Informe: Análisis de Fosfoproteómica Diferencial en MSS y PD

Introducción y Objetivos del Estudio

El análisis de la fosforilación de proteínas juega un papel central en la comprensión de los mecanismos moleculares que subyacen al cáncer. La fosforilación es una modificación postraduccional clave que regula rutas de señalización involucradas en la proliferación celular, apoptosis y migración. En este contexto, el objetivo de este estudio fue identificar diferencias en los patrones de fosforilación entre los subtipos tumorales MSS (Microsatellite Stable) y PD (Poorly Differentiated), explorando su potencial para clasificar tumores, identificar biomarcadores y revelar mecanismos moleculares específicos que podrían tener relevancia clínica.

Para abordar este objetivo, se diseñó un análisis exhaustivo basado en datos de fosfoproteómica, desde su obtención inicial hasta la interpretación funcional. Las preguntas que guiaron este trabajo incluyeron la identificación de diferencias clave en la fosforilación, la exploración de rutas de señalización asociadas y la evaluación de la utilidad de fosfopéptidos como biomarcadores para distinguir MSS de PD.

El análisis se estructuró en torno a seis áreas principales: la obtención y estructuración de los datos, su exploración inicial, la evaluación de la calidad, el análisis estadístico diferencial, la interpretación biológica mediante enriquecimiento funcional y la discusión de las limitaciones inherentes al estudio.

Datos y Diseño Experimental

Los datos utilizados en este estudio se generaron a partir de modelos de xenoinjertos derivados de pacientes (PDX), seleccionados por su capacidad para preservar las características genéticas y fenotípicas de los tumores humanos. Este enfoque permitió analizar muestras representativas de los subtipos MSS y PD, reflejando con mayor precisión los procesos biológicos que se producen en el microambiente tumoral.

Se incluyeron un total de seis muestras, tres por subtipo tumoral, con dos réplicas técnicas por muestra. Las proteínas extraídas de los tumores fueron digeridas en péptidos mediante tripsina, y posteriormente enriquecidas en fosfopéptidos utilizando técnicas específicas como TiO_2 y anticuerpos α-pTyr. Esto garantizó una mayor sensibilidad y especificidad en la detección de fosforilaciones, permitiendo centrarse en las rutas de señalización activas en cada subtipo.

Obtención, Estructuración y Exploración Inicial de los Datos

El análisis comenzó con la descarga de los datos crudos necesarios para investigar las diferencias en los patrones de fosforilación entre MSS y PD. Los datos analizados corresponden al conjunto de datos denominado **"PhosphoTumorData_MSS_PD"**, que incluye información sobre los niveles de fosfopéptidos cuantificados mediante espectrometría de masas, descripciones detalladas de los fosfopéptidos y metadatos asociados a las muestras experimentales, como el subtipo tumoral y las réplicas técnicas. Los datos fueron obtenidos directamente desde un repositorio público en GitHub utilizando la función `curl_download` del paquete **`curl`**. Este método permitió descargar los archivos de manera automatizada, garantizando la reproducibilidad

del análisis y eliminando la dependencia de rutas locales. Los datos, en formato tabular, consistían en tres elementos principales: una matriz de abundancias que representaba los niveles de fosfopéptidos cuantificados mediante espectrometría de masas, un archivo con descripciones detalladas de los fosfopéptidos y un conjunto de metadatos que contenían información clave sobre las muestras, como el subtipo tumoral y las réplicas técnicas.

Los datos fueron leídos en R utilizando funciones del paquete **`readxl`**, que permitió extraer información desde un archivo Excel estructurado en varias hojas. La matriz de abundancias y los metadatos se procesaron por separado antes de integrarse en un contenedor estándar para análisis ómicos.

Para estructurar y organizar los datos, se utilizó el contenedor `SummarizedExperiment`, parte del ecosistema de Bioconductor. Este formato facilita la integración de datos biológicos con anotaciones descriptivas y asegura la coherencia entre las dimensiones de los datos. En este análisis:

- La matriz de abundancias se almacenó en el componente `assays`.
- Las descripciones de los fosfopéptidos se integraron en el componente `rowData`.
- Los metadatos de las muestras se incorporaron en el componente `colData`.

El objeto `SummarizedExperiment` fue creado utilizando la función `SummarizedExperiment()` del paquete homónimo, permitiendo combinar estos componentes en un único objeto que facilita su manejo durante las etapas posteriores del análisis.

Control de Calidad y Preprocesado de Datos

Con el contenedor creado, se inició un control exhaustivo de la calidad de los datos para asegurar que estuvieran libres de errores técnicos y listos para su análisis estadístico. Este paso incluyó la verificación de valores nulos, la visualización de patrones de ausencia y la evaluación de la distribución de los datos mediante histogramas y boxplots.

El primer paso fue comprobar la presencia de valores nulos en la matriz de abundancias utilizando funciones nativas de R como `is.na` y `colSums`. Este análisis reveló que no existían valores nulos en las columnas de abundancia, lo que eliminó la necesidad de realizar imputaciones. Este resultado destacó la buena calidad inicial de los datos, permitiendo avanzar al siguiente nivel de evaluación.

Para complementar este análisis, se utilizó la función `missmap` del paquete **`Amelia`** para generar un mapa de calor que visualizara los valores nulos. Aunque no se identificaron valores ausentes, esta visualización fue clave para confirmar la integridad de los datos en un entorno gráfico, facilitando la detección de posibles patrones de ausencia si hubieran existido.

Tras confirmar la ausencia de valores nulos, se procedió a analizar la distribución de las abundancias en cada muestra. Los histogramas y boxplots revelaron un sesgo significativo hacia valores bajos, un fenómeno común en datos de proteómica que puede dificultar las comparaciones entre muestras. Para abordar este problema, se aplicó una transformación logarítmica `(log10(abundance + 1))` utilizando funciones nativas de R. Esta transformación estabilizó la varianza, redujo la influencia de valores extremos y mejoró la comparabilidad entre muestras.

A lo largo de este proceso, los datos se visualizaron nuevamente para confirmar las mejoras en la distribución tras la transformación. Las distribuciones ajustadas mostraron patrones más uniformes, destacando la efectividad de este preprocesamiento en la preparación de los datos para el análisis estadístico.

En conjunto, estas medidas de control de calidad garantizaron que los datos estuvieran en óptimas condiciones para avanzar hacia los pasos posteriores, incluyendo la normalización y el análisis estadístico diferencial.

Análisis Estadístico e Identificación de Fosfopéptidos Diferenciales

Una vez completado el control inicial de calidad, se procedió a realizar un análisis estadístico exhaustivo con el objetivo de identificar fosfopéptidos cuya abundancia diferencial pudiera distinguir entre los subtipos tumorales MSS y PD. Este análisis comenzó con una exploración preliminar de los datos mediante un análisis de componentes principales (PCA), que permitió evaluar patrones generales de agrupación en los datos transformados logarítmicamente. Para ello, se utilizó la función `plotPCA3`, previamente cargada desde el repositorio de GitHub. El PCA inicial mostró una separación parcial entre los grupos MSS y PD, con una clara dispersión interna en el grupo PD, lo que sugería una mayor heterogeneidad en este subtipo tumoral. Este paso fue clave para detectar las diferencias iniciales entre los grupos y guiar los ajustes necesarios en las etapas posteriores.

Tras observar estas diferencias, se procedió con el análisis estadístico utilizando el paquete **`limma`**, una herramienta del ecosistema Bioconductor ampliamente reconocida en estudios ómicos por su capacidad para manejar datos complejos y réplicas técnicas. Este análisis comenzó con la preparación de un diseño experimental, creando una matriz de diseño mediante la función `model.matrix`, que asignaba cada muestra a su grupo correspondiente (MSS o PD). Posteriormente, se utilizó la función `duplicateCorrelation` para ajustar la variabilidad técnica introducida por las réplicas, asegurando que las diferencias detectadas reflejaran variaciones biológicas reales. Para evaluar las diferencias específicas entre los subtipos, se definió una matriz de contraste utilizando la función `makeContrasts`. Esto permitió comparar directamente las abundancias de fosfopéptidos entre MSS y PD, estableciendo una relación estadística precisa entre los grupos.

Los resultados preliminares de este análisis se visualizaron mediante un Volcano Plot, que destacó algunas diferencias significativas entre los fosfopéptidos, aunque también dejó en evidencia una gran dispersión en los datos, especialmente dentro del grupo PD. Este hallazgo confirmó la necesidad de realizar ajustes adicionales para reducir la variabilidad técnica y mejorar la precisión del análisis.

Para decidir el mejor enfoque de normalización, se examinó un resumen estadístico de las abundancias transformadas logarítmicamente. Aunque la transformación había mejorado la homogeneidad de las muestras, persistía una variabilidad notable en los cuartiles superiores y valores máximos del grupo PD. Por ello, se aplicó la **normalización cuantílica** (Quantile Normalization) mediante la función `normalizeQuantiles` del paquete `limma`. Este método asegura que las distribuciones de abundancia sean comparables entre muestras, ajustando sus rangos y eliminando sesgos técnicos sin alterar las diferencias relativas entre los grupos.

Tras la normalización, se realizó un nuevo análisis PCA para evaluar el impacto de los ajustes realizados. Este análisis mostró una mejora significativa en la separación entre los subtipos MSS

y PD, con una clara reducción de la dispersión interna en el grupo PD. El primer componente principal (PC1) explicó un alto porcentaje de la variabilidad total, reflejando principalmente las diferencias entre MSS y PD, mientras que el segundo componente (PC2) capturó variaciones menores, posiblemente relacionadas con características específicas de ciertas muestras. Estos resultados validaron la efectividad de la normalización aplicada.

Con los datos normalizados, se repitió el análisis estadístico diferencial utilizando `limma`, esta vez con el ajuste adicional de pesos adaptativos mediante la función `arrayWeights`. Este enfoque permitió minimizar el impacto de muestras más variables, como M43 del grupo PD, y asignar pesos según su contribución a la variabilidad total. Además, se aplicó el ajuste bayesiano mediante la función `eBayes`, que estabilizó los estimadores de error y mejoró la precisión en la identificación de diferencias significativas. Esta técnica es particularmente valiosa en estudios con un número reducido de muestras, ya que reduce el ruido estadístico y garantiza que los resultados sean robustos.

Los fosfopéptidos significativos se seleccionaron bajo criterios estrictos: un **FDR (False Discovery Rate) < 0.05**, controlado mediante el método de Benjamini-Hochberg, y un **log Fold Change (logFC) ≥ |1|**, lo que corresponde a cambios de al menos el doble o la mitad en abundancia, considerados biológicamente relevantes. Los resultados finales se visualizaron nuevamente mediante Volcano Plots, donde los fosfopéptidos destacados mostraron una alta significancia estadística junto con cambios biológicamente relevantes.

Finalmente, los fosfopéptidos identificados como diferencialmente abundantes entre MSS y PD se almacenaron en el contenedor `SummarizedExperiment`, listos para su análisis biológico y funcional en las etapas posteriores. Este enfoque meticuloso permitió garantizar la robustez de los resultados y sentar las bases para la interpretación biológica de las diferencias observadas.

Análisis de Significación Biológica con Enriquecimiento Funcional

Una vez identificados los fosfopéptidos diferencialmente abundantes entre los subtipos tumorales MSS y PD, se procedió al análisis de significación biológica, cuya finalidad fue interpretar las diferencias funcionales y biológicas asociadas a estos fosfopéptidos. Este análisis permitió profundizar en los mecanismos moleculares que distinguen a MSS y PD, proporcionando un contexto funcional y clínico para los resultados obtenidos.

El análisis comenzó con la generación de dos listas específicas de fosfopéptidos, correspondientes a los identificados como diferencialmente abundantes en MSS y PD. Los fosfopéptidos con un **log Fold Change positivo** (>1) se asociaron al subtipo **PD**, indicando una mayor abundancia en este grupo, mientras que aquellos con un **log Fold Change negativo** (<-1) se relacionaron con **MSS**, reflejando una mayor abundancia en este subtipo. Estas listas se generaron a partir de los resultados de **limma**, filtrando los fosfopéptidos mediante las funciones `filter` y `subset`, con un umbral de **FDR < 0.05** y un **logFC ≥ |1|**. Estas listas se generaron y almacenaron en el metadato del objeto `SummarizedExperiment` para facilitar su integración en análisis posteriores.

Para obtener información descriptiva y funcional de los fosfopéptidos, se realizaron anotaciones utilizando **UniProt IDs** como identificadores principales. Estas anotaciones incluyeron consultas directas a **UniProt** para recuperar nombres de genes, funciones moleculares, localizaciones subcelulares y asociaciones con enfermedades. Estas consultas se realizaron empleando funciones específicas como fetch_uniprot_annotations, lo que permitió acceder a una descripción detallada de las proteínas asociadas a los fosfopéptidos. Además, mediante el paquete **biomaRt**, se

complementaron estas anotaciones con atributos adicionales, como símbolos genéticos y descripciones funcionales más detalladas, asegurando una base sólida para los análisis de enriquecimiento.

En ciertas etapas del análisis, fue necesario realizar conversiones de identificadores. Para el enriquecimiento funcional en **Reactome**, se utilizó el paquete **ClusterProfiler** para convertir los UniProt IDs a **ENTREZ IDs**, garantizando la compatibilidad con esta base de datos de rutas metabólicas y señalización. Asimismo, en el análisis clínico en **Human Protein Atlas**, se llevó a cabo la conversión de UniProt IDs a **ENSG IDs** mediante biomaRt, lo que permitió validar los resultados en términos de expresión proteica en tejidos tumorales.

El análisis funcional se centró en herramientas de Bioconductor como **ClusterProfiler** y **ReactomePA**, integradas con bases de datos como **Gene Ontology (GO)**, **KEGG** y **Reactome**. Estas herramientas permitieron identificar procesos biológicos, funciones moleculares y rutas de señalización enriquecidas para cada subtipo tumoral. En **PD**, los resultados mostraron un enriquecimiento en rutas como **MAPK/ERK** y **PI3K/AKT**, fundamentales en la progresión tumoral debido a su papel en proliferación celular, migración e invasión. En contraste, los fosfopéptidos de **MSS** estuvieron enriquecidos en procesos metabólicos y de regulación celular, como la homeostasis y el control del ciclo celular, reflejando un perfil menos agresivo y más estable. Los resultados se visualizaron mediante herramientas como dotplot y emapplot, que resumieron gráficamente las diferencias observadas.

Además, se realizaron análisis de interacciones proteína-proteína (PPI) utilizando **STRINGdb**, lo que permitió construir redes moleculares a partir de las proteínas derivadas de los fosfopéptidos identificados. En **PD**, las redes mostraron una alta densidad y conectividad, destacando proteínas clave en la señalización proliferativa y evasión inmune, mientras que en **MSS**, las redes presentaron una estructura más dispersa, subrayando su diversidad funcional y reguladora.

Finalmente, se realizó una validación de los biomarcadores identificados mediante consultas en bases de datos clínicas como **Human Protein Atlas** y literatura científica en **PubMed**. En Human Protein Atlas, los ENSG IDs permitieron verificar la expresión diferencial de las proteínas asociadas en tejidos tumorales, confirmando su relevancia clínica. En PubMed, se utilizaron UniProt IDs para buscar estudios relacionados con las proteínas y sus implicaciones en la progresión tumoral, automatizando este proceso mediante una API personalizada. Estas validaciones confirmaron que varios fosfopéptidos son relevantes no solo biológicamente, sino también clínicamente, posicionándolos como biomarcadores prometedores para la estratificación de pacientes y el diseño de terapias dirigidas.

En resumen, este análisis de significación biológica destacó diferencias fundamentales entre MSS y PD. Mientras que en **PD** se evidenció una activación significativa de rutas asociadas a la proliferación, invasión y evasión inmune, en **MSS** se observó un perfil más enfocado en la estabilidad metabólica y la regulación celular. Estas diferencias no solo enriquecen la comprensión molecular de los subtipos tumorales, sino que también identifican rutas y fosfopéptidos clave con potencial para aplicaciones clínicas y terapéuticas.

URL repositorio Github: https://github.com/jvaldiviaga/Valdivia-Garcia-Jessica-PEC1