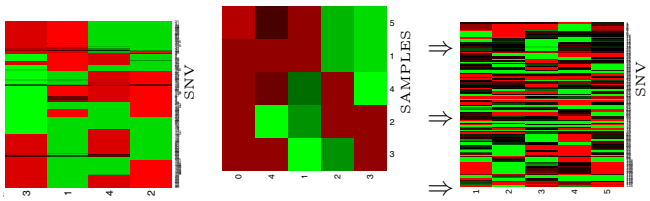
Slide 1

Modeling Tumor Heterogeneity

PETER MÜLLER and YANXUN XU, 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

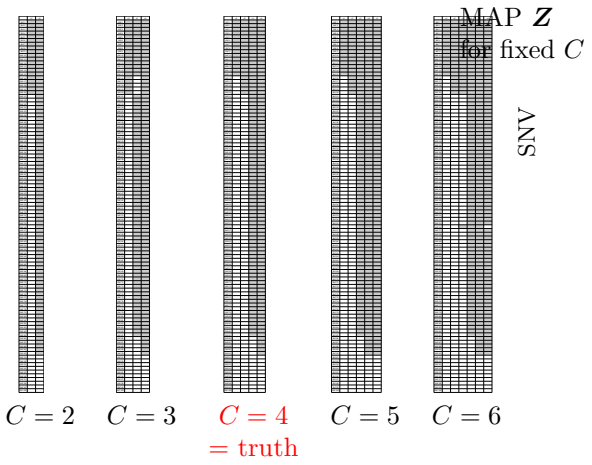
- $p(Z_{sc} = 1 \mid \pi_c) = \pi_c$, $c = 1, \ldots, C$
- $\pi_c \sim \text{Be}\left(\frac{\alpha}{C}, 1\right)$
- Drop unselected features

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

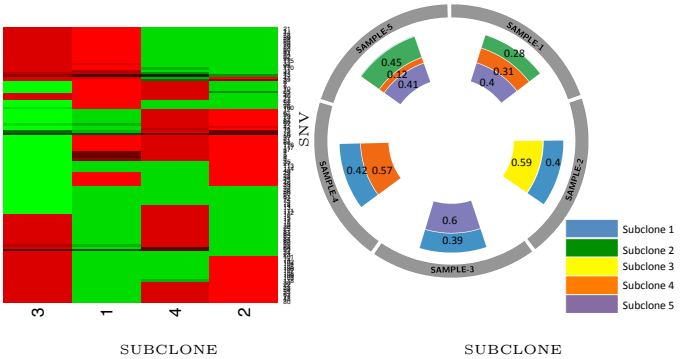
Simulation



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

MCMC: random $C \to$ transdimensional MCMC, to move across C

Sloooow mixing: challenging to define good transition prob with reasonable mixing,
like in any model comparison with informative data & high posterior correlation.
posterior simulation for given C is straightforward.

IBF model comparison: Casella & Moreno (2006 JASA) propose propose model selection with IBF (intrinsic Bayes factor). Stylized simplified version:

Slide 9

FBF Model Comparison

Casella & Moreno's IBF model comparison:

1. split data into training data $n'_{st} = \beta n_{st}; N'_{st} = \beta N_{st}$ and test data $n"_{st} = (1-\beta)n_{st}$ and $N"$ ...
2. informative prior
   $p_1(\boldsymbol{Z}, \boldsymbol{w}, p_0 \mid C) \equiv p_1(\boldsymbol{Z}, \boldsymbol{w}, p_0 \mid n', C)$
3. average over all possible splits.

We do the same, but w/o step 3 and

4. use MH proposal

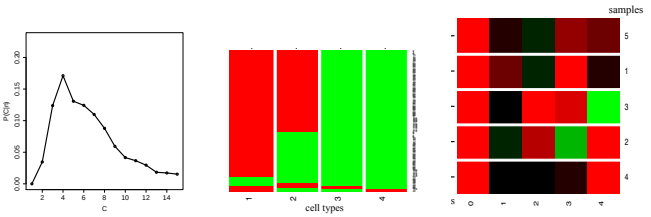
$$q(\tilde{C}, \tilde{\boldsymbol{Z}}, \tilde{\boldsymbol{w}}, \tilde{p}_0) = q(\tilde{C} \mid C)\, p_1(\tilde{\boldsymbol{Z}}, \tilde{\boldsymbol{w}}, \tilde{p}_0 \mid \tilde{C})$$

Acceptance ratio simplifies and we get a well mixing MC

In the binomial likelihood, splitting the data can also be characterized as fractional BF (O'Hagan 1995).

Slide 10

Pancreatic Cancer



$p(C \mid data)$ $\boldsymbol{Z}^\star \mid C = 4$ $\boldsymbol{w}^\star \mid C = 4, bZ^\star$

Slide 11

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 12

log posterior $\approx$ k-means like criterion:

$$-\frac{1}{\beta}\log L \approx \sum_{s=1}^{S}\sum_{t=1}^{T}\{-n_{st}\log(p_{st}) - (N_{st}-n_{st})\log(1-p_{st})\} + C\lambda^2,$$

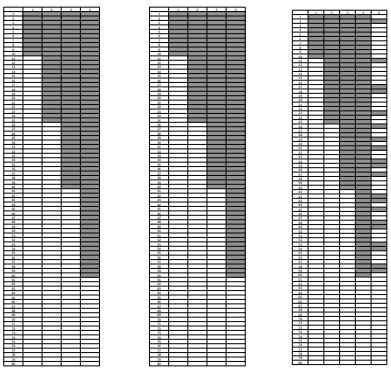
Iterative FL-means algorithm:

- min w.r.t. $\boldsymbol{Z}$ for fixed $\boldsymbol{w}$
- min w.r.t. $\boldsymbol{w}$ for fixed $\boldsymbol{Z}$
- consider incrementing $\boldsymbol{Z} \to \tilde{\boldsymbol{Z}}$ with one (randomly selected) SNV and min w.r.t. w.
keep if it lowers the criterion

Slide 13

Simulation

True and estimated $\boldsymbol{Z}$

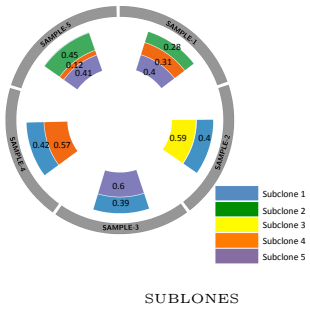


Truth $\quad \lambda^2 = 8, 10 \quad \lambda^2 = 6$

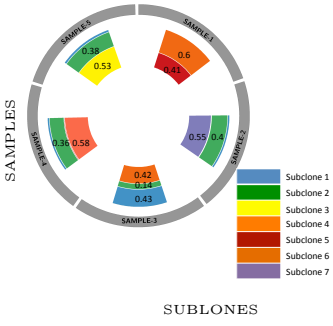
Slide 14

Results – Pancreatic Cancer

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



$S = 118$ SNV's in KEGG pathway

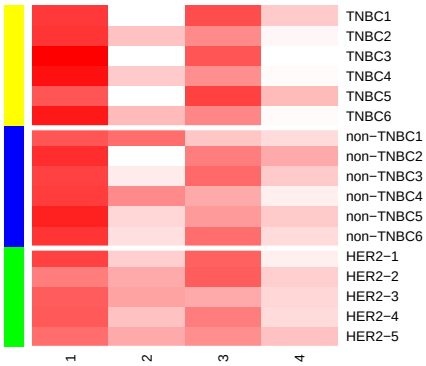


$S = 7000$ SNV's

Slide 15

Results – Breast Cancer

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



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

Slide 16

Haplotypes vs. Subclones

Haplotypes: for a diploid organism, pairs of haplotypes define a unique genome.

Prior $p(\boldsymbol{w})$**:** this implies that w_{tc} should be (a mixture of) tied in pairs, i.e., $w_{tc} = w_{tc'}$, with *pairs* defining subclones.

cIBP:

Alternatively, we use a categorical generalization of the IBP, $p(\boldsymbol{Z})$ for a random trinary matrix with $z_{sc} \in \{0, 1, 2\}$

	1	2	3	4	5
1	0.5	1	0	1	0
2	1	0.5	1	1	1
3	0.5	0	0	0	0.5
4	0.5	0	0.5	0	0.5
5	1	1	0.5	0.5	0.5
6	1	0	0.5	0	0
7	1	0	0	0	0
8	1	0.5	0	0.5	1
9	1	0.5	1	1	1
10	0.5	0	0	0	1

Slide 17

Slide 18

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.

SOFTWARE: R code at `http://www.ma.utexas.edu/users/yxu/`