Prácticas Proteómica FragPipe

Sandra Mingo Ramírez

2024/25

Índice general

FragPipe es un software que engloba varios programas. El buscador es MSFragger, pero cuenta también con otras herramientas como Philosopher, PTM-Shepher, etc. Lo descargamos de GitHub y descargamos las licencias de MSFragger, IonQuant y diaTracer.

1. Práctica TMT: Evaluation of NCl-7 Cell Line Panel as a Reference Material for Clinical Proteomics

1.1. Con un único TMT

En este experimento había 7 líneas celulares y se querían ver las fosforilaciones de las líneas cancerígenas en comparación con el Wild-Type. Se hizo una digestión TMT, marcando los péptidos para saber si la muestra es KO o WT. Se hizo un LC-MS y se buscó contra la base de datos RefSeq. De esta forma se tienen en cuenta las modificaciones fijas y variables. Se cargometilan las cisteínas, quedando marcadas. Estas modificaciones se deben poner en la búsqueda.

En FragPipe, en la pestaña de Workflow, seleccionamos el workflow de TMT10. Luego pulsamos el botón de "Add files" y cargamos el fichero "01_CPTAC_TMTS1-NCI7_P_JHUZ_20170509_LUMOS.mzML".

MSFragger detecta que es un DDA, es decir, espectro Data Dependent. Como solo hay un fichero raw, no vamos a indicar réplicas biológicas. En la siguiente pestaña (DIA Pseudo MS2) debemos asegurarnos de que no hay nada marcado. Para Database, se pueden descargar las bases de datos de humanos, ratones, levadura y Covid. Por defecto, las bases de datos son de UniProt, y se puede seleccionar si añadir solo aquellas entradas revisadas manualmente, añadir decoys (que ayuda a calcular la FDR para la validación de los péptidos), añadir proteínas contaminantes (queratina, látex, errores experimentales), isoformas, etc. Se recomienda guardar esta base de datos con los raw.

La siguiente pestaña es de MSFragger. Se divide en varias secciones:

- **Peak Matching**

  Por un lado, se especifica la tolerancia de la masa del precursor. Se pone el rango de tolerancia en Daltons o PPM para la ventana de masa del precursos o fragmento. En el paper, se especifica que se han incluido 20 ppm tolerancia al precursor y 0.06 Da tolerancia al ion del fragmento. El estándar está en -20 y 20 en PPM, pero debemos cambiar la masa de tolerancia de los fragmentos a 0.06 Da, como se indica en el paper. Para que el programa no tarde mucho, se recomienda quitar la calibración y optimización.

- **Protein Digestion**

  En el paper se especifica que los péptidos se buscaron con dos missed cleavages. Esto lo recreamos en el programa: ponemos la enzima tripsina que corta en KR y 2 misscleavage en C.

- **Modifications**

  Aquí se dividen las modificaciones variables y fijas. La publicación, al ser un TMT 10, ya hay 229.16293 masa adicional en el N-terminal de la lisina. De igual forma, hay una carbometilación en cisteína que aumenta la masa en 57.02146. Esto lo debemos reflejar en modificaciones fijas: 229.16293 en N-Term Peptide y en K (lysine), y 57.02146 en C (cysteine), aunque esto último debería estar por defecto.

  Cuando se considera una modificación variable, deben tenerse en cuenta todos los posibles sitios de localización con todas las distribuciones posibles de esa modificación. Esto conduce a un fuerte aumento del tamaño del espacio de búsqueda, que escala exponencialmente con la inclusión de modificaciones adicionales, y con el consiguiente tiempo de computación y recursos. Para esta parte, no vamos a incluir la modificación variable de fosforilación, debido al tiempo de cómputo. No obstante, la oxidación en la metionina sí la mantenemos.

- **Opciones avanzadas**

Existen más opciones avanzadas dependiendo del tipo de búsqueda, pero no vamos a profundizar en ellos, solo en **Advances Output Options**. Se permiten especificar los formatos de salida, siendo recomendada la salida más completa: TSV_PEPXML_PIN.

Con esto, hemos terminado la parte del buscador. La siguiente pestaña es de Validación. Son métodos estadísticos o de ML para validar y comprobar que las asignaciones péptido-proteína son las correctas. Vamos a quitar MSBooter, dejando todo lo demás de forma predeterminada.

La pestaña de PTMs debe estar desactivada, al igual que Glyco y Quant (MS1). En la pestaña de Quant (Isobaric) se extraen las intensidades de los espectros, llegando a cuantificarlo comparado con las muestras. TMT-Integrator extrae y combina abundancias de canales de múltiples muestras marcadas con TMT. Activamos la pestaña y pulsamos Edit/Create. Ahí seleccionamos TMT10 y pulsamos Load into Table. Se genera un fichero en el que se van a ir guardando los canales. En Basic Options, seleccionamos Quant level 2, Define Reference Virtual, Group By All y Normalization MD (median centering).

Las pestañas de Spec Lib, Quant DIA y Skyline deben estar desactivadas. En Run, seleccionamos la carpeta donde queremos que se guarden los resultados. Finalmente, pulsamos el botón "RUN". En caso de que dé error por falta de memoria, en MSFragger se pueden ampliar los splits de la base de datos.

Entre los resultados, está psm.tsv. Lo abrimos y vemos que hay distintas columnas:

- Spectrum: identificador del espectro MS/MS, sigue el formato (nombre de archivo).(scan).(scan).(charge)
- Spectrum File: espectro nombre del archivo de identificación de origen
- Peptide: secuencia de aminoácidos del péptido sin incluir ninguna modificación
- Modified Peptide: secuencia peptídica que incluye las modificaciones; los residuos modificados van seguidos de paréntesis que contienen la masa entera (en Dalton) del residuo más la modificación; en blanco si el péptido no está modificado.
- Hyperscore: puntuación de similitud entre los espectros observados y teóricso; los valores más altos indican una mayor similitud.
- Nextscore: puntuación de similitud (hyperscore) de la segunda posición más alta para el espectro
- PeptideProphet Probability: puntuación de confianza determinada por PeptideProphet, los valores más altos indican mayor confianza.
- Number of missed cleavages: número de sitios potenciales de escisión enzimática dentro de la secuencia identificada
- Protein Start: posición inicial del péptido identificado dentro de la secuencia proteica.
- Protein End: posición final del péptido identificado dentro de la secuencia de la proteína.
- Assigned Modifications: Modificaciones variables (enumeradas por masa en Da) con residuo modificado y ubicación dentro del péptido
- Is Unique: si la secuencia identificada corresponde a una única proteína identificada (FALSE si es compartida por varias proteínas identificadas en el experimento)
- Protein ID: identificador de la proteína (número de acceso primario) de la proteína seleccionada
- Gene: nombre del gen de la proteína seleccionada
- Protein Description: descripción de la proteína seleccionada
- Columnas para los canales TMT/iTRAQ si se utilizan, donde cada una contiene las abundancias relativas de iones informadores para ese PSM.

La salida se compone de tres archivos:

- psm.tsv, contienen los resultados de la búsqueda filtrados por FDR, donde cada fila contiene una coincidencia péptido-espectro (PSM). Se generará un archivo psm.tsv distinto para cada experimento.
- peptide.tsv, contienen los resultados de la búsqueda filtrados por FDR, donde cada fila es una secuencia peptídica identificada. Se generará un archivo peptide.tsv distinto para cada experimento. Las columnas son:
  - Peptide Length: número de residuos en la secuencia peptídica
  - Charges: estado(s) de carga del ion péptido
  - Probability: puntuación de confianza determinada por PeptideProphet, los valores más altos indican mayor confianza.
  - Spectral Count: número de PSM correspondientes
  - Intensity: intensidad sumada de los 3 iones más abundantes para el péptido
- protein.tsv, contienen resultados de proteínas filtrados por FDR, donde cada fila es un grupo de proteínas identificado. Se generará un archivo protein.tsv separado para cada experimento. Las columnas son:
  - Percent Coverage: porcentaje de la secuencia de la proteína observado a partir de los péptidos identificados
  - Organism: especie de la proteína identificada
  - Protein Existence: tipo de evidencia que respalda la proteína
  - Protein Probability: puntuación de confianza determinada por ProteinProphet
  - Top Peptide Probability: mejor probabilidad del péptido entre los péptidos de soporte

Los resultados de TMT-Integrator estarán en una nueva carpeta, tmt-reports. Estos resultados contienen valores normalizados transformados por el log2, con archivos separados tanto para abundancias (intensidades) como para ratios en cada nivel (gen, proteína, péptido, etc.)

1.2. Con un único TMT: validando la PTM (fosforilación en STY)

Utilizamos el mismo fichero que en la primera parte: En FragPipe, en la pestaña de Workflow, seleccionamos el workflow de TMT10. Luego pulsamos el botón de "Add files" y cargamos el fichero "01_CPTAC_TMTS1-NCI7_P_JHUZ_20170509_LUMOS.mzML".

En MSFragger, vemos que la calibración y optimización este deshabilitada (None), que haya -20 y 20 PPM y 0.06 Da. Entre las modificaciones variables, activamos STY.

En la pestaña de Validation, incluimos un programa para detectar mejor las modificaciones variables. En este caso, es Run PTMProphet en el apartado de PTM Site Localization. Dentro de la sección de PTM Prophet, en Cmd line opts, pegamos lo siguiente: `-keepold -static -fragppmtol 15 -em 1 -nions b -mods STY:79.966331,M:15.9949 -minprob 0.5`.

En la pestaña Quant (Isobaric), hay que poner `S(79.9663),T(79.9663),Y(79.9663)` en Mod tags de PTMs. También se recomienda bajar la Min site probability a -0.75.

Con esto, en Run cambiamos la carpeta de resultados para que no se nos sobreescriba y lo ejecutamos.

Si hubiéramos incluido la modificación variable de fosforilación en la serina, tirosina y treonina, el número de proteínas identificadas hubiera aumentado de 193 a 2,776 proteínas. Sin embargo, el tiempo de cómputo, ha aumentado (con un solo mzML).

Si se realizó una cuantificación específica de PTM, se generarán informes de “multi-site” y “single-site” para la modificación especificada. En los resultados de “single-site”, los péptidos identificados con múltiples modificaciones especificadas se convierten a una forma de sitio único. PTM-Prophet proporciona la localización de la modificaciones (actualmente compatible solo con flujos de trabajo enriquecidos en fosfopéptidos).

2. Práctica Label-Free Quantification (LFQ): Proteomics separates adult-type diffuse high-grade gliomas in metabolic subgroups independent of 1p/19q codeletion and across IDH mutual status

En este paper, se llegaron a usar 42 muestras en diferentes gliomas cerebrales. Había genes WT y unos con SNP. Se quería ver si había fosforilaciones u otras metilaciones. Para comprobar esto, utilizaron diferentes réplicas. No vamos a usar las 42 muestras, si no 6, 3 mutantes y 3 wild type. Los autores hacen un corte de FDR del 1 %.

En este caso, como es otra técnica y otro tipo de espectometría usada, en Workflow limpiamos los ficheros (botón clear files). La metodología de workflow a usar es LFQ-MBR (matched between ranks) y lo cargamos (load workflow). A la hora de cargar los ficheros, pulsamos add files, vamos a la carpeta y seleccionamos todos. En las columnas, vamos a etiquetar a qué condición pertenece cada fichero mzML y si son réplicas per se. Por un lado establecemos que todos son réplicas biológicas pulsando Set bioreplicates Consecutive. Para la condición o experimento asociado, ponemos IDHwt para aquellos wildtype e IDHmut para los mutantes. El mismo programa ha detectado que los datos son DDA.

En MSFragger, seleccionamos la calibración None y mantenemos los valores por defecto de tolerancia. También mantenemos la digestión proteica de tripsina. De las modificaciones fijas lo dejamos como está, y de las modificaciones variables solo mantenemos la carbometilación, por lo que hay que desclickar STY.

En la pestaña Quant (MS1), se utilizará IonQuant. Se deja todo por defecto, pero hay que dejar seguro que la opción Match between runs esté habilitada. Además, MaxLFQ también debe estar habilitada. Ahora ya podemos darle a Run.

Una vez finalizado el análisis, los resultados de identificación con los valores de cuantificación estarán en la carpeta de resultados. Si se establecieron múltiples experimentos durante el paso de entrada de archivos en la pestaña "Workflow", estos resultados de identificación estarán en una carpeta separada para cada experimento. Los ficheros de salida son:

- combined_ion.tsv, contienen iones filtrados por FDR de todos los grupos experimentales, donde cada fila representa una secuencia de péptido con una carga y un estado de modificación específicos.
- combined_modified_peptide.tsv, contienen péptidos modificados filtrados por FDR de todos los grupos experimentales, donde cada fila corresponde a una secuencia de péptido que incluye modificaciones.
- combined_peptide.tsv, contienen péptidos filtrados por FDR de todos los grupos experimentales, donde cada fila representa una secuencia de péptido (sin modificaciones).
- combined_protein.tsv, contienen proteínas filtradas por FDR de todos los grupos experimentales, donde cada fila corresponde a un grupo de proteínas.

3. Análisis y visualización usando FragPipe-Analyst con TMT: Integrated Proteogenomic Characterization of Clear Cell Renal Cell Carcinoma

Para esta práctica usaremos los resultados de proteómica cuantitativa generados por FragPipe con TMT. FragPipe-Analyst es una aplicación web interactiva y fácil de usar desarrollada para realizar análisis de expresión diferencial con «un solo clic» y para visualizar conjuntos de datos proteómicos cuantitativos analizados mediante la plataforma computacional FragPipe. Es compatible con los flujos de trabajo de cuantificación LFQ-MBR, TMT y DIA de FragPipe. La página es https://fragpipe-analyst.org/.

Se utilizaron células renales cancerígenas y se realizó un análisis exploratorio de distintas ómicas. Se sacaron datos con TMT10 DDA y DIA. Los ficheros de entrada son dos ficheros tsv con la abundancia de las proteínas y la anotación del experimento. En FragPipe-Analyst seleccionamos el tipo de dato, que para TMT es proteína o gen. Se debe subir el resultado de FragPipe a nivel de proteínas generado por TMT-

Integrator, que es un fichero TSV. Se generan dos conjuntos de resultados de cuantificación de etiquetado isobárico:

- Los archivos que contienen "ratio" en el nombre, reportan la relación o proporción con respecto al canal de referencia, si se ha especificado, o con respecto a la abundancia promedio (enfoque de referencia virtual).
- Los archivos que contienen "abundance" reportan las intensidades brutas, la abundancia de iones; es decir, el conteo de iones para cada canal de TMT en una muestra.

Hay ratios y abundancias para: gene, protein, peptide, multi-site y single-site. El TMT-Integrator no solo calcula las intensidades y ratios, sino que también aplica normalización al nivel indicado para corregir variaciones sistemáticas y mejorar la comparabilidad entre muestras. Existen dos métodos principales de normalización: Median Centering (Centrado en la Mediana) o Global Normalization (Normalización Global).

[abundance/ratio]protein[normalization].tsv contiene información de cuantificación isobárica resumida desde las tablas psm.tsv al nivel de proteínas por TMT-Integrator.

- Index, nombre de la proteína (encabezado de la secuencia FASTA).
- NumberPSM, coincidencias totales de espectro-péptido que se asignan al gen y se utilizan en la cuantificación.
- ProteinID, identificador de proteína.
- MaxPepProb, la mayor probabilidad de PeptideProphet de los PSMs que se asignan a la proteína y se utilizan en la cuantificación.
- ReferenceIntensity, se utiliza la abundancia del canal de referencia real si se ha proporcionado una, de lo contrario, es la abundancia promedio de los canales. Los valores están escalados en log2, utilizando la intensidad mínima global de referencia para imputar los valores faltantes.

muestra/canal , abundancia/ratio normalizada y transformada en log2 para el canal de ion reportero dado, resumida al nivel de proteínas.

El archivo experimental_annotation.tsv debe contener las siguientes columnas: plex, channel, sample, replicate, condition, sample name. Es generado automáticamente por FragPipe (pero es recomendable revisarlo y anotar las columnas de réplica y condición, antes de usarlo).

3.1. Parámetros del FragPipe-Analyst

3.1.1. Imputación de valores faltantes ("missing")

En FragPipe-Analyst se proporcionan múltiples opciones de imputación de valores "missing":

- **Perseus-style:** Método se basa en el procedimiento de imputación implementado en el software Perseus, desarrollado por el equipo de MaxQuant. Los valores faltantes se reemplazan por números aleatorios generados a partir de una distribución normal con un desplazamiento de 1.8 desviaciones estándar hacia abajo y un ancho de 0.3 en cada muestra.
- **knn (k-nearest neighbors):** Los valores faltantes se reemplazan mediante una técnica de promediado con los vecinos más cercanos (nearest neighbors).
- **MLE (Maximum Likelihood Estimation):** Método basado en máxima verosimilitud utilizando el algoritmo Expectation-Maximization (EM).
- **min**: Los valores faltantes se reemplazan por el valor más pequeño que no esté ausente en los datos.

- **zero**: Los valores faltantes se reemplazan por 0

Por defecto, el método "Perseus-style" se aplica a datos de LFQ (Label-Free Quantification) tanto en DDA como en DIA. En el caso de datos TMT (Tandem Mass Tag), no se realiza imputación.

3.1.2. Normalización

En FragPipe-Analyst se proporcionan múltiples opciones para normalizar:

- **Normalización de estabilización de varianza (Variance-Stabilizing Normalization, VSN):** Está disponible solo para datos LFQ obtenidos mediante DDA y DIA.
- **Centrado en la mediana (Median Centered Normalization):** Los valores de cuantificación de cada muestra son ajustados restando la mediana de cada muestra individual

Por defecto, los datos de entrada no son normalizados, ya que se asume que ya han sido normalizados por las herramientas de cuantificación de FragPipe

3.1.3. Análisis de expresión diferencial

Internamente, usa el paquete limma de Bioconductor para realizar el análisis DE en cada proteína. El ajuste por múltiples pruebas (cálculo del p-value ajustado) se realiza con las opciones especificadas por el usuario (por defecto con "BH"). Los tipos de tasa de descubrimientos falsos (FDR) son mtodo de Benjamin Hochberg (BH) y basado en área local y de cola (por fdrtools). También se tienen en cuenta los umbrales definidos por el usuario para filtrar las proteínas diferencialmente expresadas de manera significativa.

3.2. Resultados de FragPipe-Analyst

3.2.1. Tabla de resultados

La tabla contiene el nombre de genes, ID de proteínas, logarítmico del cambio (para cada comparación por pares), p-valores ajustados aplicando correcciones FDR, p-valores, valores booleanos para significancia y la intensidad media de la proteína transformada logarítmicamente en cada muestra.

3.2.2. Gráficos de resultados

El **Volcano plot** es la representación gráfica entre el logaritmo del cambio (eje x) frente al -log10 del p-valor ajustado (eje y). Las proteínas candidatas interesantes se ubican en los cuadrantes superiores izquierdo y derecho. El usuario puede mostrar o no los nombres de proteínas o usar el "p-value ajustado" como el eje y. El usuario puede resaltar proteínas de su interés (coloreadas) seleccionando la fila desde la "Results Table". El gráfico se puede descargar utilizando el botón "Save Highlighted Plot".

El **"heatmap"** ofrece una visión general de todas las proteínas significativas/diferencialmente expresadas (filas) en todas las muestras (columnas). Esta visualización permite identificar tendencias generales. El usuario también tiene la opción de descargar información de proteínas de grupos individuales.

En cuanto al **"Feature plot"**, al seleccionar una proteína/gen de la "Results Table", se mostrará un gráfico de caja o un gráfico de violín que compara la abundancia de esas proteínas entre las condiciones.

3.2.3. Gráficos de control de calidad

El **Análisis de Componentes Principales (PCA)** es una técnica utilizada para resaltar la variación y descubrir patrones fuertes en un conjunto de datos. PC1, que es una combinación lineal de todas las

características (eje x), y explica la mayor parte de la variación de los datos, seguido por el resto de los PCs.

El **sample correlation plot** es un gráfico que muestra una matriz de correlación como un mapa de calor para visualizar los coeficientes de correlación de Pearson entre las diferentes muestras.

Sample CVs plots es un gráfico que representa la distribución del coeficiente de variación a nivel de proteínas para cada condición. Cada gráfico contiene una línea vertical que representa el porcentaje de CV mediano dentro de esa condición.

Feature Numbers es un gráfico de barras que representa el número de proteínas identificadas y cuantificadas en cada TMT.

Missing values Heatmap sirve para explorar el patrón de valores missing en los datos, indicando si los valores están ausentes o no. Solo se visualizan las proteínas con al menos un valor perdido.

El **density plot** muestra la distribución de las abundancias de proteínas en los datos originales, en los datos filtrados y en el conjunto de proteínas en las que se realizó la imputación.

3.2.4. Análisis de enriquecimiento

Pathway enrichment selecciona la bases de datos de rutas metabólicas: Hallmark, KEGG, WikiPathways y Reactome, junto con la dirección de regulación (regulado al alza o regulado a la baja). La lista de genes diferencialmente expresados (DE) se compara con conjuntos de genes conocidos para identificar posibles rutas metabólicas involucradas. Se realiza una prueba hipergeométrica para evaluar el enriquecimiento de los genes en las rutas seleccionadas.

Gene Ontology utiliza el mismo algoritmo que en la parte de rutas metabólicas. Las opciones de bases de datos de GO son: Función Molecular/Componente Celular/Proceso Biológico; y la dirección.

3.2.5. Descarga de datos

Se pueden descargar distintas tablas en diversos formatos:

- **Matriz de datos sin imputar:** Intensidades originales de las proteínas antes de la imputación en cada muestra.
- **Matriz de datos imputados:** Intensidades de proteínas después de aplicar el método de imputación seleccionado.
- **Resultados completos:** Tabla combinada con todos los datos anteriores, es decir, con y sin información de imputación, junto con los cambios de pliegue y p-values.

Además, se puede descargar un informe en PDF con algunas estadísticas y gráficos.

4. Análisis no dirigido (untarget) con datos DIA: Integrated Proteogenomic Characterization of Clear Cell Renal Cell Carcinoma

Se recopilaron 110 muestras de tumor (T) y 83 muestras de tejidos normales adyacentes (NAT) de los pacientes, y sus proteomas se perfilaron mediante espectrometría de masas. Estas muestras fueron analizadas originalmente utilizando etiquetado en tándem de masas (TMT), y adquisición independiente de datos (DIA). Usaremos solo 10 ficheros de DIA de 5 pacientes con ccRCC, con una muestra de tumor y una muestra de NAT emparejada para cada paciente.

En FragPipe, el primer paso es Install/Upgrade EasyPQP. Ahora, en la pestaña de Workflow, seleccionamos DIA_SpecLib_Quant y lo cargamos. Añadimos los ficheros raws y ponemos manualmente el experimento (NAT/T) y biorréplica. Nos debemos asegurar de que pone DIA en la celda de la derecha.

En la pestaña de Database nos descargamos la base de datos de humanos revisadas con señuelos y contaminantes. En MSFragger cambiamos Calibration and Optimization a None y mantenemos lo demás por defecto. En las modificaciones variables, no vamos a incluir la fosforilación por el tiempo de cómputo elevado.