Corroboración de picos y descubrimiento de motivos en *P. aeruginosa*

Anastasia Hernández Koutoucheva y Alán Fernando Muñoz González

Introducción

Pseudomonas aeruginosa es una bacteria patógena oportunista de diversos lugares del cuerpo humano; médicamente importante por su ubiquitinidad, resistencia intrínseca a una variedad de sustancias antimicrobianas y el incremento en los casos clínicos de cepas panresistentes. Esta adquisición se atribuye a cambios mutacionales que impactan, por ejemplo, la regulación río arriba de bombas de flujo que promueven expulsión antimicrobial o la desrepresión de genes que producen mayor especificidad en permeabilidad de la membrana [1]. Pseudomonas es de las bacterias patógenas más problemáticas por un conjunto de razones: Es muy fácil que las cepas sean resistentes y sus tasas de adquisición de resistencia van en incremento, aunado a que tienden a encontrarse en infecciones graves [2]. Dados los problemas actuales con la resistencia adquirida, se ha visto la regulación de factores reguladores de virulencia como una oportunidad para controlar las infecciones sin intervenir en el crecimiento mediante intervención en el quorum sensing: Sustancias que evitan la expresión de genes de virulencia aumentan la sensibilidad a antibióticos [3]. En este trabajo usamos los datos crudos de Chip-seq de un estudio de potenciales sitios de unión de factores de virulencia. Analizamos los datos crudos de diferente manera a la usada en el artículo base (Kong, W. et al, 2015) y comparamos los picos obtenidos por ellos y por nosotros para dar mayor respaldo a los sitios resultantes en ambos e incluso agregar más sitios con potencial gracias a ciertos pasos diferentes de la metodología descritos abajo.

Obtención de secuencias propias

Para comparación con los resultados que obtuvieron en el artículo base descargamos las reads directamente. Después requeríamos hacer el mapping de (en el artículo usaron TopHat), mientras que los archivos de formato fastq a SAM para mapearlos mediante bowtie2 . De los 10 archivos con reads originales el porcentaje de secuencias mapeadas fue variable (entre 31 y 85%). Finalmente usamos MACS2 para hacer el peak calling, no fue posible que MACS generase el modelo por su cuenta (log en material suplementario), ajustamos los datos a la falta del modelo: .

Detección de motivos

Los motivos fueron detectados en dos conjuntos de 10 secuencias: Las obtenidas en el repositorio de GEO del artículo original y las obtenidas en la sección anterior de este trabajo; para observar las diferencias y la significancia de los picos encontrados.

Para eso, se usó el protocolo de detección de motivos estudiado en clase {cita} con algunos cambios. Para el primer conjunto de datos, se descargaron las secuencias en formato .txt y se convirtieron a .bed (código en material suplementario), en RSat se utilizó el servidor de procariotes para hacer uso de la herramienta sequences from bed/gff/vcf contra Pseudomonas aeruginosa pao1.ASM676v1.30, después se usa Peak motifs para obtener los motivos, únicamente con la opción de análisis con oligos. Con los resultados obtenidos en este paso se procede a usar Matrix-clustering para corroborar los datos, el cual fue realizado en el servidor de fungi para disminuir el tiempo. Y, finalmente, se realizó el control negativo con fragmentos aleatorios en el genoma para observar la presencia de falsos negativos en los resultados, para esto se regresa al servidor principal y se usan las herramientas Random Genome Fragments y Peak motifs. Para el segundo conjunto de datos, tomamos las secuencias de nuestro análisis con MACS2 y las convertimos a formato .bed y realizamos el mismo análisis descrito en el análisis anterior, usando exactamente los mismos parámetros. Cabe destacar que todos los análisis se realizaron para los 10 picos mencionados en el artículo para medir su significancia,

pero los resultados se centran en los datos de AlgR, los cuales fueron los más significativos para el artículo original.

Tras esto, se realizaron gráficas para observar los cambios en los resultados obtenidos por ambos métodos. Como última prueba, se realizó un análisis de diadas para observar si los datos obtenidos para AlgR podían ser recuperados de forma total con otro método para detección de motivos.

Resultados

Comenzando con el set de datos obtenidos directamente del artículo, se obtiene la media de los pares de bases en los picos, la cual se encuentra alrededor de los 200 pdb, esto sucedió en todos los archivos, con excepción de wspr, de la cual cabe mencionar que este archivo de picos contenía únicamente una coordenada (de la posición 2069106 a 2070194) , por lo que los datos obtenidos en el análisis no fueron buenos ni concordaban con el resto de los picos, en los cuales se observaba una mayor cantidad de coordenadas (entre 10 y 75). También, se observa la composición de las bases nitrogenadas, la cual en 8 de 10 secuencias es mayor en G/C y menor en A/T (Figura 1) en las mismas proporciones, dentro de esta clase se encuentra algr, las otras 2 secuencias conservan la proporción de las bases pero la señal entre los nucleótidos individuales no es la misma. El siguiente parámetro obtenido fue la cantidad de motivos obtenidos, el cual de nuevo, coincidió en 8 de 10 archivos, con 10 motivos en cada una. Enfocándonos en la significancia de los resultados, la mayor obtenida fue de 28 con la secuencia de motivo accgacgaacggtcg, archivo que correspondía a las coordenadas de AlgR. Para consultar las tablas con los resultados completos, puede consultarse el material suplementario. Con la secuencia mencionada anteriormente secuencia en mente, podemos hacer énfasis en la figura 2 y 3 del material suplementario, en la primera podemos observar la imagen obtenida directamente del artículo original en la que se obtiene el motivo de las coordenadas del archivo de AlgR con ayuda de MEME; en la figura 3 tenemos un logo generado por RSat con el procedimiento descrito en los métodos de descubrimiento de motivos con las secuencias descargadas directamente de GEO. Podemos observar que las secuencias en ambos motivos son iguales, tanto en tipo de nucleótidos como en frecuencia de las mismas; por lo tanto, podemos observar que al analizar las mismas secuencias por dos métodos distintos (MEME y RSat) conservamos el mismo resultado final.

Prosiguiendo el estudio con el set de datos obtenidos desde las secuencias crudas (.sar), obtenemos la media de los pares de bases en los picos, la cual está alrededor de 300 pdb, en este caso, no tenemos ningún problema con ninguna secuencia en particular ya que todas se encuentran dentro del mismo rango. La composición de pares de base es exactamente igual que en el análisis con los datos de GEO, con excepción de la secuencia wspr la cual se comporta igual que la mayoría de las secuencias en el resultado anterior (>G/C <A/T, en todos los nucleótidos). Los motivos descubiertos son totalmente consistentes, ya que en cada uno de los archivos se encontraron 10, sin ninguna excepción. Con los resultados de significancia, podemos observar que en todos los archivos de coordenadas conseguimos valores mucho más altos y los máximos globales en éstas fueron de 77 y 76 en gana y AlgR, respectivamente. La secuencia correspondiente a ambas secuencias fue la misma, ttcctcggcga. Los resultados completos pueden observarse en la tabla del material suplementario.

Con esta segunda secuencia en mente, podemos observar el logo generado en la figura 4 del material suplementario; haciendo énfasis en que la secuencia de nucleótidos obtenida es la misma que en las figuras 2 y 3 pero lo que varía un poco es la frecuencia en la que los encontramos; esto podemos atribuirlo a que la cantidad de coordenadas que obtuvimos en nuestro análisis con MACS2 fue mucho mayor a la obtenida en el artículo original; pero de forma independiente, sigue observándose claramente la conservación de la secuencia.

Para corroborar la calidad de los resultados, se realizó el control negativo de ambos análisis de los cuales pueden observarse los resultados completos en las tablas respectivas; en todos los casos descubrieron la misma cantidad de motivos que en la prueba y en la mayoría de los casos obtuvieron significancias mucho menores. Pero, haciendo énfasis en la secuencia algr, podemos observar que las significancias en estos controles bajan a 7 y a 40, para cada uno de los experimentos individuales.

También, para continuar la corroboración con Matrix-Clustering en AlgR, podemos observar que los clusters encontrados en las pruebas fueron 5(4,2,2,1,1) y 5(4,2,2,1,1), para cada uno de los sets de secuencias; por lo

que los clusters se conservaron en este sentido. En cambio, para las secuencias generadas aleatoriamente fueron de 2(6,4) y 6(2,3,2,1,1,1); siendo claro que los clusters se conservan únicamente en secuencias verdaderas.

En los resultados del análisis de diadas, puede observarse que parte de las regiones obtenidas están conservadas en las secuencias de resultados; demostrando una mayor robustez del sitio obtenido. Incluso, se descubre un sitio con una secuencia de mayor tamaño, lo que podría sugerir que se perdió cierta parte de la información del resultado con el análisis usando MEME en el artículo original.

En la parte de estadística con R, al analizar la comparación de reads entre el artículo y nuestros resultados, obtuvimos lecturas de mayor tamaño, pero la mayoría rodean los 200. Después comparamos el enriquecimiento de los sitios detectados, en rojo los sitios del artículo y en verde los nuestros. Corroboramos la mayoría de los sitios y encontramos otros potenciales a lo largo del genoma de Pseudomonas aeruginosa. Para ello colocamos primero los de ellos sobre los nuestros y viceversa.

Conclusiones

Nuestras gráficas muestran la robustez de los picos obtenidos en el artículo, pues obtuvimos una mayor cantidad de peaks debido a un mapeo de lecturas más fuerte, incluso con mayores constricciones en el peak calling.

Enfocándonos en el pico en el que se centraron los análisis, tanto en el artículo como en nuestro trabajo; podemos inferir la importancia de asegurarnos que AlgR sea un sitio de unión significativo, ya que los genes que tienen este comportamiento están envueltos en patogénesis [1] indicando su importancia en la virulencia de P. aeruginosa. Por esto, la observación de que el motivo obtenido en el artículo podría estar perdiendo información, es bastante importante.

Por último, cabe recalcar que estos datos nos otorgan pistas para entender de mejor forma los mecanismos moleculares de la regulación del organismo completo, y que para asegurarnos de que nuestras inferencias sean correctas, antes debemos confirmar que tenemos datos correctos y completos..

Bibliografía.

1. Kong, W., Zhao, J., Kang, H., Zhu, M., Zhou, T., Deng, X., & Liang, H. (2015). ChIP-seq reveals the global regulator AlgR mediating cyclic di-GMP synthesis in Pseudomonas aeruginosa. Nucleic acids research, gkv747.
2. Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome biol, 10(3), R25.
3. Contreras Moreira, B., Mondragon, JC., Rioualen, C., Cantalapiedra, CP,. Van Helden, J. (2016). RSAT::Plants: motif discovery in ChIP-seq peaks of plant genomes.
4. Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., ... & Liu, X. S. (2008). Model-based analysis of ChIP-Seq (MACS). Genome biology, 9(9), R137.
5. Livermore, D. M. (2002). Multiple mechanisms of antimicrobial resistance in Pseudomonas aeruginosa: our worst nightmare?. Clinical infectious diseases, 34(5), 634-640.
6. Poole, K. (2011). Pseudomonas aeruginosa: resistance to the max. Front Microbiol, 2(65). Hentzer, M., Wu, H., Andersen, J. B., Riedel, K., Rasmussen, T. B., Bagge, N., ... & Manefield, M. (2003). Attenuation of Pseudomonas aeruginosa virulence by quorum sensing inhibitors. The EMBO journal, 22(15), 3803-3815.