ETH zürich

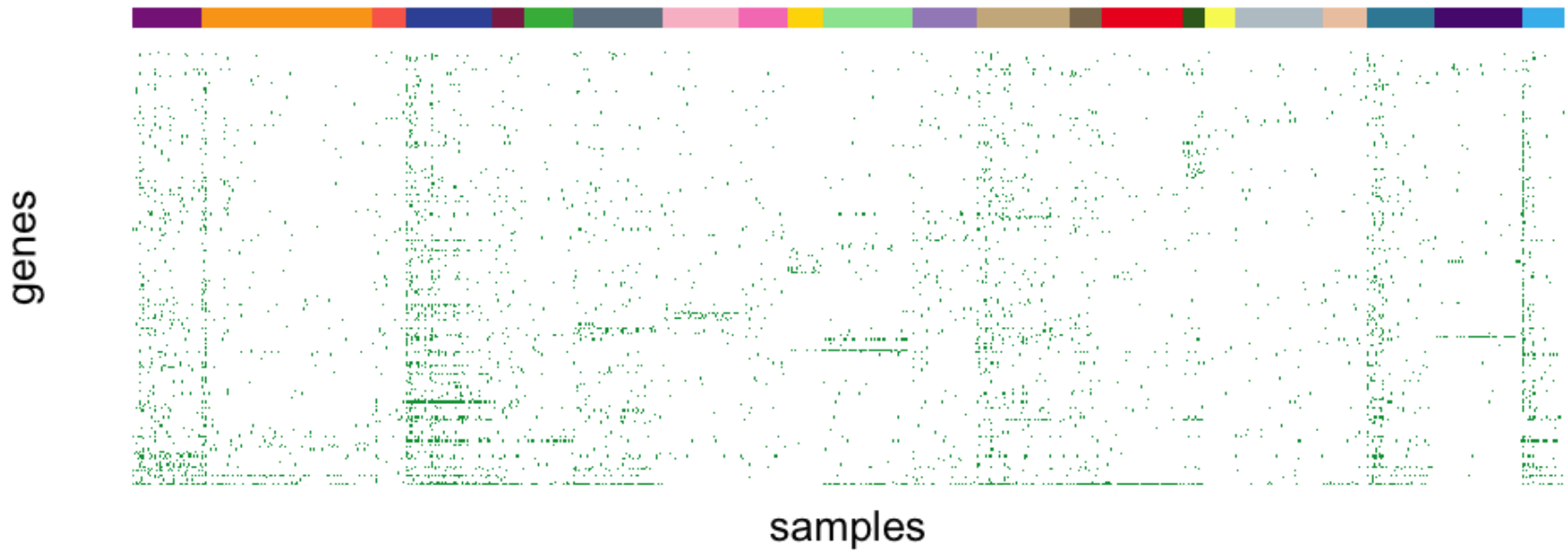
Mutational interactions define novel cancer subgroups

Can they inform precision oncology?

Jack Kuipers, ETH Zürich
Thomas Thurnherr, Giusi Moffa, Polina Suter, Jonas Behr
Ryan Goosen, Gerhard Christofori & Niko Beerenwinkel

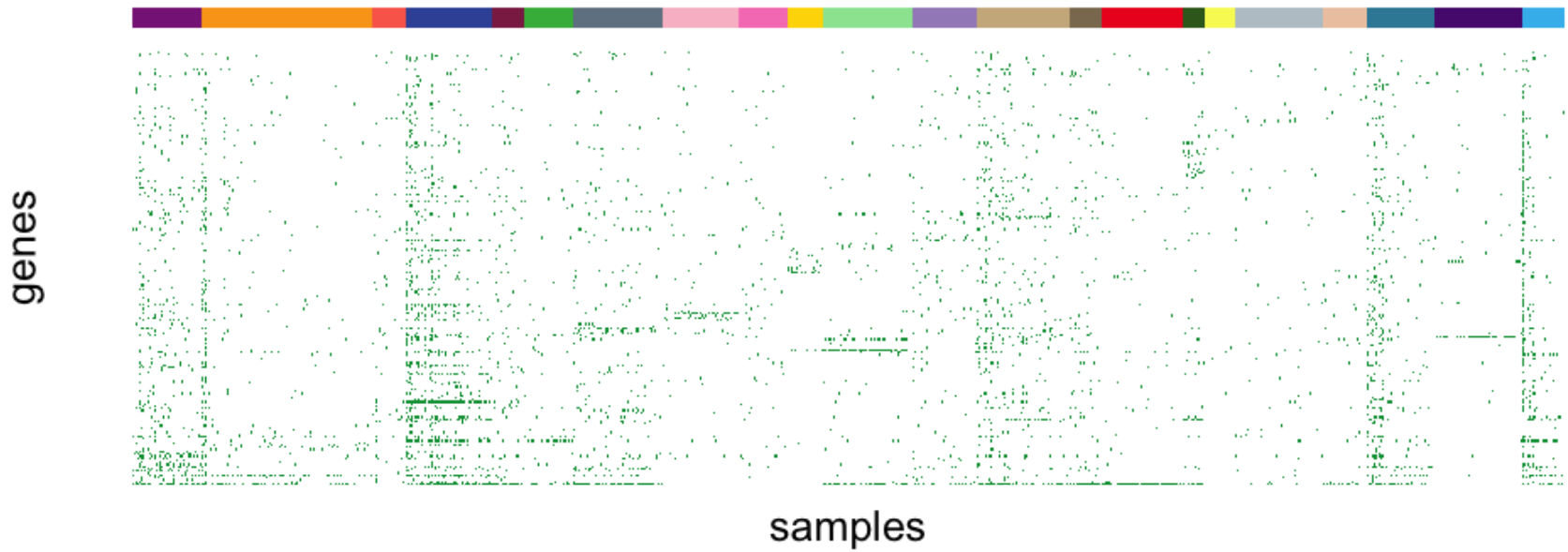
4th June 2019

TCGA mutations



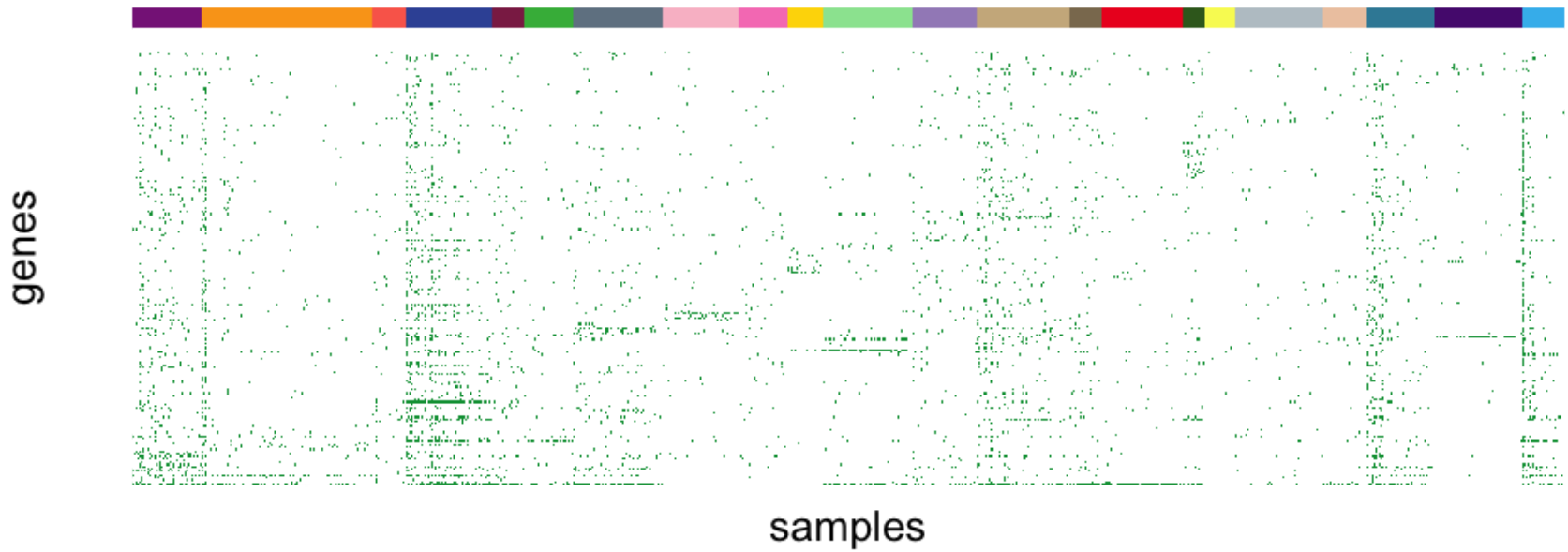
- characterize mutation patterns within and across cancer types
  - correlations and interactions
  - generative model

TCGA mutations



- characterize mutation patterns within and across cancer types
  - correlations and interactions
  - generative model
- cluster patient samples by mutation profiles
  - survival prediction
  - precision treatment?

TCGA data



- characterize mutation patterns within and across cancer types
  - correlations and interactions
  - generative model

- 22 cancer types with more than 100 samples
  - 8198 patients
- 16 most significantly mutated genes per cancer type MutSig2CV
  - 201 genes

Bayesian networks

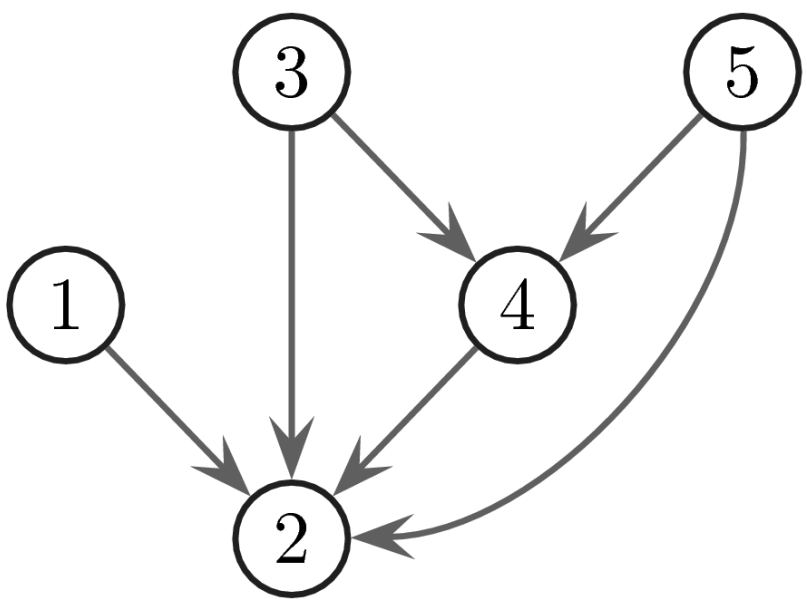
Underlying structure comprised of DAGs

Directed Acyclic Graphs

- random variable on each node

  regressed on parents

- edges encode conditional dependencies



Bayesian networks

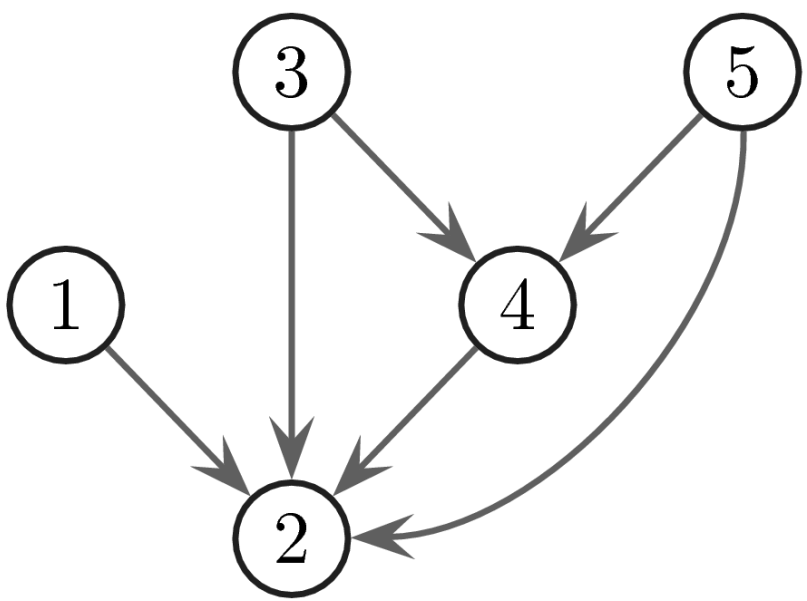
Underlying structure comprised of DAGs

Directed Acyclic Graphs

- random variable on each node

  regressed on parents

- edges encode conditional dependencies



DAGs can be:

- generated recursively

  Robinson, 1970, 1973

1	1
2	3
3	25
4	543
5	29281
	…
21	$\approx 10^{80}$

- sampled uniformly

  Kuipers and Moffa, Stats Comp 2015

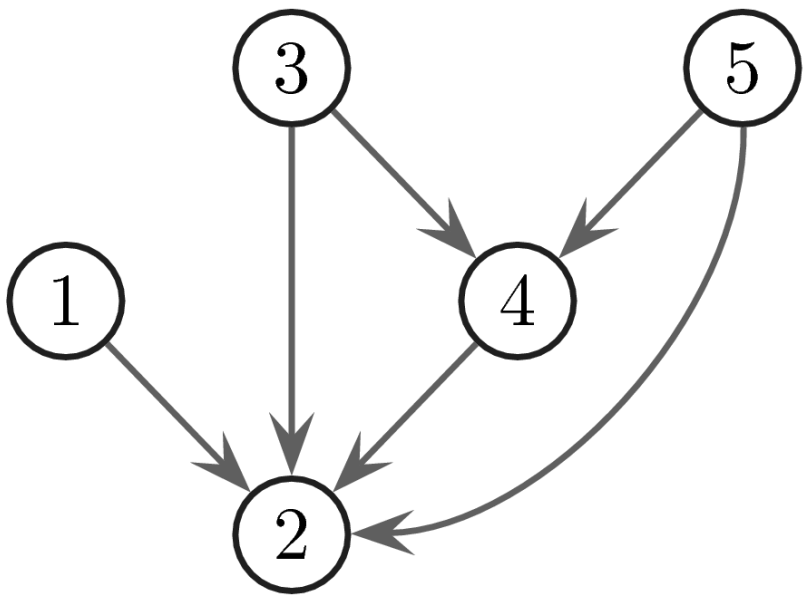
Partition MCMC

Underlying structure comprised of DAGs

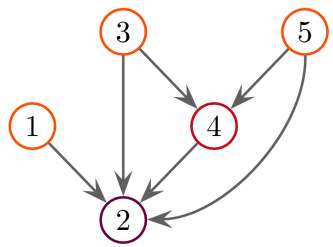
- random variable on each node

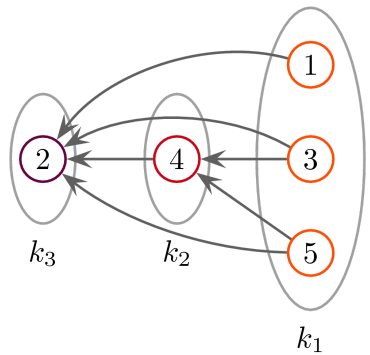
  regressed on parents

- edges encode conditional dependencies



Rearrange DAG as ordered partition





Build MCMC on space of partitions

- join or break elements
- swap nodes between them

Unbiased sampling

Kuipers and Moffa, JASA 2017

Larger Bayesian network inference

Speed up inference for large DAGs by filtering parents arXiv:1803.07859

- filter with independence tests

  PC algorithm, Spirtes, Glymour and Scheines, 2000

- allow one additional parent
- MCMC search and score

Example

- 20 nodes
- 200 observations
- 80 repetitions
- 1.4 expected number of parents

Larger Bayesian network inference

Speed up inference for large DAGs by filtering parents arXiv:1803.07859

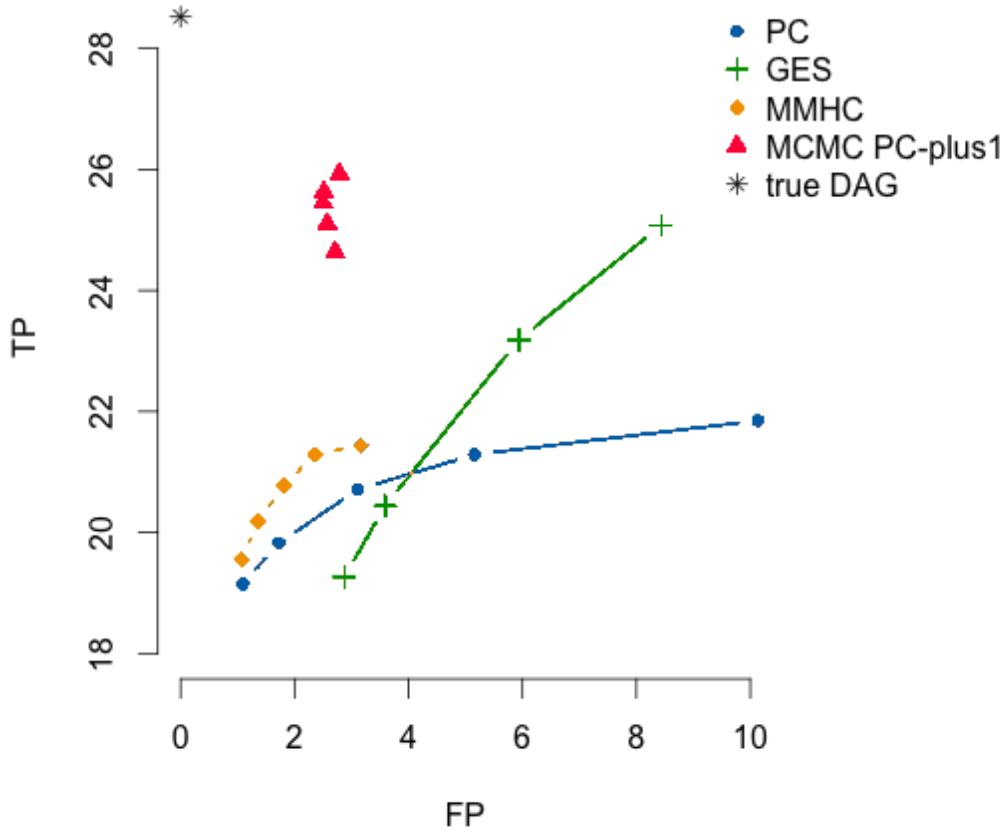
- filter with independence tests

  PC algorithm, Spirtes, Glymour and Scheines, 2000

- allow one additional parent
- MCMC search and score

Example

- 20 nodes
- 200 observations
- 80 repetitions
- 1.4 expected number of parents



GES: Greedy equivalent search
Chickering, JMLR 2002
MMHC: Max-min hill climbing
Tsmardinos, Brown and Aliferis, ML 2006

Iterative improvement

Speed up inference for large DAGs by filtering parents arXiv:1803.07859

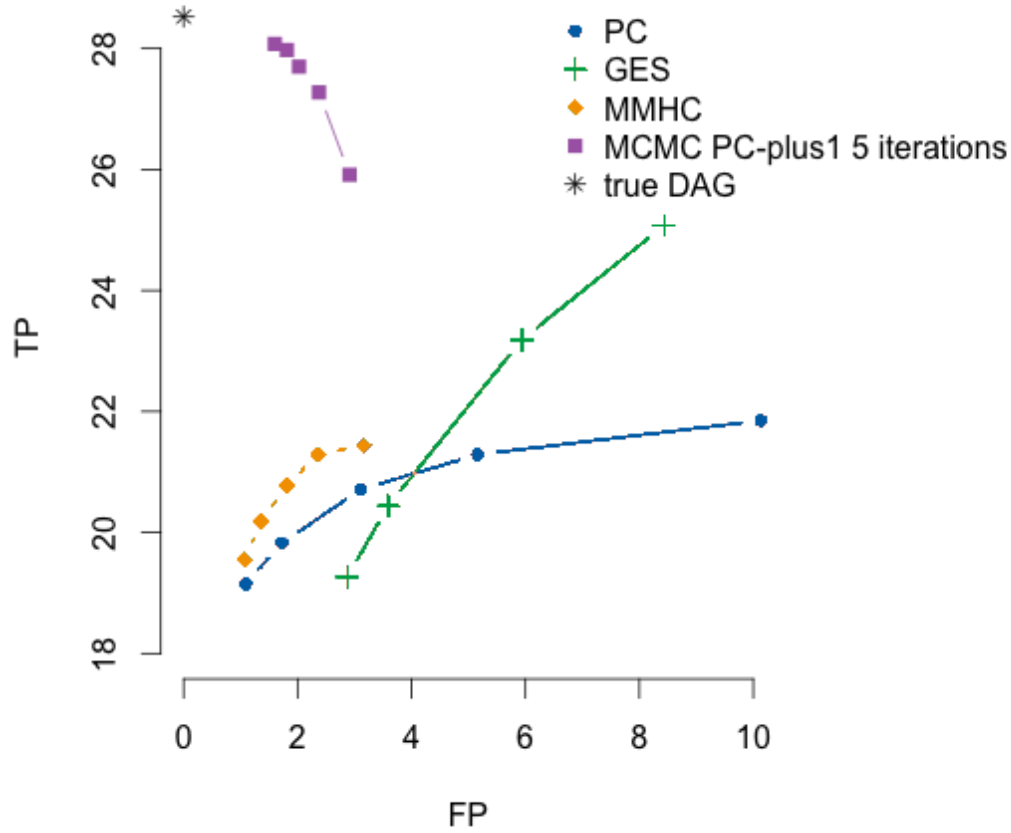
- filter with independence tests

  PC algorithm, Spirtes, Glymour and Scheines, 2000

- allow one additional parent
- iteratively improve search space

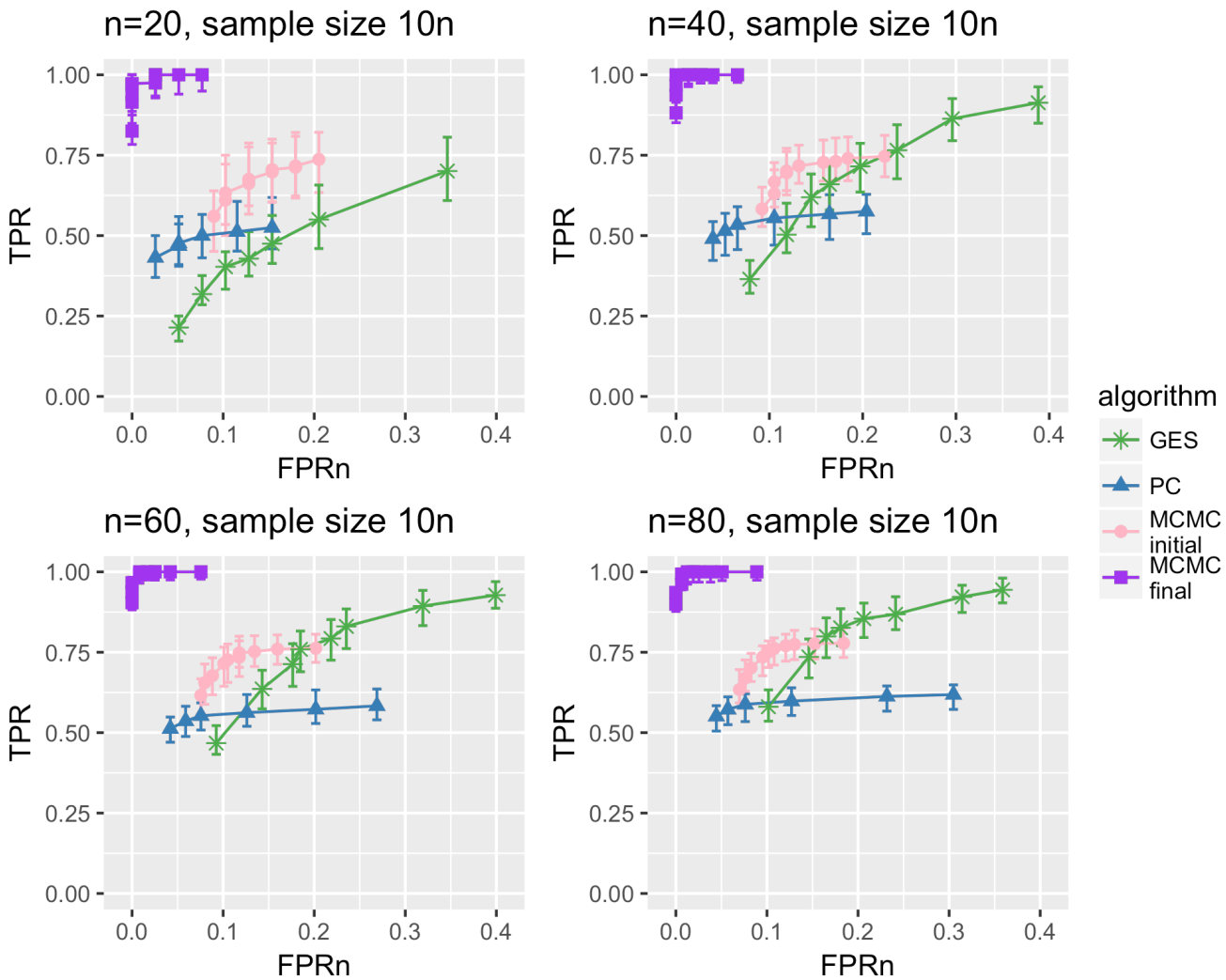
Example

- 20 nodes
- 200 observations
- 80 repetitions
- 1.4 expected number of parents



GES: Greedy equivalent search
Chickering, JMLR 2002
MMHC: Max-min hill climbing
Tsmardinos, Brown and Aliferis, ML 2006

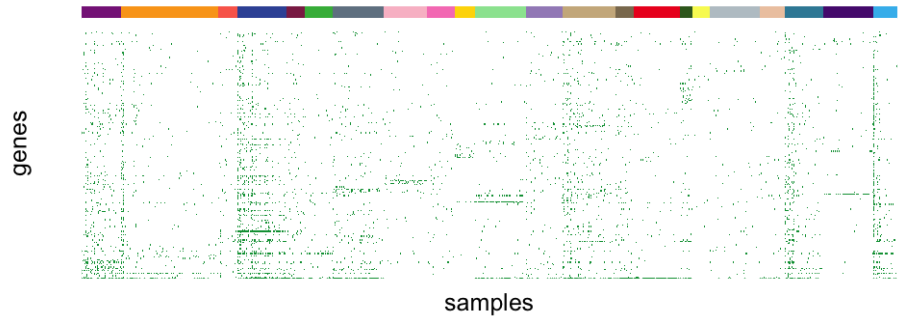
Larger DAGs


n=20, sample size 10n
n=40, sample size 10n
n=60, sample size 10n
n=80, sample size 10n
TPR
FPRn
0.00
0.25
0.50
0.75
1.00
0.0
0.1
0.2
0.3
0.4
algorithm
GES
PC
MCMC initial
MCMC final

BDe score

Bayesian Dirichlet equivalent (BDe) score

Heckerman and Geiger, UAI 1995



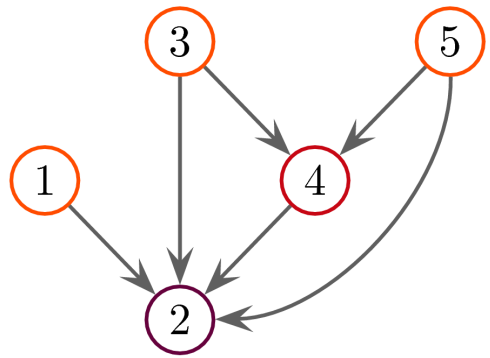
Binary case for DAG G

- node X with m parents $\boldsymbol{Y}$
- each state of $\boldsymbol{Y}$ has parameter $\theta_{\boldsymbol{Y}}$

$$P(X = 1|\boldsymbol{Y}) = \theta_{\boldsymbol{Y}}$$

- beta prior on θ with hyperparameter

$$\alpha = \beta = \frac{\chi}{2^m}$$



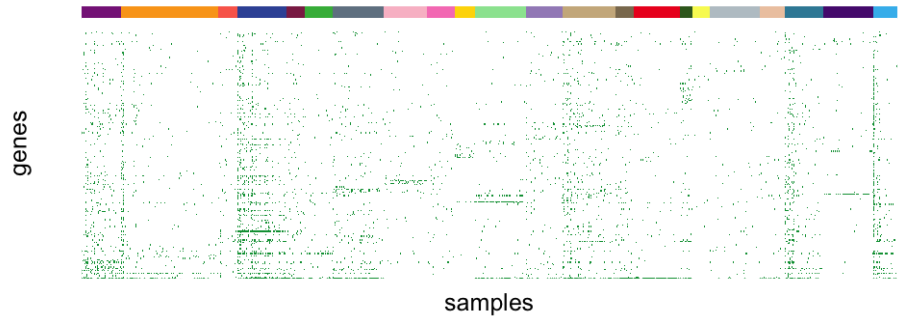
BDe metric is marginal likelihood $P(D|G)$

BDe score is posterior $P(G|D)$

TCGA graph sampling

For each cancer type

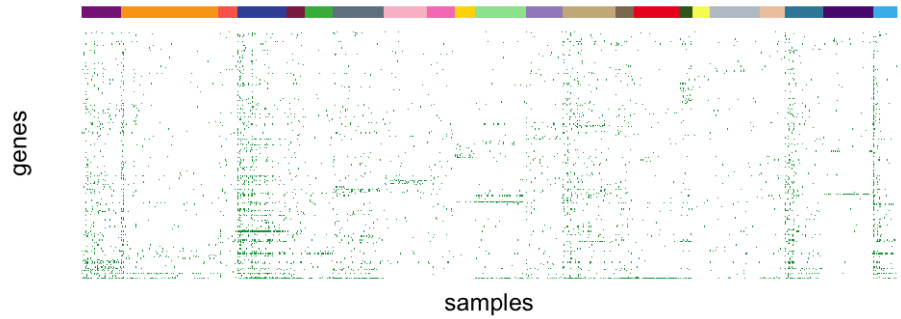
- sampled 100 DAGs from posterior
- edge prior from STRING



TCGA graph sampling

For each cancer type

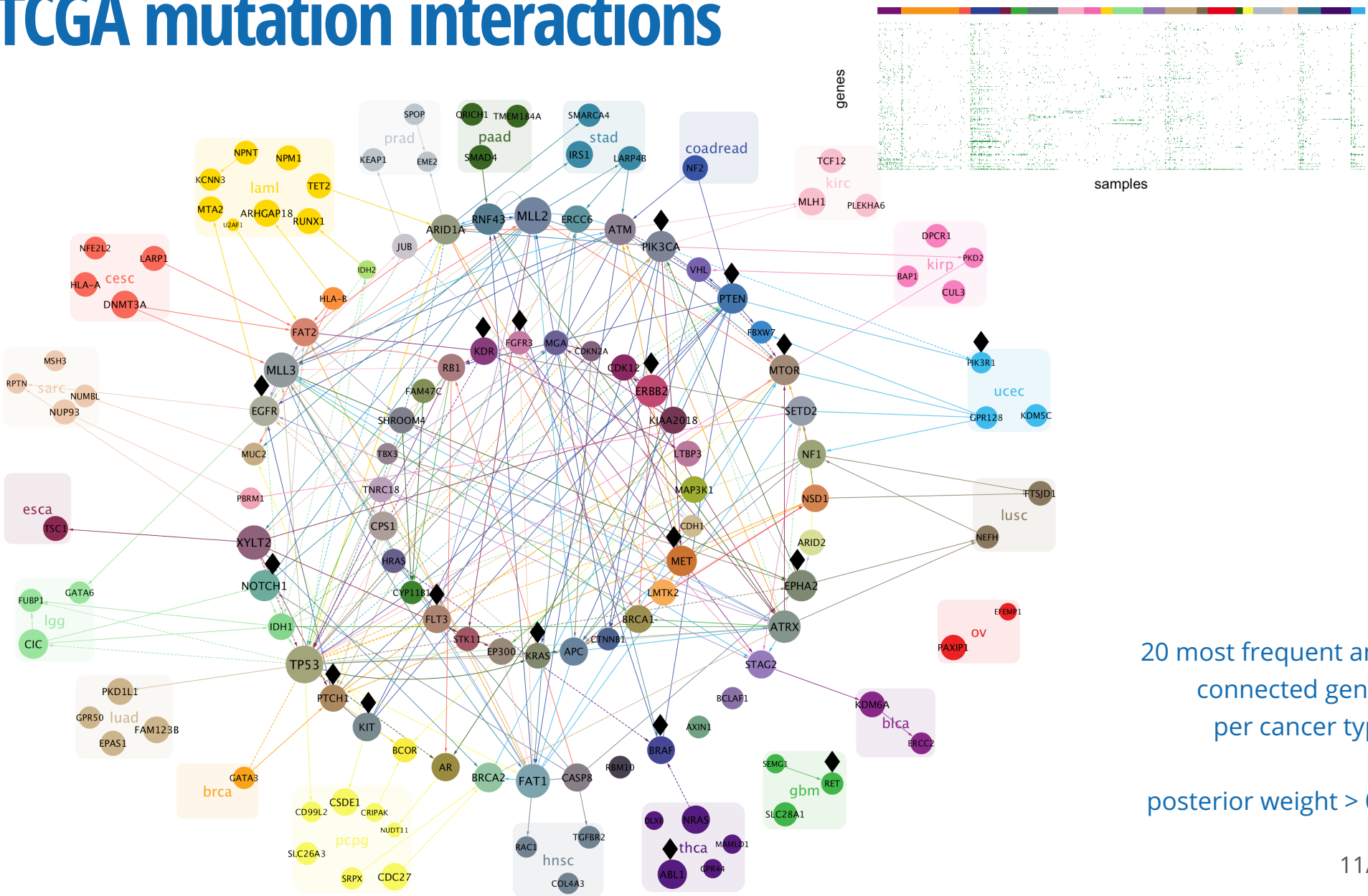
- sampled 100 DAGs from posterior
- edge prior from STRING



To visualise networks:

- Select edges
  - posterior weight > 0.5
- Select 20 most frequent and connected genes
  - frequency $\times$ edges
- Extract subnetwork
  - colour edges by cancer type
  - overlay

TCGA mutation interactions



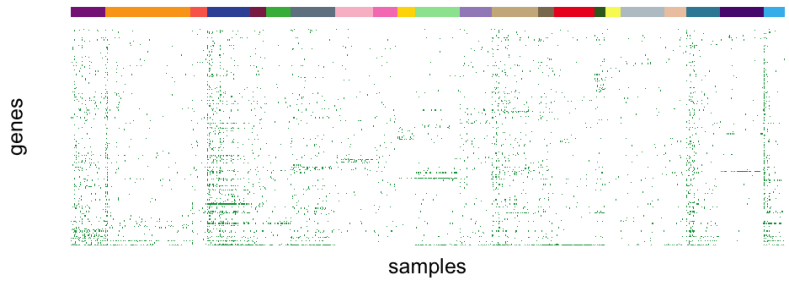
20 most frequent and connected genes per cancer type

posterior weight > 0.5

TCGA graph sampling

For each cancer type

- sampled 100 DAGs from posterior
- edge prior from STRING



TCGA graph sampling

For each cancer type

- sampled 100 DAGs from posterior
- edge prior from STRING

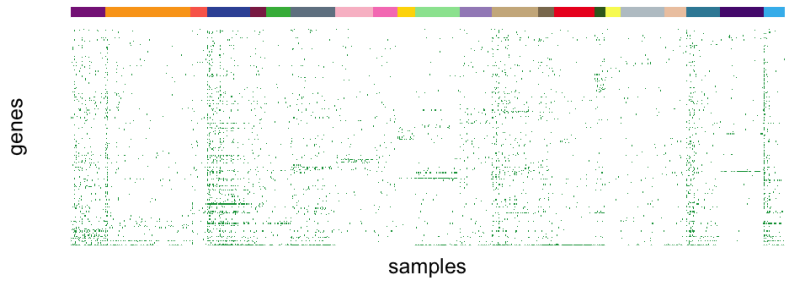
To visualise heterogeneity:

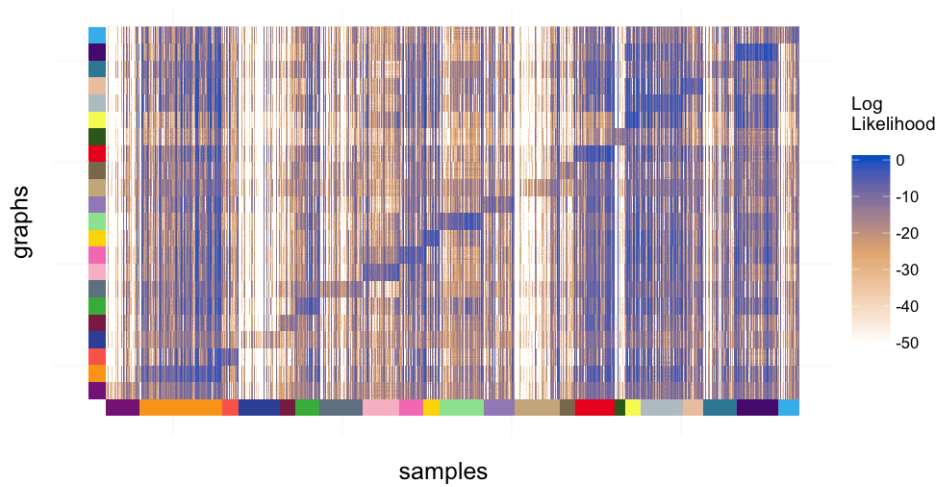
- Fit each patient sample to each graph

$$P(D_i|G_j, \bar{\theta}_j)$$

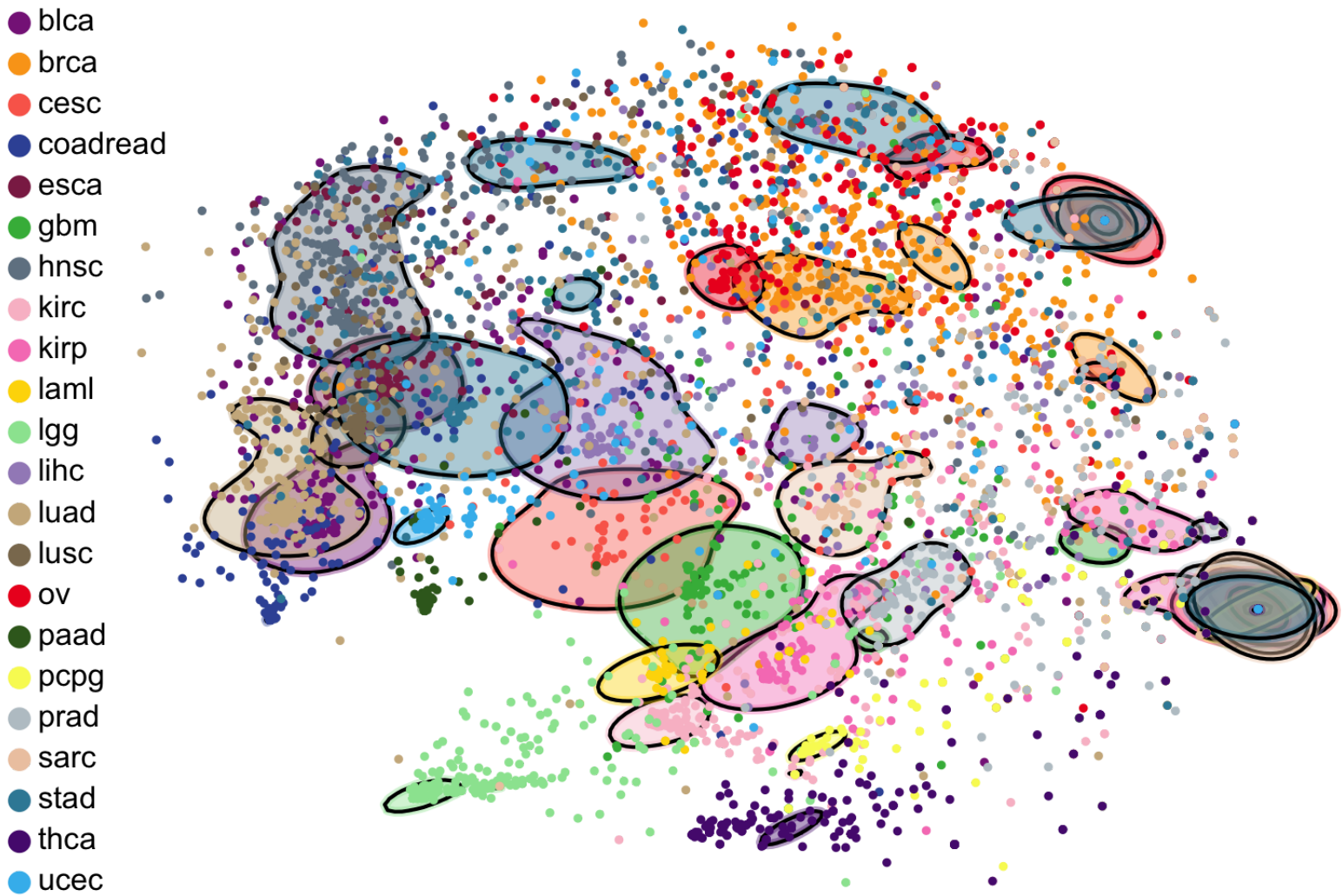
using posterior means $\bar{\theta}$

- Calculate Jenson-Shannon divergence between columns
  - distance between patient samples
  - project to 2D





Inter-patient heterogeneity



EM MAP mixture model clustering

Weight patient samples into k groups

- compute MAP relative sizes τ
- learn MAP DAG G, θ for each
- reweight patient samples

$$\propto \tau_j P(D_i | G_j, \bar{\theta}_j)$$

- repeat till convergence

Latent Z indicates which graph model each patient sample derives from

$$D_i | (Z_i = j) \sim (G_j, \theta_j)$$

$$P(Z_i = j) = \tau_j$$

EM MAP mixture model clustering

Weight patient samples into k groups

- compute MAP relative sizes τ
- learn MAP DAG G, θ for each
- reweight patient samples

$$\propto \tau_j P(D_i | G_j, \bar{\theta}_j)$$

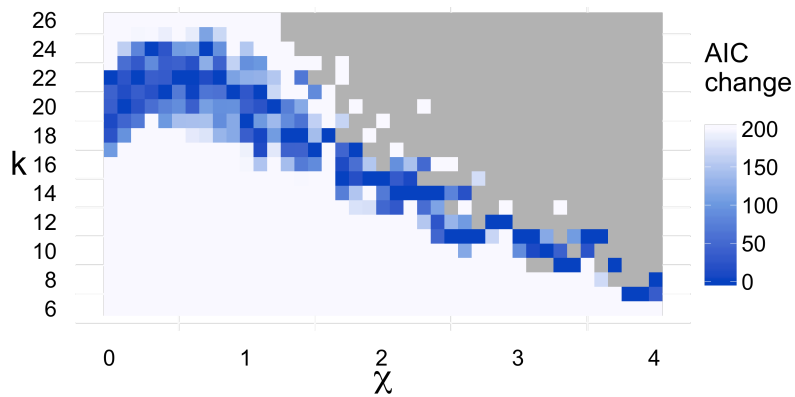
- repeat till convergence

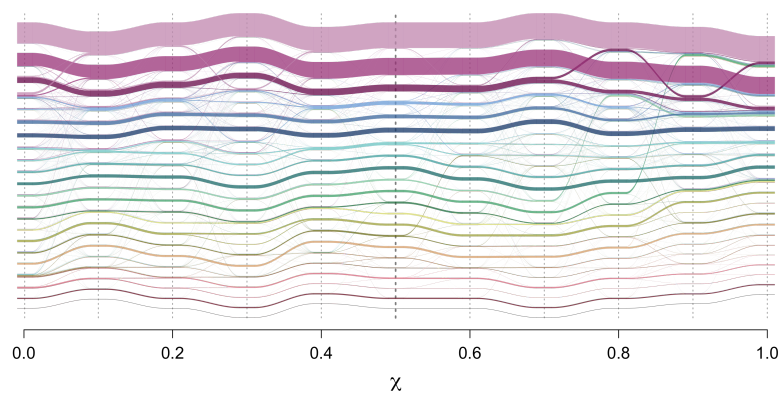
Latent Z indicates which graph model each patient sample derives from

$$D_i | (Z_i = j) \sim (G_j, \theta_j)$$

$$P(Z_i = j) = \tau_j$$

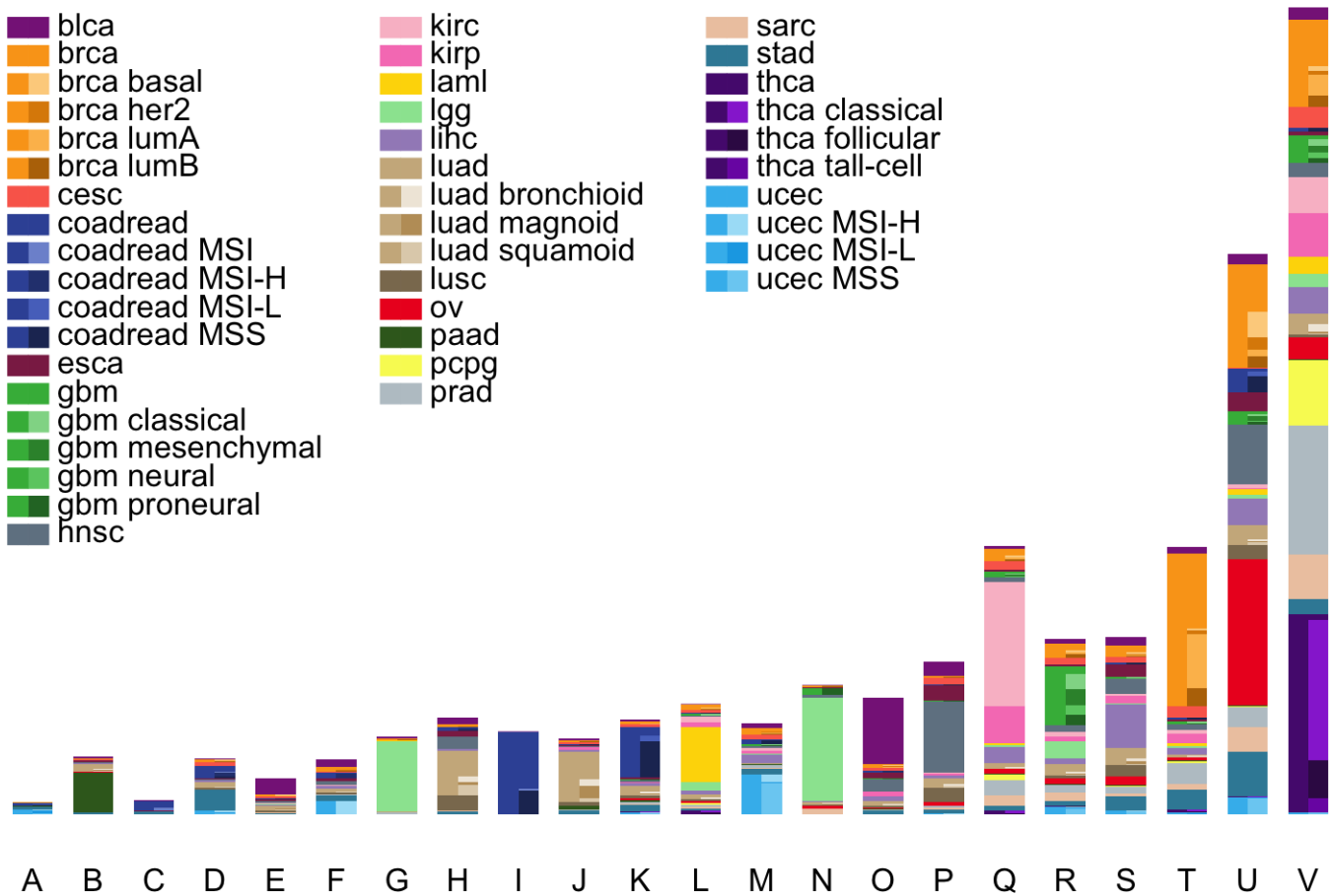
Start with no edges parameter learning





- find 22 clusters for $\chi = 0.5$

Graphical clustering



- alternative stratification?

Survival analysis

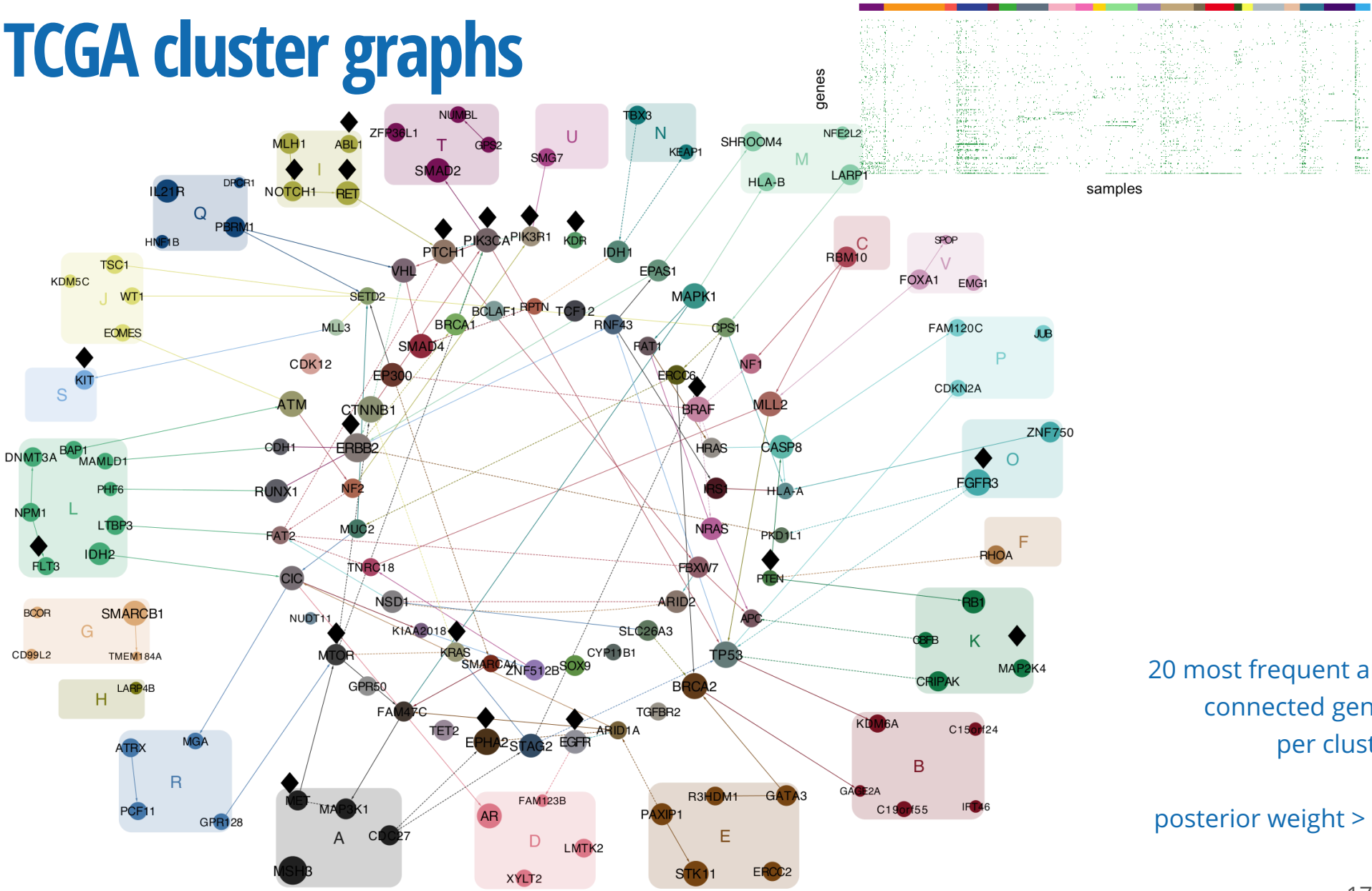
Check if clusters based on mutational profiles are a significant survival predictor

- above age, stage and cancer type

METHOD	UNCORRECTED LR	CORRECTED LR	P-VALUE
Hierarchical clustering (Hamming distance)	11.4	5.7	0.95
Non-negative matrix factorisation	104.7	12.7	0.23
K-means	172.0	29.8	1.5×10^{-5}
Gaussian mixture model (Mclust)	205.6	33.1	1.4×10^{-6}
Bernoulli mixture model (no edges, $\chi = 0$)	240.9	34.0	7.5×10^{-7}
Bernoulli mixture model (no edges, $\chi = 0.5$)	242.4	35.7	2.1×10^{-7}
Graphical clustering (edges, $\chi = 0.5$)	253.0	37.0	7.6×10^{-8}

- Likelihood ratio (LR) from Cox proportional hazard regression including cluster assignment
- number of clusters fixed to 22

TCGA cluster graphs



20 most frequent and connected genes per cluster

posterior weight > 0.5

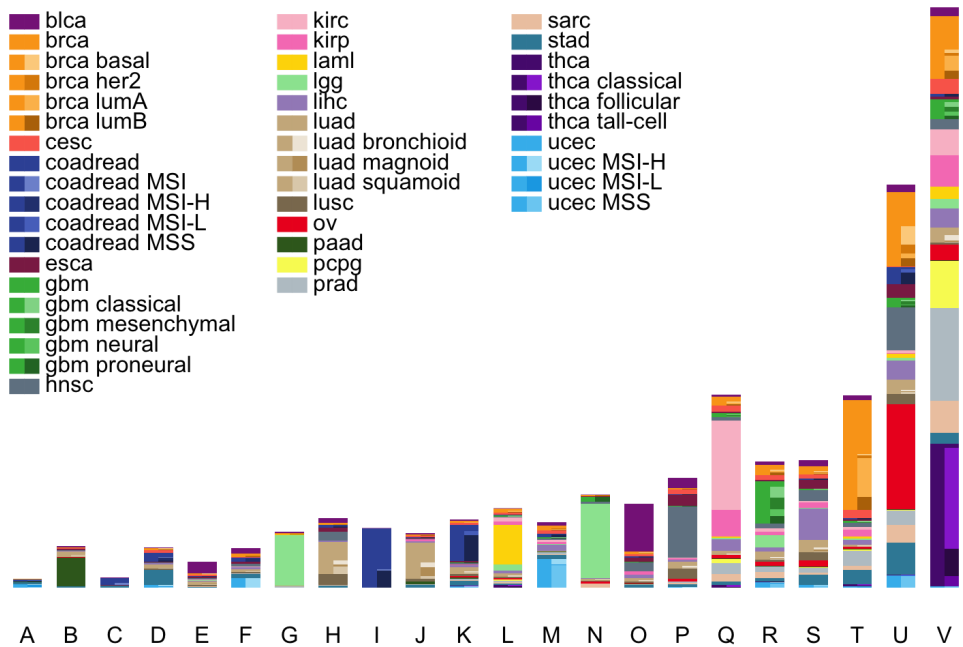
Towards precision oncology?

Can cluster mutation profiles

- recapitulate a lot of clinical information

Want to integrate clinical covariates

- cluster conditionally



Towards precision oncology?

Can cluster mutation profiles

- recapitulate a lot of clinical information

Want to integrate clinical covariates

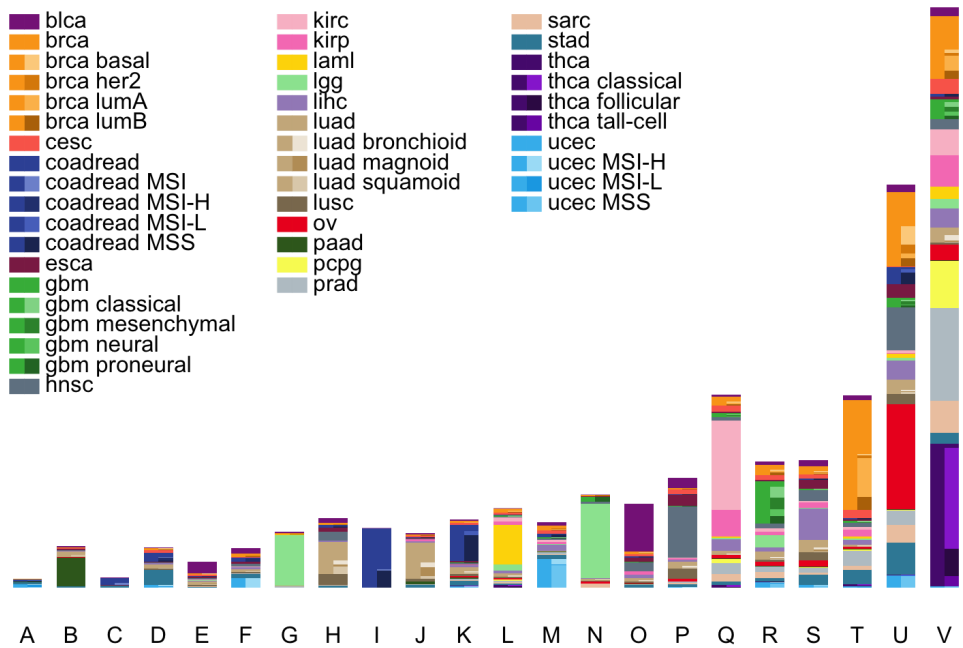
- cluster conditionally

Want to link to drug response

- cell line data GDSC

Want to integrate further data

- expression
- copy number aberations
- methylation



Tumour progression

Tumours are heterogeneous

- complex clonal structure

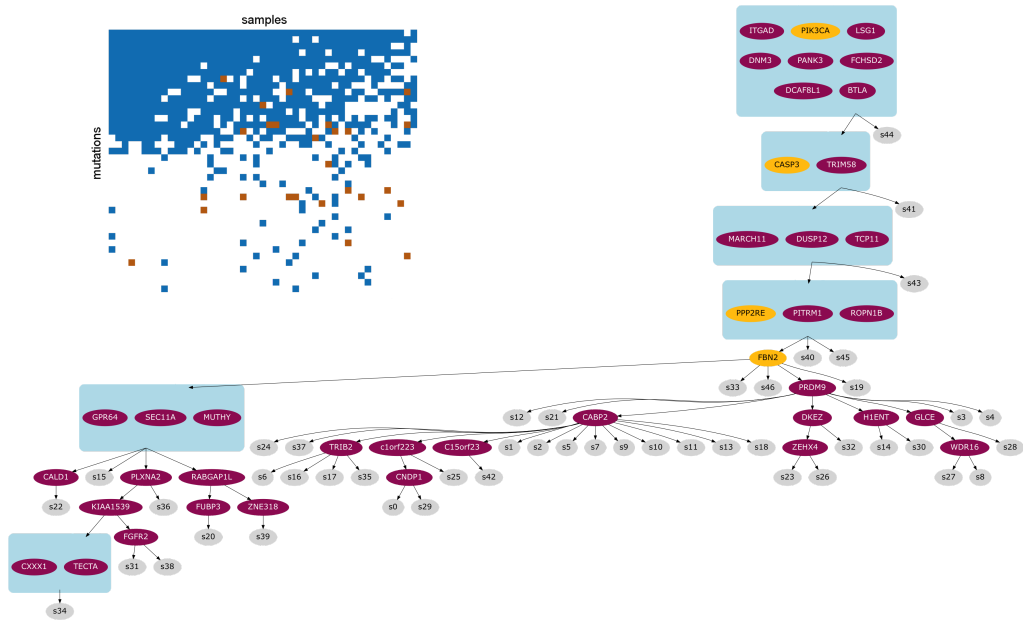
They evolve

- over time
- under treatment

Mutation profiles were a single snapshot!

Should account for

- clonal structure
- tumour progression



Data Wang … Navin, Nature 2014

Inference Jahn, Kuipers and Beerenwinkel, GB 2016

Summary

Can use partitions to sample DAGs JASA 2017

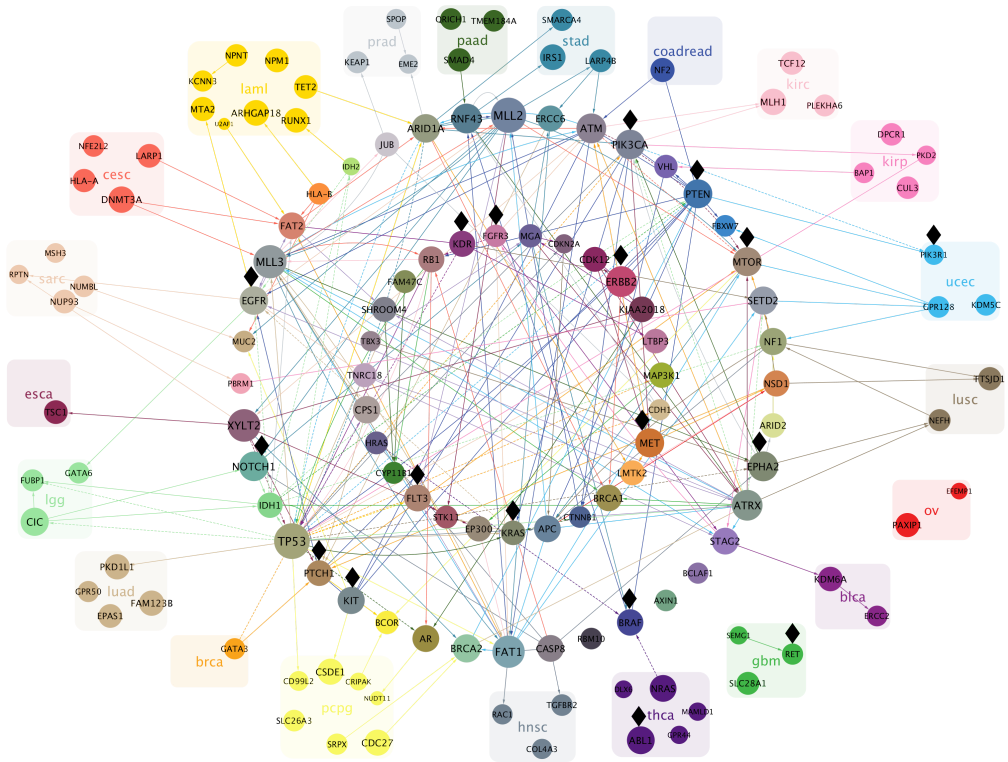
Extend to larger networks arXiv:1803.07859

Applied to pancancer mutations NC 2018

- supervised cancer type networks
- unsupervised clustering

Outlook

- integrate clinical data?
- integrate other data modalities?
- tumour progression and heterogeneity?



Computational Biology Group

ETH zürich

UCL

Giusi Moffa, Polina Suter
Thomas Thurnherr, Niko Beerenwinkel
Jonas Behr, Ryan Goosen, Gerhard Christofori