# Predicting the effects of gene knockouts from observational data

Danny Hendrix
Harm Berntsen

Supervisor Joris Mooij

Radboud University Nijmegen

# Introduction

- Cures for diseases often target particular genes
- Unexpected effect caused by gene relations
- Knockouts are difficult and time consuming

# Method

- Measured yeast genes: ±5000 genes from ±60 samples (yeast cells)

  Hughes, T. R., Marton, M. J., Jones, A. R., Roberts, C. J., Stoughton, R., Armour, C. D., ... & Friend, S. H. (2000)
  Functional Discovery via a Compendium of Expression Profiles

- Simulated data
  - Variation in amount of samples

- Compare results with ground truth included with simulated data

## Goal

- Predict the effect of knocking out gene x on gene y
- Under certain conditions, if we observe $w \perp\!\!\!\perp y \mid [x]$ then $x \Rightarrow y$
- Derive (conditional) (in-)dependencies from the data

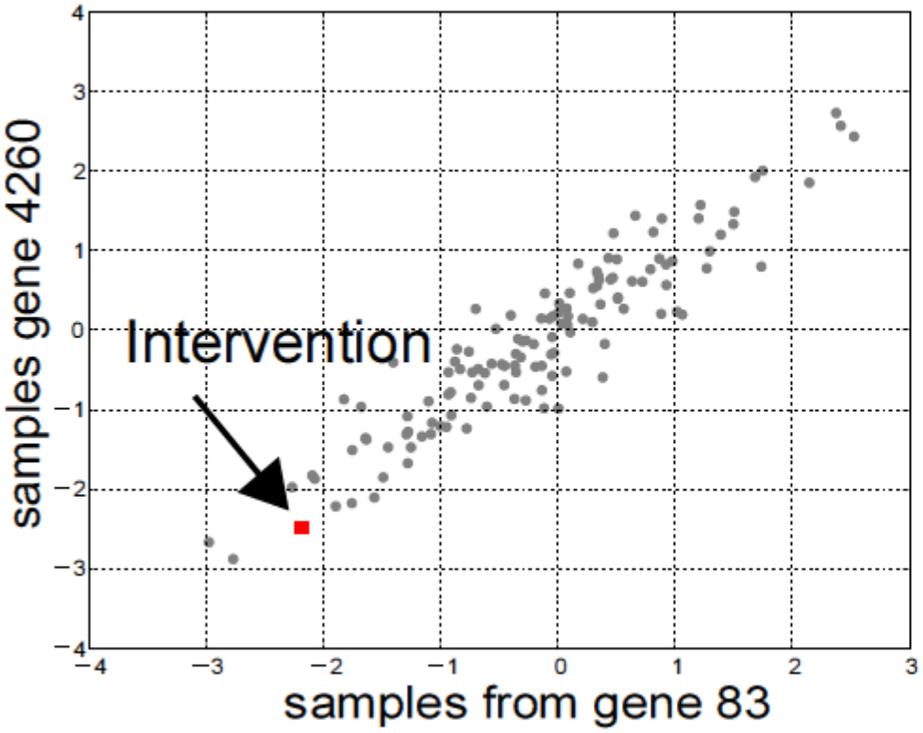
-Doris Entner, Patrik O. Hoyer, Peter Spirtes
Data-driven covariate selection for nonparametric estimation of causal effects

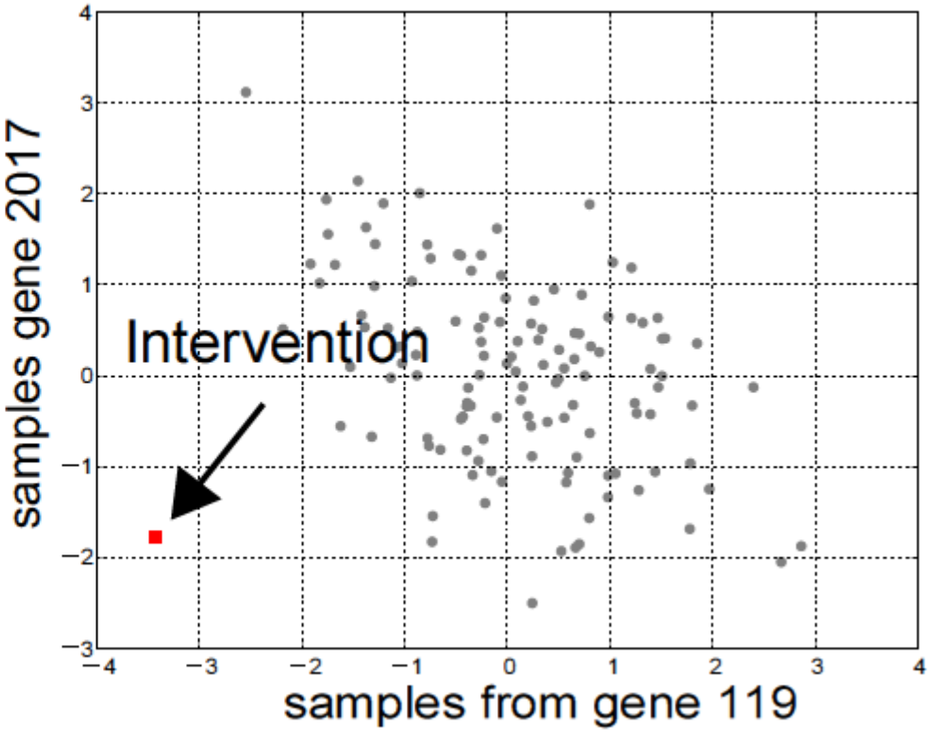
-Tom Claassen and Tom Heskes
A Logical Characterization of Constraint-Based Causal Discovery

# Software implementation

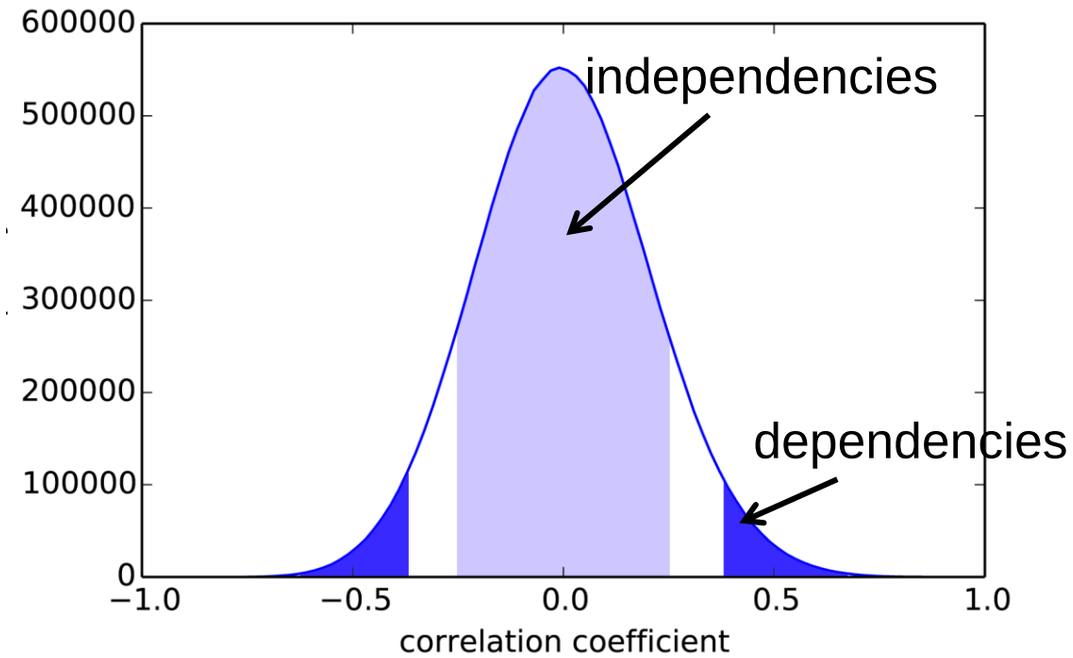
- Python calls the C++ code
  - Splits computation in steps and saves intermediate values
  - Easy generation of graphs
- C++ Python module
  - C++AMP (GPU)
  - OpenMP (Parallelism on CPU)
  - Boost
  - Eigen (Fast matrix)

# Interventions



# Defining thresholds

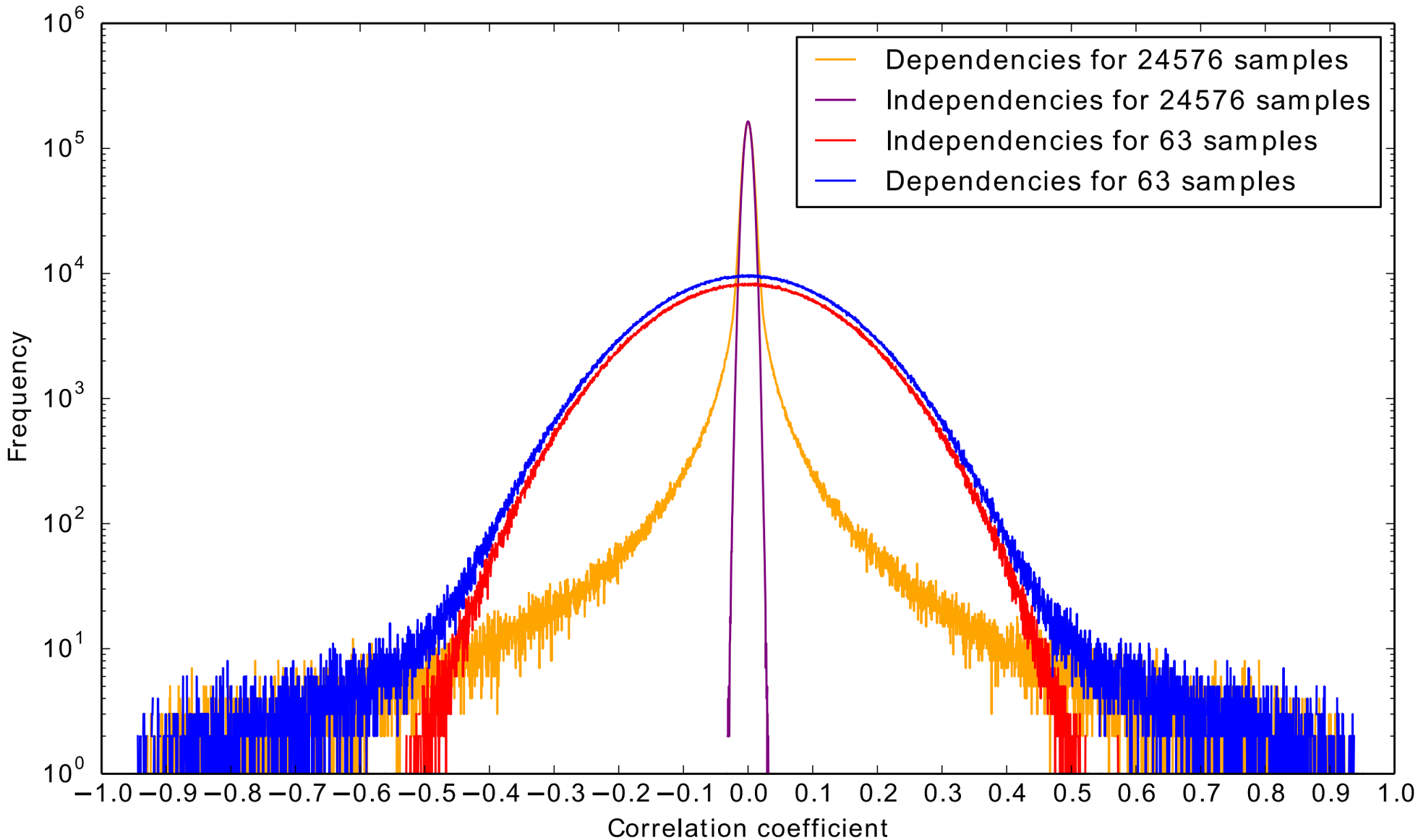



- Find a threshold combination that works in simulated data
  - Lowest amount of wrong predicted (in-)dependencies
  - Highest amount of correct predicted dependencies

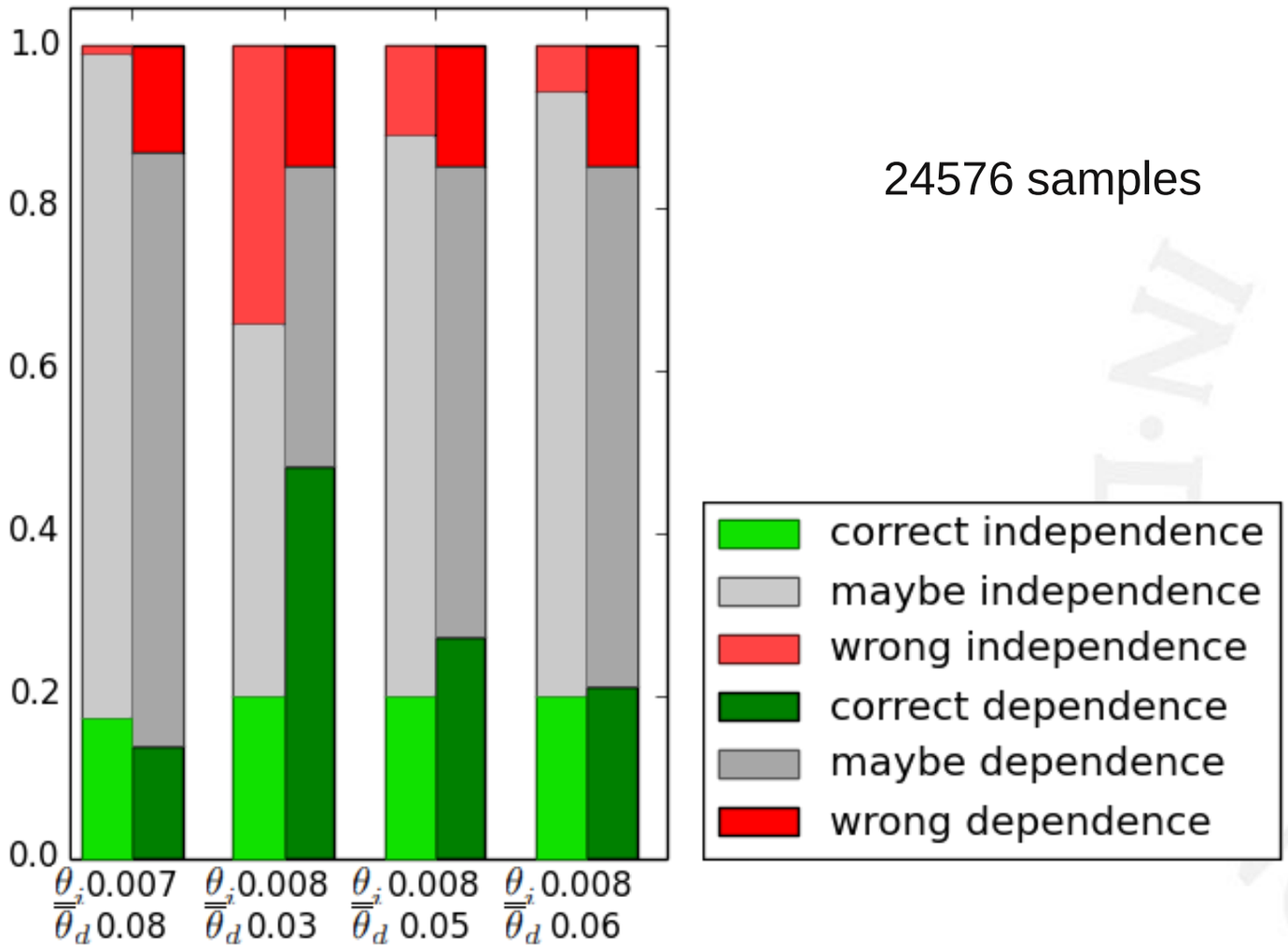
# Defining thresholds

- Analysed individual parts of algorithm
  - Correlation coefficient
  - Conditional (in-)dependencies

# Pearson correlation coefficient

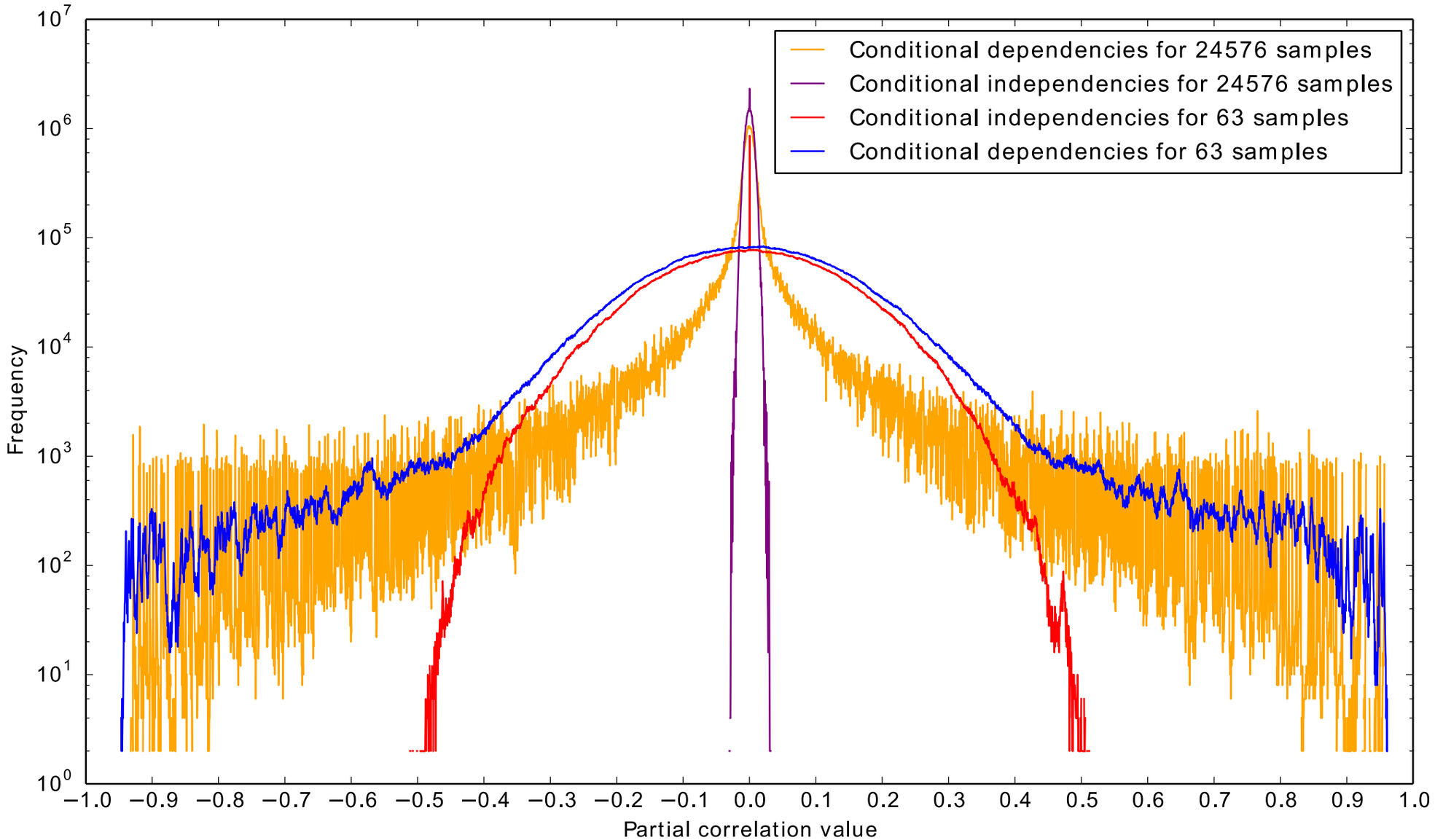


# Pearson correlation coefficient




24576 samples

# Conditional (in-)dependencies



# Defining threshold

Conditional (in-)dependencies

- 30% correct and 70% wrong dependencies
- 90% correct and 0% wrong independencies
- Limited to first 512 genes

Full algorithm

- 56% correct predictions

# Conclusion

- Unable to distinguish dependencies in conditional (in-)dependencies
- Increasing the amount of samples gives a slightly better prediction
- Unknown if the simulation did not work, possibly a different kind of simulation does work
- Unknown whether this holds for actual data

# Questions