Cake Talk
Subphenotyping in TCGA data

James McMurray

PhD Student
Department of Empirical Inference

22/10/2014

Outline of talk

- Background
  - TCGA Project
  - Subphenotyping
  - General idea
- Example study
  - Overview
  - Data
- Replication
  - Overview
  - Analysis of larger dataset
- Future Work
  - Deep models
- Conclusion

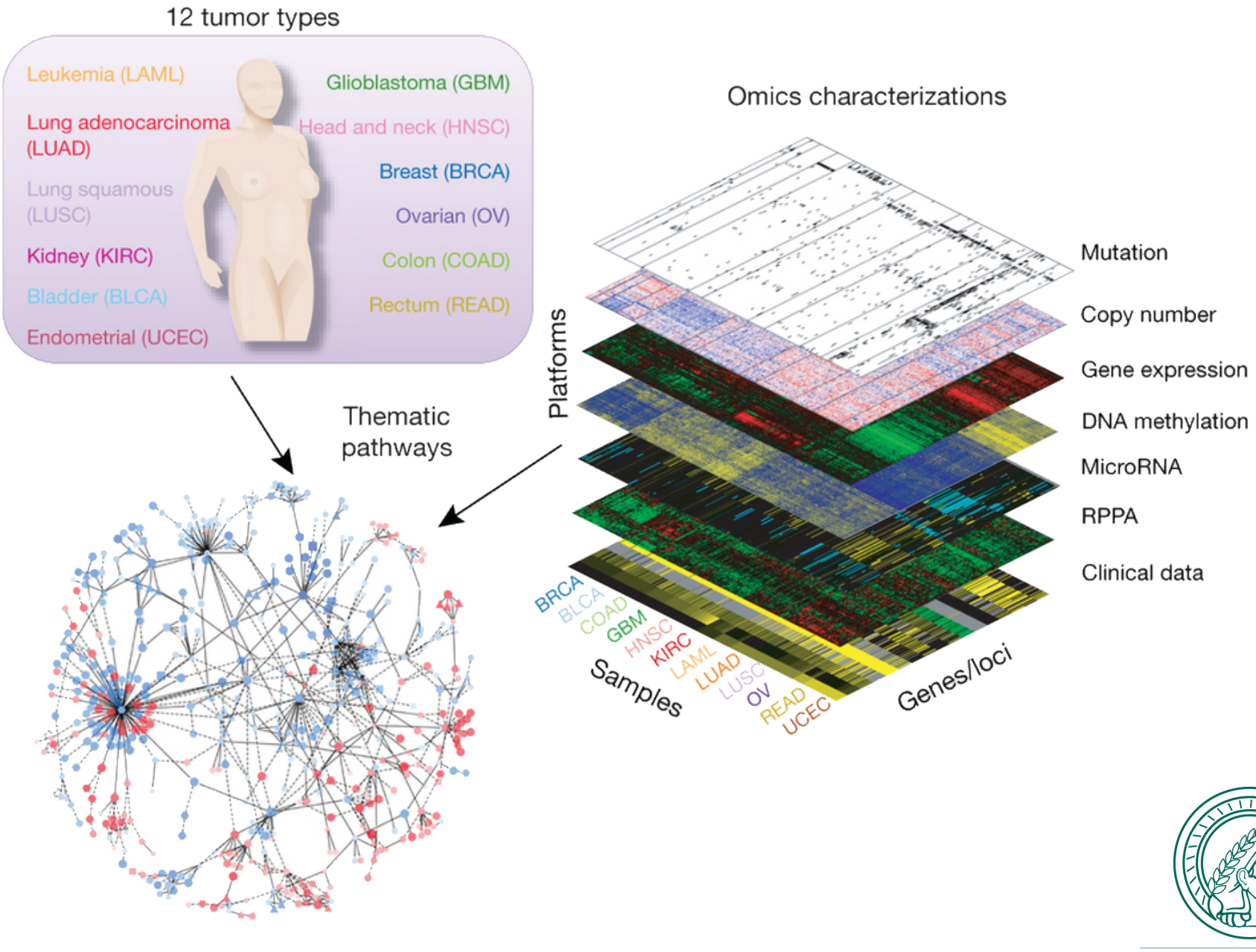
MAX-PLANCK-GESELLSCHAFT

TCGA Project

- The Cancer Genome Atlas
- Public multi-omics data:
  - SNPs (restricted)
  - Gene Expression arrays
  - RNASeq
  - Copy Number Variation
  - DNA Methylation
  - miRNASeq
  - Proteomics
- Many different types of cancer including GBM (brain), BRCA (breast cancer), KIRC/KIRP (kidney cancer), etc.
- Aim to find links between various types of cancer
- Improve understanding of molecular basis

MAX-PLANCK-GESELLSCHAFT

TCGA Overview



MAX-PLANCK-GESELLSCHAFT

What is subphenotyping?

- Identify sub-types to broad phenotypes - group patients by these
- Clustering of patients - population structure
- Sub-disease classification
- Helps to provide intuition about molecular basis
- Diagnostic biomarkers
- Provide specific candidate drug targets
- Improve precision of medicine
- Unsupervised Learning

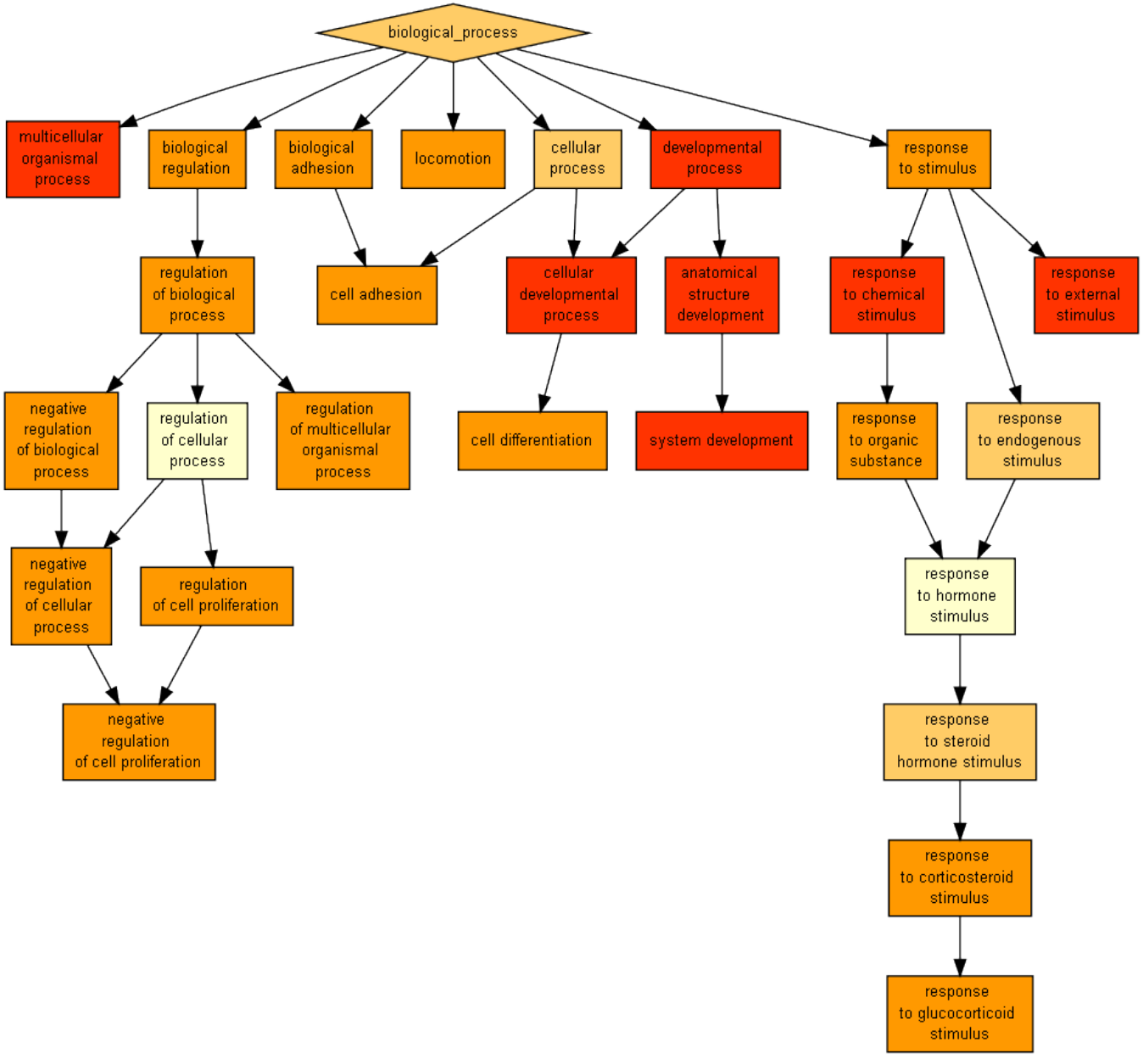
MAX-PLANCK-GESELLSCHAFT

General idea

1. Cluster tumour samples based on some biomarkers (e.g. variations in gene expression)
2. Find the most significant differences between clusters (i.e. in gene expression) and if the clusters correspond to clinical differences (i.e. in survival time)
3. Carry out a Gene Ontology Enrichment analysis (i.e. find if certain functional classes of genes are over-expressed or under-expressed in the clusters)
4. If so, investigate possible causal pathways and identify drug targets (i.e. genes which might have an effect if knocked-out in the tumour)

MAX-PLANCK-GESELLSCHAFT

GO Example



MAX-PLANCK-GESELLSCHAFT

Outline of talk

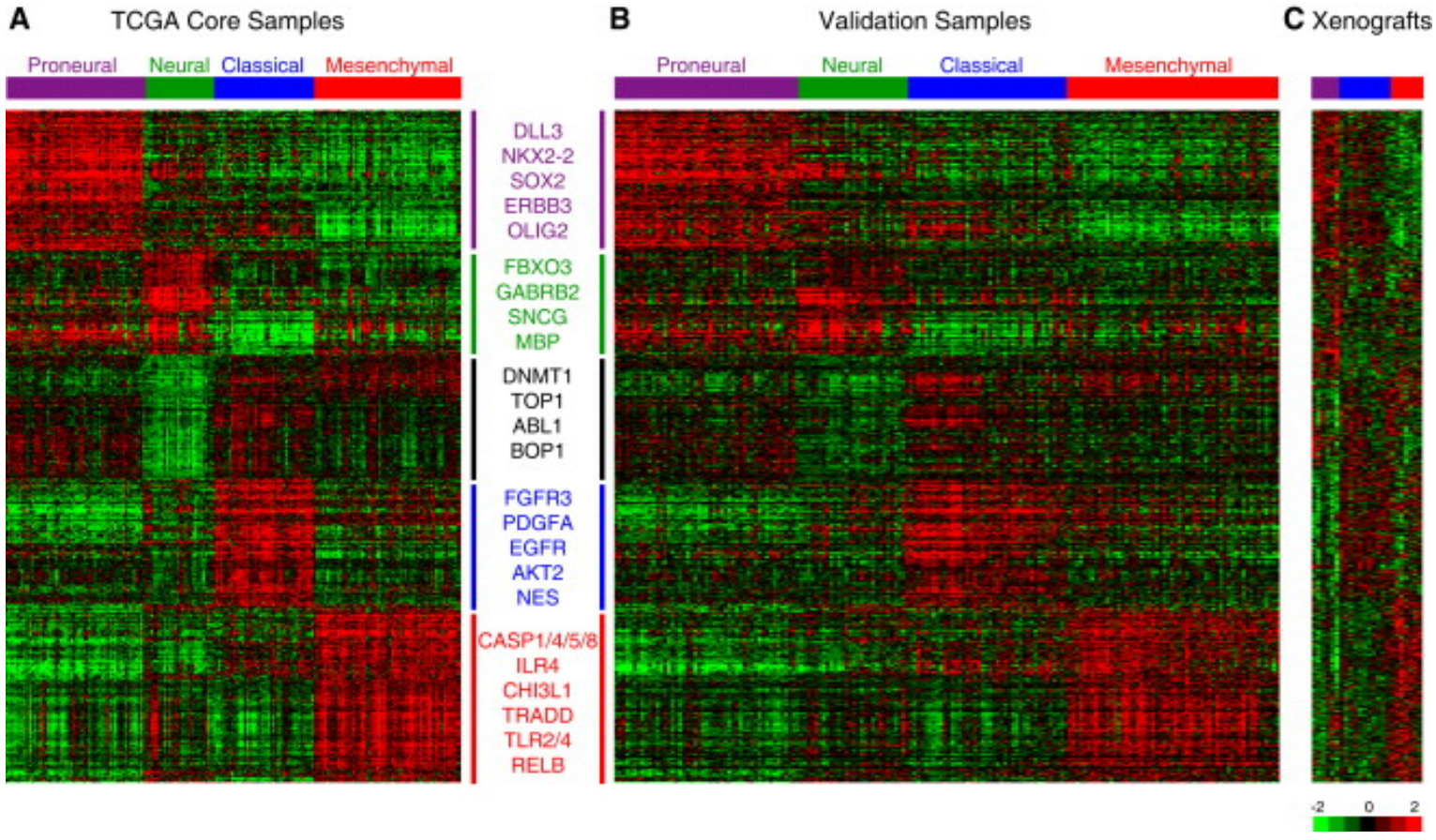
MAX-PLANCK-GESELLSCHAFT

Example study

- Verhaak, R.G., et al. (2010) *Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1*. Cancer Cell. 17(1):98-110
- Previously identified four sub-types of GBM (GlioBlastoma Multiforme) using factor analysis and consensus clustering
  - Proneural
  - Neural
  - Classical
  - Mesenchymal
- Most significant genes were PDGFRA, IDH1, EGFR, and NF1.
- Glioblastoma multiforme (GBM) is the most common form of malignant brain cancer in adults
- Affected patients have a poor prognosis with a median survival of one year

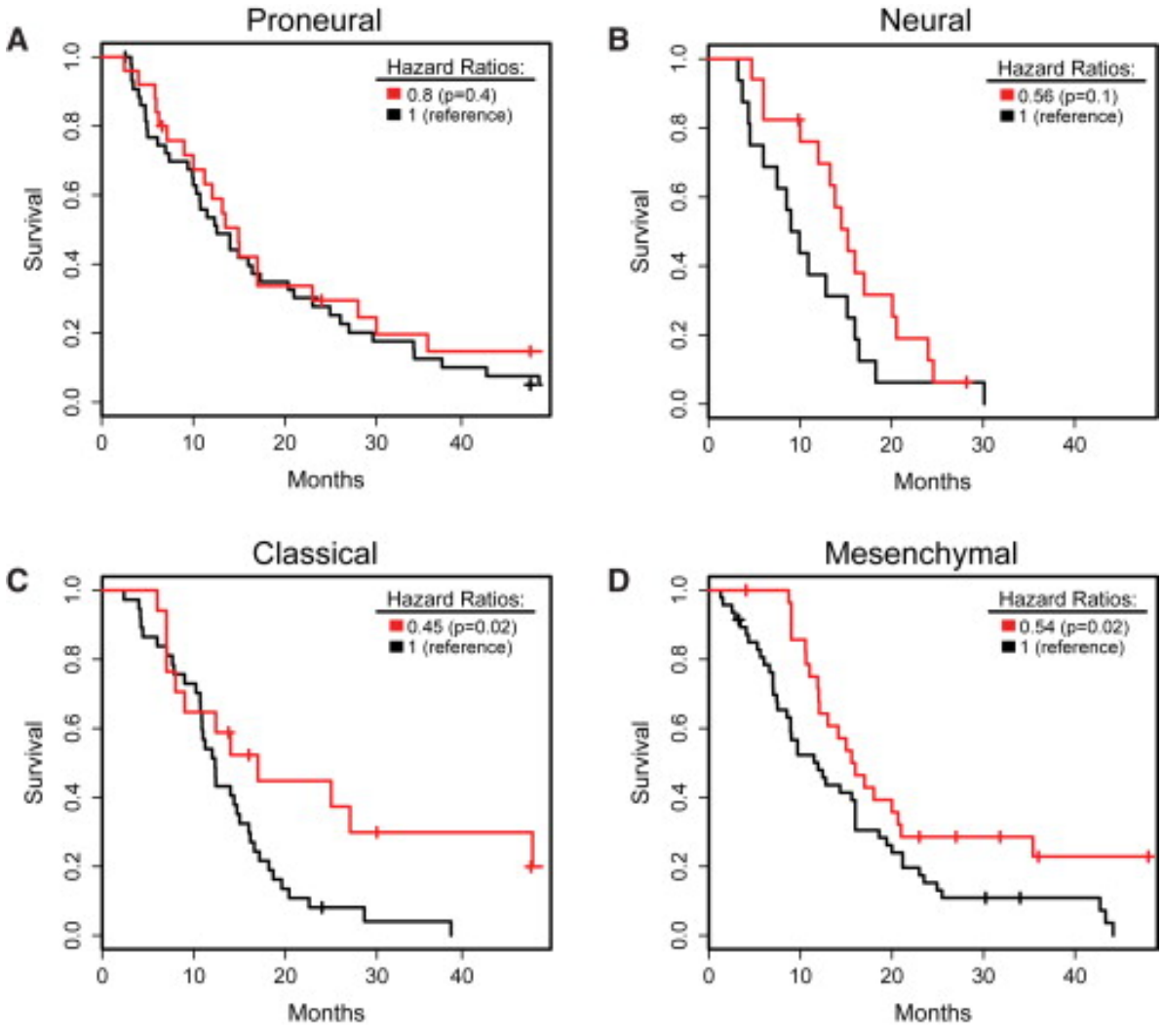
Gene Expression differences

- Gene expression differences:



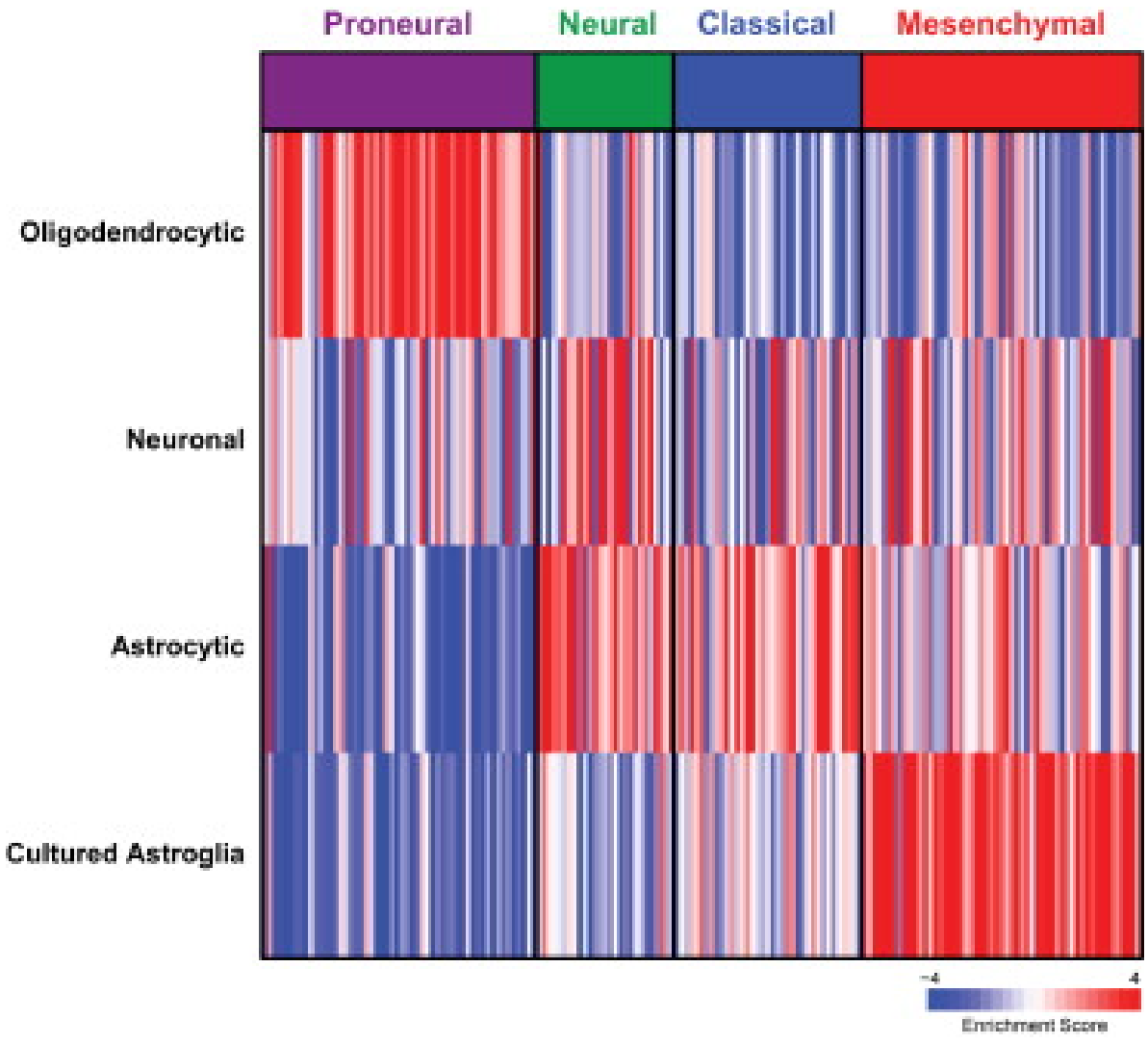
Clinical differences

- Clinical differences:



Gene Ontology Enrichment

- Gene Ontology (GO) Enrichment:



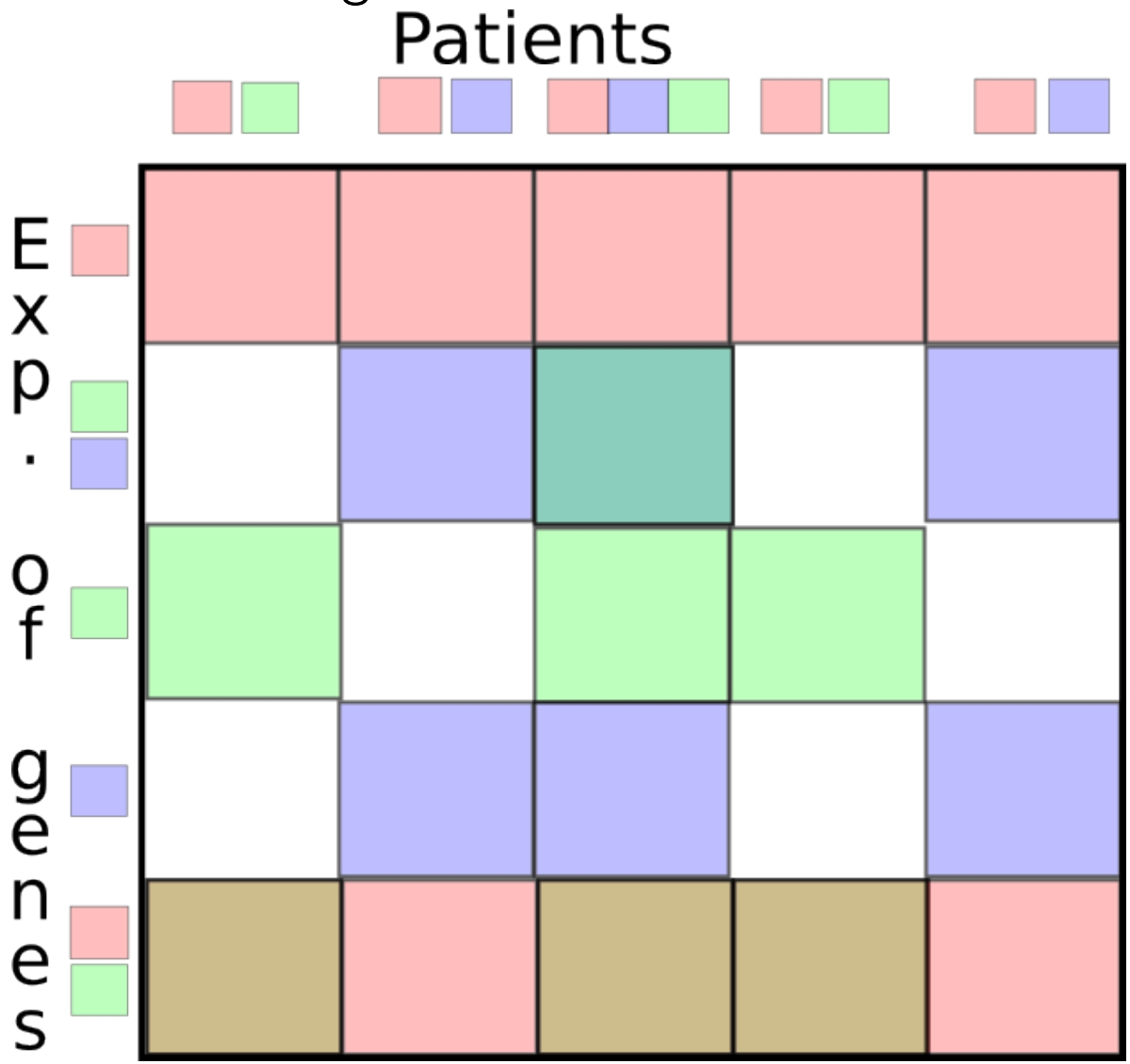
Data

- Patients with GBM cancer
- 202 samples with three gene expression measurements of 11,861 genes.
- Note we could also include RNASeq which is another measure of Gene Expression
- Neglected due to the size of the data and the available samples
- Note that not all expression arrays measure the same genes so there is some missing data
- If we wanted to use more samples we need to deal with missing gene expression measurements across samples too

MAX-PLANCK-GESELLSCHAFT

Data

- Lots of missing data:



Outline of talk

MAX-PLANCK-GESELLSCHAFT

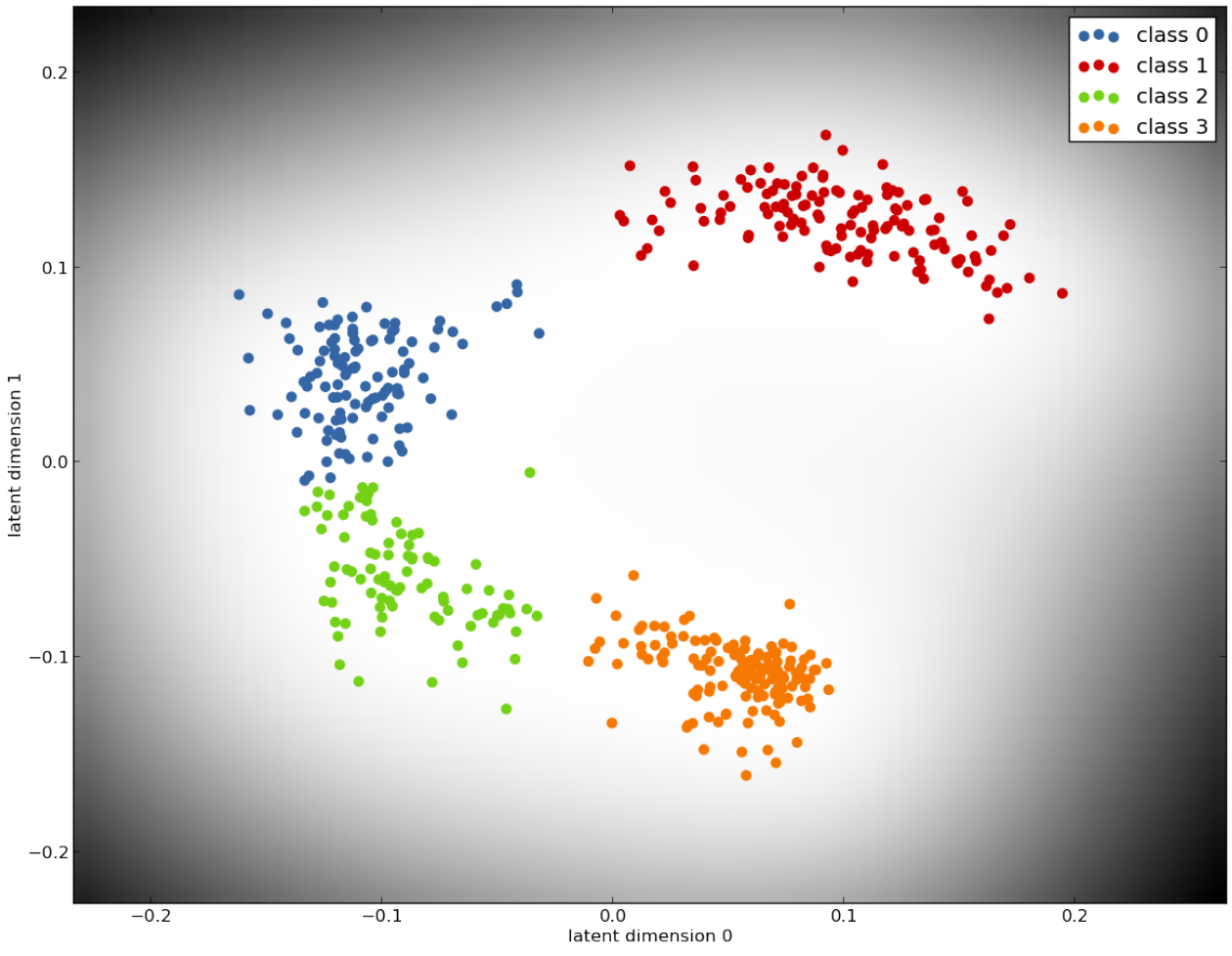
Overview

- Wanted to replicate the study using other dimensionality reduction and clustering methods to test robustness.
- Used other TCGA GBM samples, and the data of the aforementioned paper.
- Other samples: 473 samples of 17,430 genes
- Verhaak, et al.: 202 samples of 11,861 genes.
- Used GPLVM for dimensionality reduction then k-means for clustering.

MAX-PLANCK-GESELLSCHAFT

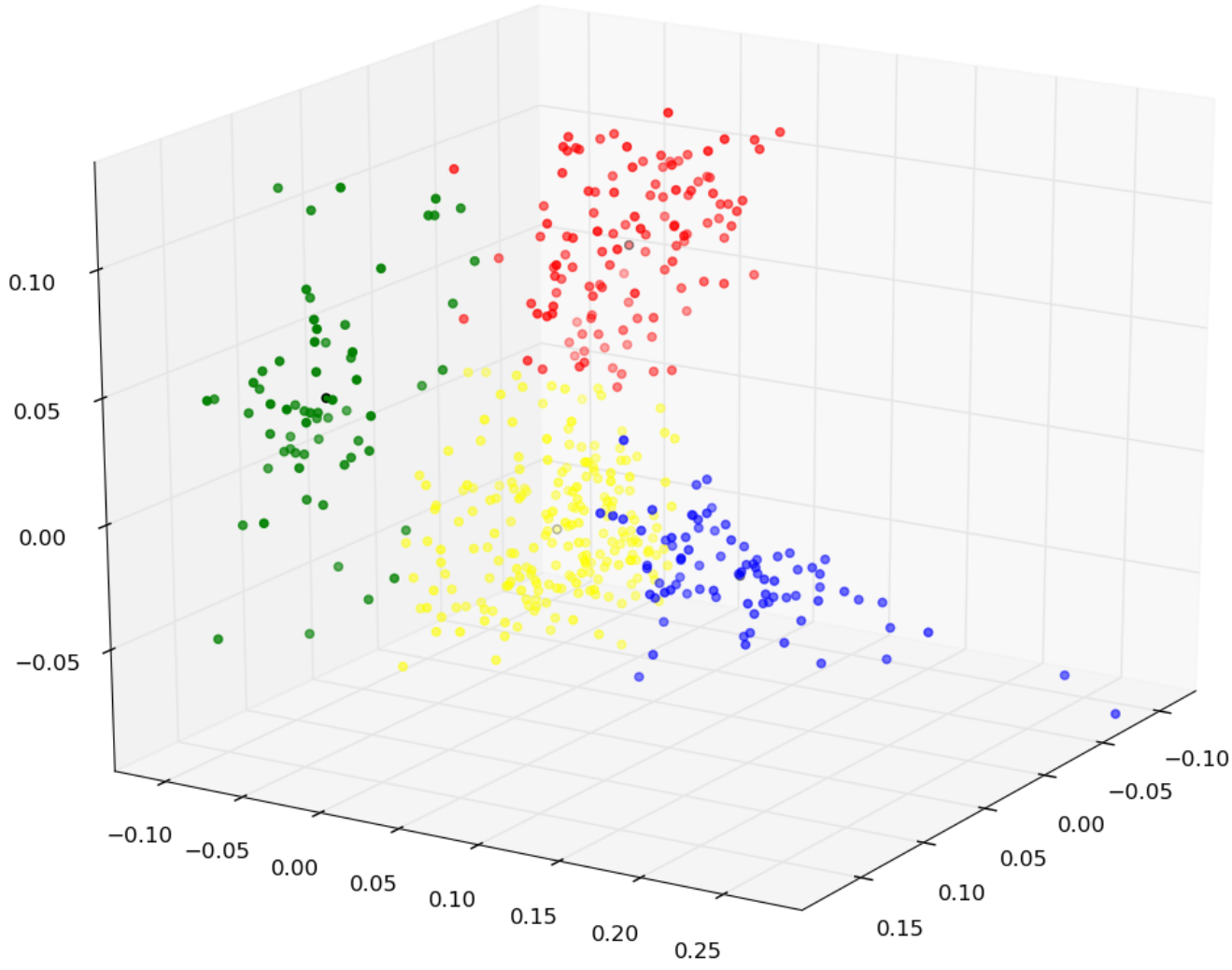
Clustering with GPLVM: 2D

- Larger dataset clustered with k-means on 2d latent space



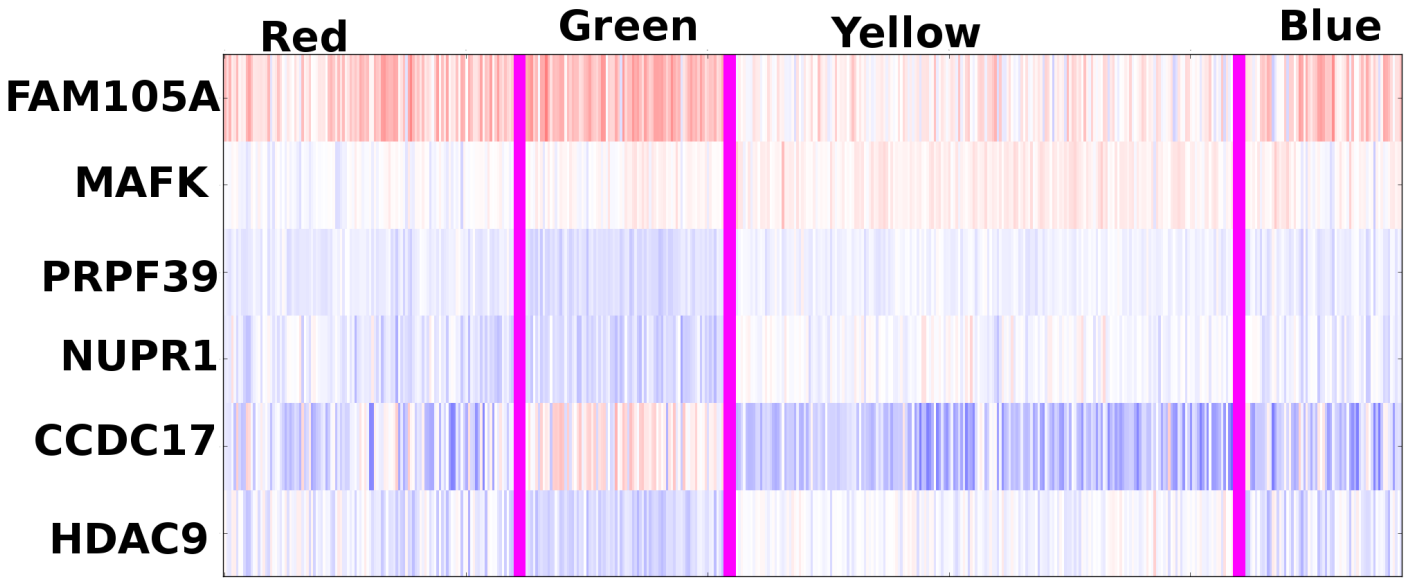
MAX-PLANCK-GESELLSCHAFT

Clustering with GPLVM: 3D



MAX-PLANCK-GESELLSCHAFT

Most significantly different genes



- The expression of the genes (rows) across the GBM samples (columns). The magenta lines delineate the clusters.
- Note different genes to Verhaak, et al.

MAX-PLANCK-GESELLSCHAFT

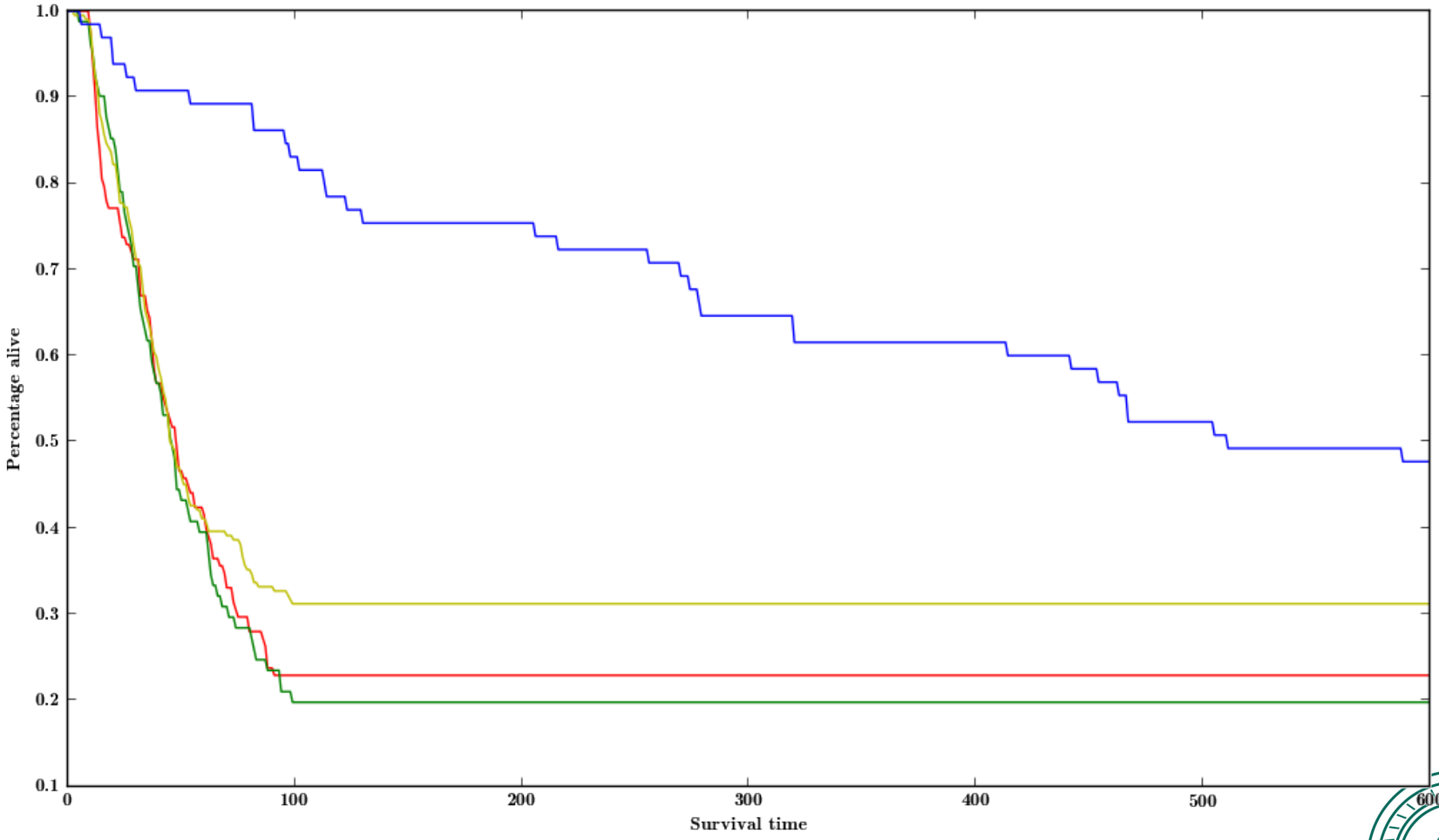
Clinical differences

Cluster	Total dead	Mean survival time of the dead (days)
Red	91/120 (75.8%)	40.2
Green	65/82 (79.3%)	41.8
Yellow	139/206 (67.5%)	37.7
Blue	47/65 (72.3%)	**528.4**

- The mean survival time of those who died demonstrates clinical differences between the clusters

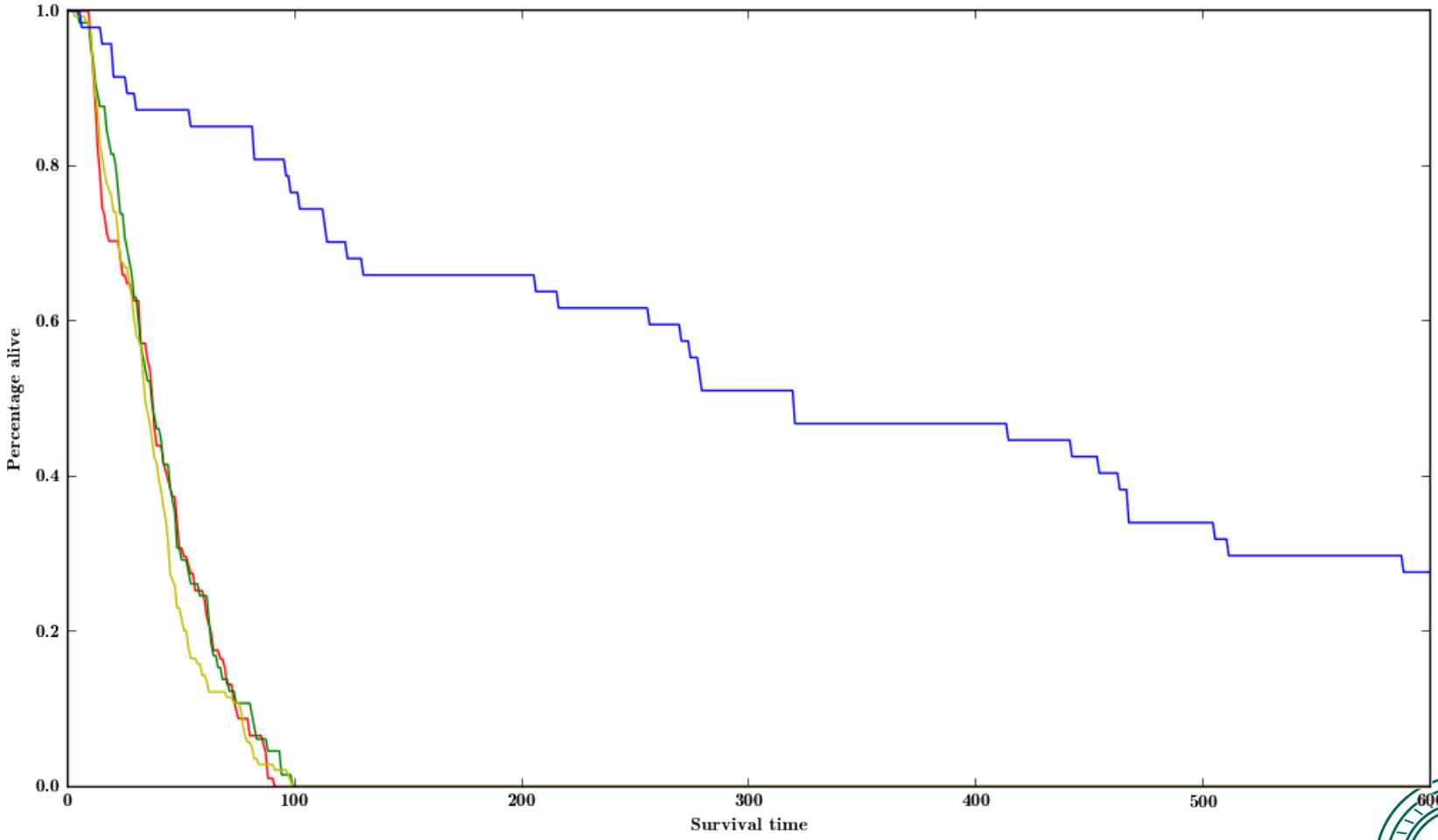
Clinical differences: Survival curves

- Also observe difference in survival curves:



Clinical differences: Survival curves

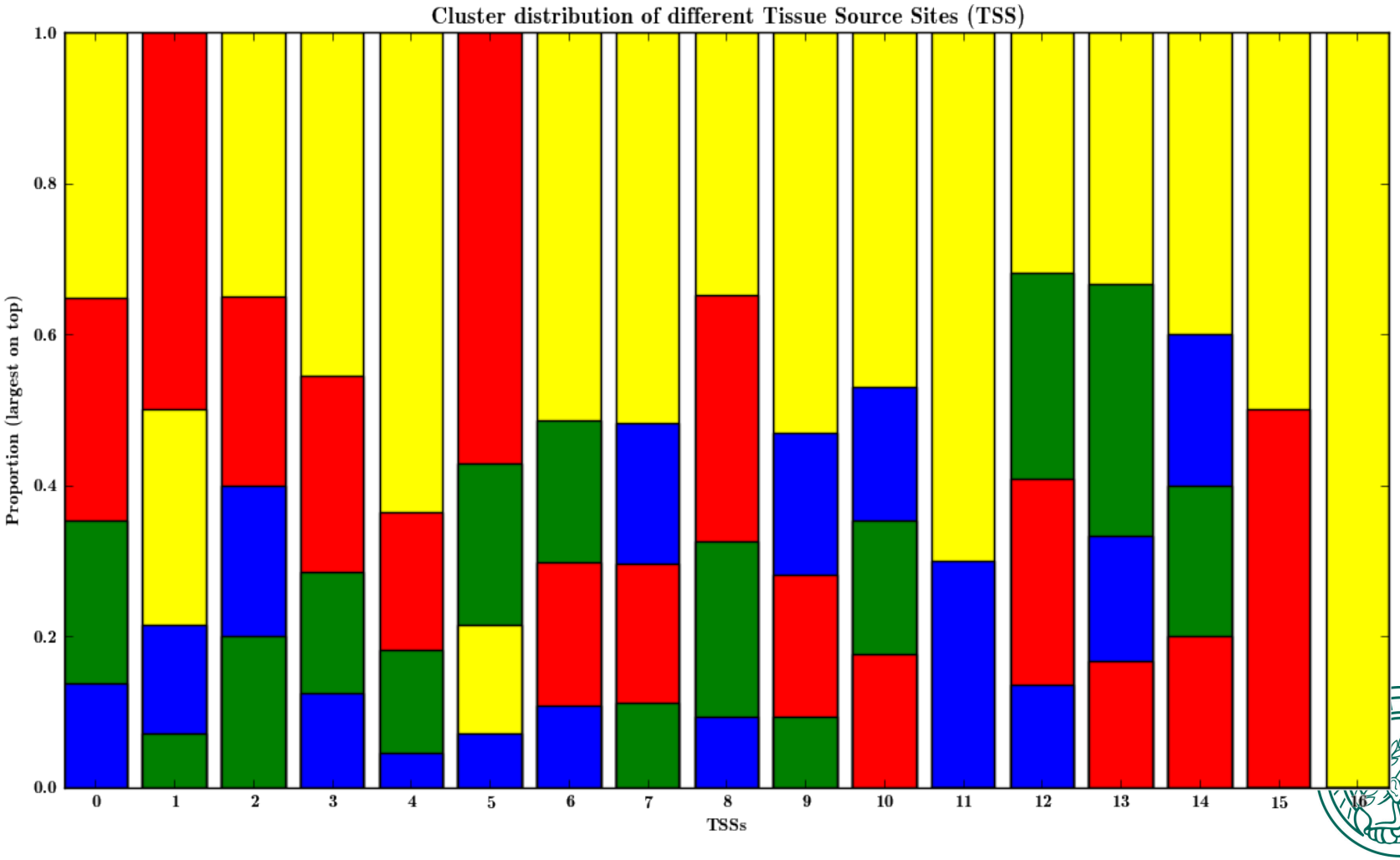
- Looking only at those who died:



MAX-PLANCK-GESELLSCHAFT

Checking for artefacts: Tissue Source Site

- Clusters do not seem to correspond solely to Tissue Source Site (source lab of sample)



Outline of talk

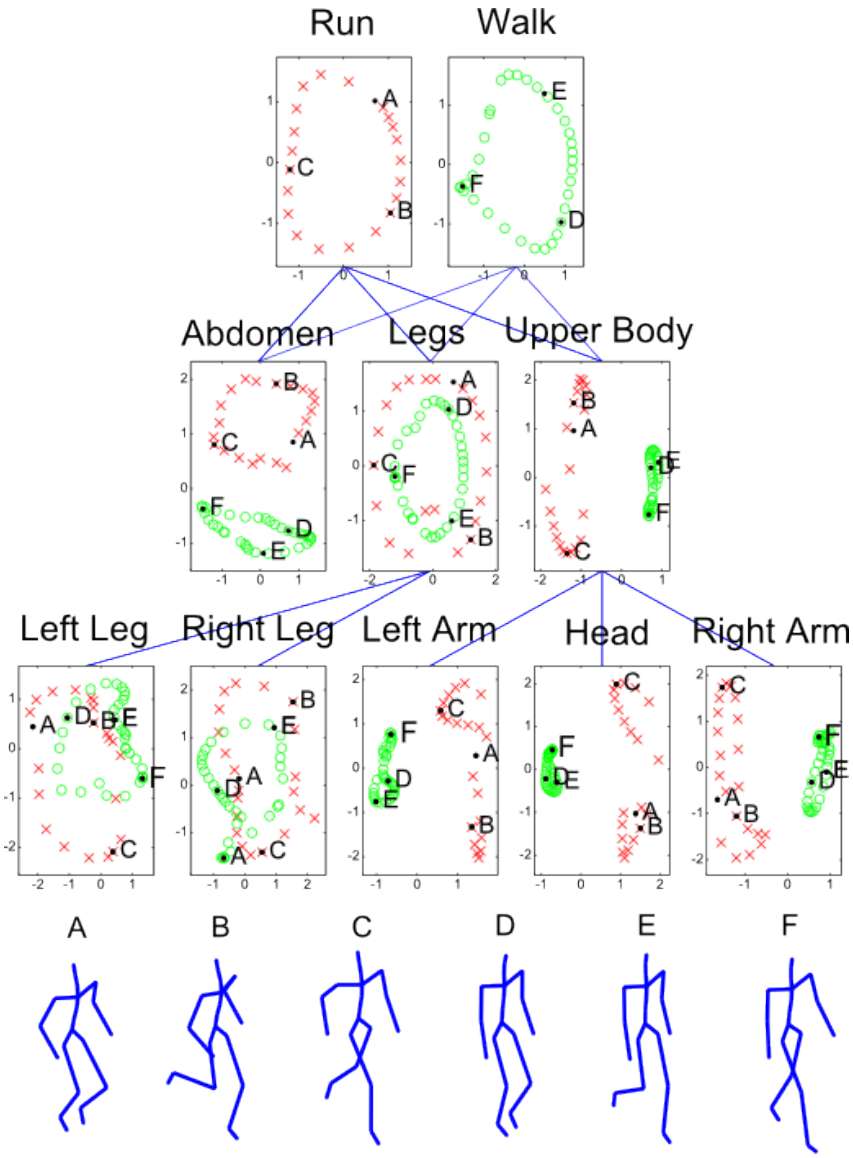
MAX-PLANCK-GESELLSCHAFT

Future Work

- Still need to carry out Gene Ontology analysis and analyse clinical data more thoroughly (e.g. producing survival graphs)
- Compare results thoroughly with the results of Verhaak, et al.
- Repeat analysis on their dataset (mostly finished but omitted here due to time constraints)
- Possible application of Deep Learning?

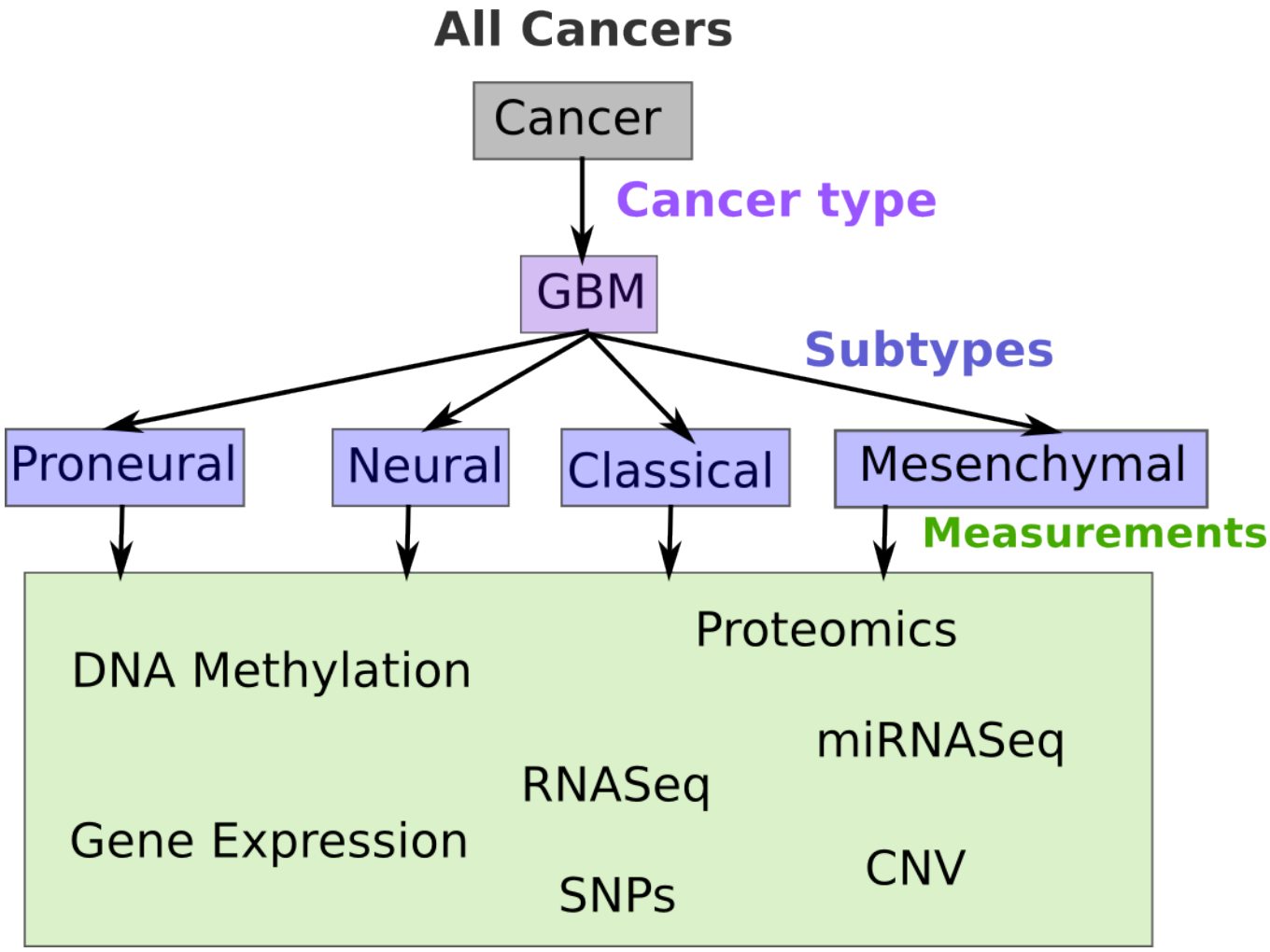
Deep Probablistic Models

- Hierarchical GPLVM example with stick figure motion:



Deep Probablistic Models

- TCGA data also has hierarchy:



Outline of talk

MAX-PLANCK-GESELLSCHAFT

Conclusion

- Sub-phenotyping of cancer is important for discovering clinically distinct sub-populations, and possible drug targets for treatments.
- Started analysis of GBM cancer data due to possible comparison with previously published work by Verhaak, et al.
- Main contributions of Machine Learning:
  - Feature selection
  - Dimensionality reduction
  - Clustering
  - Handling missing data
  - Principled data fusion
- Any suggestions for these tasks would be appreciated

Thanks for your time

Questions?

MAX-PLANCK-GESELLSCHAFT