PEC1

Rashid Babiker Sánchez

14 de abril, 2020

Contents

Resumen

En el siguiente estudio se usarán los datos de publicaciones sobre leucemia (Weniger et al. (2018), Brune et al. (2008), Giefing et al. (2013)) para medir y comparar los patrones de expresión de **células CB**, precursoras sanas de linfocitos B maduros, con dos tipos de linfocitos tumorales: **células NLPHL**, extraídas de pacientes con linfoma de Hodkin con predominio de linfocitos nodulares; y **células cLH** del linfoma de Hodkin clásico. Los datos usados están disponibles en la base de datos GEO, en el siguiente enlace: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE12453

El estudio elegido es el siguiente: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE12453

El repositorio con todos los análisis realizados se puede consultar en el siguiente enlace https://github.com/RashBabiker/PEC1

Objetivos

Con este estudio se quieren analizar 2 cuestiones:

- ¿Presentan las células tumorales un patrón de expresión distinto a las células sanas? Esta información tiene valor diagnóstico, se comprobará con mapas de calor ordenados.

En el siguiente estudio se comparan los patrones de expresión de los centroblastos (precursores de linfocitos B maduros, CB), con dos tipos de linfocitos tumorales: celulas LH, extraidas de pacientes con linfoma de Hodkin con predominio de linfocitos nodulares (NLPHL) y celulas cLH del linforma de Hodkin clásico. Los datos usados están disponibles en la base de datos GEO, en el siguiente enlace https://www.ncbi.nlm.nih.gov/geo/geo2r/?acc=GSE12453

Repositorio online de este proyecto, con el resto de códigos utilizados, material usado y resultados está en el siguiente enlace: https://github.com/RashBabiker/PEC1.git

Con este estudio se quieren analizar 2 cuestiones:

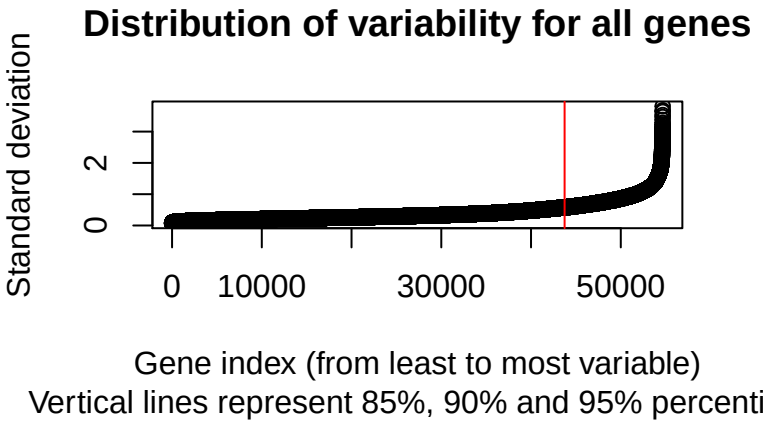
- ¿Presentan las celulas tumorales un patrón de expresión distinto a las celulas sanas? Esta información tiene valor diagnóstico, se comprobará con mapas de calor ordenados.
- ¿Se parecen las células tumorales entre sí? Si es así podrían tener un origen común.

Materiales y Métodos

Se usan 22 muestras: 5 de células sanas, 5 de células NLPHL y 12 de células cHL. Las muestras se obtuvieron de tejido de amígdala de los pacientes y donadores sanos, posteriormente se extrajo el ARN, se amplificó, se retrotranscribió a cDNA, se fragmentó e hibridó con el microarray GeneChip Human Genome U133 Plus. 2.0 de affymetrix. De 67 muestras originales, de distintos tipos de células sanas y afectadas por algún tipo de leucemia, se han elegido todas las muestras de células sanas (5 muestras), NLPHL (5 muestras) y cHL (12 muestras). Brune et al. (2008)

A continuación, se exponen los pasos seguidos, también se indica el nombre de los archivos de Rmarkdown del repositorio donde se puede acceder el código usado para la realización de cada tarea con una explicación mucho más detallada:

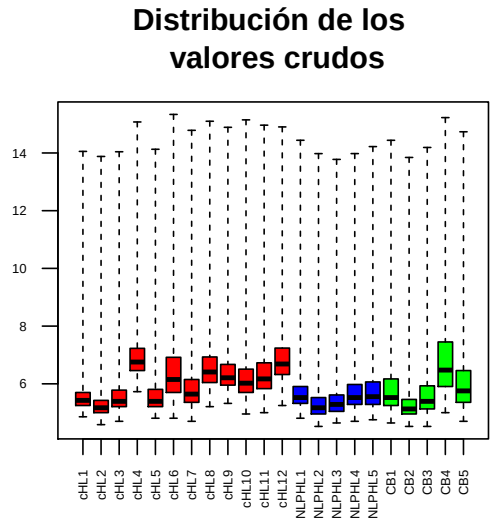
1. **Obtención de los targets (código en "01 targets.Rmd"):** Adaptando la información fenotípica de los datos obtenida usando el paquete GEOquery.
2. **Preparación de las muestras (código en "02 preparacion de las muestras.Rmd"):** Primero se descargan los archivos crudos (.cel) de https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE12453. Una vez obtenidos se combinan con los targets para preparar el ExpressionSet necesario para hacer los análisis.
3. **Control de calidad y normalización de los datos (código en "03 Control de calidad y normalizacion.Rmd"):** El análisis de calidad se realiza con el paquete arrayQualityMetrics, una función que construye boxplot, PCAs y otras medidas para analizar la variabilidad de las muestras y detectar valores atípicos, la normalización usada es rma (Robust Multichip Analysis), que sigue tres pasos: corregir ruido de fondo (background), normalizar y sumarizar.
4. **Filtrado inespecífico (código en "04 Filtrado.Rmd"):** Se eliminan los genes cuya variación se puede atribuir a la variación aleatoria para aumentar la potencia de los análisis posteriores, en este paso se han eliminado el 80% de los genes, manteniendo el 20% que presenta mayor variabilidad, a partir de la línea roja en la siguiente figura:

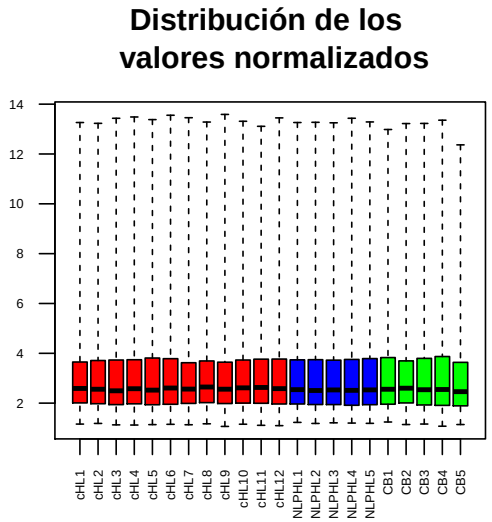


5. **Análisis de expresión (código en "05 analisis de expresion.Rmd"):** Identificación de genes diferencialmente expresados en alguna condición y comparaciones de expresión entre condiciones usando modelos lineales. representados con diagramas de Venn y mapas de calor.
6. **Análisis biológico de los resultados (código en "06 analisis biologico de los resultados.Rmd"):** Usando análisis de enriquecimiento, un método que, a partir de una lista de genes, en este caso genes diferencialmente expresados en las distintas comparaciones, localiza las funciones, procesos biológicos o pathways más frecuentes.

Resultados y discusión

La normalización reduce el error entre muestras. La distribución es más homogénea.





Y reduce el número de valores atípicos.

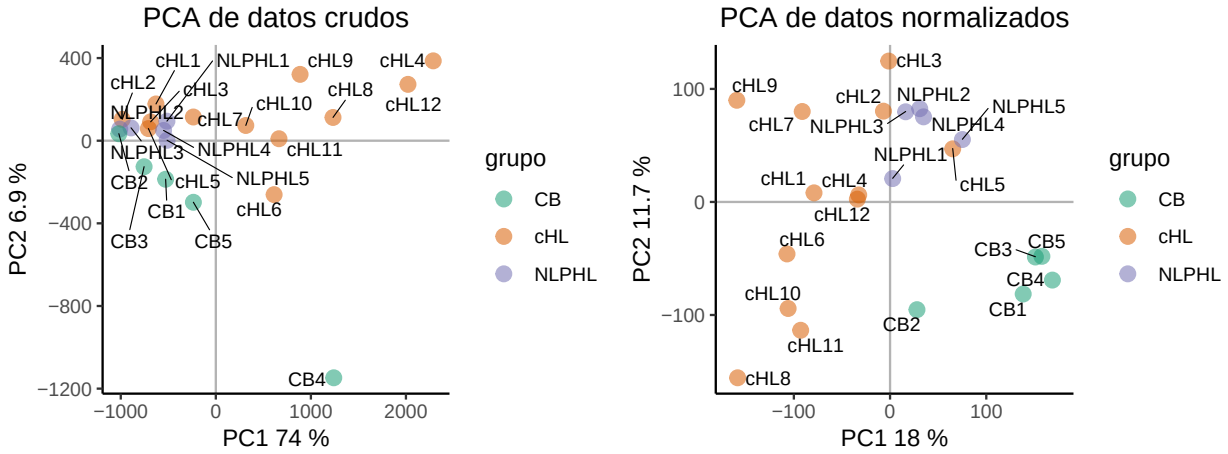
Datos Crudos

array	sampleNames	*1	*2	*3	geo_accession	group
1	cHL1				GSM312811	cHL
2	cHL2			x	GSM312812	cHL
3	cHL3				GSM312813	cHL
4	cHL4			x	GSM312814	cHL
5	cHL5				GSM312815	cHL
6	cHL6			x	GSM312816	cHL
7	cHL7				GSM312817	cHL
8	cHL8			x	GSM312818	cHL
9	cHL9			x	GSM312819	cHL
10	cHL10			x	GSM312820	cHL
11	cHL11			x	GSM312821	cHL
12	cHL12			x	GSM312822	cHL
13	NLPHL1				GSM312823	NLPHL
14	NLPHL2			x	GSM312824	NLPHL
15	NLPHL3				GSM312825	NLPHL
16	NLPHL4				GSM312826	NLPHL
17	NLPHL5				GSM312839	NLPHL
18	CB1				GSM312937	CB
19	CB2			x	GSM312938	CB
20	CB3				GSM312939	CB
21	CB4	x		x	GSM312940	CB
22	CB5			x	GSM312941	CB

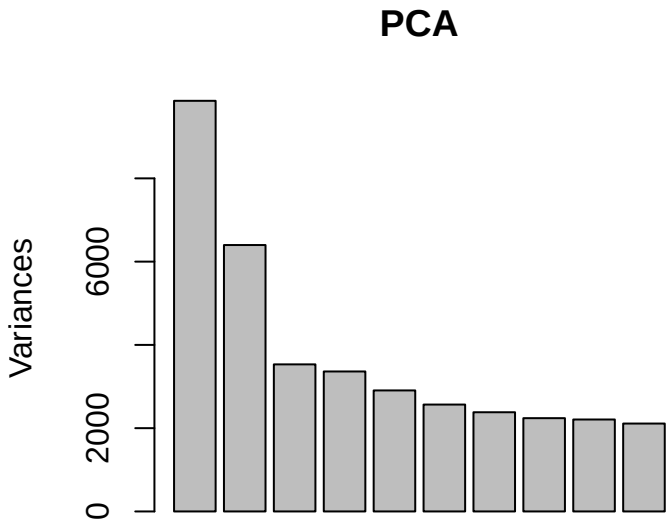
Datos normalizados

array	sampleNames	*1	*2	*3	geo_accession	group
1	cHL1		x		GSM312811	cHL
2	cHL2				GSM312812	cHL
3	cHL3				GSM312813	cHL
4	cHL4				GSM312814	cHL
5	cHL5				GSM312815	cHL
6	cHL6				GSM312816	cHL
7	cHL7				GSM312817	cHL
8	cHL8				GSM312818	cHL
9	cHL9				GSM312819	cHL
10	cHL10				GSM312820	cHL
11	cHL11				GSM312821	cHL
12	cHL12				GSM312822	cHL
13	NLPHL1				GSM312823	NLPHL
14	NLPHL2				GSM312824	NLPHL
15	NLPHL3				GSM312825	NLPHL
16	NLPHL4				GSM312826	NLPHL
17	NLPHL5				GSM312839	NLPHL
18	CB1				GSM312937	CB
19	CB2				GSM312938	CB
20	CB3				GSM312939	CB
21	CB4				GSM312940	CB
22	CB5				GSM312941	CB

Estos cambios implican una mayor diferenciación entre grupos según sus niveles de expresión, como se puede comprobar con un análisis de componentes principales



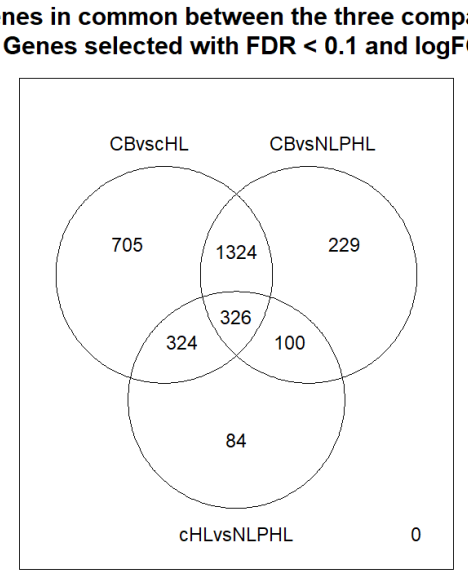
El porcentaje de varianza explicada en los datos normalizados es mucho menor. En este caso si que parece que puede ser útil una representación en 3D, para explicar el 36.2%.



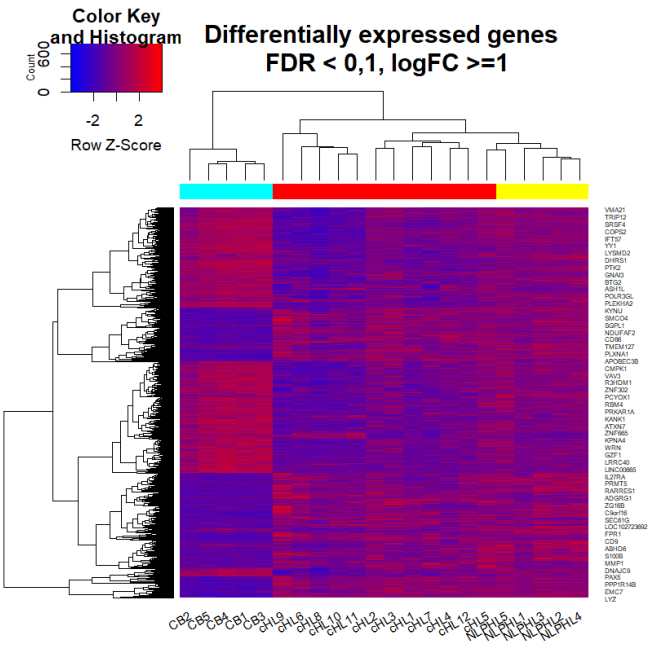
El PCA en 3D se puede consultar en el siguiente enlace, NLPHL5 y cHL5 tienen patrones de expresión similar, puede que hubiera un error en la caracterizacion de las celulas cHL5:

https://rawcdn.githack.com/RashBabiker/PEC1/483774eb655f385f96165b0fc03351ad4cad7f4d/00%20Resultados/PCA%20normalizado%203D.html#L1

El diagrama de Venn muestra diferencias en el número de genes diferencialmente expresados en las distintas condiciones, las células sanas (CB) presentan una expresión diferencial respecto a las células tumorales de 3008 genes, presentando una expresión más parecida a las NLPHL (con 1979 genes diferentes) que con las cHL (con 2679 genes diferentes). 1650 genes se expresan distinto en genes tumorales que en sanas. Al comparar los dos tipos de celulas tumorales vemos que presentan expresión diferencial en 834 genes.



La expresión más similar de las células NLPHL se confirma también con el mapa de calor, la sobreexpresión de genes como los del bloque VMA21-PRLT3GLo están menos sobreexpresados en estas células que en las cHL. También se ve una clara diferencia entre tipos de células chL.



La siguiente tabla muestra los 15 procesos donde más genes se han desregulado en células cHL respecto a las células sanas. La desregularización de genes específicos de linfocitos afecta a 3 procesos distintos: i) la función inmunológica, alterando el procesamiento de antígenos y degranulación de neutrófilos impide el correcto funcionamiento del linfocito; ii) el desequilibrio del ciclo celular por alteraciones en las vías de apoptosis (señalización de NOTCH,BCR, TP53) permite a las células proliferar sin control causando el tumo y iii) alteraciones en el metabolismo energético (Ciclo de Krebs (TCA) y transporte de electrones), permiten una estabilidad a largo plazo aun consumiendo más recursos de los habituales.

Description	Count	p.adjust
Neutrophil degranulation	61	0.0305828
Transcriptional Regulation by TP53	48	0.0305828
The citric acid (TCA) cycle and respiratory electron transport	29	0.0305828
Programmed Cell Death	27	0.0332575
Apoptosis	26	0.0409765
ABC-family proteins mediated transport	20	0.0305828
Antigen processing-Cross presentation	19	0.0305828
Signaling by the B Cell Receptor (BCR)	19	0.0409765
Signaling by NOTCH4	16	0.0332575
ABC transporter disorders	15	0.0354695
The role of GTSE1 in G2/M progression after G2 checkpoint	15	0.0354695
G1/S DNA Damage Checkpoints	14	0.0354695
Pyruvate metabolism and Citric Acid (TCA) cycle	13	0.0305828
Regulation of RUNX3 expression and activity	13	0.0305828
Stabilization of p53	13	0.0332575

Al comparar las células sanas con las células NLPHL se ve algo similar, alteraciones en el metabolismo energético, función inmunológica y ciclo celular principalmente.

Description	Count	p.adjust
Cellular responses to stress	76	0.0028151
M Phase	73	0.0015182
Diseases of signal transduction	69	0.0040967
Transcriptional Regulation by TP53	68	0.0022433
Class I MHC mediated antigen processing & presentation	64	0.0086085
Infectious disease	64	0.0141858
Metabolism of amino acids and derivatives	60	0.0306945
Cell Cycle Checkpoints	56	0.0028151
Translation	55	0.0040967
Processing of Capped Intron-Containing Pre-mRNA	54	0.0002693
Antigen processing: Ubiquitination & Proteasome degradation	54	0.0126180
Organelle biogenesis and maintenance	53	0.0095098
Deubiquitination	51	0.0162142
MAPK family signaling cascades	49	0.0365957
Mitotic Anaphase	47	0.0002693

La comparación entre los dos tipos de células tumorales muestra expresión diferencial en genes que regulan el ciclo celular y función inmunológica, el número de estos genes diferentes es mayor al observado al comparar las células tumorales por separado con las células sanas, lo que sugiere que los dos tipos de cáncer tienen el mismo efecto, pero actúan de forma muy distinta. Por otro lado, no se ven diferencias en genes del metabolismo energético, por lo que parece que en ese aspecto son similares.

Sería interesante localizar los genes más sobreexpresados relacionados con el metabolismo energético en las comparaciones con las células sanas y si coinciden en los dos tipos de cáncer podría ser una diana terapéutica para tratar estos tipos de cáncer.

Description	Count	p.adjust
Neutrophil degranulation	142	0.0000002
Signaling by Interleukins	110	0.0352388
M Phase	108	0.0003628
Transcriptional Regulation by TP53	105	0.0000920
Class I MHC mediated antigen processing & presentation	100	0.0012910
Cell Cycle Checkpoints	77	0.0113577
Interferon Signaling	68	0.0000159
Mitotic Anaphase	59	0.0044877
Mitotic Metaphase and Anaphase	59	0.0048865
SUMOylation	57	0.0029660
G2/M Transition	57	0.0071559
Mitotic G2-G2/M phases	57	0.0077754
Mitotic Prometaphase	57	0.0077754
SUMO E3 ligases SUMOylate target proteins	54	0.0067151
Separation of Sister Chromatids	53	0.0193737

Conclusión:

Las células sanas muestran un patrón de expresión distinto a las células tumorales, reconocibles mediante los microarrays GeneChip Human Genome U133 Plus. 2.0 de affymetrix, por lo que es un método válido de diagnóstico. Es un método invasivo porque requiere tomar una biopsia de las amígdalas, pero funcional.

Las células tumorales cHL y NLPHL también presentan un patrón de expresión distinto, lo que sugiere un origen distinto del cáncer. Ambos afectan principalmente a los mismos procesos biológicos (metabolismo energético, función inmunológica y ciclo celular) pero de forma distinta. A nivel de metabolismo energético parecen similares, se podría estudiar esto para encontrar dianas terapéuticas.

Bibliografía

Brune, Verena, Enrico Tiacci, Ines Pfeil, Claudia Döring, Susan Eckerle, Carel J.M. Van Noesel, Wolfram Klapper, et al. 2008. "Origin and pathogenesis of nodular lymphocyte-predominant Hodgkin lymphoma as revealed by global gene expression analysis." *Journal of Experimental Medicine* 205 (10): 2251–68. https://doi.org/10.1084/jem.20080809.

Giefing, Maciej, Supandi Winoto-Morbach, Justyna Sosna, Claudia Döring, Wolfram Klapper, Ralf Küppers, Sebastian Böttcher, Dieter Adam, Reiner Siebert, and Stefan Schütze. 2013. "Hodgkin-Reed-Sternberg cells in classical Hodgkin lymphoma show alterations of genes encoding the NADPH oxidase complex and impaired reactive oxygen species synthesis capacity." *PLoS ONE* 8 (12): 1–9. https://doi.org/10.1371/journal.pone.0084928.

Weniger, Marc A., Enrico Tiacci, Stefanie Schneider, Judith Arnolds, Sabrina Rüschenbaum, Janine Duppach, Marc Seifert, Claudia Döring, Martin Leo Hansmann, and Ralf Küppers. 2018. "Human CD30+ B cells represent a unique subset related to Hodgkin lymphoma cells." *Journal of Clinical Investigation* 128 (7): 2996–3007. https://doi.org/10.1172/JCI95993.