Calidad de los fenotipos humanos en pacientes con discapacidad intelectual con estudios de secuenciación de exoma en un hospital pediátrico peruano

Material Suplementario 1.

Fórmulas y pruebas estadísticas utilizadas para evaluar la calidad del fenotipo en pacientes con discapacidad intelectual en un hospital pediátrico

Uso del contenido de información (IC)[10],

Se utiliza como una medida para cuantificar la especificidad, el valor informativo de dicho término y comprender las características particulares de cada grupo de pacientes[10].

El IC se calcula utilizando la siguiente fórmula:

$$IC\,(t) = -\log_{10}\left(\frac{\text{frecuencia del término } t}{\text{número total de términos}}\right)$$

Donde:

- *IC (t)* representa el contenido de información del término t.
- Frecuencia del término t es el número de veces que el término t aparece en el conjunto de datos, específicamente en pacientes dentro de un grupo particular.
- Número total de términos denota el número total de ocurrencias de todos los términos en el conjunto de datos.

Índice de especificidad del conjunto de datos (DsI) [9]

Para el cálculo de Dsl se tiene la siguiente fórmula:

DsI = HsS/LsS

Donde

LsS (Puntaje general de la sección baja) = $\frac{\sum_{L=1}^{Lmax} dL*(Lmax - L + 1), \text{if } dL > 0}{LS}$

El LsS pondera los puntajes dL para la sección baja de la ontología, asignando pesos inversamente al nivel de la ontología para penalizar términos inespecíficos[9].

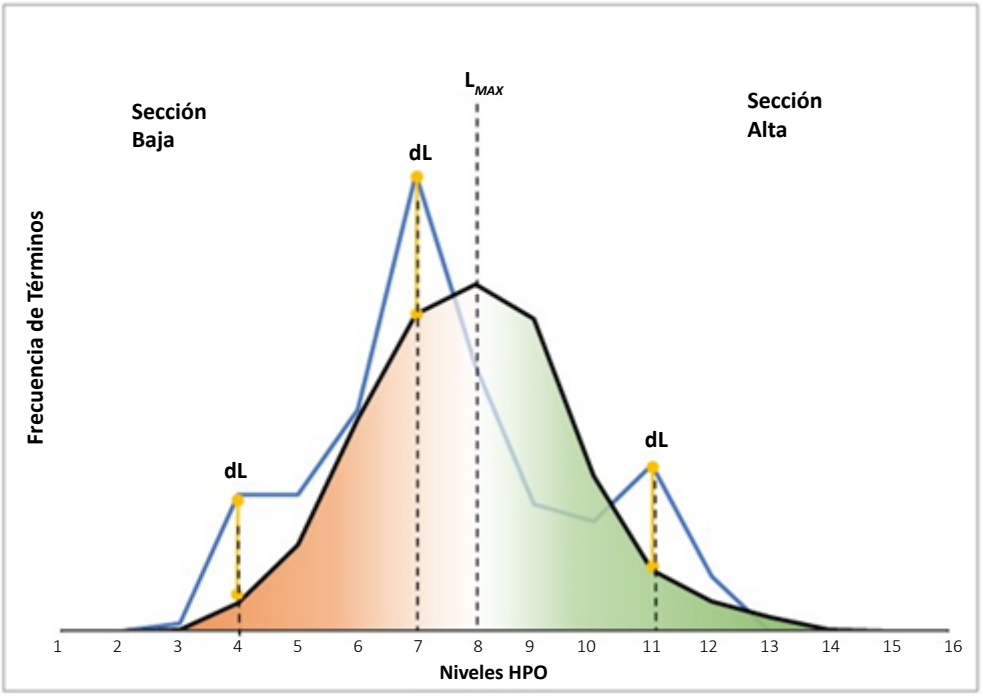
HsS (Puntaje general de la sección alta) = $\frac{\sum_{L=L=Lmax+1\ 1}^{Lo} (dL*(Lo - L)), \text{if } dL > 0}{LS}$

Lmax= el nivel más alto de HPO.

dL (puntajes para ambas secciones (alta y baja) = PobsL – PontL.

El puntaje de diferencia (dL) se calcula para cada nivel en la jerarquía. Esto representa la diferencia entre la proporción de términos observados en la cohorte (PobsL) y la proporción de términos en la ontología (PontL) para un nivel dado L. Los puntajes dL se utilizan para calcular puntajes generales para las dos secciones. Solo los puntajes dL mayores a cero contribuyen al puntaje de sección. Esto pondera los puntajes dL para la sección alta, asignando pesos proporcionales con respecto a Lmax, para recompensar términos específicos[9].

Material Suplementario 2.



Distribución a nivel de término para el HPO (línea negra) y para una cohorte hipotética de pacientes (línea azul). Los niveles de la ontología se dividen en dos secciones, la sección Baja y la sección alta. El motivo de la división se basa en la probabilidad de que se utilice un nivel, la cual está relacionada con la cantidad de niveles que lo separan del nodo raíz en la ontología, y la cantidad de términos contenidos en el nivel *Lmax* es el nivel con mayor número de términos en la ontología y es el último nivel para el cual un fenotipado más profundo conduce a un mayor número de términos posibles. Los siguientes niveles tienen recuentos de términos decrecientes, por lo que tienen menores probabilidades de aparecer en el perfil de un paciente. Al dividir la ontología en secciones, podemos ver cómo una cohorte de pacientes reales usa la ontología midiendo las diferencias de distribución en cada nivel (*dL*). Estas diferencias se ponderan teniendo en cuenta las secciones y cuántos términos pertenecen a cada nivel, y se utilizan para calcular el DsI para medir la información fenotípica de una cohorte de pacientes. Fuente: Traducido de Rojano *et al.*[9].

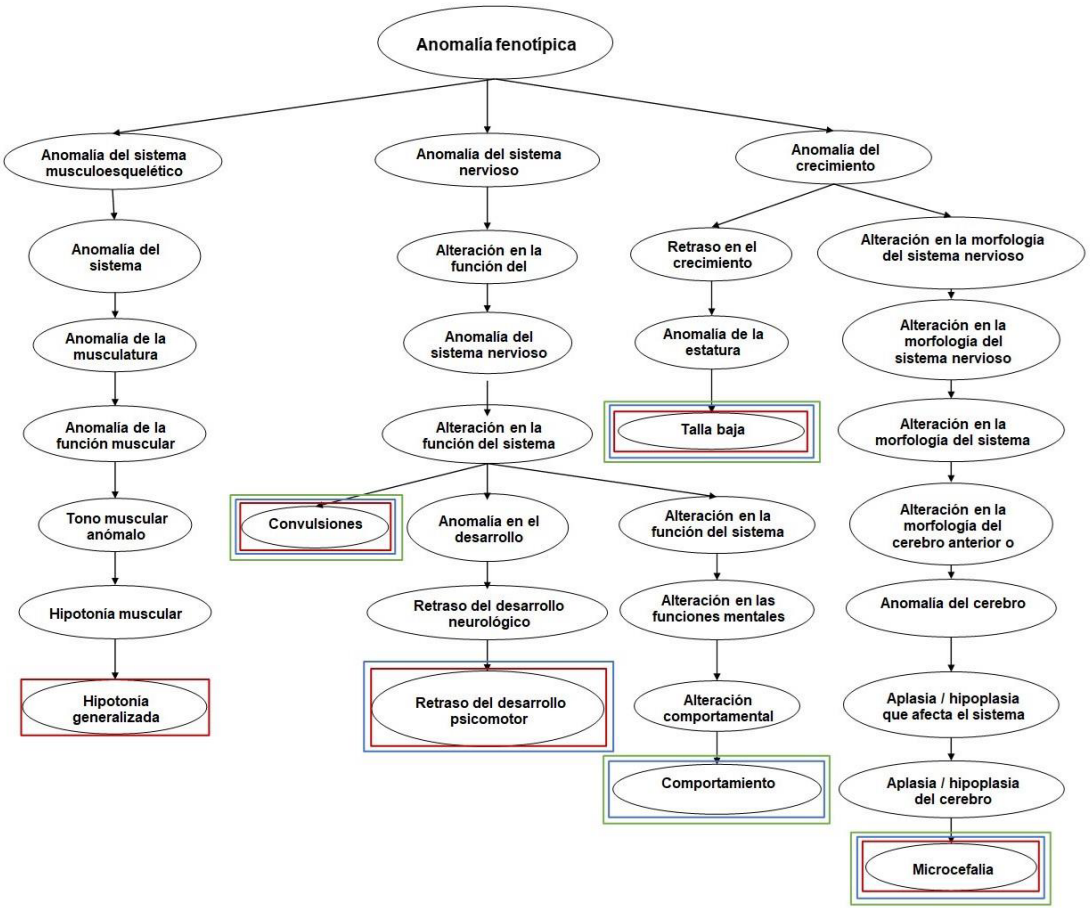
Material Suplementario 3.

Los cinco términos HP0 más frecuentes en los conjuntos de datos “general”, “sin variantes”, “patogénico o probablemente patogénico” y “VUS”.

	General	%	Sin variantes	%	Patogénico o probablemente patogénico	%	VUS	%
1	Retraso del desarrollo psicomotor	14,3%	Retraso del desarrollo psicomotor	12,9%	Retraso del desarrollo psicomotor	15,9%	Retraso del desarrollo psicomotor	13,8%
2	Microcefalia	4,9%	Microcefalia	5,2%	Convulsiones	5,5%	Microcefalia	4,7%
3	Convulsiones	4,4%	TEA	3,3%	Microcefalia	4,5%	Convulsiones	4,7%
4	Talla baja	3,4%	Talla baja	3,3%	Hipotonía generalizada	2,9%	TEA	3,7%
5	TEA	3,0%	Convulsiones	2,9%	Talla baja	2,9%	Talla baja	3,7%

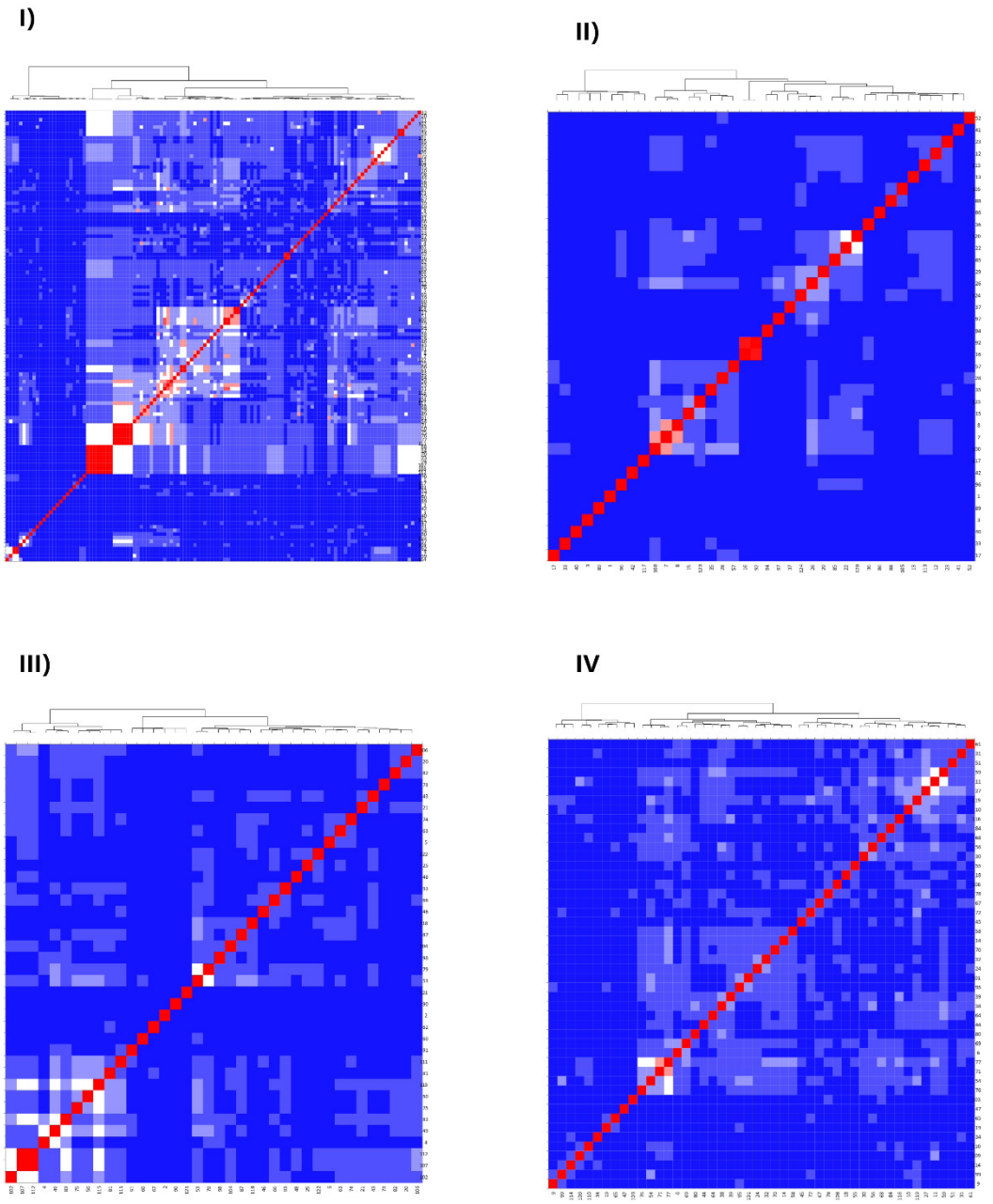
TEA= Trastorno del espectro autista

Material Suplementario 4.



Representación esquemática de los cinco términos HPO más frecuentes para cada cohorte en relación con el nodo "anomalía fenotípica". Los cuadrados de colores representan términos para cada cohorte. Rojo: patogénico o probablemente patogénico. Azul: sin variantes. Verde: variante de significado incierto (VUS). (Hipotonía generalizada= 8, convulsiones=5, retraso del desarrollo psicomotor=8, talla baja=5, Comportamiento=9, microcefalia=10).

Material Suplementario 5.



Mapas de calor representando la matriz de similitud de pacientes calculados con la medida de similitud de Jaccard para cada par de perfiles. La columna coloreada muestra los grupos de pacientes identificados por el algoritmo de agrupación. Sólo se utilizaron pacientes con tres o más términos HPO para construir la matriz de similitud. (I) general, (II) sin variantes, (III) patogénico o probablemente patogénico y (IV) VUS. Los valores en el mapa de calor varían de cero (representado por el color azul) a uno (color rojo), y el blanco en el espectro significa similitud intermedia.

Material Suplementario 6.

Gráfico de dispersión con puntuaciones factoriales del ACP.

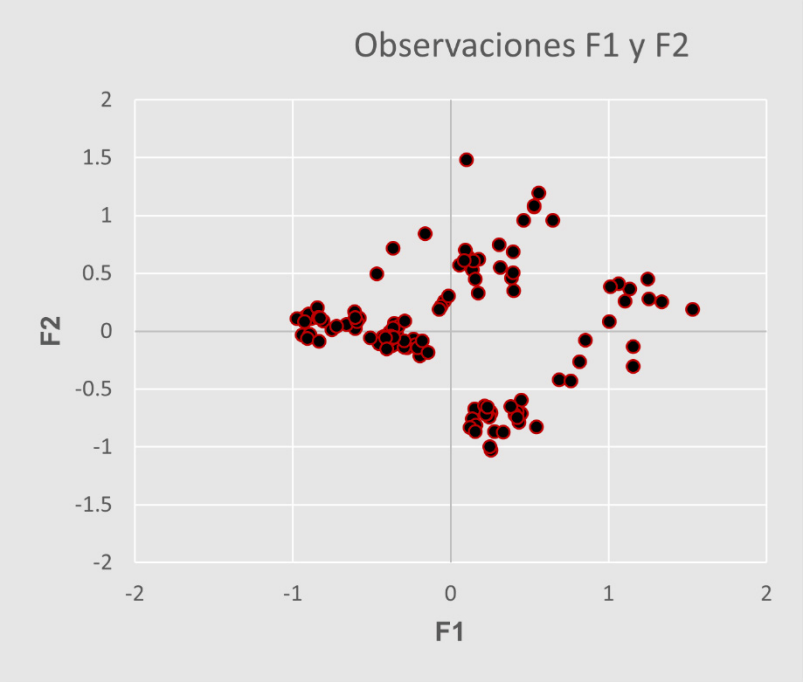
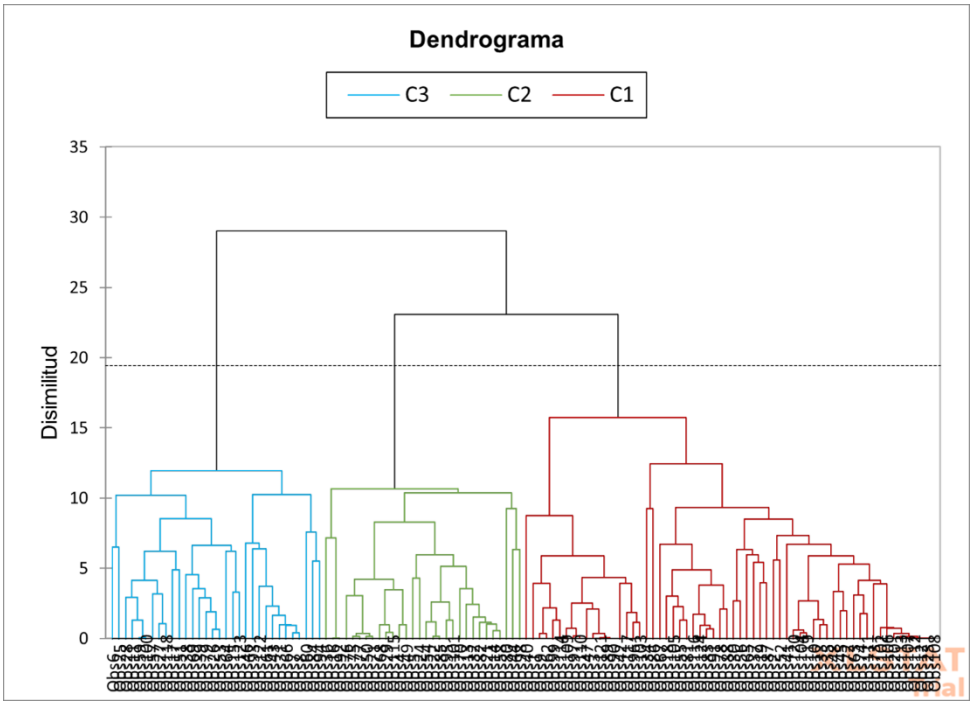


Gráfico de dispersión que representa las puntuaciones factoriales F1 y F2 resultantes del Análisis de Componentes Principales (ACP). Cada punto en el gráfico, identificado como un círculo negro con borde rojo, representa una observación o paciente en nuestro estudio.

Material Suplementario 7.



Dendrograma obtenido tras el Análisis de Clusterización Jerárquica (ACJ), representando una estructura jerárquica claramente definida de toda la cohorte. Este diagrama visualiza la relación de similitud y disimilitud entre las observaciones (pacientes), estratificando la muestra en tres grupos identificados por colores distintivos. El Clúster 1, representado en rojo, el Clúster 2 en verde y el Clúster 3 en celeste, emergen como segmentos claves de esta estructura. La altura de las ramas en el dendrograma representa la distancia entre los grupos y ofrece información sobre la similitud entre los grupos. Cuanto menor sea la unión de dos grupos en el dendrograma, mayor será la similitud entre ellos.

Material Suplementario 8.

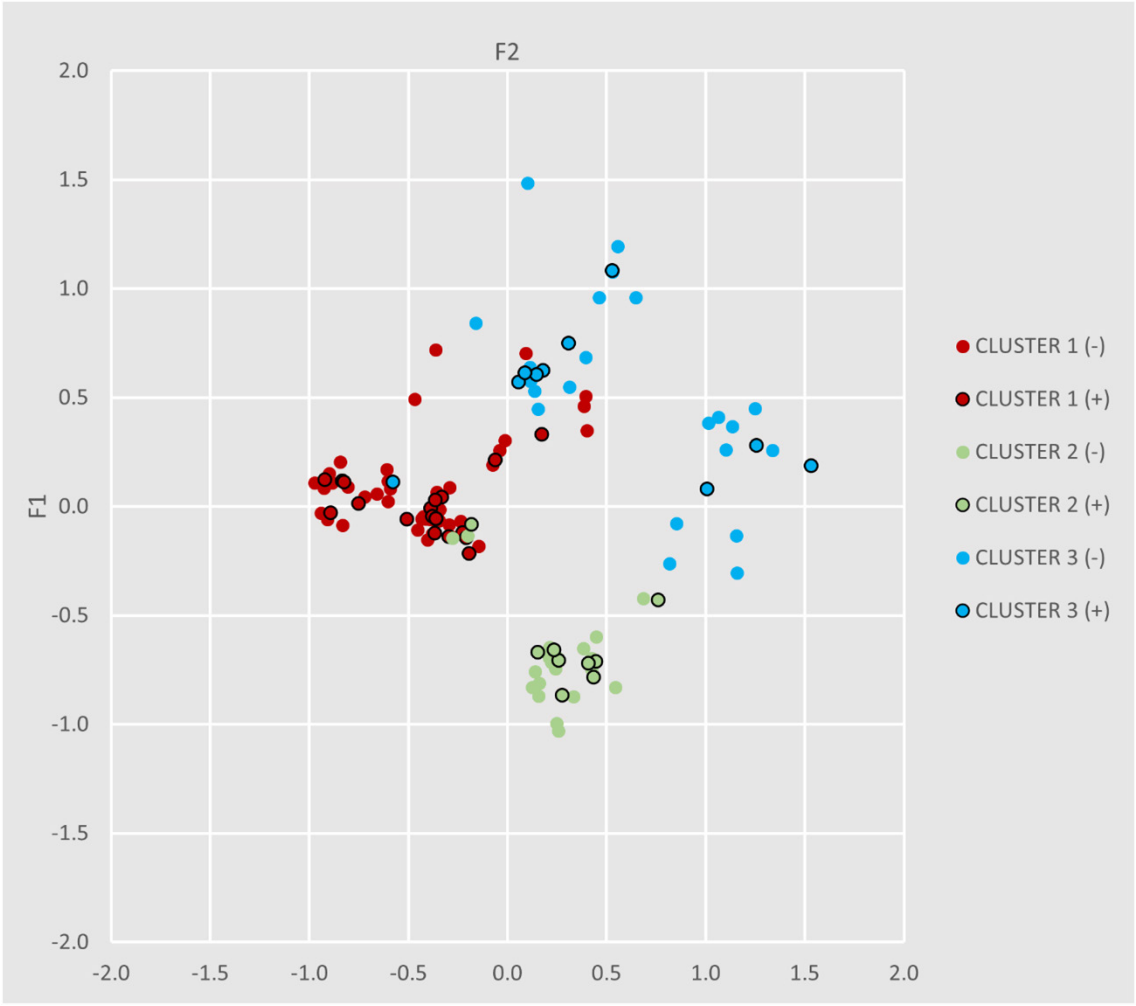
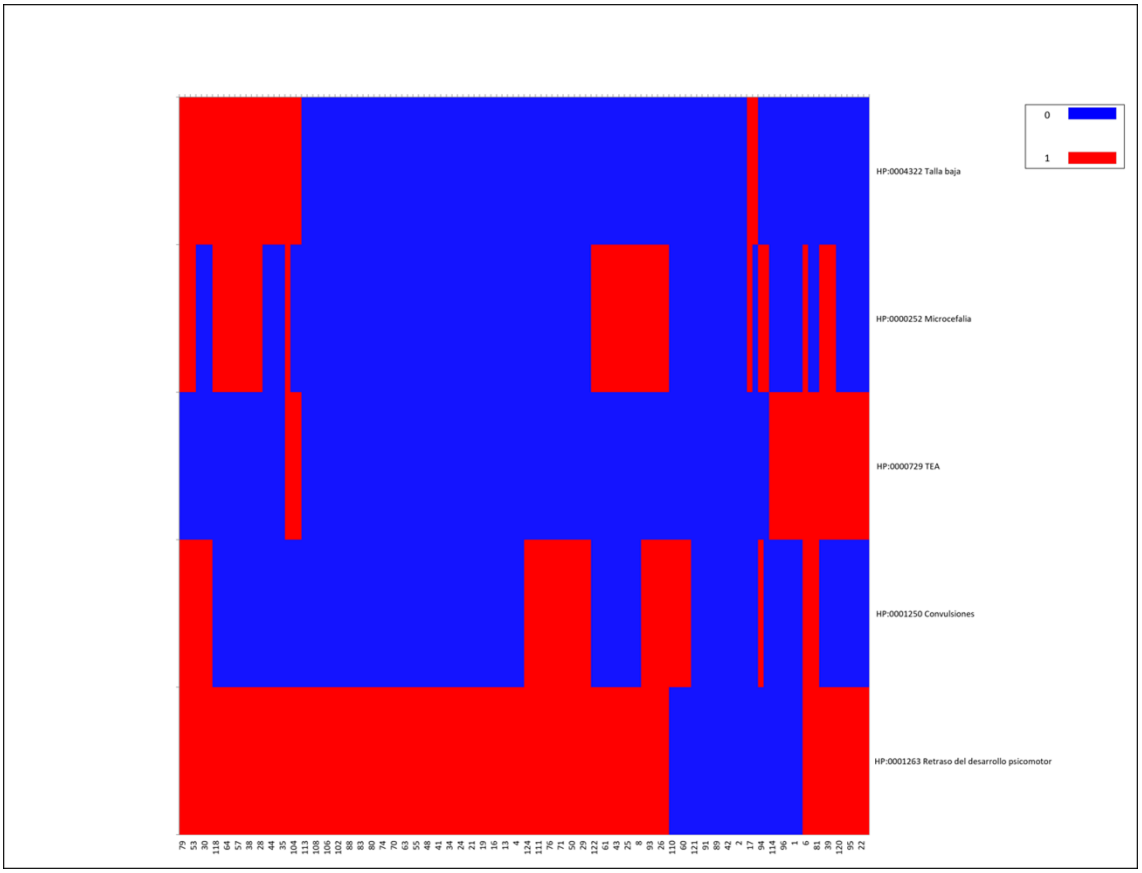


Gráfico de dispersión representando la distribución de pacientes en un espacio bidimensional. Cada punto representa un paciente, y la diferenciación entre tres clústeres se destaca con distintos colores: el Grupo 1 en rojo, el Grupo 2 en verde y el Grupo 3 en celeste. La presencia del borde negro indica que el término asociado es positivo, mientras que la ausencia de borde señala que el término es negativo.

Material Suplementario 9.



Mapa de calor con la disposición visual de los cinco fenotipos más prevalentes en la cohorte completa, dispuestos a la derecha del gráfico. La distribución de cada paciente se traza a lo largo del eje horizontal, utilizando la mitad de los individuos de la cohorte. Cada celda del mapa de calor tiene sólo dos colores: rojo, que indica la existencia del rasgo, y azul, que indica su ausencia. Esta forma simplificada permite una explicación fácil y clara de la presencia o ausencia de fenotipos específicos en cada paciente, ofreciendo una perspectiva gráfica de la distribución de ciertos rasgos fenotípicos en la cohorte.