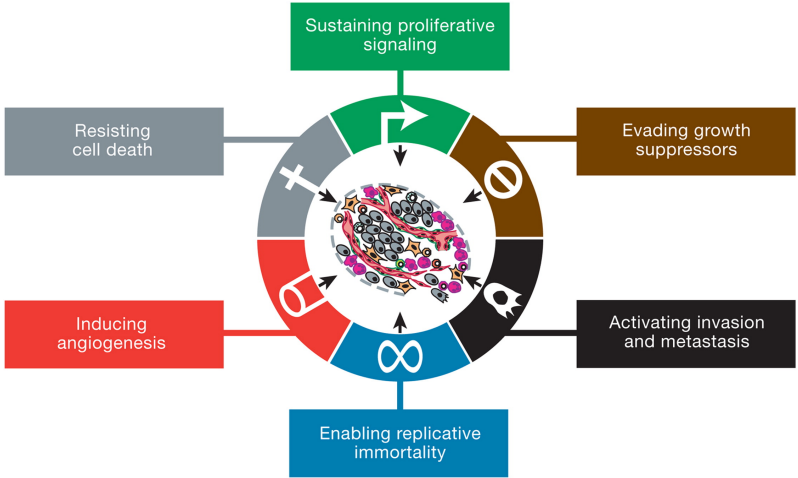
Statistical Methods for High Dimensional Cancer Biology

Andrew Roth

21st March 2022

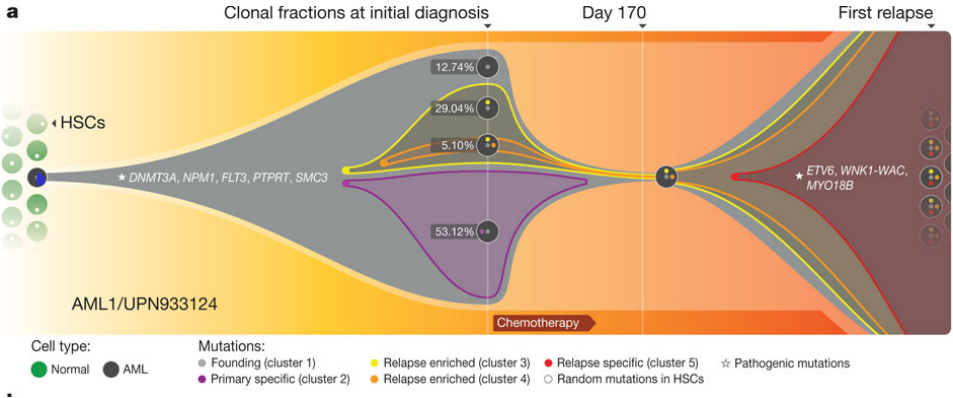
Biology

Cancer



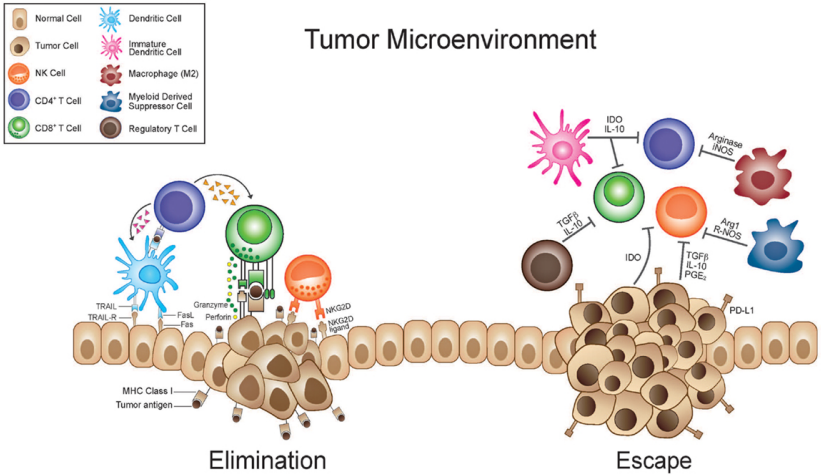
Hallmarks of Cancer: The Next Generation. Hanahan and Weinberg

Cancer evolution



Ding et al. Nature 2012

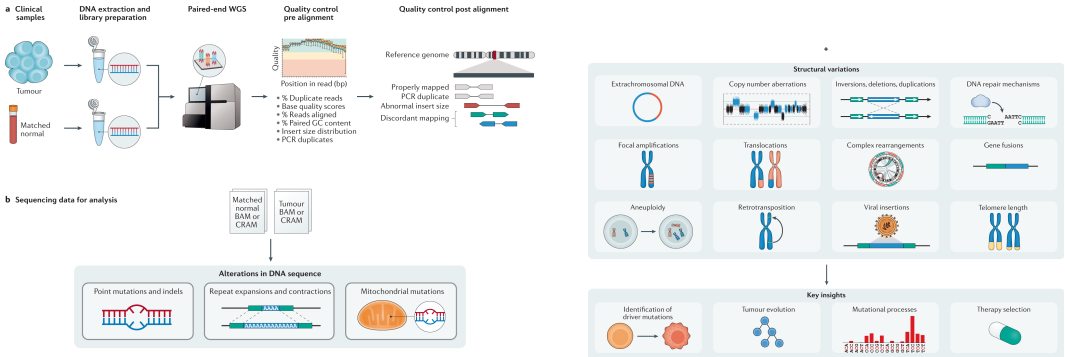
Tumour microenvironment



https://en.wikipedia.org/wiki/Tumor_microenvironment

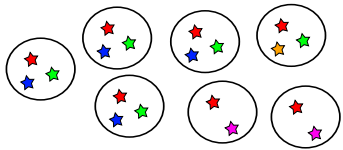
Technologies

High throughput sequencing



Computational analysis of cancer genome sequencing data. Cortés-Ciriano et. al.

Bulk sequencing

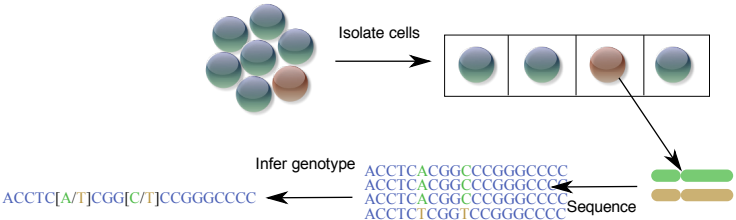


Sequence Bulk Sample

```
ACTCCCGTCGGAACCAATGCC---
-CTCCCGTCGGAACCAATGCCACC
---CCCGTCGGAACCAATGCCACG
-----CGTCGGAACCAATGCTACG
-----CATCGGAACCAATGCTACG
------GTCGGAACCAATGCCACG
--------------CAATGCCACC
--------------------CACC
```

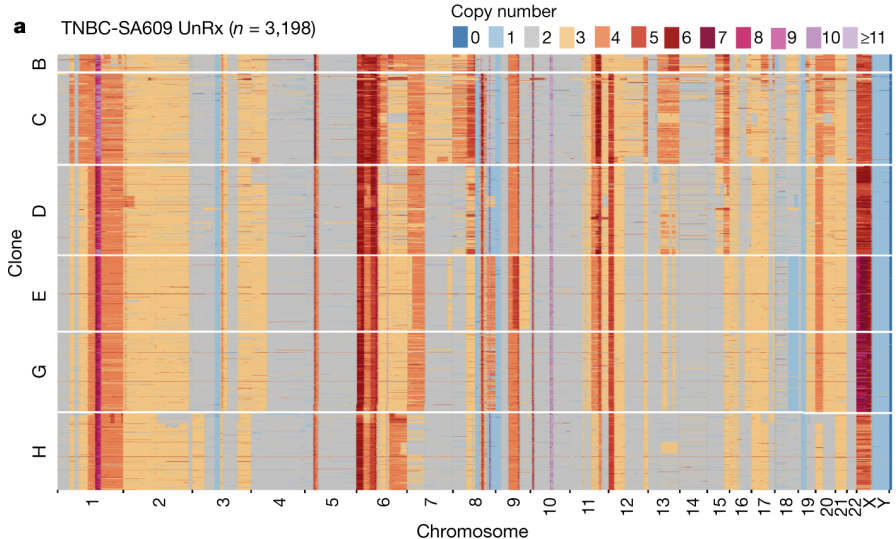
- Mature and high throughput
- Cell lysis ⇒cell of origin for reads is lost
- Proportion of reads with variants is related to prevalence of cells with variant

Single cell sequencing



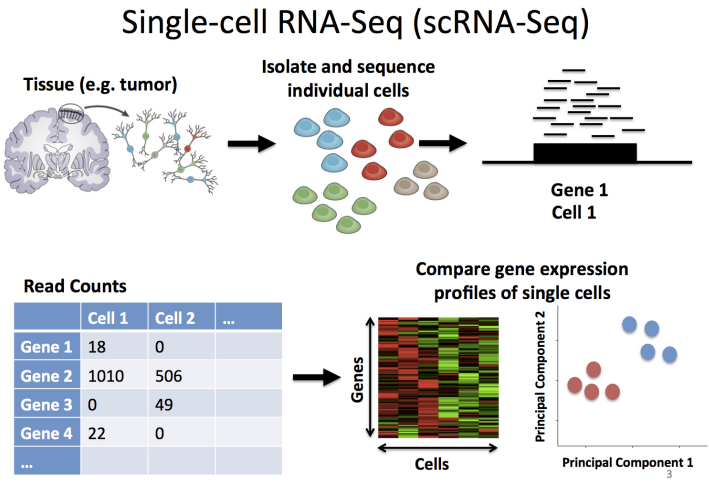
- Missing data
- Allele dropout
- Doublets

Single cell genome sequencing



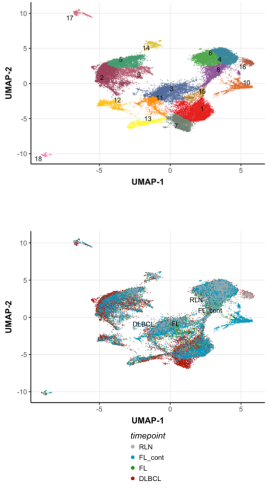
Clonal Decomposition and DNA Replication States Defined by Scaled Single-Cell Genome Sequencing

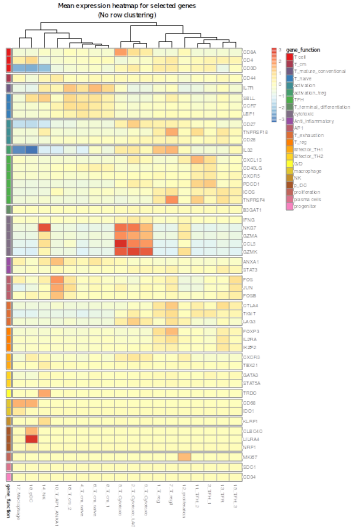
Single cell expression profiling



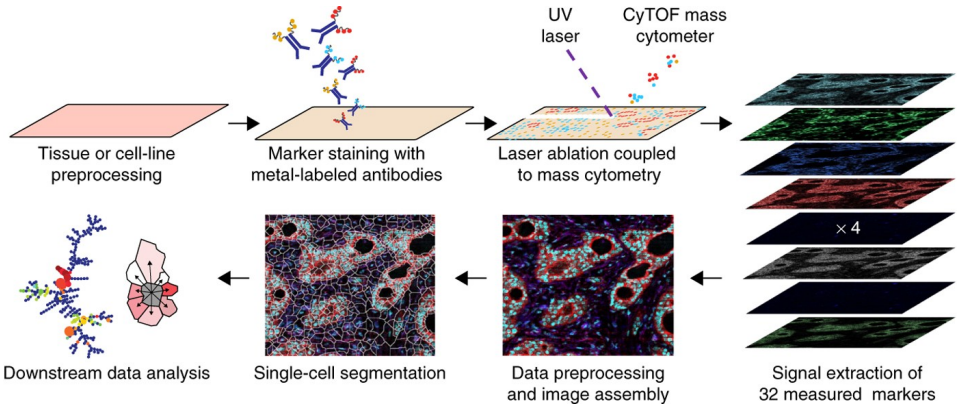
https://learn.gencore.bio.nyu.edu/single-cell-rnaseq/

Single cell expression profiling





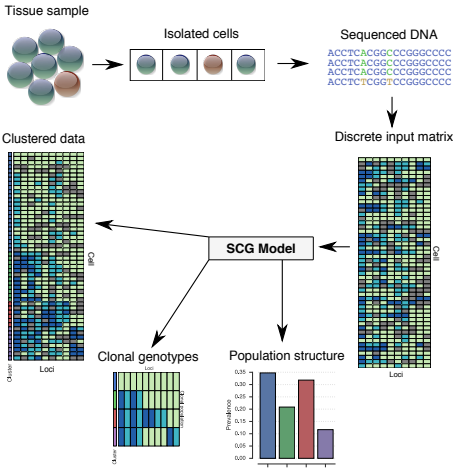
Spatial expression profiling

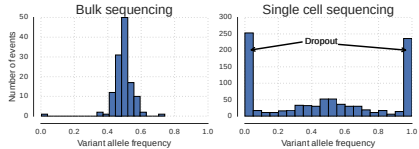


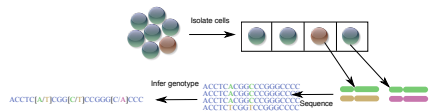
Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. Giesen et. al.

Probabilistic modelling

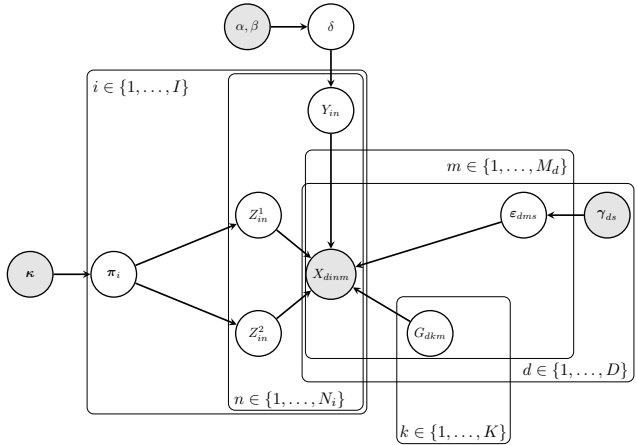
Why do we need probabilistic models?







What do I mean by probabilistic modelling?



Basic probability

- Let $A, B, \{C_i\}$ be random events.
- Then there are three basic equations that are at the centre of Bayesian probabilistic modelling.

joint — conditional — marginal

$$\text{Chain rule - } p(A,B) = p(B|A)p(A)$$

$$\text{Bayes' rule - } p(A|B) = \frac{p(B|A)p(A)}{p(B)}$$

$$\text{Law of total probability - } p(B) = \sum_i p(B|C_i)p(C_i)$$

$$\int p(x,y)\,dy = p(x)$$

Bayesian inference in a slide

- The Bayesian paradigm is one framework for probabilistic modelling.
- In the Bayesian setting model parameters are considered random like the data.
- The core quantity we need to compute in the Bayesian setting is the posterior

$$p(\theta|X) = \frac{p(x|\theta)\,p(\theta)}{p(x)}$$

- Here X is the data, θ are the parameters, $p(X|\theta)$ is the *likelihood* and $p(\theta)$ the *prior*.
- The normalisation constant is $p(X) = \int p(X|\theta)p(\theta)\mathrm{d}\theta$ and is sometimes called the *model evidence*

Bayesian inference a bit more

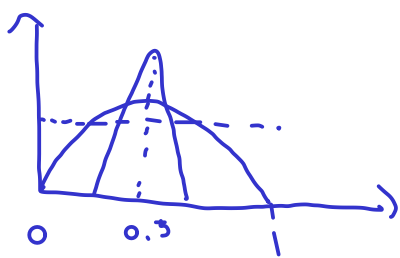
- The Bayesian approach to model fitting states that all the information about the parameters is encapsulated in the posterior.
- We begin with prior belief about θ encoded in the prior $p(\theta)$.
- Our updated belief about θ after seeing the data is given by the posterior $p(\theta|X)$.
- If we see new data X', then our previous posterior becomes our new prior and the new posterior is

$$p(\theta|X', X) = \frac{p(X'|\theta)\, p(\theta|X)}{p(X, X')}$$

Coin flipping

You flip a coin n times and observe x heads.

- How would you use this information to estimate the probability the coin comes up heads?
- Do you have prior beliefs about the coin?



Binomial distribution

- Consider performing a series of n trials where the outcomes are success or failure
- Let the probability of success be ρ
- Let X be a random variable indicating the number of successes
- Then X follows a Binomial distribution and the *probability mass function (pmf)* is

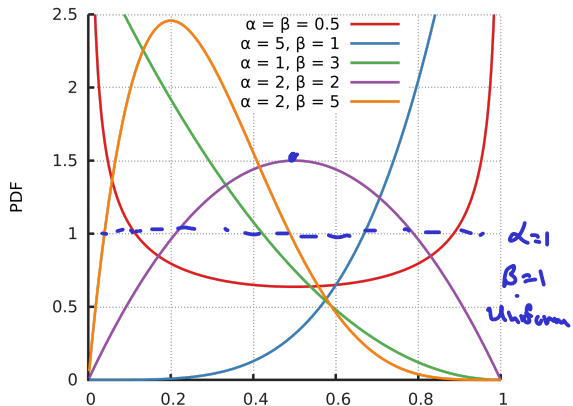
$$p(X = x|n, \rho) = \binom{n}{x} \rho^x (1-\rho)^{n-x}$$

Beta distribution

- The Beta distribution is a *continuous* distribution taking values in $(0,1)$
- The *probability density function (pdf)* is

$$p(x) = \frac{1}{\mathcal{B}(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

$$\mathcal{B}(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$



Coin flipping analysis

conjugacy $p(\theta)$ & $p(\theta|x)$ are in same family

A simple Bayesian model for the coin flipping experiment. Let ρ be the probability of success.

$$\rho|\alpha,\beta \sim \text{Beta}(\cdot|\alpha,\beta)$$
$$X|n,\rho \sim \text{Binomial}(\cdot|n,\rho)$$

$\alpha = 2$
$\beta = 2$

Now we compute

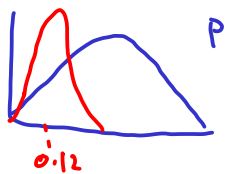
$p(\rho|X) = p(\rho|X,n,\alpha,\beta) \propto p(X|n,\rho)\cdot p(\rho|\alpha,\beta)$

$n = 100$
$x = 10$

$= \binom{n}{x}\rho^x(1-\rho)^{n-x} \cdot \frac{1}{B(\alpha,\beta)}\rho^{\alpha-1}(1-\rho)^{\beta-1}$

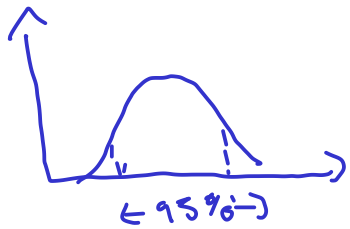
$= \rho^{x+\alpha-1}(1-\rho)^{n-x+\beta-1}$

$\rho|X \sim \text{Beta}(\cdot|x+\alpha, n-x+\beta)$



Summarising the posterior

- In principle the posterior tells us everything we could want to know about θ.
- In practice it can be hard to interpret high dimensional posteriors so we turn to point or region estimates.
- Simple point estimates include reporting summary statistics such as the mean and variance.
- Region estimates can be used to quantify uncertainty i.e. credible intervals.



Loss functions

- Loss functions provide a very general framework for summarising posteriors.
- A loss function $L(x, y)$ is a positive valued function which encodes our loss if we predict x when the true value is y.
  - L^1 loss $|x - y|$
  - L^2 loss $||x - y||^2$
- The Bayesian approach to point estimation is then to report the value with the minimum expected loss

$$\begin{aligned} \hat{\theta} &= \text{argmin}_{\theta'} \mathbb{E}_{p(\theta|X)}[L(\theta, \theta')] \\ &= \text{argmin}_{\theta'} \int L(\theta', \theta) p(\theta|X) \mathrm{d}\theta \end{aligned}$$

L_2 $\hat{\theta}$: mean

L_1 $\hat{\theta}$: median

The steps to building a probabilistic model

1. Identify a well defined problem.
2. Explore the data and qualitatively understand its properties.
3. Decompose the main problem into smaller pieces which can be iteratively extended.
4. Identify and implement a means to estimate model parameters.
5. Validate the model and inference approach.

Variant finding problem

Given n sequence reads covering a genomic position determine the genotype at that position.

Input: Integer counts representing the total number of reads d and the number of reads mis-matching the reference b.

Output: Predicted probability of each genotype (AA,AB,BB) at the position.

Data

- Let d_i denote the number of reads at the position and b_i the number that mis-match the reference.

Reference Genome

ACTCCCGTCGGAACGAATGCCACG

Reads

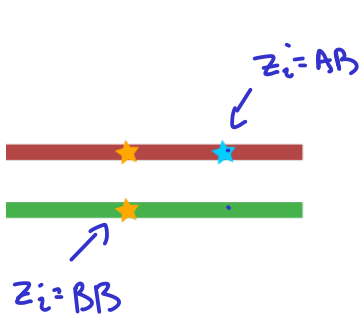
```
GCTCCCGTCGGAACCAATGCC---
-CTCCCGTCGGAACCAATGCCACC
---CCCGTCGGAACCAATGCCACG
-----CGTCGGAACCAATGCCACG
-----CATCGGAACCAATGCCACC
------GTCGGAACCAATGCCACG
--------------CAATGCCACC
----------------------CACC
```

Allelic Counts

```
b 1000001000000070000000004
d 1223356666666677777787 77
```

Latent variables

- Let $Z_i \in \{AA, AB, BB\}$ indicate the (hidden) genotype of the position.



Reference Genome

Reads

Allelic Counts

```
ACTCCCGTCGGAACGAATGCCACG

GCTCCCGTCGGAACCAATGCC---
-CTCCCGTCGGAACCAATGCCACC
---CCCGTCGGAACCAATGCCACG
-----CGTCGGAACCAATGCCACG
-----CATCGGAACCAATGCCACC
------GTCGGAACCAATGCCACG
--------------CAATGCCACC
--------------------CACC
b 10000010000000700000004
d 12233566666666777778777
```

Likelihood

- Assume the probability of sampling a read with the variant is given by θ_x for $x \in \{AA, AB, BB\}$ for example
  - $\theta_{AA} = 0.01$, $\theta_{AB} = 0.5$, $\theta_{BB} = 0.99$
  - Let $\boldsymbol{\theta} = (\theta_{AA}, \theta_{AB}, \theta_{BB})$
- If we knew the value of Z_i we could compute the probability of observing b_i out of d_i variant reads as a binomial i.e.

$$\begin{aligned} P(b_i|d_i, Z_i = x, \boldsymbol{\theta}) &= P(b_i|d_i, \theta_x) \\ &= \binom{d_i}{b_i} \theta_x^{b_i} (1-\theta_x)^{d_i - b_i} \end{aligned}$$

Predicting genotypes

- What we really want to compute is $P(Z_i = x | b_i, d_i, \theta)$
- How?

$$P(Z_i = x | b_i, d_i, \theta) = \frac{P(b_i | z_i = x, d_i, \theta) \cdot P(z_i = x)}{\sum_z P(b_i | z_i = z, b_i, \theta) \, P(z_i = z)}$$

$P(z_i = x)$: $\frac{1}{3}$ uniform

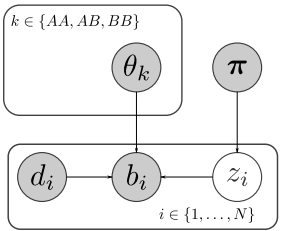
Examples

- What is the most probable genotype and what is the probability if
  - $d_i = 2$ and $b_i = 1$
  - $d_i = 20$ and $b_i = 10$
  - $d_i = 2000$ and $b_i = 1000$
  - $d_i = 100$ and $b_i = 25$

Variant calling model

$cat(z|\vec{\pi}) = \prod_k \pi_k^{I(z=k)}$ ← indicator

$P(z=\pi|\vec{\pi}) = \pi_k$ $\sum_k \pi_k$



$$
\begin{aligned}
\pi_k &= \frac{1}{K} = \tfrac{1}{3} \\
z_i|\pi &\sim \text{Categorical}(\cdot|\pi) \\
(\theta_{AA}, \theta_{AB}, \theta_{BB}) &= (0.01, 0.5, 0.99) \\
b_i|d_i, z_i = k, \theta &\sim \text{Binomial}(\cdot|d_i, \theta_k)
\end{aligned}
$$

Inference

$$
\begin{aligned}
p(z_i = k|b_i, d_i, \theta, \pi) &= \frac{p(b_i|d_i, z_i = k, \theta)p(z_i = k|\pi)}{\sum_{\ell \in \{AA,AB,BB\}} p(b_i|d_i, z_i = \ell, \theta)p(z_i = \ell|\pi)} \\
&\propto \underbrace{\theta_k^{b_i}(1-\theta_k)^{d_i - b_i}}_{p(b_i|d_i, z_i=k, \theta)} \underbrace{\pi_k}_{p(z_i=k|\pi)}
\end{aligned}
$$

Limitations

- We assume that $\boldsymbol{\theta} = (\theta_{AA}, \theta_{AB}, \theta_{BB})$ is known.
  - How could we estimate these parameters?
- In general the genotype *AA* is vastly more common that *AB* or *BB*.
  - How should we account for this?
- Is sequence data really binomial?

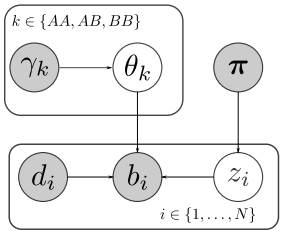
overdispersion

Binomial → BetaBinomial

Poisson → NegativeBinomial

Variant calling model - Learning θ

$\gamma_k = (\alpha_k, \beta_k)$



$$
\begin{aligned}
\pi_k &= \frac{1}{K} \\
z_i|\pi &\sim \text{Categorical}(\cdot|\pi) \\
\theta_k|\gamma_k &\sim \text{Beta}(\cdot|\alpha_k, \beta_k) \\
b_i|d_i, z_i = k, \theta &\sim \text{Binomial}(\cdot|d_i, \theta_k)
\end{aligned}
$$

Inference

$$
p(\boldsymbol{z}, \theta|\boldsymbol{b}, \boldsymbol{d}, \pi) \propto \prod_i p(b_i|d_i, z_i, \theta) p(z_i = k|\pi) \prod_k p(\theta_k|\gamma_k)
$$

$p(x)$ intractable $=$ i.e. no closed form

$$
= \prod_i \prod_k \{\textbf{Binomial}(b_i|d_i, \theta_k)\pi_k\}^{\mathbb{I}(z_i=k)} \prod_k p(\theta_k|\gamma_k)
$$

Why Bayesian inference is challenging

- Bayesian inference is conceptually simple - we apply Bayes' rule and compute the posterior.
- In practice this means computing the normalisation constant $p(X) = \int p(X|\theta)p(\theta)\mathrm{d}\theta$.
  - If θ is high dimensional this is typically intractable.
- Bayesian inference becomes hard because we need to either:
  - Avoid explicitly computing $p(X)$.
  - Use advanced methods to estimate $p(X)$.
- As a result we almost always rely on some method to compute an approximation to the posterior.

Monte Carlo methods

- In the Bayesian setting we typically want to compute expected values.
  - For example mimising the expected loss functions.
- Monte Carlo methods make the following simple observation

$$\begin{aligned} \mathbb{E}_p[h] &= \int h(x)p(x)\mathrm{d}x \\ &\approx \frac{1}{S}\sum_{s=1}^{S} h(x^{(s)}) \end{aligned}$$

  where $x^{(s)}$ are random draws from p.
- The accuracy of this estimator increases as the number of samples, S, increases.
  - This is the Law of Large Numbers in action.

MCMC methods

- Sampling from the posterior directly is typically as hard as computing the posterior.
- The basic idea of Markov Chain Monte Carlo (MCMC) methods is to construct a Markov chain that admits $p(\theta|X)$ as its invariant distribution.
- We can then sequentially draw samples from the Markov chain to obtain samples from $p(\theta|X)$.

$$p(x_n | x_{n-1}, x_{n-2}, \ldots, x_1) = p(x_n | x_{n-1})$$

1st order Markov

Gibbs sampling

$p(\theta_1, \theta_2 | X)$ - goal

- Assume $\theta = (\theta_1, \theta_2)$ and we can sample from the conditional distributions $p(\theta_1|\theta_2, X)$ and $p(\theta_2|\theta_1, X)$.
- The Gibbs sampler works by alternatively sampling from $p(\theta_1|\theta_2, X)$ and then $p(\theta_2|\theta_1, X)$.
- We use the previous values from each step to compute the conditional in the next.
- Iterate many times to draw a set $\{\theta^{(s)}\}_{s=1}^{S}$ of samples to approximate posterior.

$\theta_1' \sim p(\theta_1 | X, \theta_2)$

$\theta_2' \sim p(\theta_2 | X, \theta_1')$

save $\theta' = (\theta_1', \theta_2')$

Gibbs updates for z_i

$$
\begin{aligned}
p(z_i = k | b_i, d_i, \boldsymbol{\theta}, \boldsymbol{\pi}) &\propto p(b_i | d_i, z_i = k, \boldsymbol{\theta}) p(z_i = k | \boldsymbol{\pi}) \prod_k p(\theta_k | \gamma_k) \\
&\propto p(b_i | d_i, z_i = k, \boldsymbol{\theta}) p(z_i = k | \boldsymbol{\pi}) \cdot p(\theta_k | \gamma_k) \\
&= \text{Binomial}(b_i | d_i, \theta_k) \pi_k \\
\implies p(z_i = k | b_i, d_i, \boldsymbol{\theta}, \boldsymbol{\pi}) &= \frac{\pi_k \theta_k^{b_i} (1 - \theta_k)^{d_i - b_i}}{\sum_\ell \pi_\ell \theta_\ell^{b_i} (1 - \theta_\ell)^{d_i - b_i}}
\end{aligned}
$$

Gibbs updates for θ_k

$$
\begin{aligned}
p(\theta_k|\boldsymbol{b},\boldsymbol{d},\boldsymbol{z},\boldsymbol{\pi}) &\propto \prod_i \prod_k \{\text{Binomial}(b_i|d_i,\theta_k)\pi_k\}^{\mathbb{I}(z_i=k)} \prod_k p(\theta_k|\gamma_k) \\
&\propto \prod_i \{\text{Binomial}(b_i|d_i,\theta_k)\pi_k\}^{\mathbb{I}(z_i=k)} \\
&\propto \theta_k^{\alpha_k+\sum_i \mathbb{I}(z_i=k)\, b_i - 1} (1-\theta_k)^{\beta_k+\sum_i \mathbb{I}(z_i=k)\,(d_i-b_i)-1}
\end{aligned}
$$

$\bar{\alpha}_k$ $\bar{\beta}_k$

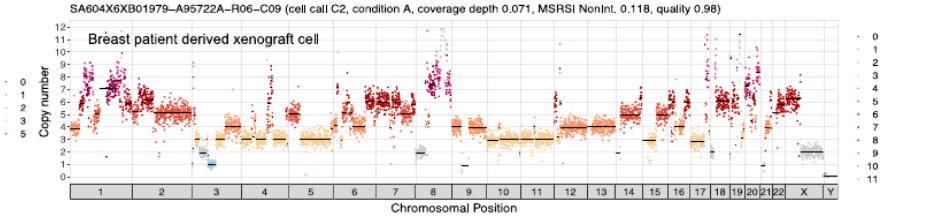
$$\theta_k \sim \text{Beta}(\cdot\,|\bar{\alpha}_k, \bar{\beta}_k)$$

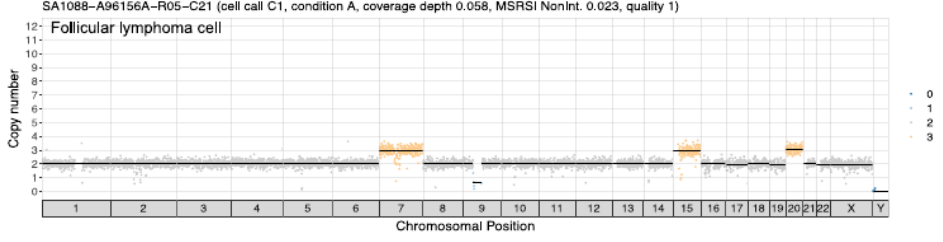
Probabilistic models in cancer biology

Copy number calling

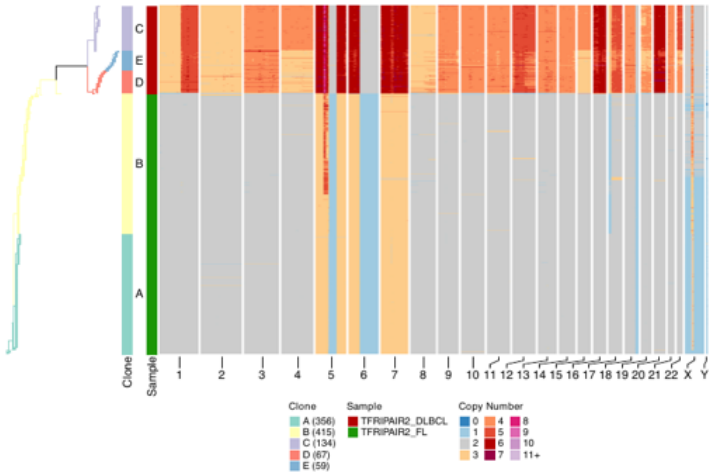
data i.i.d.

Hidden Markov Model

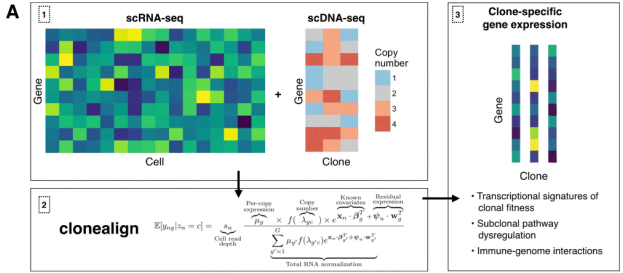




Phylogenetics



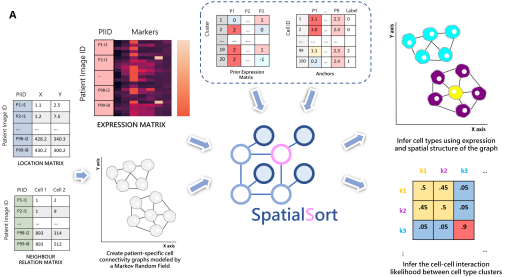
Multi-modal integration



clonealign

Spatially aware clustering

Hidden
Markov
Random
Field



Variable	Represents	Distribution
$V_{p,n,m}$	Marker m expression of cell n of patient p	Gaussian
$x_{p,n}$	Cell type label of cell n of patient p	Modified Potts
$\theta_{k,m}$	Combined mean and variance of marker m expression of cluster k	Normal-Gamma
β_c	Cell-cell interaction parameter of class c	Beta
C_p	Patient subtype of patient p	Categorical