"CONACYT, desarrollando cultura de ciencia, tecnología, innovación y calidad"

FINANCIAMIENTO DE BECAS DE INVESTIGACIÓN
(PRIMERA CONVOCATORIA)

Aplicación de Machine Learning en secuencias genómicas de monos rhesus infectados con SIV para la evaluación de la respuesta molecular a vacuna terapéutica.

Universidad Comunera

Horacio Sosa – biososa20@gmail.com

RESUMEN

La búsqueda de una vacuna contra el VIH es crucial debido a su impacto global y la gravedad de la enfermedad. Una vacuna efectiva sería un logro significativo para prevenir la infección y detener la propagación del virus. Un estudio reciente utilizó una vacuna terapéutica en monos Rhesus y reveló que aproximadamente la mitad de los monos quedaron protegidos, mientras que la otra mitad aún mostraba carga viral detectable. Este estudio reanalizó los datos utilizando análisis de expresión diferencial y aprendizaje automático. El análisis de expresión diferencial identificó 16 genes con expresión diferencial en los monos protegidos. Por otro lado, las técnicas de Machine Learning no pudieron predecir con precisión los genes asociados con la protección observada en el 50% de los individuos. A pesar de estos resultados, este estudio sienta las bases para futuras investigaciones que podrían ayudar a identificar los genes responsables y acercarnos cada vez más a una vacuna efectiva para poner fin a la enfermedad del VIH.

OBJETIVOS

1. Comprender y entender el conjunto de datos y el contexto en el que fueron seleccionados y sus características.
2. Verificar la calidad de los datos para garantizar la fiabilidad de resultados obtenidos
3. Identificar genes diferencialmente expresados en el conjunto de datos proveniente de monos rhesus que recibieron la vacuna contra el SIV.
4. Identificar funciones asociadas a genes diferencialmente expresados mediante anotaciones funcionales usando bases de datos de GSEA
5. Evaluar modelos de Machine Learning para su aplicación e identificación de patrones

APORTES DE LA INVESTIGACIÓN

Los datos recopilados en este estudio fueron analizados utilizando técnicas de secuenciación de ARN (RNA-seq) y aprendizaje automático (Machine Learning). En la semana 18 (W18), se identificaron un total de 16 genes con expresión diferencial en los monos protegidos. De estos genes, 8 mostraron un nivel de sobreexpresión en los individuos protegidos, mientras que otros 8 genes presentaron una disminución en su expresión en los individuos protegidos. A pesar de utilizar técnicas de aprendizaje automático, no se logró predecir con precisión los genes asociados con la protección contra el SIV. Este resultado resalta la complejidad de la respuesta inmunológica y la necesidad de seguir investigando para comprender mejor los mecanismos subyacentes de la protección contra el virus. Es importante destacar que aunque estos resultados no fueron completamente positivos, este estudio sienta las bases para futuras investigaciones que podrían identificar los genes responsables de la protección contra el VIH. Cada avance en este campo nos acerca un paso más a encontrar una vacuna efectiva y poner fin a la enfermedad causada por el VIH.

ACTIVIDADES REALIZADAS

1. Obtención de datos abiertos
2. Limpieza y normalización de datos.
3. Análisis exploratorio de datos
4. Análisis de expresión diferencial
5. Aplicación de métodos de machine Learning para feature selection y clasificación.

RESULTADOS OBTENIDOS

Como se observa en la figura 1, los gráficos de PCA no son capaces de capturar con su primer y segundo componentes las diferencias que permiten discernir entre individuos protegidos

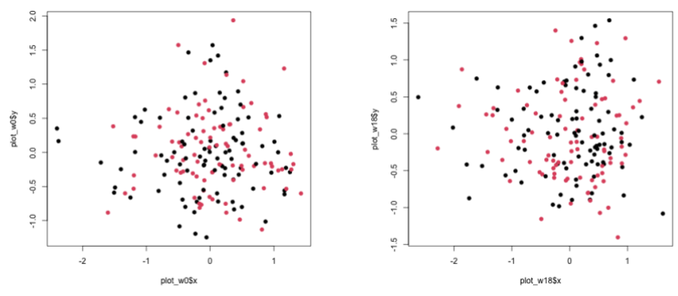

Figura 1. Gráficos de PCA de transcriptomas por semana. Los transcriptomas de monos Rhesus fueron agrupados por semana según el estado conferido por la vacuna, protegido (Rojo) y no protegido (Negro). A) Transcriptomas de monos Rhesus en la semana 0. B) Transcriptomas de monos Rhesus en la semana 18. C) Transcriptomas de monos Rhesus en la semana 88.

Los genes diferencialmente expresados (DEGs) se identificaron usando el modelo binomial negativo (glmfit) de edgeR y una comparación por pareja entre individuos protegidos (Prot) y no protegidos (Unprot) en las semanas 0, 18 y 88 (W0, W18 y W88), usando transcriptomas de individuos no protegidos (Unprot).

En la semana 18 (W18), se identificaron un total de 16 genes diferencialmente expresados, de los cuales 8 están con un nivel de sobreexpresión en los individuos protegidos y 8 genes están con una caída en la expresión en individuos protegidos.

	logFC	logCPM	F	PValue
CDH20	0.511518571428387	0.882787649525447	30.500914954815	1.14501436818639E-07
BPIFC	-0.96297464982607	0.275969352553442	22.3127865157385	4.63104789386006E-06
BICDL2	0.719465724731043	1.10659985008775	15.8733422001812	9.80361298927051E-05
TTN	0.774399328410136	0.618329544095269	15.902552609373	9.66555268834223E-05
ITLN1	-0.709256117697439	3.12426781338036	22.6395109789958	3.98096749767591E-06
EFHC1	-0.870475328837509	1.6118331891915	18.871575225724	2.32508109923697E-05
ODAD3	0.594203342954933	1.37681782856193	15.9507987128629	9.44183619112247E-05
SDK2	1.25733271779727	0.71864018024787	54.3318443919037	5.82639242216465E-12
ROR2	-0.554369711828164	2.90563462069171	15.2747275874056	0.00013120175385883
CD300E	-0.737349655301038	0.376593724657068	22.9255649443278	3.48810781300959E-06
CPE	0.549729474570332	1.50855865915272	17.766009199889	3.93750472191718E-05
ADIRF	-1.06071411384394	1.18346002295392	19.0620293871312	2.12428467955742E-05
TRDV3	0.990465305333166	0.831194628887631	32.0154122867721	5.88785953362672E-08
VSTM1	-0.527065173330209	4.98376649545927	19.5062174691226	1.72161977567765E-05
ND1	-0.930837033760994	0.45263393566475	30.5885009326595	1.10163851521826E-07
ND4	4.66698874035212	4.94431296622028	90.8199291277017	1.05181928172079E-17

Tabla 1. Genes diferencialmente expresados entre individuos protegidos y no protegidos analizados en la semana 18 (W18).

Por último, se procedió a analizar los genes diferencialmente expresados comunes entre los periodos de análisis. Entre los genes diferencialmente expresados comunes entre W0 y W18 se encuentran los siguientes genes:

```
## [1] "BICDL2" "EFHC1"  "ODAD3"  "SDK2"   "CPE"    "TRDV3"  "ND4"
```

En cuanto a los métodos de ML utilizados, ninguna de las estrategias nos permitió encontrar un grupo reducido de genes que nos permita discernir entre los individuos sanos y los infectados. Esto está en línea con lo observado el análisis de PCA, así como en el análisis de expresión diferencial, donde el número de genes diferencialmente expresado es muy bajo, lo que sugiere la poca variación génica entre sujetos protegidos y no protegidos.

VISIÓN Y PLANES FUTUROS

En trabajos futuros, se podrían aplicar nuevas técnicas de Machine Learning que nos permitan discernir entre los dos grupos de estudio y así encontrar un grupo reducido de genes, el cual permitiría el desarrollo de nuevas terapias y/o medicamentos. A pesar de que no fue posible la construcción de modelos de Machine Learning para clasificar individuos protegidos de los no protegidos, los resultados del análisis de expresión diferencial nos brindan nuevos conocimientos y posibles candidatos a ser profundizados para el desarrollo de tratamientos o medicamentos.