**An Automatic Sequencing Data Processing Pipeline**

Cheng-En Tan1,2, FangZhou Li1,2, Minseung Kim1,2 and Ilias Tagkopoulos1,2,\*

1Department of Computer Science, University of California, Davis 95616, USA

2Genome Center, University of California, Davis 95616, USA

\*Corresponding author (itagkopoulos@ucdavis.edu)

**ABSTRACT**

**INTRODUCTION**

**Description**

This pipeline can prepare the transcriptomic compendium (a normalized, format-consistent data matrix across samples from different studies) by collecting the samples in Sequencing Read Archive (SRA) database given the topic you are interested in and your target species.

There are four steps in this pipeline. First, the pipeline will read the SRA Experiment ID list provided by users and fetch the corresponded metadata in SRA database. In addition, the pipeline will read the gene annotation file in gff3 format and download the sequence of reference genome from NCBI database. Second, the alignment procedure will align the sequencing data with reference genome sequence. Third, the aligned result will be counted into gene expression level. Fourth, the counted gene expression level will be normalized to ensure each gene expression profiles from different lab and experiment platform are comparable.

**Input**

The pipeline takes two inputs: SRA Experiment ID list in raw text format and Gene Annotation Data in gff3 format.

**Procedures**

**Required packages and software**

|  |  |  |
| --- | --- | --- |
| Toolkit | Version | Note |
| sratoolkit | 2.10.8 |  |
| bowtie2 | 2.3.4 |  |
| python | 3.6 |  |

|  |  |  |
| --- | --- | --- |
| Python Packages | Version | Note |
| biopython |  |  |
| pandas |  |  |
| RSeQC |  |  |
| HTSeq |  |  |
| missingpy |  |  |
| matplotlib |  |  |
| sklearn |  |  |

**Metadata preparation**

**Sequencing data download**

**Sequencing data alignment**

**Gene expression counting**

**Data normalization**

**Output**

The pipeline will produce a metadata table and a table with gene expression levels of samples specified in SRA Experiment ID list.

**Validation**

**Unsupervised validation**

**Supervised validation: Correlation among gene expression profiles**

**Supervised validation: Knowledge capture**

**Results**

**ACKNOWLEDGEMENTS**

We appreciate the recommendations and discussions by Ameen Eetemadi, Beatriz Pereira, XiaoKang Wang and Navneet Rai. Computation resources and HPC support was provided by the college of Engineering in UC Davis. This work was supported by NSF awards ….

**METHODS**

**Data curation and data processing pipeline description**

For transcriptomic profiles, data was collected from Gene Expression Omnibus (GEO)17, ArrayExpress18 and Sequence Read Archive (SRA)19 databases. All data and supplementary information was downloaded for each entry. We collected mRNA expression profiles for microarray data and RNA sequencing (RNA-seq) data for raw sequencing data. 202 biological projects which contained 4950 transcriptomic profiles including 4169 microarray profiles and 813 RNA-seq profiles were collected. Among these 202 biological projects, were 139 projects were published.

Once raw data was downloaded, the information of each profile was manually annotated after reading the relevant literature. Then data preprocessing, gene feature extraction, background removal, missing value imputation, normalization and outlier profile removal were applied to the entire dataset (Figure 2). All functions in R packages and toolkits were run with default parameters except specified cases.

**Metadata annotation**

We annotated metadata of transcriptome profiles based on the definitions described in Table S1. For each profile, information about related projects, strains, stresses, and culture conditions were annotated based on the corresponding literature and projects. Project information includes related project or series identification number in GEO, ArrayExpress, or SRA databases, and corresponding principle investigator, publication date, and identification number in the Pubmed database20 which is critical for data integration of multiple series into super series. This information was collected from those three databases automatically via the downloaded metadata of each profile. Other information was collected from literature manually. For genetic backgrounds, information of 205 different strains including parent strains, gene perturbation, plasmid, and antibiotic resistance were annotated. For stress information, data on 47 different stress including concentration and duration of stress were annotated. For culture conditions, data on 38 different types of mediums including the ingredients of media, pH, culture temperature, culture duration and culture type (planktonic or biofilm)were annotated. Additional information such as the type of instrument for profiling, library type of RNA-seq data, and paired information are also annotated for data processing.

**Gene features extraction and mapping**

To extract and map raw features of microarray and RNA-seq data into gene expression levels, feature mapping was applied. ` information which maps names of probes/array positions into genes are downloaded for each platform from the GEO and ArrayExpress databases. For the platforms with probe sequence but without gene mapping information, nucleotide BLAST (Basic Local Alignment Search Tool)21 implement in R package rBLAST (version 0.99.1)22 are applied to align probe sequences with selected representative genome templates which appear frequently in collected dataset (Table 1). For each RNA seq profiles, LT2 strain is selected as the representative strain as template (Genbank ID:NC\_003197.2) for sequence alignment.

**RNA-seq data processing**

813 RNA-seq profiles downloaded from SRA database are read using fastq-dump toolkit (version 2.8.1)19 and aligned with selected reference genomes using bowtie2 alignment toolkit (version 2.3.1)23. 32 profiles which are not from Illumina and cannot be read by bowtie2 are discarded. There are 296 RNA-seq profiles without the annotation of stranded information. To decide whether RNA-seq data is stranded or unstranded in those profiles, ratios of sequences of two different direction in alignment results are evaluated using infer\_experiment.py in RSeQC package (version 2.6.4).24 The profiles with only unidirectional reads with significant read numbers are considered as stranded data otherwise are annotated as unstranded data. Finally, features are extracted using HTSeq toolkits (version 0.6.1)25 and gene expression levels are converted into log scale. Since there are 268 RNA-seq profiles are from samples co-cultured with mammalian tissues and there are many sequence are not from *Salmonella* which may significantly reduce the alignment rate and the ratio of reads that cannot be counted as any gene expression, all data were not filtered according any quality indicators of alignment. To maintain the quality of RNA-seq profile, correlations among replicates from the same conditions are evaluated and hierarchical clustering are applied. This procedure will be performed with microarray profiles at the last step of data processing.

**Microarray data processing**

Microarray data downloaded from GEO database and ArrayExpress database are collected. For each sample recorded in GEO database, Simple Omnibus Format in Text (SOFT)26 files which contain both background and intensities of two channels are read otherwise the raw data stored in supplementary data folder are read. Raw data are read for each sample recorded in ArrayExpress database. Among these 4169 microarray profiles, 3204 microarray profiles contain both background intensities and gene expression levels for background correction. 944 profiles which do not contain background intensities or contain only fold change levels and 21 profiles recorded in raw image format are discarded. To perform background correction, background correction function in R package limma (version 3.28.21)27 with normal and exponential convolution model (normexp) approach is used. Log conversion are applied after background correction. Data from samples are merged into a single matrix which contains background-corrected gene expressions of collected samples then missing value imputation is performed. Samples or genes with missing value ratio higher than 0.3 are discarded before imputation and then Random Forest approach implemented in R package missForest (version 1.4)28 with parallel backend implemented in R package doParallel (version 1.0.11)29 are performed. Among 3204 microarray profiles, 2653 microarray profiles whose missing value ratio lower than 0.3 are remained and 3222 gene expression levels are collected.

**Merging data and preparing the dataset for building gene expression prediction model**

Before data normalization, two data matrices of RNA-seq data with 781 log-transformed gene expression profiles and microarray data with 2653 log-transformed expression profiles are merged into one single matrix. After the merging of RNA-seq data and microarray data, quantile normalization is applied to reduce the platform bias using R package preprocessingCore (version 1.40.0).30 3206 genes which were detected in both RNA-seq and microarray are stored and the merged dataset contains 3434 profiles.

To prepare the dataset for building gene expression prediction model, the profiles with unknown strain, medium, stresses or gene perturbation annotation are discarded and the combination of these four attributes were defined as a condition. Since different profiling time may change the gene expressions significantly and the profiles with profile time annotation are limited, only the profiles extracted from exponential phase culture are used to reduce the variance contributed by different profile time. 827 exponential phase profiles with annotated strain, medium, stresses and gene perturbation information are extracted.

To remove the outlier, complete-link hierarchical clustering using one minus pairwise Pearson correlation coefficient (PCC) as distance on gene expression profiles are applied for each condition. The cutoff threshold is set to 0.2 so that the minimum pairwise PCC will be greater than 0.8 in all cluster. Only the largest cluster will be kept, and the remaining profiles will be treated as outliers and discarded.

***E. coli* dataset**

Ecomics *E. coli* compendium by Kim et. al16 was used for transfer learning. Since it is difficult to obtain *Salmonella* gene copy number information which is necessary for absolute level quantification in Ecomics16, the modified version of Ecomics compendium without absolute value quantification were used to reduce the systematic bias between two compendium contributed by different preprocessing procedure. In addition, only exponential phase profiles with annotated strain, medium, stress and gene perturbation information are used. Outlier removal and average on dataset are similar as the preprocess procedure for preparing *Salmonella* compendium. 493 conditions of 4103 genes from 1809 profiles from Ecomics compendium are extracted.

**Finding homologous genes of *Salmonella* and *E. coli***

To find the homologous genes of 4103 genes in Ecomics compendium and 3206 genes in *Salmonella* compendium, NCBI protein cluster database which contains homologous gene information are used. First, the *E. coli* and *Salmonella* gene locus are mapped into the corresponded protein in NCBI protein database, and then the protein are mapped into the identical protein group. Finally, identical protein group from different species can be mapped into the same protein cluster in NCBI protein cluster database. Genes from two different species are treated as homologous genes if their corresponded protein belongs to the same protein cluster. 2328 homologous genes are found in *Salmonella* gene set and Ecoli gene set (Figure 3).

**Dataset comparison with Ecomics compendium**

2328 Homologous gene expression in *Salmonella* compendium and Ecomics compendium are compared to observe similarity of their gene expression and regulation. Since the culture conditions are different, gene expression of two dataset cannot be compared together. To compare these two dataset, we assume that both *Salmonella* and *E.coli* have some common gene regulatory networks and therefore similarity could be observed in correlation matrix of gene expressions in these two dataset. Gene expression correlation matrix of 2328 homologous genes in both datasets are compared in two ways. First, hierarchical clustering is applied to the correlation matrix in *Salmonella* dataset, and then observe whether clusters existed in Ecomics compendium. Second, each column of correlation matrix represent how this specific gene co-expressed with other 2327 genes, and there should be similarity on co-expression between *Salmonella* and *E.coli*. The co-expression profiles of each genes in *Salmonella* and *E.coli* are compared and then the distribution of correlation of co-expression profiles are observed.

**Gene expression prediction model**

Multi-Omics Model and Analytics (MOMA) platform in transcriptomic layer16 is used for gene expression prediction. For MOMA, model and optimization parameters are specified as follows: For the model parameters, no bias and dropout are used in MOMA and the other model parameters are used as default. For the optimization parameters, the batch size is set as 32 and the patience is set as 32 and the other parameters are used as default.

**Transfer learning on MOMA**

To apply transfer learning on MOMA, the *E.coli* MOMA model using all modified Ecomics compendium are pretrained and the *Salmonella* MOMA model using training subset of *Salmonella* compendium are pretrained. After the training of these two models, the feature and gene expression conversion layer are added to *E.coli* MOMA: The feature conversion layer convert features in *Salmonella* compendium to the features in Ecomics. The initial weights of the feature conversion layer are initialized as follows: for the features annotated in both compendium, the corresponded weights are initialized as ones and all weights related to the features not annotated in both compendium are fixed as zeros during training. Linear activation function is used in feature conversion layer. The gene expression conversion layer convert *E.coli* genes to *Salmonella* genes. The initial weights of the gene conversion layer are initialized as follows: the corresponded weights of homologous genes which are annotated in both compendiums are initialized as ones, and for the weights of genes in *E.coli* only, they will be fully-connected to *Salmonella* genes. Sigmoid activation function is used in gene conversion layer. The overall design of the transfer learning model is shown in Figure 4.

**Observation of gene expression performance improvement by transfer learning**

Two approaches are used to observe the gene expression performance improvement. First, we assume transfer learning approach can generally improved gene expression performance on all conditions. To verify this assumption, cross-validation (CV) with different training set ratio (10%, 20%, 33%, 50% (2-Fold CV) and 66.7% (3-Fold CV)) are applied. Since the transfer learning model involved ensemble model structure, four different control models are used (Figure 5) to verify that *E.coli* dataset contributes performance improvement instead of ensemble model structure. Second, we assume transfer learning approach can improve gene expression performance related to specific stress responses which is annotated in both datasets. We select anaerobic stress responses to verify this assumption. In addition to four control models, one additional control that *E.coli* MOMA model pretrained with wild type conditions (without stress and gene perturbation conditions) are added to verify that anaerobic conditions in *E.coli* compendium contributes the gene expression performance improvement.

**Comparison of weights between *Salmonella* MOMA model and *E.coli* MOMA in transfer learning model**

To analyze how *E.coli* MOMA in transfer learning improve the prediction performance, the kernel weights corresponded to specific stresses in *Salmonella* MOMA and *E.coli* MOMA in transfer learning model are compared. Since feature conversion layer and gene conversion layer are involved in transfer learning model, it is necessary to convert the weights of *E.coli* MOMA in transfer learning MOMA before compared with the weights of *Salmonella* MOMA. The derivation of weight conversion is as follows: First, we have the gene expression prediction formula by *E.coli* MOMA with two conversion layers. We assume the kernel layer in MOMA contributes much more than the recurrent layer and bias therefore we have the following approximation:

|  |  |  |
| --- | --- | --- |
|  |  | (1) |

Where is the gene prediction from *E.coli* MOMA with conversion layers, is the features represent 407 different culture conditions in *Salmonella* compendium, is the weights of feature conversion layer, is the weights of kernel layer of *E.coli* MOMA and is the weights of gene conversion layer. And then we consider how *Salmonella* MOMA predict the gene expressions with approximation:

|  |  |  |
| --- | --- | --- |
|  |  | (2) |

Where is the gene prediction from *Salmonella* MOMA and is the weights of kernel layer of *Salmonella* MOMA. The average of gene prediction from *E.coli* MOMA with conversion layers and *Salmonella* MOMA will be the final gene prediction results of transfer learning model. To make can be compared with ,we rewrite the formula (1):

|  |  |
| --- | --- |
|  | (3) |

Where is the weight which merges the weights of kernel layer of *E.coli* MOMA and two conversion layers. This weight can be compared with . To obtain , we replace the terms in (1) and (3) and solve the equation:

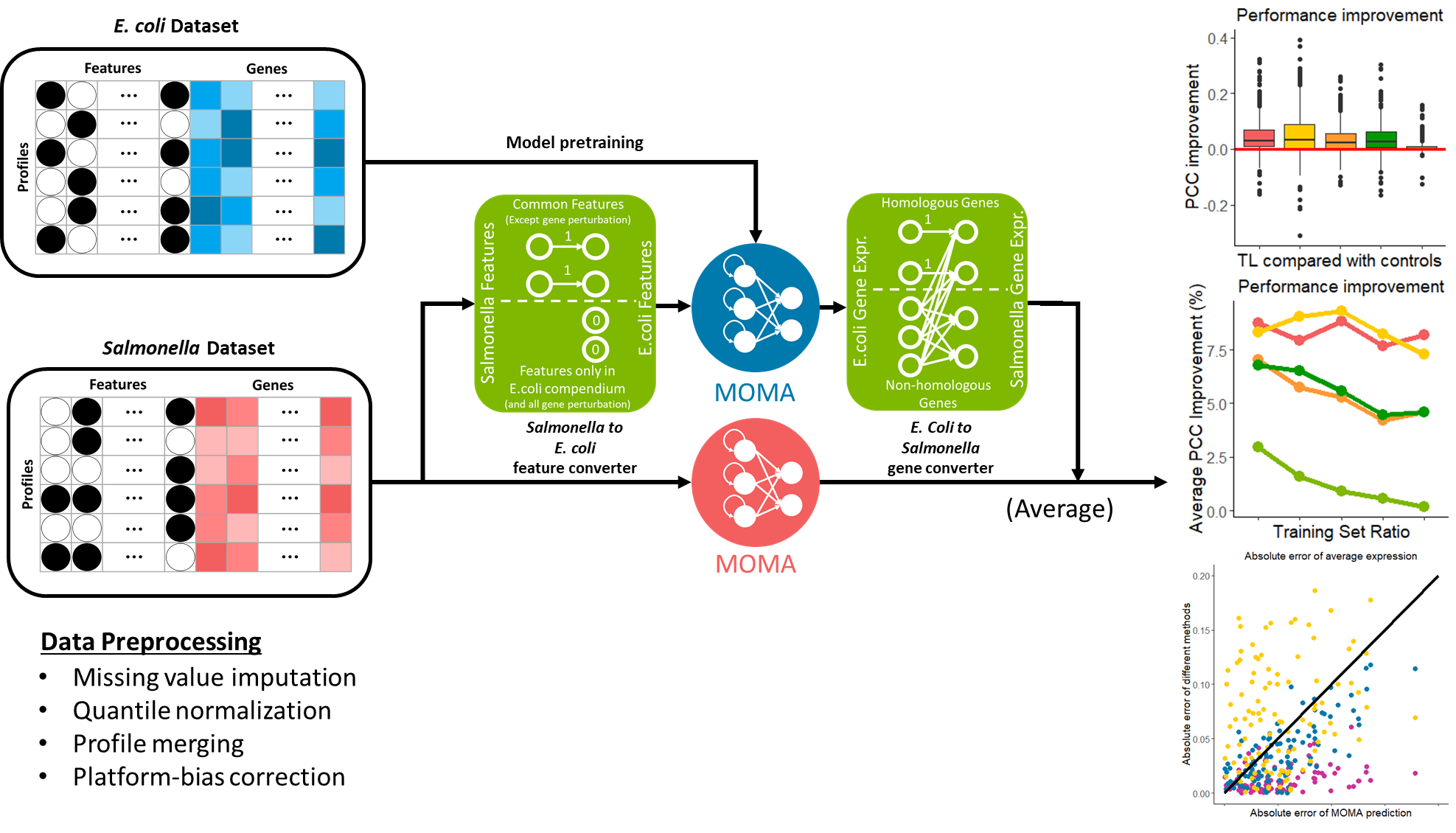
|  |  |
| --- | --- |
|  | (4) |

Where is the pseudo inverse of . All 407 conditions with 347 features in *Salmonella* compendium are used for weights conversion and comparison.

**Supplementary Information**

Table S1. Descriptions of metadata annotation. It defines features to annotate for profile, stress, medium, and strain.

Table S2. Metadata of compiled transcriptome profiles for *Salmonella*.

Figure 1. Overview of transfer learning approach. *E.coli* and *Salmonella* models are pretrained then parts of weights in *E.coli* model are transferred to *Salmonella* models.

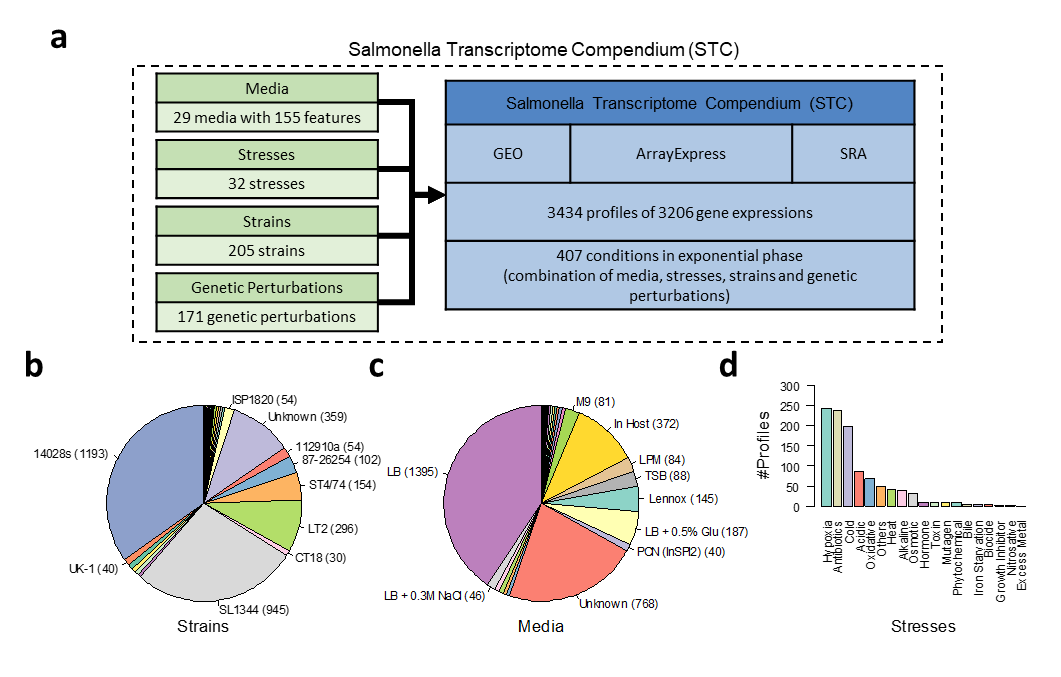


Figure 2. Analysis of the compendium of *Salmonella*. (a) Distribution of 26 media and the top 10 media (including unknown medium) are shown. Medium information of 768 profiles is missing and 372 profiles are cultured in host. (b) Distribution of to 10 strains (including unknown strain). (c) Hierarchical clustering results based on gene expressions of 407 conditions. Four different color bars represent different strains, media, stresses and gene perturbations. (d) To quantify the importance of these four features, the entropy of specific features for the groups from the results of hierarchical clustering with complete link under different cut threshold are evaluated. High entropy means conditions different strain, medium, stress or gene perturbation cannot be separate well by clustering and the highest entropy could be obtained if the conditions are randomly permuted. This plot shows the difference between the evaluated entropy and the entropy from randomly permuted conditions of four features.

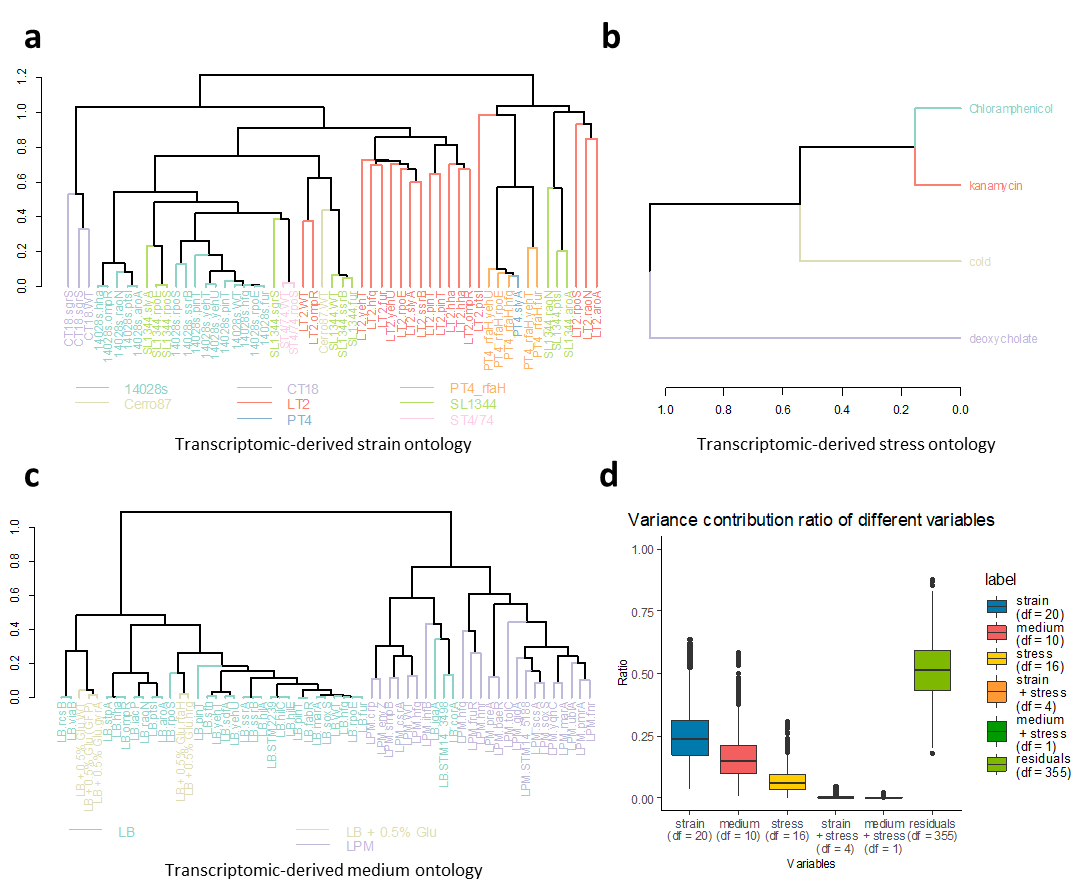


Figure 3. Heatmap of gene expression correlation in (a) *Salmonella* compendium (b) Ecomics. (c) Hierarchical clustering results of Salmonella. The color bar also shows the hierarchical clustering results of E.coli (only the clusters which can be mapped to clusters in Salmonella are colored)

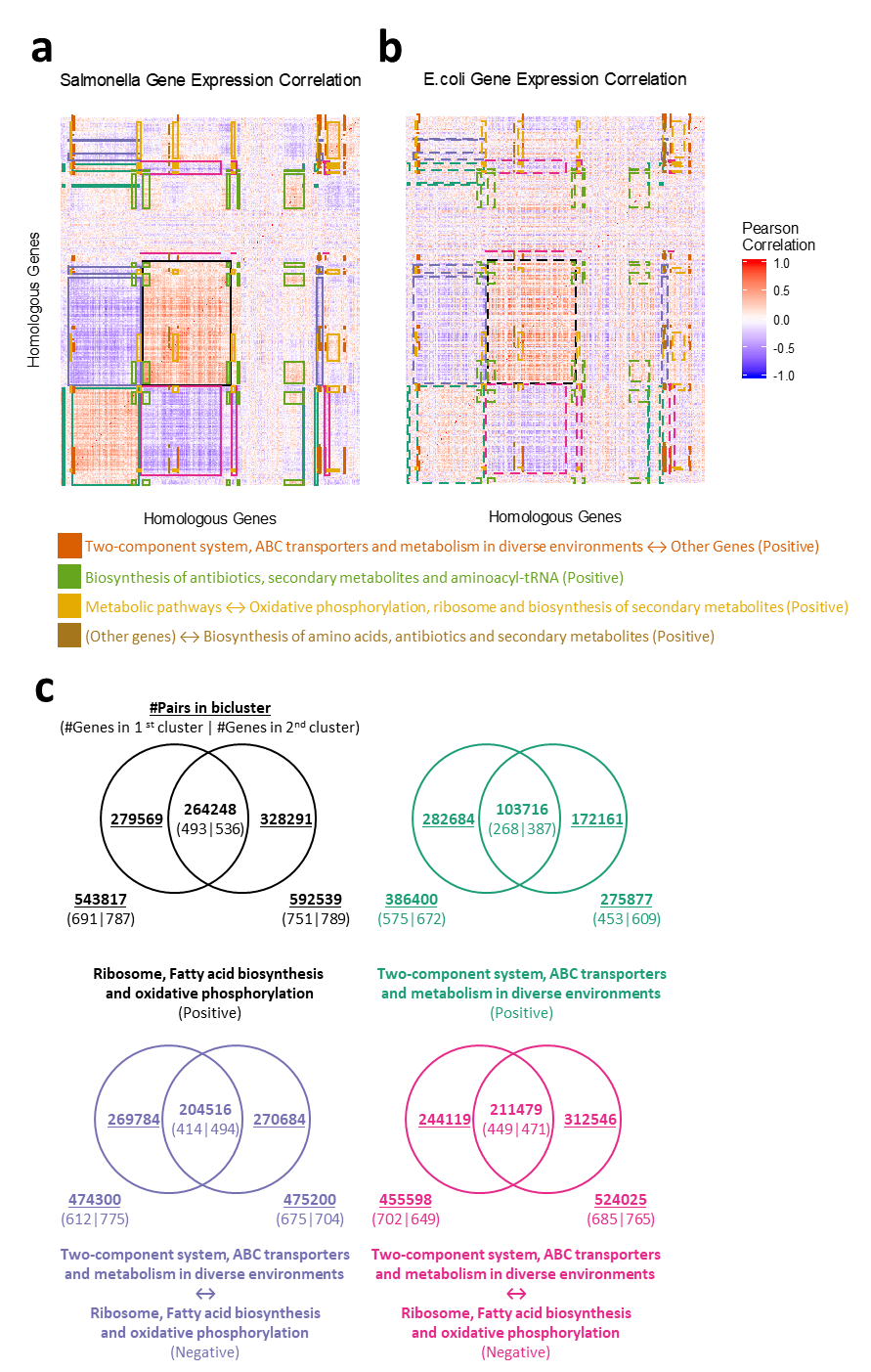


Figure 4. Correlation of 2328 homologous gene expression correlation between E.coli and Salmonella. (a) The distribution of correlation of gene expression correlation compared with control by randomly permutated gene expression correlation matrix. (b) Distribution of correlation of correlation of gene expression between E.coli and Salmonella of clusters found in Salmonella.

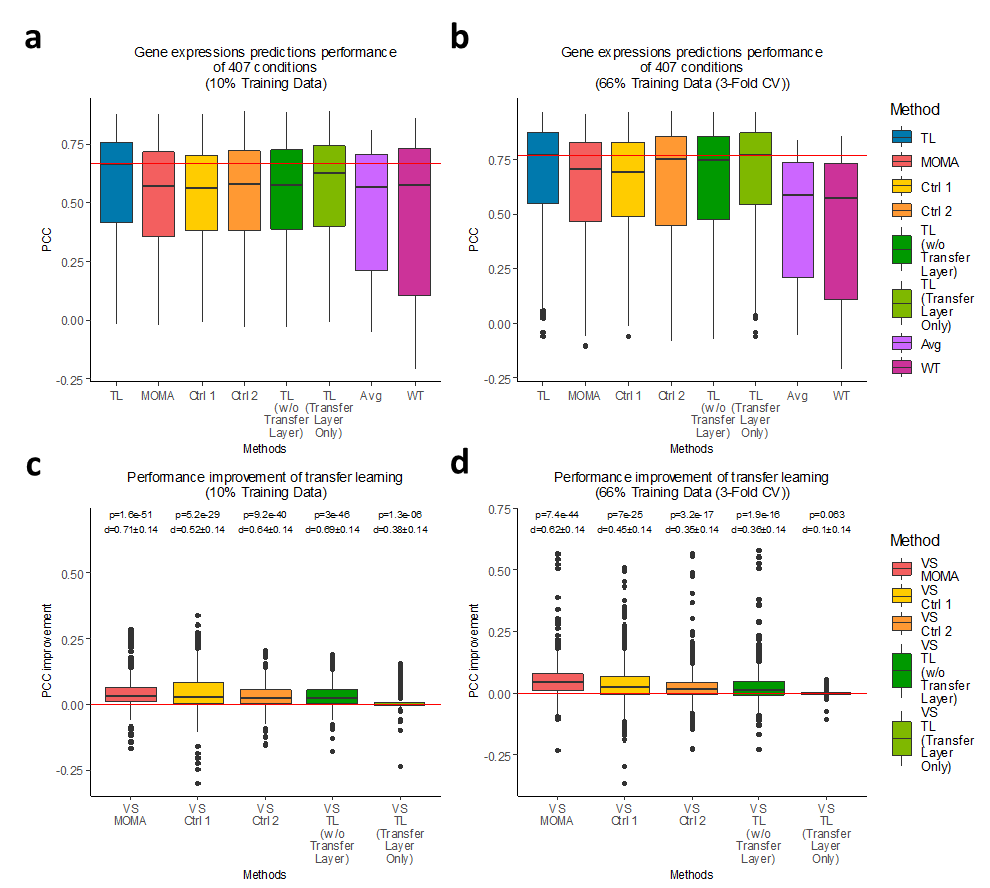


Figure 5 Gene expression performance comparison between transfer learning (TL), MOMA, 4 controls and 2 baselines in cross validation with two different training dataset ratio. (a) Results of cross validation with 10% training data. (b) Results of cross validation with 66.7% training data (3-Fold cross validation). (c) Improvement of transfer learning in cross validation with 10% training data compared with MOMA and 4 controls. (d) Improvement of transfer learning in cross validation with 66.7% training data (3-Fold cross validation) compared with MOMA and 4 controls.

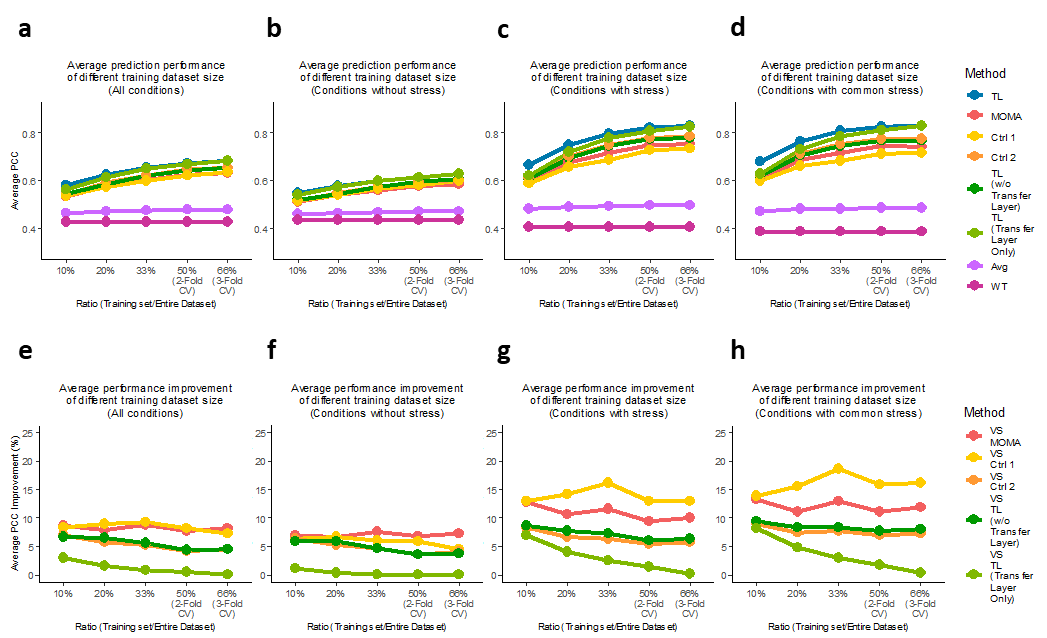


Figure 6. Performance comparison among transfer learning, MOMA, controls and baselines and improvement of transfer learning on different conditions set with different training set ratio. (a) General conditions, (b) non-stress conditions, (c) stress conditions and (d) conditions related to stress in both *E.coli* and *Salmonella* compendium. Improvement of transfer learning corresponded to these different conditions are shown in (e), (f), (g), (h).

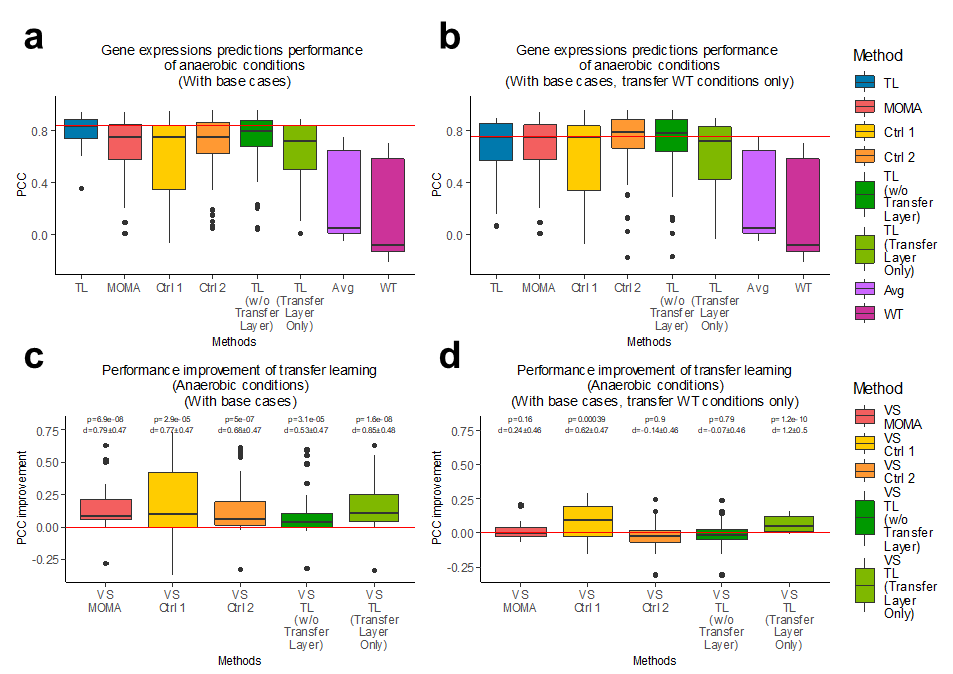


Figure 7. Gene expression performance comparison on anaerobic stress conditions among MOMA, transfer learning (TL), 4 controls and 2 baselines in cross validation. (a and c) Performance and improvement from the model trained with three selected conditions (14028s.MD141.O2-starvation.STM14\_5188\_KO, ST4/74.MD018.O2-starvation;cold.sfh\_KO and 112910a.MD018.O2-starvation;deoxynivalenol.na\_WT) in addition to non-anaerobic stress conditions and *E.coli* MOMA model is trained with all conditions. (b and d) Performance and improvement from the model trained with three selected conditions in addition to non-anaerobic stress conditions but *E.coli* MOMA model is trained with wild type conditions (without any stresses or gene perturbation) only.

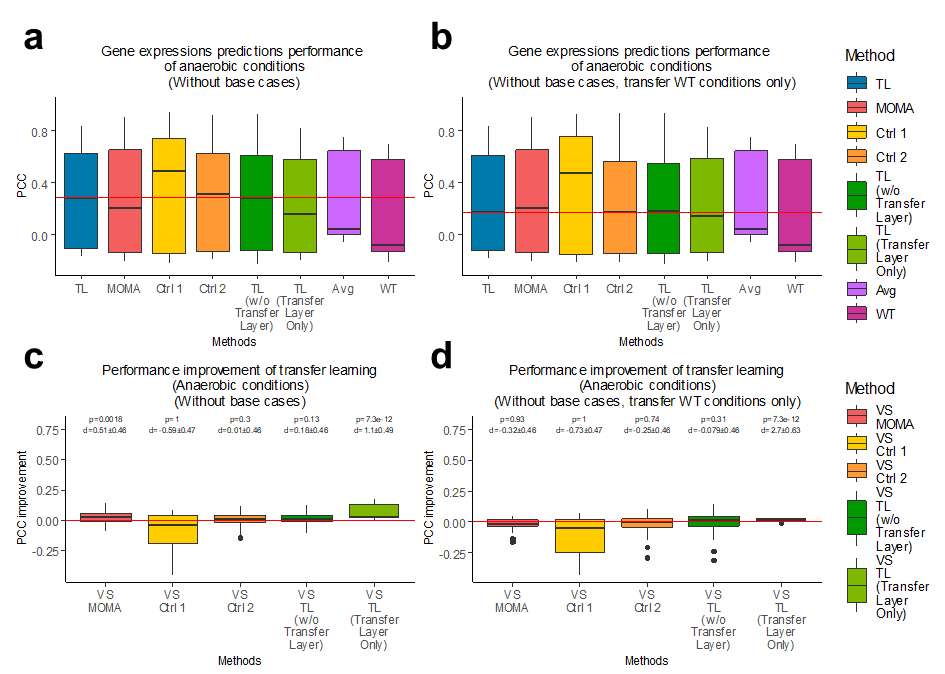


Figure 8 Gene expression performance comparison on anaerobic stress conditions among MOMA, transfer learning (TL), 4 controls and 2 baselines in cross validation. (a and c) Performance and improvement from the model trained with non-anaerobic stress conditions without additional conditions and *E.coli* MOMA model is trained with all conditions. (b and d) Performance and improvement from the model trained with non-anaerobic stress conditions but *E.coli* MOMA model is trained with wild type conditions (without any stresses or gene perturbation) only.

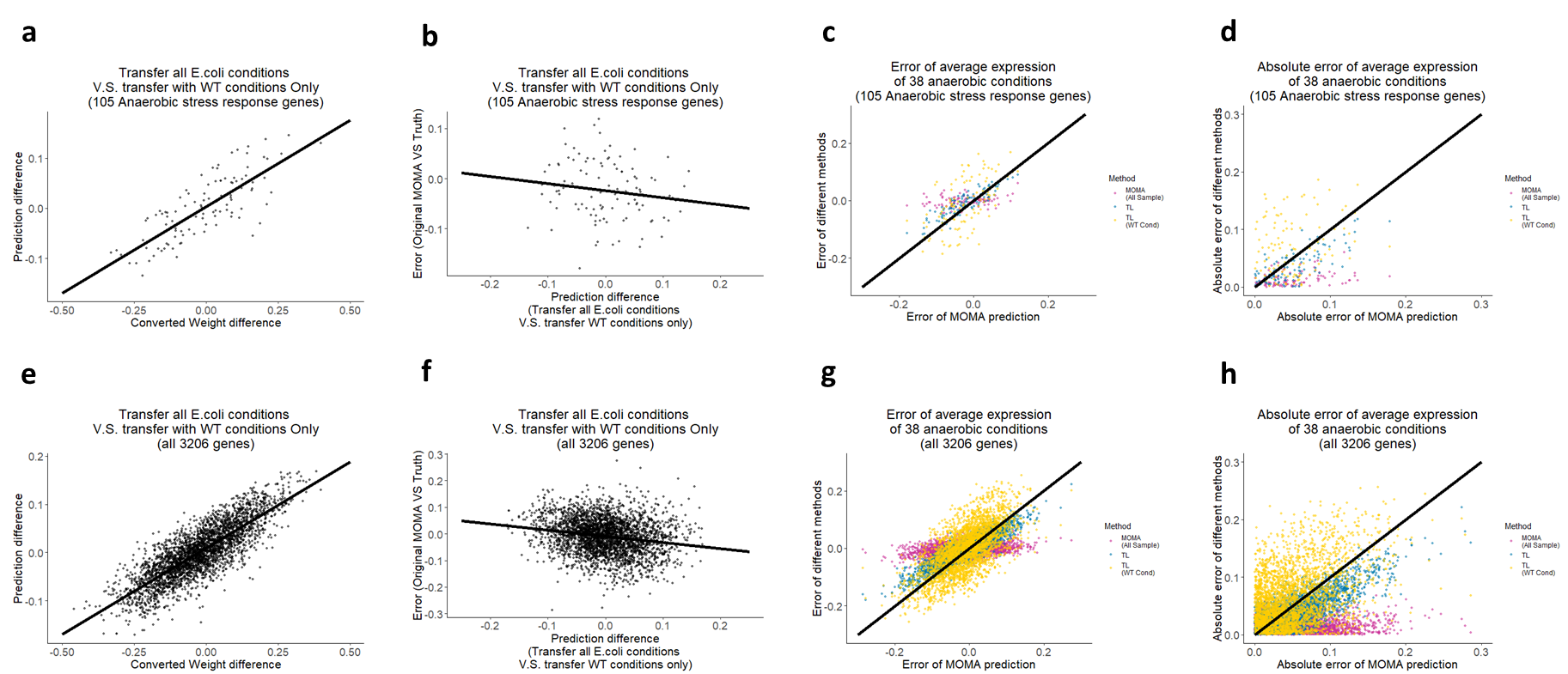


Figure 9. Anaerobic case study. (a, d) Correlation between prediction difference and difference of weights in *E.coli* MOMA trained with all conditions and trained only with wild type conditions in transfer learning model for 105 anaerobic stress response genes and all 3206 Salmonella genes. (b, f) Correlation between error in original MOMA and prediction difference between transferring all *E.coli* conditions V.S. transferring only wild type conditions for 105 anaerobic stress response genes and all 3206 Salmonella genes. (c, g) Error of original MOMA prediction and error of transfer learning with two controls: original MOMA trained with all 407 conditions (as the positive control) and transfer learning transferred with wild type conditions only for 105 anaerobic stress response genes and all 3206 Salmonella genes. (d, h) Absolute error of original MOMA prediction and error of transfer learning with two controls: original MOMA trained with all 407 conditions (as the positive control) and transfer learning transferred with wild type conditions only for 105 anaerobic stress response genes and all 3206 Salmonella genes.

**References**

(1) Parry, C. M., et al. *N Engl J Med* **2002**, *347*, 1770-82.

(2) Chin, James **2000**,

(3) Brenner, FW, et al. *Journal of clinical microbiology* **2000**, *38*, 2465-2467.

(4) Crim, Stacy M, et al. *MMWR Morb Mortal Wkly Rep* **2015**, *64*, 495-9.

(5) Frenzen, Paul D, et al. *FoodReview* **1999**,

(6) Faucher, S. P., et al. *Proc Natl Acad Sci U S A* **2006**, *103*, 1906-11.

(7) Centers for Disease, Control and Prevention *MMWR Morb Mortal Wkly Rep* **2005**, *54*, 325-8.

(8) Humphrey, T. *Nat Rev Microbiol* **2004**, *2*, 504-9.

(9) Calix-Lara, T. F., et al. *Food Microbiol* **2014**, *38*, 192-200.

(10) Arrach, N., et al. *Cancer Res* **2010**, *70*, 2165-70.

(11) Pawelek, J. M., K. B. Low, and D. Bermudes *Cancer Res* **1997**, *57*, 4537-44.

(12) Edgar, Ron, Michael Domrachev, and Alex E Lash *Nucleic acids research* **2002**, *30*, 207-210.

(13) Kodama, Yuichi, Martin Shumway, and Rasko Leinonen *Nucleic acids research* **2011**, *40*, D54-D56.

(14) Parkinson, H, et al. *Nucleic acids research* **2005**, *33*, D553-D555.

(15) Pan, S. J. and Q. A. Yang *Ieee Transactions on Knowledge and Data Engineering* **2010**, *22*, 1345-1359.

(16) Kim, M., et al. *Nat Commun* **2016**, *7*, 13090.

(17) Barrett, T., et al. *Nucleic Acids Res* **2009**, *37*, D885-90.

(18) Brazma, A., et al. *Nucleic Acids Res* **2003**, *31*, 68-71.

(19) Kodama, Y., et al. *Nucleic Acids Res* **2012**, *40*, D54-6.

(20) Coordinators, Ncbi Resource *Nucleic Acids Res* **2016**, *44*, D7-19.

(21) Altschul, S. F., et al. *J Mol Biol* **1990**, *215*, 403-10.

(22) Michael Hahsler, Anurag Nagar, rBLAST: Interface to the Basic Local Alignment Search Tool (BLAST)

(23) Langmead, B. and S. L. Salzberg *Nat Methods* **2012**, *9*, 357-9.

(24) Wang, L., S. Wang, and W. Li *Bioinformatics* **2012**, *28*, 2184-5.

(25) Anders, S., P. T. Pyl, and W. Huber *Bioinformatics* **2015**, *31*, 166-9.

(26) Barrett, T., et al. *Nucleic Acids Res* **2007**, *35*, D760-5.

(27) Smyth, Gordon K, et al. **2002**,

(28) Stekhoven, D. J. and P. Buhlmann *Bioinformatics* **2012**, *28*, 112-8.

(29) Analytics, Revolution and Steve Weston *R package version* **2014**, *1*,

(30) Bolstad, Benjamin Milo *R package version* **2013**, *1*,

(31) Kroger, C., et al. *Proc Natl Acad Sci U S A* **2012**, *109*, E1277-86.

(32) Jarvik, T., et al. *J Bacteriol* **2010**, *192*, 560-7.

(33) Cohen, Jacob *Statistical power analysis for the behavioral sciences*; L. Erlbaum Associates: 1988