Material Suplementario: Corroboración de picos y descubrimiento de motivos en *P. aeruginosa*

Anastasia Hernández Koutoucheva y Alán Fernando Muñoz González

El trabajo se basa en los datos de Chip-seq obtenidos en el artículo ChIP-seq reveals the global regulator AlgR mediating cyclic di-GMP synthesis in Pseudomonas aeruginosa.

### Detección de motivos

Se obtienen las secuencias en formato .txt de GEO Accesion GSE65356 (<http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE65356>), se convierten al formato .bed requerido para el análisis, mediante:

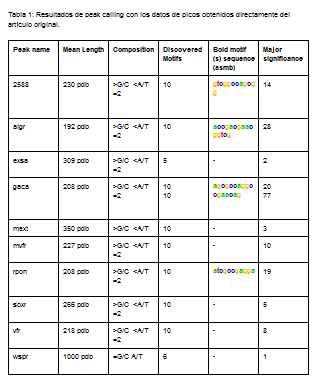
En bash: more secuencia\_DirectaDeGeo.txt | grep -v "#" | cut -f 2,3,10 > seqid.bed

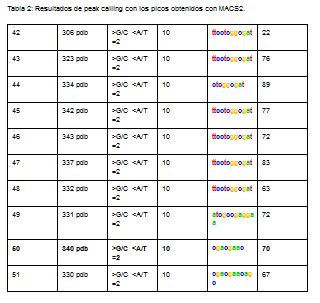
Usando R:  
 table <- read.table("../Downloads/seqid.bed") Chrname <- rep("Chromosome",length(table[,1])) ntable <- data.frame(Chrname,table) write.table(ntable,"../Downloads/seqid\_clean.bed", sep = "", row.names = F, col.names =F, quote = F)

En RSat, se cambia a la base de datos de procariotes; en sequence tools -> sequences from bed/gff/vcf y se elige el organismo Pseudomonas aeruginosa pao1.ASM676v1.30; la opción Mask repeats se deja en su forma predeterminada. Tras obtener los archivos en formato .fasta, se procede a usar la herramienta Peak motifs; con los parámetros default excepto en Compare discovered motifs with databases, en la que únicamente se elige a los procariotes en RegulonDB.

### Análisis de Peak Calling

Las secuencias tienen alrededor de 230 pdb lo cual es un buen indicio para los picos de ChIP-seq ya que son números menores a los que se obtendrían con picos de histonas. La composición en las secuencias es mayor para nucleótidos G/C y menor para la de A/T; pero en cada una de las parejas se observa el mismo patrón (Figura 1). En la composición de dinucleótidos se observa el mismo patrón de aumento en G/C sobre A/T. Para la secuencia de control, se observan los mismos patrones de composición de nucleótidos.

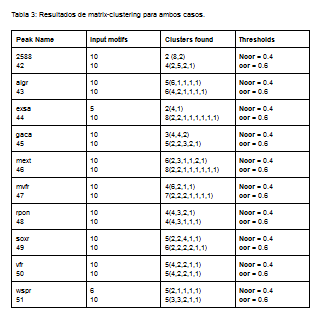




Como nota, la secuencia usada para el análisis con los datos procesados manualmente es la número 50; debido a que toda la secuencia es completamente igual a la obtenida en el artículo. Cabe resaltar la similitud con otras como la 51, en la que obtenemos la secuencia misma y además, más información. Después las matrices con los resultados generados son guardadas en un formato .transfac

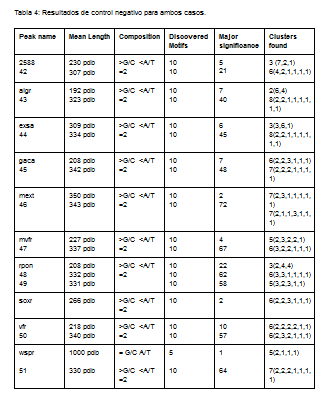
## Uso de herramientas de clustering para corroborar los resultados

Para realizar este paso, se cambia al servidor Fungi para aumentar la velocidad del proceso. Matrix tools -> Matrix-clustering, todos los parámetros se quedan en su forma predeterminada, con excepción de Motif comparison with compare-matrices-quick (100 times faster). Only for Ncor and Cor, el cual se activa.

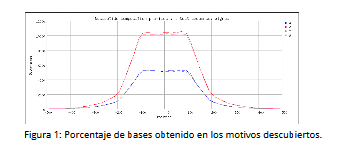


### Control Negativo

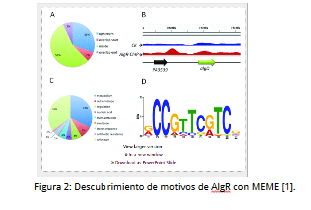
Regresas al servidor RSat de Procariotes. Se usa NGS-ChIP-seq-> random genome fragments para buscar falsos positivos en los resultados. Para esto, se utiliza el mismo organismo que en el análisis de peak calling (Pseudomonas aeruginosa pao1.ASM676v1.30) y dejas todas las opciones de forma predeterminada, con excepción de Mask repeats. Después, se usa la opción Peak-motifs en RSat de Fungi; con los mismos parámetros que en el análisis de Peak calling.

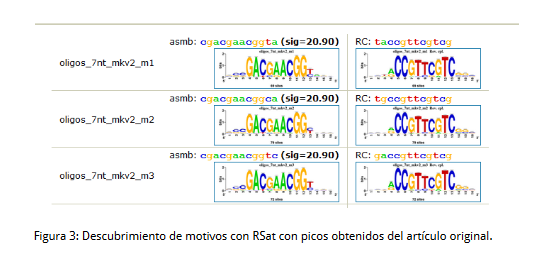


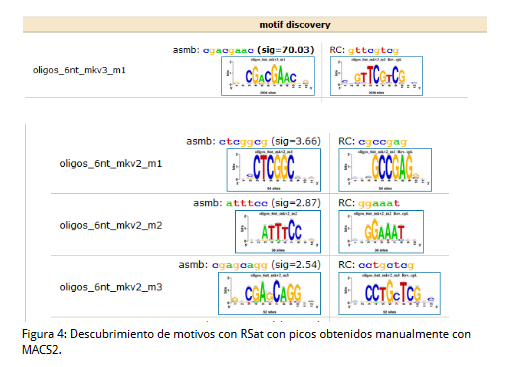
Como nota, todos los análisis de peak-motifs se realizaron con oligos, esto fue porque se intentaron usar otros enfoques (el análisis de diadas no fue usado en este punto) como el de tomar únicamente el parámetro de posición, el cual (generalmente) regresó menos motivos (con menor significancia) con las secuencias prueba. Y, en el control negativo disminuyeron aún más los motivos encontrados (pero, la significancia se mantuvo). Por ejemplo, para el caso de la secuencia 2588 la secuencia regresó 4 motivos con significancia menor a 1 y el control negativo en este experimento retornó 1 motivo con significancia de 1. Además, únicamente con este enfoque, podría usarse la secuencia completa del organismo como control.



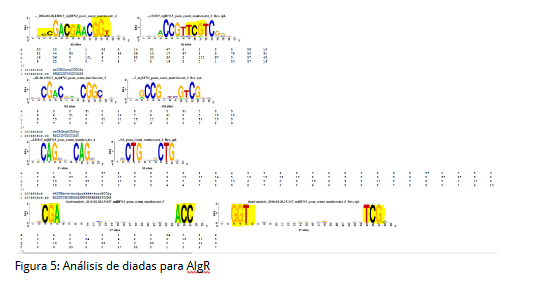
### Motivos de Algr







### Análisis de diadas

El análisis de diadas se realizó con la herramienta dyad-analysis de RSat, en el servidor de procariotes. Para esto, se usó el archivo en formato fasta del pico de AlgR y P. aeruginosa como organismo para modelo de Background, el resto de los parámetros se quedaron en su forma predeterminada. 

### Workflow para nuestros peaks

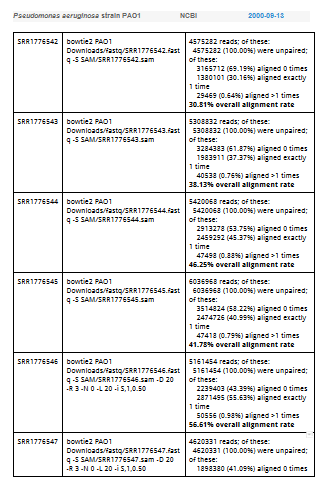
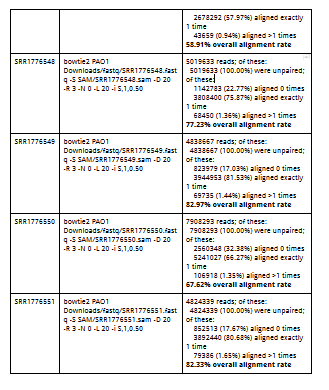
Descargar Secuencias de lecturas cortas <http://www.ebi.ac.uk/ena/data/view/SRP052880>

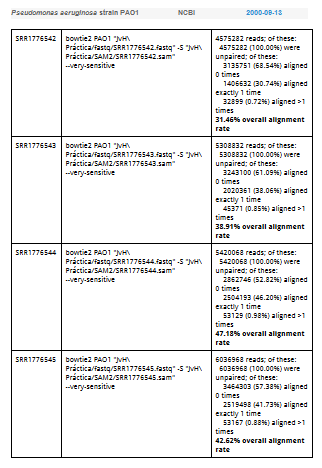
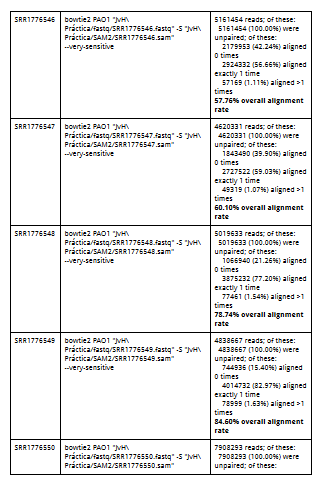
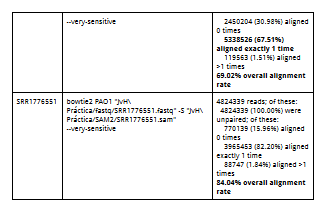
Mapear con genoma de referencia en bowtie2 Index de genoma de Pseudomonas aeruginosa a bowtie2 <http://support.illumina.com/sequencing/sequencing_software/igenome.html> bowtie2 build [Genome\_fasta] PAO1

Mapear Lecturas de Illumina con genoma de referencia bowtie2 PAO1 seq\_ name.fastq -S seq\_name.sam -N 1 bowtie2 PAO1 seq\_ name.fastq -S seq\_name.sam --very-sensitive

El cambio daba ~1% de más secuencias, por lo que decidimos usar en RSAT los mapeos de la línea inicial

Pseudomonas aeruginosa strain PAO1 NCBI 2000-09-13

### MACS2 Peak Call

macs2 callpeak -t SAM/$ bname.sam -f SAM -n peaks/$bname -g 6.3e6 -q 0.01 --nomodel --shiftsize 100

### Comandos en el servidor:

Peak Calling: $RSAT/perl-scripts/peak-motifs -v 1 -title '2688vsmacs2' -i $RSAT/public\_html/tmp/apache/2016/02/20/peak-motifs.2016-02-20.195401\_2016-02-20.195401\_PP1XpT/peak-motifspeak\_seq -max\_seq\_len 1000 -markov auto -disco oligos,positions -nmotifs 5 -minol 6 -maxol 7 -no\_merge\_lengths -2str -origin center -motif\_db regulonDB tf $RSAT/public\_html/motif\_databases/REGULONDB/regulonDB\_2015-08-07.tf -scan\_markov 1 -source getfasta -task purge,seqlen,composition,disco,merge\_motifs,split\_motifs,motifs\_vs\_motifs,timelog,archive,synthesis,small\_summary,motifs\_vs\_db,scan -prefix peak-motifs -noov -img\_format png -outdir $RSAT/public\_html/tmp/apache/2016/02/20/peak-motifs.2016-02-20.195401\_2016-02-20.195401\_PP1XpT)

Matrix-clustering: $RSAT/perl-scripts/matrix-clustering -v 1 -max\_matrices 300 -matrix\_format transfac -i $RSAT/public\_html/tmp/www-data/2016/02/22/matrix-clustering\_2016-02-22.014846\_eMBEdG/matrix-clustering\_query\_matrices.transfac -hclust\_method average -title 'testfun2588' -metric\_build\_tree 'Ncor' -lth w 5 -lth cor 0.6 -lth Ncor 0.4 -quick -label\_in\_tree name -return json,heatmap -o $RSAT/public\_html/tmp/www-data/2016/02/22/matrix-clustering\_2016-02-22.014846\_eMBEdG/matrix-clustering 2> $RSAT/public\_html/tmp/www-data/2016/02/22/matrix-clustering\_2016-02-22.014846\_eMBEdG/matrix-clustering\_err.txt

Control negativo: $RSAT/perl-scripts/random-genome-fragments -template\_format len -i $RSAT/public\_html/tmp/apache/2016/02/21/random-genome-fragments\_2016-02-21.202329\_U9mlWz.lengths -org Pseudomonas\_aeruginosa\_pao1.ASM676v1.30 -return seq -rm -o $RSAT/public\_html/tmp/apache/2016/02/21/random-genome-fragments\_2016-02-21.202329\_U9mlWz\_fragments.fasta 2> $RSAT/public\_html/tmp/apache/2016/02/21/random-genome-fragments\_2016-02-21.202329\_U9mlWz\_error\_log.txt

$RSAT/perl-scripts/peak-motifs -v 1 -title '2588\_negativectrol' -i $RSAT/public\_html/tmp/www-data/2016/02/22/peak-motifs.2016-02-22.032359\_2016-02-22.032359\_flPJWK/peak-motifspeak\_seq -ctrl $RSAT/public\_html/tmp/www-data/2016/02/22/peak-motifs.2016-02-22.032359\_2016-02-22.032359\_flPJWK/peak-motifscontrol\_seq -max\_seq\_len 1000 -markov auto -disco oligos,positions -nmotifs 5 -minol 6 -maxol 7 -no\_merge\_lengths -2str -origin center -motif\_db regulonDB tf $RSAT/public\_html/motif\_databases/REGULONDB/regulonDB\_2015-08-07.tf -scan\_markov 1 -task purge,seqlen,composition,disco,merge\_motifs,split\_motifs,motifs\_vs\_motifs,timelog,archive,synthesis,small\_summary,motifs\_vs\_db,scan -prefix peak-motifs -noov -img\_format png -outdir $RSAT/public\_html/tmp/www-data/2016/02/22/peak-motifs.2016-02-22.032359\_2016-02-22.032359\_flPJWK

$RSAT/perl-scripts/retrieve-seq -org Pseudomonas\_aeruginosa\_PAO1\_uid57945 -feattype gene -type upstream -format fasta -label name -from -1000 -to 200 -noorf -rm -all (?)

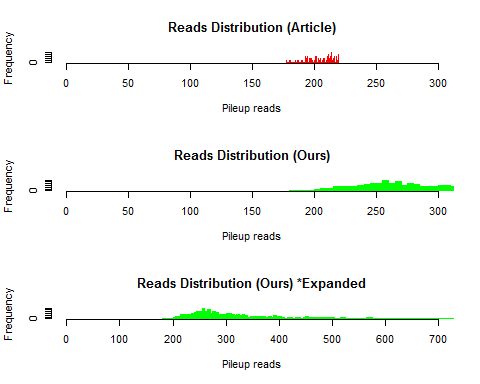
## Peak Calling Statistics

Llamamos datos:

chip\_art <- read.table("GSE65356\_algrmacs2e15\_peaks.txt", header = T)  
chip\_ours <-read.table("SRR1776550\_peaks.xls", header=T)

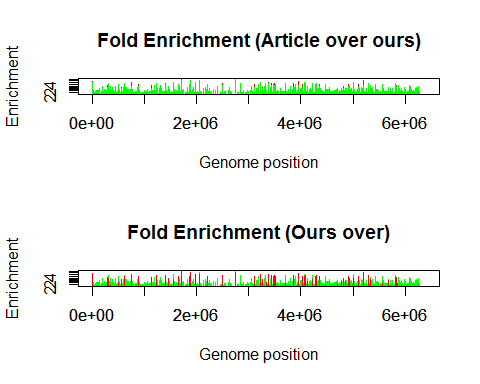
Ploteamos la distribución de reads de acuerdo al pileup:

par(mfrow=c(3,1))  
hist(chip\_art$pileup,  
 main="Reads Distribution (Article)",  
 ylab="Frequency",  
 xlab="Pileup reads",  
 xlim = c(0,300),  
 col="red",  
 border = "red",  
 breaks = 500)  
  
hist(chip\_ours$pileup,  
 main="Reads Distribution (Ours)",  
 ylab="Frequency",  
 xlab="Pileup reads",  
 xlim = c(0,300),  
 col="green",  
 border = "green",  
 breaks = 500)  
  
hist(chip\_ours$pileup,  
 main="Reads Distribution (Ours) \*Expanded",  
 ylab="Frequency",  
 xlab="Pileup reads",  
 xlim = c(0,700),  
 col="green",  
 border = "green",  
 breaks = 500)



Ploteamos el enriquecimiento de picos uno sobre otro y viceversa, para ver dónde coinciden y con cuánta precisión.

par(mfrow=c(2,1))  
plot(chip\_art $abs\_summit,chip\_art$fold\_enrichment,  
 main="Fold Enrichment (Article over ours)",  
 ylab="Enrichment",  
 xlab="Genome position",  
 xlim = c(0,6400000),  
 col="red", type = "h")  
  
par(new=T)  
  
plot(chip\_ours$abs\_summit,  
 chip\_ours$fold\_enrichment,  
 xlim = c(0,6400000),  
 ylab = "",  
 xlab = "",  
 col="green", type = "h")  
   
  
plot(chip\_ours $abs\_summit,chip\_ours$fold\_enrichment,  
 main="Fold Enrichment (Ours over)",  
 ylab="Enrichment",  
 xlab="Genome position",  
 xlim = c(0,6400000),  
 col="green", type = "h")  
  
par(new=T)  
  
plot(chip\_art$abs\_summit,  
 chip\_art$fold\_enrichment,  
 xlim = c(0,6400000),  
 ylab = "",  
 xlab = "",  
 col="red", type = "h")



summit <- chip\_ artabs\_ summit & end>chip\_ art$abs\_summit)))