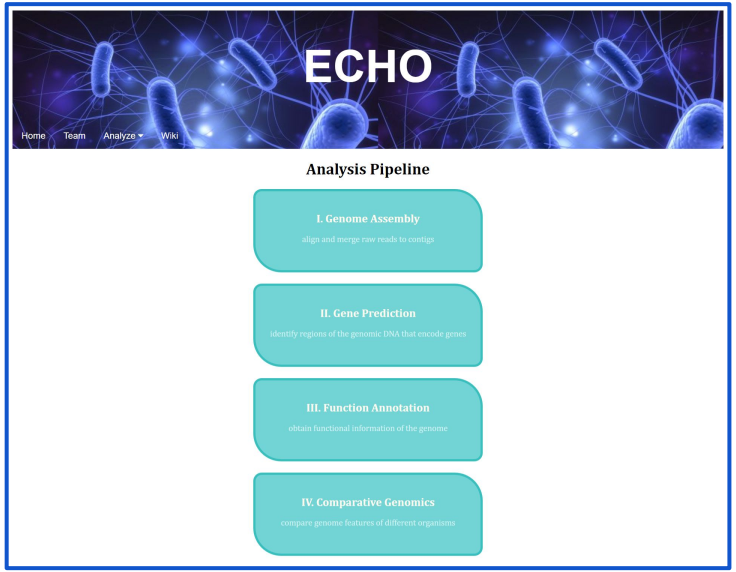
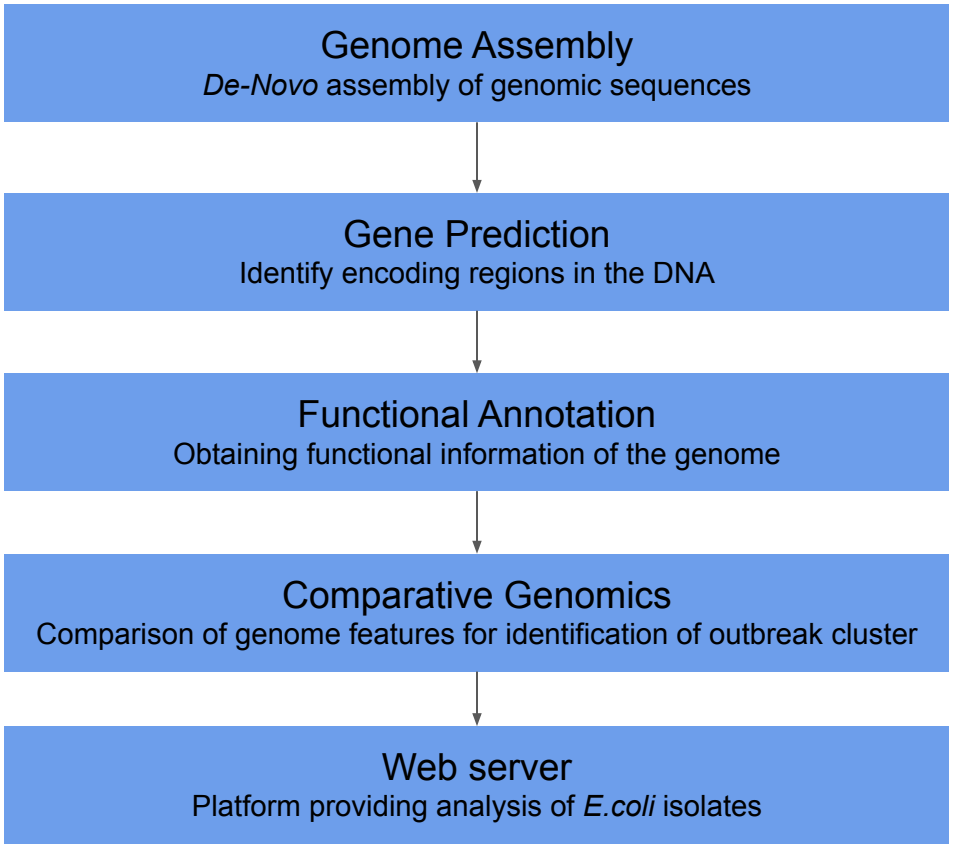
Team 1: Web server

Devishi Kesar, Shuheng Gan, Winnie Zheng, Priya Narayanan, Aaron Pfennig

E.Coli analysation web server (ECHO)




Genome Assembly
De-Novo assembly of genomic sequences
Gene Prediction
Identify encoding regions in the DNA
Functional Annotation
Obtaining functional information of the genome
Comparative Genomics
Comparison of genome features for identification of outbreak cluster
Web server
Platform providing analysis of *E.coli* isolates

Aim

- Provide a comprehensive, automated platform to analyze *E.coli* isolates in order to predict virulence factors and outbreak cluster
- Functionalities of the web server:
  - Identify virulence factors/microbial resistance and outbreak response for provided isolates
  - Allow data upload at each step of outline pipeline
  - Visualize findings in an comprehensible way
- Design
  - Intuitive usage
  - Provide only essential options

Genome Assembly

- Performs *de-novo* assembly with FastQ files as input
- Runs following tools by default:
  - fastp : read pre processing
  - Unicycler : Genome assembly
- Options:
  - Perform read preprocessing
  - Kmer-size
  - Spades as alternative assembly method
- The input FastQ files must be paired end reads
- Outputs FASTA file
- Visualisation : Quast output

Gene Prediction

- Gene finding in assembled isolates or provided FASTA fileTakes FastQ files as input
- Runs following tools by default:
  - CDS: Prodigal
  - tRNA: Aragorn
  - rRNA: barrnap
- Options:
  - GeneMarkS-2 as alternative tool for CDS predictions
  - tRNAscan-SE as alternative tool for tRNA predictions
  - RNAmmer as alternative tool for rRNA predictions
- Outputs *.gff file, *_cds.fna file, *_protein.faa file and *_rna.fna file

Functional Annotation

- Obtain functional information about predicted genes
- Input: FASTA file
- Cluster Tool: usearch
  - Output: centroid.fasta
- Homology Tools:
  - General annotation: InterProScan, EggNOGmapper
  - Antibiotic resistance gene: DeepARG
- Abinitio Tools:
  - Signal Peptides: SignalP 5.0
  - Transmembrane Proteins: TMHMM
  - CRISPR Sites: PilerCR
- Output: *.tsv file

Comparative Genomics

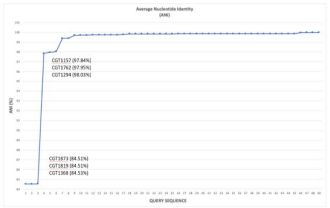
- Comparison of genomic features of input files to identify outbreak cluster
- Input: FASTA file, prodigal training file(chewBBACA)
- Tools used:
  - MUMmer 4.0
  - chewBBACA
  - kSNP 3.0
  - FigTree
- Options:
  - Parsimony tree, maximum likelihood and neighbour joining trees as option for kSNP
  - k-mer size option for kSNP
- Output: .tsv file(for chewBBACA, MUMmer), .png(kSNP)
- Visualisation: Phylogenetic tree for identified SNP’s, phylogenetic tree for MLST, graph for epidemiological data visualisation

Visualization Results

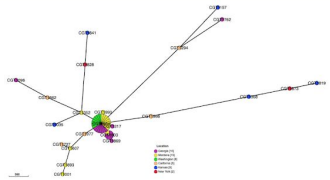
Class results

We have analyzed 50 E.coli isolates, including epi data, with respect to a foodborn outbreak. The data has been analyzed using this web server and the final results are presented below. We analyzed the data by performing de-novo genome assembly, gene prediction, functional annotation and comparative genomics. We have paid special attention to the virulence factors, possible food sources and the outbreak location.

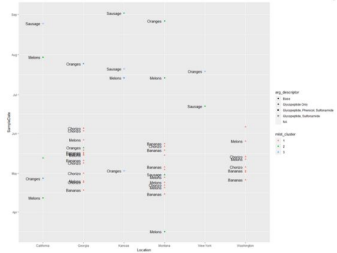
First, we have chosen one our isolates as reference and have determined the Average Nucleotide Identity of all other isolates with respect to our selected reference. Three isolates have a relatively low ANI of approximately 84%. Three other isolates have an ANI between 97-98% signifying some differences in regions of the genome with respect to the reference. All other 44 isolates are closely related to the reference genome with an ANI of approximately 99%. To determine the ANI MUMMER-4.0 has been used which has low resolution and does not discriminate more details about differences between highly similar genomes. The plot below shows the ANIs with respect to the reference genome:



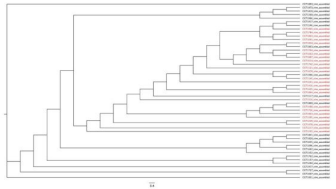
Subsequently, we have performed MLST analysis using chewBBACA to create a schema and do allele calling on the assembled genomes of the 50 isolates. Thereby, cluster cluster outbreak isolates have been identified. These preliminary results combined with epidemiological data allowed to narrow down outbreak locations. The results were visualized using Grapetree and are shown below:



In our strain analysis we incorporated functional annotation results. The results supported with the results from the MLST analysis yielding hints on three food sources: melons, chorizo and bananas which are all served at certain brunch places. Futhermore, we asked the question whether these clear strains, possessed of clear genetic relatedness, are treatable in similar fashion. Therefore, these strains have been analyzed using deepARG and (fortunately) they are identical on this basis and vulnerable to phenicol and sulfonamides. The insights gained from the MLST analysis, strain analysis and epidemiological data is depicted below:



In addition SNP analysis has been performed using kSNP 3.0. The optimal k-mer size has been determined using Kchooser which yields, as a nice feature, the fraction of kmers that are present in all genomes. The FCK value is a measure of sequence diversity and hence a measure of relatedness. The lower the FCK, the more diverse and hence the more distantly related. The FCK is 0.422. Studies have shown when FCK is ≥ 0.1 SNP detection efficiency is adequate, and the accuracy of parsimony trees estimated by kSNP3 is > 97%; i.e. the trees can be considered to be reliable. The tree is shown below:



References

1. Maiden MC, Jansen van Rensburg MJ, Bray JE, et al. MLST revisited: the gene-by-gene approach to bacterial genomics. Nat Rev Microbiol. 2013;11(10):728-36.
2. Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, et al. (2018) MUMmer4: A fast and versatile genome alignment system. PLOS Computational Biology 14(1): e1005944. https://doi.org/10.1371/journal.pcbi.1005944
3. Perez-Losada M, Arenas M, Castro-Nallar E. Microbial sequence typing in the genomic era. Infection, Genetics and Evolution. 2018;63:346-359. http://dx.doi.org/10.1016/j.meegid.2017.09.022
4. Streckbine N, Bopp C, Fields P, Kasper J, Nataro J. 2015. Escherichia, Shigella, and Salmonella, p 685-713. In Jorgensen J, Pfaller M, Carroll K, Funke G, Landry M, Richter S, Warnock D (ed), Manual of Clinical Microbiology, Eleventh Edition. ASM Press, Washington, DC. doi: 10.1128/9781555817381.ch37
5. Sultan, I., Rahman, S., Jan, A. T., Siddiqui, M. T., Mondal, A. H., & Haq, Q. M. R. (2018). Antibiotics, Resistome and Resistance Mechanisms: A Bacterial Perspective. Frontiers in Microbiology, 9(2066). doi:10.3389/fmicb.2018.02066
6. Trees E, Reta P, Maccannell D, Gerner-smidt P. Molecular Epidemiology, p 131-159. In Jorgensen J, Pfaller M, Carroll K, Funke G, Landry M, Richter S, Warnock D (ed), Manual of Clinical Microbiology, Eleventh Edition. ASM Press, Washington, DC. 2015. doi: 10.1128/9781555817381.ch10
7. Silva M, Machado M, Silva D, Rossi M, Moran-Gilad J, Santos S, Ramirez M, Carriço J. 15/03/2018. M Gen 4(3): doi:10.1099/mgen.0.000166
8. Z Zhou, NF Alikhan, MJ Sergeant, N Luhmann, C Vaz, AP Francisco, JA Carrico, M Achtman (2018) "GrapeTree: Visualization of core genomic relationships among 100,000 bacterial pathogens", Genome Res; doi: https://doi.org/10.1101/gr.232397.117

Demo