

CS4330: Combinatorial Methods in Bioinformatics

Genome characteristics estimation using K-mers

Wong Limsoon

Acknowledgement: This set of slides were adapted from Ken Sung's



NUS
National University
of Singapore

National University of Singapore

Genome characteristics



$$\text{Percentage of Repeat Content} = \left(\frac{\text{Total Length of Repeats}}{\text{Total Genome Length}} \right) \times 100$$

$$\text{Heterozygous Rate} = \left(\frac{\text{Number of Heterozygous Positions}}{\text{Total Number of Analyzed Positions}} \right) \times 100$$

Exercise

Paternal TTCGGAAGCTACAGTCACACACAGACGTCGATCAGCTTCATGGACAGCTTCAGTAA

Maternal TTCGGAAGCTTCAGTCACACACAGACGCCGATCAGCTTCATGGACAGCTTCAGTAA

Green is repeat region

Red is heterozygous bases

Compute the followings:

Genome size

Percentage of repeat content

Heterozygous rate

Homozygous repeat-free genome

In a homozygous repeat-free genome, most K-mers occurring in it have similar counts

Unique K-mers of a genome are K-mers occurring exactly once in the genome / in each read

Example

In this genome, all 4-mers are “unique”

```
TACTGCATGCCGCACT
TACT
ACTG
CTGC
TGCA
GCAT
CATG
ATGC
TGCC
GCCG
CCGC
CGCA
GCAG
CAGT
```

Genome size estimation

Suppose...

G = genome size, i.e. length of the haploid genome

L = mean read length

N = # of reads

Then,

If C = sequencing coverage is known, then $G \approx N L / C$

But estimating C is resource demanding as the reads have to be aligned and then get the average number of reads aligned to each position in the consensus genome

Genome size estimation by K-mers

Suppose..

G = size of a homozygous repeat-free genome

L = mean read length

N = # of reads

μ = mean K -mer count in the reads covering a base

Then,

$\mu G \approx \text{Total } K\text{-mer counts in the reads} = N (L - K + 1)$

$\Rightarrow G \approx N (L - K + 1) / \mu$



Estimating genome size was easy for homozygous repeat-free genomes

Real genomes are hardly ever homozygous repeat-free

Needs modelling ... using K-mer spectrum

K-mer spectrum

K-mer spectrum is distribution of K-mer counts in a given set of DNA sequences

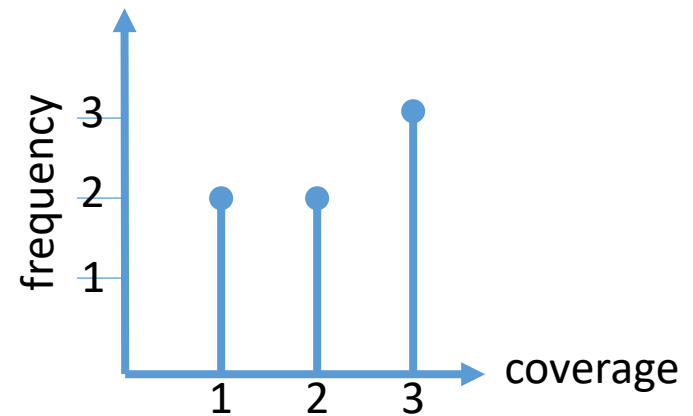
K-mer spectrum is often visualized as a histogram
x-axis = counts of diff K-mers
y-axis = # of K-mers with a specific count

ACGTC
CGTCA
GTCAA
TCAAG
CAAGT

Reads

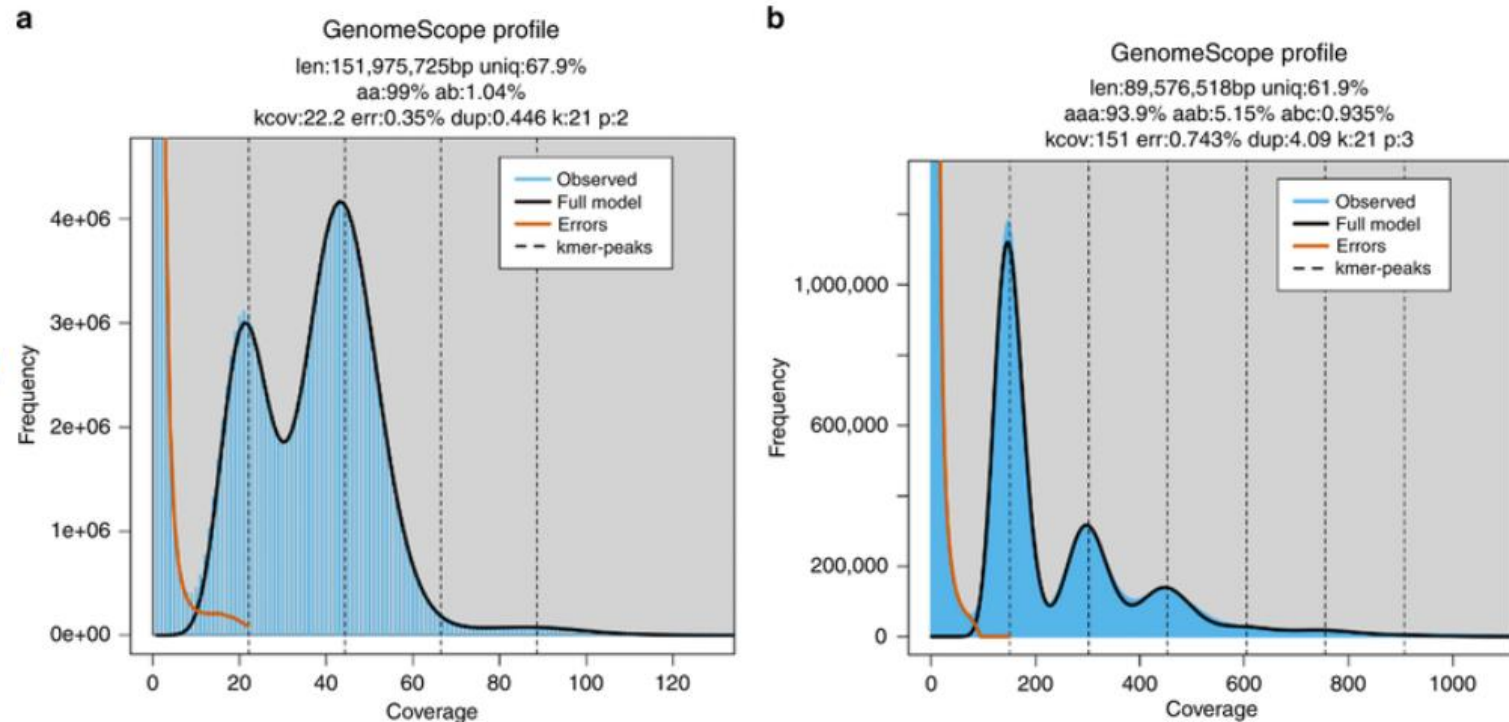
k-mer	count
ACG	1
CGT	2
GTC	3
TCA	3
CAA	3
AAG	2
AGT	1

K-mer counts



K-mer spectrum

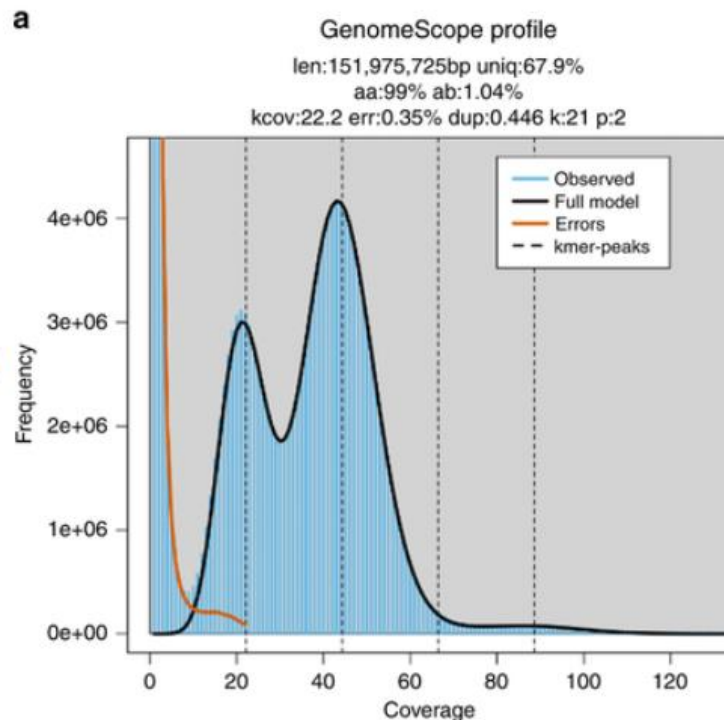
K-mer spectra of heterozygous diploid & triploid genomes



GenomeScope plots for heterozygous species K-mer spectra and fitted models for (a) diploid *Arabidopsis thaliana* and (b) triploid *Meloidogyne enterolobii*. Note that the diploid plot has two major peaks, while the triploid plot has three major peaks. Both also have high frequency putative error k-mers with coverage near 1.

Exercise

Given this K-mer spectrum for a diploid genome

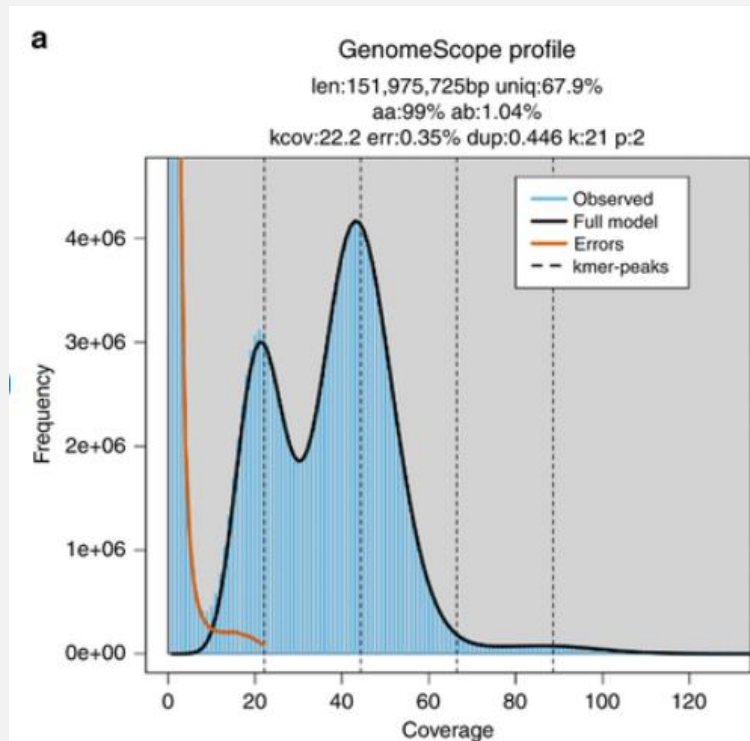


Which peak corresponds to K-mers covering homozygous bases?

Which peak corresponds to K-mers covering heterozygous bases?

What is the sequencing coverage?

Modelling observed K-mer spectrum



GenomeScope fits a theoretical model (black curve) to the observed K-mer spectrum (blue histogram)

Genome size (~152B), heterozygous rate (1.04%), etc. are then extracted from parameters of the fitted model

Let's see how this is done...

K-mer spectrum of a homozygous repeat-free genome

Suppose ...

And also no
sequencing bias

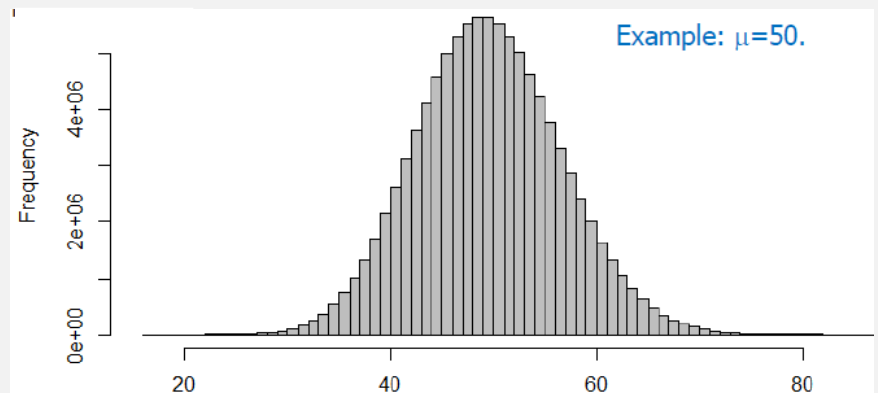
*No sequencing error, no
heterozygosity, no repeat*

*K-mers are randomly
extracted from the genome*

Then,

*K-mer spectrum is a
Poisson distribution having
 μ = the mean K-mer count*

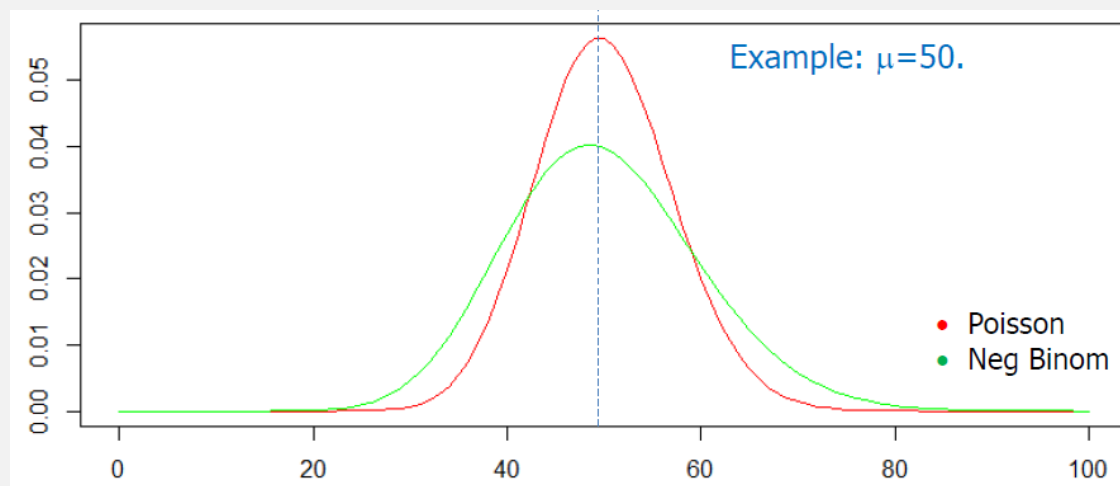
Lander & Waterman, "Genomic mapping by fingerprinting
random clones: a mathematical analysis", *Genomics*
2(3):231-239, 1988



Sometimes, $\text{Poisson}(\mu)$ does not fit well...

Real sequencing data is a bit over-dispersed compared to Poisson, due to e.g. GC bias in sequencing

Negative binomial $\text{NB}(\mu, \mu / \rho)$ is used instead, where ρ is a variant parameter that controls over-dispersion



Negative binomial

Imagine a sequence of independent [Bernoulli trials](#): each trial has two potential outcomes called "success" and "failure." In each trial the probability of success is p and of failure is $1 - p$. We observe this sequence until a predefined number r of successes occurs. Then the random number of observed failures, X , follows the **negative binomial** (or **Pascal**) distribution:

$$X \sim \text{NB}(r, p)$$

Probability mass function [\[edit \]](#)

The [probability mass function](#) of the negative binomial distribution is

$$f(k; r, p) \equiv \Pr(X = k) = \binom{k + r - 1}{k} (1 - p)^k p^r$$

where r is the number of successes, k is the number of failures, and p is the probability of success on each trial.

Taken from Wikipedia

Use the pmf, $f(c; \mu, \mu / \rho)$, of a negative binomial to model the prob of a random K-mer having coverage c , where μ is the observed mean K-mer coverage and ρ a fitted parameter

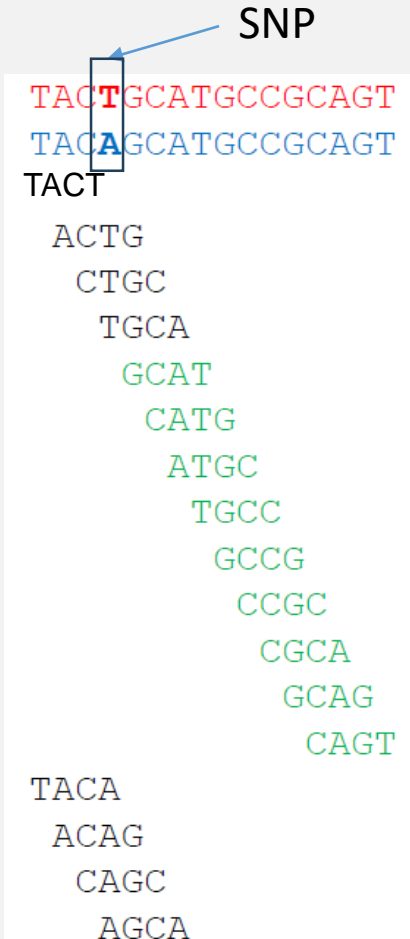
Do this separately for each kind of K-mers: homozygous, heterozygous, 2-copy repeats, 3-copy repeats

Repeat-free diploid genome

This is a diploid genome
where all K-mers are unique

One heterozygous base gives
2K heterozygous K-mers

K = 4
The SNP creates 8 (= 2K) 4-mers



K-mer spectrum of repeat-free diploid genome

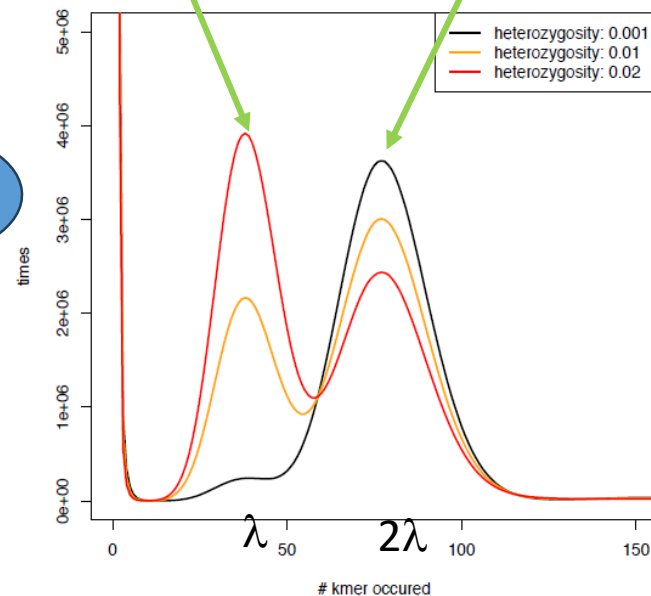
If a genome is heterozygous and repeat-free, there are two peaks at K-mer coverage λ and 2λ

Why?

As one heterozygous base creates $2K$ heterozygous K-mers, the heterozygous peak grows fast

heterozygous K-mers

Homozygous K-mers



Supplementary Figure 1. Impact of heterozygosity on the *k*-mer profile. *K*-mer profiles were drawn from 100x sequencing coverage of simulated reads with 0.1%, 1% and 2% heterozygosity embedded into the *D. melanogaster* reference genome.

Vurture et al., "GenomeScope", *Bioinformatics* 33(14):2202-2204, 2017

Homozygous vs heterozygous K-mers

Consider a repeat-free diploid genome

Let r = heterozygosity rate

Then,

$(1 - r)^K$ = prob that a random K-mer is homozygous

$1 - (1 - r)^K$ = prob that a random K-mer is heterozygous

Homozygous vs heterozygous K-mers

Let α = proportion of heterozygous K-mers wrt genome size

Let β = proportion of homozygous K-mers wrt genome size

Then,

$$\alpha = 2 (1 - (1 - r)^K)$$

$$\beta = (1 - r)^K$$

If instead the diploid genome has a non-zero heterozygosity rate r , then those heterozygous bases will create additional *k-mers* beyond the original G *k-mers*. Note that if r is the probability that a given base is heterozygous, then $1-r$ is the probability that a given base is not heterozygous (i.e. homozygous). Furthermore, $(1-r)^K$ is the probability that a given *k-mer* is homozygous, and $1-(1-r)^K$ is the probability that a *k-mer* is heterozygous in at least once nucleotide. As a result, there will be $G \cdot (1-r)^K$ homozygous *k-mers* and $2 \cdot G \cdot (1-(1-r)^K)$ heterozygous *k-mers*. Of the heterozygous *k-mers*, $G \cdot (1-(1-r)^K)$ will originate on the maternal haplotype and an additional $G \cdot (1-(1-r)^K)$ *k-mers* will originate on the paternal haplotype. Consequently, the total number of *k-mers* present in the diploid genome will no longer be G , but rather will depend on the rate of heterozygosity and equal $(1+(1-(1-r)^K)) \cdot G$. At high rates of heterozygosity near 100%, the total number of *k-mers* present in the diploid genome will equal $2 \cdot G$ meaning that that every *k-mer* in the maternal and paternal haplotypes is different.

Vurture et al., "GenomeScope", *Bioinformatics* 33(14):2202-2204, 2017

A model of K-mer spectrum for repeat-free diploid genome

$$F(X) = \alpha \text{NB}(X; \lambda, \lambda / \rho) + \beta \text{NB}(X; 2\lambda, 2\lambda / \rho)$$

X coverage values

λ mean heterozygous K-mer coverage

ρ dispersion parameter

- Example: $r=0.01, \rho=0.5$.

- Heterozygous k-mers: $\alpha = 2(1-(1-0.01)^{21})=0.38$.

- Homozygous k-mers: $\beta = (1 - 0.01)^{21}=0.81$.

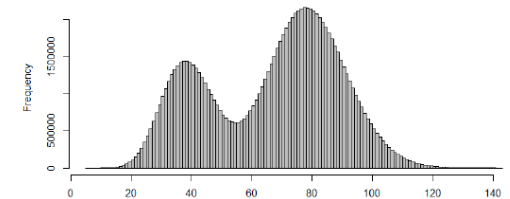
100x sequencing coverage, $k=21$

$0.38 * \text{NB}(40, 80) + 0.81 * \text{NB}(80, 160)$

- Let the base coverage be $C=100, L=100, k=21$.

- k-mer coverage = $C(L-k+1)/L = 80$

- Hence, $\lambda = 80/2 = 40$.



- Example: $r=0.02, \rho=0.5$.

- Heterozygous k-mers: $\alpha = 2(1-(1-0.02)^{21})=0.69$.

- Homozygous k-mers: $\beta = (1 - 0.02)^{21}=0.65$.

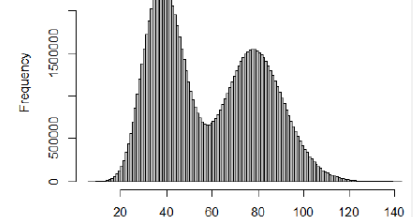
100x sequencing coverage, $k=21$

$0.69 * \text{NB}(40, 80) + 0.65 * \text{NB}(80, 160)$

- Let the base coverage be $C=100, L=100, k=21$.

- k-mer coverage = $C(L-k+1)/L = 80$

- Hence, $\lambda = 80/2 = 40$.



Estimating genome characteristics

Once the model is fitted to the observed K-mer spectrum

Heterozygous rate is obtained as the value of r used in defining α and β

Genome size is obtained by summing total # of K-mers and dividing by 2λ , the estimated mean coverage of homozygous K-mers



Why?

GenomeScope

In general, a genome may have repeats

GenomeScope fits a mixture of four evenly spaced negative binomial distributions to the K-mer spectrum to model the relative abundances of heterozygous, homozygous, and two-copy repeats of various types

GenomeScope only models 2-copy repeats

For non-repeats:

α = proportion of unique heterozygous K-mers

Each K-mer has 1 copy

β = proportion of unique homozygous K-mers

Each K-mer has 2 copies

r = heterozygosity rate

For 2-copy repeats:

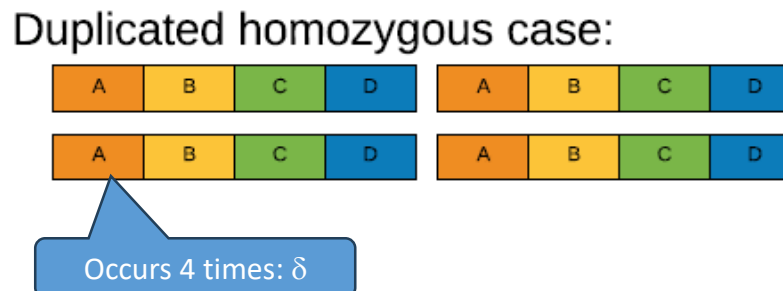
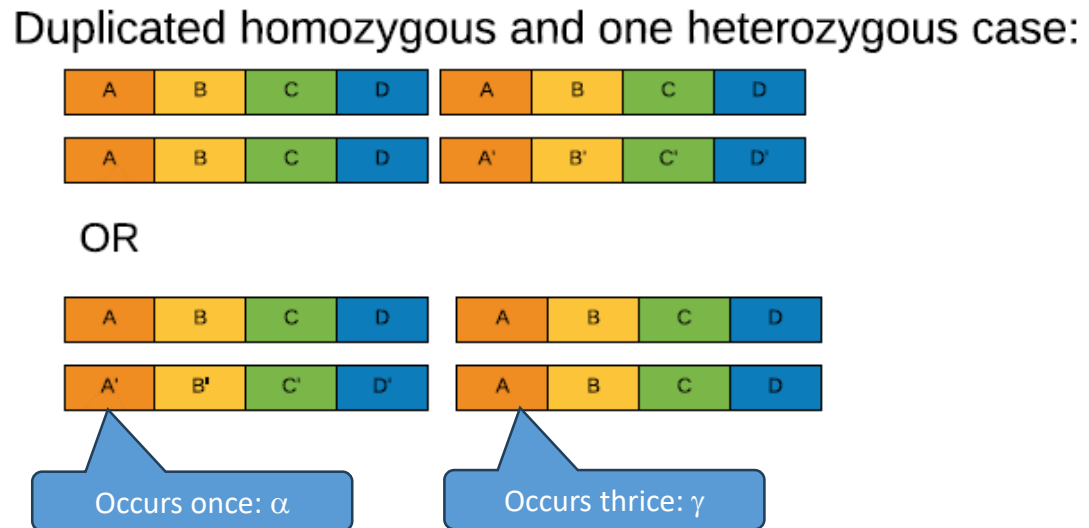
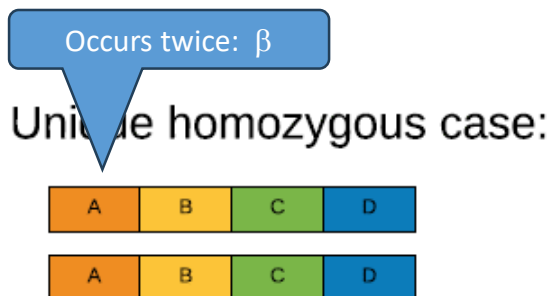
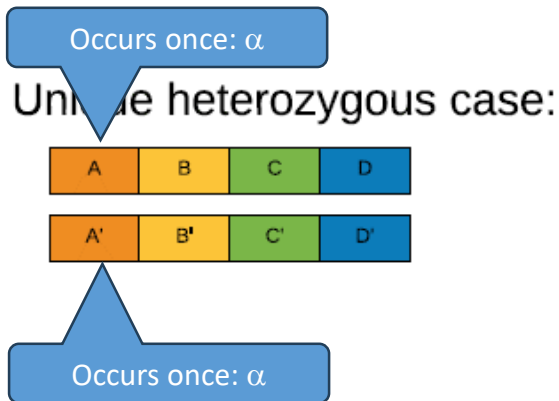
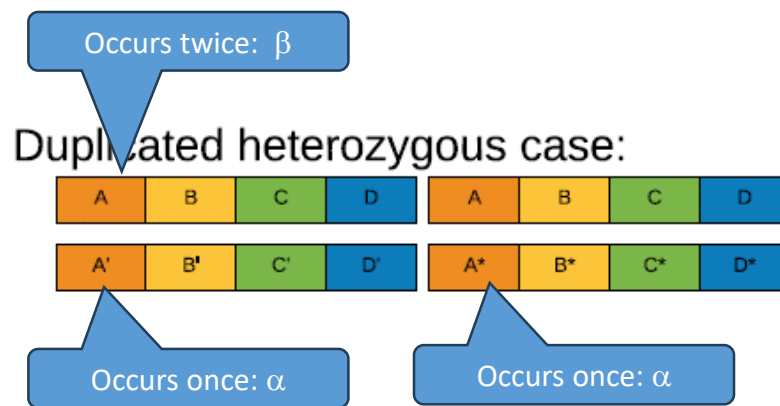
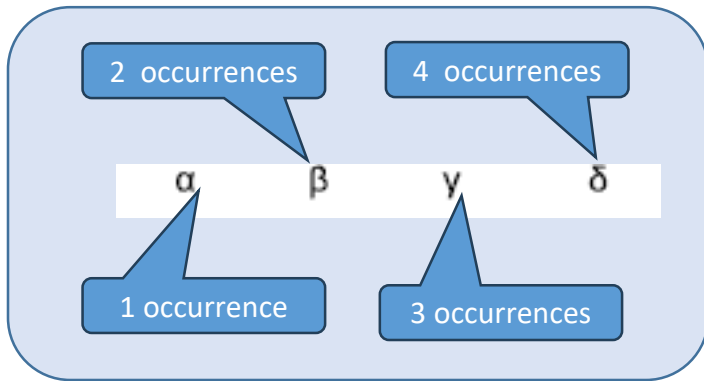
γ = proportion of duplicated heterozygous K-mers

Each K-mer has 3 copies

δ = proportion of duplicated homozygous K-mers

Each K-mer has 4 copies

d = proportion of repeat regions in the genome



Unique heterozygous K-mers

Unique heterozygous case:



total contribution to α peak: $2(1-d)(1-(1-r)^k)$

$$\alpha = 2 (1 - d) (1 - (1 - r)^K) + \dots$$

Non-repeat

Heterozygous



α β γ δ

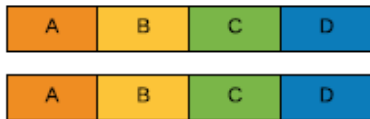
1 occurrence

Legend:

kmer: X mutated kmer: X' X*

Unique homozygous K-mers

Unique homozygous case:

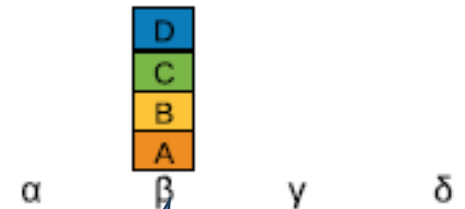


total contribution to β peak: $(1-d)((1-r)^k)$

$$\beta = (1 - d) (1 - r)^K + \dots$$

Non-repeat

Homozygous



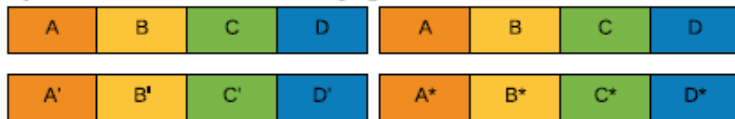
2 occurrences

Legend:

kmer: X mutated kmer: X' X*

Duplicated heterozygous K-mers

Duplicated heterozygous case:



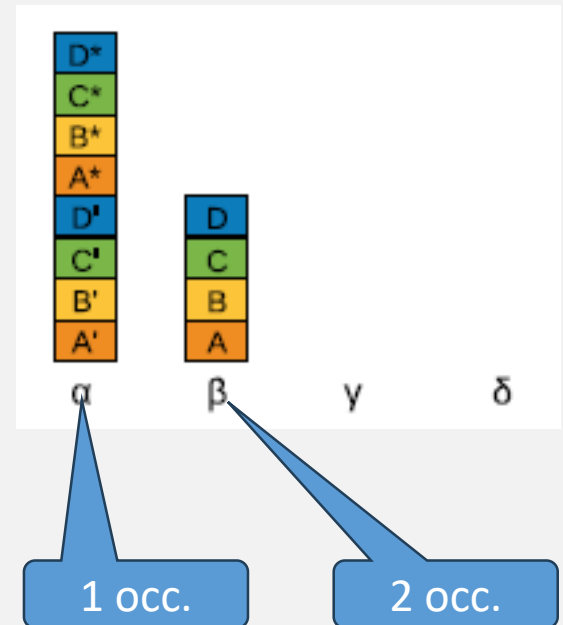
total contribution to α peak $2d(1-(1-r)^k)^2$ and β peak $d(1-(1-r)^k)^2$

$$\alpha = 2 d (1 - (1 - r)^K)^2 + \dots$$

$$\beta = d (1 - (1 - r)^K)^2 + \dots$$

Repeat

Heterozygous

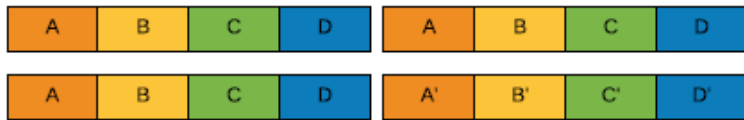


Legend:

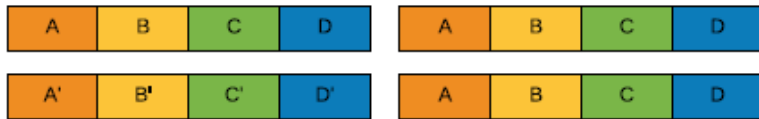
kmer: X mutated kmer: X' X*

Duplicated mixed homozygous heterozygous K-mers

Duplicated homozygous and one heterozygous case:



OR



total contribution to α peak $2d((1-r)^k)(1-(1-r)^k)$ and γ peak $2d((1-r)^k)(1-(1-r)^k)$

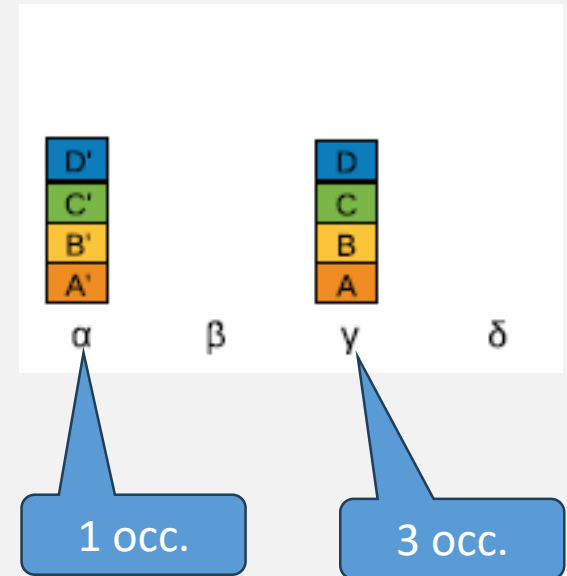
$$\alpha = 2 d (1 - r)^K (1 - (1 - r)^K) + \dots$$

$$\gamma = 2 d (1 - r)^K (1 - (1 - r)^K) + \dots$$

Repeat

Homozygous

Heterozygous



Legend:

kmer:

X

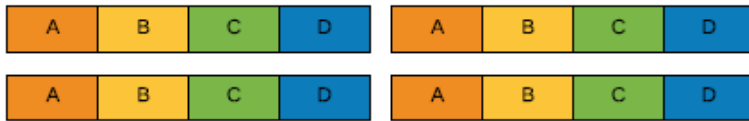
mutated
kmer:

X'

X*

Duplicated homozygous K-mers

Duplicated homozygous case:



total contribution to δ peak: $d(1-r)^{2k}$

$$\delta = d (1 - r)^{2K} + \dots$$

Repeat

Homozygous



4 occurrences

Legend:

kmer:

X

mutated
kmer:

X'

X''

In summary

GenomeScope fits the K-mer spectrum by a mixture of four negative binomials spaced at λ , 2λ , 3λ , and 4λ :

$$F(X) = G * (\alpha \text{NB}(X; \lambda, \lambda / \rho) + \beta \text{NB}(X; 2\lambda, 2\lambda / \rho) + \gamma \text{NB}(X; 3\lambda, 3\lambda / \rho) + \delta \text{NB}(X; 4\lambda, 4\lambda / \rho))$$

G is scaling parameter corresponding to genome size

$$\alpha = 2(1-d)(1-(1-r)^K) + 2d(1-(1-r)^K)^2 + 2d(1-r)^K(1-(1-r)^K)$$

$$\beta = (1-d)(1-r)^K + d(1-(1-r)^K)^2$$

$$\gamma = 2d(1-r)^K(1-(1-r)^K)$$

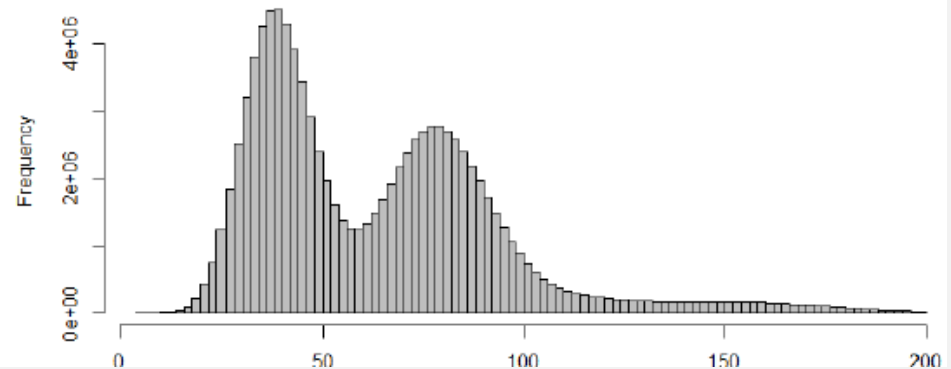
$$\delta = d(1-r)^{2K}$$

Example

- Example: $r=0.02$, $d=0.1$, $\rho=0.5$.
- $\alpha = 0.6914884$
- $\beta = 0.6007841$
- $\gamma = 0.04524103$
- $\delta = 0.6007841$
- Let the base coverage be $C=100$.
 $L=100$. $k=21$.
- k-mer coverage = $C(L-k+1)/L = 80$
- Hence, $\lambda = 80/2 = 40$.

100x sequencing coverage, $k=21$

$$0.691 * \text{NB}(40, 80) + 0.397 * \text{NB}(80, 160) + 0.05 * \text{NB}(120, 240) + 0.04 * \text{NB}(160, 320)$$

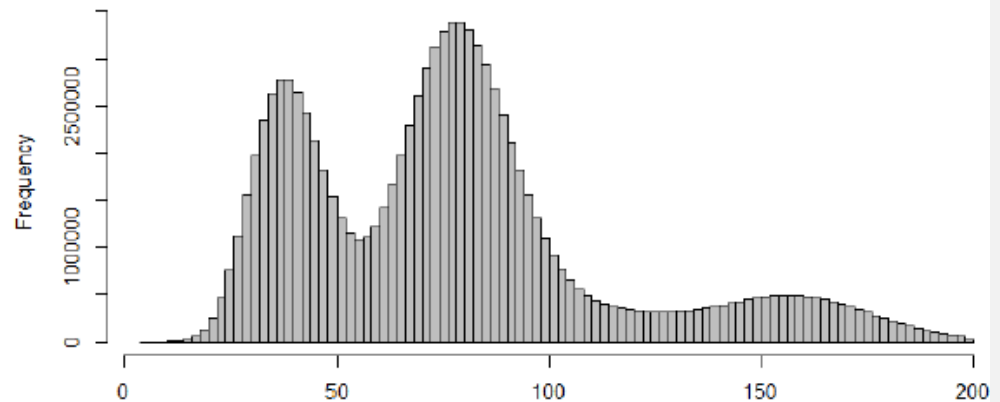


Example

- Example: $r=0.01$, $d=0.2$, $\rho=0.5$.
- $\alpha = 0.3805443$
- $\beta = 0.655023$
- $\gamma = 0.06162746$
- $\delta = 0.1311318$
- Let the base coverage be $C=100$.
 $L=100$. $k=21$.
- k -mer coverage = $C(L-k+1)/L = 80$.
- Hence, $\lambda = 80/2 = 40$.

100x sequencing coverage, $k=21$

$$0.344 * \text{NB}(40, 80) + 0.655 * \text{NB}(80, 160) + 0.06 * \text{NB}(120, 240) + 0.131 * \text{NB}(160, 320)$$

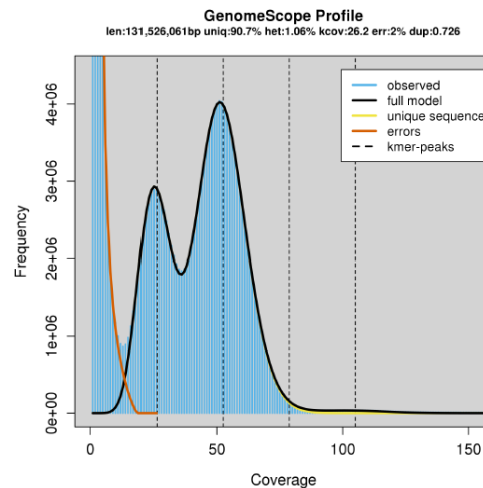


How genome characteristics are estimated

Perform K-mer counting to get empirical K-mer spectrum

Estimate d , r , λ , G to fit $F(X)$ to the empirical distribution

Genome size= G .
Percentage of repeat content= r
Heterozygous rate= d
Coverage of haplotype= λ



$G=131,526,061$
 $(1-d)=90.7\%$
 $r=1.06\%$
 $\lambda=26.2$

GenomeScope
modelling
results on *D.
melanogaster*

Supplementary Figure 4. Modeling results on *D. melanogaster*. The sequencing errors are identified by low coverage *k*-mers not explained by the model (shown in orange). This way a single cutoff value does not need to be used nor does it assume a particular shape to the distribution of the error *k*-mers. See below for more details on the *D. melanogaster* analysis.

Vurture et al., *Bioinformatics* 33(14):2202-2204, 2017

Estimation of parameters

Initial model

$d = 0, r = 0, \rho = 0.5, \lambda = \text{estKmerCov}, G = \text{estGenomeSize}$

estKmerCov is coverage w/ max height in K-mer spectrum, after excluding low-coverage sequencing errors and K-mers with coverage > CovMax

estGenomeSize = # of observed K-mers / estKmerCov

Iterate

Based on previous model, remove low-coverage error K-mers & K-mer with coverage > CovMax

Minimize least square error to optimize d, r, ρ, λ

Set $G = \# \text{ of K-mers excluding errors} / 2\lambda$

Limitations of GenomeScope

Require decent sequencing coverage, $> 25x$

Require low error rate \Rightarrow cannot support long-read sequencing like ONT

Cannot support polyploid genomes (this is fixed in GenomeScope2.0)

Cannot support genomes having non-uniform copy number of their chromosomes (e.g. leukemia patients)

Good to read

The GenomeScope paper, esp. its supplementary material

G. W. Vurture et al, “GenomeScope: Fast reference-free genome profiling from short reads”, *Bioinformatics* 33(14):2202-2204, 2017.

<https://doi.org/10.1093%2Fbioinformatics%2Fbtx153>