Salmonella enterica is a bacterium known to cause salmonellosis. In enteric bacteria, DNA supercoiling is responsive to environmental conditions and different antibiotics can be used to relax supercoiling and alter the expression of supercoiling-sensitive genes. However, *S. enterica* shows significant resistance to novobiocin antibiotic and relatively small variability of supercoiling response.

Here we analyzed the novobiocin effect on *S. enterica* gene expression and tried to reveal different mechanisms of *S. enterica* antibiotic resistance and transcription-supercoiling coupling.

This repository contains data and code which were used for analysis. Description of code files and their output files briefly summarized in Table 1.

Methods

Data

For our study, we used RNA-seq data obtained after incubation of *S. enterica* strain 14028S with various novobiocin concentrations (0, 100 and 500 μg/ml) at different time points (10, 20 and 60 min) [1].

Data Preprocessing

Reads were mapped onto the genome sequence of the Salmonella enterica subsp. enterica serovar Typhimurium strain 14028S assembly GCA_000022165.1 [2] using HISAT2 version 2.1.0 [3]. For conversion BAM-formatted files into SAM-formatted files, we used SAMtools [4]. Code for data processing can be found in the "Code_for_alignment.sh" file.

Mapped sequencing reads to genomic features were assigned using featureCounts (R package) [5]. Code for this step can be found in the "Feature_counts.r" file and this file contains commands which should be implemented in R.

Analysis of differentially expressed genes

To perform analysis of differentially expressed genes we used R package DESeq2 [6].

Principal component analysis (PCA) was used to assess the variance between sample groups and sample replicates. Several approaches were implemented to find genes with a significant difference in expression level in different time points or treatment conditions. Formulas which were used as a design for comparison can be found in scripts "FGSEA.r", "Enrichment_analysis_TopGO.r", "Clusterization.r".

Position Related Data Analysis

We used WoPPER web server [7] for position-related data analysis of gene expression in prokaryotes. Tables with log2Foldchange were obtained after DESeq2 analysis and were used as an input.

GO enrichment analysis

We also performed enrichment analysis in GO terms for differentially expressed genes. Since Salmonella enterica and especially this strain is non-model organism there was no available ready-to-use database with a mapping of gene names and GO-terms we created our own custom database. Web-server Quick GO [8] was used to retrieve data about mapping and the resulting database can be found as "Salmonella_enterica_14028S_gene_to_GO" file.

This database can be used for GO-enrichment analysis in topGO package [9] and an example of the code for such analysis can be found in the "Enrichment_analysis_TopGO.r" file.

Clusterization using expression data

In order to identify gene clusters with similar expression dynamics in several time points, we performed k-means clusterization method. This analysis was also implemented in R and required packages and code can be found in the "Clusterization.r" file.

Gene Set Enrichment Analysis (GSEA) and Over-Representation Analysis (ORA)

To perform GSEA analysis R package fgsea was used. This package requires a list of ranked genes and a gmt file with gene sets. In our case, custom gene sets with converging and diverging genes were used. To create these sets and resulting gmt file we transformed GTF file which contains information about gene name, coordinates and strand were used and Python code for creating sets can be found in "Converging_diverging.ipynb" file.

Code for fgsea analysis can be found in the "FGSEA.r" file. It also contains code for visualization.

Over-Representation analysis was used in order to determine if converging and diverging genes is enriched by genes from clusters which were received at the Clusterization step. We used a hypergeometric test and built-in R function phyper for this analysis.

Software requirements

HISAT2 (version 2.1.0), SAMtools (version 1.9), R (version 3.0.1), additional R packages (Rsubread, DESeq2, topGO, fgsea), Python3 (version > 3.6).

Table 1. A brief summary of code files and their output files.

Method	**Code File**	**Output Files**
Data Preprocessing	Code_for_alignment.sh	Set of bam files which can be used for featureCounts
Reads assignment	Feature_counts.r	Tables with raw counts
Differential expression	FGSEA.r, topGO_enrichment.r, Clusterization.r	Visualization of data
Gene Ontology	Web-server Quick GO	Table with gene to GO mapping which can be used as data for creating gene2GO or GO2gene files for topGO
GO-enrichment analysis	Enrichment_analysis_TopGO.r	Enrichment of differentially expressed genes by genes associated with different biological processes in GO-terms and plots for visualization of these data
Clusterization	Clusterization.r	Sets of genes with similar expression

		dynamics and plots for visualization of these data
GSEA	Converging_diverging.ipynb FGSEA.r	Sets of converging and diverging genes and plots for visualization of these data

References:

1. Gogoleva, N. E., Konnova, T. A., Balkin, A. S., Plotnikov, A. O., & Gogolev, Y. V. (2020). Transcriptomic data of Salmonella enterica subsp. enterica serovar Typhimurium str. 14028S treated with novobiocin. Data in Brief, 29, 105297.
2. ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/022/165/GCF_000022165.1_ASM2216v1/GCF_000022165.1_ ASM2216v1_genomic.fna.gz - genome sequence of the Salmonella enterica subsp. enterica serovar Typhimurium strain 14028S assembly GCA_000022165.1
3. Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. Nature Methods, 12(4), 357–360.
4. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., … Homer, N. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics, 25(16), 2078–2079.
5. Liao, Y., Smyth, G. K., & Shi, W. (2013). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics, 30(7), 923–930.
6. Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology, 15(12).
7. Puccio, S., Grillo, G., Licciulli, F., Severgnini, M., Liuni, S., Bicciato, S., Peano, C. (2017). WoPPER: Web server for Position Related data analysis of gene Expression in Prokaryotes. Nucleic Acids Research, 45(W1), W109–W115.
8. https://www.ebi.ac.uk/QuickGO/
9. Alexa A, Rahnenfuhrer J (2020). *topGO: Enrichment Analysis for Gene Ontology*. R package version 2.41.0.