LAB NOTEBOOK

ANALYSIS AND OPTIMIZATION OF THE SEQUENCING DATA PROCESSING FOR THE EFFECTIVE LOCALIZATION OF MUTATIONS RESPONSIBLE FOR THE ANTIBIOTIC RESISTANCE PROPERTY IN E.COLI ON THE EXAMPLE OF AMPICILLIN

SIDORENKO OKSANA
ILYA OLKHOVSKY

Getting the raw data

Terminal command	Result file	Comment
wget https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/005/845/GCF_000005845.2_ASM584v2/GCF_000005845.2_ASM584v2_genomic.fna.gz	ecoli_parental_ref.fna.gz.fna.gz	Reference sequence of the parental (unevolved, not resistant to antibiotics) E. coli strain: .fasta
wget https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/005/845/GCF_000005845.2_ASM584v2/GCF_000005845.2_ASM584v2_genomic.gff.gz	ecoli_parental_ref_annotation.gff.gz	Annotation : .gff
wget https://figshare.com/ndownloader/files/23769689	amp_res_1.fastq.gz	Illumina sequencing reads from shotgun sequencing of an E. coli strain that is resistant to the antibiotic ampicillin: forward
wget https://figshare.com/ndownlo	amp_res_2.fastq.gz	Illumina sequencing reads from shotgun sequencing

ader/files/23769692		of an E. coli strain that is resistant to the antibiotic ampicillin: reverse
gzip -dk amp_res_1.fastq.gz	amp_res_1.fastq	Decompress and keep amp_res_1.fastq.gz
gzip -dk amp_res_2.fastq.gz	amp_res_2.fastq	Decompress and keep amp_res_2.fastq.gz

Inspecting raw sequencing data manually

zcat ecoli_parental_ref.fna.gz | head -20

zcat ecoli_parental_ref_annotation.gff.gz | head -20

zcat amp_res_1.fastq.gz | head -20

zcat amp_res_2.fastq.gz | head -20

Terminal command	Result	Comment
wc -l amp_res_1.fastq	1823504	There are 1823504 lines in amp_res_1.fastq = 455876 reads
wc -l amp_res_2.fastq	1823504	There are 1823504 lines in amp_res_2.fastq = 455876 reads

Inspecting raw sequencing data with fastqc. Filtering the reads

Terminal command	Result	Comment
sudo apt-get install fastqc		Installing fastqc

sudo fastqc -o . /home/oxana/Project1/raw_data/amp_res_1.fastq /home/oxana/Project1/raw_data/amp_res_2.fastq	amp_res_1_fastqc.html amp_res_1_fastqc.zip amp_res_2_fastqc.html amp_res_2_fastqc.zip	Running fastqc on the two fastq files: amp_res_1.fastq and amp_res_2.fastq
conda install -c bioconda trimmomatic		Installing Trimmomatic
java -jar /home/oxana/miniconda3/pkgs/trimmomatic-0.39-hdfd78af_2/share/trimmomatic-0.39-2/trimmomatic.jar PE -phred33 amp_res_1.fastq.gz amp_res_2.fastq.gz output_forward_paired.fq.gz output_forward_unpaired.fq.gz output_reverse_paired.fq.gz output_reverse_unpaired.fq.gz ILLUMINACLIP:/home/oxana/miniconda3/pkgs/trimmomatic-0.39-hdfd78af_2/share/trimmomatic-0.39-2/adapters/TruSeq3-PE.fa:2:30:10 LEADING:20 TRAILING:20 SLIDINGWINDOW:10:20 MINLEN:20	Using PrefixPair: 'TACACTCTTTCCCTACACGACGCTCTTCCGATCT' and 'GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT' ILLUMINACLIP: Using 1 prefix pairs, 0 forward/reverse sequences, 0 forward only sequences, 0 reverse only sequences Input Read Pairs: 455876 Both Surviving: 445524 (97,73%) Forward Only Surviving: 9951 (2,18%) Reverse Only Surviving: 271 (0,06%) Dropped: 130 (0,03%) ● output_forward_paired.fq.gz	Running Trimmomatic in paired end mode, with following parameters: ● Cut bases off the start of a read if quality below 20 ● Cut bases off the end of a read if quality below 20 ● Trim reads using a sliding window approach, with window size 10 and average quality within the window 20. ● Drop the read if it is below length 20.

	● output_forward_unpaired.fq.gz ● output_reverse_paired.fq.gz ● output_reverse_unpaired.fq.gz	
zcat output_forward_paired.fq.gz | wc -l	1782096	Checking the count of the trimmed paired reads (forward) manually: 1782096/4 = 445524
zcat output_reverse_paired.fq.gz | wc -l	1782096	Checking the count of the trimmed paired reads (reverse) manually: 1782096/4 = 445524
sudo fastqc -o . /home/oxana/Project1/BI_Project_1/raw_data/output_forward_paired.fq.gz /home/oxana/Project1/BI_Project_1/raw_data/output_reverse_paired.fq.gz	output_forward_paired_fastqc.html output_forward_paired_fastqc.zip output_reverse_paired_fastqc.html output_reverse_paired_fastqc.zip	Running fastqc on the two fastq files: output_forward_paired.fq and output_reverse_paired.fq
What happens if we increase the quality score at all steps to 30?		
/home/oxana/miniconda3/pkgs/trimmomatic-0.39-hdfd78af_2/share/trimmomatic-0.39-2/trimmomatic.jar PE -phred33 amp_res_1.fastq.gz amp_res_2.fastq.gz		Running Trimmomatic in paired end mode, with following parameters: ● Cut bases off the start of a read if quality below 30

test30_forward_paired.fq.gz test30_forward_unpaired.fq.gz test30_reverse_paired.fq.gz test30_reverse_unpaired.fq.gz ILLUMINACLIP:/home/oxana/miniconda3/pkgs/trimmomatic-0.39-hdfd78af_2/share/trimmomatic-0.39-2/adapters/TruSeq3-PE.fa:2:30:10 LEADING:30 TRAILING:30 SLIDINGWINDOW:10:30 MINLEN:30		● Cut bases off the end of a read if quality below 30 ● Trim reads using a sliding window approach, with window size 10 and average quality within the window 30. ● Drop the read if it is below length 30.
zcat test30_forward_paired.fq.gz | wc -l	1439764	Checking the count of the trimmed paired reads (forward) manually: 1782096/4 = 359941
zcat test30_reverse_paired.fq.gz | wc -l	1439764	Checking the count of the trimmed paired reads (reverse) manually: 1782096/4 = 359941
sudo fastqc -o . /home/oxana/Project1/BI_Project_1/raw_data/test30_forward_paired.fq.gz /home/oxana/Project1/BI_Project_1/raw_data/test30_reverse_paired.fq.gz	test30_forward_paired_fastqc.html test30_forward_paired_fastqc.zip test30_reverse_paired_fastqc.html test30_reverse_paired_fastqc.zip	Running fastqc on the two fastq files: test30_forward_paired.fq and test30_reverse_paired.fq

Aligning sequences to reference

5.1 Indexing the reference file

Terminal command	Result	Comment
apt-get install bwa	-	bwa installation
bwa index ecoli_parental_ref.fna.gz &	ecoli_parental_ref.fna.gz.amb ecoli_parental_ref.fna.gz.ann ecoli_parental_ref.fna.gz.bwt ecoli_parental_ref.fna.gz.pac ecoli_parental_ref.fna.gz.sa	Indexing the reference file (background)

5.2 Aligning reads

Terminal command	Result	Comment
bwa mem -t2 ecoli_parental_ref.fna.gz output_forward_paired.fq.gz output_reverse_paired.fq.gz > alignment.sam	alignment.sam	Aligning trimmed, paired sequences to the reference genome
samtools view alignment.sam | head -20	-	Checking the format manually

5.3. Compressing SAM file

Terminal command	Result	Comment

samtools view -S -b alignment.sam > alignment.bam	alignment.bam	Converting a sam file to a bam file
samtools view alignment.bam | head -20	-	Checking the format manually
samtools flagstat alignment.bam	891306 + 0 in total (QC-passed reads + QC-failed reads) 891048 + 0 primary 0 + 0 secondary 258 + 0 supplementary 0 + 0 duplicates 0 + 0 primary duplicates 890190 + 0 mapped (99.87% : N/A) 889932 + 0 primary mapped (99.87% : N/A) 891048 + 0 paired in sequencing 445524 + 0 read1 445524 + 0 read2 887122 + 0 properly paired (99.56% : N/A) 888962 + 0 with itself and mate mapped 970 + 0 singletons (0.11% : N/A) 0 + 0 with mate mapped to a different chr 0 + 0 with mate mapped to a different chr (mapQ>=5)	Getting some basic statistics: we have 890190 (99.87%) mapped reads.

5.4 Sort and index BAM file

Terminal command	Result	Comment
samtools sort alignment.bam -o alignment_sorted.bam	alignment_sorted.bam	Sorting bam file by sequence coordinate on reference
samtools index alignment_sorted.bam	alignment_sorted.bam.bai	Indexing bam file for faster search
sudo ./igv.sh	Beautiful pictures :)	Visualization with IGV browser with ecoli_parental_ref.fasta and alignment_sorted.bam (we need alignment_sorted.bam.bai as well).

6. Variant calling

Terminal command	Result	Comment
samtools mpileup -f ecoli_parental_ref.fasta alignment_sorted.bam > my.mpileup	my.mpileup	Making an mpileup intermediate file
varscan mpileup2snp my.mpileup --min-var-freq 0.2 --variants --output-vcf 1 > VarScan_results_snp.vcf	Only SNPs will be reported Warning: No p-value threshold provided, so	Running VarScan to reporn SNPs with option --min-var-frequency 0.2

	p-values will not be calculated Min coverage: 8 Min reads2: 2 Min var freq: 0.2 Min avg qual: 15 P-value thresh: 0.01 Reading input from my.mpileup 4641514 bases in pileup file 9 variant positions (6 SNP, 3 indel) 0 were failed by the strand-filter 6 variant positions reported (6 SNP, 0 indel) VarScan_results_snp.vcf	(20%). This sets the minimum % of non-reference bases at a position required to call it a mutation in the sample. The --variants flag tells VarScan to only output positions that are above our threshold. The --output-vcf 1 option tells we want the output in yet another kind of data format called vcf (variant call format).
varscan mpileup2indel my.mpileup --min-var-freq 0.2 --variants --output-vcf 1 > VarScan_results_indel.vcf	Only indels will be reported Warning: No p-value threshold provided, so p-values will not be calculated Min coverage: 8 Min reads2: 2 Min var freq: 0.2 Min avg qual: 15 P-value thresh: 0.01 Reading input from my.mpileup 4641514 bases in pileup file	Running VarScan to detect indels with option --min-var-frequency 0.2 (20%). This sets the minimum % of non-reference bases at a position required to call it a mutation in the sample. The --variants flag tells VarScan to only output

	9 variant positions (6 SNP, 3 indel) 0 were failed by the strand-filter 3 variant positions reported (0 SNP, 3 indel) VarScan_results_indel.vcf	positions that are above our threshold. The --output-vcf 1 option tells we want the output in yet another kind of data format called vcf (variant call format).

7. Variant effect prediction

Terminal command	Result	Comment
sudo ./igv.sh	Beautiful pictures :)	Visualization with IGV browser with ecoli_parental_ref.fasta, alignment_sorted.bam, ecoli_parental_ref_annotation.gff.gz, VarScan_results_snp.vcf

Exploring mutations, find whether each mutation occurs in a gene, whether it is missense (changes the amino acid sequence), nonsense (introduces a frameshift or early stop codon), or synonymous (no amino acid change). For missense or nonsense mutations finding out what that gene name is.

SNPs

Reference	Mutation	Type of mutation	Gene
GCC (A)	GGC (G)	missence	ftsI
CAG (Q)	CTG (L)	missence	acrB

TTT	TCT	not protein coding	rybA
GGT (G)	GAT (D)	missence	mntP
GTA (V)	GGA (G)	missence	envZ
GCC (A)	GCA (A)	synonymous	rsgA

Automatic SNP annotation

Terminal command	Result	Comment
conda install -c bioconda snpeff		snpeff installation
wget https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/005/845/GCF_000005845.2_ASM584v2/GCF_000005845.2_ASM584v2_genomic.gbff.gz	GCF_000005845.2_ASM584v2/GCF_000005845.2_ASM584v2_genomic.gbff.gz	Downloading the file that contains both annotation and sequence
Database creating		
vim snpEff.config	snpEff.config	Creating text file snpeff.config with one string: k12.genome ecoli_K12
mkdir -p data/k12	data/k12	Creating folder for the database

gunzip GCF_000005845.2_ASM584v2_genomic.gbff.gz cp GCF_000005845.2_ASM584v2_genomic.gbff data/k12/genes.gbk	data/k12/genes.gbk	Putting the .gbk file (unzipped and renamed to genes.gbk) into data/k12
sudo java -jar /home/oxana/miniconda3/pkgs/snpeff-5.1-hdfd78af_2/share/snpeff-5.1-2/snpEff.jar build -genbank -v k12	sequence.NC_000913.3.bin snpEffectPredictor.bin	Creating database
sudo java -jar /home/oxana/miniconda3/pkgs/snpeff-5.1-hdfd78af_0/share/snpeff-5.1-0/snpEff.jar ann k12 VarScan_results_snp.vcf > snp_ann.vcf	snpEff_summary.html snp_ann.vcf	Annotation
sudo ./igv.sh	Beautiful pictures :)	Visualization with IGV browser with ecoli_parental_ref.fasta, alignment_sorted.bam, ecoli_parental_ref_annotation.gff.gz, snp_ann.vcf