LA INVESTIGACIÓN EN BIOMEDICINA COMO CIENCIA DE DATOS ÓMICOS

El Doctor Pedro Carmona Sáez, experto en bioinformática, análisis de datos *ómicos* y desarrollo de técnicas estadísticas y computacionales para la extracción de información a partir de grandes volúmenes de datos biomédicos, remarcó 4 mensajes durante su intervención en las V Jornadas de Bioinformática de la Universidad de Granada [1]:

1.-Las tecnologías *ómicas* han revolucionado la investigación biomédica. Esta revolución comenzó en 1998 con el uso de los microarrays, acentuándose, en 2001, con la publicación del primer borrador del genoma humano [2] y, en 2005, con la aparición de las técnicas de secuenciación masiva, que han permitido analizar muchos más genes en muy poco tiempo. Por tanto, se hizo imprescindible la figura del "bioinformático", persona encargada de analizar el gran volumen de datos que generan estas tecnologías.

En el ámbito de la Biomedicina, el manejo de este tipo de datos supone un reto, ya que los sistemas biológicos son complejos y dinámicos. Un gen puede realizar diferentes funciones en diferentes tipos de células o tejidos, la relación señal-ruido en los análisis es baja y existen muchas variables que dificultan su estudio. Además, se disponen de pocas muestras y decenas de miles de variables en cada capa *ómica*, que se relacionan entre sí (*multiómicas*) aumentando su dimensión y complejidad (figura 1). Esto ha implicado cambios de paradigmas en las investigaciones biomédicas, ya que en la era *preómica* la mayor parte de los recursos humanos y económicos se empleaba en realizar experimentos que generaban datos de forma discreta, mientras que ahora los experimentos generan datos de forma masiva y la mayor inversión se emplea en analizarlos durante meses.

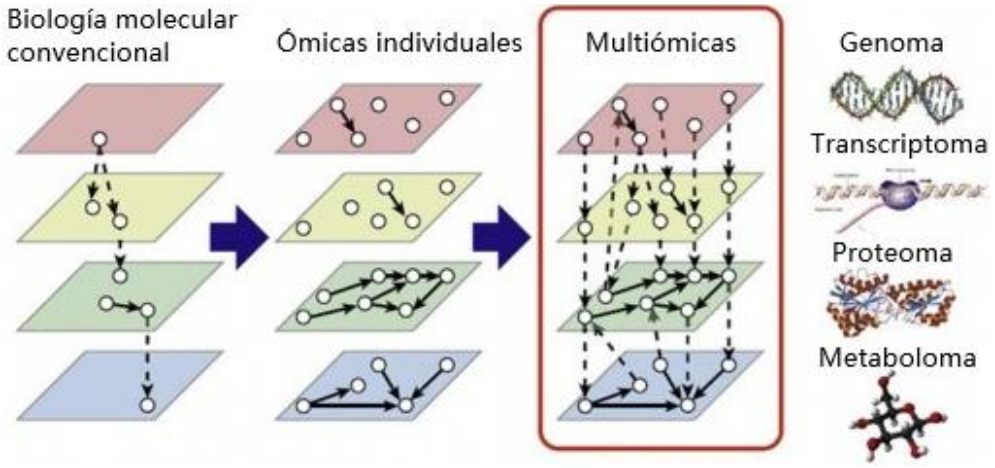

Figura 1. La evolución de la investigación biomédica, desde la biología molecular convencional hasta las *multiómicas*, ha aumentado su dimensión y complejidad.

2.- Las tecnologías basadas en *ómicas* han liderado el desarrollo de nuevos campos y disciplinas, como la medicina de precisión o la farmacogenómica. La secuenciación masiva se está implementando en el ámbito clínico, principalmente para el diagnóstico [3], y ha permitido averiguar, por ejemplo, que las alteraciones moleculares que llevan a un tipo de cáncer son diferentes en cada paciente y, por tanto, es posible aplicar un tratamiento específico para cada uno (figura 2). Con el avance de estas tecnologías, cada vez es mayor el número de pacientes elegibles para estas terapias personalizadas [4, 5].

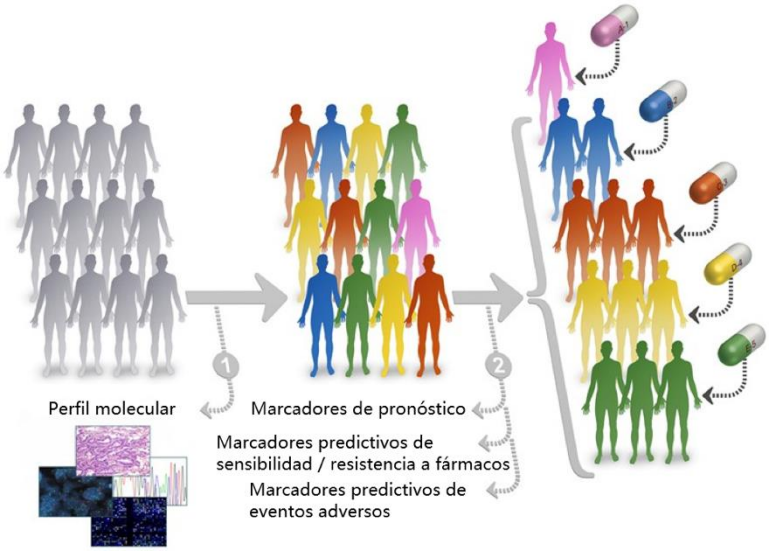

Figura 2. Etapas de la terapia personalizada en cáncer.

3.- La enorme expansión del uso de las tecnologías *ómicas* ha provocado la reducción de sus costes y la acumulación exponencial de datos [6, 7], así como de bases de acceso público que los almacenan (más de 1600) [8]. Estas bases de datos biológicos pueden ser generales, como el ArrayExpress del EBI (Instituto Europeo de Bioinformática) [9] y las del NCBI (Centro Nacional de Información Biotecnológica) [10], o específicas, las cuales han impulsado la medicina de precisión, como TCGA (Atlas del Genoma del Cáncer) [11], que recoge datos *multiómicos* de casi todos los tipos de cáncer útiles para su diagnóstico, ICGD (Consorcio Internacional del Genoma del Cáncer) [12], CCLE (Enciclopedia de líneas celulares de cáncer) [13] y CLUE (anteriormente Lincscloud) [14, 15, 16], que está relacionada con el reposicionamiento de fármacos y almacena los perfiles transcripcionales de células tratadas con cada uno de los fármacos usados en la clínica. Esto se consiguió analizando todos los datos públicos disponibles y se determinó mediante un modelo matemático que, analizando la expresión de mil genes, es posible estimar la expresión del resto del genoma por inferencia estadística, lo cual tiene limitaciones, pero es asequible.

4- Para gestionar, analizar y extraer conocimiento de esa gran cantidad de datos complejos se requieren expertos formados en los tres pilares fundamentales de la Bioinformática: la Estadística, que ofrece los métodos para diseñar los experimentos y manejar los datos, la Ciencia de la Computación, que aporta las técnicas y herramientas para llevar a cabo el análisis, y la Biología, que permite realizar las preguntas adecuadas a los datos e interpretar los resultados (figura 3). Estos investigadores pueden especializarse en una de esas ramas, pero siempre deben poseer conocimientos interdisciplinares.

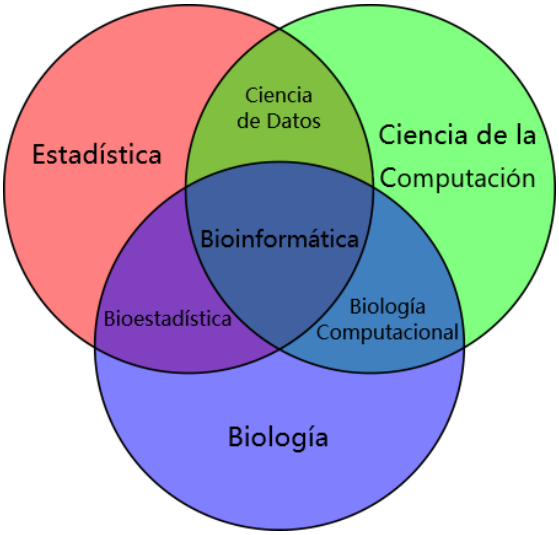

Figura 3. Pilares de la Bioinformática.

Trabajos de la Unidad de Bioinformática del GENyO

Entre otros cargos, Pedro Carmona es director de la Unidad de Bioinformática del Centro de Investigación Genómica y Oncológica (GENyO) en Granada. Su grupo de investigación está interesado en el desarrollo de **técnicas de metaanálisis**, es decir, técnicas estadísticas para integrar datos de diferentes estudios, resolver discrepancias entre ellos, identificar patrones comunes y obtener resultados más robustos al aumentar el tamaño muestral (figura 4). Sin embargo, los investigadores no expertos en el campo a veces comenten errores al aplicar estas técnicas [17], de modo que su grupo se ha dedicado a crear **herramientas** que les faciliten esa tarea. Es el caso de **MetaGenyo** [18, 19], una web que permite realizar un metaanálisis guiado que integre estudios de asociación genética heterogéneos, en los que se analiza si una variante de un polimorfismo está asociada a una patología y, por tanto, es un posible biomarcador para el diagnóstico o el diseño de fármacos. Este metaanálisis puede aplicarse a un único genotipo [20, 21] o a patologías distintas que tengan características clínicas similares para encontrar un biomarcador común [22].

También es posible detectar genes que expliquen por qué los pacientes de ciertas enfermedades tienen menos predisposición a padecer otras, lo cual se consigue invirtiendo las firmas de expresión génica de esas patologías para encontrar genes diferencialmente expresados [23]. A partir esta metodología, en el GENyO crearon otra herramienta web llamada **ImaGEO** (Integrative Meta-Analysis of GEO Data) [24, 25] que permite realizar este tipo de metaanálisis a partir de los códigos GEO de los conjuntos de datos, tras definir las etiquetas de las muestras.

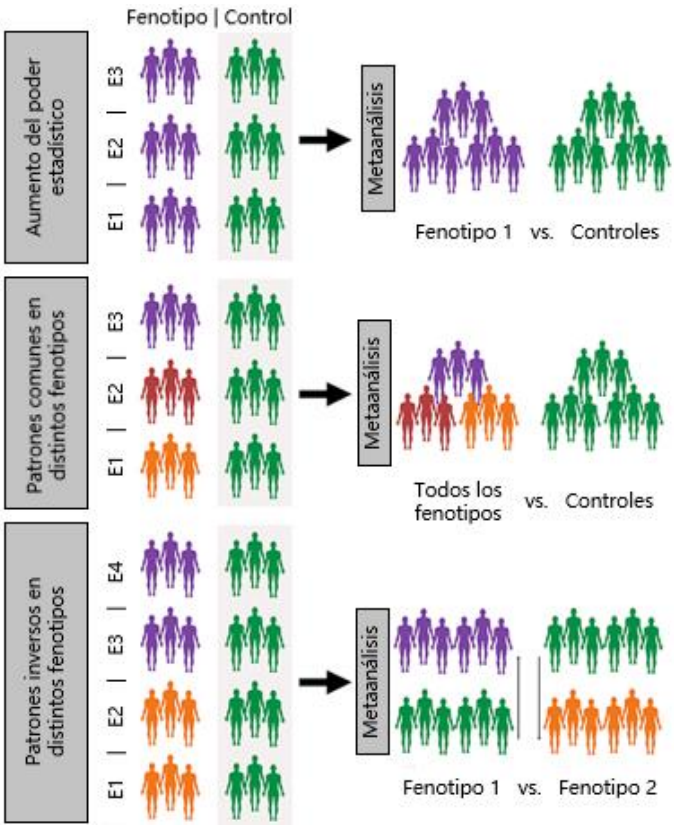

Figura 4. Aplicaciones del metaanálisis de estudios (E) de expresión génica: aumento del poder estadístico y búsqueda de patrones comunes o inversos entre diferentes fenotipos [26].

Otra posibilidad es comparar la firma génica de una patología con los datos de Lincscloud mediante el método no paramétrico GSEA (gene set enrichment analysis) [27]. De esta forma, es posible encontrar fármacos que expresen los mismos genes que la enfermedad y, por tanto, ayuden a identificar su mecanismo de acción o, por el contrario, fármacos con una firma génica inversa que sean interesantes como tratamiento [21]. Para aplicar esta metodología al reposicionamiento de fármacos en enfermedades autoinmunes, Pedro Carmona participó en la creación de una herramienta de priorización de fármacos para la inmunomodulación llamada **DREIMT** [28, 29]. Así mismo, su grupo del GENyO contribuyó al desarrollo de la plataforma **ADEx**

(Autoimmune Diseases Explorer) [30] que integra toda la información disponible en las bases de datos públicas sobre estas enfermedades, para permitir a los investigadores analizar patrones de expresión de esos genes y buscar los que actúen en toda la ruta.

Por último, movidos por la alta dimensionalidad de los datos biomédicos, desarrollaron una metodología para analizarlos basada en una técnica de factorización no negativa de matrices. En ella, la matriz original se descompone en una submatriz de factores y otra donde cada muestra está representada en el espacio factorial. Esos factores codifican partes locales de los datos y en expresión génica corresponden a biclusters o submatrices (conjuntos de variables coexpresadas en subconjuntos de muestras experimentales). Esto permite integrar distintos tipos de datos y llevar a cabo estudios de **clustering** (agrupamientos). Se ha empleado, por ejemplo, para caracterizar una línea celular inmune por un tipo de marcador, eliminando la problemática de multiplataforma y otros efectos que incluyan confusión al análisis [31, 32].

Podemos concluir por tanto que las tecnologías *ómicas* han abierto todo un campo de nuevas posibilidades para la investigación biomédica y que, para obtener conclusiones útiles a partir de la enorme cantidad de información que ofrecen, es indispensable la ayuda de herramientas bioinformáticas complejas. Por ello, la labor de profesionales como Pedro Carmona y su grupo es tan relevante hoy en día, al poner estas herramientas al alcance de cualquier investigador que las necesite.

BIBLIOGRAFÍA

1. https://vjornadas.ugrbioinformatics.com/
2. International Human Genome Sequencing Consortium, Whitehead Institute for Biomedical Research, Center for Genome Research, Lander, E., Linton, L.M., Birren, B., ... , Szustakowki, J. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921. https://doi.org/10.1038/35057062
3. Ng, S., Buckingham, K.J., Lee, C., Bigham, A.W., Tabor, H.K., Dent, K.M., ... Bamshad M.J. (2010). Exome sequencing identifies the cause of a mendelian disorder. *Nat. Genet.*, 42, 30–35. https://doi.org/10.1038/ng.499
4. Colomer, R., Mondejar, R., Romero, N., Alfranca, A., Sanchez, F. & Quintela M. (2020). When should we order a next generation sequencing test in a patient with cancer? *EClinicalMedicine*, 25:100487. https://doi.org/10.1016/j.eclinm.2020.100487

5. Marquart, J., Chen, E.Y. & Prasad, V. (2018). Estimation of the percentage of US patients with cancer who benefit from genome-driven Oncology. *JAMA Oncol.*, 4(8):1093–1098. https://doi.org/10.1001/jamaoncol.2018.1660
6. Cook, C., Stroe, O., Cochrane, G., Birney, E., & Apweiler R. (2020). The European Bioinformatics Institute in 2020: Building a global infrastructure of interconnected data resources for the life sciences. *Nucleic acids research*, 48(D1):D17–D23. https://doi.org/10.1093/nar/gkz1033
7. https://www.ncbi.nlm.nih.gov/geo/
8. https://digitalworldbiology.com/blog/bio-databases-2020-viruses-and-covid-19
9. https://www.ebi.ac.uk/arrayexpress/
10. https://www.ncbi.nlm.nih.gov/
11. https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga
12. https://daco.icgc.org/
13. https://portals.broadinstitute.org/ccle
14. https://clue.io/
15. https://portals.broadinstitute.org/cmap/
16. Subramanian, A., Narayan, R., Corsello, S.M., Peck, D.D., Natoli, T.E., Lu, X., ... Golub, T.R. (2017). A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell*, 171(6):1437-1452.e17. https://doi.org/10.1016/j.cell.2017.10.049
17. Park, J.H, Eisenhut, M., van der Vliet, H.J. & Shin, J.I. (2017). Statistical controversies in clinical research: overlap and errors in the meta-analyses of microRNA genetic association studies in cancers. *Annals of Oncology*, 28: 1169–1182. https://doi.org/10.1093/annonc/mdx024
18. http://bioinfo.genyo.es/metagenyo/
19. Martorell, J., Toro, D., Alarcón, M.E. & Carmona, P. (2017). MetaGenyo: A web tool for meta-analysis of genetic association studies. *BMC Bioinformatics*, 18:563. https://doi.org/10.1186/s12859-017-1990-4
20. O'Mara, T., Zhao, M. & Spurdle, A. (2016). Meta-analysis of gene expression studies in endometrial cancer identifies gene expression profiles associated with aggressive disease and patient outcome. *Sci. Rep.*, 6, 36677. https://doi.org/10.1038/srep36677

21. Toro, D., Carmona, P. & Alarcón, M.E. (2017) Support for phosphoinositol 3 kinase and mTOR inhibitors as treatment for lupus using in-silico drug-repurposing analysis. *Arthritis Res. Ther.*, 19, 54. https://doi.org/10.1186/s13075-017-1263-7
22. Toro D, Carmona P, Alarcon ME. (2014) Shared signatures between rheumatoid arthritis, systemic lupus erythematosus and Sjögren's syndrome uncovered through gene expression meta-analysis. Arthritis Res Ther, 16(6):489. https://doi.org/10.1186/s13075-014-0489-x
23. Sánchez, J., Tejero, H., Ibáñez, K., Portero, J.L., Krallinger, M., Al-Shahrour, F., … Valencia, A. (2017). A molecular hypothesis to explain direct and inverse co-morbidities between Alzheimer's Disease, Glioblastoma and Lung cancer. *Sci. Rep.*, 7, 4474. https://doi.org/10.1038/s41598-017-04400-6
24. http://bioinfo.genyo.es/imageo/
25. Toro, D., Martorell, J., López, R., García, A., González, V., Alarcón, M.E. & Carmona P. (2019). ImaGEO: Integrative Gene Expression Meta-Analysis from GEO database. *Bioinformatics*, 35(5):880–882. https://doi.org/10.1093/bioinformatics/bty721
26. Toro, D., Villatoro, J.A., Martorell, J., Román, Y., Alarcón, M. & Carmona, P. (2020). A survey of gene expression meta-analysis: methods and applications. *Briefings in Bioinformatics*, bbaa019(3). https://doi.org/10.1093/bib/bbaa019
27. www.gsea-msigdb.org
28. http://www.dreimt.org/
29. Troulé, K., López, H., García, S., Reboiro, M., Carretero, C., Martorell, J., … Gómez, G. (2020). DREIMT: a drug repositioning database and prioritization tool for immunomodulation. *Bioinformatics*, 1367-4803. https://doi.org/10.1093/bioinformatics/btaa727
30. https://adex.genyo.es/
31. Carmona, P., Varela, N., Luque, M.J., Toro, D., Martorell, J., Alarcón, M.E. & Marañón C. (2017). Metagene projection characterizes GEN2.2 and CAL-1 as relevant human plasmacytoid dendritic cell models. *Bioinformatics*. 33(23):3691-3695. https://doi.org/10.1093/bioinformatics/btx502
32. Barturen, G., Babaei, S., Català, F., Martínez, M., Makowska, Z., Martorell, J., … Alarcón, M.E. (2020). Integrative Analysis Reveals a Molecular Stratification of Systemic Autoimmune Diseases. *Arthritis Rheumatol*. https://doi.org/10.1002/art.41610

ENTREVISTA A PEDRO CARMONA SÁEZ

***My Scientific Journal* - Durante su intervención en las V Jornadas de Bioinformática de la Universidad de Granada explicó que los profesionales de su área pueden especializarse en alguno de los pilares fundamentales de la Bioinformática, pero deben poseer conocimientos básicos de todos ellos. Usted, por ejemplo, se licenció en Biología, especializándose en Bioquímica y Biología molecular, y realizó su Doctorado en la unidad de Biocomputación del CNB (Centro Nacional de Biotecnología). ¿Si tuviera que recomendarle a un futuro Bioinformático una especialidad, cuál sería y por qué?**

Pedro Carmona Sáez - Independientemente de lo que estudies, necesitarás formación complementaria en otras áreas, las personas que estudien bioquímica, biomedicina o disciplinas afines necesitarán aprender estadística y computación, y viceversa. Es por esto que cualquier formación en estos campos en principio es válida. Lo que sí hay hoy en día son diversos másteres específicos de bioinformática y análisis de datos, que sí son una muy buena vía de formación para la gente que luego se quiere dedicar profesionalmente a la bioinformática. En este tipo de másteres se suelen plantear diferentes itinerarios de forma que se da una formación en programación y estadística a biólogos y bases de biología molecular a estudiantes que provienen de grados más técnicos.

***MSJ* - Dado el enorme potencial de las tecnologías *ómicas* y su creciente extensión, ¿cabría esperar su pronta integración en la práctica de la medicina cotidiana?**

PCS - Sí, sin duda. De hecho, las tecnologías de secuenciación masiva se están usando ya para el análisis de las alteraciones en el genoma de pacientes de diferentes patologías. En nuestro país aún no están integradas como una prueba diagnóstica que el clínico pueda solicitar de forma "rutinaria", pero sin duda lo veremos en un corto espacio de tiempo. Esto requerirá también de personal con formación específica para llevar a cabo el análisis de estos resultados y poder elaborar los informes que den la información necesaria a los especialistas para la toma de decisiones en la clínica, tanto para diagnóstico como para poder proponer tratamientos más específicos.

***MSJ* - Las metodologías y herramientas bioinformáticas que han desarrollado usted y sus compañeros son de gran utilidad en el ámbito biomédico, lo cual se refleja en los múltiples estudios que las han empleado. Sin embargo, ¿considera que se esté explotando al máximo su potencial tanto en investigación como en la práctica clínica?**

PCS - Las herramientas que hemos desarrollado están teniendo un gran impacto, pero más en el ámbito de la investigación que en la clínica. Aplicaciones como MetaGenyo, que se usa para integrar estudios de asociación genética, o IMAGEO, para integrar estudios de expresión génica, están siendo ampliamente utilizadas por cientos de investigadores de diversas partes del mundo. Estos investigadores las usan como herramientas para llevar a cabo el análisis de sus datos y los resultados que ayudan a obtener se están publicados en numerosas revistas científicas de primer nivel.

***MSJ* - ¿Podrían, además, algunas de estas herramientas aplicarse al manejo de otro tipo de datos o servir de base para crear nuevas herramientas con ese propósito? En caso afirmativo, ¿puede darnos algunos ejemplos?**

PCS - Sí podrían en algunos casos usarse para otro tipo de datos, como por ejemplo la aplicación de MetaGenyo. Esta herramienta implementa un conjunto de técnicas estadísticas para llevar a cabo estudios de meta-análisis que se basan en integrar información generada en diferentes trabajos para buscar patrones y efectos comunes. Aunque en principio está automatizada para el análisis de datos de asociación genética, podría adaptarse para cualquier otro ámbito. En este contexto, es importante mencionar que las herramientas las desarrollamos con código abierto y las funcionalidades y los códigos los hacemos disponibles para que otros investigadores los puedan usar y adaptar para sus estudios.