Trabajo_final

Juan Esteban Rodriguez

April 13, 2016

Metodología (como fue secuenciado, que organismo es y si fue secuenciado en pares o no)

Hay datos en algunos de los documentos? Se puede averiguar lo del organismo con blasts, pero y los datos de la secuenciacion y los pares??? Repasar practicas

Análisis y filtros de calidad de lecturas

cd /home/usuario/Documents/Biologia_computacional/Proyecto_final/ fastqc -O Quality/ RNA-Seq_Data_PBI/Sample*

Despues de revisar los resultados de fastqc se decidieron filtros para cada set de datos.

java -jar /home/usuario/Downloads/Trimmomatic-0.36/trimmomatic-0.36.jar SE RNA-Seq_Data_PBI/Sample-A_Rep1.R1.fastq.gz Trimmed_data/Trimmed_Sample-A_Rep1.R1.fastq.gz HEADCROP:10 SLIDINGWINDOW:4:15 MINLEN:90 fastqc -O Quality/ Trimmed_data/Trimmed_Sample-A_Rep1.R1.fastq.gz

Se arregla el sesgo de bases por posicion de las lecturas, sin embargo, hay una sobrerrepresentacion de secuencias y mucha duplicacion. Cuando se busca en blast el origen de estas secuencias sale: Synthetic construct external RNA control ERCC-00074 sequence.

Falta revisar si son de una sola cadena para dejar el parametro SE en todos los comandos y ver la plataforma para quitar los adaptadores (que no se si sean la causa de las repeticiones y secuencias sobrerrepresentadas). SUpongo que si los datos estan divididos en R1 y R2 entonces solo son de una cadena

2do: java -jar /home/usuario/Downloads/Trimmomatic-0.36/trimmomatic-0.36.jar SE RNA-Seq_Data_PBI/Sample-A_Rep1.R2.fastq.gz Trimmed_data/Trimmed_Sample-A_Rep1.R2.fastq.gz HEADCROP:10 SLIDINGWINDOW:4:15 MINLEN:90 fastqc -O Quality/ Trimmed_data/Trimmed_Sample-A_Rep1.R2.fastq.gz

Sale la misma sobrerrepresentacion: Synthetic construct external RNA control ERCC-00074 sequence & Synthetic construct external RNA control ERCC-00096 sequence

3ro: java -jar /home/usuario/Downloads/Trimmomatic-0.36/trimmomatic-0.36.jar SE RNA-Seq_Data_PBI/Sample-A_Rep2.R1.fastq.gz Trimmed_data/Trimmed_Sample-A_Rep2.R1.fastq.gz HEADCROP:10 SLIDINGWINDOW:4:15 MINLEN:90 fastqc -O Quality/ Trimmed_data/Trimmed_Sample-A_Rep2.R1.fastq.gz

La misma sobrerrepresentacion: Synthetic construct external RNA control ERCC-00074 sequence

4to: java -jar /home/usuario/Downloads/Trimmomatic-0.36/trimmomatic-0.36.jar SE RNA-Seq_Data_PBI/Sample-A_Rep2.R2.fastq.gz Trimmed_data/Trimmed_Sample-A_Rep2.R2.fastq.gz HEADCROP:10 SLIDINGWINDOW:4:15 MINLEN:90 fastqc -O Quality/ Trimmed_data/Trimmed_Sample-A_Rep2.R2.fastq.gz

Igual con algunos organismos en los lugares que siguen: Synthetic construct external RNA control ERCC-00074 sequence.

5to: java -jar /home/usuario/Downloads/Trimmomatic-0.36/trimmomatic-0.36.jar SE RNA-Seq_Data_PBI/Sample-A_Rep3.R1.fastq.gz Trimmed_data/Trimmed_Sample-A_Rep3.R1.fastq.gz HEADCROP:10 SLIDINGWINDOW:4:15 MINLEN:90 fastqc -O Quality/ Trimmed_data/Trimmed_Sample-A_Rep3.R1.fastq.gz

Igual: Synthetic construct external RNA control ERCC-00002 sequence

6to: java -jar /home/usuario/Downloads/Trimmomatic-0.36/trimmomatic-0.36.jar SE RNA-Seq_Data_PBI/Sample-A_Rep3.R2.fastq.gz Trimmed_data/Trimmed_Sample-A_Rep3.R2.fastq.gz HEADCROP:10 SLIDINGWINDOW:4:15 MINLEN:90 fastqc -O Quality/ Trimmed_data/Trimmed_Sample-A_Rep3.R2.fastq.gz

++ ME QUEDE AQUI Falta hacer los mismos filtros para los datos B y aclarar las dudas ++++++++++++++++++++++++++++++++

Ensamble de novo

cd Trimmed_data/Trimmed_Sample-A_Rep1.R1.fastq.gz Trimmed_data/Trimmed_Sample-A_Rep2.R1.fastq.gz Trimmed_data/Trimmed_Sample-A_Rep3.R1.fastq.gz Trimmed_data/Trimmed_Sample-B_Rep1.R1.fastq.gz Trimmed_data/Trimmed_Sample-B_Rep2.R1.fastq.gz Trimmed_data/Trimmed_Sample-B_Rep3.R1.fastq.gz

Trinity –seqType fq –SS_lib_type RF
–left Trimmed_data/Trimmed_Sample-A_Rep1.R1.fastq.gz Trimmed_data/Trimmed_Sample-A_Rep2.R1.fastq.gz Trimmed_data/Trimmed_Sample-A_Rep3.R1.fastq.gz Trimmed_data/Trimmed_Sample-B_Rep1.R1.fastq.gz Trimmed_data/Trimmed_Sample-B_Rep2.R1.fastq.gz Trimmed_data/Trimmed_Sample-B_Rep3.R1.fastq.gz
–right Trimmed_data/Trimmed_Sample-A_Rep1.R2.fastq.gz Trimmed_data/Trimmed_Sample-A_Rep2.R2.fastq.gz Trimmed_data/Trimmed_Sample-A_Rep3.R2.fastq.gz Trimmed_data/Trimmed_Sample-B_Rep1.R2.fastq.gz Trimmed_data/Trimmed_Sample-B_Rep2.R2.fastq.gz Trimmed_data/Trimmed_Sample-B_Rep3.R2.fastq.gz
–CPU 2 –max_memory 1G
–trimmomatic
–output Trinity_output

Hay que indicar la direccion con SS_lib_type. . . ..

Alineamiento de lecturas al transcriptoma

. . . .ALGO ASI for i in All_comparison_trinity/*.qtrim.gz ; do zcat $i | head ; done ln -s ./All_comparison_trinity/*.P.qtrim.gz .

Al momento de hacer el indice hay que ver cuantos reads se mapearon. Puede ser que no se haya descargado bien un documento

o que el numero del k-mero no sea adecuado.

gmap_build -d genome -D . -k 13 Sp_genome.fa gmap -n 0 -D . -d genome All_comparison_trinity/Trinity.fasta -f samse > trinity_gmap.sam

En este caso usamos single end (con samse) porque no estamos alineando reads, sino los TRANSCRITOS/CONTIGS

secuencias largas que surgieron del alineamiento de las lecturas

more trinity_gmap.sam

samtools view -Sb trinity_gmap.sam > trinity_gmap.bam samtools sort trinity_gmap.bam trinity_gmap
samtools index trinity_gmap.bam ...

Con el output de trinity... trinity.fasta

y luego...

MAPEAR AL TRANSCRIPTOMA

bowtie2-build All_comparison_trinity/Trinity.fasta Sp_transcript

tophat2 -T Sp_transcript
Sp_log.left.fq.gz.P.qtrim.gz,Sp_hs.left.fq.gz.P.qtrim.gz,Sp_ds.left.fq.gz.P.qtrim.gz,Sp_plat.left.fq.gz.P.qtrim.gz
Sp_log.right.fq.gz.P.qtrim.gz,Sp_hs.right.fq.gz.P.qtrim.gz,Sp_ds.right.fq.gz.P.qtrim.gz,Sp_plat.right.fq.gz.P.qtrim.gz

...

Visualización de transcritos en el genoma (opcional)

Anotaciones via Trinotate de genes sobreexpresados

Análisis de expresión diferencial