Bayesian Nonparametric Statistics for Fostering Innovation and Discovery in Biomedical Research

Yanxun Xu
Department of Applied Math and Statistics
Mathematics Institute for Data Science
Division of Biostatistics and Bioinformatics
The Sidney Kimmel Comprehensive Cancer Center
Johns Hopkins University

April, BNP 2022

Outline

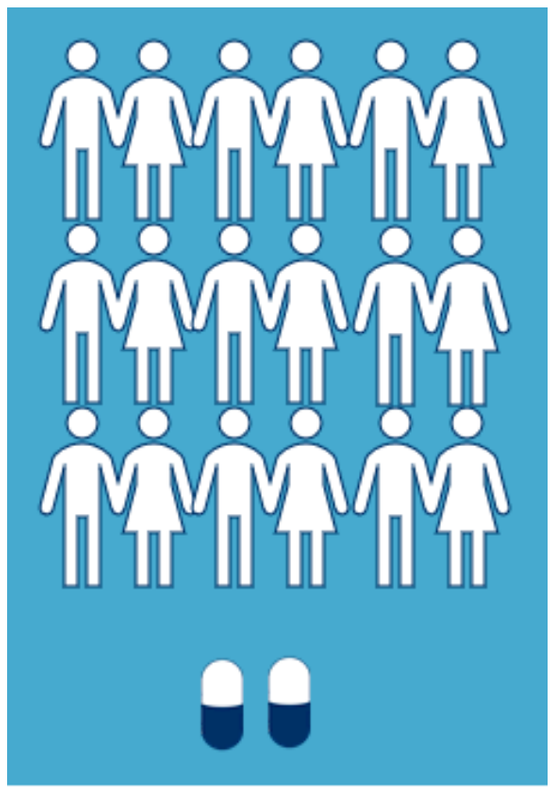
- **Part 1:** *Monday*
  - **Density estimation** for efficient clinical trial designs
  - **Regression** for precision dosing

- **Part 2:** *Wednesday*
  - **Clustering** for subgroup finding
  - **Latent feature models** for tumor heterogeneity

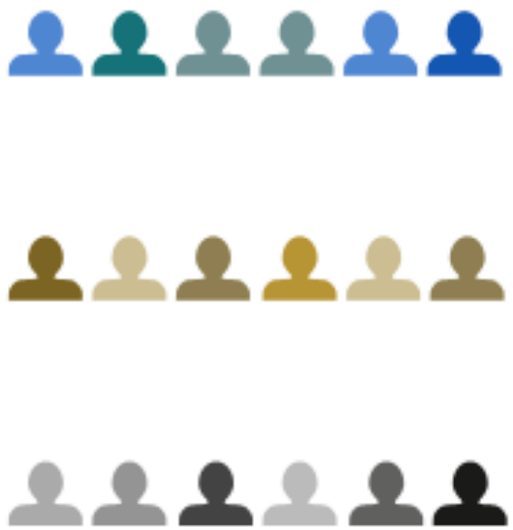
- **Part 3:** *Friday*
  - Estimating treatment effects from observational data

Treatment Effect

Clinical Trials



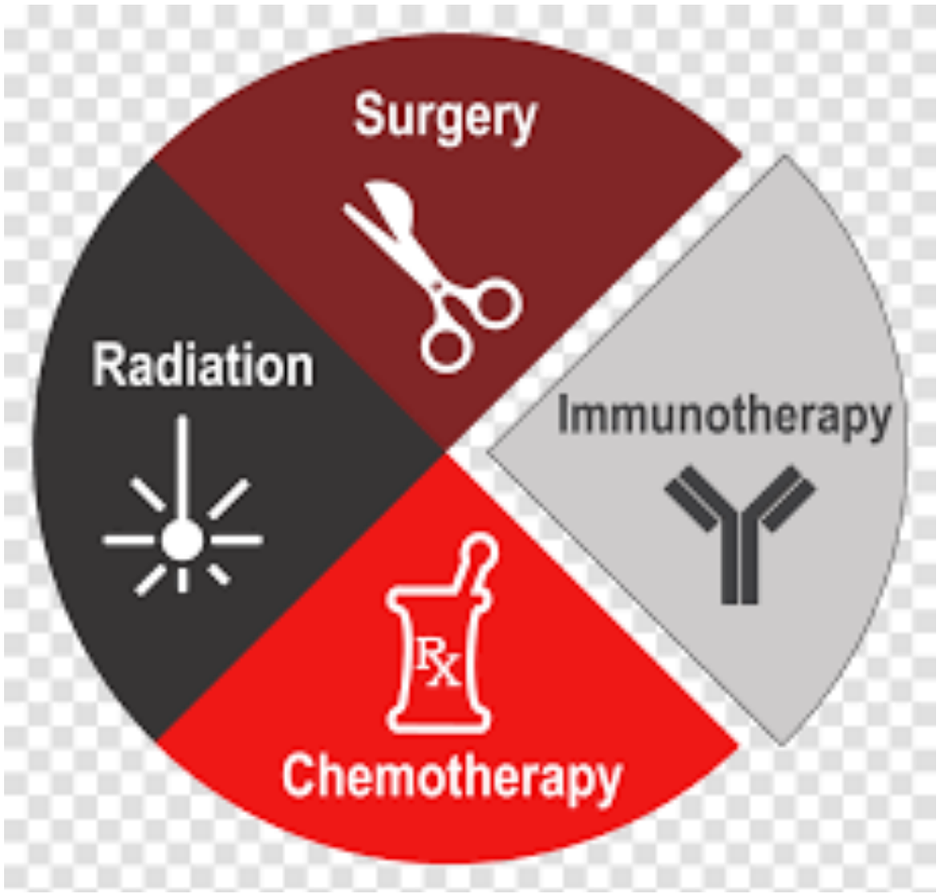
Biomarker-based Trials



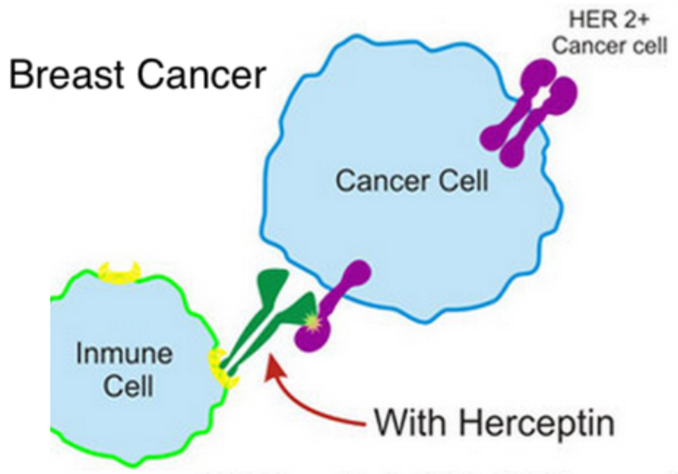
Observational Data

Population → Subgroup → Personalization

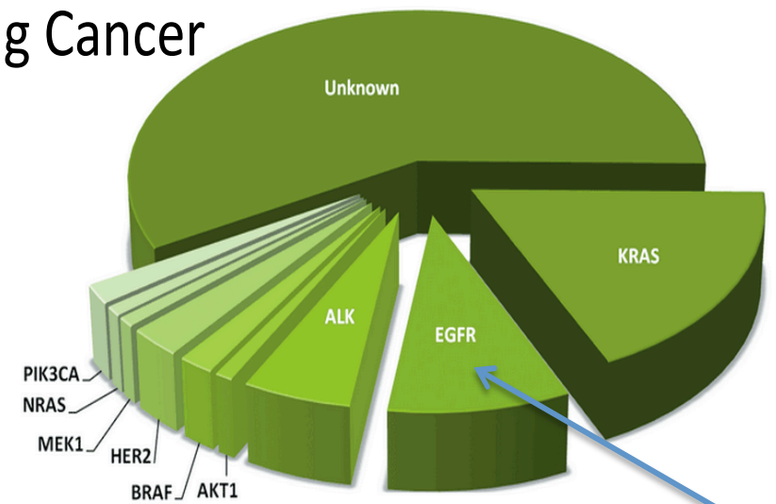
One-size Fits All Cancer Treatment



Targeted Therapy



Lung Cancer



Courtesy of Lung Cancer Foundation of America

Erlotinib

Genomic-driven Cancer Trials

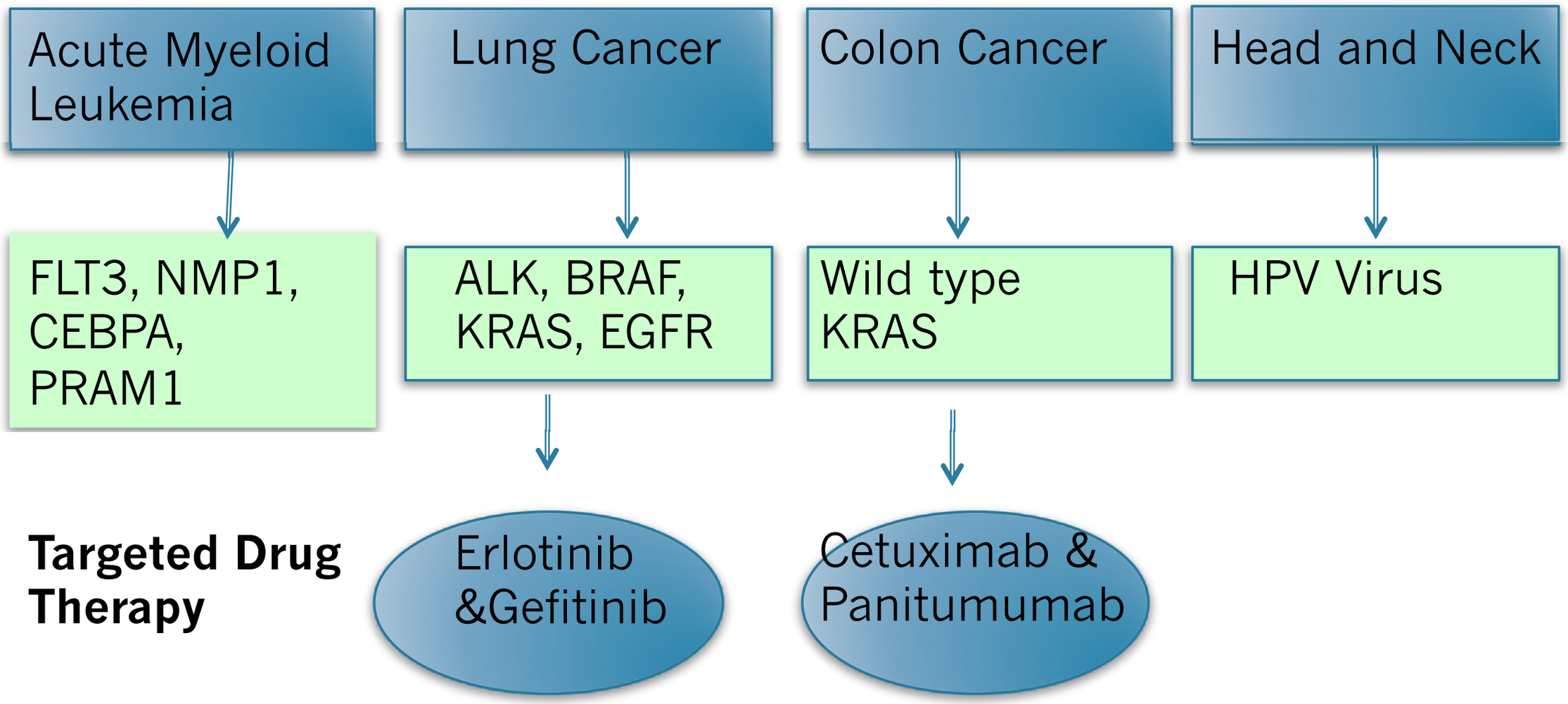
Umbrella Trials

in one **single** cancer type, test the effect of targeted agents on different alterations.

Basket Trials

across **multiple** cancer types, test the effect of targeted agents on the same genomic alternations.

Basket Trial



Motivation Trial: IMPACT II

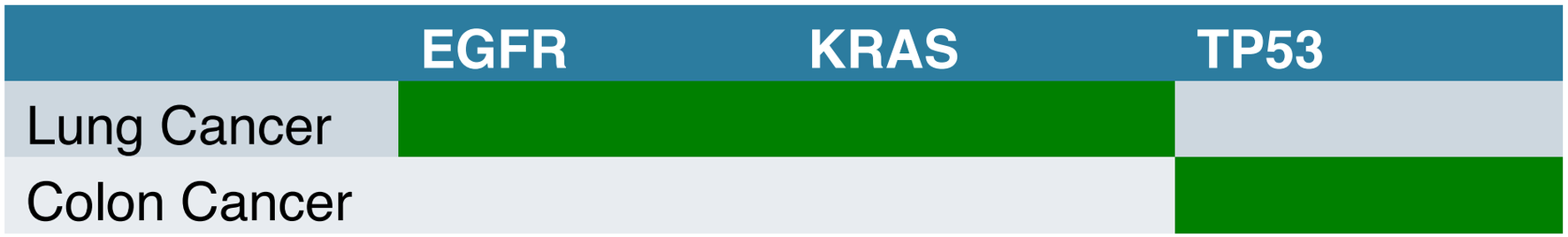
- **Clinical Trial**: study of targeted agents in metastatic cancers.
- **Patients:** with metastatic cancer (thyroid, ovarian, melanoma, lung, breast, CRC and other)
- **Treatments**: therapy that targets particular molecular aberrations (TT) vs. standard of care (S)
- **Population**: heterogeneous population; different mutations; different cancers; baseline covs . . . Treatment might be effective in a sub-population

Data:

TRT	TUMOR	PFS	CENS	MUTATIONS					
				m1	m2	m3	m4	m5	m6
TT	THYROID	2.6	0	NA	NA	NA	NA	NA	NA
TT	THYROID	3.6	0	NA	0	0	0	NA	0
S	OVARIAN	4.2	1	0	NA	0	0	0	0
S	MELANOMA	5.8	1	NA	0	0	0	NA	0
. . .	. . .	. . .	. . .						

Motivation Trial: IMPACT II

Objective: determine the subpopulation that achieves the maximum benefit from TT.



We will cast this goal as a **decision problem**.

Subpopulation Finding: Decision Problem

- Outcome: progression free survival (PFS) time, $y_i, i = 1,\ldots,n$
- Action: report a subgroup of patients who might benefit from the TT. A set of **mutation-tumor pairs**,

$$A = \{a : a = (j_a, c_a)\}$$

- $j_a = \{1,\ldots,q\}$: Molecular aberration
- $c_a \in \{1,\ldots,n_c\}$: tumor type

{(KRAS, Lung), (TP53, Breast)}

Subpopulation Finding: Decision Problem

- Action: report a subgroup of patients who might benefit from the TT. A set of **mutation-tumor pairs**,

$$A = \{a : a = (j_a, c_a)\}$$

Bayes Rule: $A^* = \text{argmax}_A \int u(A, \theta) p(\theta \mid y, X) d\theta$

Utility: we favor a subpopulation with difference in log hazards ratio (LR) and large size

Data from IMPACT

- Outcome: progression free survival times, y_i
- Covariates: $x_i = (c_i, m_i, b_i)$
  - Tumor type c_i (categorical)
  - Molecular aberrations $m_i = (m_{i1}, \ldots, m_{iM})$ (binary)
  - Other baseline covariates b_i (age, # prior therapies, etc)

Challenges

Probability model needs to allow for:

- interactions of covariates
- heterogeneous population
- missing data
- Extrapolation with small # observations

BNP!

Random Partition

- $s = (s_1, \ldots, s_n)$ be cluster membership indicators, $s_i \in \{1,\ldots,J\}$
- $S_j = \{i : s_i = j\}$

Product partition model: $p(s) \propto \prod_{j=1}^{J} c(S_j)$

For DP, $c(S_j) = \alpha(|S_j| - 1)!$

Random Partition

- $s = (s_1, \ldots, s_n)$ be cluster membership indicators, $s_i \in \{1,\ldots,J\}$
- $S_j = (i : s_i = j)$
- x_j^* by cluster

Product partition model with covariates (PPMx):

$$p(s \mid x) \propto \prod_{j=1}^{J} c(S_j) g(x_j^*)$$

Favors clusters homogeneous in x_i with $g(x_j^*)$ scoring similarity of $x_j^* = \{i : s_i = j\}$.

Mueller et al. (2011 JCGS), Quintana et al. (2015 StandJS)

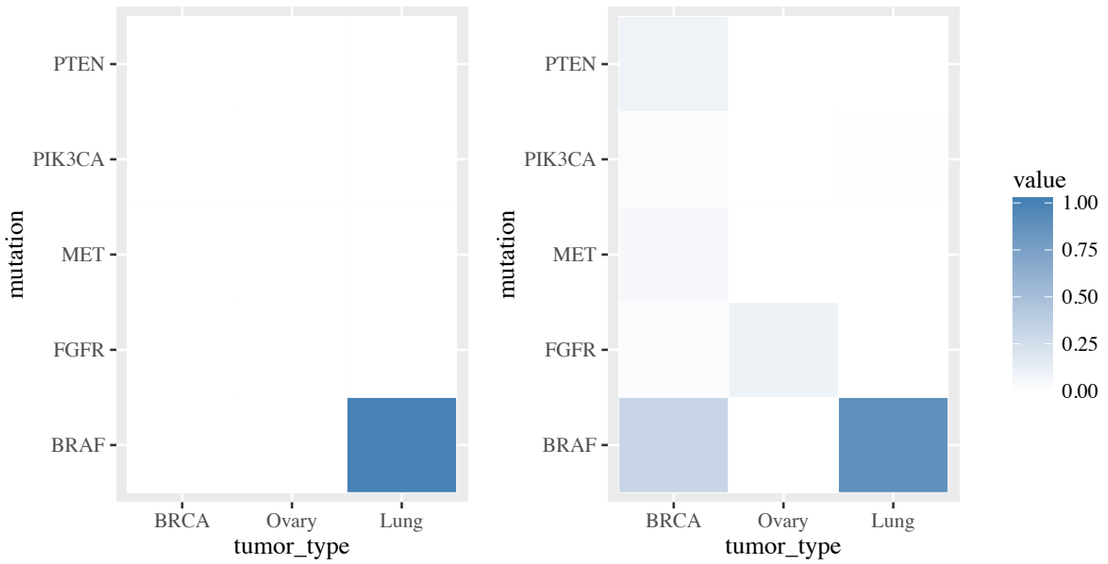
Similarity function: over observed covariates only

$$g(x_j^*) = \prod_{l=1}^{p} g_l(\{x_{il}, i \in S_j \text{ and } x_{il} \text{ observed}\}$$

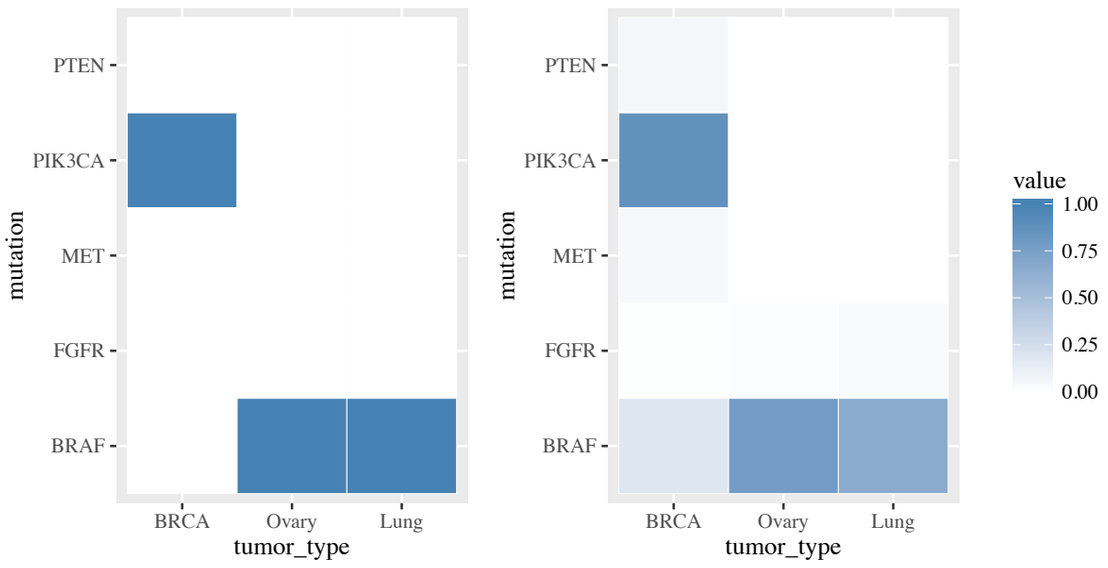
Sampling model: exchangeable within clusters (e.g., lognormal regression model)

$$p(y \mid s, x, \eta) = \prod_{j=1}^{J} \prod_{i \in S_j} p(y_i \mid \eta_j)$$

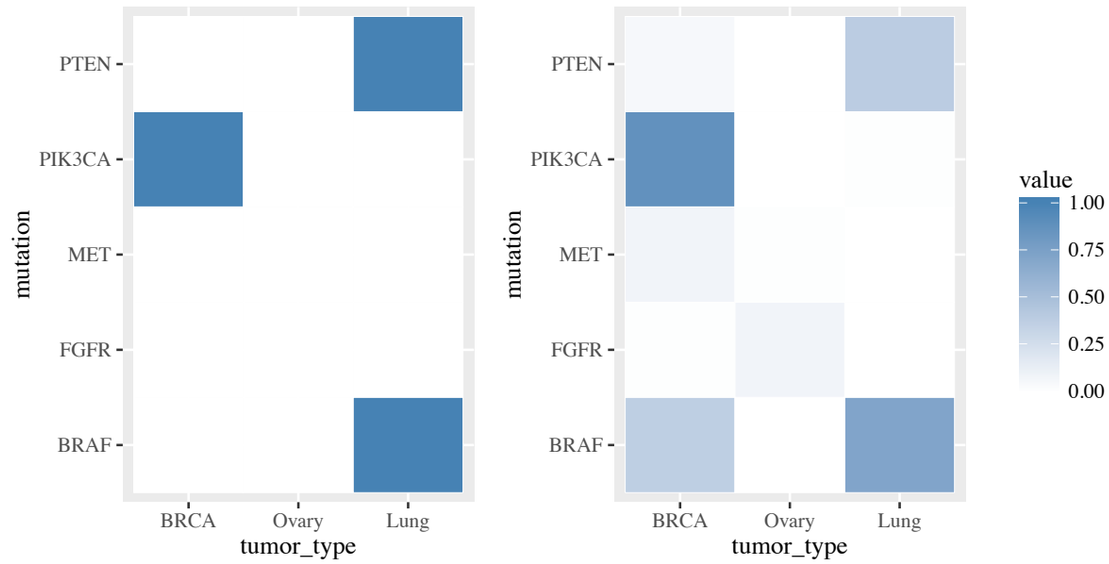
Results



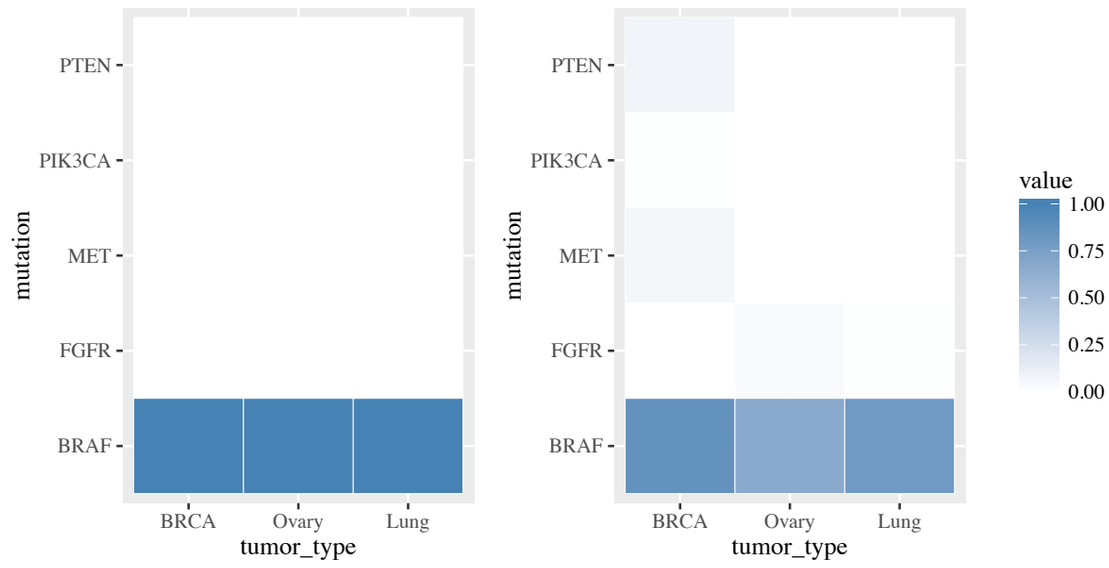
Scenario 3



Scenario 4



Scenario 5



Scenario 6

Take-home Messages

- A general class of probability models that allow for interactions and missing data
- Subgroup finding can be casted as a decision problem.
- Separate the decision problem with probability model
- Can be used in clinical trial designs to adaptively assign patients

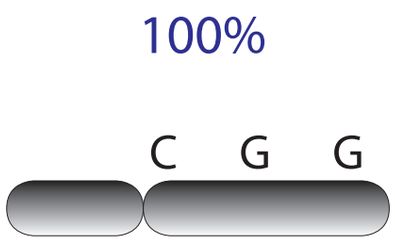
Outline

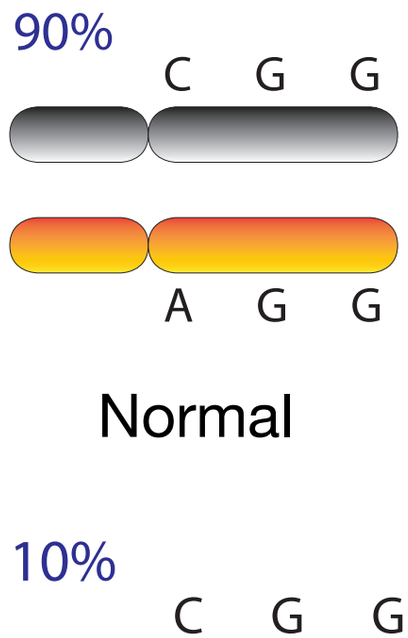
- **Part 1:** *Monday*
  - **Density estimation** for efficient clinical trial designs
  - **Regression** for precision dosing

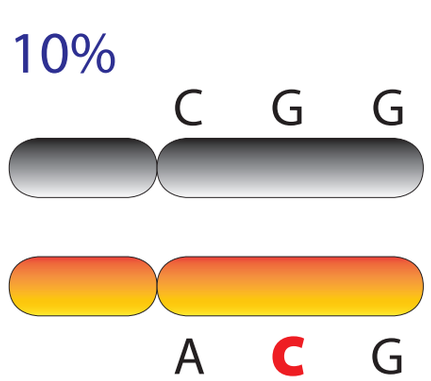
- **Part 2:** *Wednesday*
  - **Clustering** for subgroup finding
  - **Latent feature models** for tumor heterogeneity

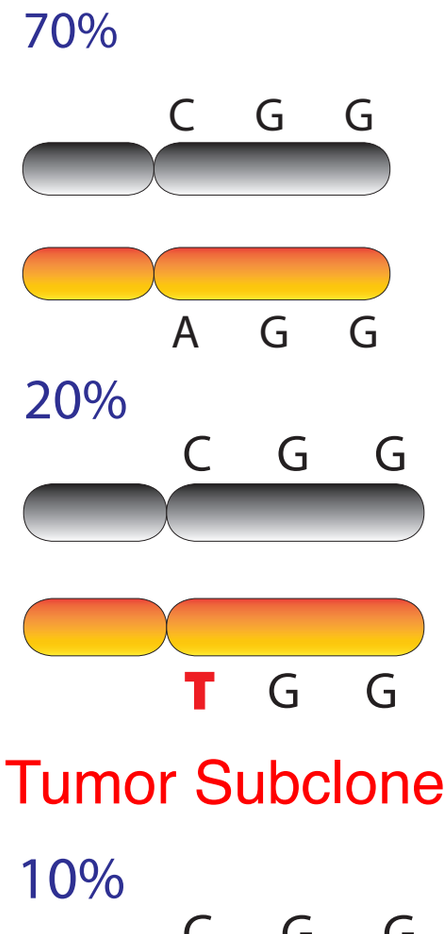
- **Part 3:** *Friday*
  - Estimating treatment effects from observational data

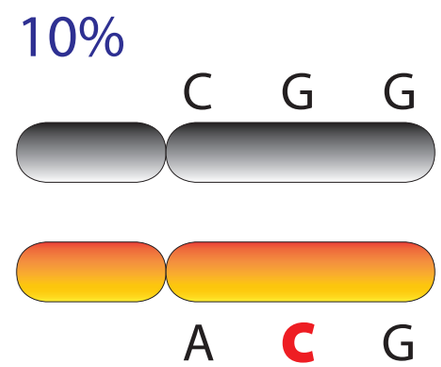
Tumor Heterogeneity (TH)

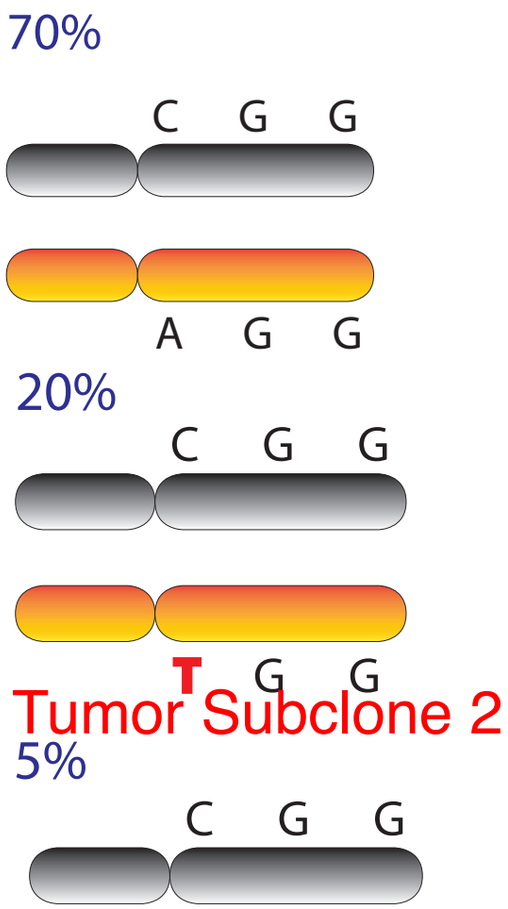


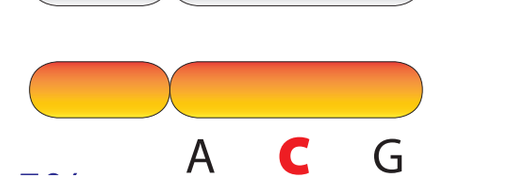






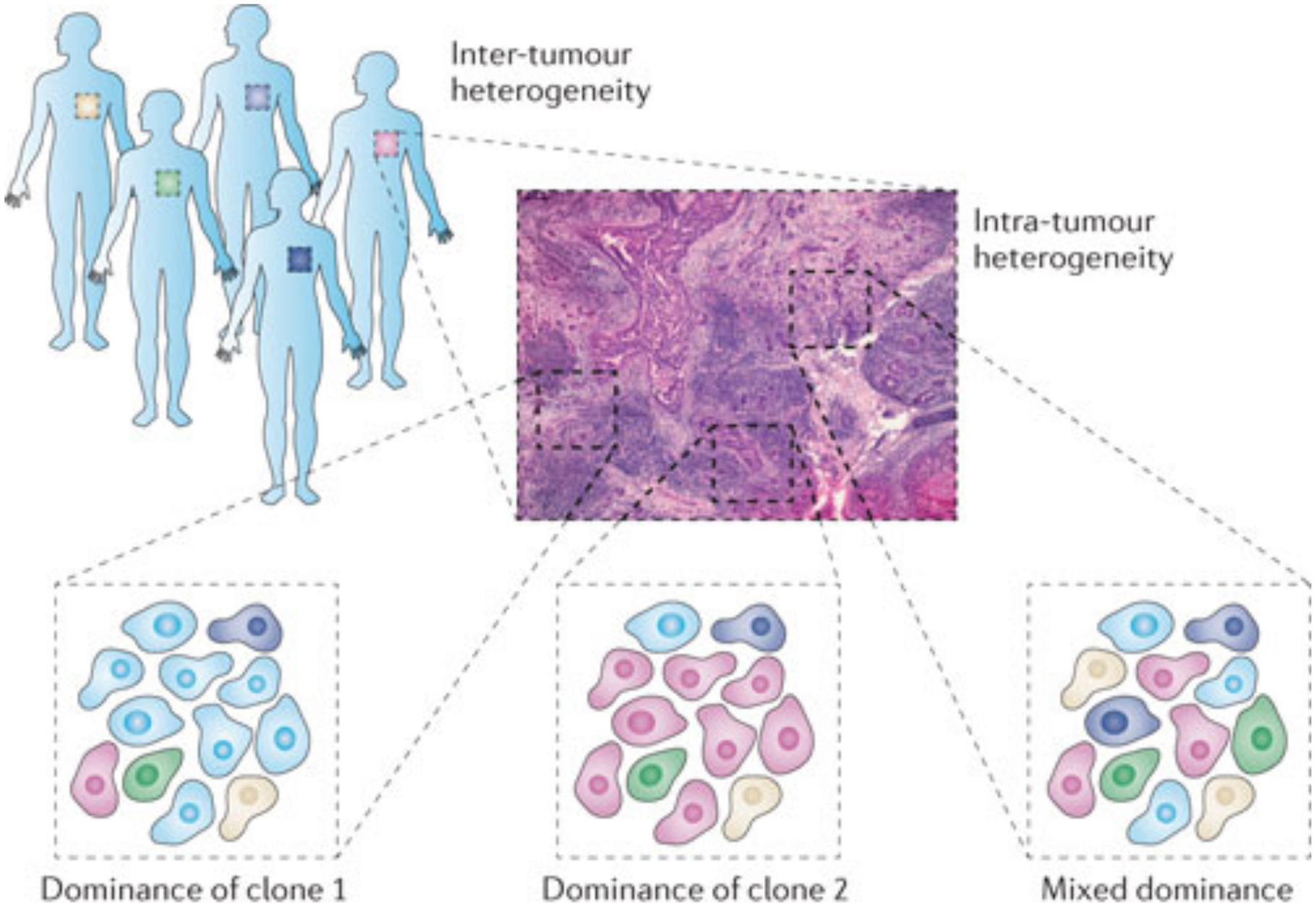






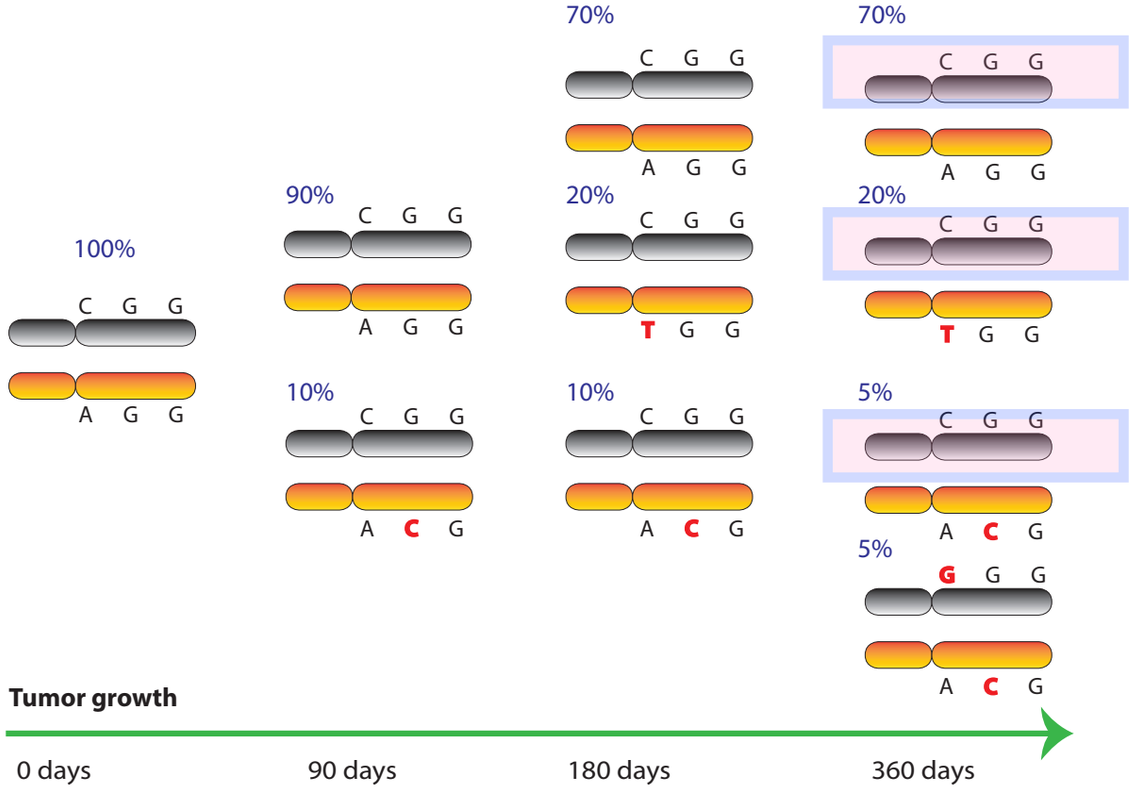
Life History of a Tumor

Clinical Utility of TH

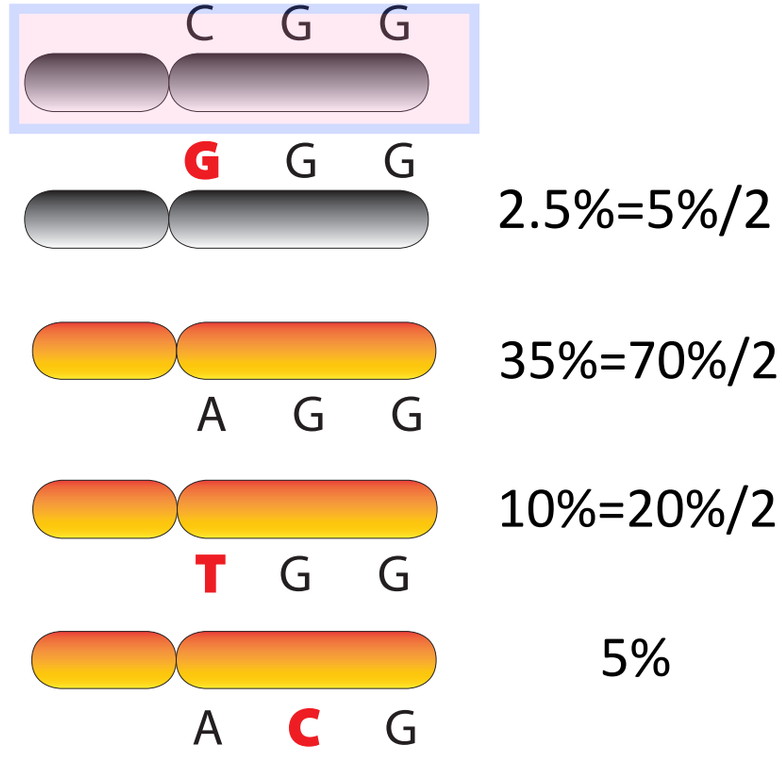


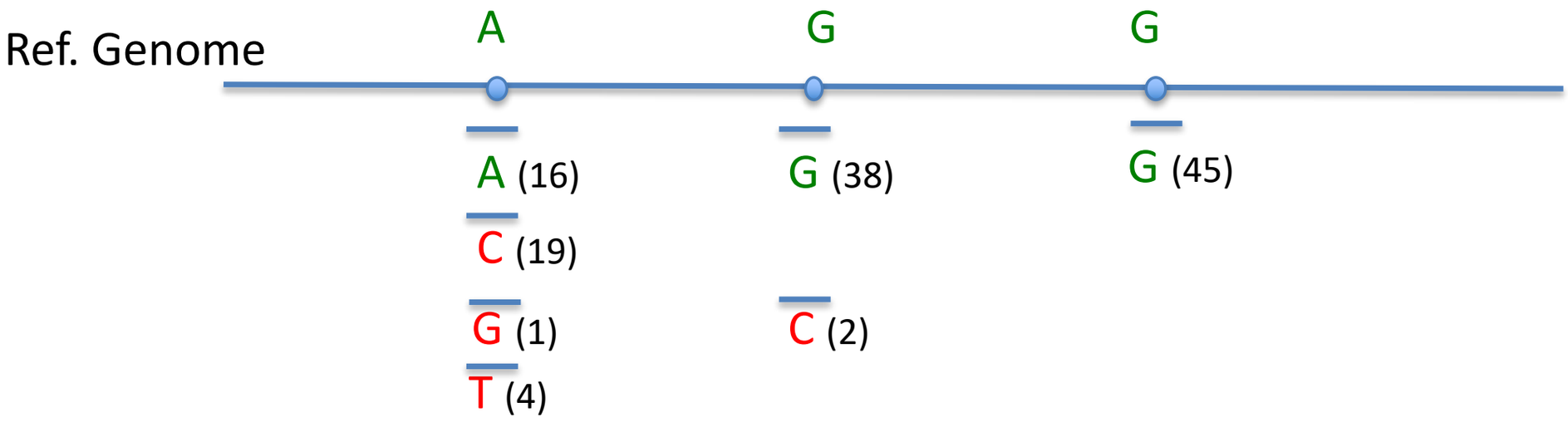
Nature Reviews | Cancer

Haplotype



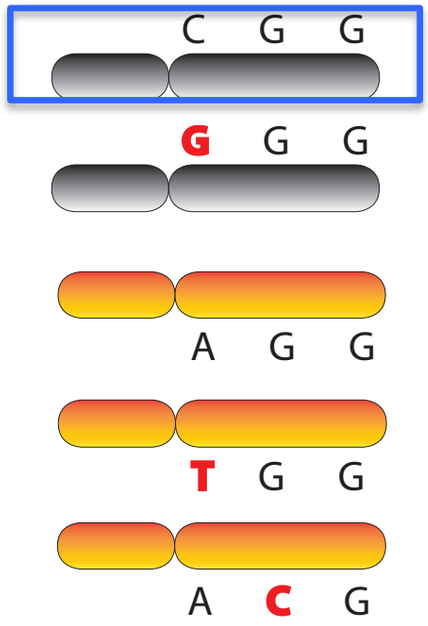
47.5%=(70%+20%+5%)/2





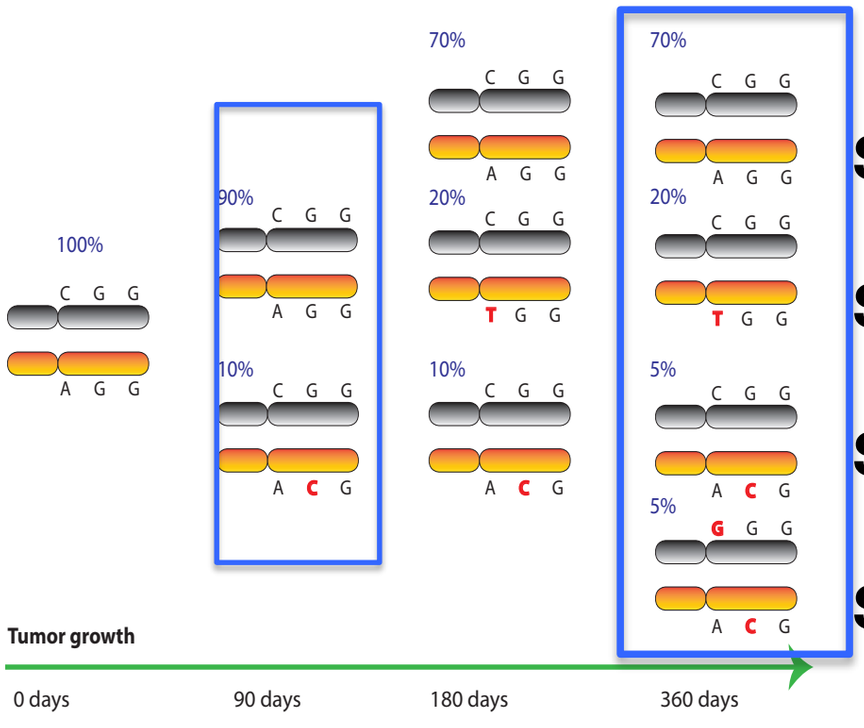
47.5% (CGG) + **2.5** (GGG) + **35%** (AGG) + **10%** (TGG) + **5%** (ACG)

Tumor Heterogeneity in Terms of Haplotype Genome (Z) and Cellular Fractions (W)



	Hap1	Hap2	Hap3	Hap4	Hap5
SNV1	1	1	0	1	0
SNV2	0	0		0	1
SNV3	0	0	0	0	0

The ***Z*** Matrix



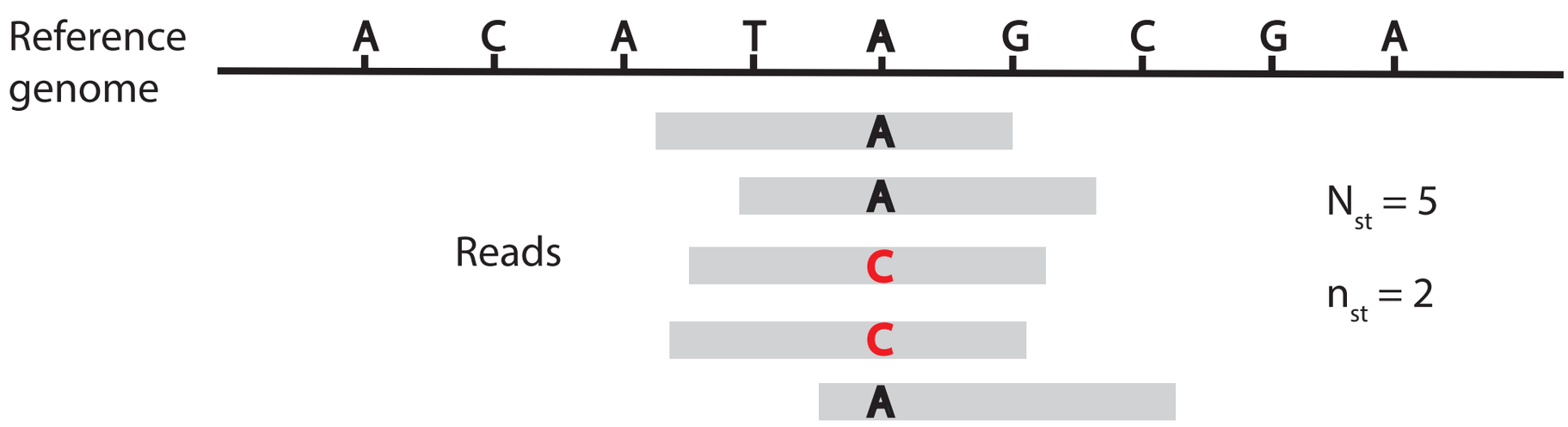
	Hap1	Hap2	Hap3	Hap4	Hap5
Sample 1	47.5%	2.5%		10%	5%
Sample 2	50%	0	35	10%	5%
Sample 3	50%	0	%	0	5%
Sample 4	50%	0	%	0	0

The ***W*** Matrix

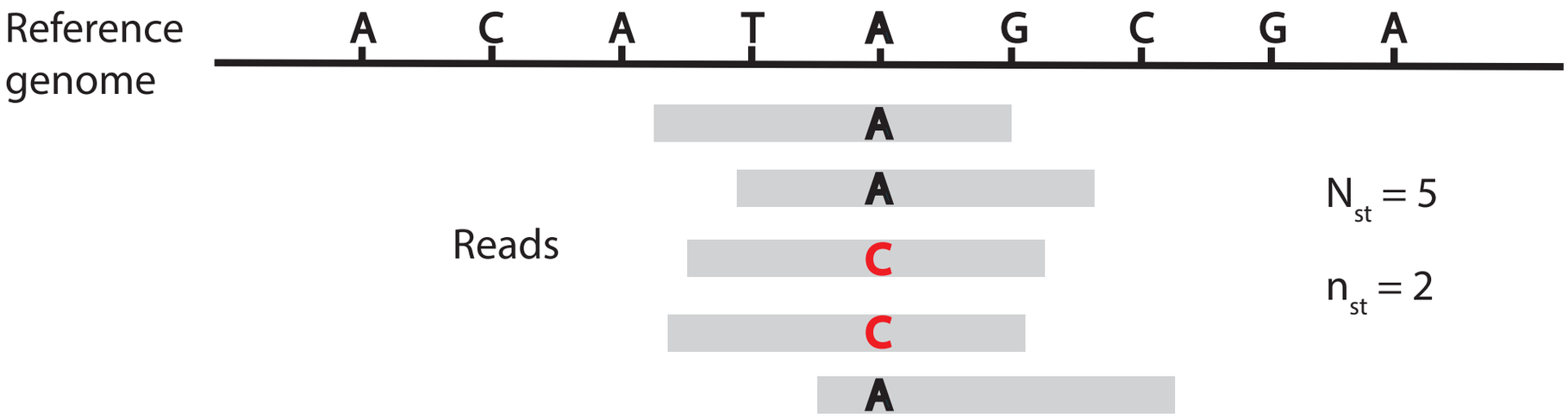
Notations

- SNV: point mutations, $s = 1,\ldots,S$
- Sample: $t = 1,\ldots,T$
- Data: N_{st} = # reads mapped to locus of SNV s in sample t

  n_{st} = # of them with SNV.



Sampling Model

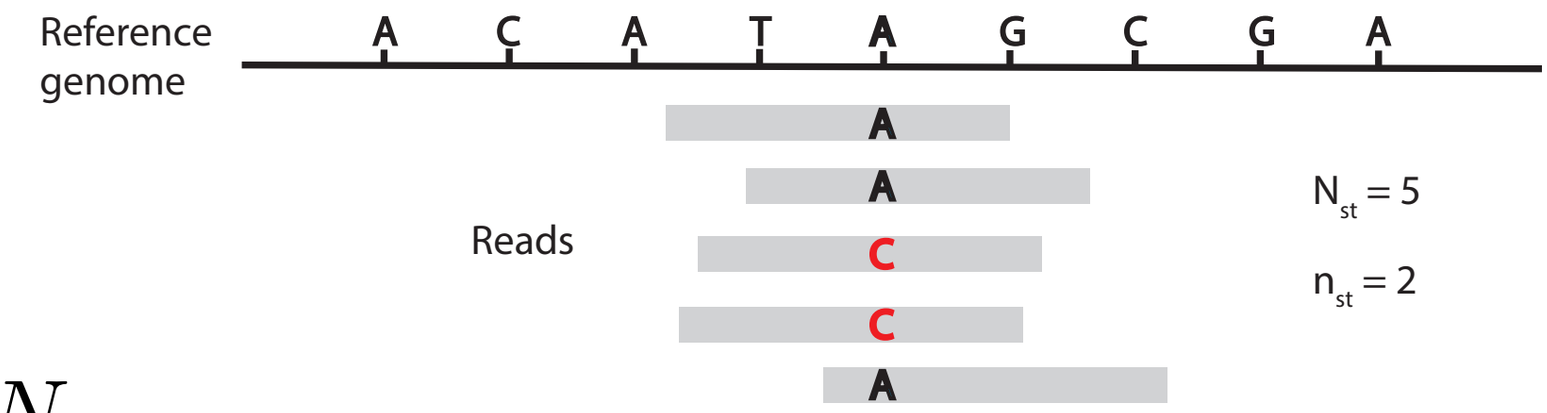


$$n_{st} \sim \text{Binomial}(N_{st}, p_{st})$$

VAF: variant allele fraction

$$\text{Observed VAF: } n_{st}/N_{st}$$
$$\text{Expected VAF: } p_{st} = E(n_{st}/N_{st})$$

Link VAFs with Haplotypes



Observed VAF: n_{st}/N_{st}

Expected VAF: $p_{st} = E(n_{st}/N_{st})$

Key Idea: A variant read must be from a haplotype with variant.

Link VAFs with Haplotypes

Key Idea: A variant read must be from a haplotype with variant.

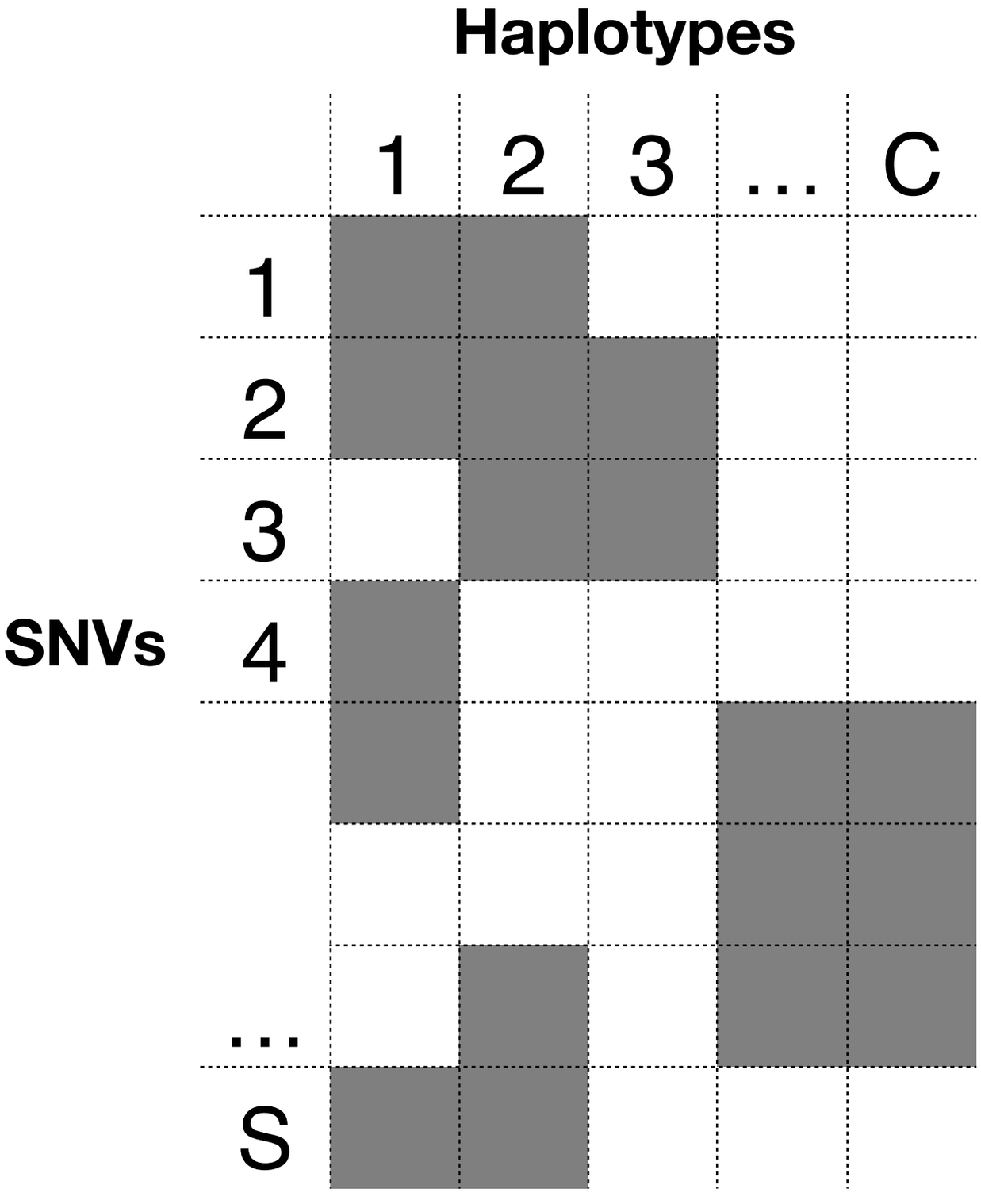
s: SNV; c: haplotype (latent); t: sample

$z_{sc} = 1$: haplotype c has a variant on SNV s.
$z_{sc} = 0$: haplotype c has no variant on SNV s.
w_{tc}: fraction of haplotype c in sample t.

Linking Equation:

$$p_{st} = \sum_c w_{tc} z_{sc}$$

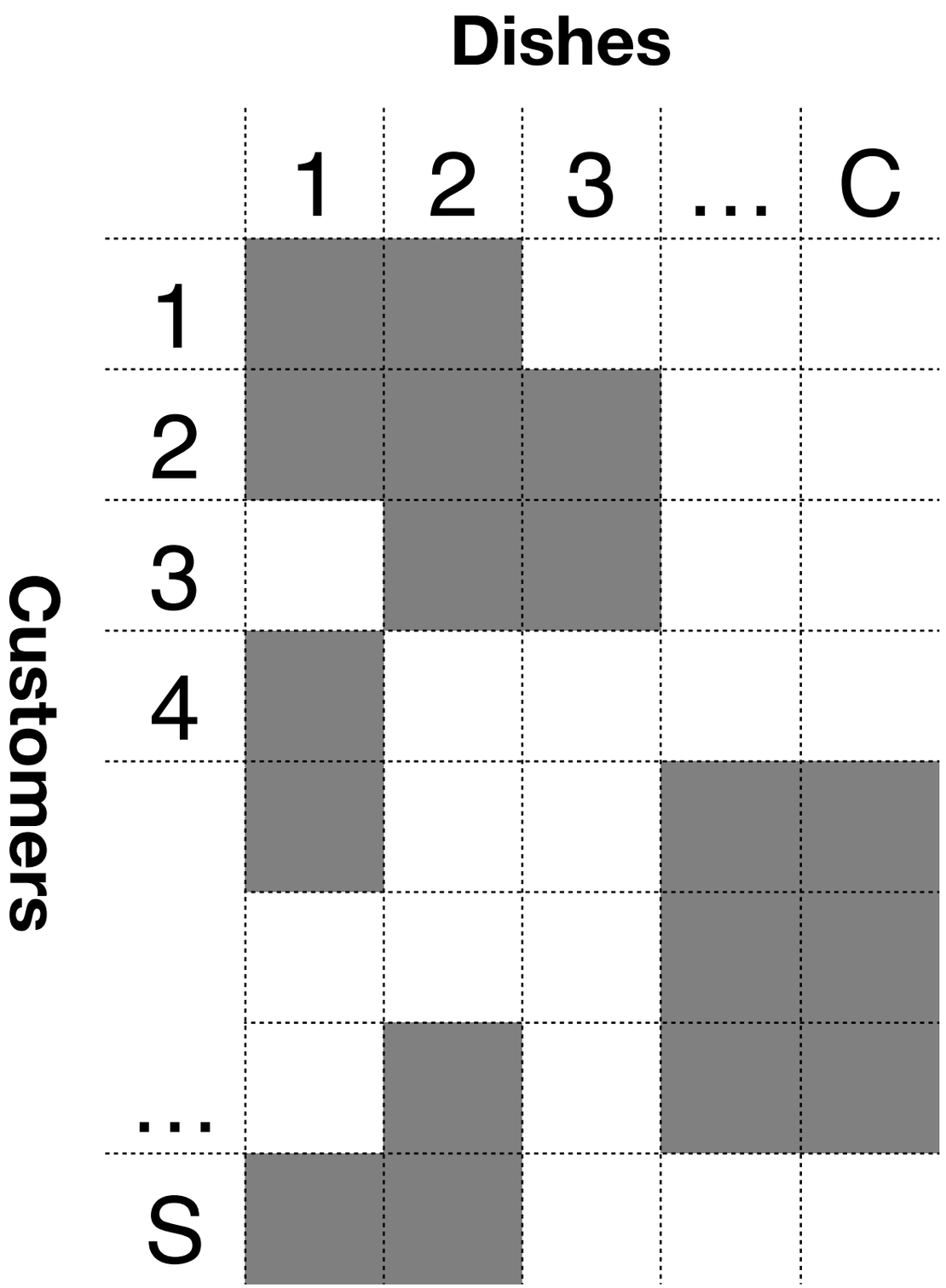
Haplotype Genotype Z



$p(Z)$ on $(S \times C)$ binary matrix

Indian Buffet Process (IBP)

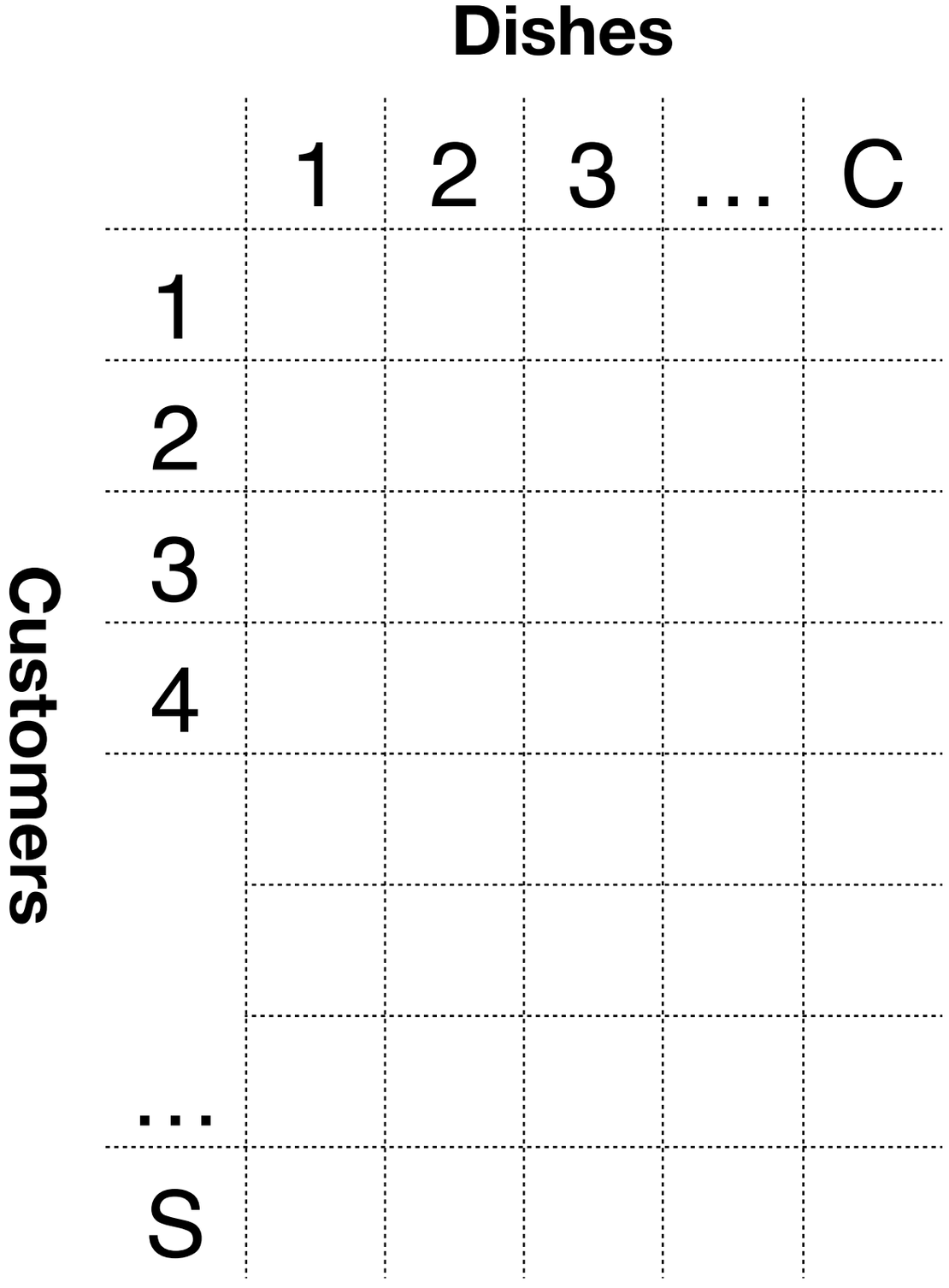
Indian Buffet Process (IBP)



For $s = 1, \ldots, S$

- Customer s chose dish c that has been already chosen m_k time with probability m_k / s
- Number of new dishes: $K_s \sim \text{Poisson}(\gamma / \text{s})$

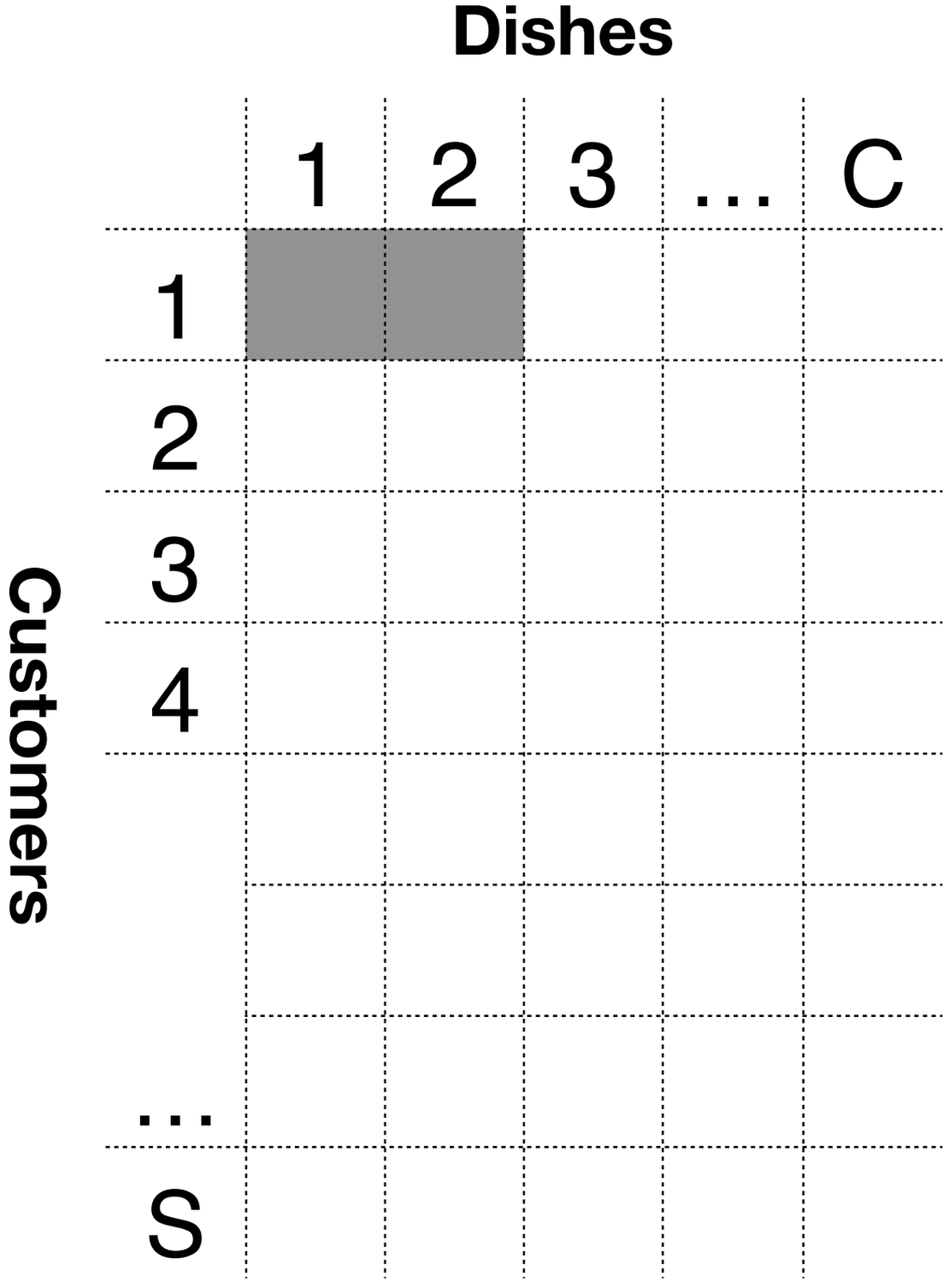
Indian Buffet Process (IBP)



For $s = 1, \ldots, S$

- Customer s chose dish c that has been already chosen m_k time with probability m_k / s
- Number of new dishes: $K_s \sim \text{Poisson}(\gamma / \text{s})$

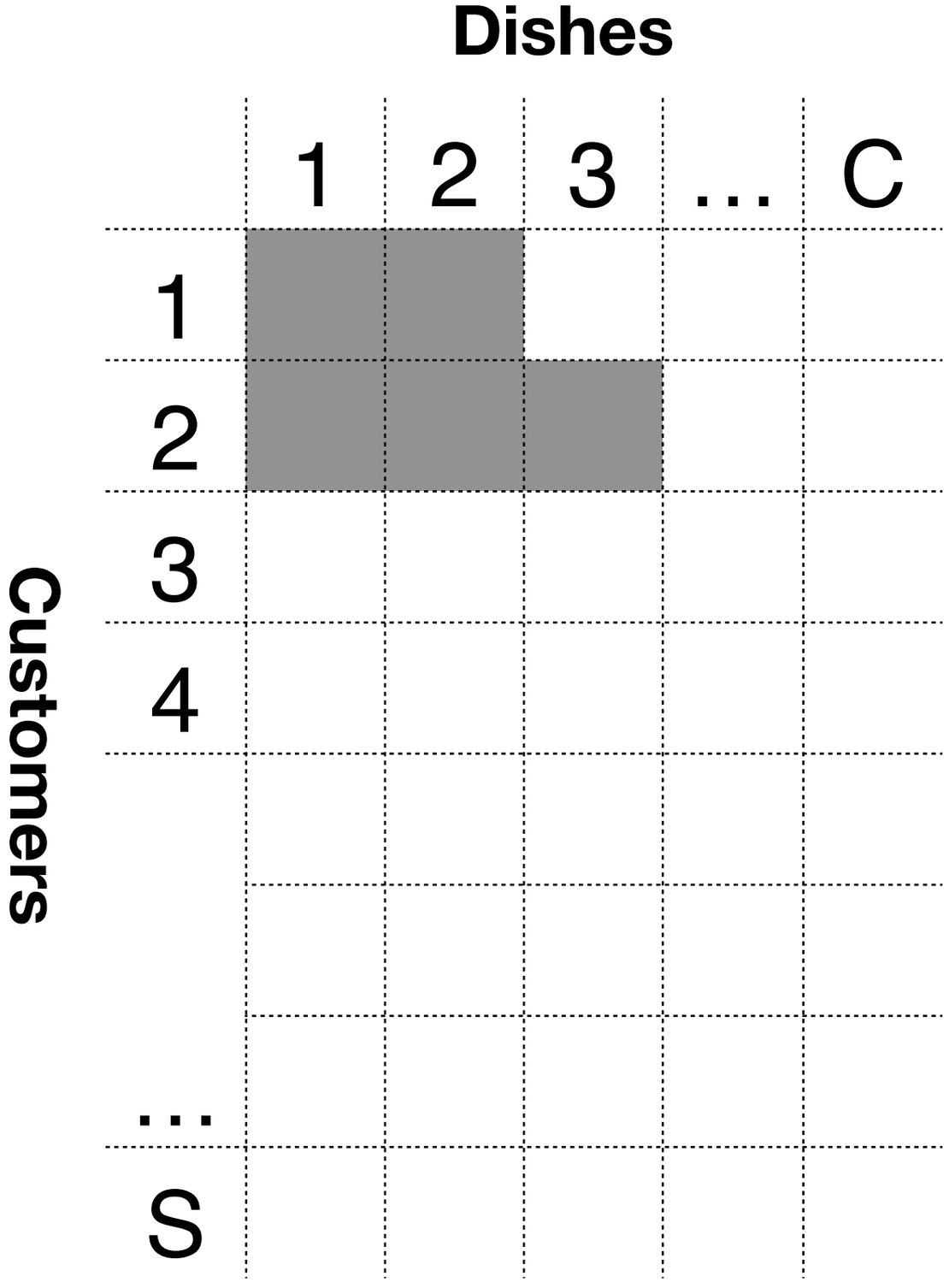
Indian Buffet Process (IBP)



For $s = 1, \ldots, S$

- Customer s chose dish c that has been already chosen m_k time with probability m_k / s
- Number of new dishes: $K_s \sim \text{Poisson}(\gamma / \text{s})$

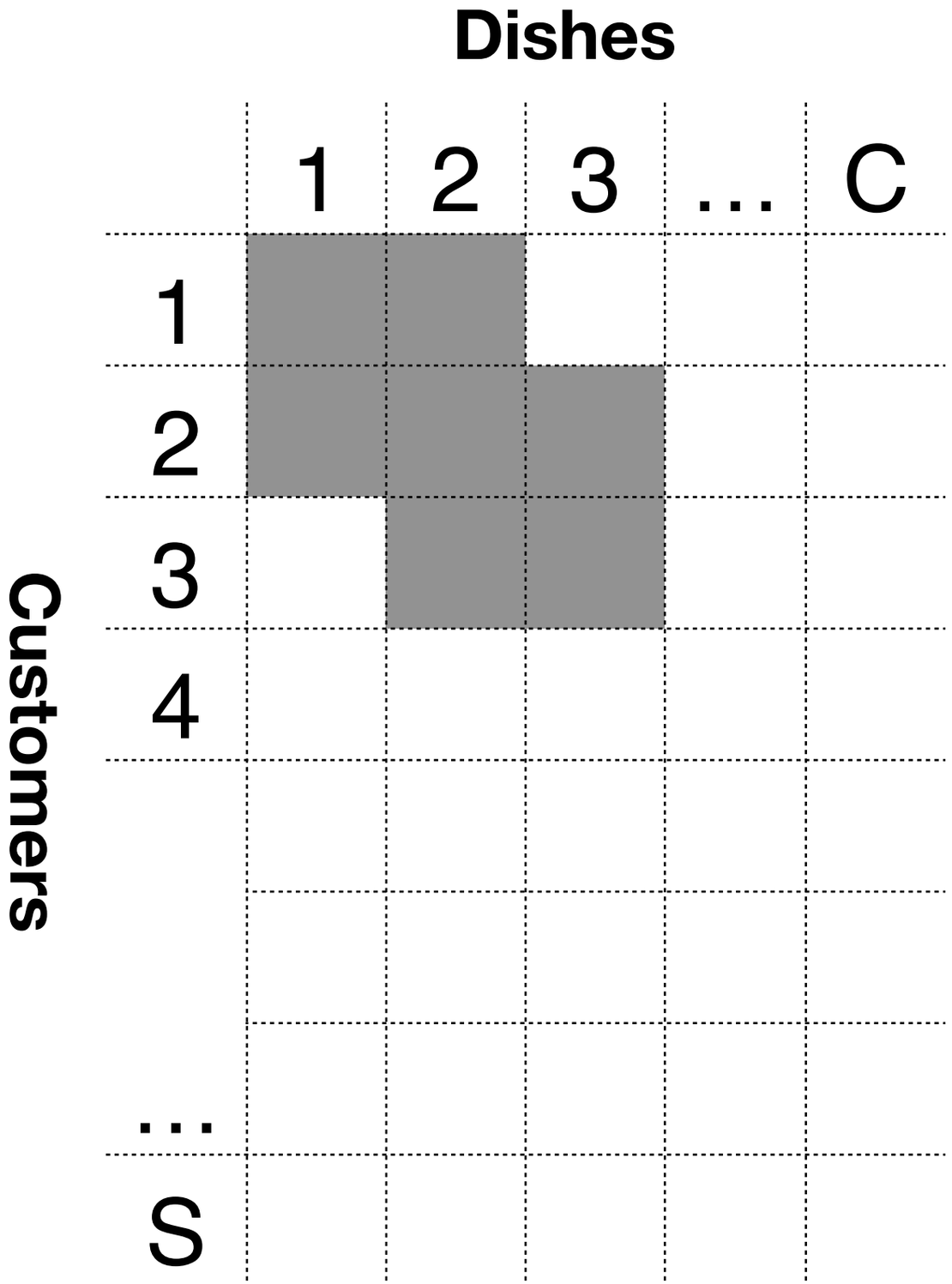
Indian Buffet Process (IBP)



For $s = 1, \ldots, S$

- Customer s chose dish c that has been already chosen m_k time with probability m_k / s
- Number of new dishes: $K_s \sim \text{Poisson}(\gamma / \text{s})$

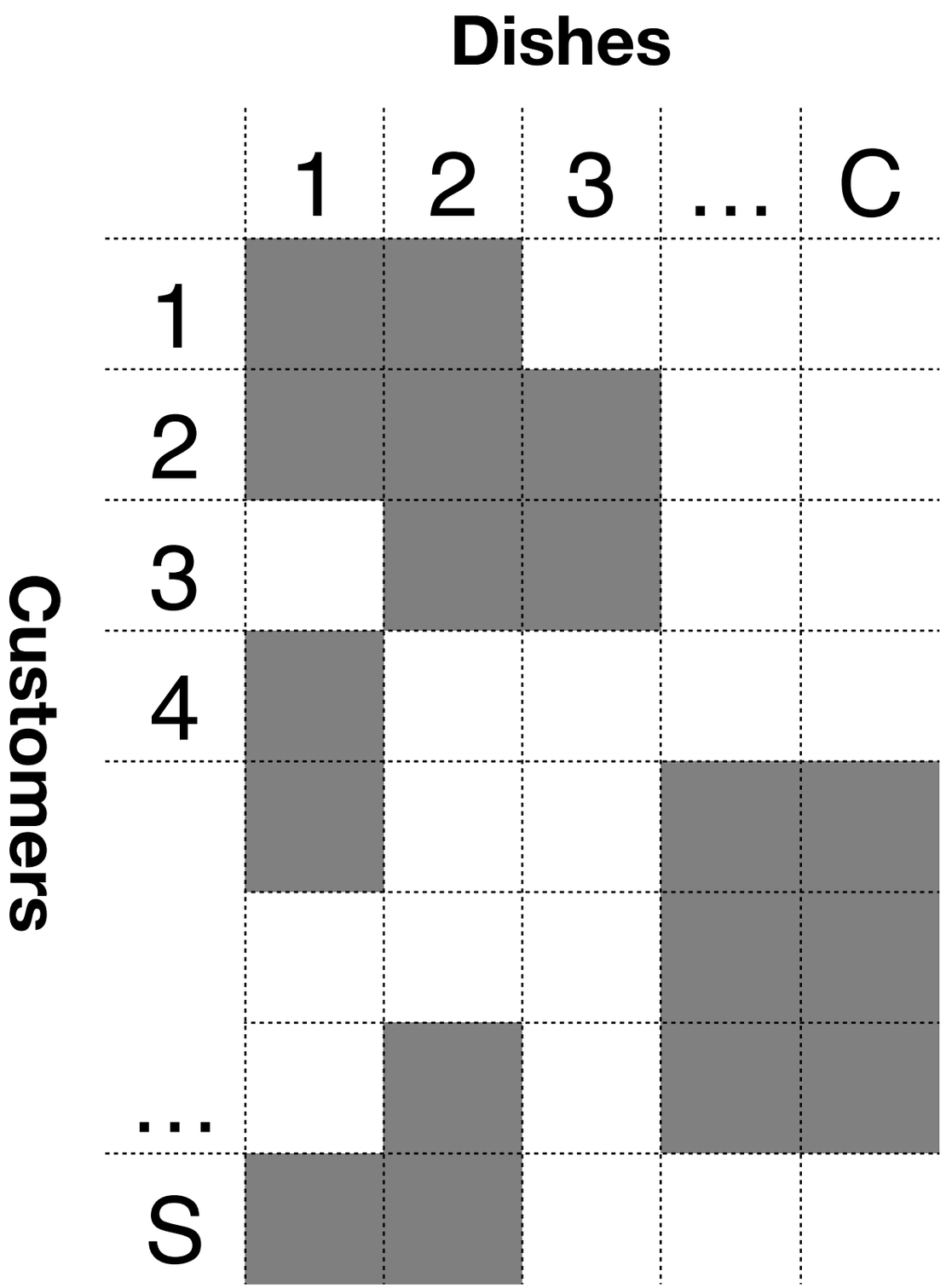
Indian Buffet Process (IBP)



For $s = 1, \ldots, S$

- Customer s chose dish c that has been already chosen m_k time with probability m_k / s
- Number of new dishes: $K_s \sim \text{Poisson}(\gamma / \text{s})$

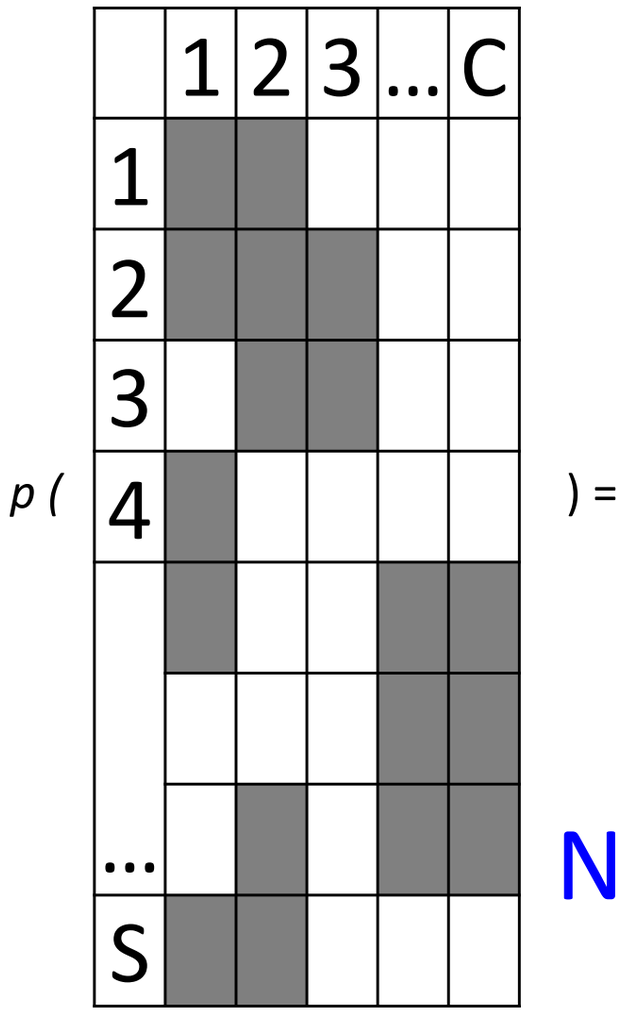
Indian Buffet Process (IBP)



For $s = 1, \ldots, S$

- Customer s chose dish c that has been already chosen m_k time with probability m_k / s
- Number of new dishes: $K_s \sim \text{Poisson}(\gamma / \text{s})$

IBP Prior



$p($ $)=$

customers (SNVs) who chose dish c (haplotype)

Number of customers (SNVs)

Number of dishes (haplotypes)

$$p(\mathbf{Z} \mid \gamma) = \frac{\gamma^C \, e^{-\gamma H_S}}{C!} \prod_{c=1}^{C} \frac{(S - m_c)!(m_c - 1)!}{S!}$$

Model Summary

$$p(Z, w, n \mid N) = \underbrace{p(Z)}_{\text{IBP}}\, p(w \mid Z)\, \underbrace{p(n \mid Z, w, N)}_{\text{Binomial}}.$$

$$p_{st} = \sum_{c} w_{tc} z_{sc}$$

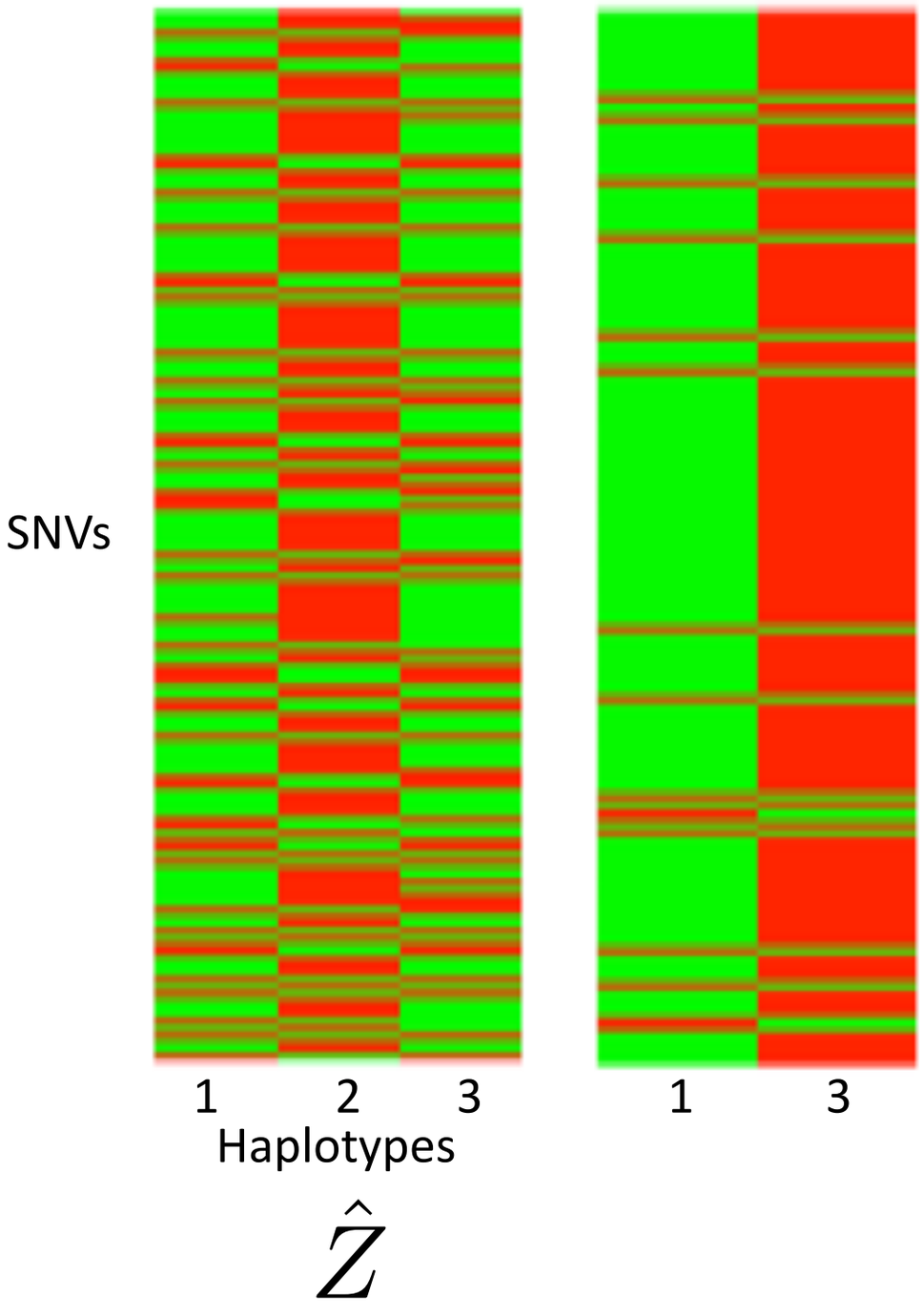
$$p(w_t) \sim \text{Dir}(a_1, \ldots, a_C),\ t = 1, \ldots, T.$$

$$p(Z, w \mid N, n)$$

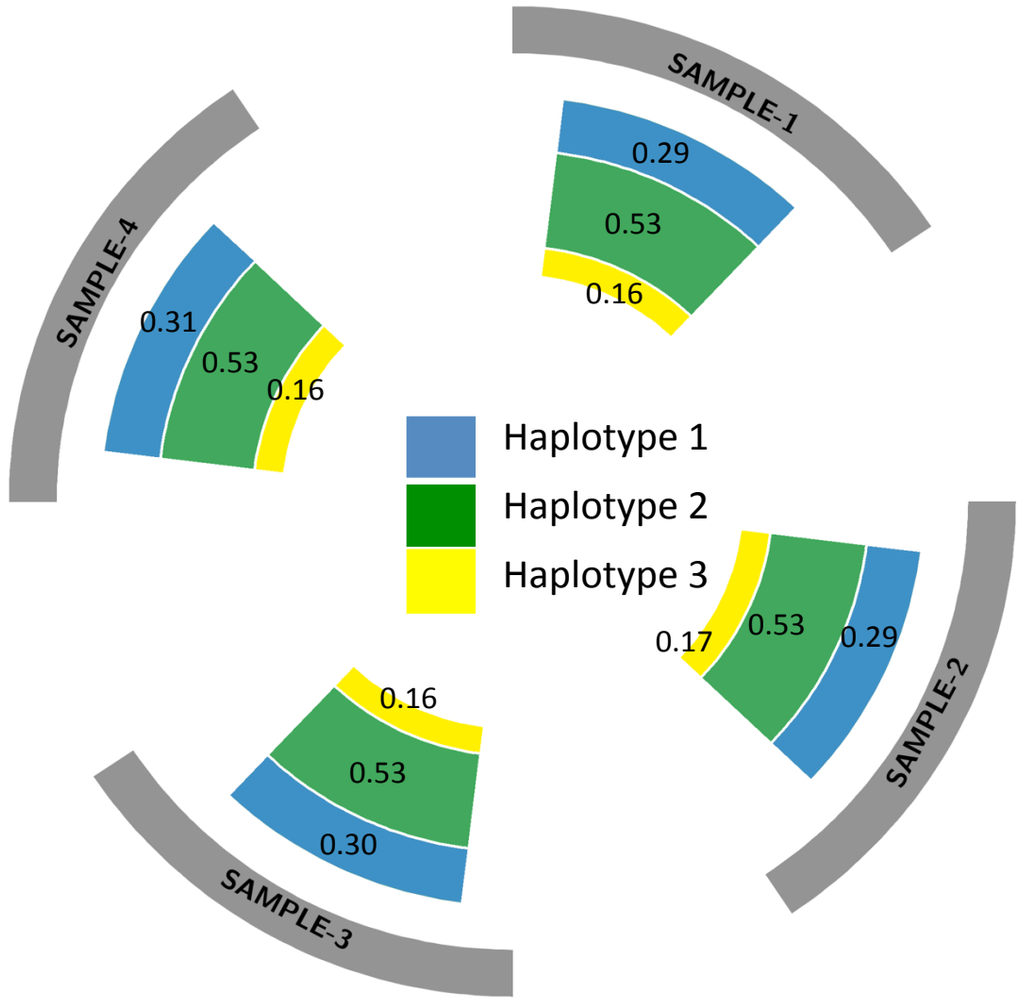
Application: Intra-Tumor Heterogeneity

- One tumor from lung cancer; 4 samples surgically dissected
- Each sample generates a whole-genome sequencing data set
- Bio-X pipeline (BWA, Samtools, GATK) for data preprocessing: coverage ~ 100X.
- Selected S=17,160 SNVs

Application: Intra-Tumor Heterogeneity



$\hat{Z}$

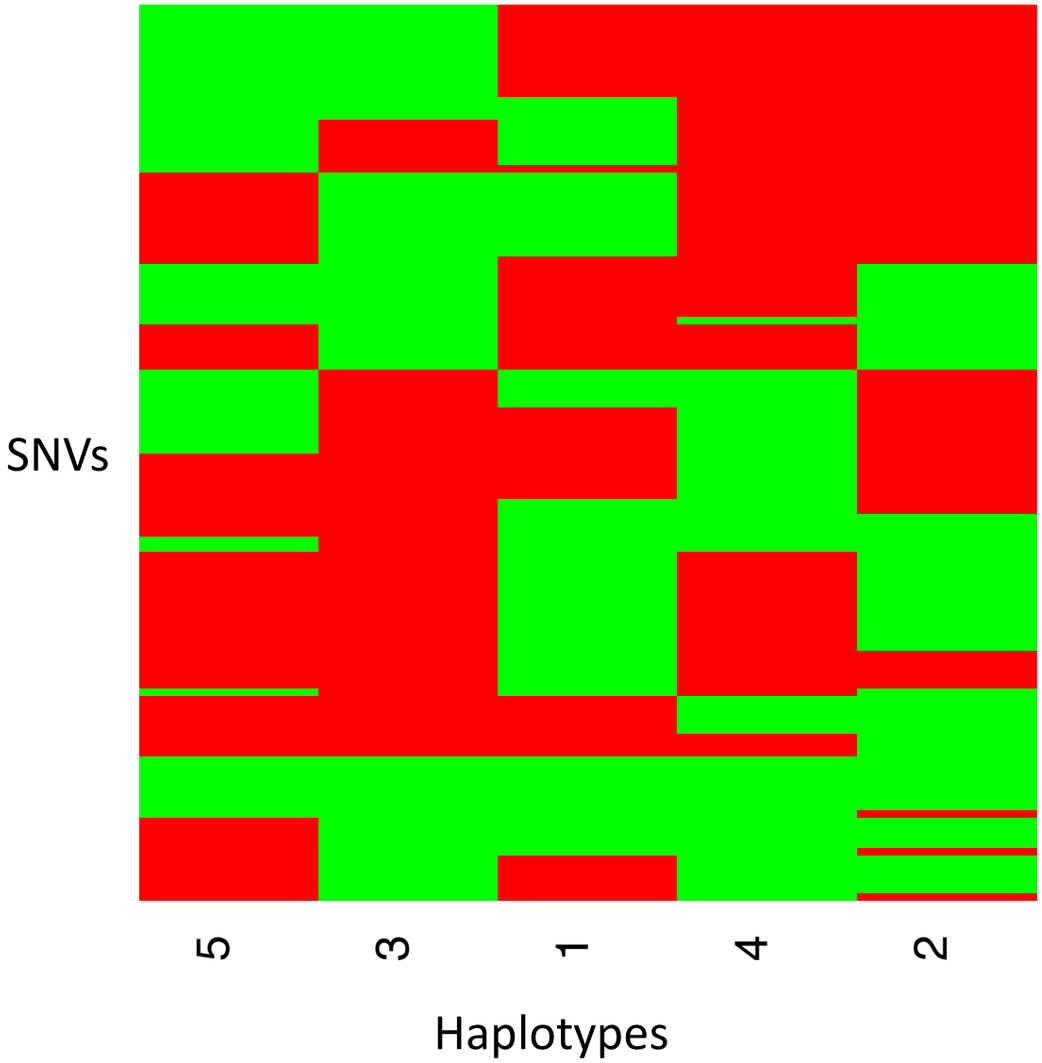


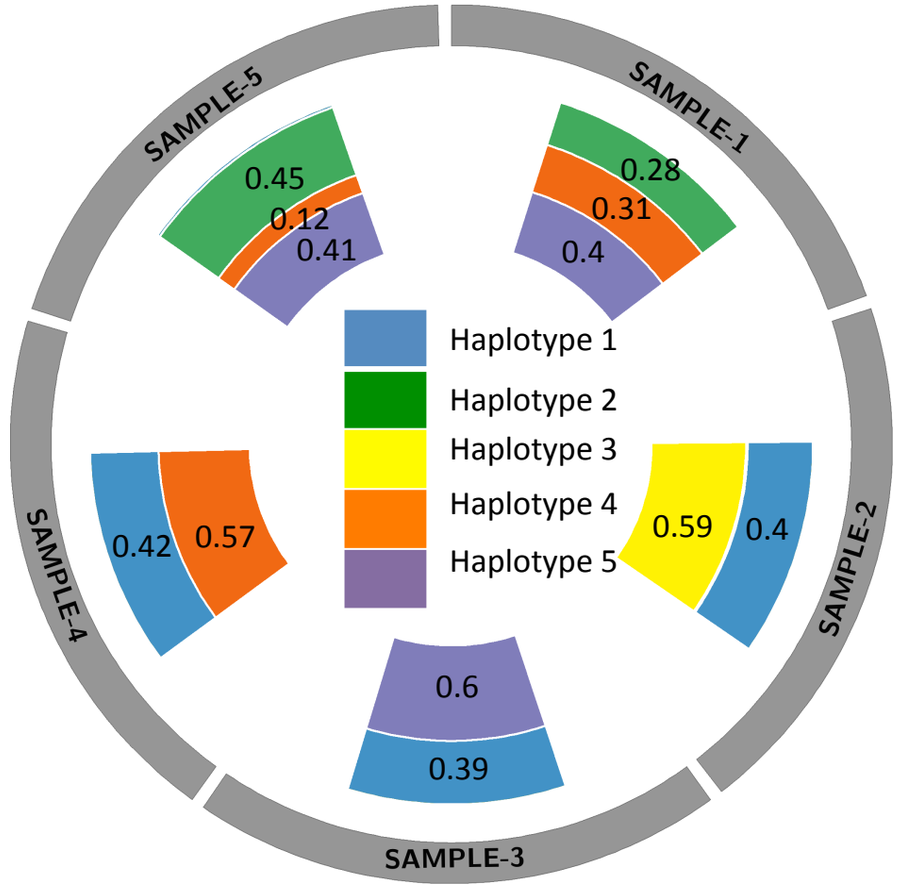
$\hat{w}$

Application: Inter-Tumor Heterogeneity

- Exome-sequencing data for five tumor samples from four different pancreatic ductal adenocarcinoma (PDAC) patients
- Bio-X pipeline (BWA, Samtools, GATK) for data preprocessing: coverage ~ 70X.
- Selected 118 SNVs: 1) significant coverage in all samples; 2) related to PDAC in the KEGG pathway database; 3) are nonsynonymous

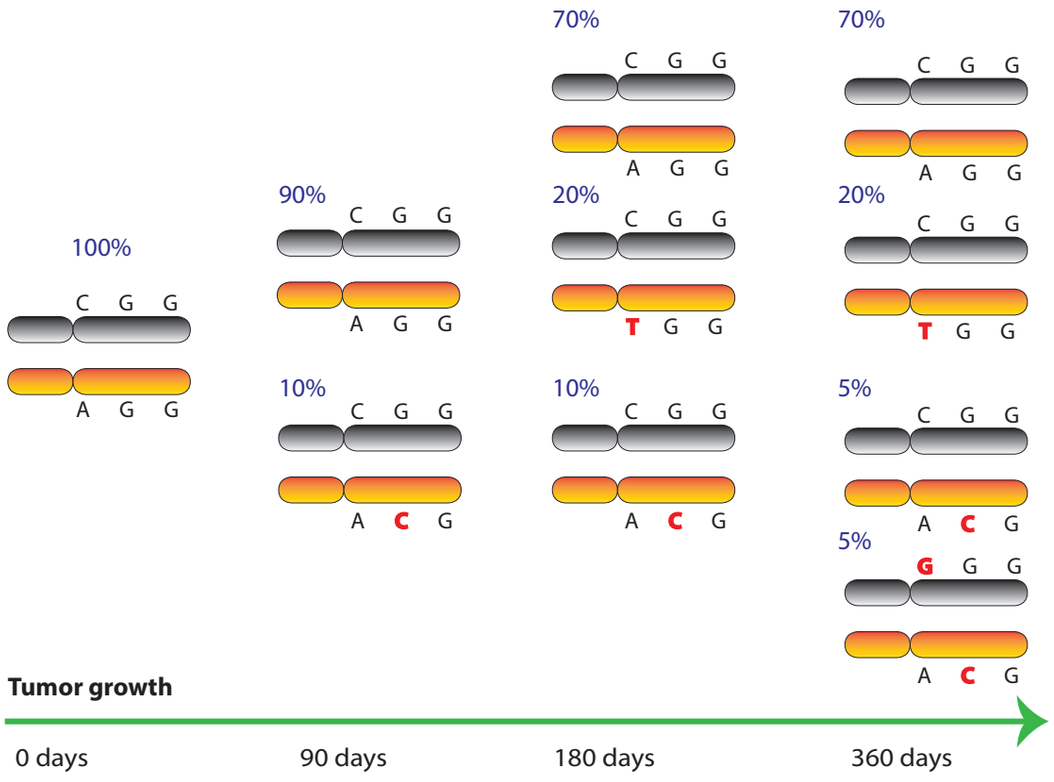
Application: Inter-Tumor Heterogeneity





$\hat{w}$

Extension: Categorial IBP



Subclone

	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

Clinical Trial Based on TH

The PANGEA -IMBBP Trial

Personalized ANtibodies for Gastro-Esophageal Adenocarcinoma: A Pilot 1st Metastatic Trial of Biologics Beyond Progression

Historical Control (Arm A) → FOLFOX → PFS_1 6m → FOLFIRI 60% → PFS_2 4m → FOLTAX 30% → PFS_3 2m

Biomarker Evaluation in all samples to allow for treatment assignment

Diagnosis: metastatic cancer

i) Primary Endpoints:
Feasibility:
Time to treatment assignment
Safety: toxicity
ii) mOS (HR 0.67) in HER2+/MET+
Historic Arm A v Arm B
(Improve to 18 months from 12 months)
HR: 0.67 of combined HER2+ & MET+
{12 month survival rate ~63%}
N= 68 (80% power)

•Secondary Endpoints:
$PFS_{1,2,3}$, PFS_{1+2+3} for HER2+/MET+
mOS for all 5 groups
2nd/3rd line treatment rates, RR
-Arm A_1 v A_2/ B_1 v B_2
-Compared to Historic Controls
Tissue correlates
- primary tumor to metastatic lesion
- baseline v $PFS_{1,2,3}$

Anticipated Incidence

Primary mOS Endpoint (N=68)

15-20%	HER2 amplified	Arm B1	FOLFOX -Trastuzumab	PD →	FOLFIRI + T	PD → FOLTAX + T
30-50%	MET amplified/Hi	Arm B2	FOLFOX -METab	PD →	FOLFIRI + M	PD → FOLTAX + M
5-10%	FGFR2 amplified	Arm B3	FOLFOX -FGFR2ab	PD →	FOLFIRI + F	PD → FOLTAX + F
25%	EGFR amplified/Hi KRAS wild type NI HER2,FGFR2,RON,MET	Arm B4	FOLFOX -EGFRab	PD →	FOLFIRI + EGFRab	PD → FOLTAX + EGFRab
20-25%	KRAS/BRAF/ PIK3CA/AKT/PTENdel mt/amplified NI HER2,FGFR2,RON,MET	Arm B5	FOLFOX -VEGFR2ab	PD →	FOLFIRI + V	PD → FOLTAX + V

ARM B: Therapy based on molecular profile

Secondary mOS Endpoint (N~104)