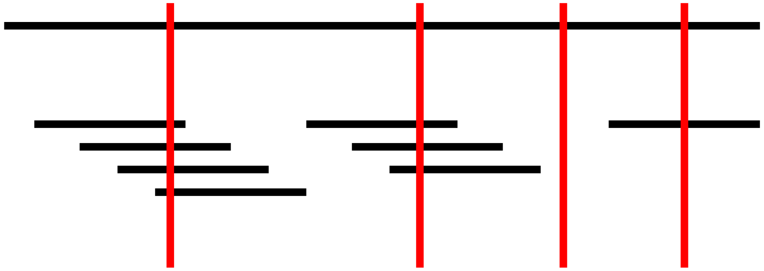
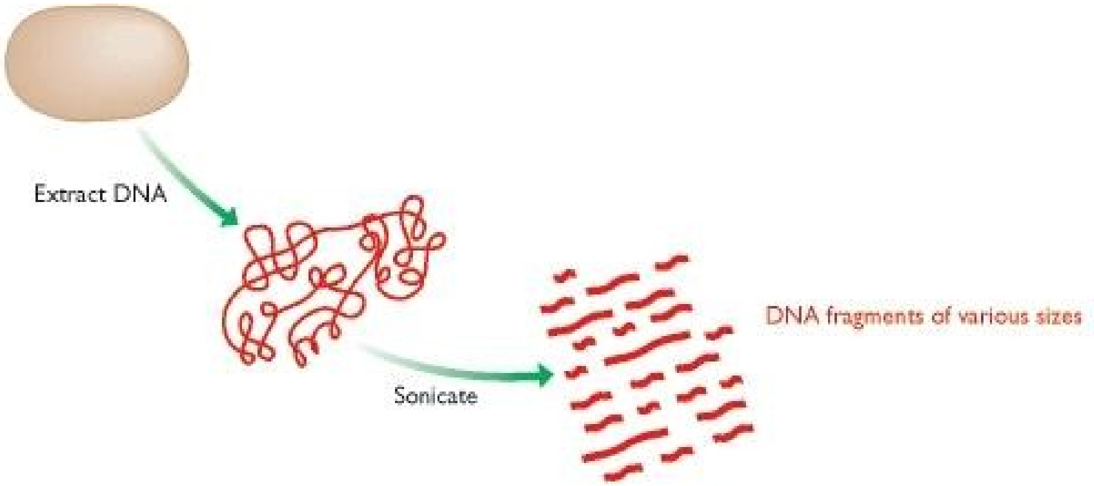
Genome Coverage



Genome

Local coverage: 4 3 0 1

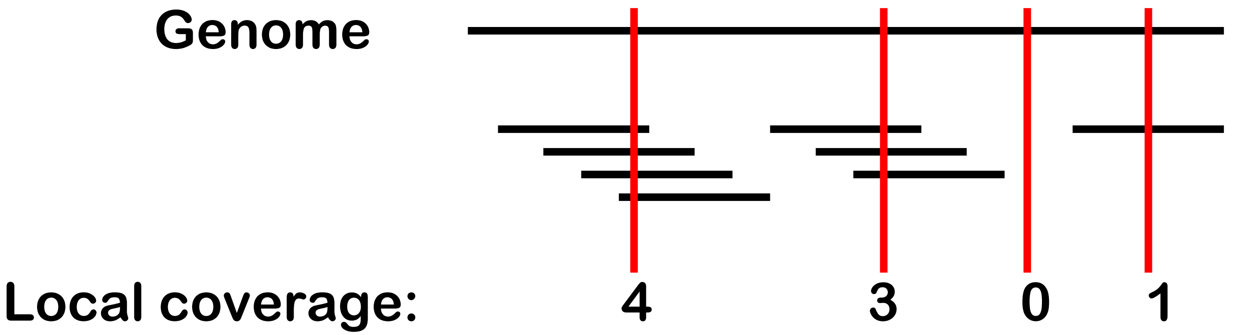
Shotgun sequencing involved random fragmentation of the genome



The first 36 bases from each fragment are sequenced

Each read is a sequence of length 36 randomly extracted from the genome

The global genome coverage of the reads is the average number of reads that cover one specific position in the genome



G: Length of genome
L: Length of reads
N: Number of reads

Global coverage: $a = NL/G$

The answers to the following questions all depend on the outcome of random fragmentation

How many contigs will I get?
How big will the contigs be?
What percentage of the genome will remain uncovered?
How long can gaps be?

How many reads do I need to get 99% of the genome covered?

Random fragmentation is a random experiment like …

Flipping a coin, Throwing a die, Playing roulette

Generating a random number with the computer's random generator

The mutation of a specific base in the genome

The collision of diffusing molecules in the cell (and hence their binding)

Finding a mutation in a certain gene in 100 randomly chosen individuals

Outcomes of random experiments are associated with probabilities

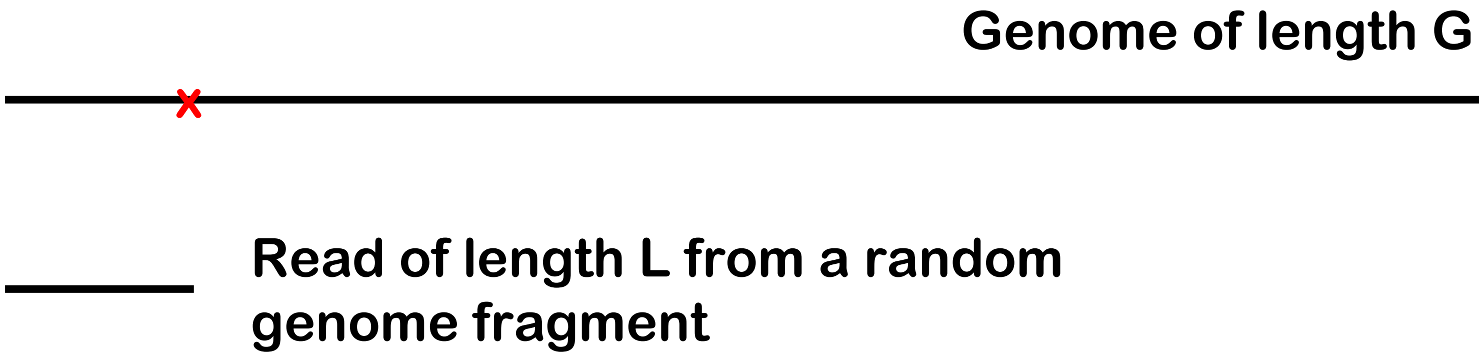
The probability that a coin falls “head” is 1/2
The probability that a die falls “3” is 1/6

Assume that the random generator of your computer generates every number between 0 and 1 with the same probability:

The probability that this number is smaller than 0.341 is 0.341

The probability that a newborn is a boy is roughly 0.5

How can we get a grip on random genome fragmentation?



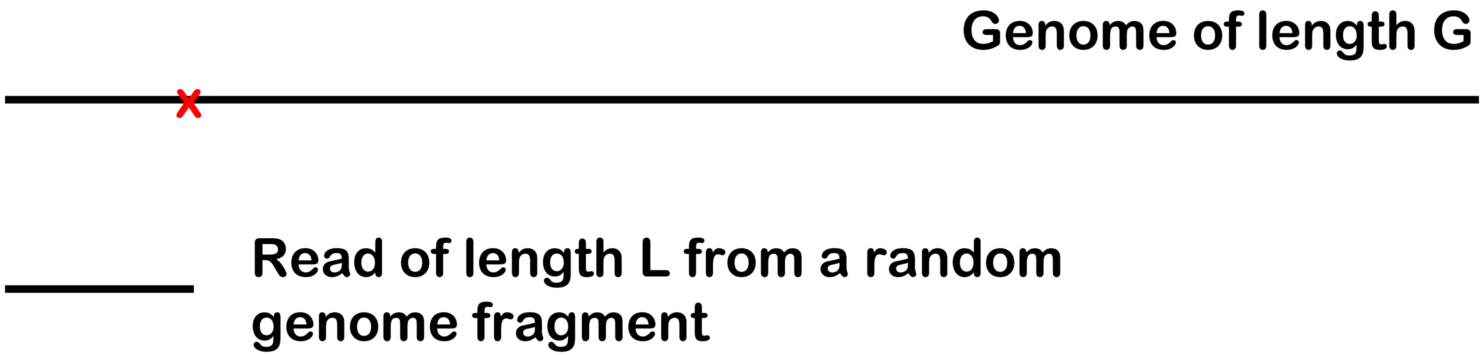
Genome of length G

Read of length L from a random genome fragment

What is the probability that the read starts at x ?

What is the probability that the read covers x ?

The read can begin at any position of the genome



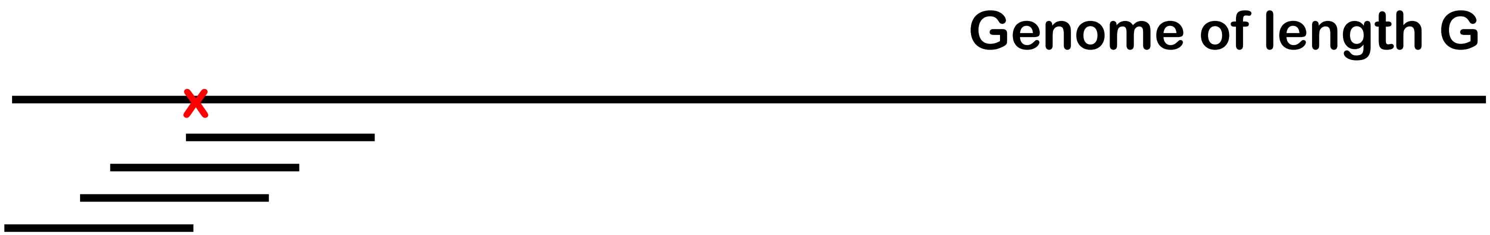
Genome of length G

Read of length L from a random genome fragment

What is the probability that the read starts at x ?

- The genome has G bases
- This leaves G-L+1 possible start positions for the read
- We assume that all start positions are equally likely
- The probability for each is 1/(G-L+1)
- Since G is large and L is small, this probability is about 1/G

There are L possible start positions for a read that covers x



What is the probability that the read covers x ?

L out of G-L+1 start positions lead to coverage of x

Ignoring end effects the probability is L/G

Which formalism are we following in these computations?

Events are sets of possible outcomes of a random experiment

Ω is the set of all possible outcomes

Die: Ω={1,2,3,4,5,6}

Start position of a read: Ω ={1,2,...,G-L+1}

Outcomes can be combined to events

Die: An even number

A={2,4,6}

Start position of a read covering x

A={x, x-1, x-2, … ,x-L+1}

Events are subsets of Ω

Random experiments can be easily simulated using the random generator of a computer

Ω

```
> sample(c(1,2,3,4,5,6),size=1, replace=TRUE)
[1] 6
> sample(c(1,2,3,4,5,6),size=1, replace=TRUE)
[1] 1
> sample(c(1,2,3,4,5,6),size=1, replace=TRUE)
[1] 6
```

The Laplace Model assumes that all outcomes have the same probability

For an event A the Laplace Model defines its probability by

$$P(A) = |A| / |\Omega|$$

where

|A| is the number of outcomes in A

$|\Omega|$ is the number of all possible outcomes

Die:{1,2,3,4,5,6} Probability: 1/6

Start position of read {1,2,…,G-L+1}

Probability: 1/(G-L+1)

Any Ω with N elements Probability: 1/N

The logical OR can be implemented by the union of events

If A and B are events, then we can define the event

$$C = A \cup B$$

which occurs if either A or B occur (or both A and B occur)

Reads: A=read starts at x

B=read starts at x-1 or x+1

C= the start of the read is at most 1 base away from x

The probability of the union of events can be calculated from the probabilities of the individual events

A and B are events and $C = A \cup B$

Laplace: $P(C) = |C| / |\Omega|$

$P(A) = |A| / |\Omega|$

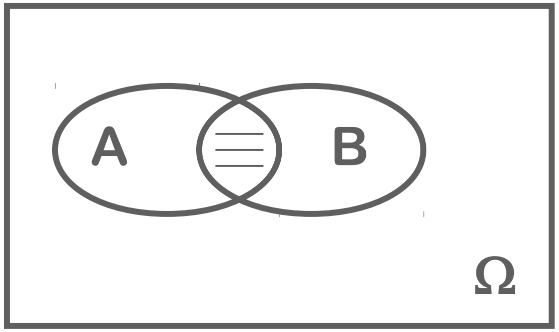
$P(B) = |B| / |\Omega|$

$P(C) = |A \cup B| / |\Omega|$

$= P(A)+P(B)-P(A \cap B)$

If A and B are disjoint (mutually exclusive):

$P(A \cup B) = P(A) + P(B)$



The subtracted term $P(A \cap B)$ corrects for counting the intersection twice

Mathematicians like to use the same language for trivialities then they use for relevant results

Ω is a set of outcomes too. Hence it is an event.

Its probability is: $P(\Omega) = |\Omega| / |\Omega| = 1$

Ω is a certain event. It occurs with probability 1.

The empty set $\varnothing$ is a set of outcomes too.

It has zero elements.

Its probability is: $P(\varnothing) = |\varnothing| / |\Omega| = 0$

$\varnothing$ is an impossible event. It occurs with probability 0.

The logical “but not” can be implemented by the difference of sets

If A and B are events, then we can define the event

$$C = A \setminus B$$

which occurs if A but not B occurs

Reads: A= read starts at most 2 bases away from x

B= read starts at x-1 or x+1

C= read start is in {x-2, x, x+2}

Probabilities of event differences follow directly from the Laplace model

$P(A \setminus B) = P(A) - P(A \cap B)$

Reads: A= read starts at most 2 bases away from x

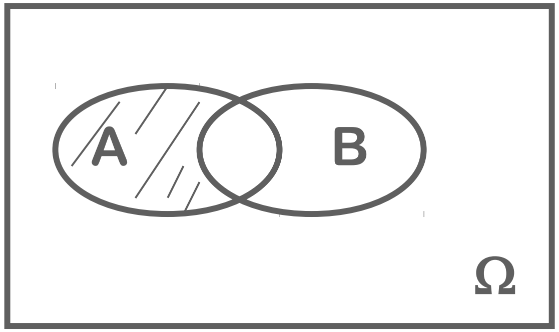
B= read starts at x-1 or x+1

C= read start is in {x-2,x,x+2}

P(A) = 5/G

P(B) = 2/G

P(C) = (5-2)/G = 3/G



The logical NOT can be implemented by the complement of events

If A is an event, we can define the event

$$A^c = \Omega \setminus A$$

which occurs if A does not occur

Reads: A = read covers x

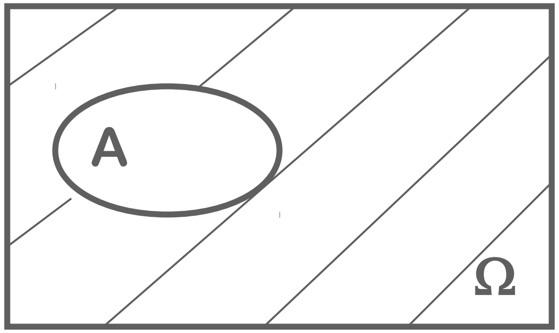
A^c = read does not cover x

Probabilities of complementary events follow from the Laplace Model

$P(A^c) = P(\Omega \setminus A)$

$= P(\Omega) - P(\Omega \cap A)$

$= 1 - P(A)$



Reads: A= read covers x

A^c = read does not cover x

$P(A) = L/G$

$P(A^c) = 1 - L/G$

If events depend on several independent random experiments we can use vectors as elements from Ω

Two Dice: Ω = {(1,1), (1,2), ..., (6,5), (6,6)}

Outcome = (5,3): Die1 = 5 and Die2 = 3

A: Die1 = 6: {(6,1), (6,2), ..., (6,6)}

B: Die2 = 6: {(1,6), (2,6), ..., (6,6)}

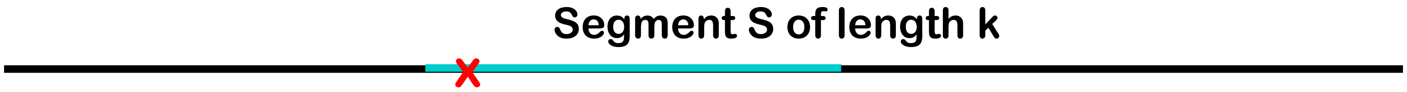
C: Two sixes = A $\cap$ B

P(A) = 6/36 = 1/6

P(B) = 6/36 = 1/6

P(A $\cap$ B) = 1/36 = P(A) P(B)

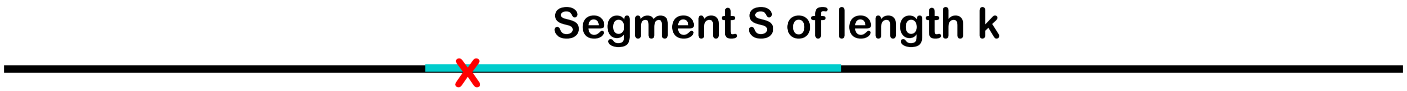
From two dice to millions of reads



G: Length of genome
L: Length of reads
N: Number of reads

What is the probability that no read begins in S?

From two dice to millions of reads



G: Length of genome
L: Length of reads
N: Number of reads

What is the probability that no read begins in S?

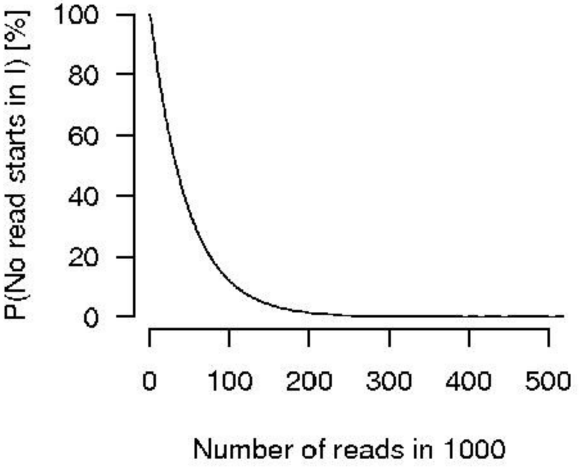
P(first read starts in S) = k/G

P(first read does not start in S) = 1-k/G

P(no read starts in S) = $(1 – k/G)^N$

If no read starts in a segment of the size of the read length it is impossible to extend contigs across this gap

k=L: P(no read starts in S) = (1 – L/G)^N



H.acinonychis:
***G =* 1.5 Mbps**
***L =* 32**

A Bernoulli variable X indicates whether an event occurred or not

X = 1 if A occurred

X = 0 else

P(X = 1) = P(A) := p

Notation:

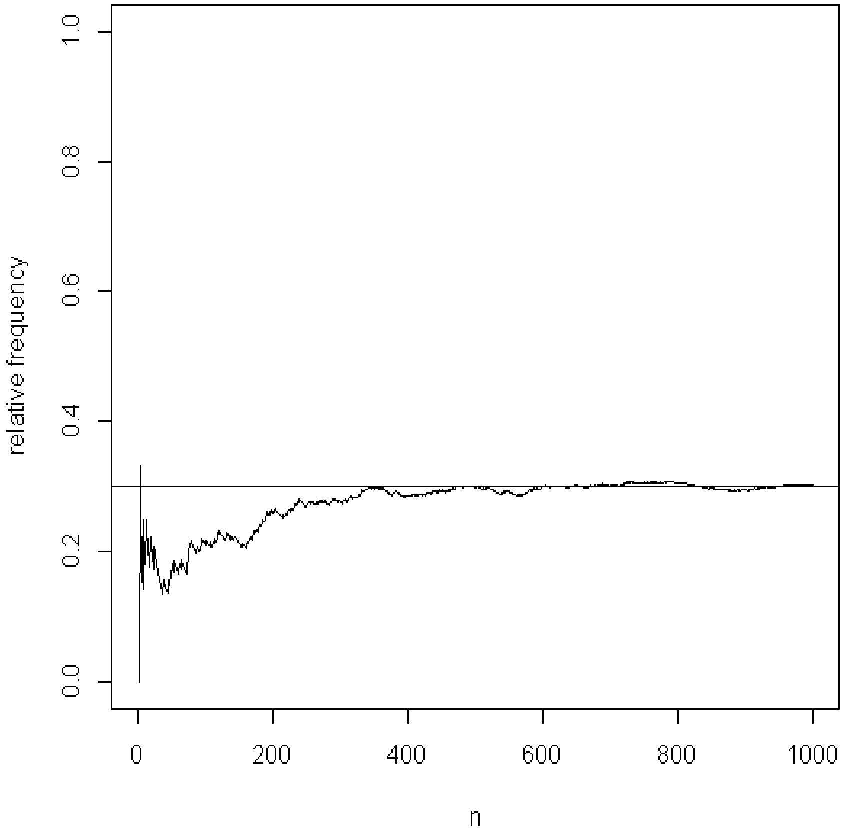
X ~ Bernoulli(p)

Note: To simulate X we only need to know p and not A

Sequences of independent Bernoulli variables can be easily simulated

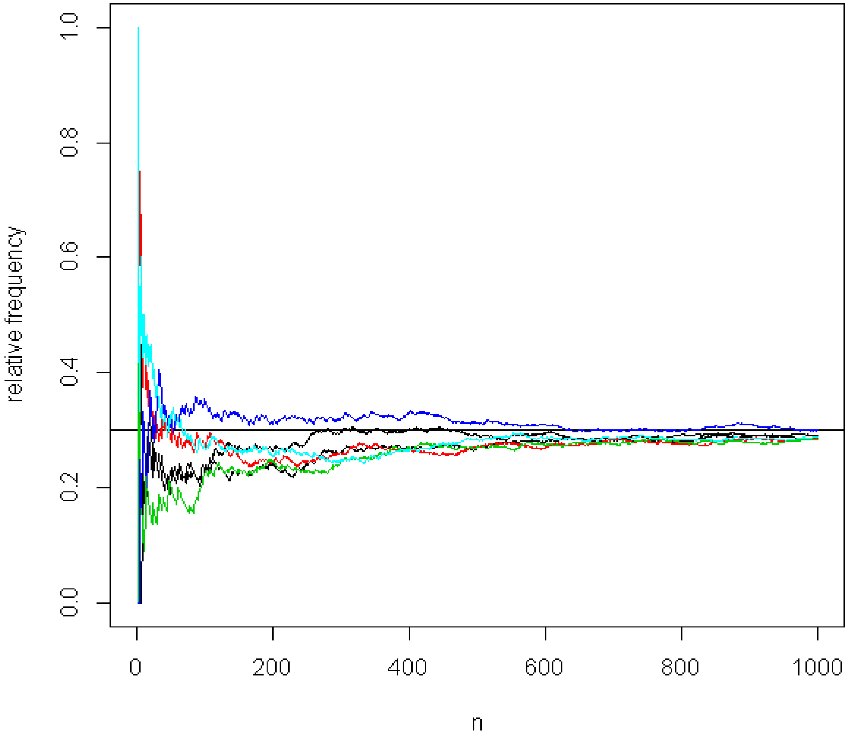
```
> p=0.3       # p=P(X=1)
> N=10        # Number of draws
> sample(c(1,0),size=N, prob=c(p,1-p),replace=TRUE)
 [1] 0 0 0 1 1 0 1 1 0 1

> p=0.3       # p=P(X=1)
> N=10        # Number of draws
> sample(c(1,0),size=N, prob=c(p,1-p),replace=TRUE)
 [1] 0 1 0 0 0 0 0 0 0 0
>
```



The relative frequencies of 1s converge to p, if the number of independent draws becomes large

```
> p = 0.3          # p=P(X=1)
> N = 1000         # Number of draws
>
> draws = sample(c(1,0),size=N, prob=c(p,1-p),replace=TRUE)
> relfreq = cumsum(draws)/1:N
> plot(1:N,relfreq,type="l",xlab="n",ylab="relative frequency",ylim=c(0,1))
> abline(p,0)
```

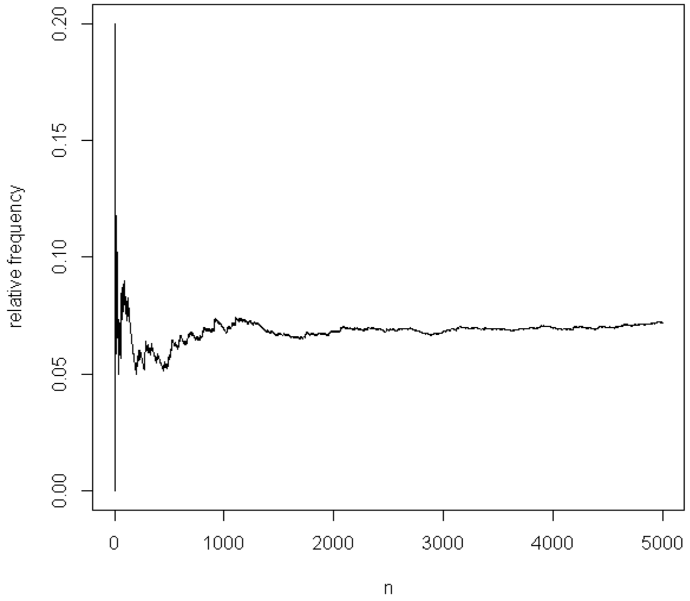


The trajectories of relative frequencies differ early on, but the limit is always the same

```
...
>
> # more runs
> for(i in 1:5){
+    draws =sample(c(1,0),size=N, prob=c(p,1-p),replace=TRUE)
+    relfreq=cumsum(draws)/1:N
+    points(1:N,relfreq,type="l",xlab="n",ylab="relative frequency",ylim=c(0,1),col=i)
+ }
```

We can use this convergence to calculate probabilities via simulation

Given a genome of length 1000 and reads of length 36, what is the probability that two reads overlap?



```
> N=5000, G=1000, L=36
> d1= sample(1:G-L+1,N, replace=TRUE) # random start points
> d2= sample(1:G-L+1,N, replace=TRUE) # random start points
> x=as.numeric(abs(d1-d2)<L)          # x=1 if the reads overlap
> relfreq=cumsum(x)/1:N
> relfreq[N]
[1] 0.072
```

The distribution of a Bernoulli variable is summarized in the following table

Outcome	Probability
0	p
1	1-p

We can generalize this concept to variables that have more then 2 possible real valued outcomes

Outcome	Probability
0.3	0.2
0.7	0.2
1.2	0.3
1.8	0.1
2.0	0.2

Outcome	Probability
x1	p1
x2	p2
x3	p3
...	...
x_n	p_n

p1 + p2 + …+ p_n = 1

Note that the outcomes do not have equal probabilities any more. We have left the Laplace Model

Discrete random variables can be easily simulated

Outcome	Probability
0.3	0.2
0.7	0.2
1.2	0.3
1.8	0.1
2.0	0.2

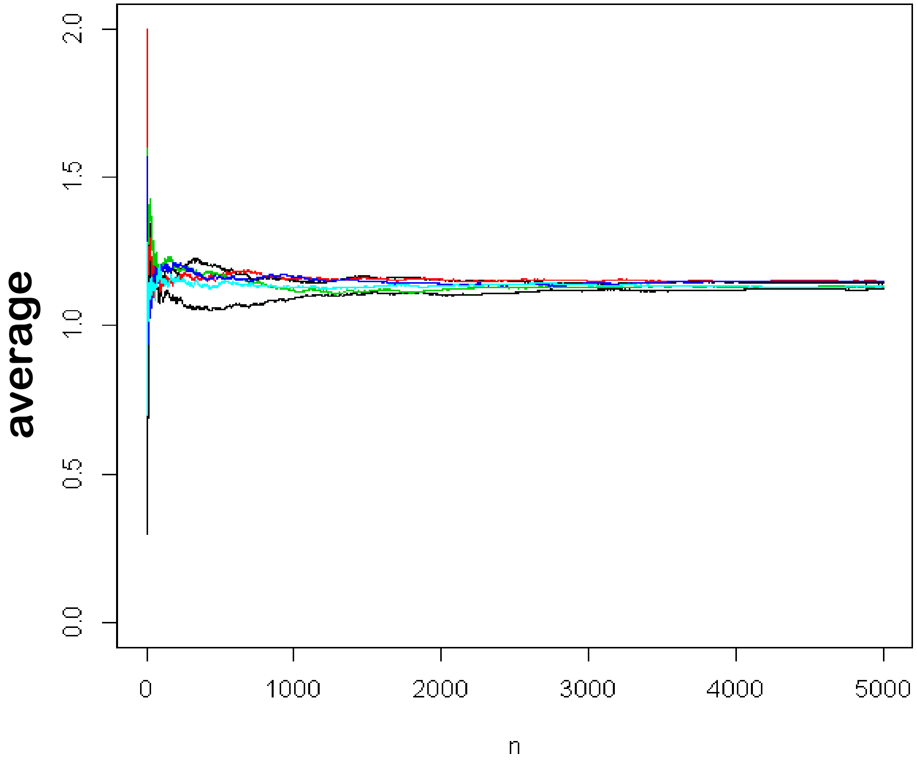
```
> N=10
> omega = c(0.3,0.7,1.2,1.8,2.0)
> p     = c(0.2,0.2,0.3,0.1,0.2)
> sample(omega, size=N, prob=p, replace=TRUE)
 [1] 0.7 1.2 0.7 1.2 1.2 0.7 1.8 0.7 1.2 2.0

> sample(omega, size=N, prob=p, replace=TRUE)
 [1] 1.2 1.2 0.3 0.3 2.0 1.8 1.2 1.2 2.0 0.3

> sample(omega, size=N, prob=p, replace=TRUE)
 [1] 1.2 1.8 0.7 0.3 2.0 1.2 0.3 0.3 0.3 0.7
>
```

The averages of simulated data converge

Outcome	Probability
0.3	0.2
0.7	0.2
1.2	0.3
1.8	0.1
2.0	0.2



What do they converge to?

The expected value E(X) of a discrete random variable X is a weighted average of the outcomes

Outcome	Probability
x1	p1
x2	p2
x3	p3
...	...
x_n	p_n

$$E(X) = \sum_i p_i \, x_i$$

$$p_i = P(X = x_i)$$

The expected value is a linear operator

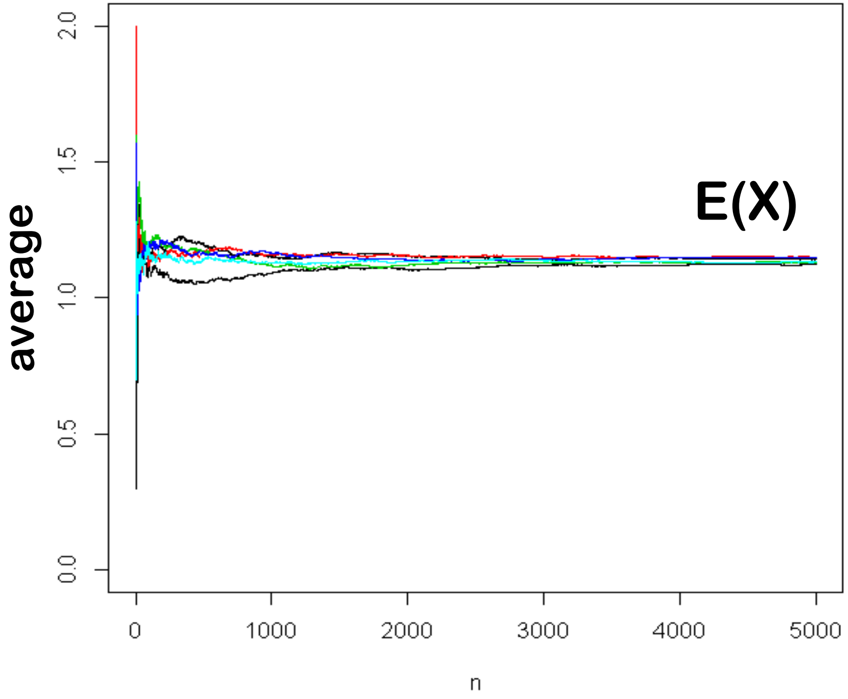
$$E(\alpha\, X + \beta\, Y) = \alpha\, E(X) + \beta\, E(Y)$$

This follows immediately from the definition and holds true for arbitrary random variables X and Y

The Law of Large Numbers:

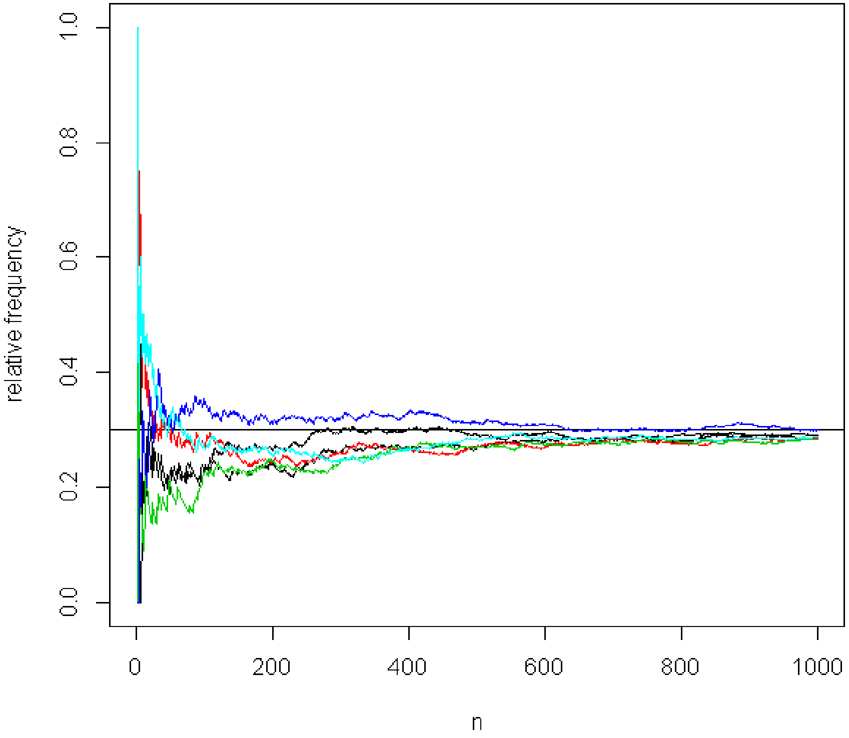
The averages of simulated random variables converge towards the expected value

Outcome	Probability
0.3	0.2
0.7	0.2
1.2	0.3
1.8	0.1
2.0	0.2



E(X) = 0.3*0.2 + 0.7*0.2 + 1.2*0.3 + 1.8*0.1 + 2.0 *0.2 = 1.14

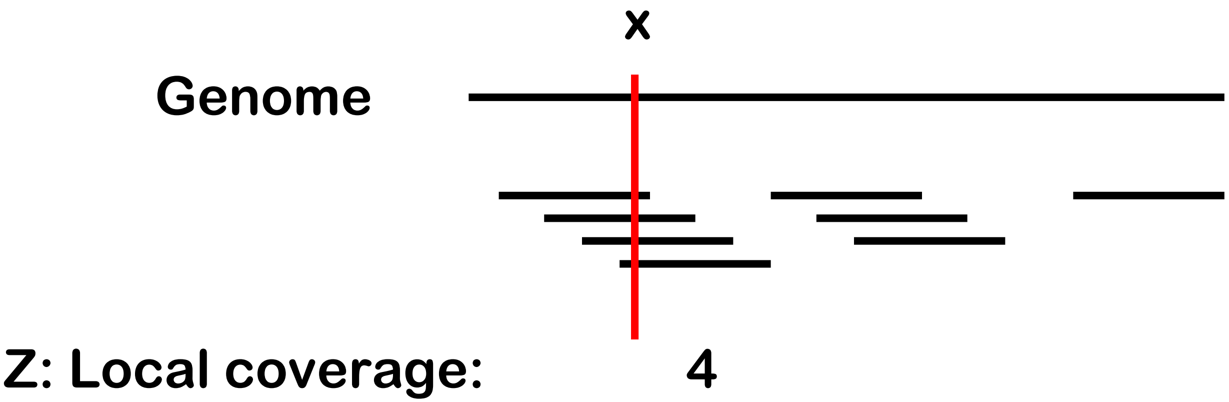
The expected value of X~Bernoulli(p) is p



For 0-1-data the relative frequency of 1s is the average of the data

E(X) = 1*p + 0*(1-p)

We can model the local coverage of a fixed genome position x by the sum of independent Bernoulli variables



x

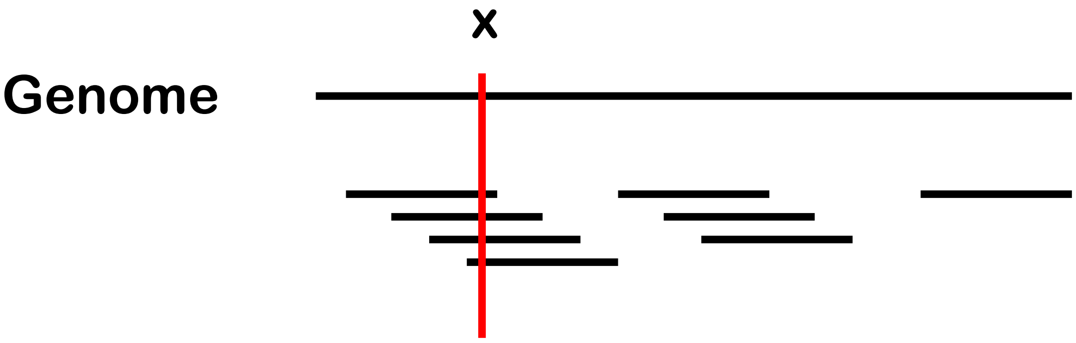
Genome

Z: Local coverage: 4

X_i = 1 if read i covers position x
X_i = 0 else

$$Z = \sum_i X_i$$

We can model the local coverage of a fixed genome position x by the sum of independent Bernoulli variables



x

Genome

Z := Local coverage: 4

X_i = 1 if and only if the read starts in the interval [x-L+1, x]

X_i ~ Bernoulli(p)

p := P(X_i=1) ≈ L/G

P(Z = k) can be computed via combinatorics

A: The first k reads cover x $P(A) = p^k$

B: Only the first k reads cover x $P(B) = p^k\,(1-p)^{n-k}$

C: The local coverage of x = k $P(C) = \binom{n}{k}\,p^k\,(1-p)^{n-k}$

The sum of n Bernoulli(p) variables Z is a Binomial(n,p) variable

Outcome	Probability
k=0	p0
k=1	p1
k=2	p2
...	...
k=n	p_n

$$p_k = \binom{n}{k} p^k (1-p)^{n-k}$$

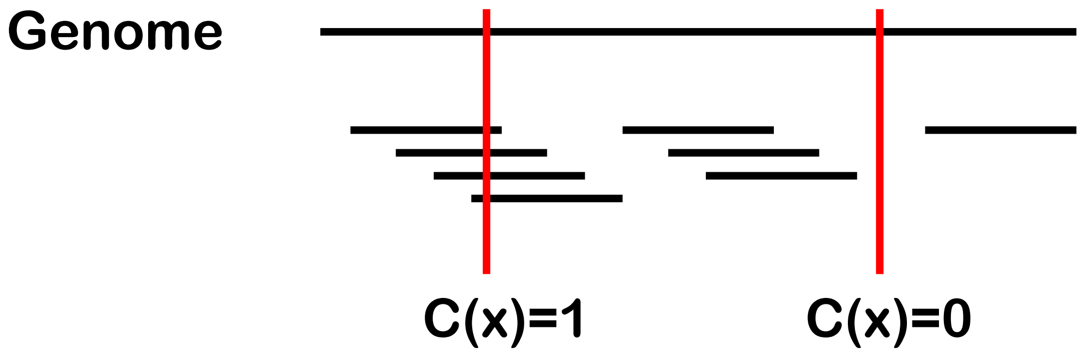
Z is a standard discrete random variables with n+1 possible outcomes. The corresponding probabilities sum to 1.

The expected value of a Binomial(n,p) variable is np

$$E(Z) = E(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} E(X_i) = n\,p$$

… by linearity of the operator E( )

We can model sequenced positions in a genome as dependent Bernoulli variables



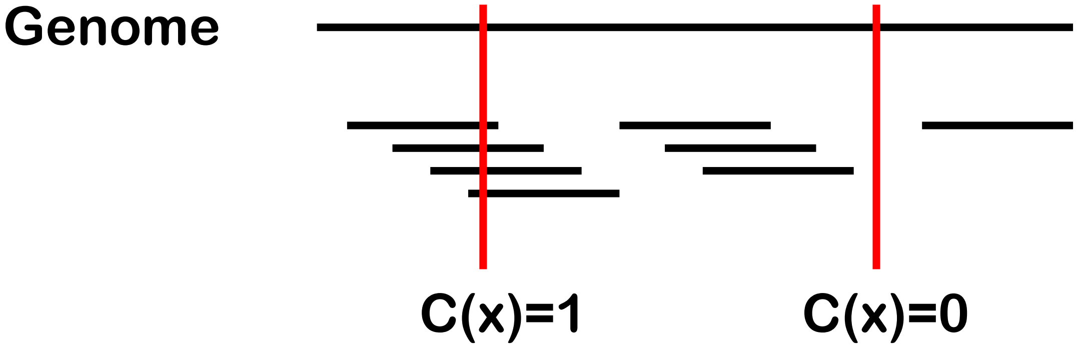
C(x)=1 if and only if some read starts in the interval [x-L+1, x]

G: Length of genome
L: Length of reads
N: Number of reads

$P(C(x)=1) = 1 - (1 - L/G)^N$

The C(x) are Bernoulli $(1 - (1 – L/G)^N)$

The sum of the C(x) gives us the total number of covered genome positions

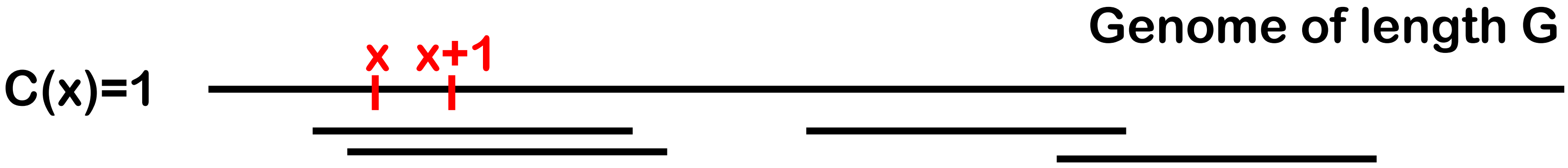


$$Z = \sum_{x=1}^{G} C(x)$$

The C(x) are Bernoulli (1 - (1 – L/G)^N)

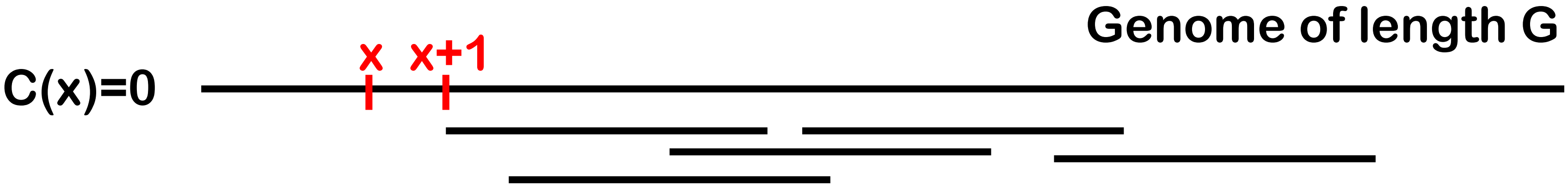
The C(k) are not independent

C(x)=1 if and only if some read starts in the interval [x-L+1, x]



Genome of length G

C(x)=1

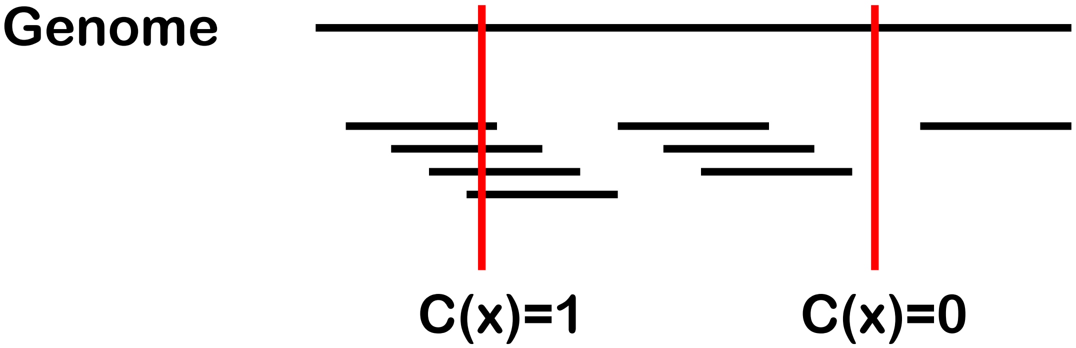


Genome of length G

C(x)=0

If C(x)=0, x+1 can only be covered by a read that starts at x+1, since the other start points (x+1)-1, (x+1)-2, …(x+1)-L+1 would also cover x

Linearity of E( ) gives us the expected amount of covered genome



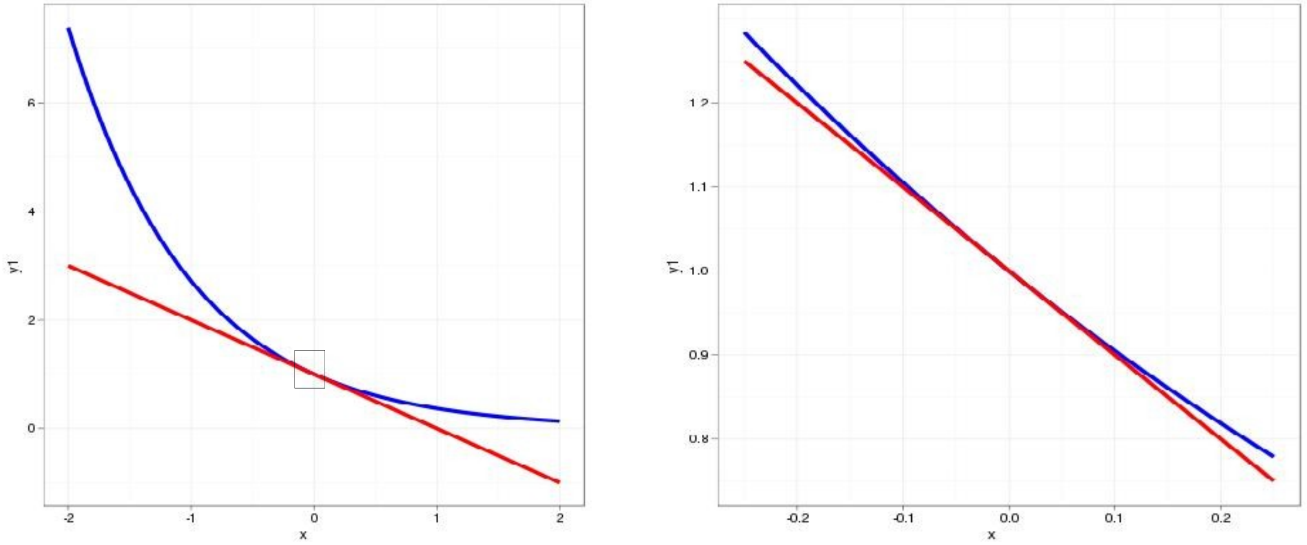
The C(x) are Bernoulli (1 - (1 – L/G)^N)

X: Number of sequenced positions in the genome

$Z = \sum_{x=1}^{G} C(x)$ $E(Z) = \sum_{x=1}^{G} E(C(x))$

Expected amount of covered genome: *E(Z)= G (1 - (1 – L/G)^N)*

f(x) = 1-x is the tangent to g(x)= exp(-x) at x=0



1/G is very small since genomes are long

$$(1 - 1/G) \approx e^{-1/G}$$

The Lander-Waterman-Formula calculates the expected amount of sequenced genome from the global coverage and the length of the genome

Expected amount of sequenced genome: $E(Z) = G\,(1 - (1 - L/G)^{N})$

$(1-L/G) \approx \exp(-L/G)$

$G\,(1-(1-L/G)^{N}) \approx G\,(1-\exp(-a))$

$a = \frac{NL}{G}$ **a = global coverage**

Lander-Waterman-Formula

Sequenced genome = length(genome) x (1-exp(-coverage))

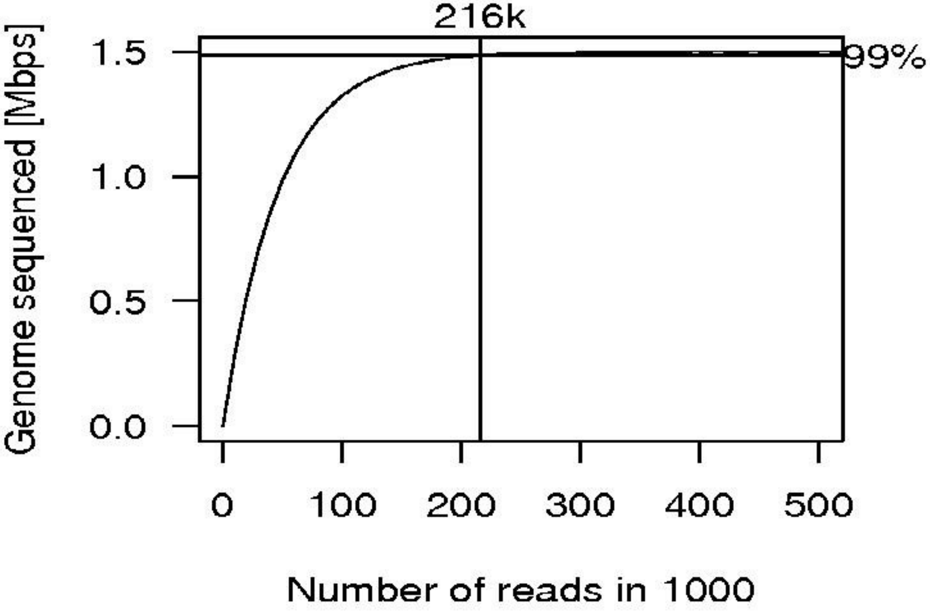
How many random reads do we need to cover 99% of the helicobacter acinonychis genome in average?

X: Number of sequenced positions in the genome

Lander-Waterman: E(Z) = G (1 – exp (-a))

G: about 1.5 mbps

216K reads of length 32 are needed for 99% genome coverage of a 1.5mbps genome



In average we will get 99% coverage, but we can still be unlucky and get less than average

Lander-Waterman: E(Z) = G (1 – exp (-a))

We want to use enough reads such that we can almost guarantee that we will have 99% coverage or more

How do we calculate: P(99% coverage)

We can simulate the probability that a random fragmentation will cover 99% of the genome

1. Randomly select start points for N reads
2. Calculate percentage of covered genome
3. Repeat this simulation F times

C(f) = 1 if fragmentation f covered more then 99% of the genome
C(f) = 0 else

The average of C(f) converges towards P(99% coverage)

This simulation can be used to find the number of reads that give us a high coverage almost surely

The average of C(f) converges towards P(99% coverage)

We can tune the number of reads N such that
P(99% coverage) > 0.95

For this number of reads we have a very good chance to end up with 99% coverage or more

End of Chapter 6