

Metagenome Analysis of Premature Birth

Jaewoong Lee Semin Lee

Department of Biomedical Engineering
Ulsan National Institute of Science and Technology

jwlee230@unist.ac.kr

2021-09-24

Overview

- 1 Introduction
- 2 Materials
- 3 Methods
- 4 Results
- 5 Discussion

Introduction

Microbiome

- Microbiota: the microorganisms which live inside & on humans (Turnbaugh et al., 2007)
- Microbiome: 10^{13} to 10^{14} microorganisms whose collective genome (Gill et al., 2006)



Figure: Concept of a core human microbiome (Turnbaugh et al., 2007)

- Ribosomal RNA
- Well-known as a key to phylogeny (Olsen & Woese, 1993)

Premature Birth (Preterm Birth; PTB)

PTB:

- 1 PTB < 37 GW (Gestational week)
- 2 Normal ≥ 37 GW

Detailed PTB:

- 1 Extremely PTB < 28 GW
- 2 $28 \text{ GW} \leq \text{Very PTB} < 34 \text{ GW}$
- 3 $34 \text{ GW} \leq \text{Late PTB} < 37 \text{ GW}$
- 4 Normal $\geq 37 \text{ GW}$

(J. Tucker & McGuire, 2004; Voronkov, Solonovych, Liashenko, & Revenko, 2018)

Materials

16S rRNA Sequencing

16S rRNA sequencing is the *reference method* for bacterial taxonomy & identification (Mignard & Flandrois, 2006)

Three main reasons (Janda & Abbott, 2007):

- 16S rRNA exists in almost all bacteria
- Functions of the 16S rRNA has not changed over time
- 16S rRNA is large enough for bioinformatics

Train/Test Data vs. Validate Data

- JBNU/Helixco data
 - First data
 - Second data
 - Stool data

Table: Sample Information

Data	Participants	Samples	Remarks
First	24	107	-
Second	35	288	-
Third	10	106	-
Stool	63	126	Stool

Methods

Methods

Qiime 2 Workflow

Qiime 2 Workflow



Figure: QIIME 2 workflow (Bolyen et al., 2019; Mandal, Van Treuren, White, Eggesbø, et al., 2015; McDonald et al., 2012)

Filtering with Quality Score

Drawback between:

- Longer sequence read
- Higher quality value

∴ Select the maximum length n where:

$$\begin{aligned} \forall n_i \in \{n_k | \text{MedianQualityScore} \geq 30\} \\ \exists ! n \in \{n_i\} : n \geq n_i \end{aligned} \tag{1}$$

Denoising Techniques

- DADA2: Amplicon Sequence Variants (ASVs) (Callahan et al., 2016)
- Deblur: Operational Taxonomic Units (OTUs) (Amir et al., 2017)



Figure: Denoising Algorithms

Taxonomy Classification

- Greengenes (GG) (DeSantis et al., 2006)
- SILVA (Pruesse et al., 2007; Quast et al., 2012)



Figure: Taxonomy Classifications

“A **higher** performance at taxonomic levels above *genus level*;
but performance appears to **drop** at *species level*” (Gihawi et al., 2019)

Merging Denoising/Taxonomy

Merging multiple IDs (ASVs or OTUs) into one, which have

- Different IDs
- Identified as same taxonomy

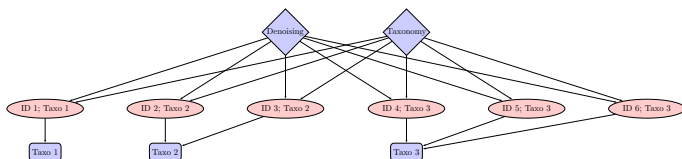


Figure: Example Diagram for Merging Denoising/Taxonomy

Methods

Abundance Test

- Analysis of composition of microbiome (Mandal, Van Treuren, White, Eggesbø, et al., 2015)
- ANCOM detects significantly abundant taxa, while maintain high statistical power
- Find taxa that can divide each classes

Methods

Diversity Indices

Diversity Indices



Figure: Three dimensions of phylogenetic information (C. M. Tucker et al., 2017)

- A quantitative measure that shows richness, divergence, and regularity (C. M. Tucker et al., 2017)
- Alpha diversity indices: the richness of taxa **at a single community**
- Beta diversity indices: the taxonomic differentiation **between communities**

Alpha Diversity Indices

- Evenness index
- Faith's Phylogenetic Diversity (Faith PD) index
- Observed Features index
- Shannon's Diversity index

Beta Diversity Indices

- Bray-Curtis distance index
- Jaccard distance index
- Unweighted UniFrac distance index
- Weighted UniFrac distance index

Methods

Miscellaneous

t-distributed Stochastic Neighbor Embedding (t-SNE)

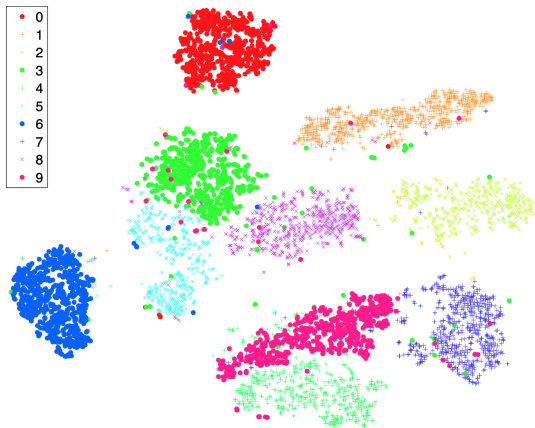


Figure: t-SNE with handwritten data (Maaten & Hinton, 2008)

- Pandas (McKinney et al., 2011)
- Scikit-Learn (Pedregosa et al., 2011)
- SciPy (Virtanen et al., 2020)
- Matplotlib (Hunter, 2007)
- Seaborn (Waskom et al., 2020)
- Statannot

Results

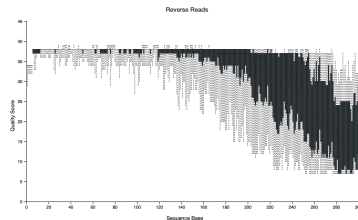
Results

Filtering Results

Quality Score from JBNU/Helixco Data



(a) Forward



(b) Reverse

Figure: Quality Score Plot

- $\ell_{Forward} = 300$
- $\ell_{Reverse} = 245$

Results

t-SNE with Clinical Information

Workflow for t-SNE with Site/Premature Information



Figure: Workflow of t-SNE for Site/Premature Information

- Diseases
 - ① Gestational Diabetes
 - ② Maternal Overweight/Obesity
 - ③ Maternal Weight Gain
 - ④ Hypertension
 - ⑤ PROM
 - ⑥ Antibiotic
 - ⑦ Steroid
- Probing sites
 - ① Maternal Mouth

Selected t-SNE Plots I

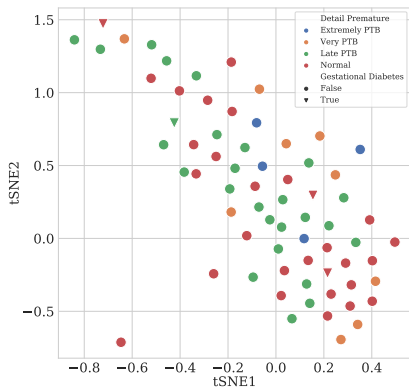


Figure: t-SNE about Gestational Diabetes

Selected t-SNE Plots II



(a) Overweight



(b) Weight gain



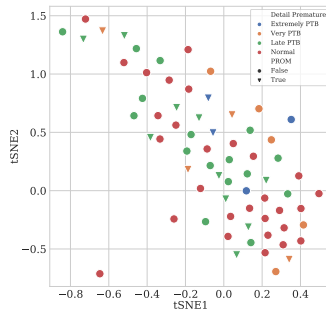
(c) Too much gain

Figure: t-SNE about Maternal Weight

Selected t-SNE Plots III



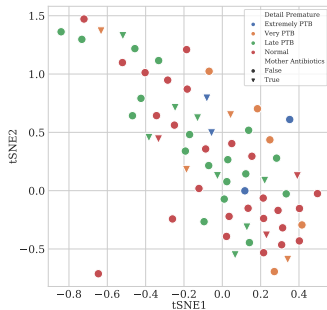
(a) Hypertension



(b) PROM

Figure: t-SNE about Disease

Selected t-SNE Plots IV



(a) Mother



(b) Neonate

Figure: t-SNE about Antibiotics Usage

Selected t-SNE Plots V

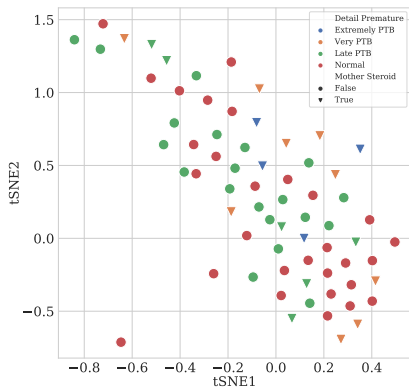


Figure: t-SNE about Steroid Usage

Results

Bacterial Abundance Test with ANCOM

- Analysis of composition of microbiomes
- ANCOM can be used for analyzing the composition of microbiomes in multiple populations (Mandal, Van Treuren, White, Eggesbø, et al., 2015)
- Differential abundance testing
- ① clr: Centered log(*Ratio*)
- ② W: a count of the number of sub-hypothesis which have passed for given species

Site selection:

- 1 Neonatal mouth: 1-day, 3-day, and 5-day
- 2 Cervix
- 3 Maternal mouth
- 4 Vagina

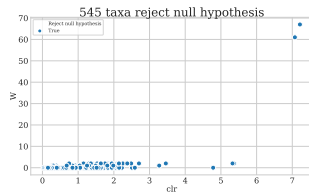
PTB:

- 1 Premature
- 2 Detail premature

ANCOM with Neonatal Mouth



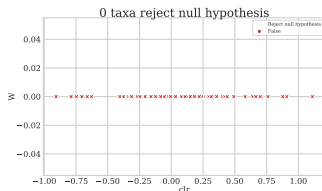
(a) PTB



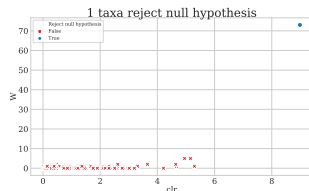
(b) Detailed PTB

Figure: ANCOM with Neonatal Mouth

ANCOM with Maternal Mouth



(a) PTB

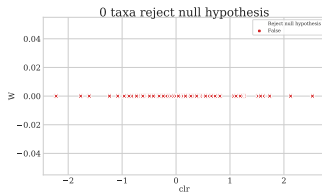


(b) Detailed PTB

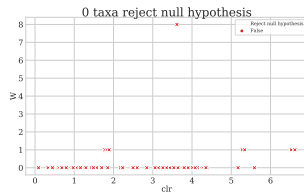
Figure: ANCOM with Maternal Mouth

- *Bacteria Proteobacteria Alphaproteobacteria Rickettsiales mitochondria* family

ANCOM with Cervix



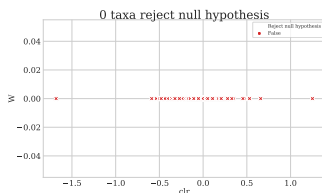
(a) PTB



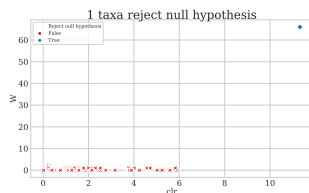
(b) Detailed PTB

Figure: ANCOM with Cervix

ANCOM with Vagina



(a) PTB



(b) Detailed PTB

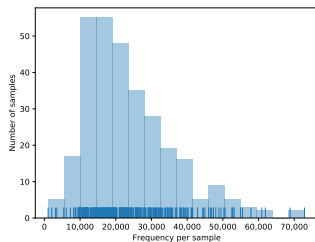
Figure: ANCOM with Vagina

- *Bacteria Proteobacteria Epsilonproteobacteria Campylobacterales Campylobacteraceae Campylobacter genus*

Results

Diversity Index

Rarefaction



(a) DADA2



(b) Deblur

Figure: Rarefaction from the Data

- $\min(\ell_{DADA2})$: 1046
- $\min(\ell_{Deblur})$: 4864

Diversity Indices

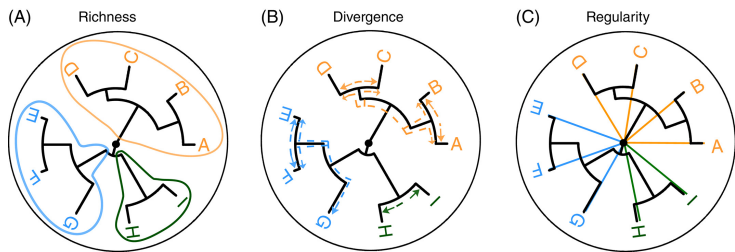
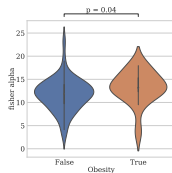


Figure: Dimensions of phylogenetic information (C. M. Tucker et al., 2017)

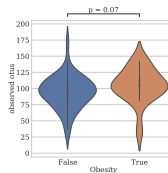
- Alpha-diversity: the species diversity *in* a local scale
- Beta-diversity: the species diversity *between* local scales

- Alpha diversity indices
 - ① Faith PD
 - ② Observed Features
 - ③ Pielou Evenness
 - ④ Shannon Entropy
- Diseases/Conditions
 - ① Gestational Diabetes
 - ② Too much Weight Gain
 - ③ Overweight/Obesity
 - ④ Hypertension
 - ⑤ PROM
 - ⑥ Antibiotics
 - ⑦ Steroid
- Site Selection
 - ① Maternal Mouth
 - ② Neonatal Mouth - 3 day

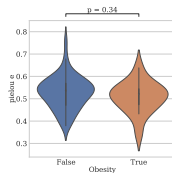
Alpha-diversity Violin Plots I



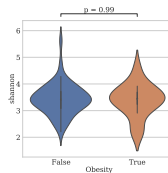
(a) Fisher alpha



(b) Observed OTUs



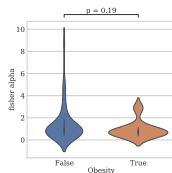
(c) Pielou e



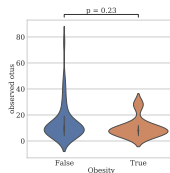
(d) Shannon

Figure: Alpha-diversities by Overweight/Obesity in Maternal Mouth

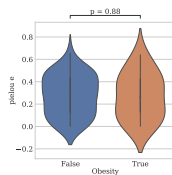
Alpha-diversity Violin Plots II



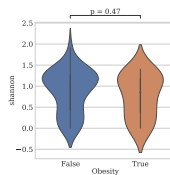
(a) Fisher alpha



(b) Observed OTUs



(c) Pielou e



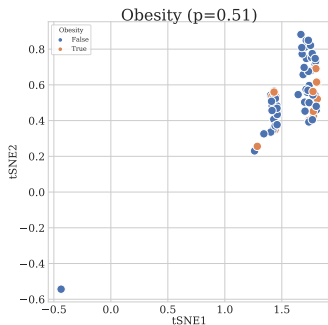
(d) Shannon

Figure: Alpha-diversities by Overweight/Obesity in Neonatal-3day

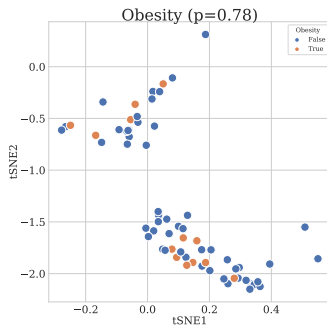
Beta-diversity

- Beta diversity indices
 - ① Bray-Curtis distance
 - ② Jaccard distance
 - ③ Unweighted UniFrac distance
 - ④ Weighted UniFrac distance
- Diseases/Conditions
 - ① Gestational Diabetes
 - ② Too much Weight Gain
 - ③ Overweight/Obesity
 - ④ Hypertension
 - ⑤ PROM
 - ⑥ Antibiotics
 - ⑦ Steroid
- Site Selection
 - ① Maternal Mouth
 - ② Neonatal Mouth - 3 day

Beta-diversity Scatter Plots I



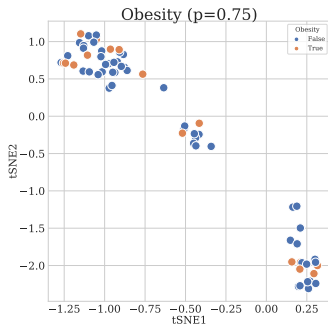
(a) Unweighted UniFrac



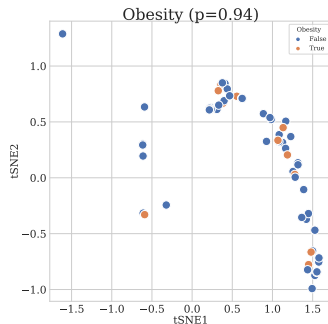
(b) Weighted UniFrac

Figure: Beta-diversities by Overweight/Obesity in Maternal Mouth

Beta-diversity Scatter Plots II



(a) Unweighted UniFrac



(b) Weighted UniFrac

Figure: Beta-diversities by Overweight/Obesity in Neonate 3-day

Results

Metagenomic Biomarker Discovery

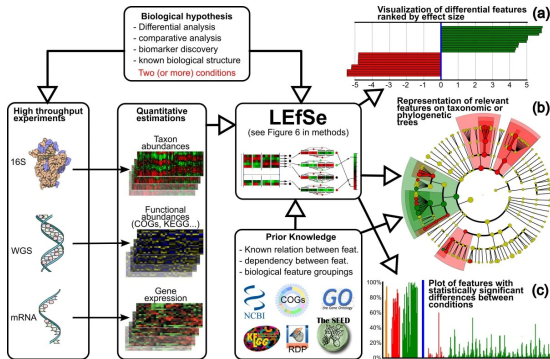
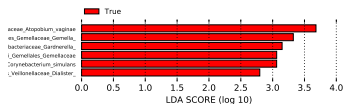


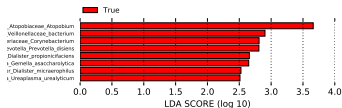
Figure: LEfSe workflow (Segata et al., 2011)

- Analysis Pipeline
 - Algorithm: DADA2 and Deblur
 - Databases: GG, Silva, and HOMD
- Subject Groups
 - PTB
 - Normal
- Sites
 - Cervix
 - Mouth
 - Neonatal Mouth
 - Vagina

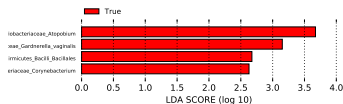
LEfSe Results with Obesity I



(a) Deblur+GG



(b) Deblur+Silva



(c) Deblur+HOMD

Figure: LEfSe in Cervix

LEfSe Results with Obesity II

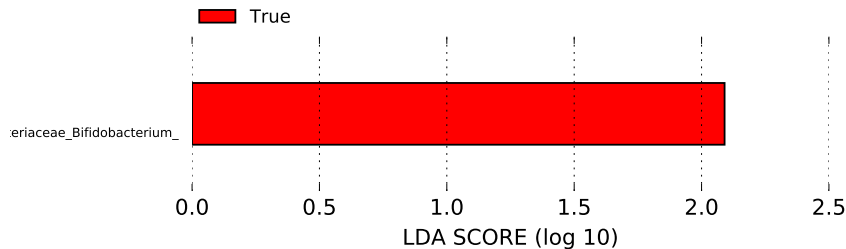


Figure: DADA2+GG in Vagina

Discussion

To-do List

- Amir, A., McDonald, D., Navas-Molina, J. A., Kopylova, E., Morton, J. T., Xu, Z. Z., ... others (2017). Deblur rapidly resolves single-nucleotide community sequence patterns. *MSystems*, 2(2).
- Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., ... Caporaso, J. G. (2019). Reproducible, interactive, scalable and extensible microbiome data science using qiime 2. *Nature Biotechnology*, 37(8), 852-857. Retrieved from <https://doi.org/10.1038/s41587-019-0209-9> doi: 10.1038/s41587-019-0209-9
- Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). Dada2: high-resolution sample inference from illumina amplicon data. *Nature methods*, 13(7), 581-583.

References II

- DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., ... Andersen, G. L. (2006). Greengenes, a chimera-checked 16s rRNA gene database and workbench compatible with ARB. *Applied and environmental microbiology*, 72(7), 5069–5072.
- Gihawi, A., Rallapalli, G., Hurst, R., Cooper, C. S., Leggett, R. M., & Brewer, D. S. (2019). Sepath: benchmarking the search for pathogens in human tissue whole genome sequence data leads to template pipelines. *Genome biology*, 20(1), 1–15.
- Gill, S. R., Pop, M., DeBoy, R. T., Eckburg, P. B., Turnbaugh, P. J., Samuel, B. S., ... Nelson, K. E. (2006). Metagenomic analysis of the human distal gut microbiome. *science*, 312(5778), 1355–1359.
- Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in science & engineering*, 9(3), 90–95.

- Janda, J. M., & Abbott, S. L. (2007). 16s rrna gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. *Journal of clinical microbiology*, 45(9), 2761–2764.
- Maaten, L. v. d., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov), 2579–2605.
- Mandal, S., Van Treuren, W., White, R. A., Eggesbø, M., Knight, R., & Peddada, S. D. (2015). Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microbial ecology in health and disease*, 26(1), 27663.
- Mandal, S., Van Treuren, W., White, R. A., Eggesbø, M., Knight, R., & Peddada, S. D. (2015). Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microbial ecology in health and disease*, 26(1), 27663. doi: 10.3402/mehd.v26.27663

- McDonald, D., Clemente, J. C., Kuczynski, J., Rideout, J. R., Stombaugh, J., Wendel, D., ... Caporaso, J. G. (2012). The biological observation matrix (biom) format or: how i learned to stop worrying and love the ome-ome. *GigaScience*, 1(1), 7. doi: 10.1186/2047-217X-1-7
- McKinney, W., et al. (2011). pandas: a foundational python library for data analysis and statistics. *Python for High Performance and Scientific Computing*, 14(9).
- Mignard, S., & Flandrois, J.-P. (2006). 16s rna sequencing in routine bacterial identification: a 30-month experiment. *Journal of microbiological methods*, 67(3), 574–581.
- Olsen, G. J., & Woese, C. R. (1993). Ribosomal rna: a key to phylogeny. *The FASEB journal*, 7(1), 113–123.

- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... others (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12, 2825–2830.
- Pruesse, E., Quast, C., Knittel, K., Fuchs, B. M., Ludwig, W., Peplies, J., & Glöckner, F. O. (2007). Silva: a comprehensive online resource for quality checked and aligned ribosomal rna sequence data compatible with arb. *Nucleic acids research*, 35(21), 7188–7196.
- Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., ... Glöckner, F. O. (2012). The silva ribosomal rna gene database project: improved data processing and web-based tools. *Nucleic acids research*, 41(D1), D590–D596.
- Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W. S., & Huttenhower, C. (2011). Metagenomic biomarker discovery and explanation. *Genome biology*, 12(6), 1–18.

- Tucker, C. M., Cadotte, M. W., Carvalho, S. B., Davies, T. J., Ferrier, S., Fritz, S. A., ... others (2017). A guide to phylogenetic metrics for conservation, community ecology and macroecology. *Biological Reviews*, 92(2), 698–715.
- Tucker, J., & McGuire, W. (2004). Epidemiology of preterm birth. *Bmj*, 329(7467), 675–678.
- Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., & Gordon, J. I. (2007). The human microbiome project. *Nature*, 449(7164), 804–810.
- Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., ... others (2020). Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3), 261–272.

- Voronkov, L., Solonovych, A., Liashenko, A., & Revenko, I. (2018). Prognostic value of cognitive tests and their combination in patients with chronic heart failure and reduced left ventricular ejection fraction. *Eureka: health sciences*(6), 36–45.
- Waskom, M., Botvinnik, O., Ostblom, J., Gelbart, M., Lukauskas, S., Hobson, P., ... Brian (2020, April). *mwaskom/seaborn: v0.10.1 (april 2020)*. Zenodo. Retrieved from <https://doi.org/10.5281/zenodo.3767070> doi: 10.5281/zenodo.3767070