatados; de otra forma, cuente 0,5. Y para aquellos pares donde $T_i = T_j$ y al menos 1 está muerto, cuente 1 si la muerte tiene peor resultado predicho; de otra forma, cuente 0,5. Se dirá que la suma de estas cuentas sobre todos los pares permisibles será la concordancia.

4. El índice de concordancia, C, será definido como C = Concordancia/Permisible.

La estimación de C obtenida a partir de los datos OOB será denominada C^{**}. Luego, el error de predicción OOB, será $EP^{**}= 1- C^{**}$, donde $0 \leq EP^{**} \leq 1$. Un valor $EP^{**} = 0,5$ indicará que la predicción no es mejor que el azar.

4.5.2.3 IMPORTANCIA DE LAS VARIABLES

La importancia de las variables (*Variable Importance, VIMP*) será calculada restando el error de predicción del ensemble cuando se ignora la variable x menos el error de predicción del ensemble original. Para ignorar la variable x, se utilizarán observaciones OOB en el árbol de sobrevida construido con data in-bag. Cuando dicha observación se encuentre con una división donde x maximiza la diferencia de sobrevidas, se asignará a dicha observación un nodo de manera aleatoria. Luego, el CHF de cada uno de esos árboles se calculará y se promediará. [41]

Valores altos de importancia indican variables con habilidad predictiva, mientras que valores cero o negativos identificarán variables no predictoras que pueden ser

filtradas. VIMP para una variable x mide el aumento o disminución en el error de predicción en los datos de testeo si x no estuviese disponible. Considerando que el árbol original fue construido considerando la variable x, es probable que el error de predicción no varíe mucho cuando ésta no sea considerada, pero sí su valor VIMP.

4.5.3 Evaluación y validación del modelo

Con el propósito de validación de los modelos de RSF y Cox-PH los datos se dividirán de manera aleatoria en sets de entrenamiento (70%) y de testeo (30%). La validación interna del modelo en RSF se realizará utilizando el error de predicción basado en el índice de concordancia y su validación externa, se realizará utilizando los datos de testeo. Luego, para el modelo de Cox-PH, se verificará el supuesto de riesgos proporcionales. La comparación de ambos modelos se realizará de manera cualitativa.

4.6 Aspectos éticos

No hay conflictos éticos que declarar. Se declara que esta tesis reporta resultados originales, contiene suficientes detalles en su extensión y apropiadas referencias a la información obtenida por terceros. Se proveerá en un anexo los scripts/git-hub para transparentar el trabajo realizado y reproducibilidad. Los datos son de tipo secundario y han sido sometidos a un comité de ética. No existen conflictos de tipo financiero que puedan influir en los resultados o su interpretación.

5. RESULTADOS

Survival prognosis and resistance to treatment in Glioblastoma: Focus on intercellular communication systems.

5.1 ABSTRACT

Currently, most of the prognostic genetic, translational, proteomics and phosphor-proteomic approaches to discover biomarkers of disease are based on large databases that identify high- and low-profile gene expression based on the abundance of transcripts or proteins within the tumor. However, biomarkers expressed in small cell subpopulations are “diluted” in most of the large-scale analysis. This is the case of factors expressed in tumor stem cells or particular structures required for survival, such as tunneling nanotubes (TNTs). Throughout several previous experiments in human glioblastoma stem cells, HIV infected macrophages and isolated TNT , we have found a remarkable coincidence of genes. Our study will assess the predictive accuracy of this unique gene proposal by using Random Survival Forest (RSF), a machine learning technique for predicting survival time based on an ensemble of decision trees.

Using descriptive analysis, our genes demonstrate similarity around mixed biological processes such as, differentiation/metabolism/methylation/TNTs. We could also confirm the lack of representation our genes had in large-scale bulk-tumor data. We expect

to continue assessing these genes under a more robust biological context by adding genomic information to compensate signal-to-noise limitations in the available data.

Key words: cancer, glioblastoma, prognosis, random survival forest, tunneling nannotubes.

Funding: This work was funded by The National Institute of Mental Health, grant MH096625, the National Institute of Neurological Disorders and Stroke, NS105584, and UTMB internal funding (to E.A.E).

5.2 INTRODUCTION

Glioblastoma (GB, astrocytoma grade IV, WHO classification) is one of the most frequent and aggressive malignant brain tumor among infiltrative gliomas with a worldwide incidence rate of less than 10 cases per 100,000 people per year.[56] Current combined temozolamide treatment plus radiotherapy has shown improvement of survival from 3 months [57] when no treatment is performed to a median of 14.6 months.[58] Despite large amounts of data collected from the analysis of healthy and GB brain tissue by genetics, genomics, transcriptomics, proteomics, and metabolomics (summarized by the term "omics"), no new treatments or biomarkers have been yet discovered. Most studies consider bulk of the tumor genomic analyses, which have allowed to identify the association between phenotype and molecular markers, such as genome wide analysis. Most of these studies consider up or downregulated genes, usually related to differentiation processes. On the contrary, subcellular populations, such as glioblastoma stem cells are underrepresented due to low abundance or susceptibility of low gene expression signals.

We propose to analize a unique set of genes identified in tunneling nanotubes (TNTs) present during the active stages of GB colonization and development. The biological functions of these genes candidates can be classified into stem cell development, differentiation, metabolism, and DNA methylation. Thus, the aim of this study is to analize the predictive power of these genes using transcriptomic and clinical data from "TCGA-GBM" project. We are aware of the relevance these gene candidates have in small cell populations and connecting structures such as TNTs, yet it is important that we first

understand the impact these genes have in tumor bulk-data, in order to proceed to analyze further data sources.

Despite the risk of only using bulk-tumor transcriptomic data, as we suspect that the signal of our transcripts could be lost because of dilution among higher expressed transcripts, the benefit of positive results would be higher (e.g. not having to purify the cellular subpopulation of the excised tumor, or continue an investigation of these candidate genes in peripheral blood). In this study, we expect to link the heterogeneity of the tumor with the signatures of "stemness" of the tumor (i.e. stem cell-like phenotype). Understanding the role these genes accomplish in tumor resistance, our results can lead to better understand of patients survival.

5.3 METHODS

We have chosen 16 genes from a remarkable coincidence, of 24 genes, between HIV reservoir and GB. In particular, we identified a subset of genes that have coincided in several experiments in HIV-infected macrophages, human GB cell lines and isolated TNT using mRNA transcripts and proteomics analysis. [4–7] These experiments used isolated stem cell lines or cell-to-cell anatomic substructures such as, TNT mithocondria obtained between connecting cells.[4]

In order to analyze these small cell population and subcellular structures, living in a proportion of 1 cell in 10^6-10^{12} in HIV and less than 1% of the total tumor in GB, we have used "TCGA-GBM" data collection of RNA-seq data, to analyze the predictive power of these biomarkers in the patients' survival using a machine learning algorithm called Random Survival Forest (RSF).

RSF splitting rule was based on Log-Rank statistic and sampling with replacement was considered as the resampling method. In order to control the variance from the OOB error and improve VIMP measures, the number of trees was 5,333 considering the default rule reported in `randomSurvivalForest` R package.[59] Missing gene expression values were imputed using the Adaptative Tree Imputation algorithm in Ishwaran, 2008, which is already implemented in the aforementioned package.[41] In addition, Cox proportional-hazards (CPH) model was fitted to compare qualitatively its performance to RSF. Harrell concordance index was used as validation practice. Subsequently, variable importance analysis was performed to identify relevance of variables in predicting patients' survival.

All analyses were performed in R version 3.6.0 and RStudio 1.2.1335, using packages survival 2.44-1.1 version and randomSurvivalForest 2.9.1 version.

5.4 RESULTS

5.4.1 SAMPLE AND CLINICAL DATA

Our dataset contains publicly available molecular and clinical data from The Cancer Genome Atlas (TCGA) participants with newly diagnosed GB between 1997 and 2011. We have only considered open access clinical and FPKM normalized and harmonized to GRCh38 RNA-Seq sample collection. Combined TMZ chemotherapy and radiation treatment is documented for all patients in this cohort. Expression values of multiple ensemble gene IDs that mapped to a single gene name were averaged, so each gene name had only one expression value. This resulted in a total of 51,716 profiled genes from 151 independent tumor samples. **Table 1** summarizes phenotypical and clinical data of the total sample, as well as for the train and test subsets.

Table 1. Clinical characteristics of our sample.

Variables	Datasets		
	Total Data	**Train Data**	**Test Data**
Gender			
• **Female**	54 (35.8%)	35 (33.3%)	19 (41.3%)
• **Male**	97 (64.2%)	70 (66.7%)	27 (58.7%)
Age (years)			
• **Median**	61	62	60
• **Frequency**			
◦ **(20-40]**	13 (8.6 %)	7 (6.6%)	6 (13%)
◦ **(40-60]**	62 (41.1%)	44 (41.9%)	18 (39.1%)
◦ **(60-80]**	69 (45.7%)	49 (46.7%)	20 (43.5%)
◦ **>80**	7 (4.6%)	5 (4.8 %)	2 (4.3%)
Gene expression-based tumor group			
• **Classical**	37 (24.5%)	26 (24.8%)	11 (23.9%)
• **Neural**	26 (17.2%)	18 (17.1%)	8 (17.4%)
• **Proneural**	28 (18.5%)	19 (18.1%)	9 (19.6%)
• **Mesenchymal**	50 (33.1%)	35 (33.3%)	15 (32.6%)
• **G-CIMP**	8 (5.3%)	5 (4.8%)	3 (6.5%)
• **Non defined**	2 (1.3%)	2 (1.9%)	-

Descriptive analysis showed comparative profiling of our total data indicating unequal distribution of expression values among gene candidates (**Fig. S1.**). Our data revealed only 23.8% of null expression values of the transcript NANOG, which we interpreted as missing values due to transcript low abundance in the bulk of the tumor. Heatmap analysis of pairwise correlations between genes of interest showed, under the commonly used 5% significance threshold, 37 significant associations (**Fig. S2.C**), being positively correlated FH and SDHB, GAP43 and WWTR1, and negatively correlated WWTR1 and OLIG2. In addition, the results of principal component (PC) analysis of the candidate gene expression values revealed 16 PCs, of which the first 3 contained most of the original variability in the data (Fig. S2A), being PC1-C2 the ones explaining 83.9% of the variance, with no apparent differences between train and test samples (**Fig. S2.B**). Heatmap between PC information and expression values of candidates genes (**Fig.S2.D**) also reflected mostly a positively correlation of PC1 with GJA1 expression and negatively with GAP43 expression. PC2 was mostly positively correlated with GAP43, GLUL, and WWTR1. From the resulting dendrogram (**Fig. S3.A**), transcripts grouped in three major clusters according to the elbow method (**Fig. S3.B**). First cluster grouped most of our genes, while second cluster only considered GJA1, and last cluster grouped GAP43 and GLUL. The clustering results were good based in the analysis made by using the Cophenetic correlation coefficient, which revealed 94.2% of correlation between Euclidian distance of input observations and cophenetic distance derived from the dendrogram.[34] Survival analysis estimated by Kaplan-Meier curves stratified by tumor subtype showed how G-CIMP profile had better survival than other subtype gene-expression based tumors (**Fig. S4.**), which should be carefully considered due to its small sample size.

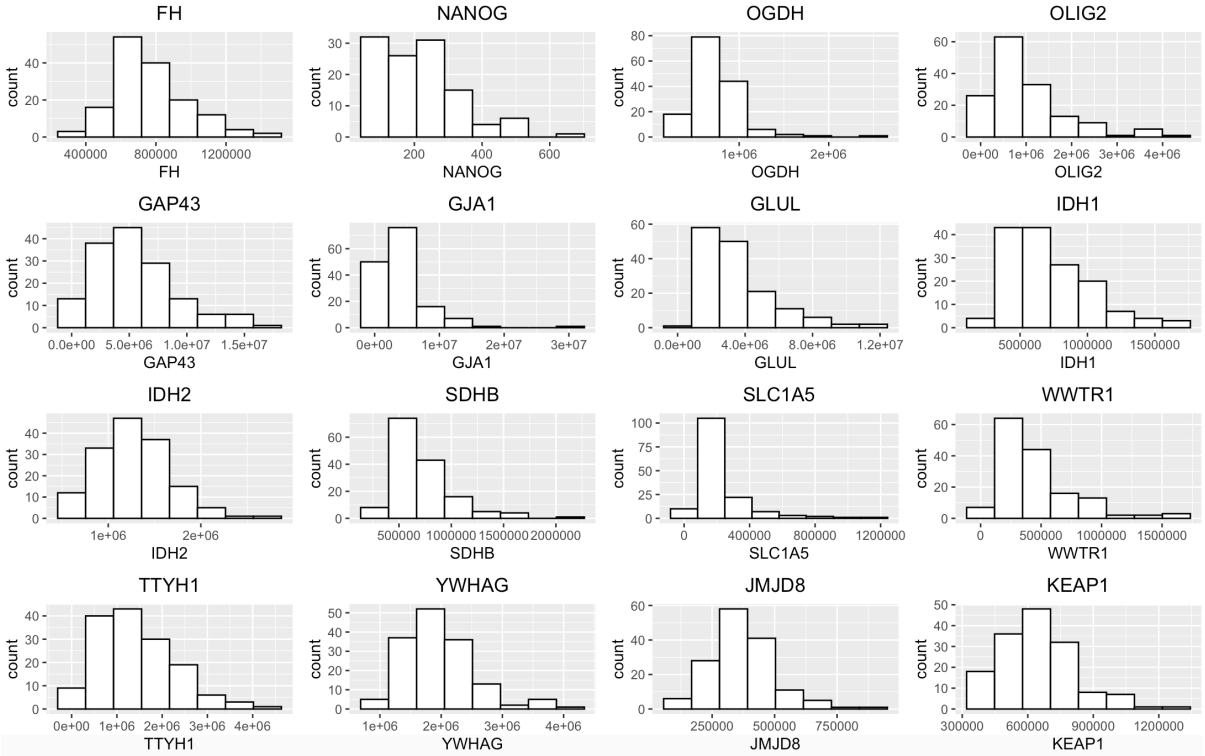
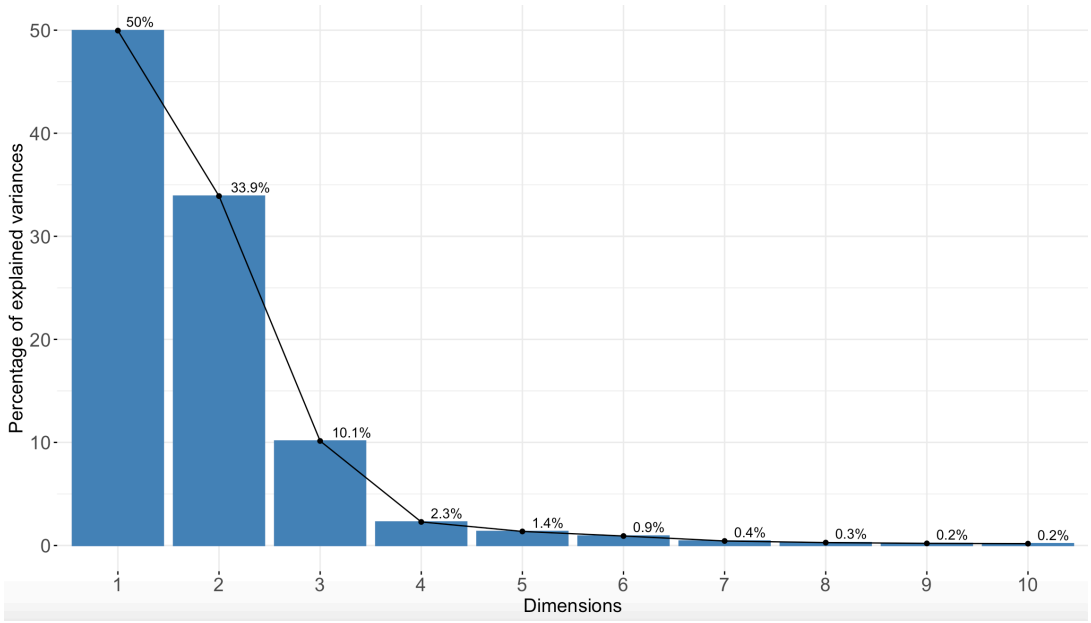
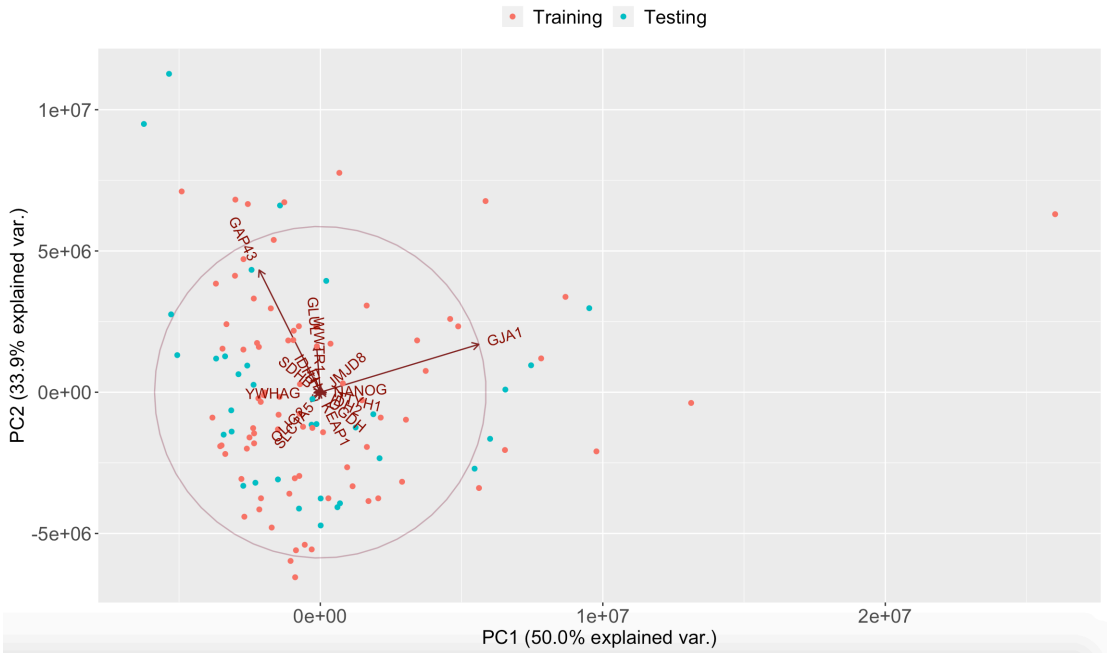


Fig. S1. Sample distribution of expression values of 16 selected transcripts in 151 glioblastoma samples. Gene expression histograms. Horizontal axis shows expression values for each gene, while vertical axis shows their frequency. Notice the heterogeneity in gene expression distributions.

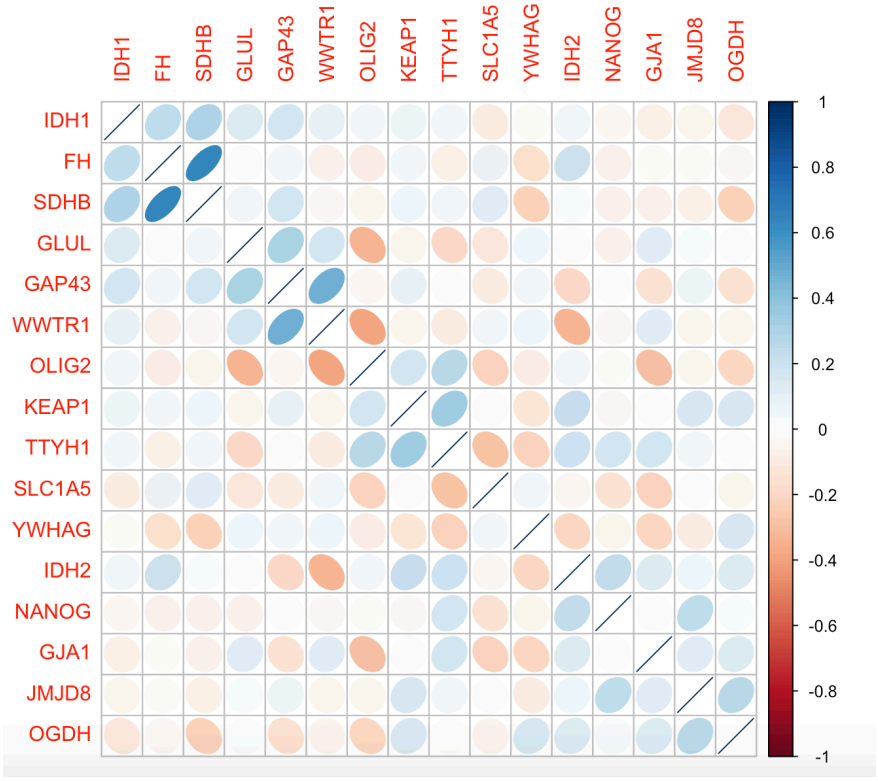
(A)


Percentage of explained variances
Dimensions
50%
33.9%
10.1%
2.3%
1.4%
0.9%
0.4%
0.3%
0.2%
0.2%

(B)


Training
Testing
PC2 (33.9% explained var.)
PC1 (50.0% explained var.)
GAPA3
GJA1
YWHAG
NANOG
XEAP1

(C)


IDH1
FH
SDHB
GLUL
GAP43
WWTR1
OLIG2
KEAP1
TTYH1
SLC1A5
YWHAG
IDH2
NANOG
GJA1
JMJD8
OGDH
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

(D)

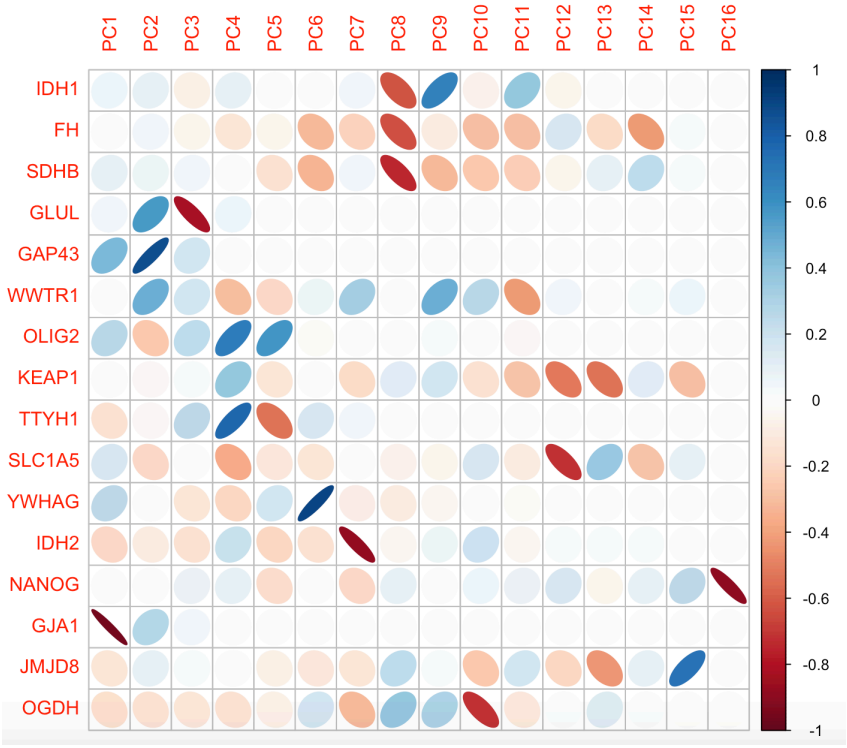
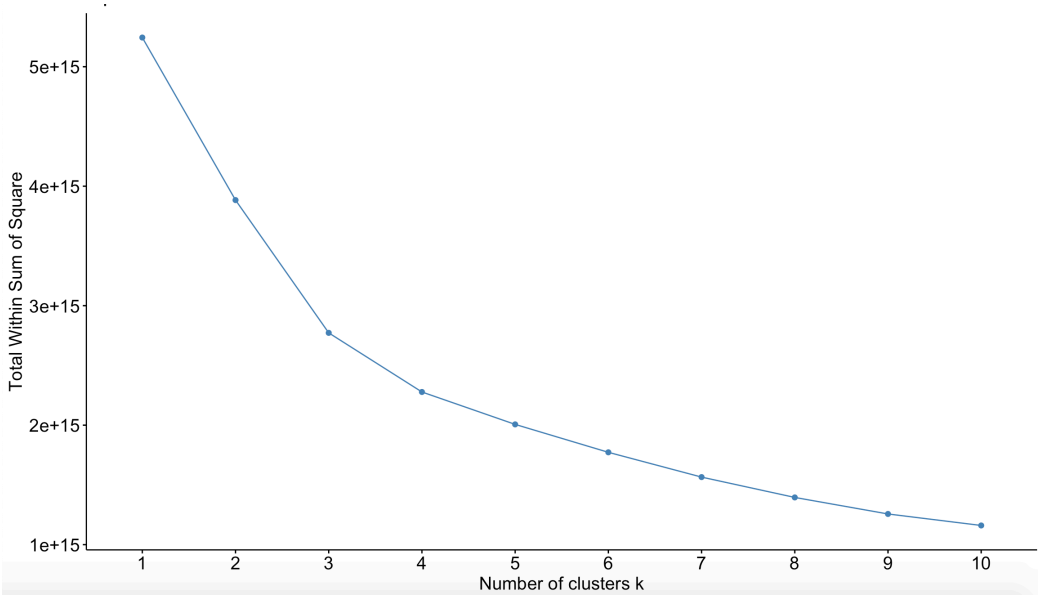

PC1
PC2
PC3
PC4
PC5
PC6
PC7
PC8
PC9
PC10
PC11
PC12
PC13
PC14
PC15
PC16
IDH1
FH
SDHB
GLUL
GAP43
WWTR1
OLIG2
KEAP1
TTYH1
SLC1A5
YWHAG
IDH2
NANOG
GJA1
JMJD8
OGDH
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

Fig. S2. Multidimensional exploratory data analysis of expression values of 16 selected transcripts in 151 glioblastoma samples. **(A)** Scree plot display the proportion of the total variance of the data that is explained by each component obtained during Principal component analysis (PCA). **(B)** PCA of 16 transcripts in 151 glioblastoma samples. First two principal components (PC) account for 83.3% of variance. GJA1 has large negative correlation on component 1, while GAP43 and GLUL have large positive loading on component 2. **(C)** Heatmap of 151 glioblastoma samples based on 16 transcripts, correlation matrix was considered for this purpose. Strong positive and negative correlations are represented by blue and red colores ellipses, respectively, whereas weak correlations are shown as faded and faded circles. **(D)** Heatmap between PC analysis and transcripts. Higher correlation values between transcripts and PC are displayed in blue, and the lower gradually fading toward red, accordingly to the strength of their correlation.

(A)



(B)

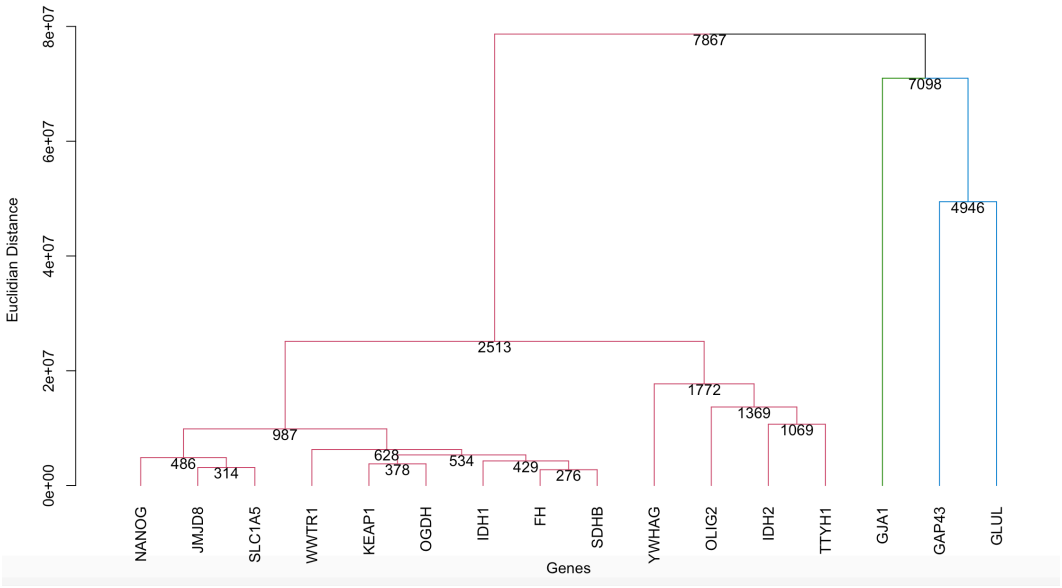
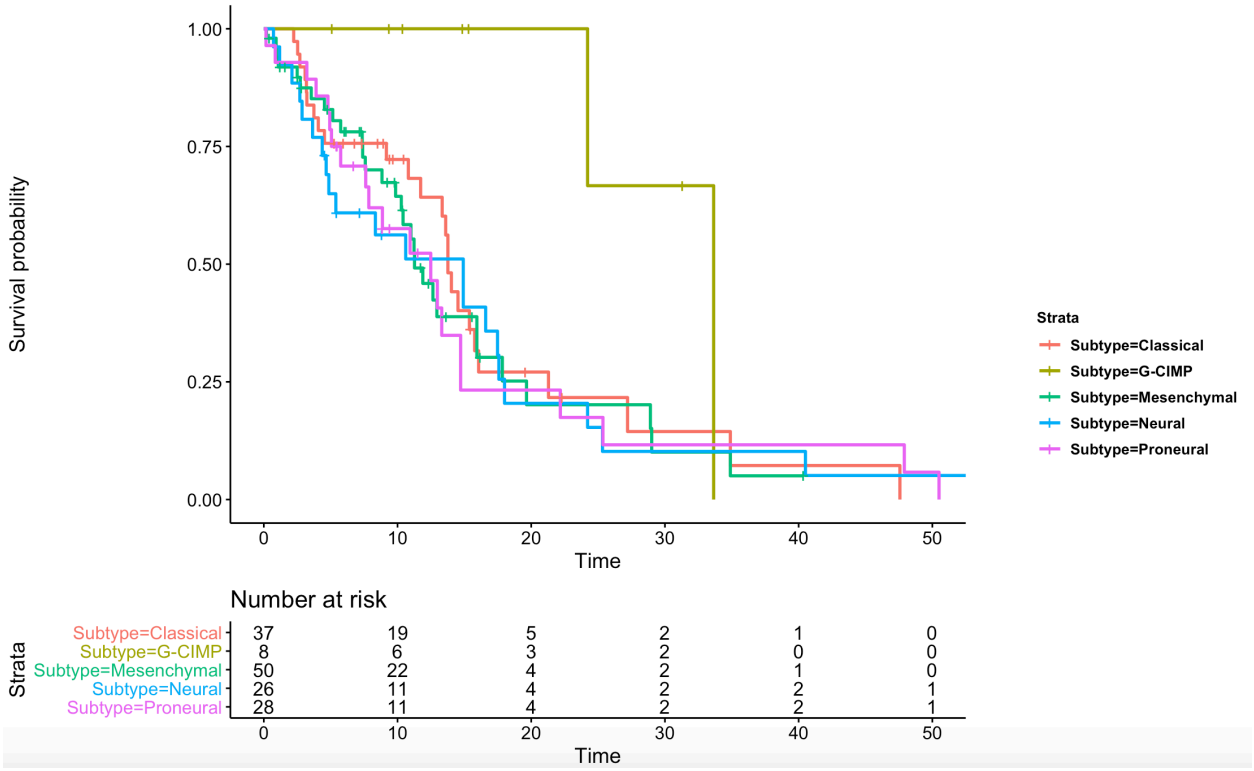


Fig. S3. Dendrogram of 16 selected transcripts using 151 samples of glioblastoma. **(A)** Elbow method to determine optimal number of clusters. Axis-X shows number of clusters k, while axis-Y shows total within-group variance **(B)** Hierarchical agglomerative clustering using dendrogram structure in a p-variable space showing 3 major clusters. Axis-Y shows scaled (values shown are 0.001 times the actual Euclidean distance) height obtained for either individual data points or clusters.



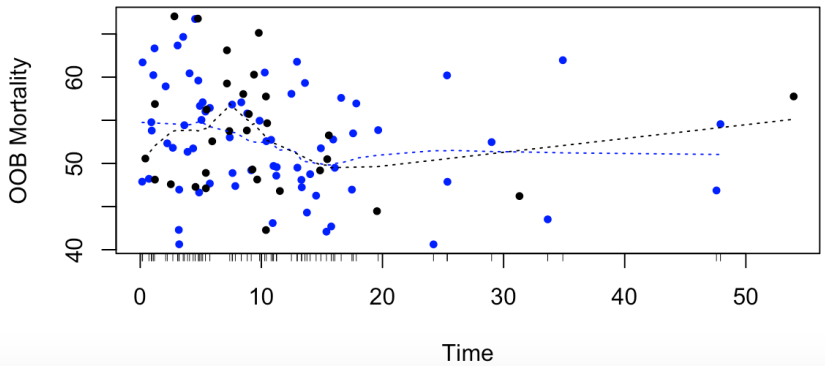
	Classical	Neural	Proneural	Mesenchymal	G-CIM
Death	12	5	7	20	6
Censored	25	21	21	30	2
Percentage of censoring	67.6%	19.2%	25.0%	40.0%	25.0%

Fig. S4. Kaplan-Meier curves from 151 glioblastoma patients. In colors, curves of survival for gene expression-based tumor subtype. Below, a risk table displaying the number of subjects at risk every 10 months by tumor-subtype. Also underneath, a table showing the frecuency of death and censored data within each subtype.

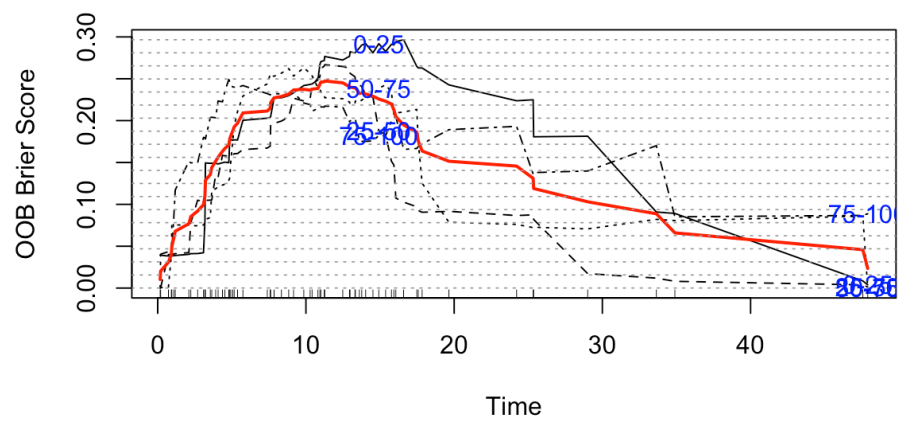
5.4.2 MODELS FOR SURVIVAL PREDICTION

In order to achieve reproducibility in our results, we set a seed across our pipeline. RSF error rate showed to stabilize around 5,000 trees. In order to analize the addition of clinical predictors to our model, three different input subsets were used. The first one included only our selected transcripts and survival information; the second input set, contained the selected transcripts plus other relevant clinical variables; and the third input set, contained only the clinical predictors last added. Training showed a high error rate for input set 1 (47.86%), following a slightly better performance when clinical data was included (43.07%). Clinical data only demonstrated an error rate of 42.37%. After, testing the data over our fitted model, input sets 1 and 2 got higher prediction error (62.29%; 47.69%), while input set 3 kept its error close to the one obtained during training (43.35%). In order to assess the performance of our trained model, we used out-of-bag (OOB) Brier score and predicted mortality, both displayed in **Fig. S5.A-B**, showing better time-dependent score in 25-50 and 75-100 quantiles of OOB ensemble mortality. Also, OOB survival showed good curve fitting by Nelson-Aalen estimator with ensemble survival function for each individual giving a good idea of the survival shape for GB survival outcome (**Fig. S5.C**). Curves of survival outcome and predicted outcome in test set are depicted in **Fig.S6.** The median survival for the predicted outcome reflects higher confidence bands for the subtype groups in colors.

(A)



(B)



(C)

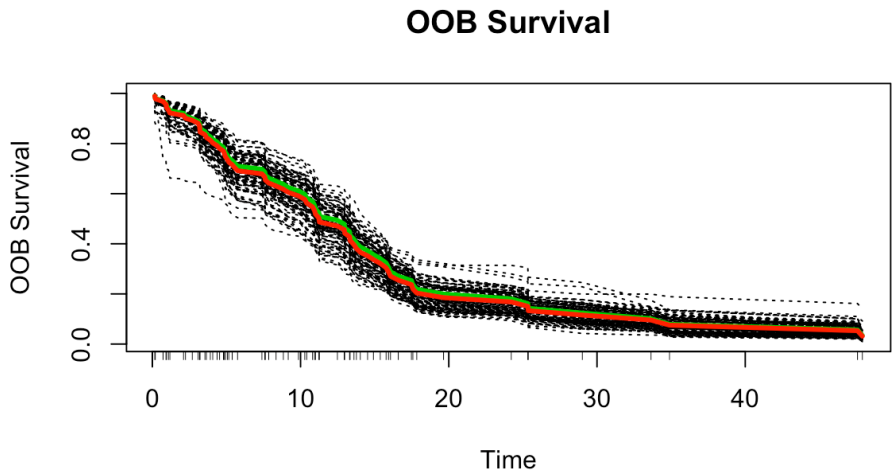
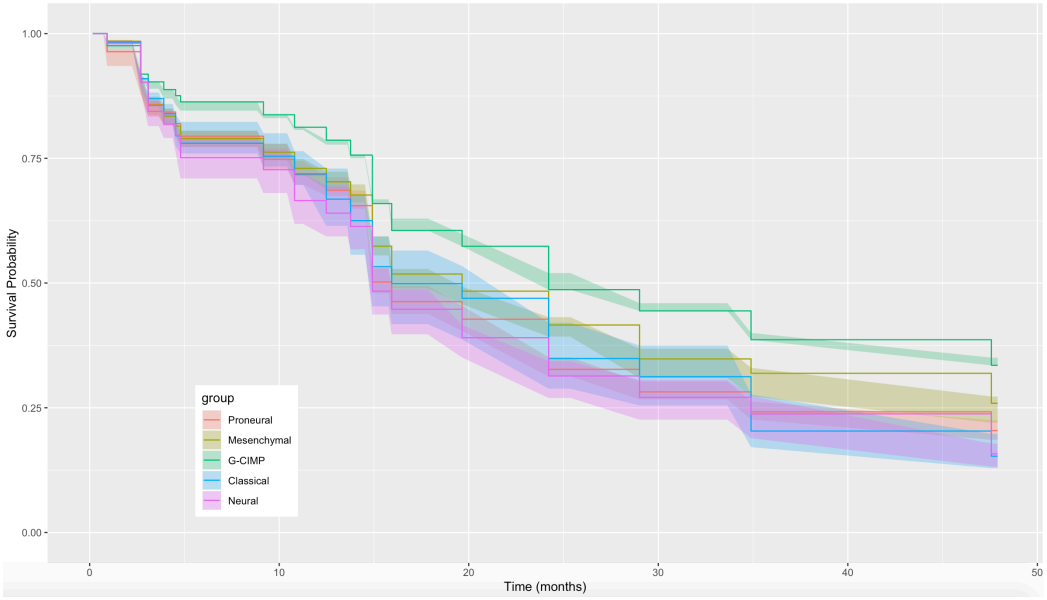


Fig. S5. (A) Mortality of each individual versus observed time. Points in blue and black correspond to event and censoring times, respectively. **(B)** Brier score stratified as 0-25, 25-50, 50-75 and 75-100 percentile values of ensemble mortality, being the Brier score 0=perfect, 1=poor, and 0.25=guessing). Model reflect better time-dependent score during exteme time points of study. **(C)** Random forest estimated survival function for each individual; thick red line represents overall ensemble survival; thick green line represents

Nelson-Aalen estimator.

(A)



(B)

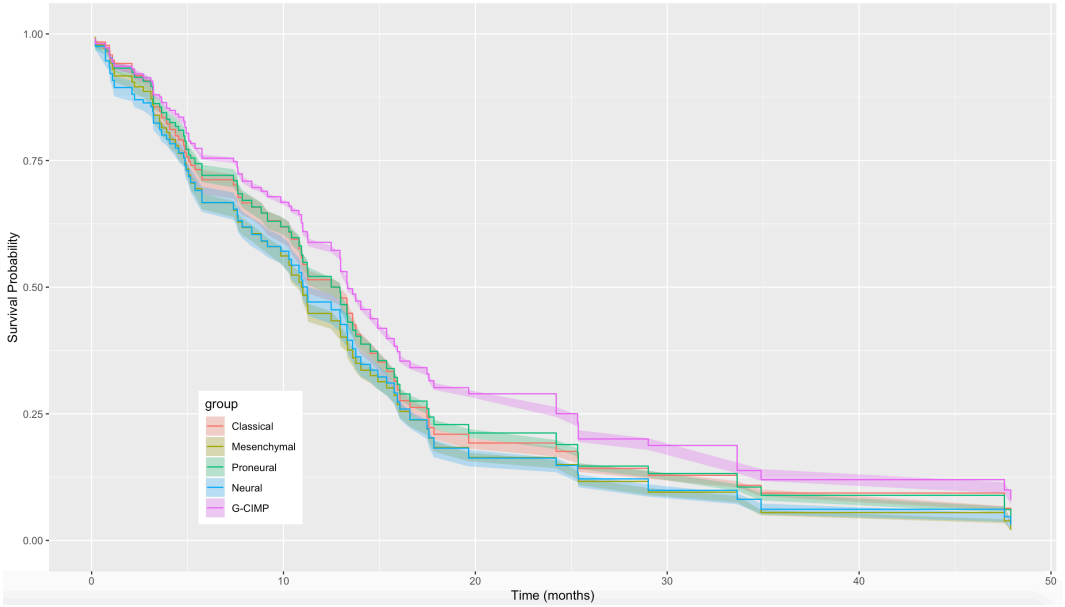
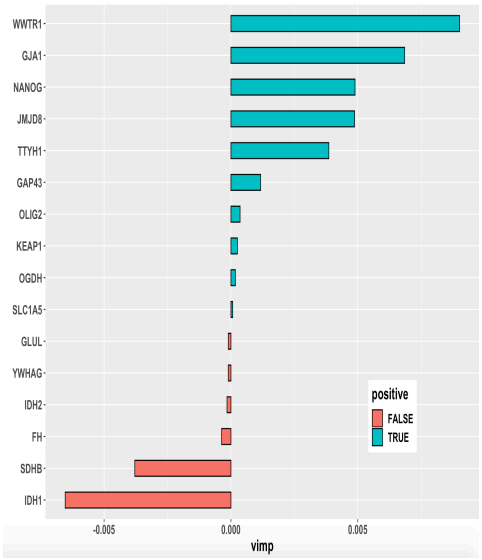


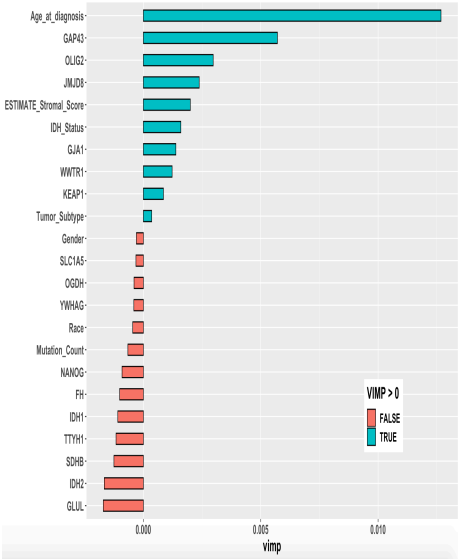
Fig.S6. Survival time by tumor subtype. **(A)** Survival outcome depicted as the median survival with a 95%

bootstrap shaded confidence bands around the median survival line, according tumor subtype groups. **(B)** Predicted survival outcome depicted as the median survival with a 95% bootstrap confidence bands around the median survival line, according tumor subtype groups. Horizontal axis reflects time points in months, while vertical axis shows the survival probability.

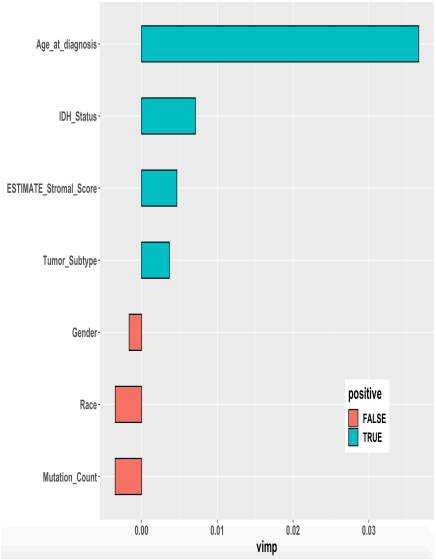
(A)



(B)



(C)



(D)

	VIMP	Relative VIMP
Age at diagnosis	0.0127	1.0000
GAP43	0.0057	0.4504
OLIG2	0.0030	0.2346
JMJD8	0.0024	0.1876
Stromal	0.0020	0.1578
IDH Status	0.0016	0.1258
GJA1	0.0014	0.1092
WWTR1	0.0012	0.0966
KEAP1	0.0009	0.0674
Tumor subtype	0.0004	0.0279
Gender	-0.0003	-0.0234
SLC1A5	-0.0003	-0.0250
OGDH	-0.0004	-0.0314
YWHAG	-0.0004	-0.0321
Race	-0.0005	-0.0359
Mutation count	-0.0007	-0.0521
NANOG	-0.0009	-0.0717
FH	-0.0010	-0.0797

IDH1	-0.0011	-0.0860
TTYH1	-0.0012	-0.0919
SDHB	-0.013	-0.0990
IDH2	-0.017	-0.1313
GLUL	-0.017	-0.1347

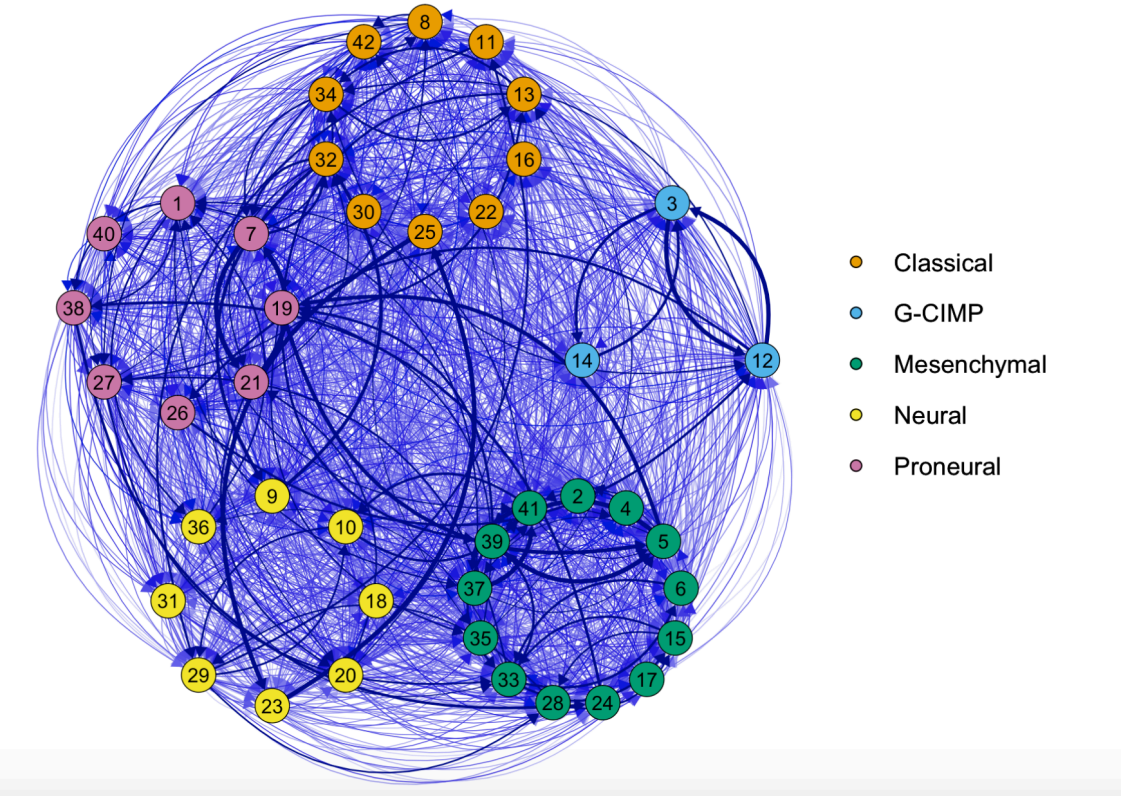
Fig. S7. Comparative information of input sets 1, 2, and 3 used in RSF. **(A) – (C)** Figures exhibit variable importance (VIMP) values; in turqoise, positive VIMP values are displayed; in salmon, negative or zero VIMP values can be displayed. **(D)** The following table presents VIMP and relative VIMP outcome values for all variables within input set 2. Age at diagnosis corresponds to the highest VIMP with 0.0127 and GLUL to the least important with -0.017.

Variable importance (VIMP) was also tested in the last three input sets (**Fig. S7.A-C**), reporting hold out VIMP and relative VIMP for clinical information and our transcripts in input set 2 (**Fig. S7.D**). Larger importance values for transcripts GAP43, OLIG2, JMJD8, GJA1, WWTR1, and KEAP1 are reported, also for clinical variables Age at diagnosis, Stromal score, IDH status and Tumor subtype. While the least important variables corresponded with GLUL, IDH2 and, SDHB reflecting zero or negative values. Further analyses, such as interactions, were not included for discussion considering that changes in VIMP values were irrelevant when running an interaction analysis using the VIMP methodology.

In order to study the similarity between our test observations, we used the symmetric proximity matrix to construct a weighted and directed graph (**Fig.S8**), visualizing (statistical) relationships between variables as weighted edges.[38] In this graph (**Fig.S8.A**), stronger relationships between observations appeared between classical and neural subtypes and mesenchymal and proneural subtypes. G-CIMP also showed similar relationships but just within his group observations, although considerations regarding

sample size should also be considered for this subtype. Regarding the direction of wider edges, mesenchymal subtype mostly depicted his edges towards proneural subtype and viceversa, whereas classical subtype depited his edges towards neural subtype and viceversa. Also classical and proneural subtypes seemed to have directions between their nodes. Shortests paths (**Fig.S8.B**) also coincide with the width of the edges in **Fig.S8.A**.

(A)



(B)

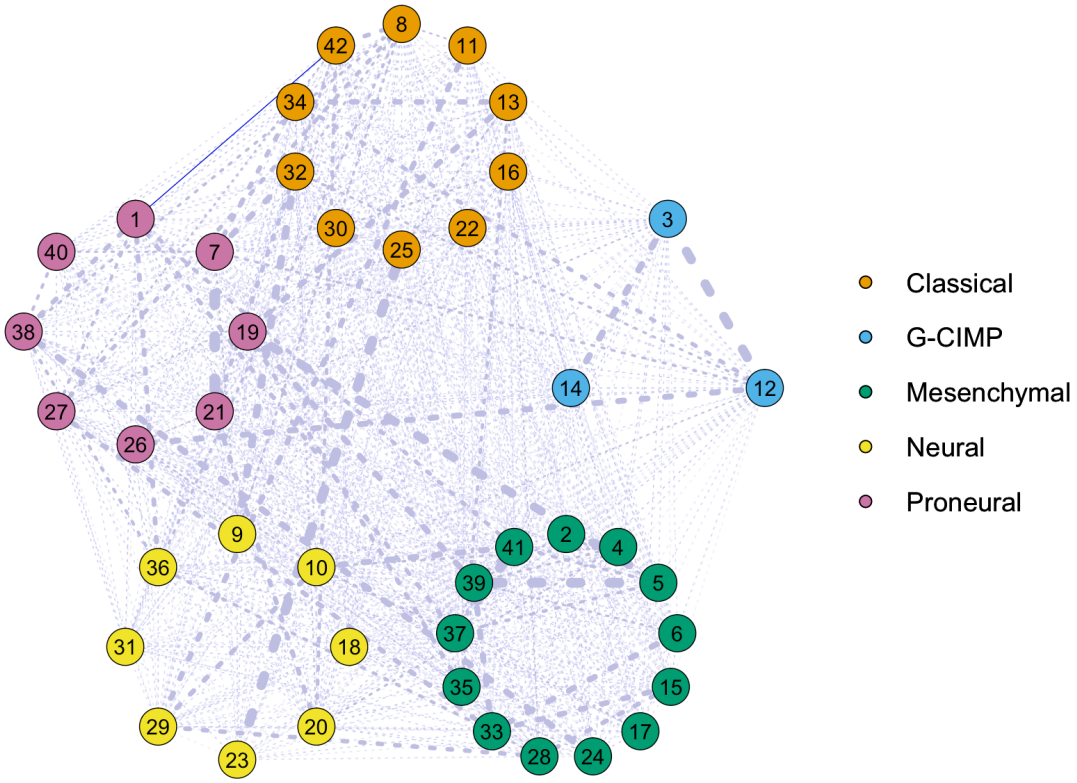


Fig.S8. Network graph showcasing connections between samples in the test set. **(A)** Directed-weighted network graph depicting similarity among test set observations. Tumor subtypes are displayed in five different colors. Placement of nodes is determined by a circular-group layout and direction between test observations is showed in arrows. Wider lines corresponds to the absolute weight and scale relative to the strongest weight in the graph. **(B)** Highlight of the shortests paths between nodes considering in the whole network.

For semi-parametric CPH model, there is no evidence against the proportional hazards assumption based on the absence of a significant relationship between Schoenfeld residuals and time. CPH models used imputed gene expression data from adaptitve tree imputation. These analyses showed that gene expression information by its own was not significant overall in the model (LR=23.72; df=16; p=0.095) and addition of clinical information improved overall assessment of the model (LR=58.01;df=27;p=0.0004), being variables race white (Z=-2.392; p=0.016), years at diagnosis (Z=3.129; p=0.0017), mutation count (Z=-2.003; p=0.045), GLUL (Z=-2.856; p=0.0042), GAP 43 (Z=2.081; p=0.037) and OLIG2 (Z=2.056;P=0.039) significant within the model. For clinical data only, the model showed overall significance (LR=36.7; df=11; p=0.0001), and variables race white (Z=-2.106; p=0.035), age at diagnosis (Z=3.079; p=6.92e-05), and mutation count (Z=-2.409); p=0.0160) were also significant within the model. In order to validate CPH models performed in our datasets, we used Harrell concordance index obtaining non significative C-Index for data set 1 (p=0.19) with 42.8% (s.e.= 0.05; C.I.= [0.32;0.53]) and for data set 2 (p=0.51) with 54.6% (s.e.= 0.06; C.I.= [0.41;0.68]). And significant (p= 0.004) C-Index for data set 3 with 66.7% (s.e.= 0.06; C.I.=0.55;0.78]).

5.5 DISCUSSION

The study of markers in gliomas throughout basic and clinical sciences has proved to have a marginal effect in patient's survival. [25,60] So far, good prognostic factors haven been usually related to tumor grade and patients' age.[13] It is not surprising that highly promising genes remain ignored due to large-scale information gathered today, experimental biases, lack of reproducible robust results, and the increase of data processing technologies, and by that the overwhelming cumulative methodology behind appropriate analysis. Most studies usually look for biological insight of cancer markers through genome-wide association studies [7,15,61–63], which implies that down regulated genes usually are misrepresented due to the difficulty of capturing signals from noise in downstream analyses.

Our gene candidates represent a sparse subpopulation of cell-to-cell communication structures, and we believe that tackling the biology behind their encoding, we are going to be able to understand the mechanisms of resistance to treatment and to better assess their potential predictive power. In this study, we presented a first approach to introduce our gene proposal of tumor resistance and to assess their relevance in bulk-tumor data. We could verify through descriptive analysis that our transcripts dilucidate a hierarchical structure. Our resulting dendrogram used 3 clusters and the outcome structure was confirmed through CCC. For most of our transcripts, this structure suggested non other than similarity among a mixed group of biological processes, that is, differentiation-metabolism-methylation-TNT. Yet, third cluster only considered action of TNT-Metabolism processes. This last, reflects an important relationship between biologically involved

predictors GLUL and GAP43. This is an imprortan findind among our transcripts considering the metabolic adaptative mechanisms that TNT formation require.

Correlation between our genes and principal components, PC1 and PC2, which accounted for most of the variability of our data, showed that GJA1 had a large positive correlation on component 1, while GAP43 and GLUL had a large positive correlation on component 2. This suggested that there is an effect behind the expression of TNT mediation mechanisms in the dimension with the largest variance out of the overall variance in our data. And secondly, transcript GAP43 had also an important role with PC2, reflecting that glioma cell growth suppression (i.e. increasing infiltrative growth)[64] is relevant among the overall variance in our data.. Regarding GLUL transcript, many of our genes participate in the tricarboxylic acid cycle, thus it is relevant to continue exploring how metabolic routes are mediating survival outcome such as, glutamate transport. Again, it seems like TNT-Metabolism group takes an important role within our data.

Regarding binary tree analysis using RSF, OOB Brier score against time points reflected better performance of RSF in 25-50 and 75-100 quantiles of OOB ensemble mortality. This could reflect that for participats with higher or lower expected mortality (i.e. higher or lower risk of event) our model, based on genetic and clinical variables, was capable to better predict survival. Yet, considering high prediction error (47.69%), our model was not able to predict patients' survival; using our gene candidates plus clinical data. This could be due to signal-to-noise limitations of transcriptomic data or due to non relevant predictive power of our transcripts in patient's survival. This could only be discover through a different methodological approach which will be discussed below. We rescue the relevant

result that clinical data improved the performance of the model because, when testing over our fitted model, dataset 1 and 2 got higher error (62.29%; 47.69%), while dataset 3 kept its error close to the one obtained during training (43.35%). It seemed that using exclusively clinical data, our model had a better performance than when adding our gene candidates. Thus, by adding our transcripts as predictors, we could have overfitted the model unnecessarily, obtaining worse generalizability than when using only clinical data.

Also, as expeted, RSF test set predictions reflected poorer performance (higher prediction error) than the one obtained using OOB data during training. Remember that prediction error measures how well the predictors correctly rank (classify) two random individuals in terms of survival, being 0 perfect and 0.5 no better than random guessing.

Trained CPH models of datasets that included clinical variables, were significant, while dataset 1 which included only transcript information was not significant. After, using test data, validation of our predictions showed higher C-Index for datasets that included clinical variables. This last is good because, although not significant for dataset 1 and 2, it reflects that if we randomly select pairs of individuals with and without the event, the 4-year (our study gets up to 54 months) probability of death outcome due to GB disease, estimated by risk tables, will be higher ("C-Index" percent of times) in individuals who have died than in those who have not died yet of GB. More precisely, the C-Index represents the proportion of pairs of subjects (with opposite outcome, i.e. who has and who has not died of GB), where the one who actually experiences the adverse outcome (i.e. death), GB, has a higher (predicted) probability of event. Only clinical data showed to be significant, considering

95% confidence interval did not include 0.5. This shows that clinical information is relevant by itself to predict patients' survival.

Our network graphs showed us the relevance of connections between some of the tumor subtypes. We highlight the connections between the mesenchymal and proneural subtypes. This last similarity could reflect the relationship of proneural-mesenchymal transition equivalent of the epithelial-mesenchymal transition associated with the most aggressive cancers.[65] There was also stronger connections between neural and classical subtypes, yet it is important that we consider how recently[18] neural subtype could possibly correspond to a contamination of the original samples with non-tumor cells.

Using only transcriptomic data for such polygenic processes makes prediction a difficult task. In this work, we have presented an exploration of novel genes and we expect to continue adding polygenic experiment schemes that allow us to unveil the true role these genes accomplish in patients' outcome. Although GB characterization becomes difficult considering important intra tumor heterogeneity. We need to connect with the reality of facing new methodologies to tackle highly promising genes unrepresented in genome wide sequencing, which by random fluctuations in signal from noise could be presented as inaccessible relevant and biologically fundamented predictors.

Using longitudinal data of GB could be the key to understand how therapy resistance works in this disease [18] and it has been proved in other cancer types that by adding molecular signatures to well-established conventional prognostic markers could add predicitive value to them. [66] So far insufficient understanding of the biology has been a limitation for the success of novel therapies, because preclinical research has been focused

mostly in experimentally well accessible genes. [67] Thus, including biologically fundamented genes, we could better predict and understand the associations within our data.[68]

In contrast with Cox-PH, RSF allowed us to impute the data, manage a high dimensional space of predictors and to approximate non-linear functions due to the average over the trees plus the randomization used in growing these trees. This model did not require to assess any type of previous statistical assumption within the data nor we had the risk of misspecifying the model. On the contrary, Cox-PH, although a semi-parametric model, it could lead to poor predictions when the model is misspecified in terms of the effect of the chosen predictors because this model still relies on linear relationships when making inference on the coefficients, which is not always a realistic assumption to be made.[69] In addition, Cox-PH required a complete data set as input, which also had to be address through RSF via ATI.

RSF is a good model for predicting survival,[40,41,51,70] thus continue using this model we can compare our results and reproduce them in the future. It is of our interest to also understand the rationale behind the model's predictions in order to contribute to better understanding of cancer mechanisms. For this, adding other methods (e.g. Local Interpretable Model-Agnostic Explanations, LIME)[39] that explain such predictions would be highly valuable for assisting clinicians at the patient's bedside. This type of methodology[71] could help us understand the predictions of our fitted models in a better way or give us an idea on how to quantify and assess each subpopulation of cells in GB, so we can eventually disentangle the signals in each one of them.

5.6 CONCLUSIONS

Previous experimental findings have presented a remarkable coincidence of genes in small cell population measures. Their involvement in metabolic routes and cell-to-cell mechanisms dilucidate a major discovery of potential mechanism for tumor resistance. By only using transcriptomic data we have showed that current practice for addressing this fatal disease it is not enough, due to poor representation of cellular sub populations of interest within the abundance of highly represented proliferative or differentiative processes in the ongoing development of GB. Yet, our genes are of interest because of biological reasons. Thus it is important that we continue studying them and to continue to enrich them using different sources of justified information such as, metabolomic data available due to the role many od these genes accomplish in metabolic routes. By giving context to these genes through more genetic information, we could be able to assess the relevance they have in GB outcome.

6. BIBLIOGRAFÍA

1. Urbanska K, Sokolowska J, Szmidt M, Sysa P. Glioblastoma multiforme - An overview [Internet]. Vol. 18, Wspolczesna Onkologia. 2014. p. 307–12. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4248049/

2. Ostrom QT, Bauchet L, Davis FG, Deltour I, Fisher JL, Langer CE, et al. The epidemiology of glioma in adults: A state of the science review. Neuro Oncol. 2014;16(7):896–913.

3. Thakkar JP, Dolecek TA, Horbinski C, Ostrom QT, Lightner DD, Barnholtz-Sloan JS, et al. Epidemiologic and molecular prognostic review of glioblastoma. Cancer Epidemiol Biomarkers Prev [Internet]. 2014;23(10):1985–96. Available from: https://www.ncbi.nlm.nih.gov/pubmed/25053711

4. Ariazi J, Benowitz A, De Biasi V, Den Boer ML, Cherqui S, Cui H, et al. Tunneling Nanotubes and Gap Junctions–Their Role in Long-Range Intercellular Communication during Development, Health, and Disease Conditions. Front Mol Neurosci [Internet]. 2017;10(October):1–12. Available from: http://journal.frontiersin.org/article/10.3389/fnmol.2017.00333/full

5. Valdebenito S. The Novel Roles of Connexin Channels and Tunneling Nanotubes in Cancer Pathogenesis. Int J Mol Sci. 2018;19.

6. Osswald M, von Deimling A, Weil S, Jung E, Horstmann H, Häring P, et al. Brain tumour cells interconnect to a functional and resistant network. Nature [Internet]. 2015;528(7580):93–8. Available from: https://www.nature.com/articles/nature16071

7. Colman H, Zhang L, Sulman EP, McDonald JM, Shooshtari NL, Rivera A, et al. A multigene predictor of outcome in glioblastoma. Neuro Oncol [Internet]. 2010;12(1):49–57. Available from: https://www.ncbi.nlm.nih.gov/pubmed/20150367

8. Tripepi G, Jager KJ, Dekker FW, Zoccali C. Statistical methods for the assessment of prognostic

biomarkers (Part I): Discrimination. Nephrol Dial Transplant. 2010;25(5):1399–401.

9. Louis DN, Ohgaki H, Wiestler OD, Cavenee WK, Burger PC, Jouvet A, et al. The 2007 WHO classification of tumours of the central nervous system. Acta Neuropathol. 2007;114(2):97–109.

10. Wick W, Kessler T. New glioblastoma heterogeneity atlas — a shared resource. Nat Rev Neurol [Internet]. 2018;1–2. Available from: http://dx.doi.org/10.1038/s41582-018-0038-3

11. Behnan J, Finocchiaro G, Hanna G. The landscape of the mesenchymal signature in brain tumours. Brain [Internet]. 2019;142(4):847–66. Available from: https://www.ncbi.nlm.nih.gov/pubmed/30946477

12. Verhaak RGW, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, et al. An Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell [Internet]. 2010;17(1):98–110. Available from: https://www.ncbi.nlm.nih.gov/pubmed/20129251

13. Phillips HS, Kharbanda S, Chen R, Forrest WF, Soriano RH, Wu TD, et al. Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. Cancer Cell [Internet]. 2006;9(3):157–73. Available from: https://www.ncbi.nlm.nih.gov/pubmed/16530701

14. Wang Q, Hu X, Muller F, Kim H, Squatrito M, Mikkelsen T, et al. Tumor Evolution of Glioma Intrinsic Gene Expression Subtype Associates With Immunological Changes in the Microenvironment. Neuro Oncol [Internet]. 2016;18(suppl_6):vi202–vi202. Available from: https://www.ncbi.nlm.nih.gov/pubmed/28697342

15. Brennan CW, Verhaak RGW, McKenna A, Campos B, Noushmehr H, Salama SR, et al. The Somatic Genomic Landscape of Glioblastoma. Cell [Internet]. 2013;155(2):462–77. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3910500/

16. Malta TM, De Souza CF, Sabedot TS, Silva TC, Mosella MS, Kalkanis SN, et al. Glioma CpG island

methylator phenotype (G-CIMP): Biological and clinical implications. Neuro Oncol [Internet]. 2018;20(5):608–20. Available from: https://www.ncbi.nlm.nih.gov/pubmed/29036500

17. Teo WY, Sekar K, Seshachalam P, Shen J, Chow WY, Lau CC, et al. Relevance of a TCGA-derived Glioblastoma Subtype Gene-Classifier among Patient Populations. Sci Rep [Internet]. 2019;9(1):1–10. Available from: https://www.nature.com/articles/s41598-019-43173-y

18. Sidaway P. CNS cancer: Glioblastoma subtypes revisited. Nat Rev Clin Oncol [Internet]. 2017;14(10):587. Available from: http://dx.doi.org/10.1038/nrclinonc.2017.122

19. Lathia JD, Mack SC, Mulkearns-Hubert EE, Valentim CLL, Rich JN. Cancer stem cells in glioblastoma. Genes Dev [Internet]. 2015;29:1203–17. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4495393/

20. William A. Freije; F. Edmundo Castro-Vargas; Zixing Fang; Steve Horvath; Timothy Cloughesy; Linda M. Liau; Paul S. Mischel; and Stanley F. Nelson. Gene Expression Profiling of Gliomas Strongly Predicts Survival. Cancer Res. 2005;64(18):6503–10.

21. Haas-Kogan DA, Prados MD, Lamborn KR, Tihan T, Berger MS, Stokoe D. Biomarkers to predict response to epidermal growth factor receptor inhibitors. Cell Cycle. 2005;4(10):1369–72.

22. Colman H, Chen M, Nigro JM, Ozburn N, Pan E, Feuerstein BG, et al. Integrated Array-Comparative Genomic Hybridization and Expression Array Profiles Identify Clinically Relevant Molecular Subtypes of Glioblastoma. Cancer Res. 2005;65(5):1678–86.

23. Zuo S, Zhang X, Wang L. A RNA sequencing-based six-gene signature for survival prediction in patients with glioblastoma. Sci Rep [Internet]. 2019;9(1):2615. Available from: http://www.nature.com/articles/s41598-019-39273-4

24. Hu N, Cheng H, Zhang K, Jensen R. Evaluating the Prognostic Accuracy of Biomarkers for Glioblastoma Multiforme Using The Cancer Genome Atlas Data. Cancer Inform. 2017;16.

25. Ludwig K, Kornblum HI, Behavior H, Angeles L, Pharmacology M, Angeles L. Molecular Markers in Glioma. 2018;134(3):505–12.

26. Nowell PC. The clonal evolution of tumor cell populations. 1976;194:23–2.

27. Bonavia R, Inda MDM, Cavenee WK, Furnari FB. Heterogeneity maintenance in glioblastoma: A social network. Cancer Res. 2011;71(12):4055–60.

28. Bhat KPL, Salazar KL, Balasubramaniyan V, Wani K, Heathcock L, Hollingsworth F, et al. The transcriptional coactivator TAZ regulates mesenchymal differentiation in malignant glioma. Genes Dev [Internet]. 2011;25(24):2594–609. Available from: https://www.ncbi.nlm.nih.gov/pubmed/22190458

29. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A. 2001;98(9):5116–21.

30. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet [Internet]. 2009;10(1):57–63. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2949280&tool=pmcentrez&rendertype=abstract

31. Chu M-KM. Statistical methods for the analysis of RNA sequencing data [Internet]. Electronic Thesis and Dissertation Repository. 2014. Available from: http://ir.lib.uwo.ca/etd/1935

32. Baghfalaki T, Ganjali M, Berridge D. Missing Value Imputation for RNA-Sequencing Data Using Statistical Models: A Comparative Study. J Stat Theory Appl [Internet]. 2016;15(3):221. Available from: https://www.atlantis-press.com/journals/jsta/25862105

33. Dündar F, Skrabanek L, Zumbo P. Introduction to differential gene expression analysis using RNA-seq [Internet]. Applied Bioinformatics Core. 2018. 1–86 p. Available from: http://chagall.med.cornell.edu/RNASEQcourse/Intro2RNAseq.pdf

34. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. Data Mining, Inference, and Prediction [Internet]. Second edi. New York, NY: Springer US; 2017. 764 p. Available from: https://web.stanford.edu/~hastie/ElemStatLearn/

35. Wikipedia. Cophenetic correlation [Internet]. 2019. Available from: https://en.wikipedia.org/wiki/Cophenetic_correlation

36. Dattorro J. Euclidean Distance Matrix. Convex Optim Euclidian Distance Geom [Internet]. 2001;(886):385–485. Available from: https://ccrma.stanford.edu/~dattorro/EDM.pdf

37. Wikipedia. Euclidian distance. 2019;

38. Epskamp S, Giulio C, Jonas H, Adela I, Angelique O J C, Denny B. Package ' qgraph ' [Internet]. 2019. p. 63. Available from: https://cran.r-project.org/web/packages/qgraph/qgraph.pdf

39. Epskamp S. Descriptive Analysis of Network Graph Characteristics. 2014. p. 43–67.

40. Tang F, Ishwaran H. Random forest missing data algorithms. Stat Anal Data Min. 2017;10(6):363–77.

41. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. Ann Appl Stat [Internet]. 2008;2(3):841–60. Available from: https://arxiv.org/pdf/0811.1645.pdf

42. Breiman L. Statistical Modeling: The Two Cultures. Stat Sci. 2001;16(3):199–231.

43. Beam AL, Kohane IS. Big Data and Machine Learning in Health Care. J Am Med Assoc [Internet]. 2018;319(13):1317–8. Available from: https://www.ncbi.nlm.nih.gov/pubmed/29532063

44. Korfiatis P, Kline TL, Coufalova L, Lachance DH, Parney IF, Carter RE, et al. MRI texture features as biomarkers to predict MGMT methylation status in glioblastomas. Med Phys. 2016;43(6):2835–44.

45. Gollapalli, Kishore. Ray, Sandipan. Srivastava R et al. Investigation of serum proteome alterations in human glioblastoma multiforme. Proteomics [Internet]. 2012;12:2378–2390. Available from: https://www.ncbi.nlm.nih.gov/pubmed/22684992

46. Chang K, Zhang B, Guo X, Zong M, Rahman R, Sanchez D, et al. Multimodal imaging patterns predict survival in recurrent glioblastoma patients treated with bevacizumab. Neuro Oncol [Internet]. 2016;18(12):1680–7. Available from: https://www.ncbi.nlm.nih.gov/pubmed/27257279

47. Kickingereder P, Bonekamp D, Nowosielski M, Kratz A, Sill M, Burth S, et al. Radiogenomics of glioblastoma: Machine Learning–based Classification of Molecular Characteristics by Using Multiparametric and Multiregional MR Imaging Features. Radiology [Internet]. 2016;281(3):907–18. Available from: https://pubs.rsna.org/doi/10.1148/radiol.2016161382

48. Kleinbaum DG, Klein Mitchel. Survival Analysis [Internet]. Vol. 36, Statistics for Biology and Health David. 2011. 712 p. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24815723

49. Shmueli G. To Explain or To Predict? Stat Sci [Internet]. 2010;25(3):289–310. Available from: https://www.stat.berkeley.edu/~aldous/157/Papers/shmueli.pdf

50. Janitza S, Hornung R. On the overestimation of random forest's out-of-bag error. PLoS One. 2018;13(8):1–31.

51. Mogensen UB, Ishwaran H, Gerds TA. Evaluating Random Forests for Survival Analysis Using Prediction Error Curves. J Stat Softw. 2012;50(11):1–23.

52. Graf E, Schmoor C, Sauerbrei W, Schumacher M. Assessment and comparison of prognostic classification schemes for survival data. Stat Med. 1999;18(1718):2529–45.

53. Klein J, Moeschberger M. Survival analysis: Techniques for censored and truncated data. Vol. 2, Springer. 2003. 542 p.

54. Yoshihara K, Shahmoradgoli M, Martínez E, Vegesna R, Kim H, Torres-Garcia W, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. Nat Commun. 2013;4.

55. Reeb PD, Bramardi SJ, Steibel JP. Assessing dissimilarity measures for sample-based hierarchical clustering of RNA sequencing data using plasmode datasets. PLoS One. 2015;10(7):1–18.

56. Hanif F, Muzaffar K, Perveen K, Malhi SM, Simjee SU. Glioblastoma Multiforme: A Review of its Epidemiology and Pathogenesis through Clinical Presentation and Treatment. Asian Pacific J Cancer Prev [Internet]. 2017;18:1–9. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5563115/

57. Malmström A, Grønberg BH, Marosi C, Stupp R, Frappaz D, Schultz H, et al. Temozolomide versus standard 6-week radiotherapy versus hypofractionated radiotherapy in patients older than 60 years with glioblastoma : the Nordic randomised , phase 3 trial. Lancet Oncol [Internet]. 2012;13(9):916–26. Available from: http://dx.doi.org/10.1016/S1470-2045(12)70265-6

58. Cohen MH, Johnson JR, Pazdur R. Food and drug administration drug approval summary: Temozolomide plus radiation therapy for the treatment of newly diagnosed glioblastoma multiforme. Clin Cancer Res [Internet]. 2005;11(19 I):6767–71. Available from: https://clincancerres.aacrjournals.org/content/11/19/6767

59. Ishwaran H, Kogalur UB. Package "randomForestSRC" [Internet]. 2019. p. 97. Available from: https://github.com/kogalur/randomForestSRC

60. Freije WA, Castro-Vargas FE, Fang Z, Horvath S, Cloughesy T, Liau LM, et al. Gene expression profiling of gliomas strongly predicts survival. Cancer Res [Internet]. 2004;64(18):6503–10. Available from: https://www.ncbi.nlm.nih.gov/pubmed/15374961

61. Tang J, He D, Yang P, He J, Zhang Y. Genome-wide expression profiling of glioblastoma using a large combined cohort. Sci Rep [Internet]. 2018;8(1):1–12. Available from: https://www.nature.com/articles/s41598-018-33323-z

62. López P, López L. Gene-Expression Profiling in Pancreatic Cancer. 2010;10(5):591–601. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5013537/

63. Van't Veer LJ, Dai H, Van de Vijver MJ, He YD, Hart AAM, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature [Internet]. 2002;415(6871):530–6. Available from:

https://www.ncbi.nlm.nih.gov/pubmed/11823860

64. Osswald M, Jung E, Wick W, Winkler F. Tunneling nanotube-like structures in brain tumors. Cancer Rep [Internet]. 2019;(December 2018):1–7. Available from: https://onlinelibrary.wiley.com/doi/full/10.1002/cnr2.1181

65. Fedele M, Cerchia L, Pegoraro S, Sgarra R, Manfioletti G. Proneural-mesenchymal transition: Phenotypic plasticity to acquire multitherapy resistance in glioblastoma. Int J Mol Sci [Internet]. 2019;20(11). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6600373/

66. Glinsky G V., Higashiyama T, Glinskii AB. Classification of Human Breast Cancer Using Gene Expression Profiling as a Component of the Survival Predictor Algorithm. Clin Cancer Res [Internet]. 2004;10(7):2272–83. Available from: https://www.ncbi.nlm.nih.gov/pubmed/15073102

67. Stoeger T, Gerlach M, Morimoto RI, Nunes Amaral LA. Large-scale investigation of the reasons why potentially important genes are ignored. PLoS Biol [Internet]. 2018;16(9):1–25. Available from: https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.2006643

68. Valdebenito J, Medina F. Machine learning approaches to study glioblastoma: a review of the last decade of applications. Reports, Cancer. 2019;

69. Lin CY, Halabi S. On model specification and selection of the Cox proportional hazards model. Stat Med [Internet]. 2013;32(26):4609–23. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3795916/pdf/nihms492895.pdf

70. Ishwaran H, Kogalur UB, Chen X, Minn AJ. Random Survival Forests for High-dimensional data. Stat Anal Data Min [Internet]. 2011;4(January):115–32. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/sam.10103

71. Clancy T, Malmberg KJ, Dannenfelser R, Troyanskaya O, Hovig E, Kristensen V. Bioinformatics approaches to profile the tumor microenvironment for immunotherapeutic discovery. Curr Pharm Des [Internet]. 2017;23(32):4716–25. Available from:

https://www.ncbi.nlm.nih.gov/pubmed/28699527