Graph-based data merging and high dimensional reduction, clustering model

Duncan Murray
Wayne Croft
Gokhan Tut
Dave Bone

Introduction

Introduction

Challenge: Platform for discovery and analysis of correlates of protection

Introduction

Challenge: Platform for discovery and analysis of correlates of protection

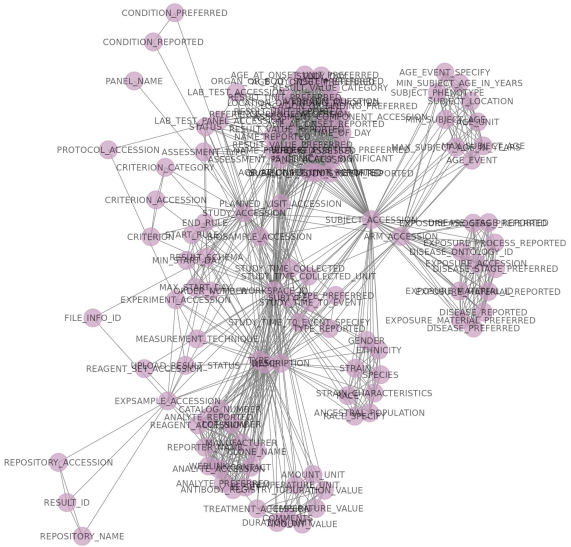
Why?

Subchallenge 1

Tools to help with understanding and preparing data for analysis

Subchallenge 1

Tools to help with understanding and preparing data for analysis

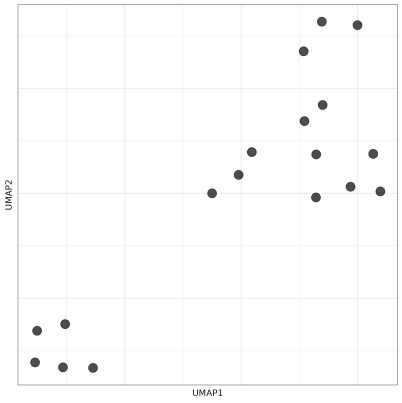


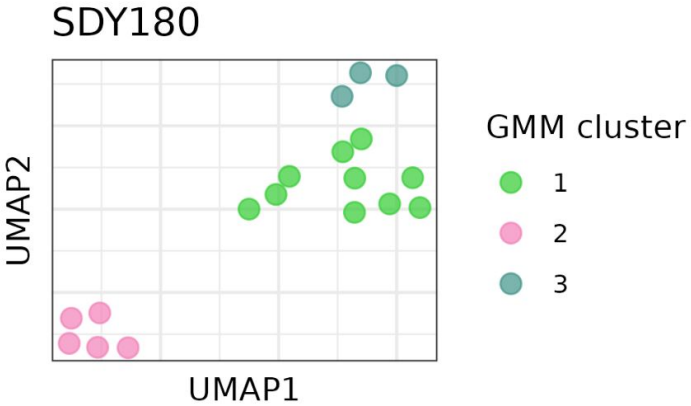
Subchallenge 2

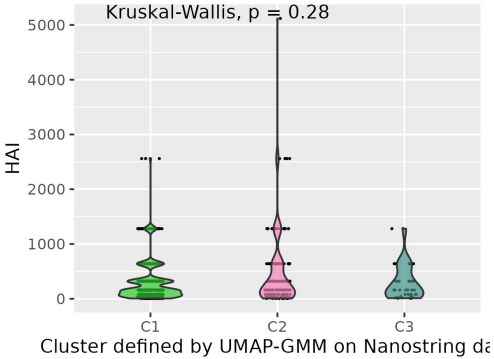
Identification of at-risk populations that is transferable across studies

Subchallenge 2

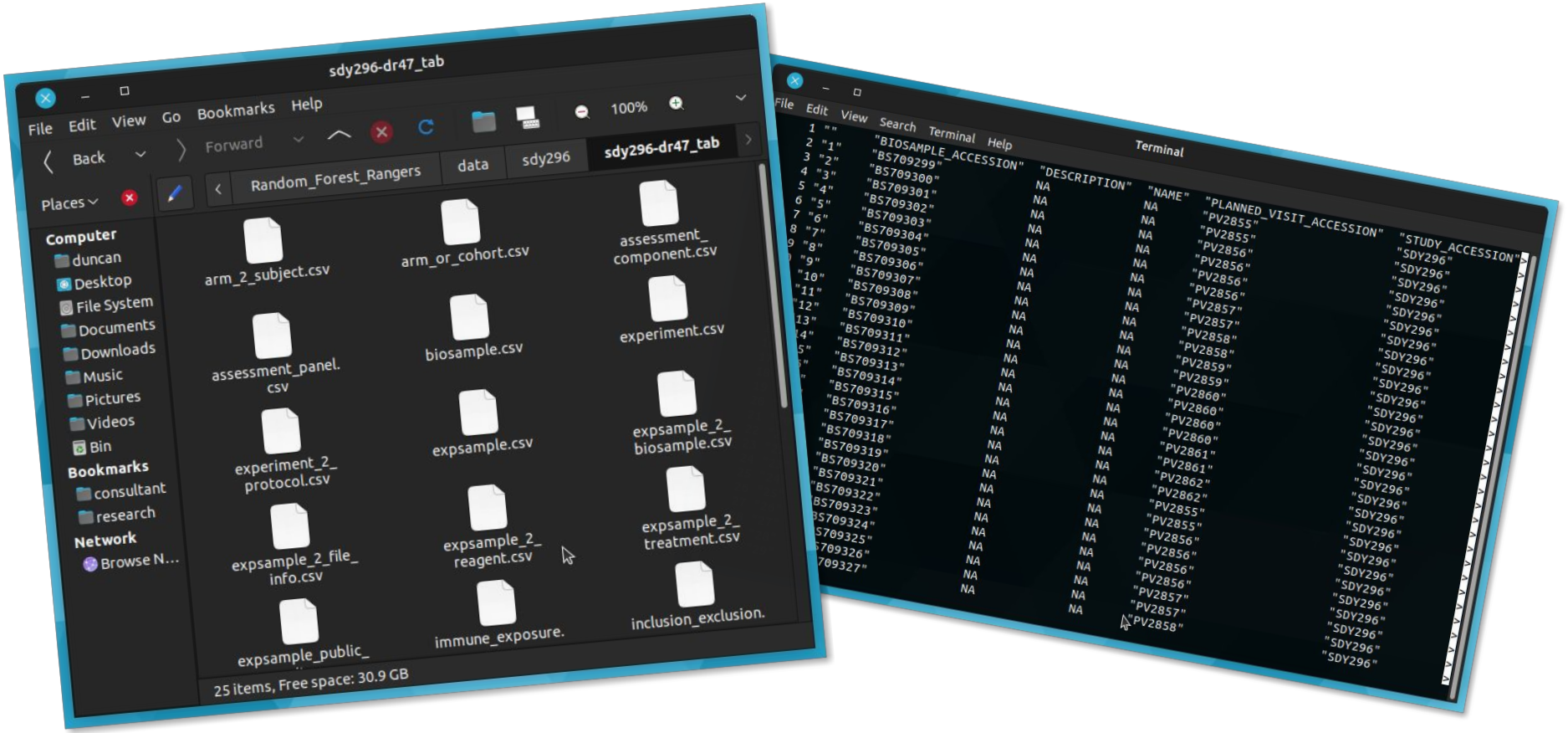
Identification of at-risk populations that is transferable across studies



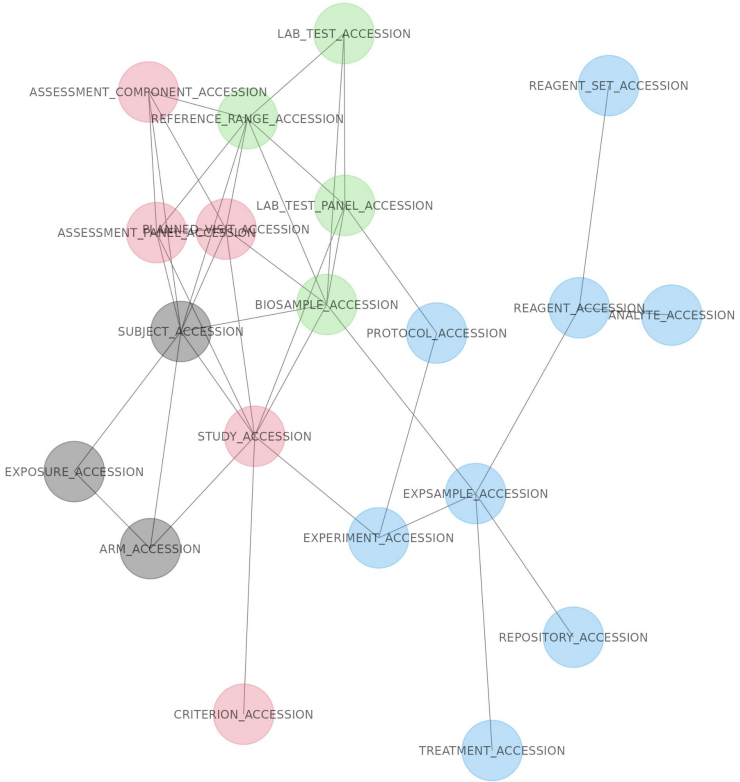




Graph-based data structure tool



Graph-based data structure tool

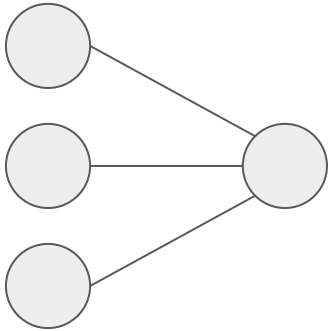


Graph-based data structure tool

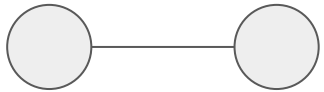
run code

Graph-based data structure tool

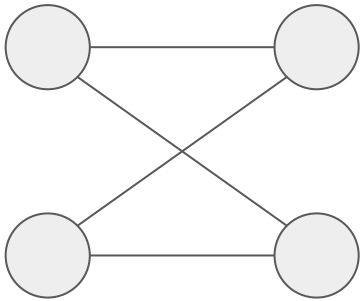
analysing the relationship between data fields



many-to-one



one-to-one



many-to-many

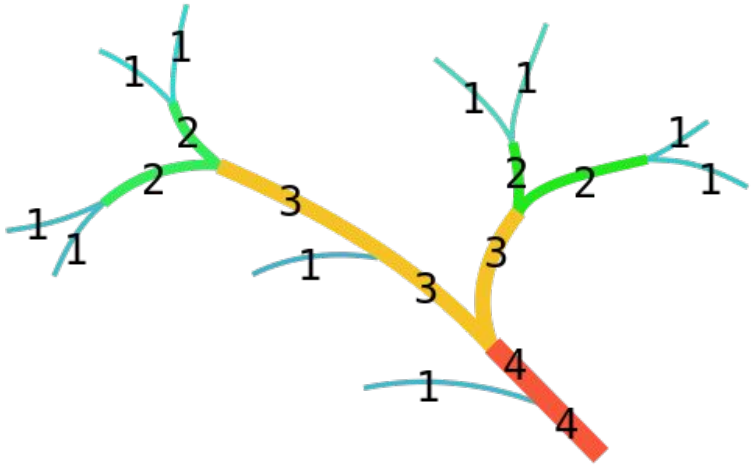
Graph-based data structure tool

directed graph

(code)

Graph-based data structure tool

iterative, random walks to score nodes



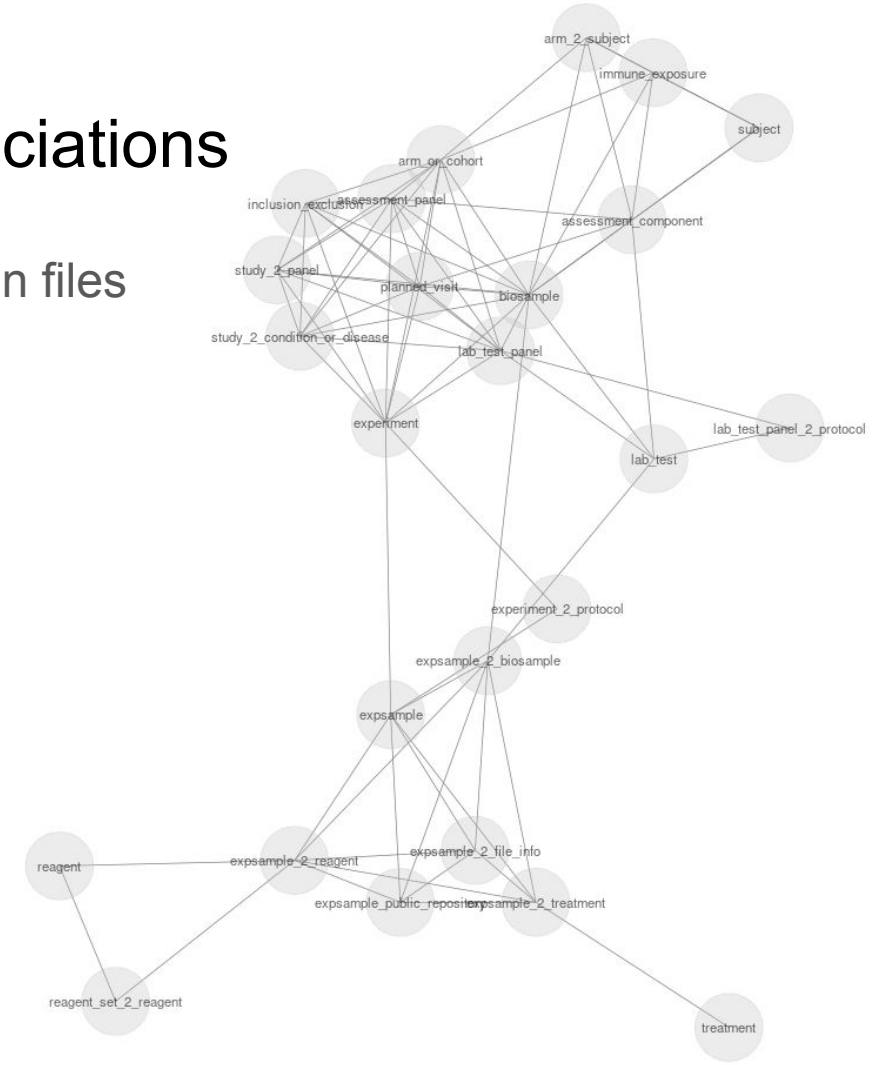
Graph-based data structure tool

directed graph with scores

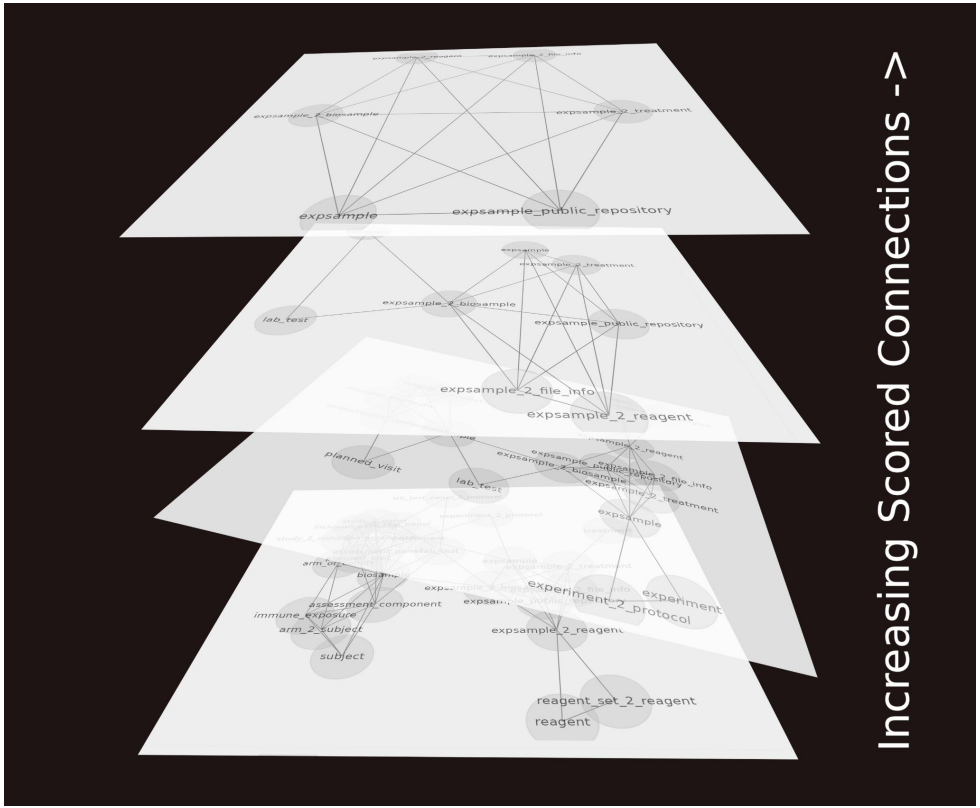
(code)

Graph based file associations

elucidates connections between files



Graph based file associations

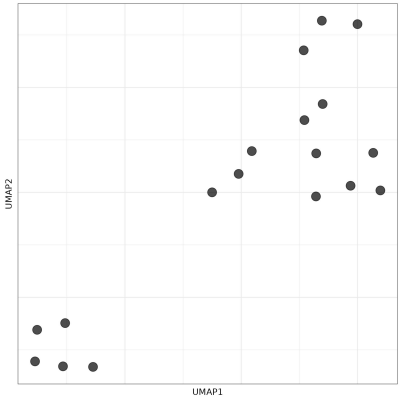


Graph-based data structure tool

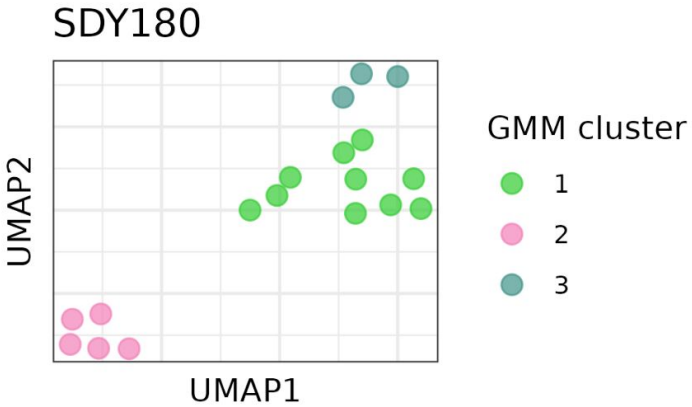
example data preparation

(code)

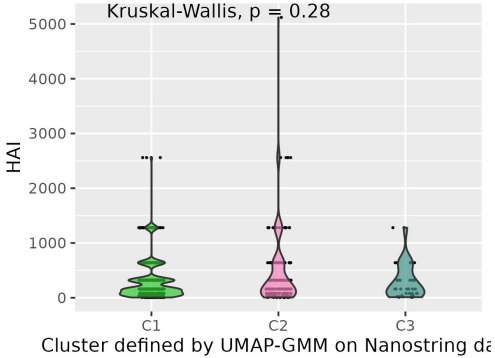
Identifying populations at risk



UMAP dimension reduction

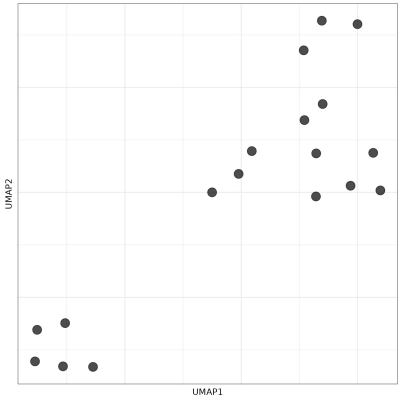


GMM clustering

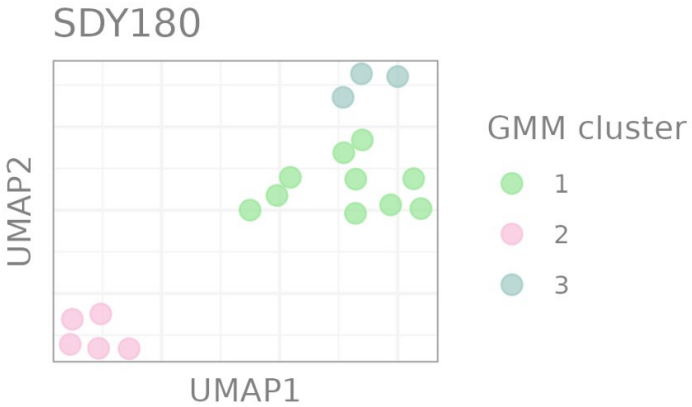


Cluster identification

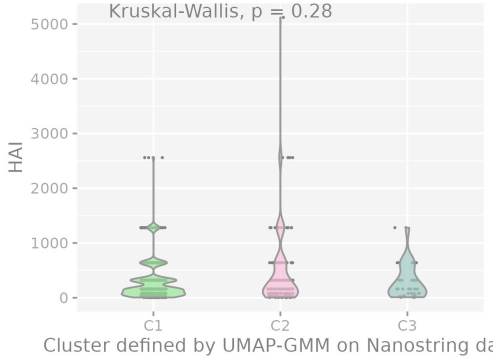
Identifying populations at risk



UMAP dimension reduction

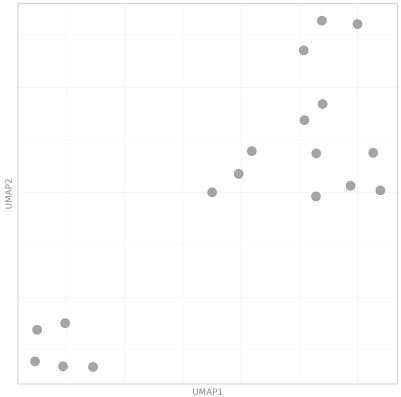


GMM clustering

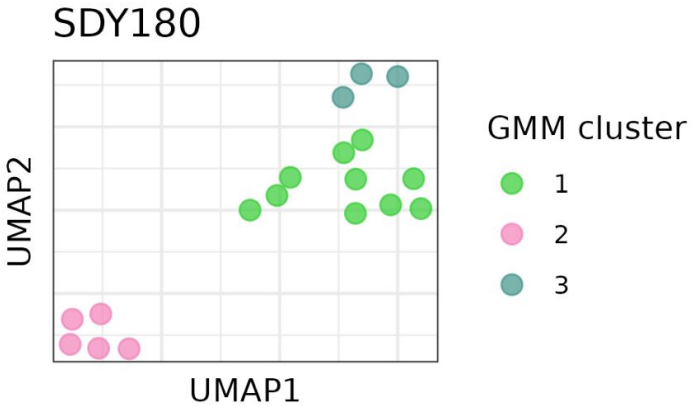


Cluster identification

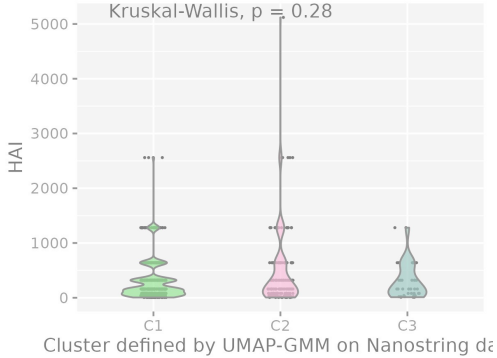
Identifying populations at risk



UMAP dimension reduction

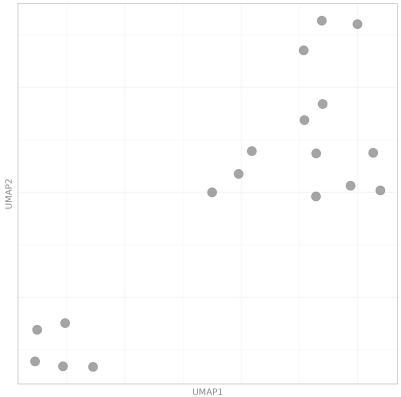


GMM clustering

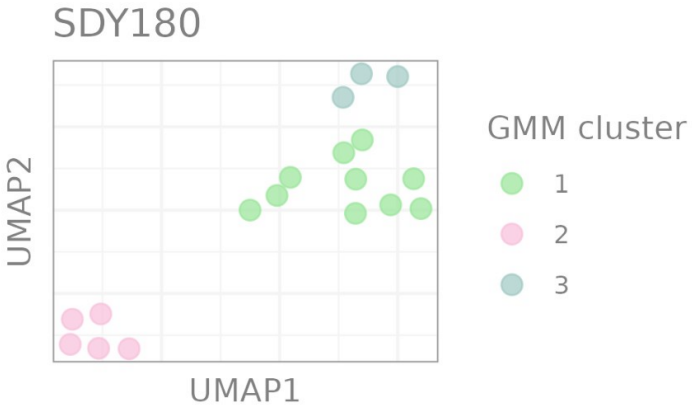


Cluster identification

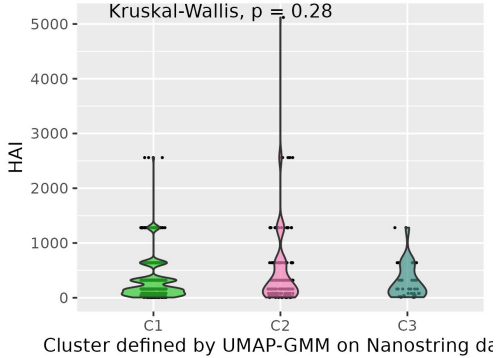
Identifying populations at risk



UMAP dimension reduction

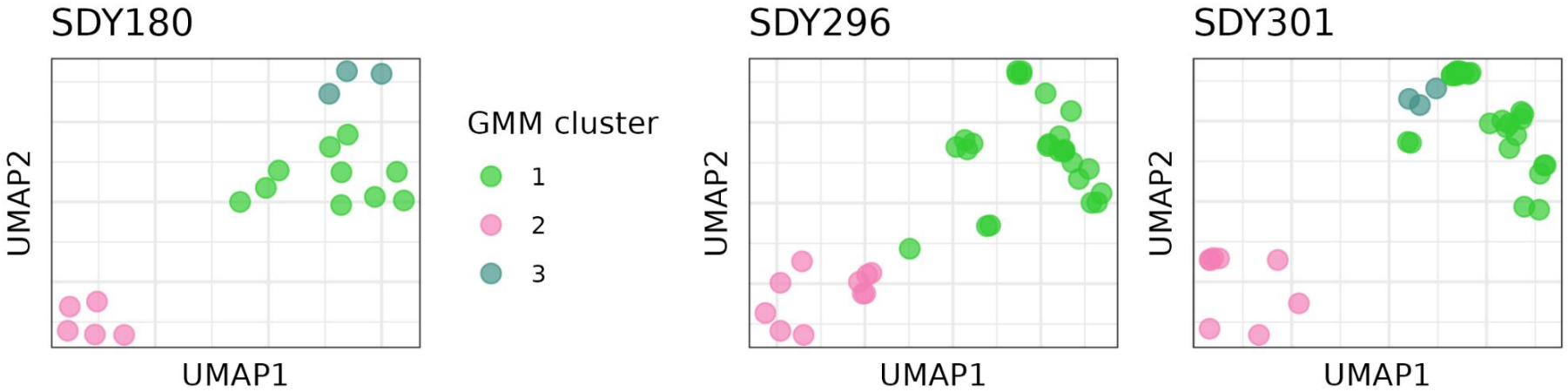


GMM clustering



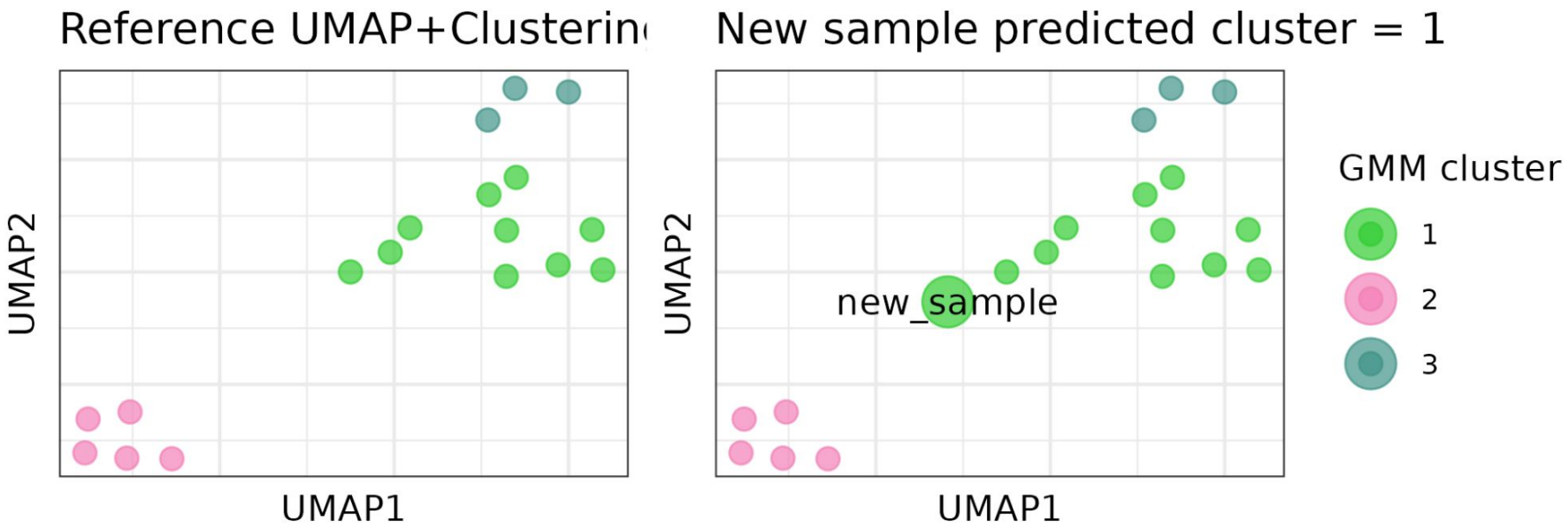
Cluster identification

Identifying populations at risk



validate on other cohorts / disease/ vaccines

Identifying populations at risk



Ensuring Impact

Engagement with biostatistician early on

Using non-traditional modelling techniques, but validating using tried and tested traditional methods

Ensuring Impact

Engagement with biostatistician early on

Using non-traditional modelling techniques, but validating using tried and tested traditional methods

Engagement with policymakers

Plan to integrate modelling as part of pandemic planning

Public buy in

Involve lay group regularly throughout development

Seek advice from lay stakeholders to understand how outcomes from our models will influence policy and by communicated with public

Conclusions

Conclusions

Novel approach to analysing and preparing complex data

Conclusions

Novel approach to analysing and preparing complex data

Model which can handle high dimensional data with different data types

Conclusions

Novel approach to analysing and preparing complex data

Model which can handle high dimensional data with different data types

Scalable, repeatable models which can non-linearly identify specific patient subsets

Future Work

Graph-based data structure analysis has potential to save significant time

Plan to develop machine learning feedback to improve model by adjust tunable parameters in dimension reduction and clustering