Statistical Data Analysis 2, Cancer phylogenetics

Ewa Szczurek

Faculty of Mathematics, Informatics and Mechanics, University of Warsaw

Useful function: gamma

You may think of the gamma function as a continuous generalization of the factorial $n!$ function.

Gamma function $\Gamma(x)$

Let $\mathbb{X}$ be the set of all real and complex numbers, except for the negative integers and zero. The gamma function $\Gamma(x)\colon \mathbb{X} \to \mathbb{R}$ is defined as

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt.$$

Since $\Gamma(x+1) = x\Gamma(x)$ (integration by parts) and $\Gamma(1) = 1$, for positive integer x we have

$$\Gamma(x) = (x-1)!$$

Useful function: beta

Beta function $B(\alpha, \beta)$

The beta function $B(\alpha, \beta) : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}$ is defined as

$$B(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1} dt = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

For α and β positive integer

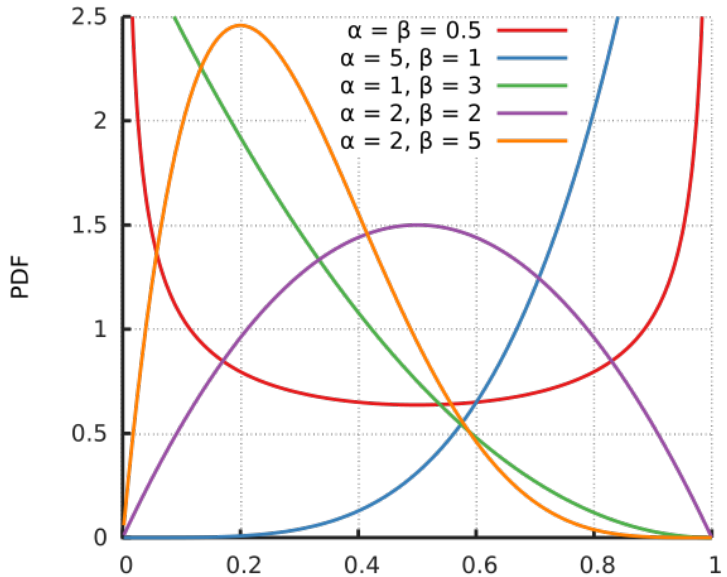
$$B(\alpha, \beta) = \frac{(\alpha-1)!(\beta-1)!}{(\alpha+\beta-1)!}.$$

Beta distribution (Beta(α, β))

Beta(α, β)

$$f(x) = \frac{x^{\alpha}(1-x)^{\beta-1}}{B(\alpha,\beta)}, \text{ for } x \in [0,1]$$

$$E(X) = \frac{\alpha}{\alpha+\beta}, \quad Var(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$



Dirichlet distribution (Dirichlet(α))

Generalization of the Beta distribution. α is a vector $\alpha \in \mathbb{R}_+^K$, with α_i called the concentration parameters and K the number of categories.

Dirichlet(α)

Let $B(\alpha) = \frac{\prod_{i=1}^K \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^K \alpha_i)}$.

$$f(x_1, \ldots, x_K) = \frac{1}{B(\alpha)} \prod_{i=1}^K x_i^{\alpha_i - 1},$$

for $x \in \Delta_{K-1}$, i.e., $\sum_{i=1}^K x_i = 1, x_i \geq 0$ for all $i \in 1, \ldots, K$.

$$E(X) = \frac{\alpha_i}{\sum_i \alpha_i}$$

Dirichlet distribution is the conjugate prior of the categorical distribution and multinomial distribution.

Illustration of the Dirichlet distribution

Illustration on a gif how the log of the density function changes when $K = 3$ as we change the vector α from $\alpha = (0.3, 0.3, 0.3)$ to $(2.0, 2.0, 2.0)$, keeping all the individual α_i's equal to each other.

Bayesian mixture model with K components

$$
\begin{aligned}
\theta_k &\sim H \\
\pi|\alpha &\sim \text{Dirichlet}(\frac{\alpha}{K}, \dots \frac{\alpha}{K}) \\
z_i|\pi &\sim \text{Multinomial}(\pi) \\
x_i|z_i = k, \{\theta_k\} &\sim F(\theta_k)
\end{aligned}
$$

where

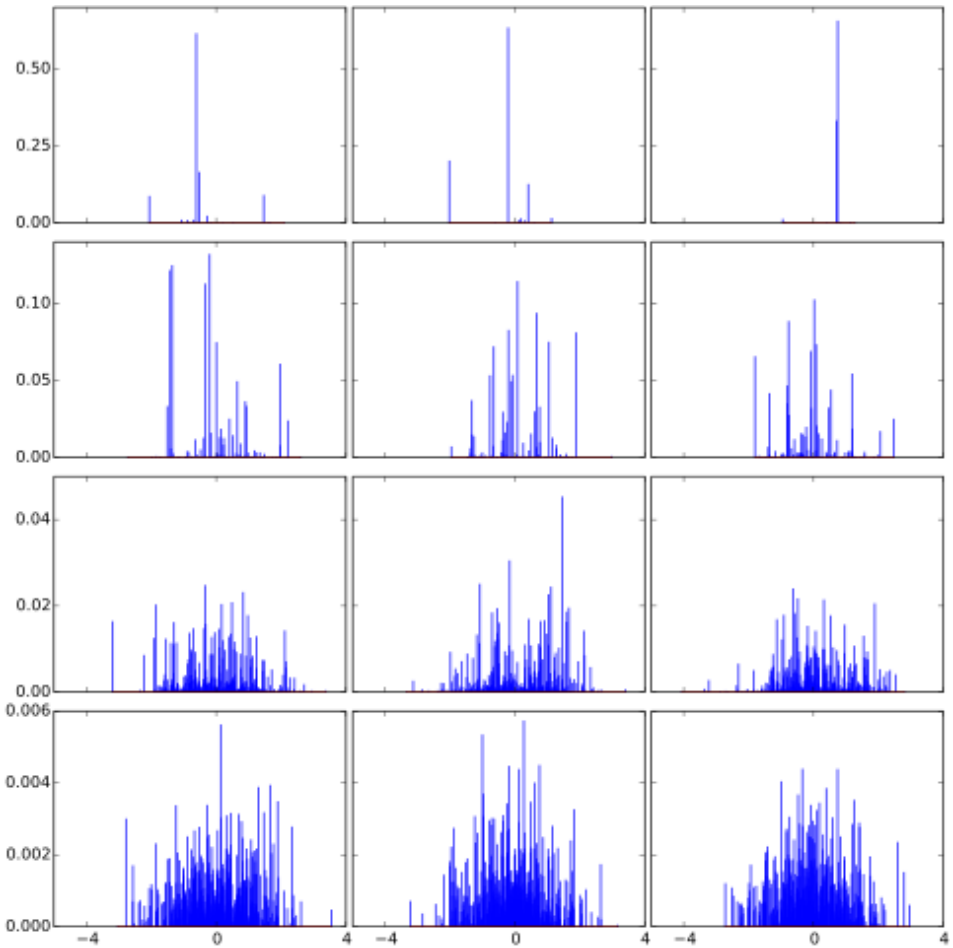
- K is the number of mixture components,
- $\pi_{k=1,\dots K}$ is the mixture weight of component k, with $\sum_k \pi_k = 1$,
- $z_{i=1,\dots N}$ is the component of observation i,
- θ_k is the parameter of distribution associated with component k,
- H is the prior probability of parameters θ.

This model induces a clustering of observations x_i.

Dirichlet process (DP(H, α))

- A stochastic process whose realizations are probability distributions
- H - base distribution - the expected value of the process
  - Although H can be any, also continuous distribution, the distributions drawn from DP are almost surely discrete
- α concentration parameter
  - For $\alpha \to 0$, the distributions are concentrated at single value, for $\alpha \to \infty$ approach continuous distribution
- Infinite dimensional ($K = \infty$) generalization of the Dirichlet distribution.

Draws from the Dirichlet process



Draws from the Dirichlet process $\mathrm{DP}(\mathrm{N}(0,1),\alpha)$. Each row uses a different α: 1, 10, 100 and 1000. A row contains 3 repetitions of the same experiment.

Dirichlet process- intuition

- A model of data $x_1, x_2, \ldots$ that tends to repeat previous values ("rich get richer" fashion) according to the following algorithm

Input: H, α

1. Draw x_1 from the distribution H.
2. For $n > 1$:
   a) With probability $\frac{\alpha}{\alpha+n-1}$ draw x_n from H.
   b) With probability $\frac{n_x}{\alpha+n-1}$ set $x_n = x$, where n_x is the number of previous observations of x.

- Equivalently, the data can be obtained by

1. Draw distribution P from $\mathrm{DP}(H, \alpha)$.
2. Draw observations $x_1, x_2, \ldots$ from P

Dirichlet process- formally

Given

- Measurable set S
- Base probability distribution H
- Concentration parameter $\alpha \in \mathbb{R}_+$

Dirichlet process $\mathrm{DP}(H, \alpha)$

Stochastic process, whose realization (set of random variates drawn from the process) is a probability distribution over S such that the following holds. For any measurable finite partition of S, denoted $\{B_i\}_{i=1}^n$, if

$$X \sim \mathrm{DP}(H, \alpha)$$

then

$$(X(B_1), \ldots, X(B_n)) \sim \mathrm{Dirichlet}(\alpha H(B_1), \ldots, \alpha H(B_n)).$$

Finite versus infinite mixture model

Example mixture model with K components

$$
\begin{aligned}
\mu_k &\sim H \\
\pi|\alpha &\sim \text{Dirichlet}(\frac{\alpha}{K}, \dots \frac{\alpha}{K}) \\
z_i|\pi &\sim \text{Multinomial}(\pi) \\
x_i|z_i = k, \mu_k &\sim \text{N}(\mu_k, \sigma^2)
\end{aligned}
$$

Equivalent model

$$
\begin{aligned}
\mu_k &\sim H \\
\pi|\alpha &\sim \text{Dirichlet}(\frac{\alpha}{K}, \dots \frac{\alpha}{K}) \\
\mu_i &\sim G = \sum_{k=1}^{K} \pi_k \delta_{\mu_k}(\mu_i) \\
x_i|\mu_i &\sim \text{N}(\mu_i, \sigma^2)
\end{aligned}
$$

Finite versus infinite mixture model

Example mixture model with K components

$$\begin{aligned} \mu_k &\sim H \\ \pi|\alpha &\sim \text{Dirichlet}(\frac{\alpha}{K}, \ldots \frac{\alpha}{K}) \\ z_i|\pi &\sim \text{Multinomial}(\pi) \\ x_i|z_i = k, \mu_k &\sim \text{N}(\mu_k, \sigma^2) \end{aligned}$$

Equivalent model

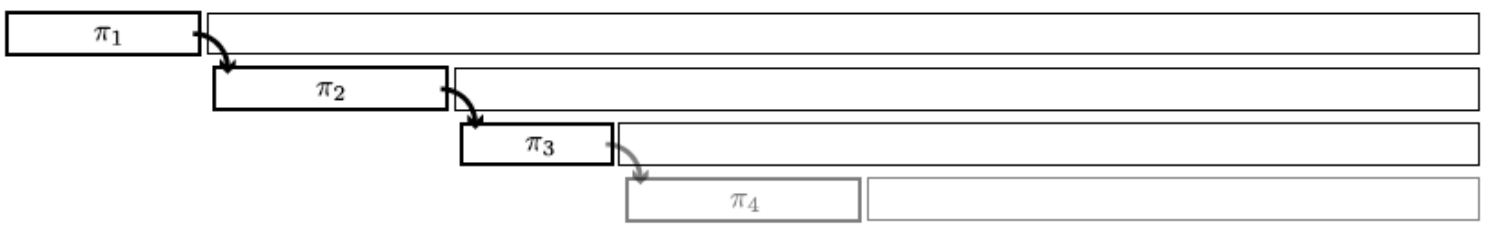
$$\begin{aligned} \mu_k &\sim H \\ \pi|\alpha &\sim \text{Dirichlet}(\frac{\alpha}{K}, \ldots \frac{\alpha}{K}) \\ \mu_i &\sim G = \sum_{k=1}^{K} \pi_k \delta_{\mu_k}(\mu_i) \\ x_i|\mu_i &\sim \text{N}(\mu_i, \sigma^2) \end{aligned}$$

Extended model without pre-specifying K ($G(\mu_i) = \sum_{k=1}^{\infty} \pi_k \delta_{\mu_k}(\mu_i)$).

$$\begin{aligned} G &\sim \text{DP}(H, \alpha) \\ \mu_i &\sim G \\ x_i|\mu_i &\sim \text{N}(\mu_i, \sigma^2) \end{aligned}$$

Stick-breaking construction of the Dirichlet process

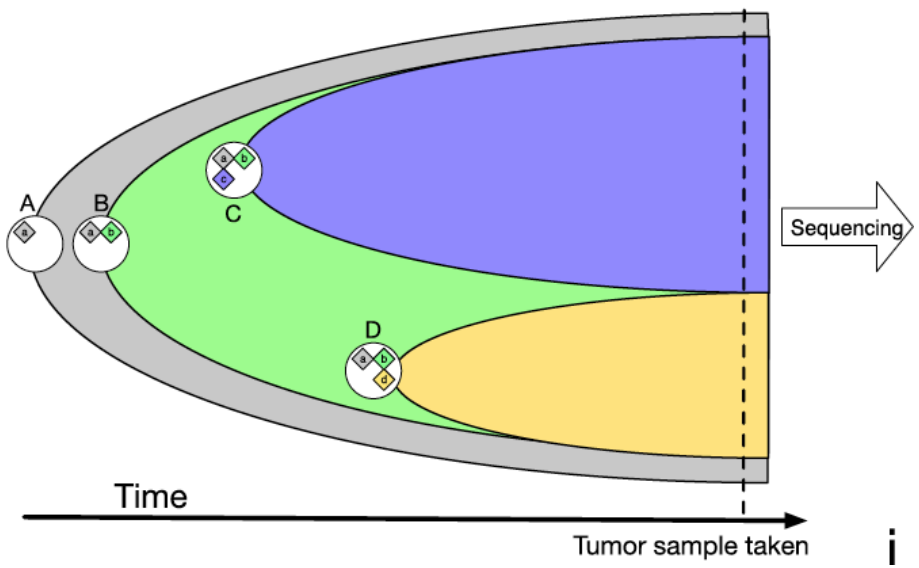
$$
\begin{aligned}
\beta_k &\sim \text{Beta}(1, \alpha) \\
\pi_k &= \beta_k \prod_{l=1}^{k-1} (1 - \beta_k) \\
\theta_k &\sim H \\
G(\theta) &= \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}(\theta)
\end{aligned}
$$



Tumor cell populations - clones

Tumor cell populations are

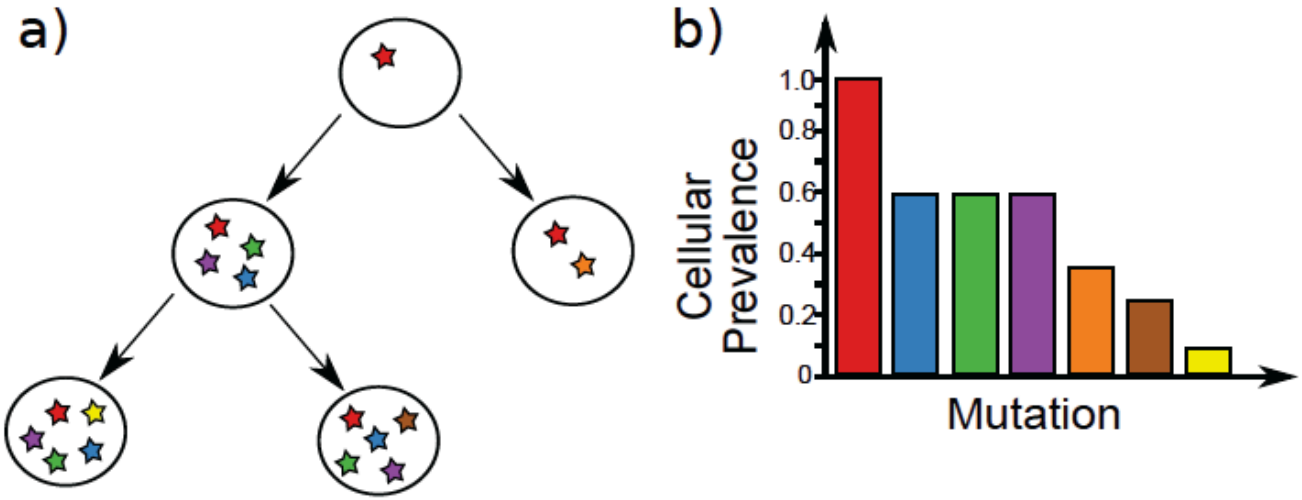
- heterogeneous
- shaped by evolution
- clone: a subpopulation of tumor cells with the same genotype
- samples taken for sequencing are 'contaminated' with normal cells.
- **How to reconstruct the clones from tumor sequencing data?**



Cellular prevalence of mutation

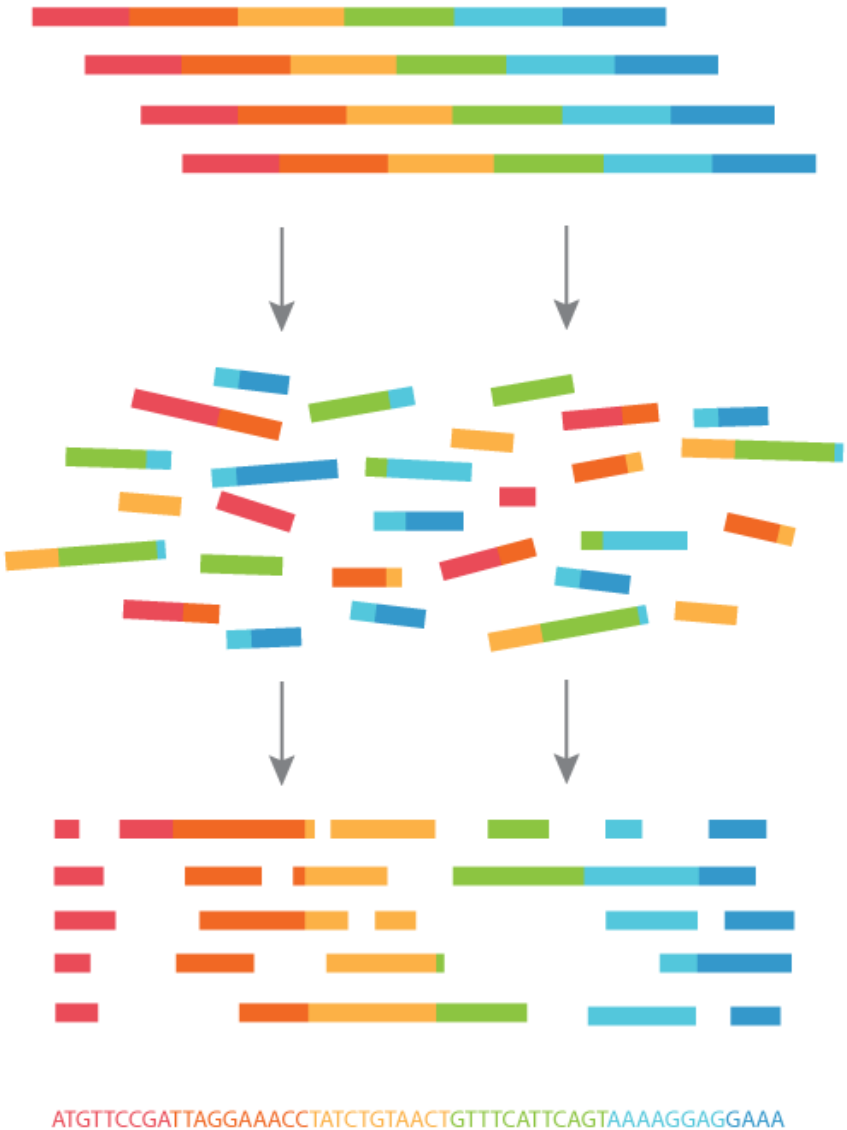
Cellular prevalence ϕ_n of mutation n

Fraction of tumor cells that carry mutation i.

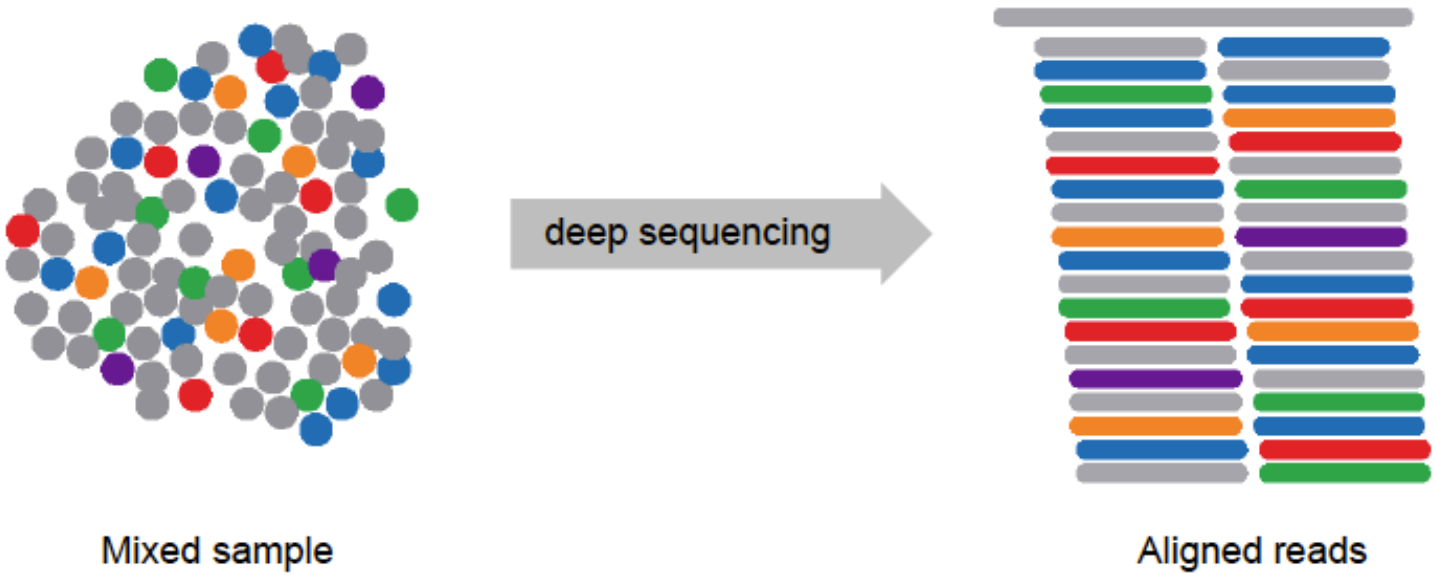


- Mutations occurring higher up the tree always have a greater cellular prevalence than their descendants.
- Green, blue and purple mutations occur at the same cellular prevalence because they co-occur in the clones of the tree.
- **Clustering of cellular prevalences $\rightarrow$ clones.**

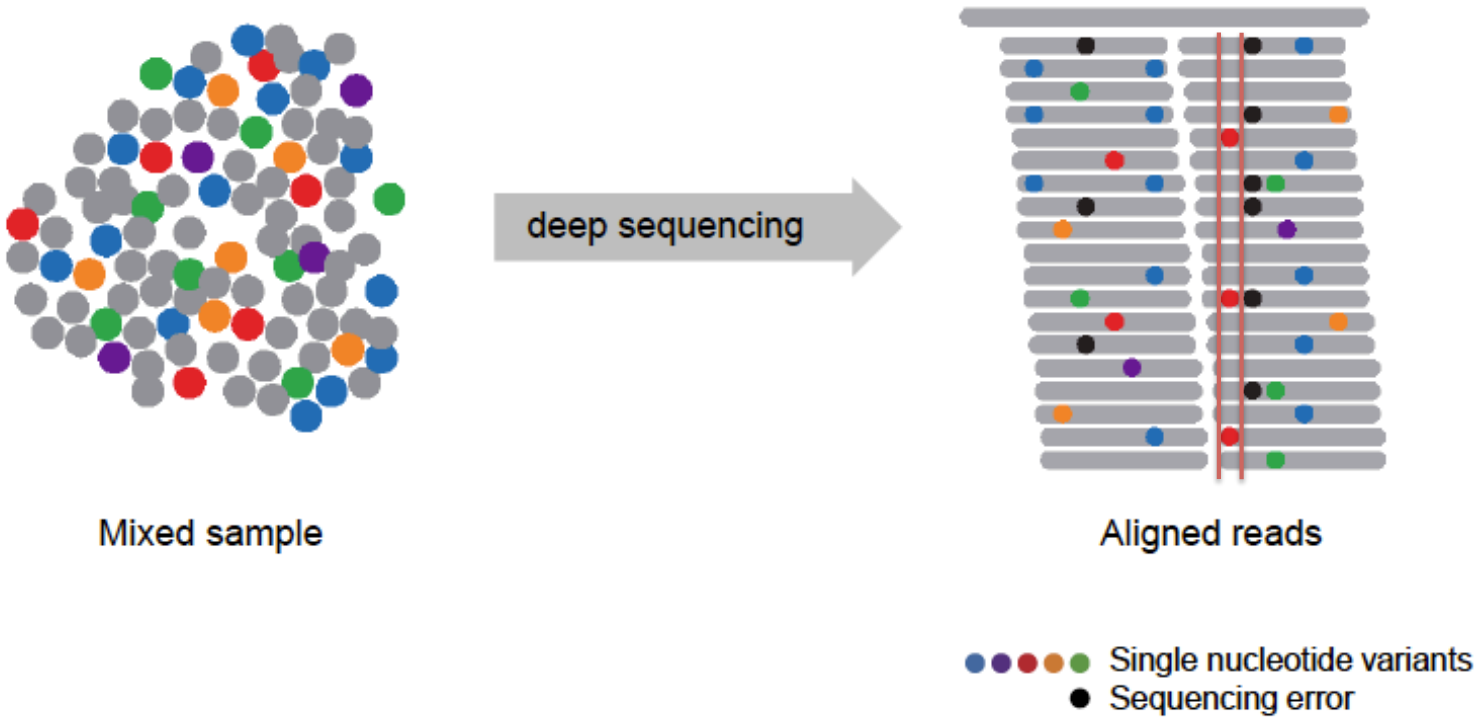
How does DNA sequencing work?



Deep sequencing of heterogeneous samples



Allelic variant prevalence



For a variant at position (locus) *i*

- fraction of reads at this position that carry the variant allele
- also called variant allele frequency (VAF)
- in tumor samples it is not equal to cellular prevalence!

Why allelic variant prevalence $\neq$ cellular prevalence?

Allelic prevalence is a compound measure depending on

- the proportion of tumor cells harboring the mutation (cellular prevalence), but also
- proportion of 'contaminating' normal cells
- number of allelic copies of the mutation in each cell
- sequencing errors and other technical noise

PyClone: a graphical model of clonal populations in cancer

Input:

- allelic prevalence of each variant
- allele-specific copy number estimates

Output:

- posterior denisties for cellular prevalence of each mutation and for other model parameters
- clustering over mutations

Roth et al., Nat Methods (2014)

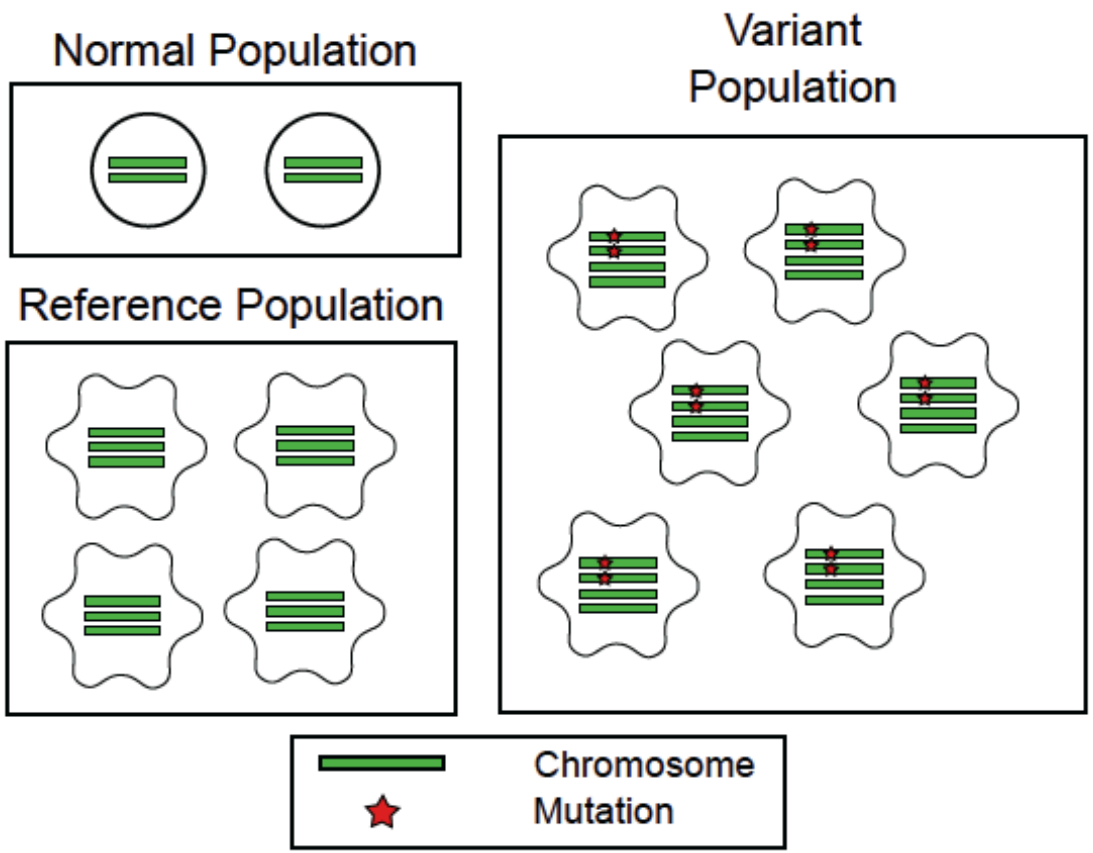
PyClone assumptions

Clonal populations follow

- a perfect phylogeny: no site mutates more than once in the evolutionary history, each site harbors at most one somatic mutation
- persistent phylogeny: mutations do not dissapear or reverse

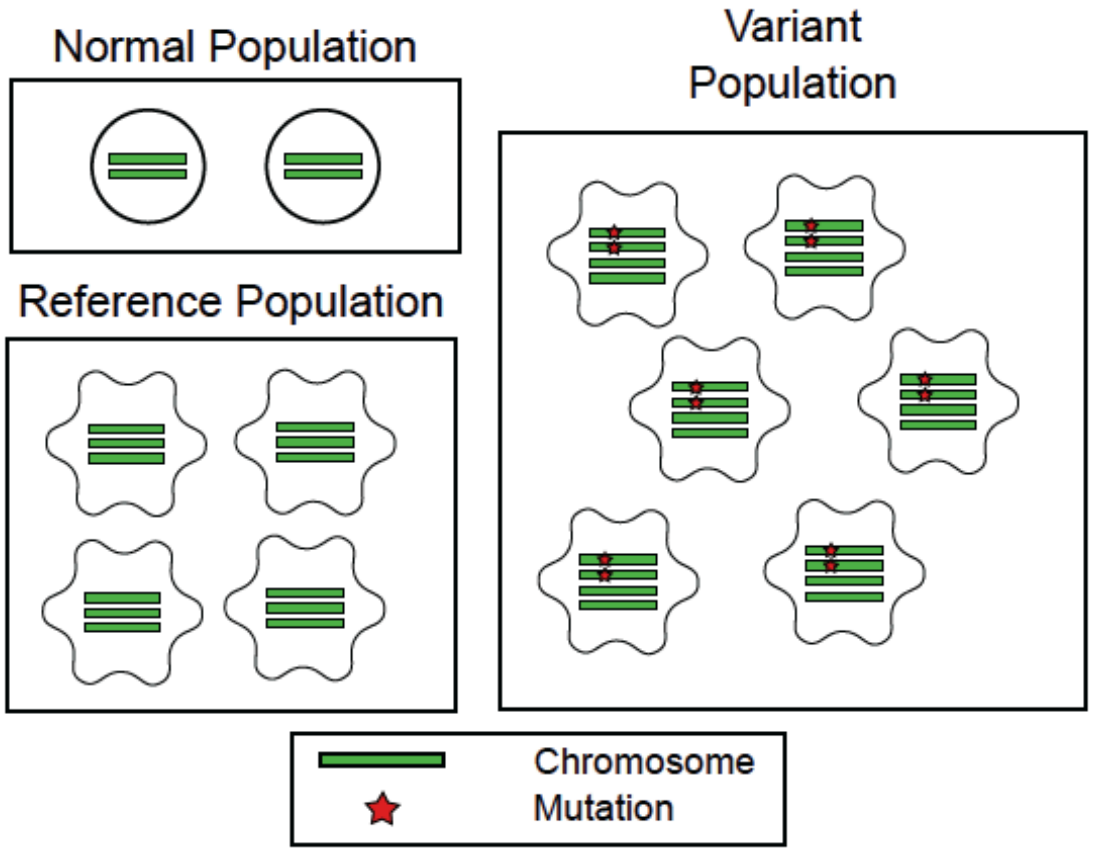
Consequently, **clusters of mutations occuring at the same point in the clonal phylogeny are present at shared cellular prevalences.**

Division of the sample into populations w.r.t mutation *n*



- 'normal population': all normal cells (circular)
- 'reference population': cancer cells (irregular) which do not contain the mutation
- 'variant population': all cancer cells with the mutation

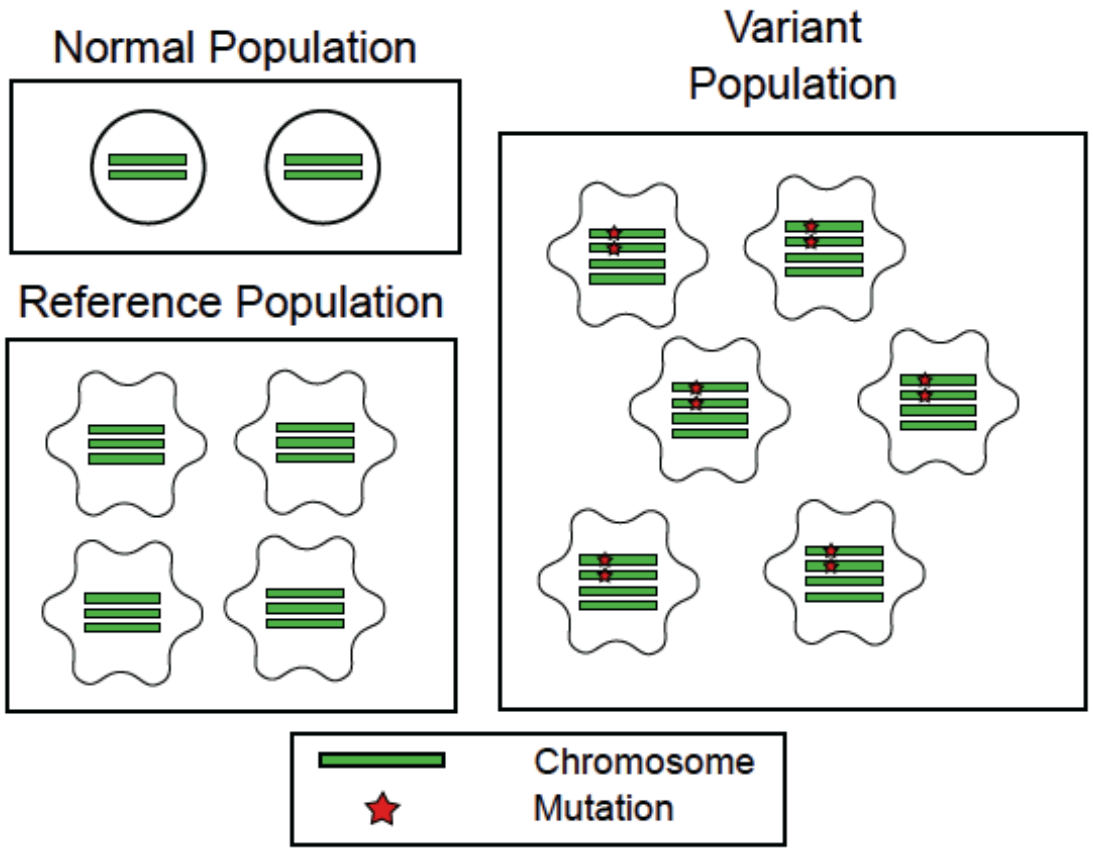
Division of the sample into populations w.r.t mutation n



Again, cellular prevalence $\neq$ variant allelic prevalence

- Cellular prevalence: $6/10{=}0.6$
- Variant allelic prevalence: $6 \cdot 4 \cdot \frac{2}{4}/(2 \cdot 2 + 4 \cdot 3 + 6 \cdot 4) = 0.3$

Division of the sample into populations w.r.t mutation n



Simplifying assumption:

- All cells within each population share the same genotype.
- Here the variant population has genotype AABB (two copies of the reference allele A, and two copies of the variant allele B)

Population genotypes w.r.t mutation n

- All cells within each population share the same genotype,
- but different populations may have different genotypes.

Population genotypes w.r.t mutation n

- All cells within each population share the same genotype,
- but different populations may have different genotypes.

Formally

- A collection of categorical random variables g_N^n, g_R^n, g_V^n $\in \mathcal{G} = \{-, A, B, AA, AB, BB, AAA, AAB, \ldots\}$, - the genotype of the normal, reference and variant populations w.r.t. mutation n.
  - AAB - genotype with two reference alleles and one variant allele.
  - $-$ genotype with no alleles, a homogeneous deletion of the locus.
- In practice, we choose finite $\mathcal{G}$.
- Vector $\psi^n = (g_N^n, g_R^n, g_V^n) \in \mathcal{G}^n$ - state for mutation n
- π^n - vector of prior probabilities over all possible states ψ^n for the nth mutation.

The probability of sampling a variant allele

- For a genotype $g \in \mathcal{G}$, $c(g) : \mathcal{G} \to \mathbb{N}$ - copy number of the genotype.
  - $c(AAB) = 3$
- $b(g) : \mathcal{G} \to \mathbb{N}$ - number of variant alleles in the genotype
  - $b(AAB) = 1$
- The probability of sampling a variant allele from a cell with genotype g is given by $\mu(g) = \frac{b(g)}{c(g)}$.
- When $b(g) = 0$, $\mu(g) = \epsilon$, where ϵ is the false positive probability of observing allele B when the true allele was A (sequencing error).
- When $b(g) = c(g)$, $\mu(g) = 1 - \epsilon$.

The probability of sampling a read

- Let t - fraction of tumor cells (tumor content).
- The probability of sampling a read
  - containing the variant allele
  - covering a mutation with
    - state $\psi = (g_N, g_R, g_V)$ and
    - cellular prevalence ϕ

$$\zeta(\psi, \phi, t) = \frac{(1-t)c(g_N)}{Z}\mu(g_N) + \frac{t(1-\phi)c(g_R)}{Z}\mu(g_R) + \frac{t\phi c(g_V)}{Z}\mu(g_V)$$

where

$$Z = (1-t)c(g_N) + t(1-\phi)c(g_R) + t\phi c(g_V)$$

Posterior distribution of the prevalences

Let

- b^n number of reads observed with the B allele of the nth mutation
- d^n total number of reads covering the locus of the nth mutation
- $b^n \sim \text{Binomial}(d^n, \zeta(\psi_n, \phi_n, t))$.

Posterior distribution for prevalences $\phi = (\phi^1, \ldots, \phi^N)$

$$
\begin{aligned}
p(\phi|b, d, \pi, t) &\propto p(\phi) \prod_{n=1}^{N} p(b^n|\phi^n, d^n, \pi^n, t) \\
&= p(\phi) \prod_{n=1}^{N} \sum_{\psi^n \in \mathcal{G}^3} p(b^n|\phi^n, d^n, \psi^n, t) p(\psi^n|\pi^n) \\
&= p(\phi) \prod_{n=1}^{N} \sum_{\psi^n \in \mathcal{G}^3} \text{Binomial}(d^n, \zeta(\psi_n, \phi_n, t)) \pi^n_{\psi^n}
\end{aligned}
$$

Dirichlet process prior on cellular prevalences

- Cellular prevalences ϕ should cluster by the clones.
- The prior on prevalences chosen as

$$\phi \sim DP(\text{Uniform}(0, 1), \alpha)$$

- If $\text{Uniform}(0, 1)$ (or any continuous probability) used as a prior, the prevalences of all mutations would be different with probability one.
- DP prior converts $\text{Uniform}(0, 1)$ into a discrete distribution - mutations share the same prevalence with non-zero probability.

Dirichlet process prior on cellular prevalences

- Cellular prevalences ϕ should cluster by the clones.
- The prior on prevalences chosen as

$$\phi \sim DP(\mathrm{Uniform}(0,1), \alpha)$$

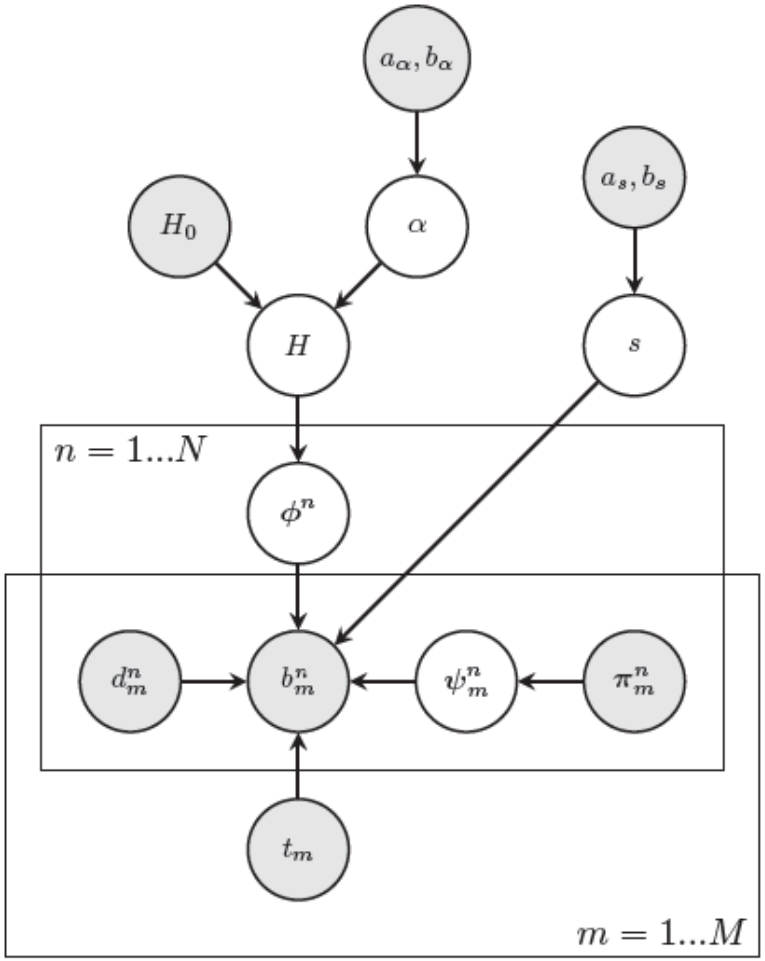
- If $\mathrm{Uniform}(0,1)$ (or any continuous probability) used as a prior, the prevalences of all mutations would be different with probability one.
- DP prior converts $\mathrm{Uniform}(0,1)$ into a discrete distribution - mutations share the same prevalence with non-zero probability.
- The prior on concentration parameter α is chosen as

$$\alpha \sim \mathrm{Gamma}(a, b)$$

with the density

$$p(\alpha|a,b) = \frac{b^a \alpha^{\alpha-1} \exp(-b\alpha)}{\Gamma(a).}$$

The PyClone graphical model



$$
\begin{aligned}
\alpha &\sim \text{Gamma}(a_\alpha, b_\alpha) \\
H_0 &= \text{Uniform}([0,1]^M) \\
H|\alpha, H_0 &\sim \text{DP}(\alpha, H_0) \\
\phi^n|H &\sim H \\
\psi_m^n|\pi_m^n &\sim \text{Categorical}(\pi_m^n) \\
\psi_m^n &= (g_{m,\text{N}}^n, g_{m,\text{R}}^n, g_{m,\text{V}}^n)
\end{aligned}
$$

either

$$
b_m^n|d_m^n, \psi_m^n, \phi_m^n, t_m \sim \text{Binomial}(d_m^n, \xi(\psi_m^n, \phi_m^n, t_m))
$$

or

$$
\begin{aligned}
s|a, b &\sim \text{Gamma}(a_s, b_s) \\
b_m^n|d_m^n, \psi_m^n, \phi_m^n, t_m, s &\sim \text{BetaBinomial}(d_m^n, \xi(\psi_m^n, \phi_m^n, t_m), s)
\end{aligned}
$$

where

$$
\begin{aligned}
\xi(\psi, \phi, t) &= \tfrac{(1-t)c(g_\text{N})}{Z}\mu(g_\text{N}) + \tfrac{t(1-\phi)c(g_\text{R})}{Z}\mu(g_\text{R}) + \\
&\quad \tfrac{t\phi c(g_\text{V})}{Z}\mu(g_\text{V}) \\
Z &= (1-t)c(g_\text{N}) + t(1-\phi)c(g_\text{R}) + \\
&\quad t\phi c(g_\text{V})
\end{aligned}
$$

Further reading

- Yee Whye Teh, Dirichlet Process
- Wikipedia
- Roth *et al.*, PyClone: Statistical inference of clonal population structure in cancer, Nat Methods, 2016