Múltiples enfoques para el análisis de datos de *chIP-seq* y evaluación de *peak calling*.

Ciro Ramírez Suástegui, José Damián Martínez Reyes, Marlet Morales Franco

05/03/2016

Introduccion

Los Factores Transcripcionales (TF) son proteínas que se unen a DNA y ayudan a reclutrar a la RNA Polimerasa para transcribir genes. Los TF tienen sitios particulares de pegado y conocerlos se ha vuelto una importante tarea para los biólogos. En la actualidad hay diferentes maneras para elucidar estos sitios, pero uno de los más usados es ChIP-seq. Desafortunadamente, esta no es una tarea trivial, ya que en el proceso deben hacerse muchos análisis, los cuales no siempre dan resultados constantes. La motivación principal de este proyecto es analizar las diferencias que hay entre diferentes tipos de procesamiento de datos de Chip-Seq, con el fin de identificar cómo afectan los parámetros de los *peak-callers* a los resultados.

Los acercamientos que utilizamos son la significancia, la consistencia y el enriquecimiento de motivos. Definimos la significancia como la importancia de recuperar un motivo dado; la consistencia como recuperar un motivo dado a través de diferentes análisis; y el enriquecimiento de motivos como encontrar el motivo de referencia dentro de los resultados obtenidos.

En el presente trabajo se analiza la información obtenida de un experimento de ChiP-seq, con el factor de Transcripcion FNR de *Escherichia Coli*. Más información sobre los datos utilizados puede ser consultada en GSM1010219 y en GSM1010224, donde puede verse la información referente a las muestras experimentales y las muestras de input.

Metodos

Herramientas: RSAT: *peak motifs*. Se utilizó para analizar los picos obtenidos de MACS2 y SWEMBL.

RSAT: *random genome sequence*. Se utilizó para obtener regiones aleatorias del genoma de *E. coli*, para que actuaran como control negativo en las pruebas de significancia.

RSAT: *convert matrix*. Se utilizó para obtener la matriz TRANSFAC del motivo de referencia FNR.

RegulonDB: se utilizó para encontrar la matriz .tab del motivo FNR de *E. coli*.

Resultados

Significancia

Se analizo con RSAT-Peak motifs las secuencias arrojadas durante el análisis de MACS2, tras haber procesados los datos con diferentes q-values. Los valores de q-value fueron 0.1, 0.05, 0.01, 0.001 y 0.0001.

Q.value	Motifs.experiment	Motifs.negative	Highest.sign.experiment	Highest.sign.negative	Ratio.sign
0.1	25	25	54.31	26.78	2.028006
0.05	24	22	53.42	32.19	1.659522
0.01	24	25	36.46	23.67	1.540346
0.001	23	25	29.66	21.35	1.389227
0.0001	23	25	16.19	13.42	1.206408

Los *peaks* se analizaron utilizando la funcion de *Spaced word pairs (dyads)* de peak motifs. Para los datos del experimento, la significancia mas alta era obtenida por el análisis de oligonucleotidos *(position bias)*. Para el control negativo, las sigificancias obtenidas por el analisis de diadas diferían demasiado, por lo que los datos obtenidos por el analisis de diadas se muestran de manera separada.

Q.value	Highest.sign.experiment	Highest.sign.negative
0.1	20.66	258.60
0.05	16.43	206.43
0.01	17.57	179.25
0.001	16.98	153.54
0.0001	18.05	142.65

En la **tabla 1** se indica el número de motivos que obtuvo cada análisis de peak motifs, además de la significancia más alta obtenida en el análisis (para el control negativo, se utiliza la mayor significancia no obtenida por diadas). En la **tabla 2** se muestra la significancia mas alta obtenida por diadas. En los resultados pueden observarse diferentes patrones.

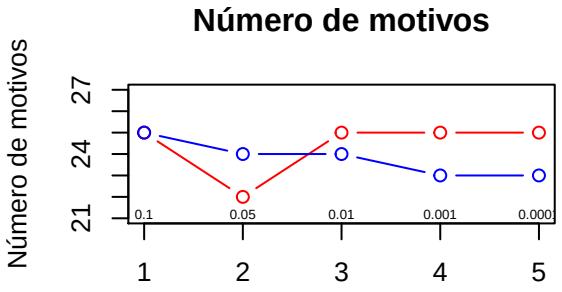


Figura 1

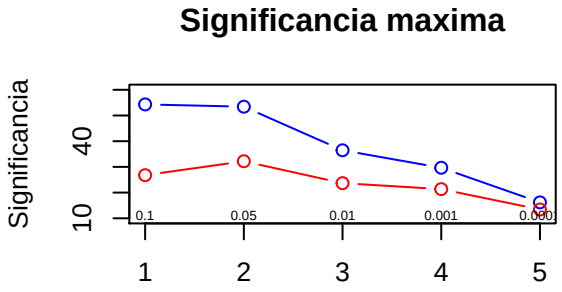


Figura 2

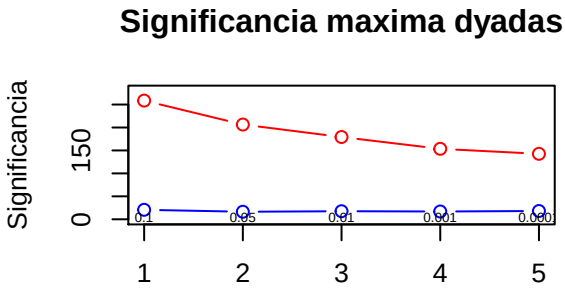


Figura 3

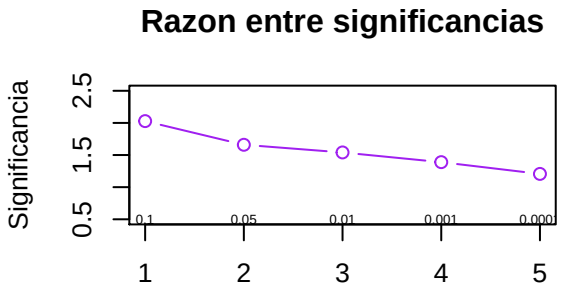


Figura 4

Los numeros por arriba del eje x son los valores de q-value.

Primeramente, el número de motivos encontrados es menor conforme disminuye el *q-value* en los experimentos (**Figura 1**). Esta diferencia no es observable en los controles negativos, pero esto puede deberse a la naturaleza de las secuencias. Podemos ver que, exceptuando el análisis de *q-value* = 0.05, el control negativo obtiene igual o mayor número de motivos que el control positivo. A pesar de obtener mas motivos, la significancia de estos es menor.

Segundo, la significancia va disminuyendo conforme disminuye el q-value en los experimentos, tendencia que puede apreciarse también en el control negativo (**Figura 2 y 3**).

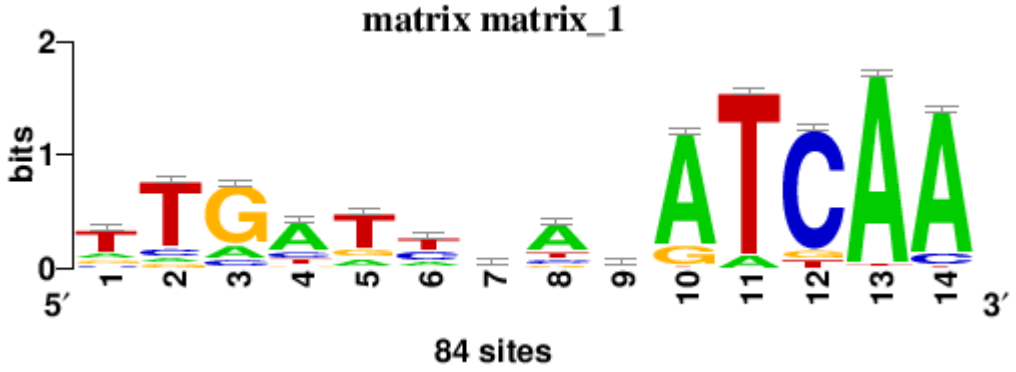
Tercero, en el analisis de diadas, las significancias de los motivos encontrados en el experimento son congruentes con las significancias obtenidas sin diadas. Sin embargo, en el control negativo, las significancias obtienen valores muy altos, lo que hace que esta herramienta no sea confiable al evaluar el control negativo (**Figura 3**). El análisis de diadas es incapaz de encontrar mejores resultados en el experimento y llena de ruido el control negativo.

Cuarto, la razón entre significancias va disminuyendo conforme disminuye el q-value. Con $q\text{-}value = 0.1$, el motivo más significante en el experimento tiene el doble de significancia que el motivo más significante del control negativo, pero esta razón es cercana a 1 cuando $q\text{-}value = 0.0001$ (**Figura 4**).

El q-value y el FDR (*False Discover Rate*) se relacionan con la siguiente ecuación: $FDR = -log_{10}(qvalue)$. A menores q-values, mayor es el valor del FDR, y viceversa. El FDR es una medida de la proporción de descubrimientos falsos entre los rechazos de hipotesis nula totales. Entre mayor el FDR, mayor es el número de hipótesis nulas que no debieron ser rechazadas. Esto se traduce en que el programa es más severo y arroja menor número de resultados, con menor significancia. Este fenómeno puede observarse en nuestros datos.

Enriquecimiento de motivo

El motivo de referencia es el FNR.



Se analizaron0 con RSAT-*Peak motifs* las secuencias arrojadas por el análisis de SWEMBL, tras haber procesados los datos con diferentes valores de gradiente (-R), también llamado valor relativo del background. Los valores R fueron: 0.5, 0.2, 0.1, 0.07, 0.06 y 0.05. La comparación de resultados, ordenados por el gradiente (-R), se encuentra en la siguiente tabla. Se pueden apreciar el valor R, la cadena, las columnas sobrelapantes con el motivo de referencia, el valor de correlación de Pearson, la correlación de Pearson normalizada y la significancia de cada motivo.

R.Motivo	Cadena	Cols	X.Al	r	Nr	Sig
0.5-dyadm1	R	14	0.7778	0.935	0.727	15.16
0.2-6ntmkv3m3	D	8	0.5000	0.846	0.423	0.44
0.2-dyadm1	R	14	0.7000	0.929	0.650	17.59
0.1-6ntmkv3m3	D	8	0.5000	0.844	0.422	1.16
0.1-dyadm1	R	14	0.7000	0.899	0.629	17.14
0.07-dyadm4	D	14	0.7778	0.942	0.732	14.26
0.06-NO	-	0	0.0000	0.000	0.000	0.00
0.05-NO	-	0	0.0000	0.000	0.000	0.00

R.Motivo	Cadena	Cols	X.Al	r	Nr	Sig

Table 3: Tabla 2

En primera instancia podemos observar (**Tabla 2**) que con un parámetro de R inicial de 0.5 se encuentra el motivo de referencia dentro de los resultados y conforme decrece esta aparece un motivo devuelto por el oligo-análisis (calcula las ocurrencias de oligonucleótidos en un set de secuencias y detecta los sobrerepresentados) hasta llegar una R de 0.06 que no devuelve ninguna coincidencia con el motivo de referencia.

Con esto podemos ver que los valores fluctuan poco entre el diferente parametrizado al obtener los picos con SWEMBL:

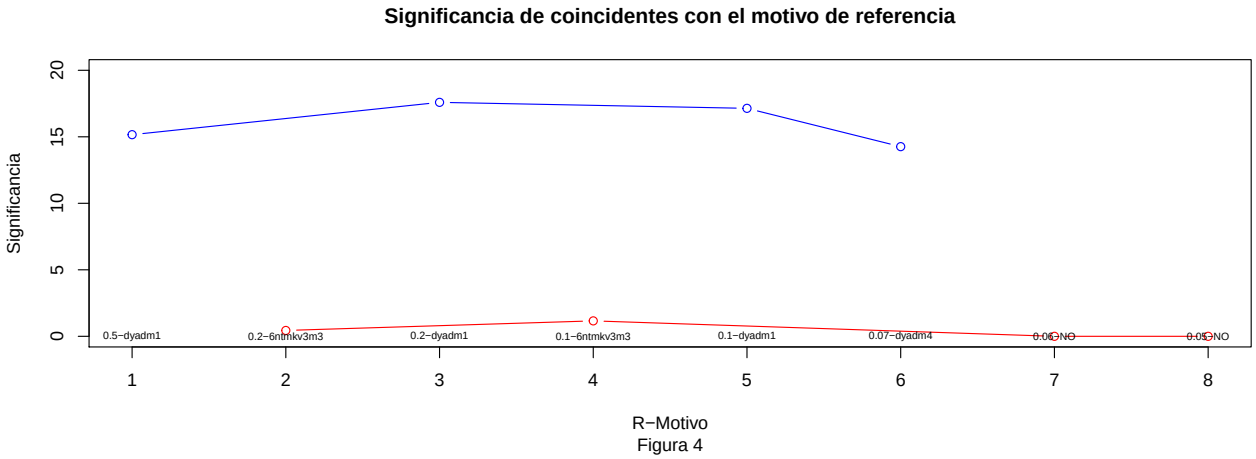


Figura 4

Respecto a la fluctuación de la significancia, como era de esperarse para los motivos coincidentes con el de referencia, se encontraró un mayor valor en los motivos encontrados por análisis de diadas (detecta diadas sobrerrepresentadas en un conjunto de secuencias de ADN; las diadas son un par de oligonucleótidos del mismo tamaño separados por un número fijo de bases). Se seleccionó el análisis de diadas porque el motivo de referencia no podría ser detectado por oligo-análisis debido, probablemente, a su naturaleza dimérica.
Sin embargo se observa (**Figura 4**) que la significancia es muy baja desde un inicio lo que hace que sospechemos de los datos obtenidos con este experimento, ya que estamos utilizando precisamente al motivo de referencia para corroborar, desde la obtención de la muestra hasta el procesamiento bioinformático, que haya salido bien el protocolo de *chIP-seq*.

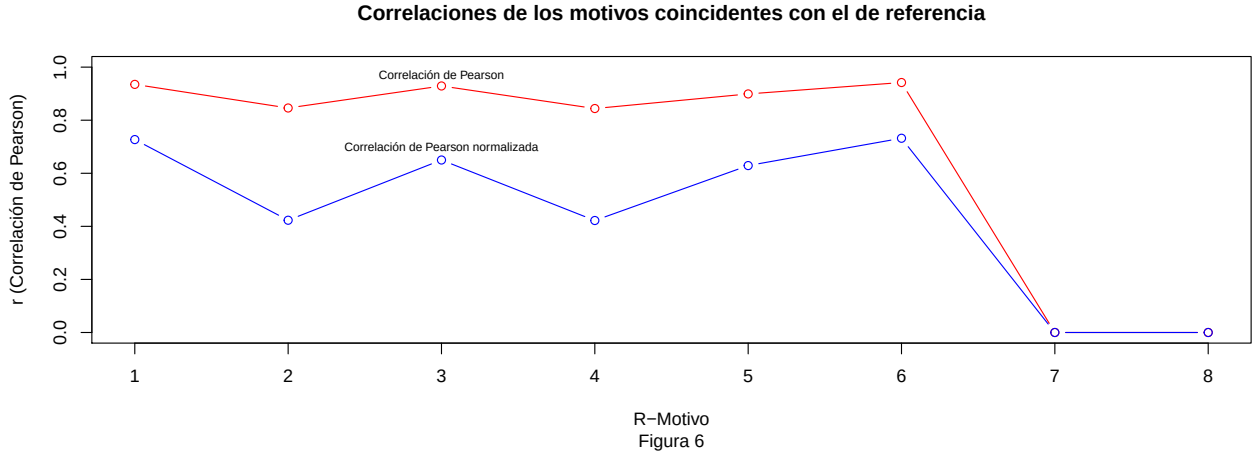


Figura 6

Para complementar con lo ya observado con la significancia tenemos las correlaciones de Pearson (medida de la relación lineal entre dos variables cuantitativas, en este caso son las matrices del metovio de referencia y las de los motivos que encontramos que coniciden con FNR) junto a las normalizadas; puede haber correlaciones altas engañosa obtenida de un alineamiento parcial por lo que se normaliza para evitar este efecto mediante añadir peso a la correlación de acuerdo con la cobertura mutua de los dos motivos comparados. Observamos (**Figura 5**) que las correlaciones son buenas aunque tomando en cuenta las normalizadas tienen un valor pobre para pensar que son lo suficientemente significativas. Esto podría justificarse con las características y el procesamiento del conjunto de datos, donde posiblemente se obtuvo una pobre recuperación en el protocolo de chIP-seq.

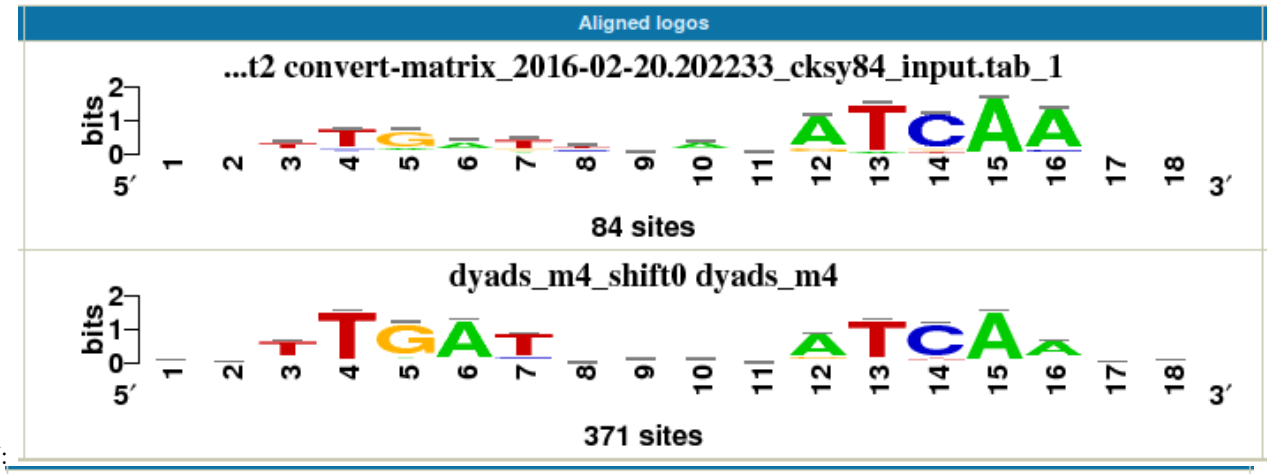
Los mejores valores tanto para r y Nr (es superado en significancia) son los correspondientes a la diada 0.07-dyadm4 seguido de la 0.5-dyadm1 con mayor significancia. Esto tiene sentido porque evidentemente es nuestro motivo de referencia, sin embargo, las correlaciones normalizadas bajan considerablemente. Esto se ve influenciado por es soporte que se le da al motivo por parte de las secuencias y específicamente los picos. Tal vez no se logró recuperar bien el motivo debido a alguna falla en la afinidad del anticuerpo o algún otro paso en el análisis.

Cabe manecionar también que el número de sitios encontrados que muestran los logos de Schneider son muchísimos; 147, 230, 166 y 371, cuando en RegulonDB muestra que solo son 84 en E. coli. Esta discrepancia nos indica de nuevo que algo anda mal con los datos que tenemos de la estracción con FNR. Tengo entendido que este es el número de sitios que esperas encontrar en el genoma de *E. coli*, por lo que con estos valores podría suponer que tiene sentido que haya valore de significancia muy bajos.

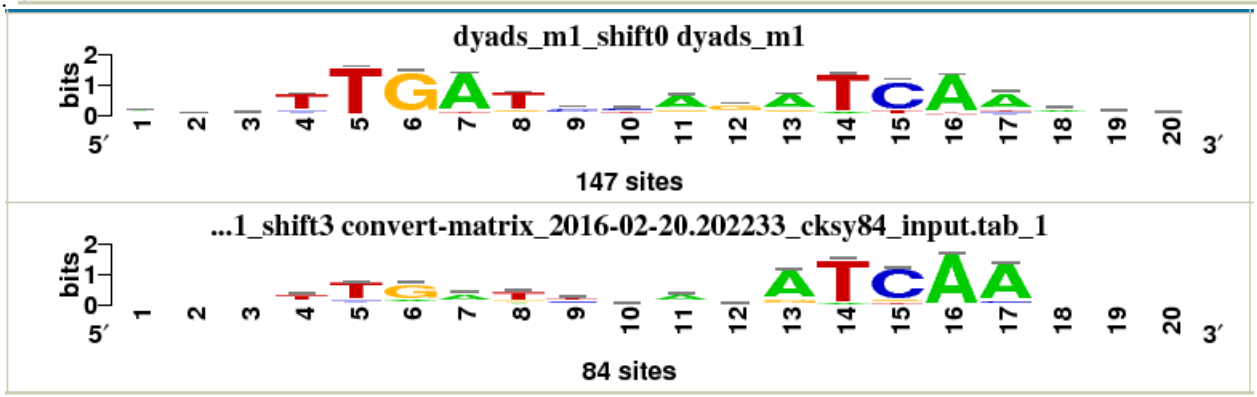
Los análisis con MACS2 son muy parecidos a estos, pero nos pareció mejor no ponerlos aquí debido que no aportaría mucho y, además, se hace un análisis más extendido con los otros dos enfoques; significancia y consistencia.

Alineamiento de logos:

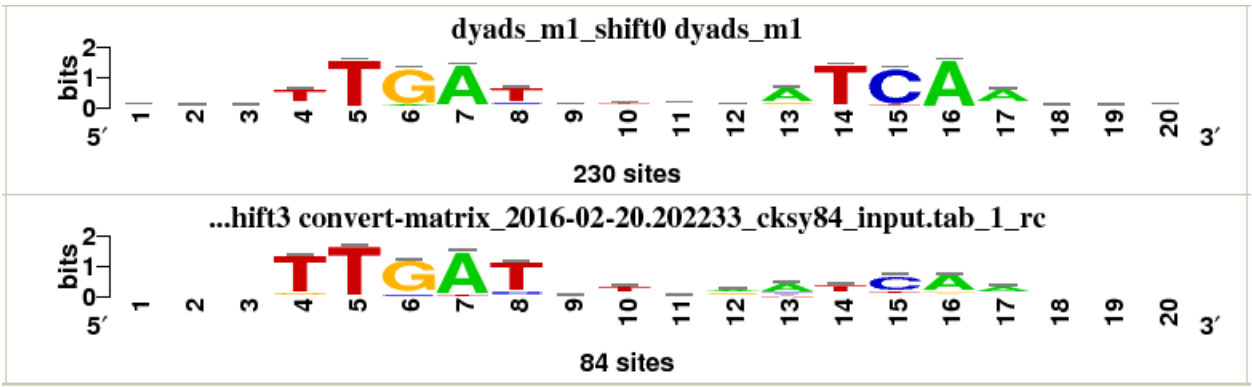
R - 0.07:



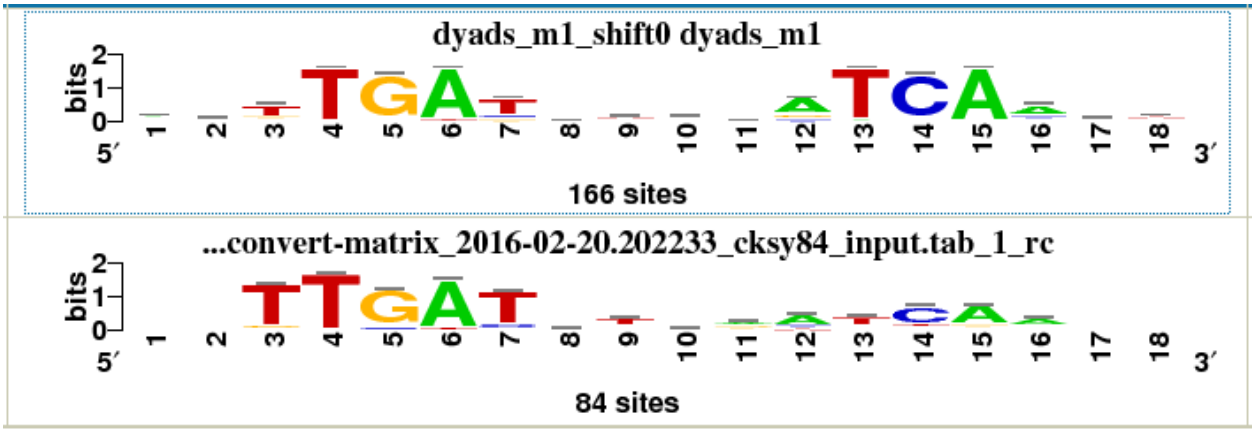
R - 0.1:



R - 0.2:



R - 0.5:

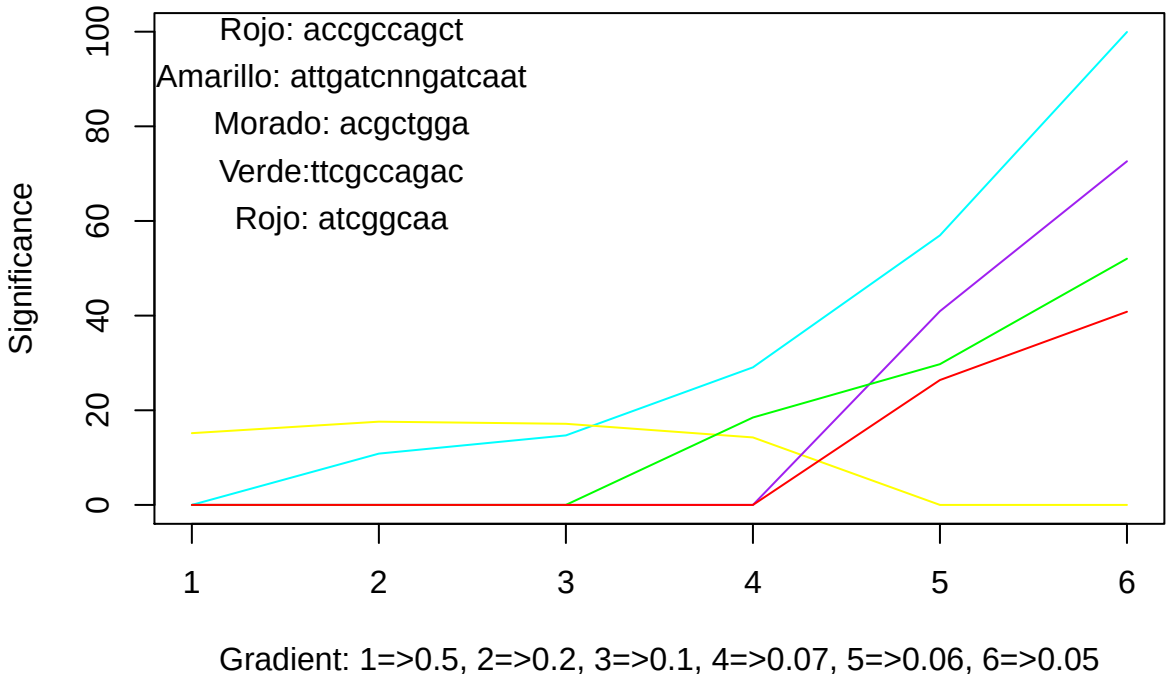


Podemos observar que los sitios, para el motivo encontrado como coincidente en los datos, van de 147 a 371 y que en el motivo de referencia, en regulonDB es de tan solo 84 en *E. coli*.

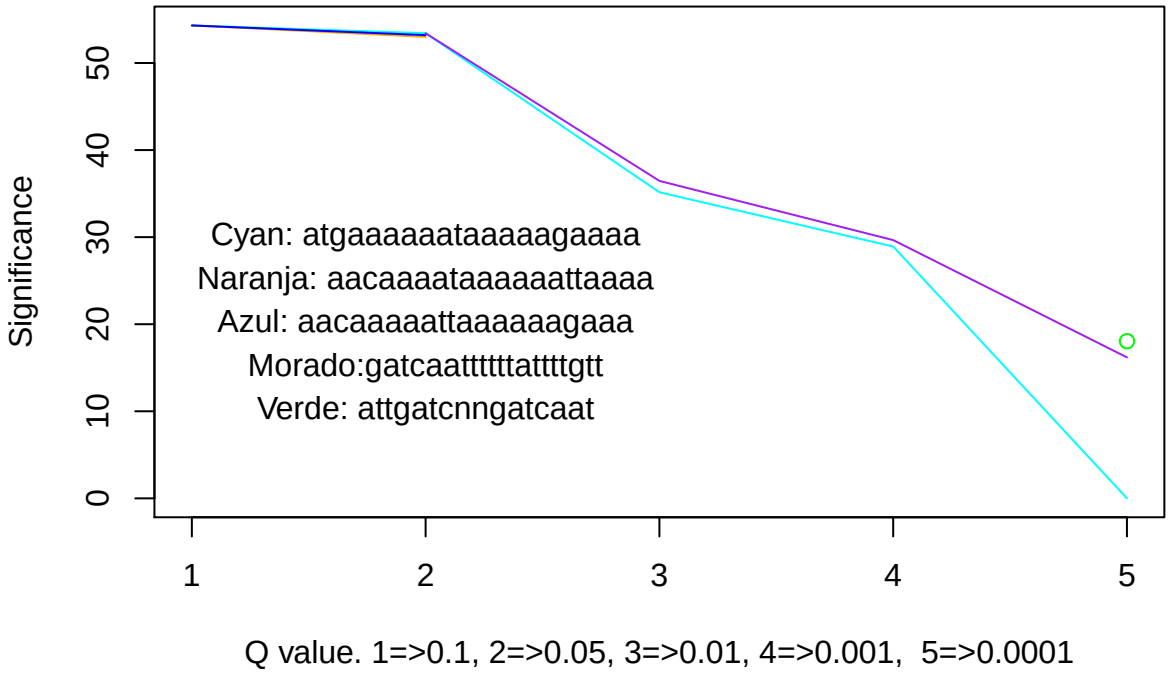
Consistencia

La diversidad en los métodos, y parámetros, para análisis de motivos supone ventajas y desventajas en cuanto a la confiabilidad y eliminación de ruido en los resultados. Es necesario entonces analizar la validez de los resultados en diferentes contextos (parámetros). Escencialmente, el análisis de consistencia trata de este tipo de validación. En las gráficas podemos observar cómo cambia la significancia de algunas secuencias conforme cambiamos los parámetros indicados para cada *peak-caller* (sólo se grafican las secuencias de mayor significancia).

A continucación los resultados arrojados por **SWEMBL**: Para este algoritmo el parámetro que cambia es el de R, que es el valor relativo del background o gradiente. En este podemos ver que a medida que decrece, la significancia aumenta por secuencia.



Estos son los resultados de **MACS2**: El parámetro variable es el valor Q, o *q-value*, que se refiere al *cutoff* o valor de corte para la toma de secuencias como motivos. En sentido estricto se trata del mínimo *false discovery rate* (tasa de descurimento de los falsos positivos). En este caso, cuando el valor de corte se hace más pequeño también lo hace la significancia. Los resultados de las gráficas son consistentes en que los valores más restrictivos del corte se reflejan en su significancia.



Observaciones: Para esta última gráfica se hicieron dos análisis, con o sin diadas. La diferencia en resultados fué únicamente el punto verde, aunque con significancia mayor a los resultados sin diadas. Además, se recuperaron más motivos en el análisis por diadas, pero sus significancias seguían siendo, en general, bajas. Otro punto importante es que las significancias en los *q-value* de 0.1 y 0.05 son las mismas y por motivos de visualización se redondearon unos valores para ver las líneas Naranja y Azul.

Conclusiones

Significancia

Tras comparar los resultados obtenidos, podemos ver que los resultados entre el experimento y el control negativo no difieren considerablemente (**Figura 3**). La **significancia más alta** obtenida es de **54.31**, lo que está *lejos de ser una significancia deseada en un análisis de motivos*. Al ver la similitud de los resultados y los bajos valores de significancia obtenidos en ambas situaciones, podemos intuir que los datos obtenidos del experimento **no son de buena calidad**. Esto podría haber sucedido por alguna complicación entre la fase de *wet-lab* y el análisis bioinformático.

Enriquecimiento de motivo

Después de comparar los resultados obtenidos con el SWEMBL (y algunos con MACS2) podemos decir que sí encontramos el motivo de referencia dentro de nuestro conjunto de datos, por lo que, solo basados en este creiterio, podría validarse el buen seguimiento del protocolo chIP-seq; sin embargo, dado que tenemos valor de significacia y **correlación** poco concluyentes se refuerza la idea (como en el análisis de significacia) de que pudo haber salido algo mal en la fase de laboratorio o en el procesamiento de los datos; apostaría más por la **obtención de datos del *input***, aunque evidentemente debería revisarse cada aspecto. Para ver cómo salió el procesamiento de datos o qué *peak caller* deberíamos usar, tenemos que tomar en cuenta el análisis de consistencia, pero no debemos pensar que este nos va a ayudar a determinar si un posterior análisis debe hacerse con qué parámetros, debido a que es más específico para este cojunto de datos.

Cabe manecionar también que el número de sitios encontrados que muestran los logos de Schneider son muchísimos; 147, 230, 166 y 371, cuando en RegulonDB muestra que solo son 84 en E. coli. Esta discrepancia nos indica de nuevo que algo anda mal con los datos que tenemos de la estracción con FNR.

Consistencia

Dados los resultados obtenidos podemos decir que ambos algoritmos presentan una buena aproximación al obtener motivos y que, en cada uno, se recuperan las mismas secuencias (o motivos muy similares) con significancias diferentes. En SWEMBL el cambio en el valor relativo del background nos dice que entre más pequeño es el background las secuencias se verán sobrerepresentadas lo que nos daría una alza en la significancia, y esto se puede observar bien. En cuanto a MACS, a medida que el valor de corte se hace más restrictivo la probabilidad de encontrar un motivo baja y por tanto su significancia. Estos resultados son consistentes con lo que uno esperaría observar en los análisis.

Al final, lo que nos debe interesar es que hay motivos que se recupera en cada paso del cambio de parámetro (o en la mayoría), y que las excepciones (punto verde en MACS2) pueden deberse a otros factores dependientes del set de datos.

Se sugeriría que el ChIP-seq con el factor de transcripcion, así como el *input*, se repitieran para obtener datos de mejor calidad. En general se plantea que deba llevarse a cabo la repetición de todos los procedimientos así como la utilización de material de mejor calidad o más reciente en la fase de laboratorio.

Material Suplementario

Data sources

- Datos experimentales:

  - Escherichia coli str. K-12 substr. MG1655star.
  - Anticuerpo: Policlonal contra FNR purificado por afinidad.
  - Cepa: Wild Type K-12.
  - Condiciones de crecimiento: Culturo Anaeróbico.
  - GEO accession: GSM1010219.

- Datos de Input:
  - Escherichia coli str. K-12 substr. MG1655star.
  - Anticuerpo: INPUT.
  - Cepa: Wild Type K-12.
  - Condiciones de crecimiento: Culturo Anaeróbico.
  - GEO accession: GSM1010224.

Se utilizaron los archivos .fasta ubicados en Tepeu, que contenían las secuencias de los picos arrojados por MACS2 y SWEMBL.

- Se sugiere ver el resultado desglosado de los análisis:
  - Resultados MACS2.
  - Resultados SWEMBL.

Recursos Bioinformaticos

Herramientas y bases de datos utilizadas para este analisis.

Acronym	Description	URL
RSAT	Regulatory Sequence Analysis Tools	http://rsat.eu/
RegulonDB	Electronically-encoded Regulatory Network (for reference motif)	http://regulondb.ccg.unam.mx/
R	Free software environment for statistical computing and graphics	https://www.r-project.org/

Lista completa de comandos y parametros

Aquí mostramos la lista completa de las **herramientas y comandos** usados durante el análisis. Esto es para hacer posible la reproduccion los resultados.

Convert Matrix: Reference motif Nombre del motivo de referencia: FNR.

Parámetros

- Background model estimation method
  - Estimate from input matrix

- Comando del servidor:

```
convert-matrix  -from tab -to transfac -i $RSAT/public_html/tmp/apache/2016/02/20/convert-matrix_2016-0
```

Peak Motif Los RSAT-peak motifs fueron corridos con parámetros por default, exceptuando:

- Cut peak sequences: +/- *empty* bp on each side of peak centers.
- Discoverover-represented spaced word pairs [dyad analysis] fue seleccionado para todos los análisis.
- RegulonDB prokaryotes fue seleccionada como base de datos para comparar motivos, y JASPAR fue deseleccionada.