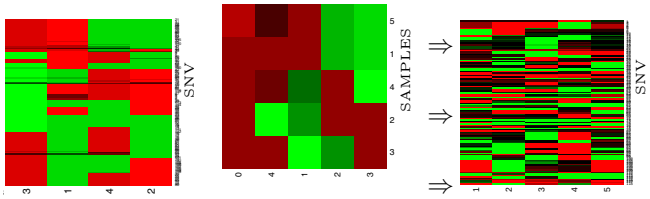
Slide 1

Modeling Tumor Heterogeneity

PETER MÜLLER, UT Austin

JUHEE LEE, UCSC, YUAN JI, U Chicago & NorthShore, K. GULUKOTA, NorthShore Health System



SUBCLONES
binary matrix
$\boldsymbol{Z} = [Z_{sc}]$
SNV s, clone c

SUBCLONES
composition
weights
$\boldsymbol{w}_t = (w_{tc})$
sample t

SAMPLES
data: n_{ts}
mutations

Slide 2

Tumor Heterogeneity

- Mutations acquired over a tumor's life history
- Every new mutation gives rise to a new subpopulation of cells ("subclone")
- $\longrightarrow$ heterogeneous population of cells, composed of subpopulations with varying numbers of mutations.
- Tumor history imprinted in each sample as the mosaicism of mutations.

Slide 3

Data

SNV: point mutations, $s = 1, \ldots, S$

Data: N_{st} = # reads mapped to locus of SNV s in sample t.
n_{st} = # of these *with* SNV.

Sampling model: $n_{st} \sim \text{Bin}(N_{st}, p_{st})$

Prior: in words,

(i) p_{st} arises as a composition of sample t as a mixture of C latent cell subclones.

(ii) Mutation s in subclone c is either present ($Z_{sc} = 1$) or not ($Z_{sc} = 0$).
$\boldsymbol{Z}_c = (Z_{sc},\ s = 1, \ldots, S)$ defines subclone, c.

(iii) Prior $p(\boldsymbol{Z})$ on $(S \times C)$ binary matrix $\boldsymbol{Z}$,
prior $p(\boldsymbol{w})$ on mixture weights w_{tc} for composition (i).

Slide 4

Inference

Goal: Reconstruct cell subpopulations = estimate $\boldsymbol{Z}$ and C.

Problem: Deconvolution of p_{st} as a mixture of binary indicators Z_{sc}

$$p_{st} = \textstyle\sum_c w_{tc} Z_{sc} \quad + w_{t0} p_{s0}$$

plus "background noise"

Real problem: $\boldsymbol{Z}$ is latent, need to infer $\boldsymbol{Z}$ from the data.

Identifiability: In principle even feasible with one sample. Weights are identified across mutations s.

Slide 5

Prior

Latent cell types: $p(\boldsymbol{Z})$ on $(S \times C)$ binary matrix, w. random C.

Feature allocation: Think of SNV s selecting cell types c
Features (dishes) = c; experimental units (customers) = s

Random feature allocation: define $p(\boldsymbol{Z})$ as

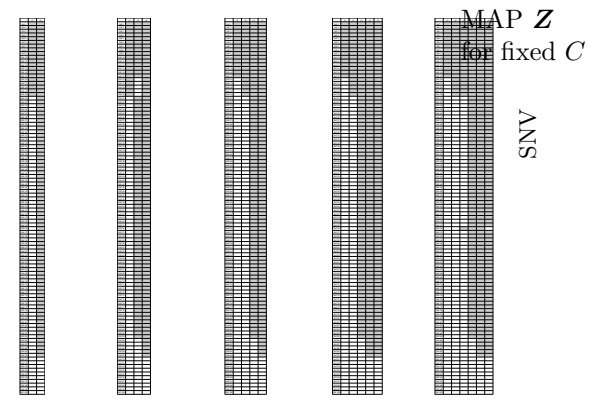
- $p(Z_{sc} = 1 \mid \pi_c) = \pi_c$, $c = 1, \ldots, C$
- $\pi_c \sim \text{Be}\left(\frac{\alpha}{C}, 1\right)$
- Drop unselected features

IBP as $C \to \infty$.

Composition of sample t as mix of cell types:
$(w_{tc},\ c = 1, \ldots, C) \sim \text{Dir}(\cdot)$.

Slide 6

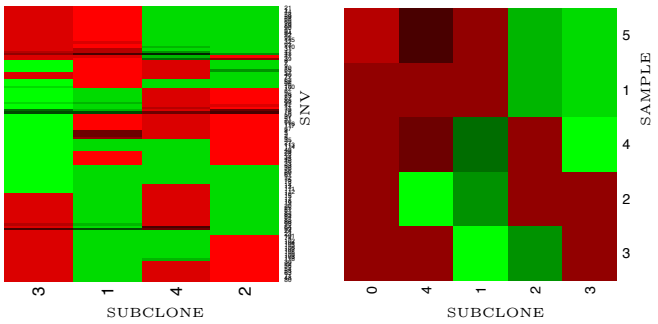
Simulation



Slide 7

Results – Pancreatic Cancer

$n = 5$ samples of pancreatic cancer (PDAC, pancreatic ductal adenocarcinoma).



$p(z_{sc} = 1 \mid data)$ $E(w_{tc} \mid data)$

Slide 8

Computation

. . . is a pain.

Slide 9

MAD Bayes for TH

with Yanxun XU, UT Austin; Yuan YUAN, Baylor C.of Med.;
Yuan JI and Kamalakar Gulukota, NorthShore Hospital.

DP mixture: Kulis & Jordan (2012) recognize log posterior $\approx$ criterion function in k-means – voila!

This is for normal sampling, asymptotically for small variance and shrinking total mass.

IBP: Broderick et al. (2013) extend a similar argument to the IBP, with normal sampling and small variance and shrinking rate of new features,

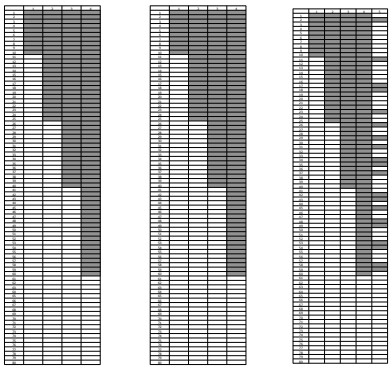
IBP with binomial sampling: same argument can be made :-)
using increasing scaling of Bin with β and shrinking IBP par γ, using $\gamma = \exp(-\beta\lambda^2)$

Approx posterior: use k-means with different starting values to characterize posterior.

Slide 10

Simulation

True and estimated $\boldsymbol{Z}$

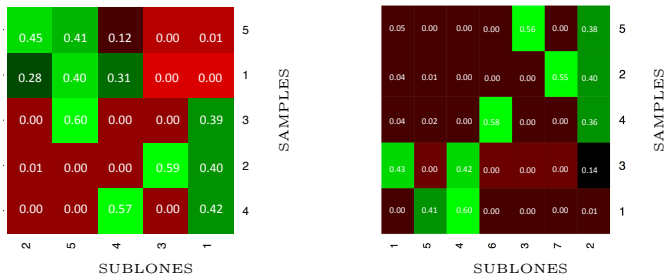


Truth $\lambda^2 = 8, 10$ $\lambda^2 = 6$

Slide 11

Results – Pancreatic Cancer

$n = 5$ samples of pancreatic cancer (PDAC, pancreatic ductal adenocarcinoma). Estimated w_{tc}:

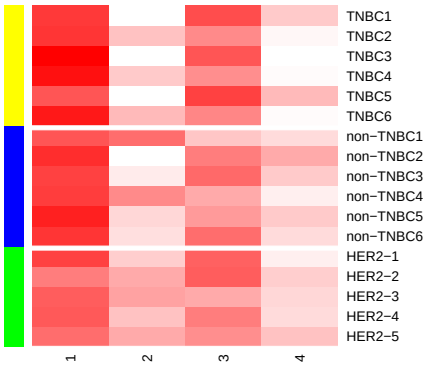


$S = 118$ SNV's in KEGG pathway $S = 7000$ SNV's

Slide 12

Results – Breast Cancer

Horvath et al. (2013): $n = 17$ BC patients, $S = 329$ SNV's.



Estimated w_{tc} with $C = 4$.

Slide 13

Summary

TH: Model-based estimation of cell subpopulations is possible – and seems to work.

Big data: MCMC is not feasible anymore – alternative approaches remain feasible.

Limitations: and extensions

Tumor phylogenetics: Without condition on phylogenetic tree of subclones

A priori independent cell types: indpendent $\boldsymbol{z}_c = (Z_{1\ldots S,c})$, with $p(\boldsymbol{z}_c = \boldsymbol{z}_{c'}) > 0$, a priori (i know — arrgh!)
Alternative dependent prior using DPP or others.

CNV: we conditioned on N_{st}.
Could use N_{st} to learn about CNV.