Directional association inference challenged by severe dropout in single-cell RNAseq data

Eliška Dvořáková

dvorael1@fel.cvut.cz

supervisor: Prof. Joe Song

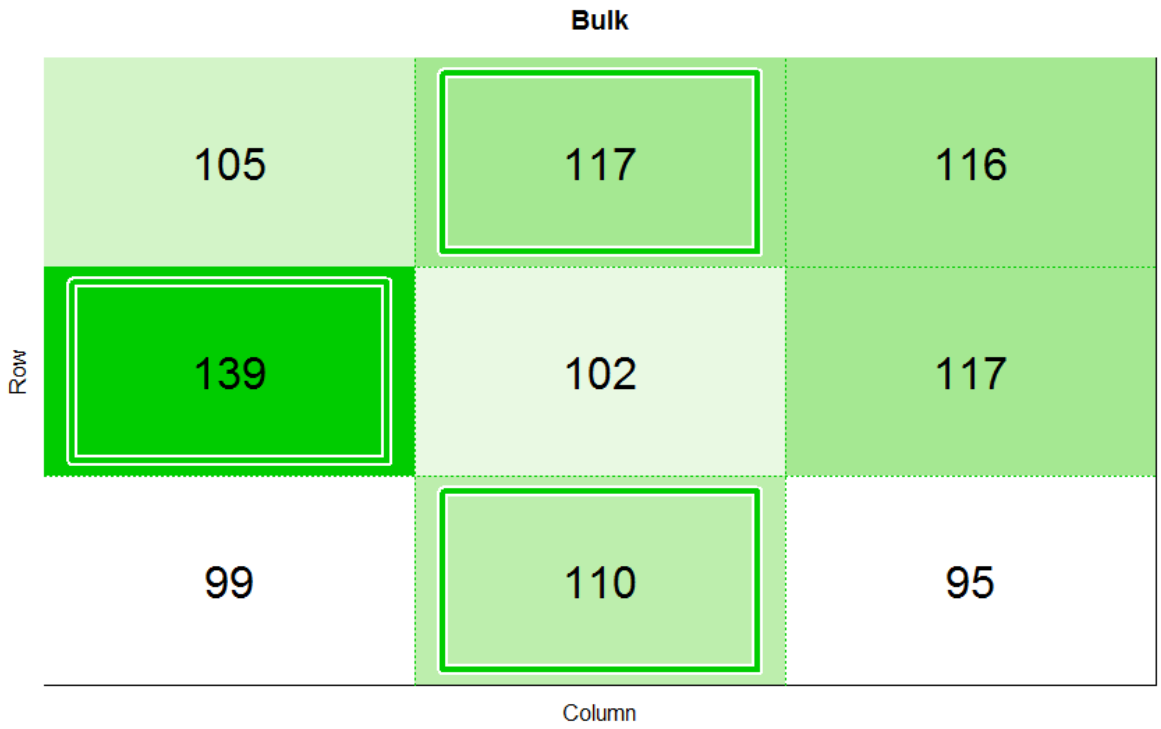
ČVUT
ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

Motivation

- **Single-cell RNA sequencing (scRNA) allows to gain RNA-seq from small amount of initial material. (single cell)**
- **New field of study focus on data from single cell.**
- **scRNA suffers from 0.9 dropout.**
- **Methods for Bulk RNA-seq performs close to random**
- **New methods FunChisq [1].**

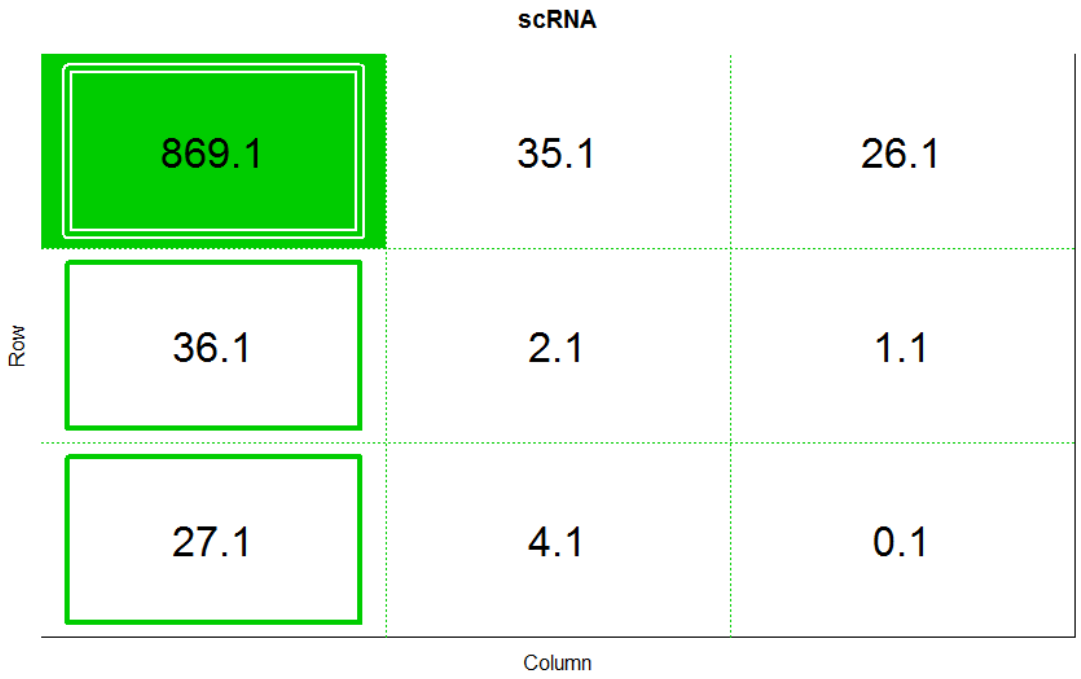
Bulk RNA-seq

- **Input material: multiple cell**
- **No dropout**
- **1000 samples**



Single-cell RNA-seq

- **Input material: single cell**
- **0.9 dropout**
- **1000 samples**



ČVUT
ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

Methods

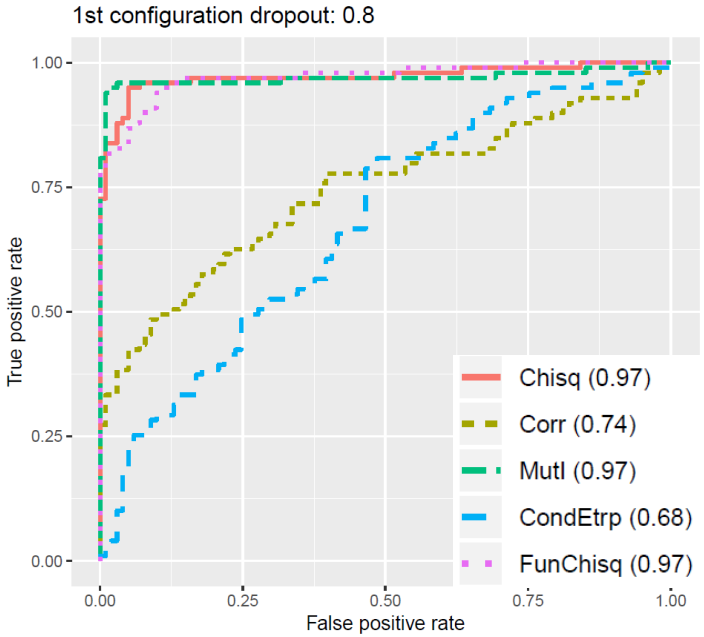
1. **Pearson's Correlation test**
2. **Mutual Information**
3. **Conditional Entropy (directional)**
4. **Pearson's χ^2-test**
5. **Functional χ^2-test (directional) [1]**

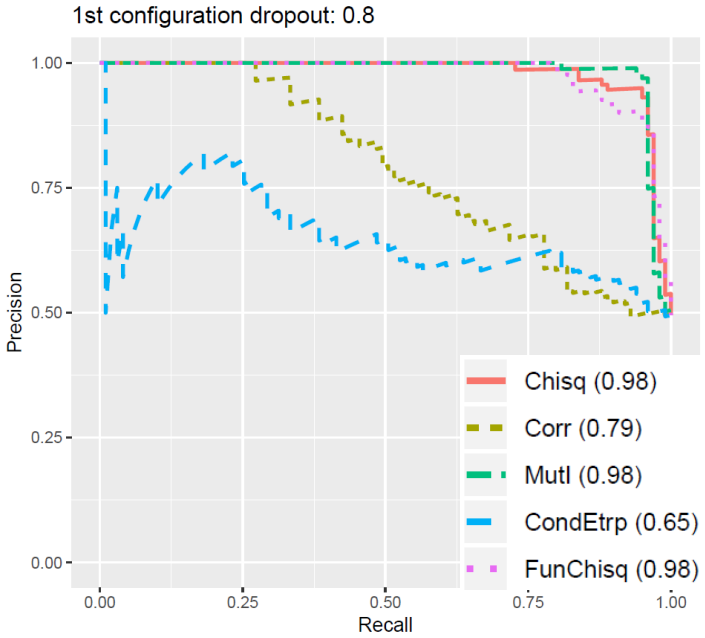
ČVUT
ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

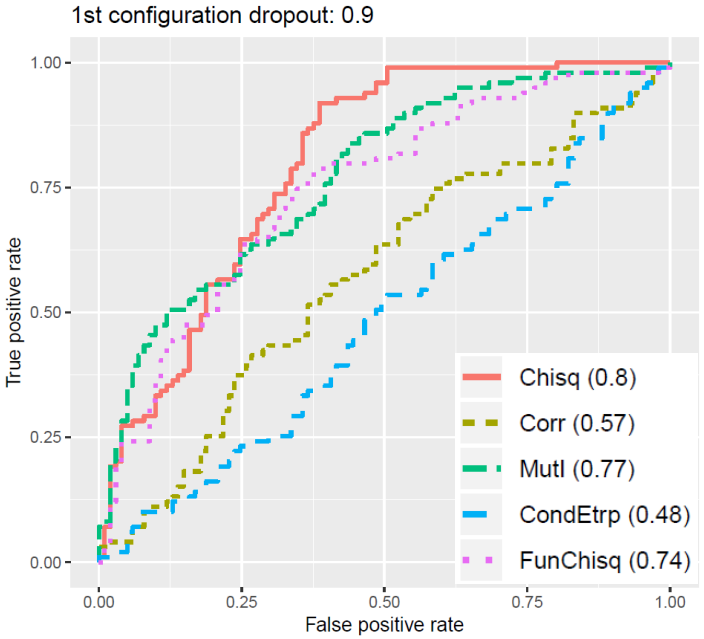
Simulated dataset

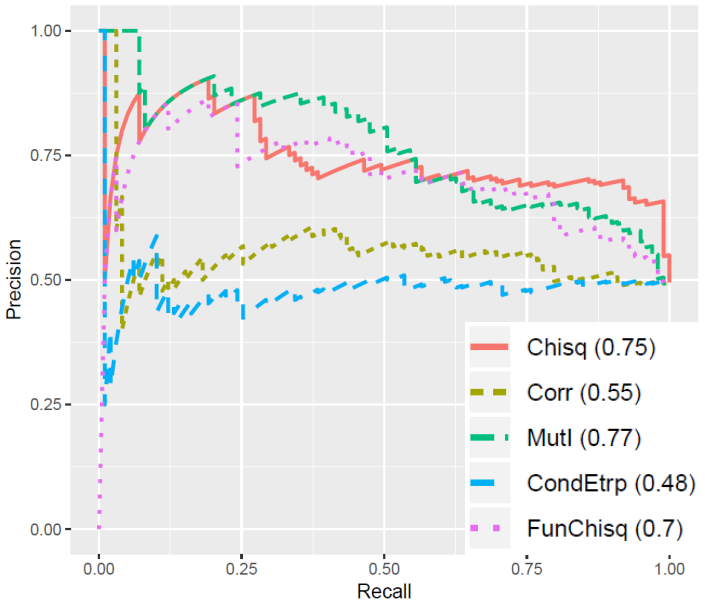
- **Size: 200 tables**
- **Dimension : 3x3**
- **Samples: 1000**
- **Noise: 0.0, 0.2**
- **Dropout: 0.2, 0.8, 0.9, 0.99**
- **Configurations:**
    1. **Detection of relationship**
    2. **Detection of relationship direction**

Results 1st Configuration

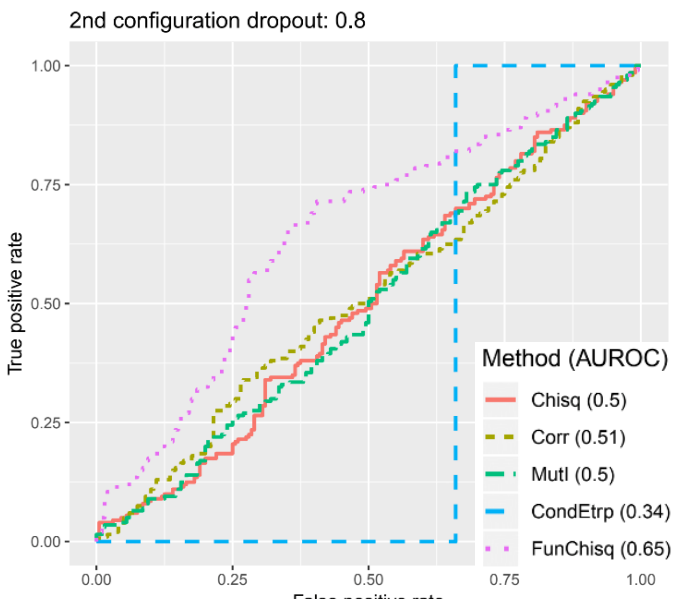


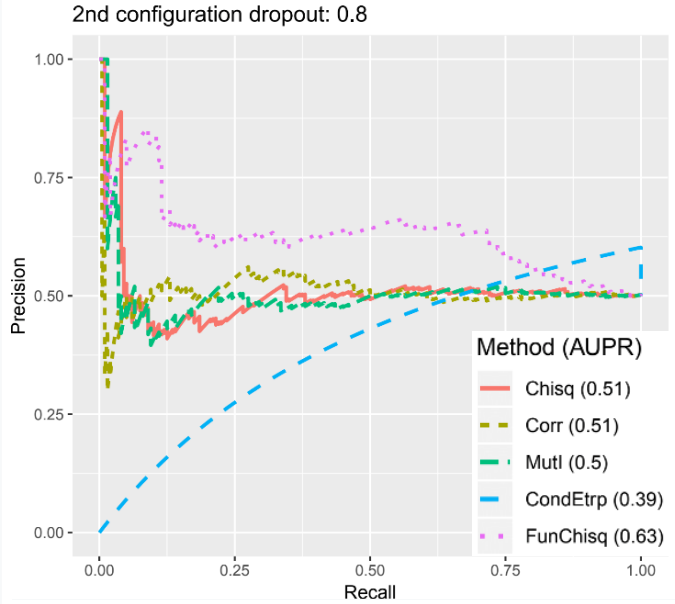


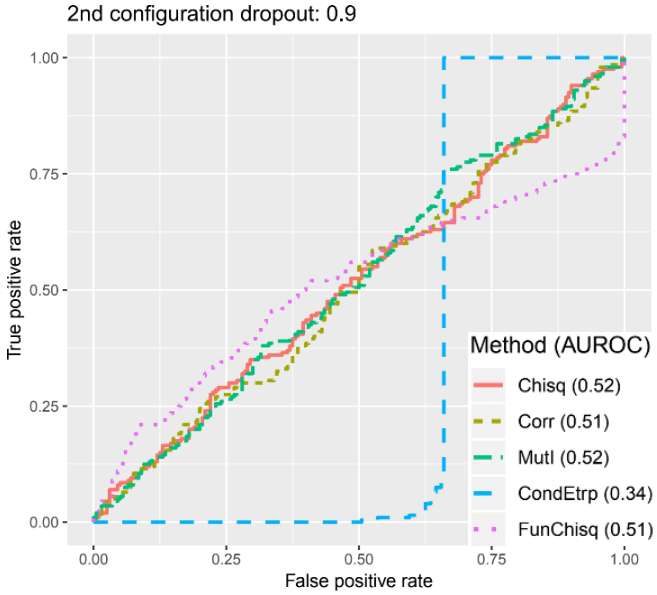


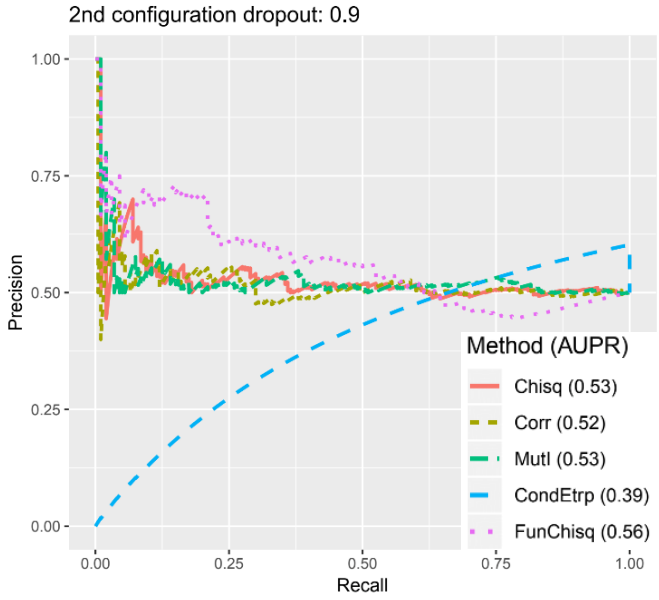


Results 2nd Configuration









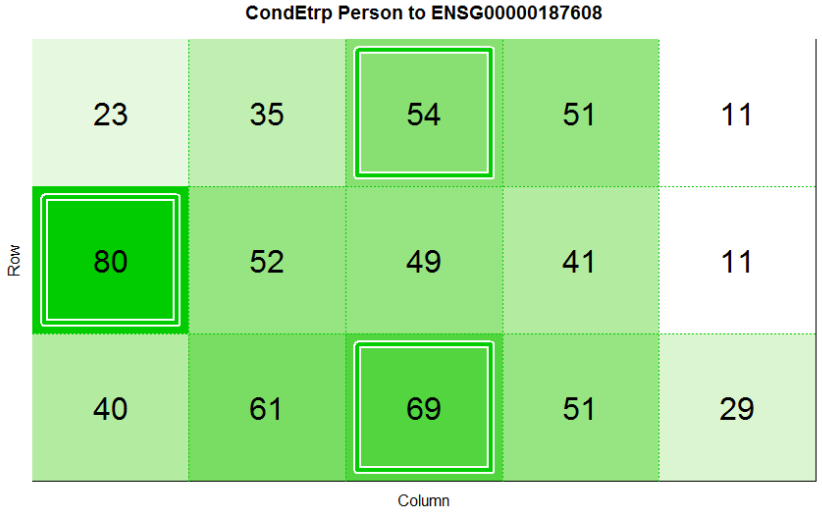
ČVUT
ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

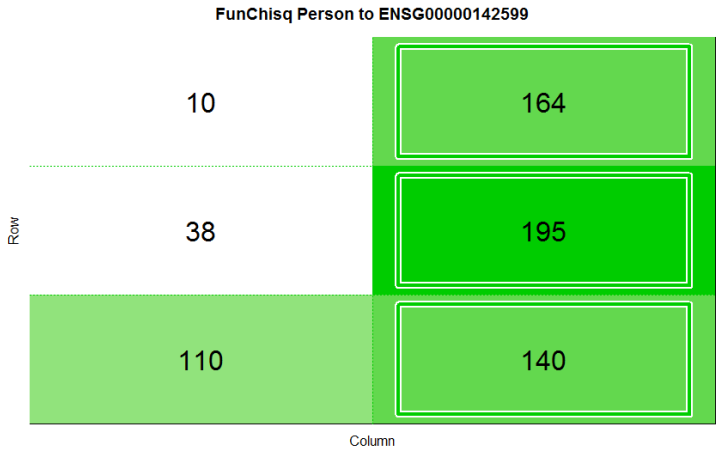
Real dataset biological background

- Dataset gained from Tung et al. [2]
- single-cell Fluidigm C1 platform
- three C1 replicates from three human induced pluripotent stem cell (iPSC) lines
- unique molecular identifiers (UMI) to all samples
- The paper is focused on finding source of variation in gene expression data.
    - Genotype
    - UMI counts are biased

Real dataset

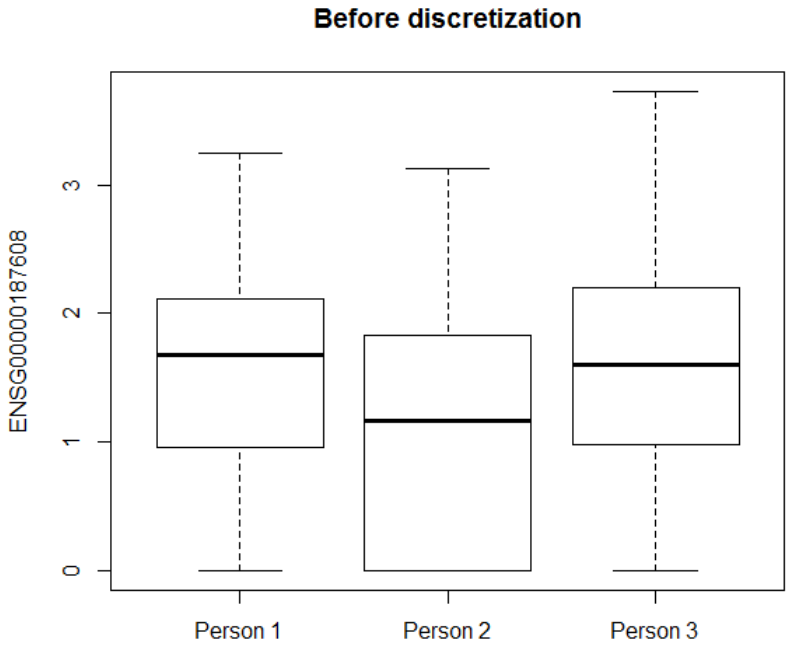
- **Subset of 40 genes**
- **Configurations:**
  1. **Person vs Gene (true) and Gene vs Person (false) only for differentially expressed genes.**
  2. **Person vs Gene (true for differentially expressed genes)**

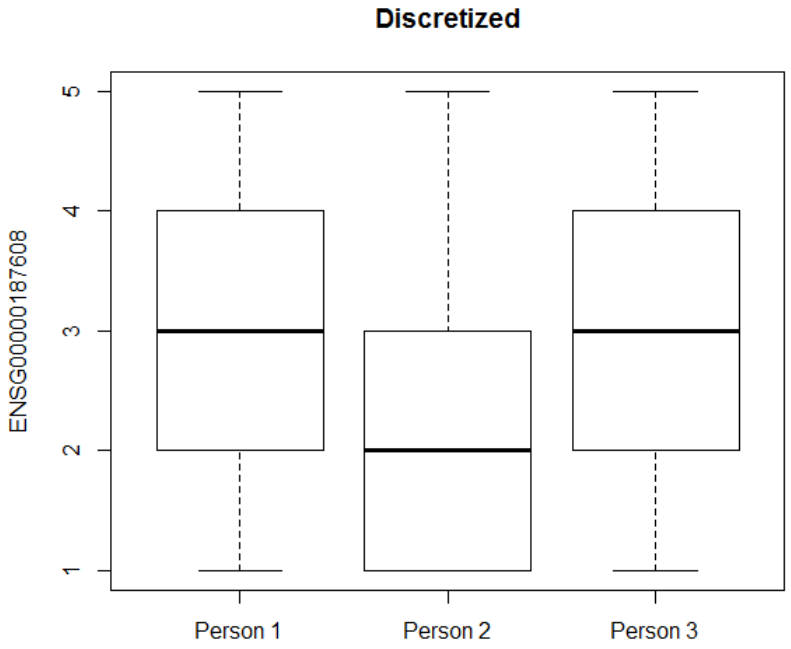




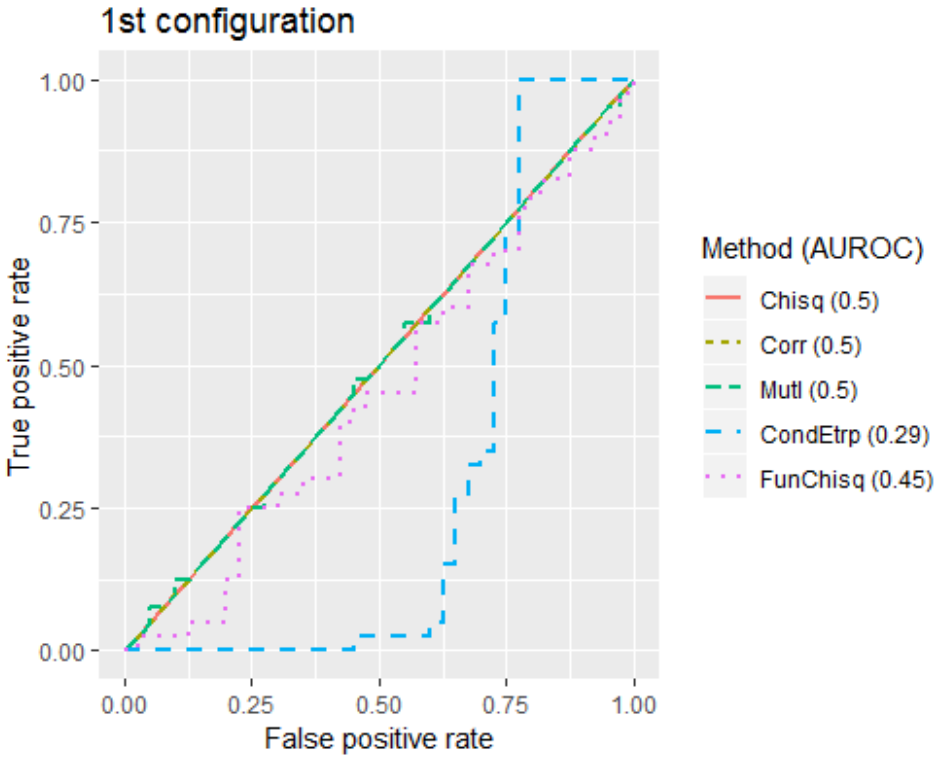
ČVUT
ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

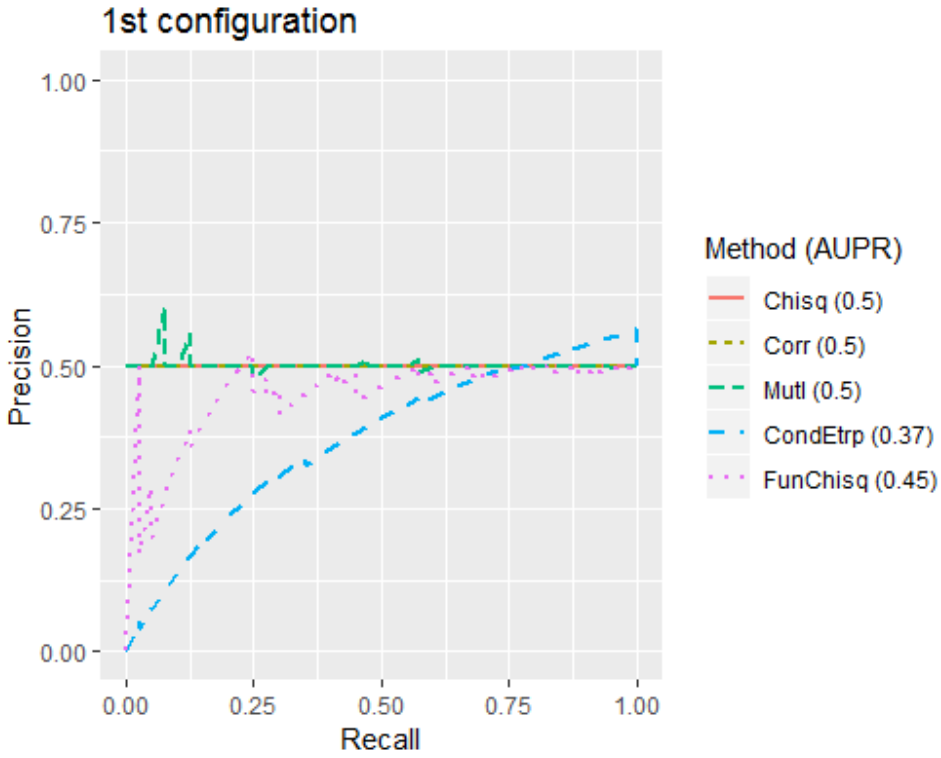
Data discretization



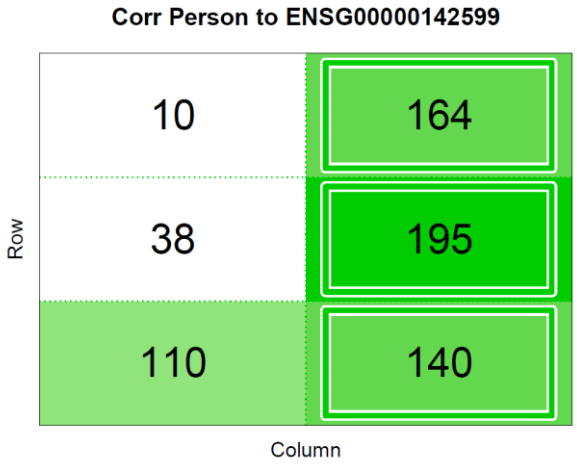


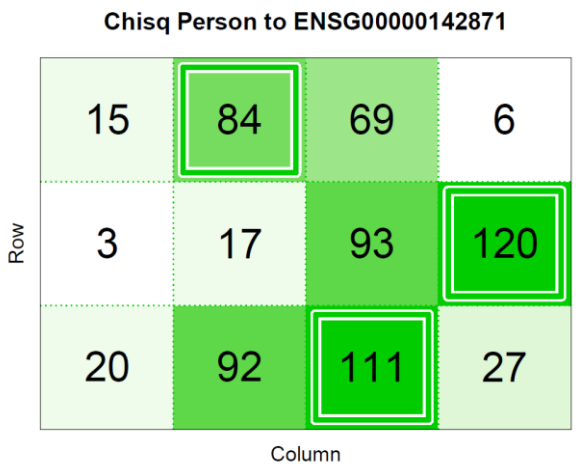
Results 1st configuration

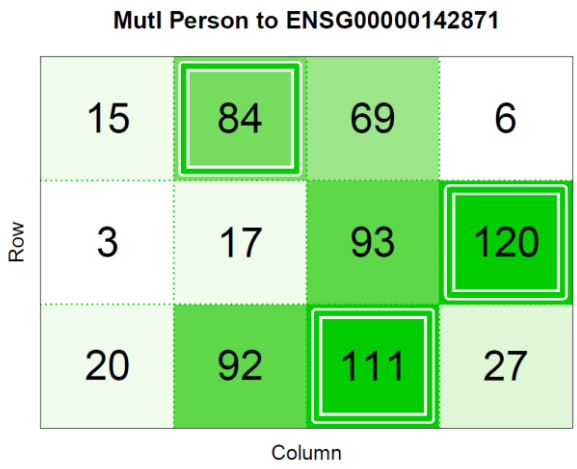




Tables with strongest inference for each method

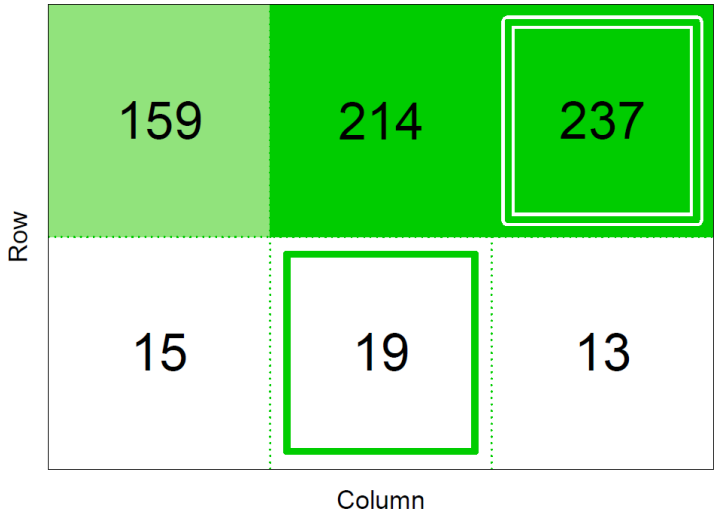




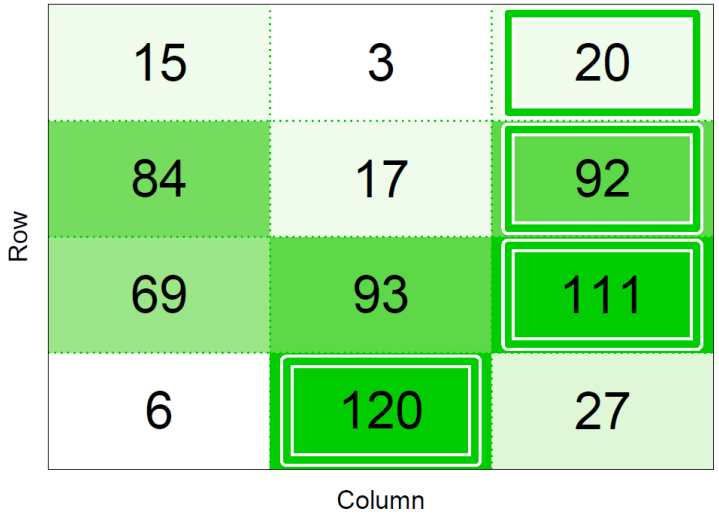


Tables with strongest inference for each method

CondEtrp ENSG00000215910 to Person

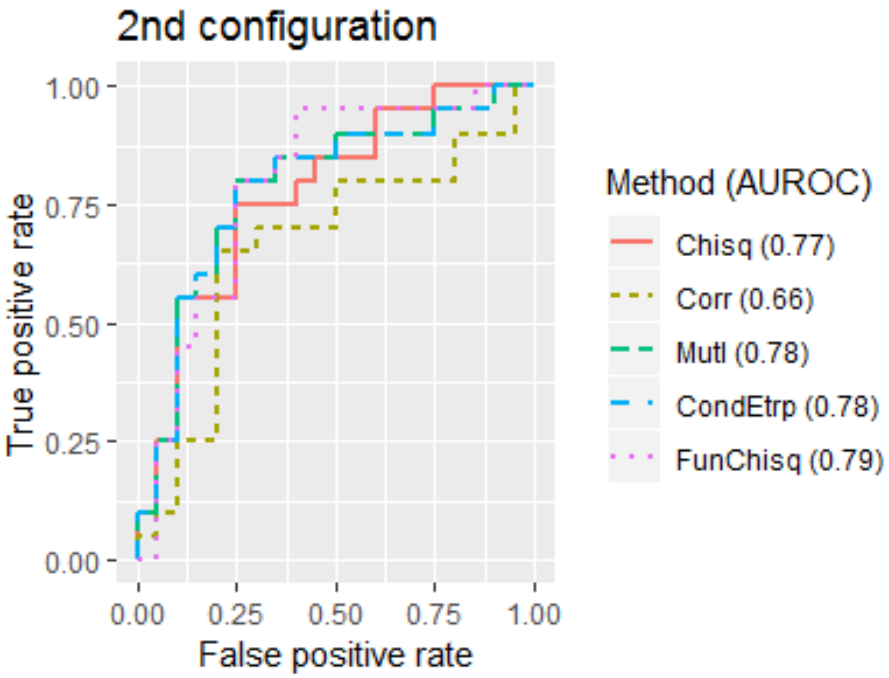


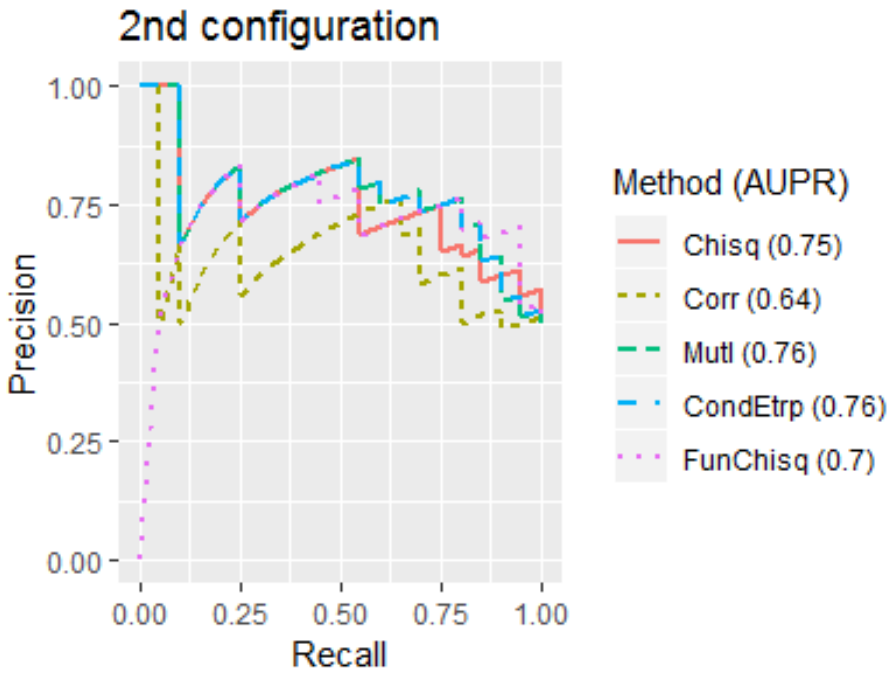
FunChisq ENSG00000142871 to Person



ČVUT
ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

Results 2nd configuration





ČVUT
ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

Conclusion

- **Simulated dataset**
- **Real dataset**
- **Future work**
  - **Answer How many samples are needed to gain at least 70% accuracy?**
  - **Add test for multiple noise**
  - **Process whole real dataset**

ČVUT
ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

References

[1] Zhang, Y., & Song, M. (2013). Deciphering interactions in causal networks without parametric assumptions. *arXiv preprint arXiv:1311.2707*.

[2] Tung, P.Y., Blischak, J. D., Hsiao, C. J., Knowles, D. A., Burnett, J. E., Pritchard, J. K., &Gilad, Y. (2017). Batch effects and the effective design of single cell gene expression studies. Sci Rep, 7, 39921.