# CS CM124/224 & HG CM124/224

# DISCUSSION SECTION (MAY2, 2013)

TA: Farhad Hormozdiari

# Reminder

- Update your your wikidot
- HW3 is due on May 13

# Agenda

- Re-sequencing
- Sequence Mapping Coverage
- Tumor Genome Reconstruction
- HW3

# Re-sequencing

- I want to sequence my genome (know my DNA sequence). How?
- Several sequencing technologies
- One is called next generation sequencing
  - Cheaper than other sequencing technologies
  - Generate many short reads from my genome
- A short read is a short DNA segment from my genome of length 30bp ~ ?
- Re-sequencing is mapping these short reads to known DNA sequence (called reference genome)
  - Assume that my genome is very close to reference genome
  - Require short reads from my genome and reference genome (constructed by other sequencing technologies)
- Why don't we just use sequencing technologies that were used to construct the reference genome?
  - Because they are more expensive and takes more time

# Problems with Re-sequencing

- Repeated sequences in reference genome – reads from target map to multiple positions
- Insertion, deletion, or inversion in target genome (or any target sequence that is significantly different from reference genome) – reads do not map to any position
- If reads have random errors
  - Solution – Collect many reads that map to the same position
  - Coverage – the number of times each position is mapped by different reads. For example, 10x coverage means there are on average 10 reads mapping to the same position
  - We then take consensus among reads that map to the same position
  - Since only few reads have error at that position, this solves error of reads. And, more reads we have at the position (higher coverage), less likely incorrect prediction is made

# Consensus Algorithm for SNP calling

- Take majority vote.

**My Genome:**

TACATGAGATC**G**ACATGAGATC**G**GTAGAGC**C**GTGAGATC

**Sequence Reads:**

TCGACATGAGATCGGTAGA**A**CCGT
GACA**A**GAGATCGGTAGAGCCGTGA
TGAGATCGG**T**AGAGCCGTGAGATC

**The Human Genome:**

```
TACATGAGATCCACATGAGATCTGTAGAGCTGTGAGATC
         TCGACATGAGATCGGTAGAACCGT
           GACAAGAGATCGGTAGAGCCGTGA
               TGAGATCGGTAGAGCCGTGAGATC
```

**Recovered Sequence:**

TACATGAGATC**G**ACATGAGATC**G**GTAGAAC**C**GTGAGATC

# Problems with Re-sequencing

- Example – If error rate is e, and we are going to predict the consensus sequence, what is the error rate if the coverage is 3.

# Problems with Re-sequencing

- Example – If error rate is e, and we are going to predict the consensus sequence, what is the error rate if the coverage is 3.
  - We will make a prediction with an error if two out of three reads or three out of all three reads have an error in the same place.

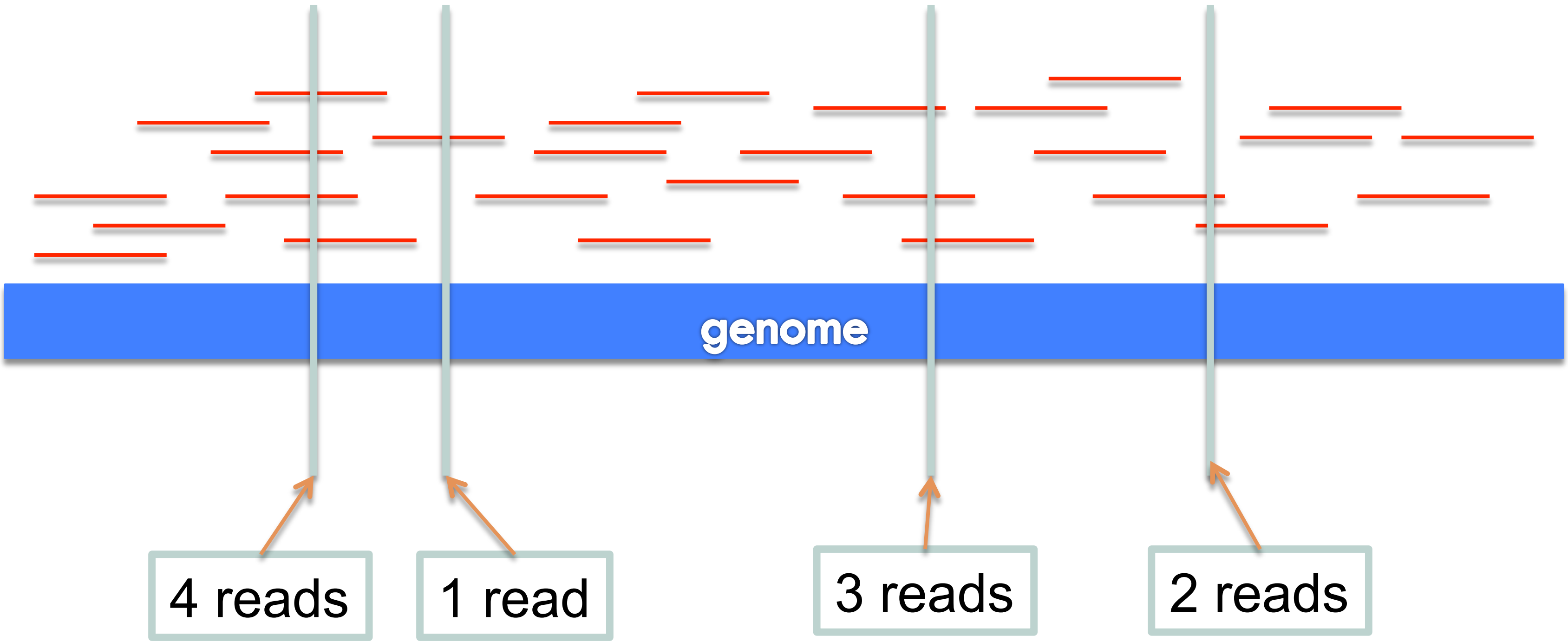
$$\text{Probability of 3 reads having error} \rightarrow \boxed{e^3} + \boxed{\binom{3}{2}(1-e)e^2} \leftarrow \text{Probability of 2 out of 3 reads having error}$$

# Sequence Mapping Coverage

- If a genome is length N (human is 3,000,000,000), and the total length of all sequence reads collected is M, the coverage (ratio) is defined as M/N
- Often written with an "x". For example, 10x or 20x coverage
- 10x coverage means there are 10 reads **on average** mapping to the same position; can be less than 10, more than 10, or exactly 10 depending on the position of genome

# Coverage Example

- Assume we have 3x overall coverage



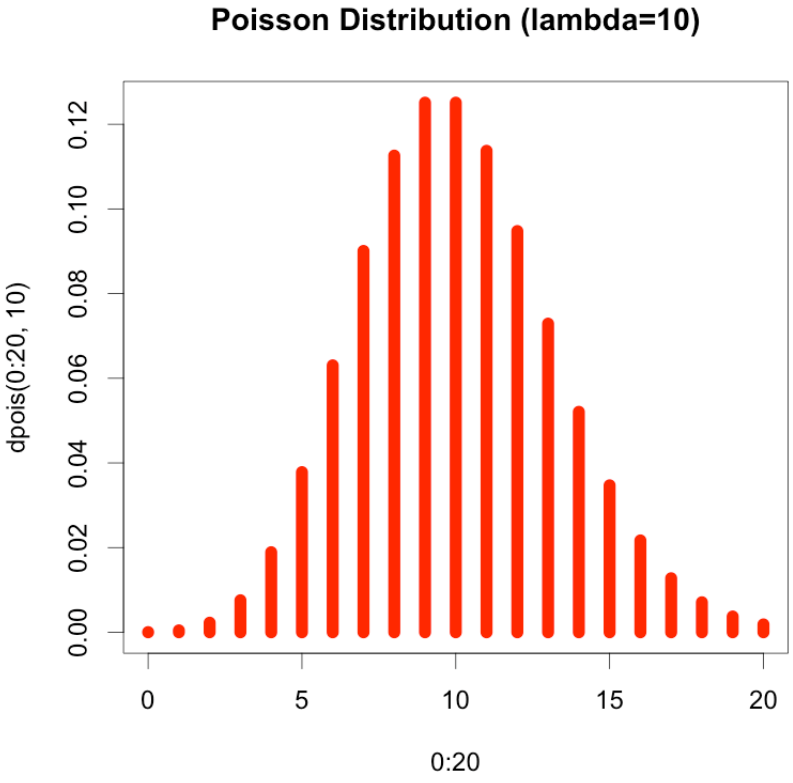
- We assume that coverage (# of reads at a specific position of genome) follows the Poisson distribution whose mean is the overall coverage (e.g. 3x)

# Poisson Distribution

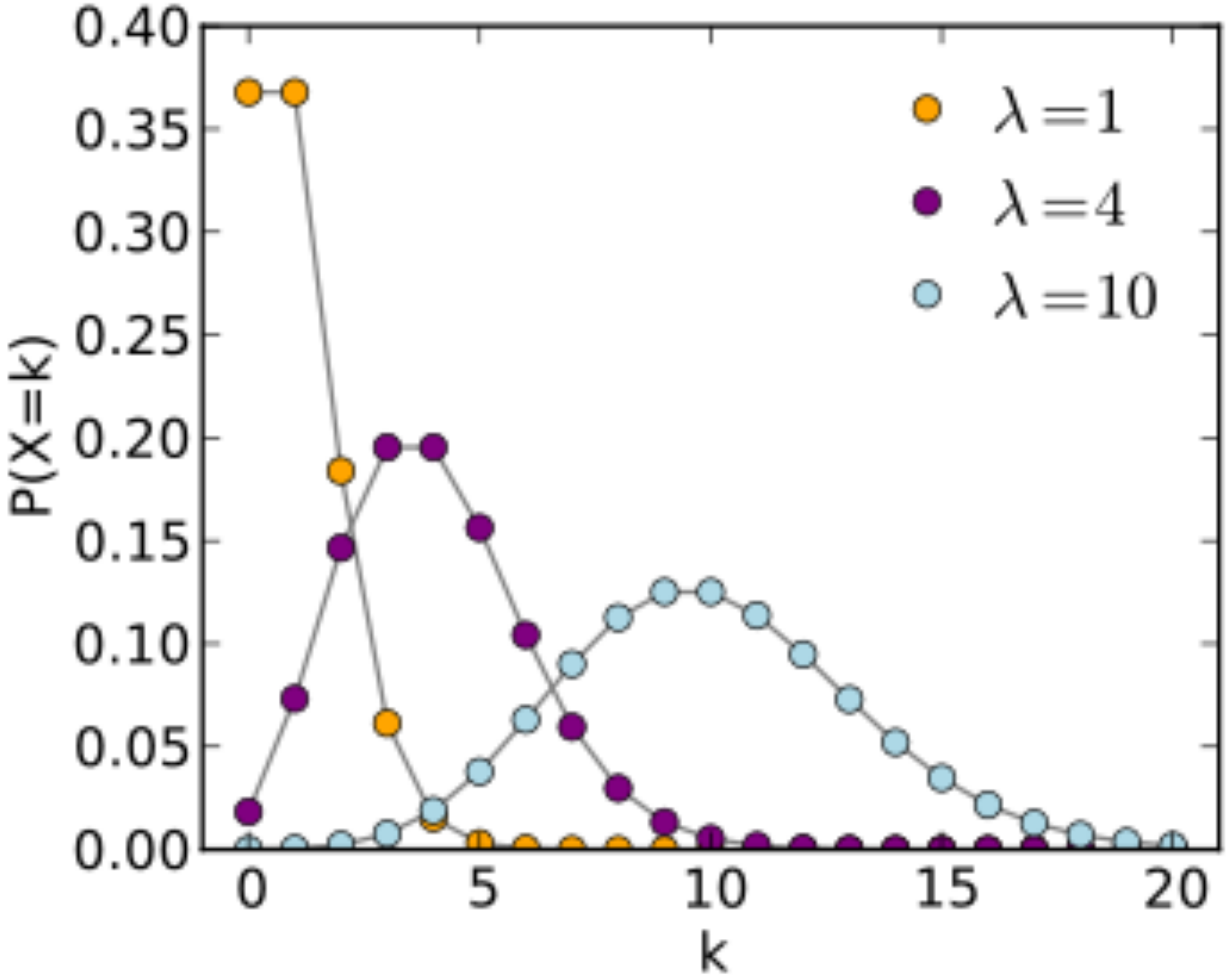
- Discrete probability distribution to compute probability of (rare) events given known mean
- Only one parameter: $\lambda$, mean of distribution
- Probability Mass Function

$$\Pr(N_t = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

- Mean = $\lambda$
- Variance = $\lambda$



# Poisson Distribution



# Poisson Distribution to Sequence Coverage

- Overall coverage = $\lambda$
- Probability that exactly X reads span a certain position (percentages of genome that have coverage equal to X)
  - **dpois(X, $\lambda$)**
- Probability that X or fewer reads span a certain position (percentages of genome that have coverage equal to or less than X)
  - **ppois(X, $\lambda$)**
- At least Y% of the genome have at least $\lambda$ coverage
  - **qpois(Y, $\lambda$)**

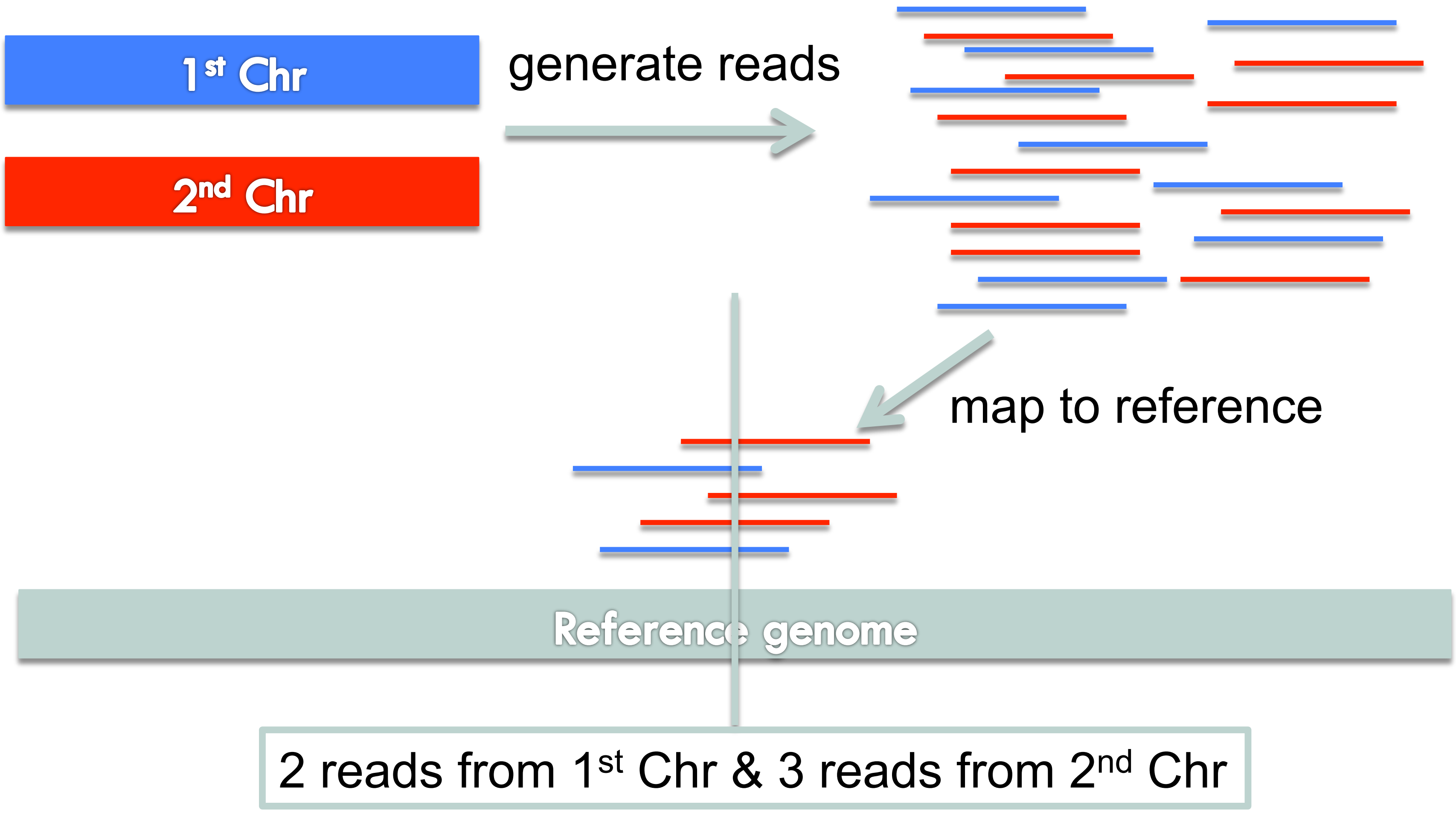
# Coverage examples

- For human genome (L=3,000,000,000) sequenced at 30x coverage, what is the probability that a specific location has exactly 30 coverage?
- **λ=30 dpois(30,λ)=dpois(30,30)=0.072**
- What is the probability that a specific location has at least 30 coverage?
- **1-ppois(29,λ)=1-ppois(29,30)=0.524**
- What is the probability that a specific location has at least 10 coverage?
- **1-ppois(9,30)=0.9999929**

# Diploid Coverage

□ Humans have 2 chromosomes

□ Each read comes from one chromosome at random



1st Chr

2nd Chr

generate reads

map to reference

Reference genome

2 reads from 1st Chr & 3 reads from 2nd Chr

# Diploid Coverage

- Assume a position in the reference genome is covered by Y reads
- The probability that X of those Y reads come from the first chromosome follows the binomial distribution with .5 probability
  - **dbinom(X, Y, 0.5)**
  - Same as the probability of observing X heads when we toss the fair coin Y times

# Diploid Coverage

- Given that we have Y reads mapped to **a specific position** of reference genome, what is the probability of having at least X reads (or coverage) for each chromosome?
- Let's assume Y = 10, X = 3
- We want to add the following probabilities
  - The probability of having 3 reads from 1st Chr and 7 reads from 2nd Chr
  - The probability of having 4 reads from 1st Chr and 6 reads from 2nd Chr
  - The probability of having 5 reads from 1st Chr and 5 reads from 2nd Chr
  - The probability of having 6 reads from 1st Chr and 4 reads from 2nd Chr
  - The probability of having 7 reads from 1st Chr and 3 reads from 2nd Chr
- dbinom(3,10,0.5)+dbinom(4,10,0.5)+dbinom(5,10,0.5)+dbinom(6,10,0.5)+dbinom(7,10,0.5)
- Or

$$\sum_{i=X}^{Y-X} \mathrm{dbinom}(i, Y, 0.5)$$

# Diploid Coverage Examples

- If a position is covered by 10 reads, what is the probability that exactly 3 reads come from the first chromosome?
- **dbinom(3,10,.5)=.117**
- If a position is covered by 10 reads, what is the probability that at least 4 reads come from the first chromosome?
- **1-pbinom(3,10,.5)=.828**
- If a position is covered by 10 reads, what is the probability that at least 4 reads come from each chromosome?
- **dbinom(4,10,.5)+dbinom(5,10,.5)+dbinom(6,10,.5)=.656**

# Another Diploid Coverage

- We assume that the **overall coverage** is $\lambda$
- What is the probability of having at least X coverage for each chromosome over the whole genome?
- First, we want to compute the probability of having *i* coverage at **a specific position** of genome given the overall coverage $\lambda$
  - **dpois(i, $\lambda$ )**
- Given we have *i* coverage at a specific position, what is the probability of having at least X coverage for each chromosome?

$$\sum_{j=X}^{i-X} \mathrm{dbinom}(j,i,0.5)$$

- Then, given the overall coverage (λ), what is the probability of having *i* reads (or coverage) at a specific position and having at least X coverage for each chromosome?

$$\mathrm{dpois}(i,\lambda)\sum_{j=X}^{i-X} \mathrm{dbinom}(j,i,0.5)$$

# Another Diploid Coverage

- We only computed the probability when there are i reads (or coverage) at a specific position
- The minimum value of i is 2X (2 times X)
  - We want at least X coverage for each chromosome, so we need to have at least 2X coverage at a specific position
  - For example, if we want at least 5 coverage for each chromosome, we need to have at least 10 reads mapped to the reference genome
- The maximum value of i is infinitity
- Hence, i increases from 2X to infinity, and we sum the probabilities for each i value

$$\sum_{i=2X}^{\infty}\left(\mathrm{dpois}(i,\lambda)\sum_{j=X}^{i-X}\mathrm{dbinom}(j,i,0.5)\right)$$

- Note that this is a nested loop (double loop). Not just multiplying two separate loops

# Diploid Coverage Examples

- If genome is covered with coverage 30, what is the probability that a position will have at least 10 reads from each chromosome?

$$\sum_{i=20}^{\infty} \mathrm{dpois}(i,30) \sum_{j=10}^{i-10} \mathrm{dbinom}(j,i,0.5)$$

# Tumor Genome Reconstruction

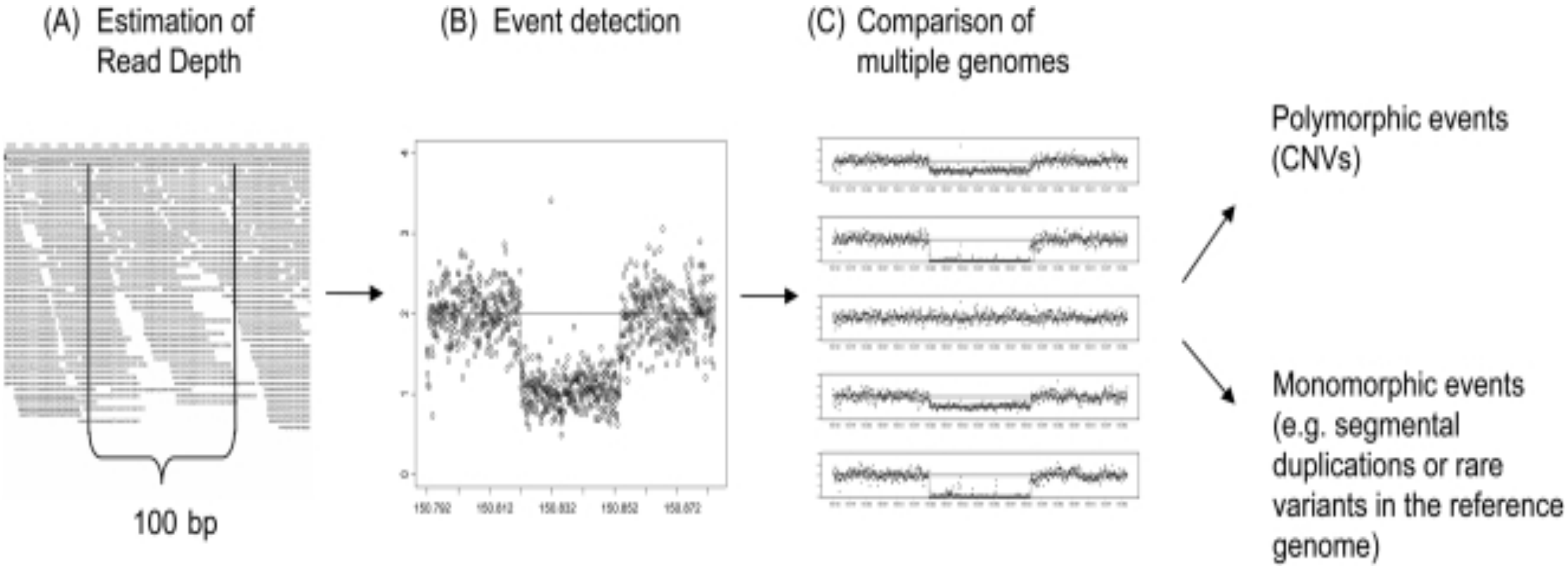
- We have the reference genome (known)
- We have the tumor genome (unknown)
- We have single-end reads from tumor genome and map them to the (known) reference genome
- By observing how those paired-end reads map to the reference genome, we can reconstruct the tumor genome
- Parts of tumor genome can be the same as the reference genome, or some regions may be inverted, deletion.

# Copy Number Variation (CNV)

- We have the reference genome (known)
- We have the donor genome (unknown)
- We have single-end reads from donor genome and map them to the (known) reference genome
- By observing how those paired-end reads map to the reference genome, we can reconstruct the donor genome
- Parts of tumor genome can be the same as the reference genome, or some regions may be copied, deletion.

# Read depth methods

- Map the single end reads and look at the coverage in each positions.

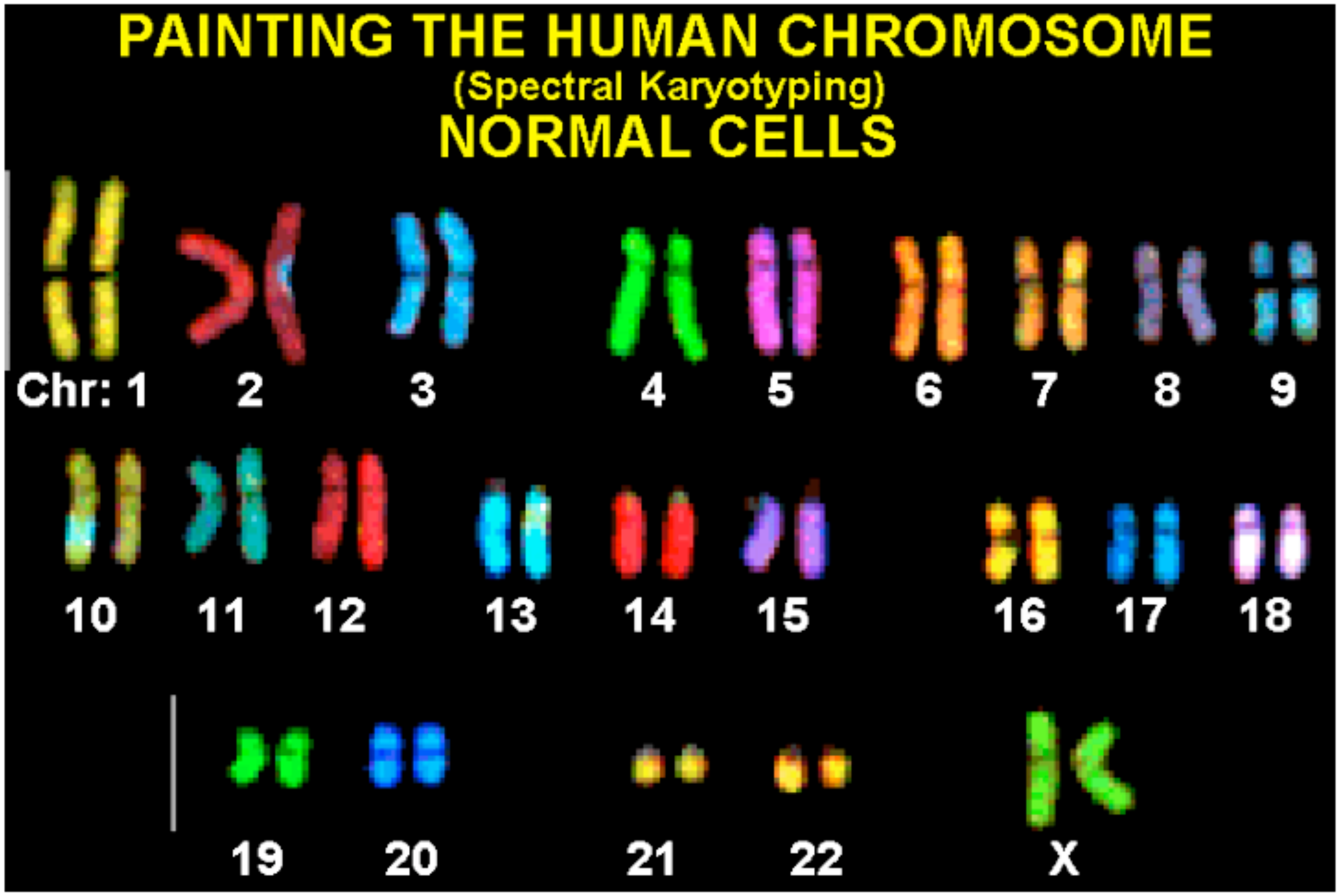

(A) Estimation of Read Depth
100 bp
(B) Event detection
(C) Comparison of multiple genomes
Polymorphic events (CNVs)
Monomorphic events (e.g. segmental duplications or rare variants in the reference genome)

Yoon et. al 2009

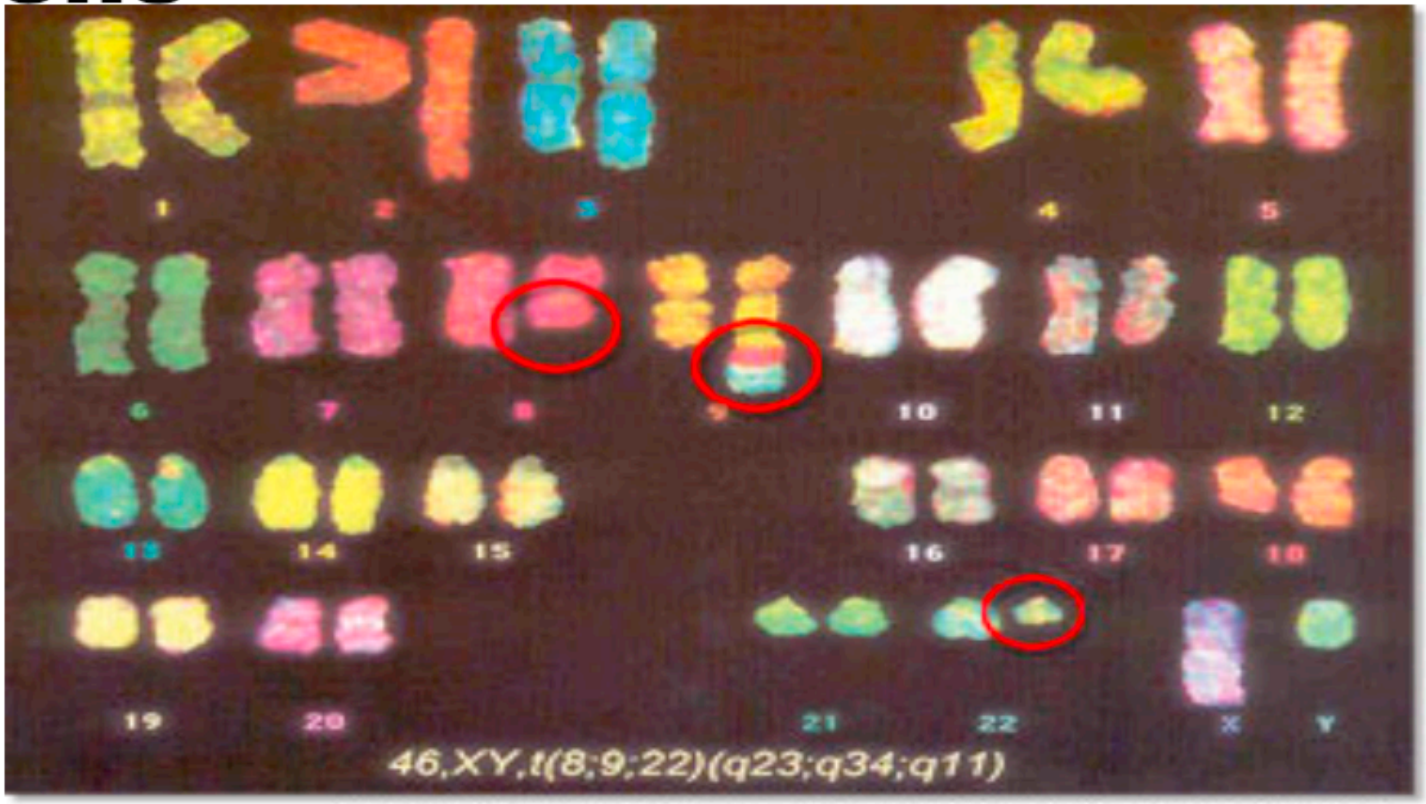
# Tumor Genome Reconstruction

- We have the reference genome (known)
- We have the tumor genome (unknown)
- We have paired-end reads from tumor genome and map them to the (known) reference genome
- By observing how those paired-end reads map to the reference genome, we can reconstruct the tumor genome
- Parts of tumor genome can be the same as the reference genome, or some regions may be inverted, duplicated, or translocated

# Chromosome Painting: Normal Cells



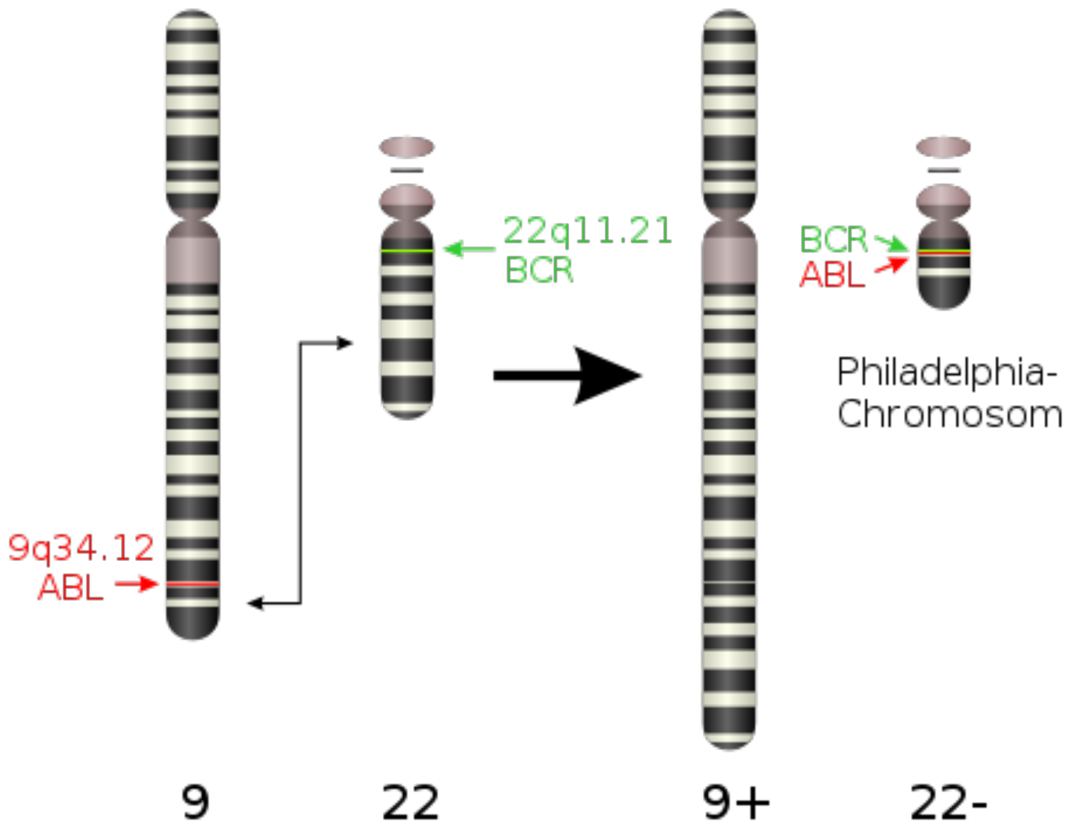
# Chromosome Painting: Tumor Cells



Note: This karyotype was prepared using a FISH technique known as "chromosome painting". As well as having a translocation from chromosome 22, chromosome 9 also has translocated material from chromosome 8.

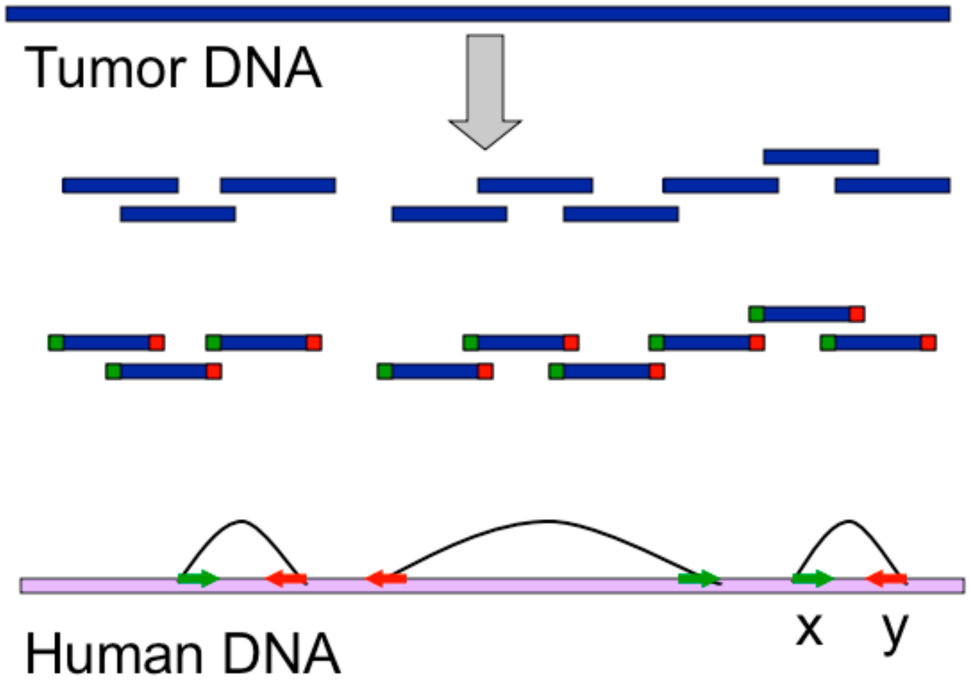
# Why do we care about rearrangement?

chronic myelogenous leukemia



# Paired End Sequencing (PE)

## C. Collins et al. (UCSF Cancer Center)



1) Pieces of tumor genome: clones (100-250kb).

2) Sequence ends of clones (500bp).

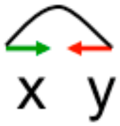
3) Map end sequences to human genome.

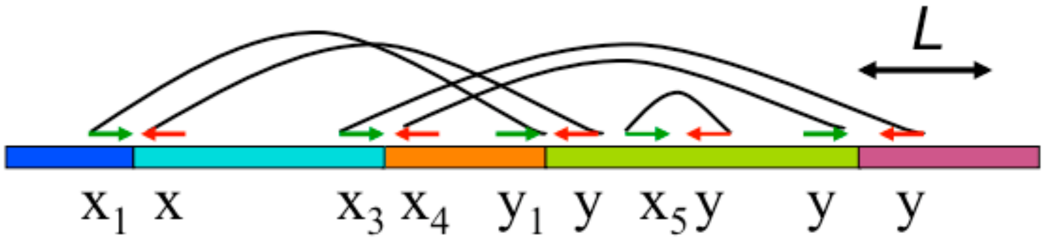
Each clone corresponds to pair of end sequences (***PE pair***) *(x,y).*

Typical Next Generation Sequencing read lengths are shorter.

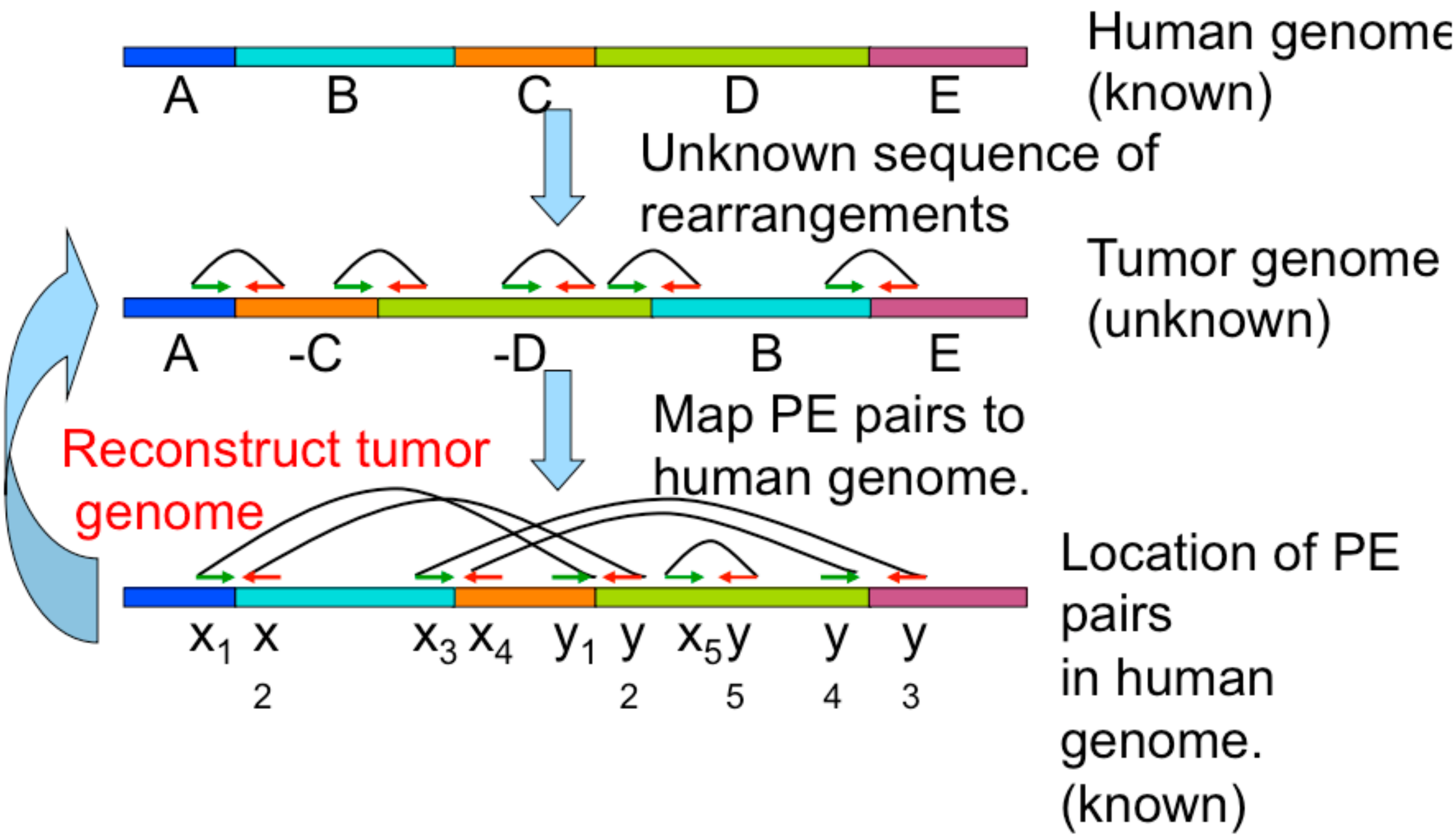
# PE Pairs

- Order PE pair such that $x < y$.
- PE pair $(x,y)$ is
  - *valid* if
    - $x,y$ on same chromosome. and
    - $l \leq y - x \leq L$, min (max) size of clone.
    - $x$, $y$ have opposite, convergent orientations
  - *invalid*, otherwise.
    - Results from rearrangement or experimental "noise".





# Tumor Genome Reconstruction Puzzle



Human genome (known)

Unknown sequence of rearrangements

Tumor genome (unknown)

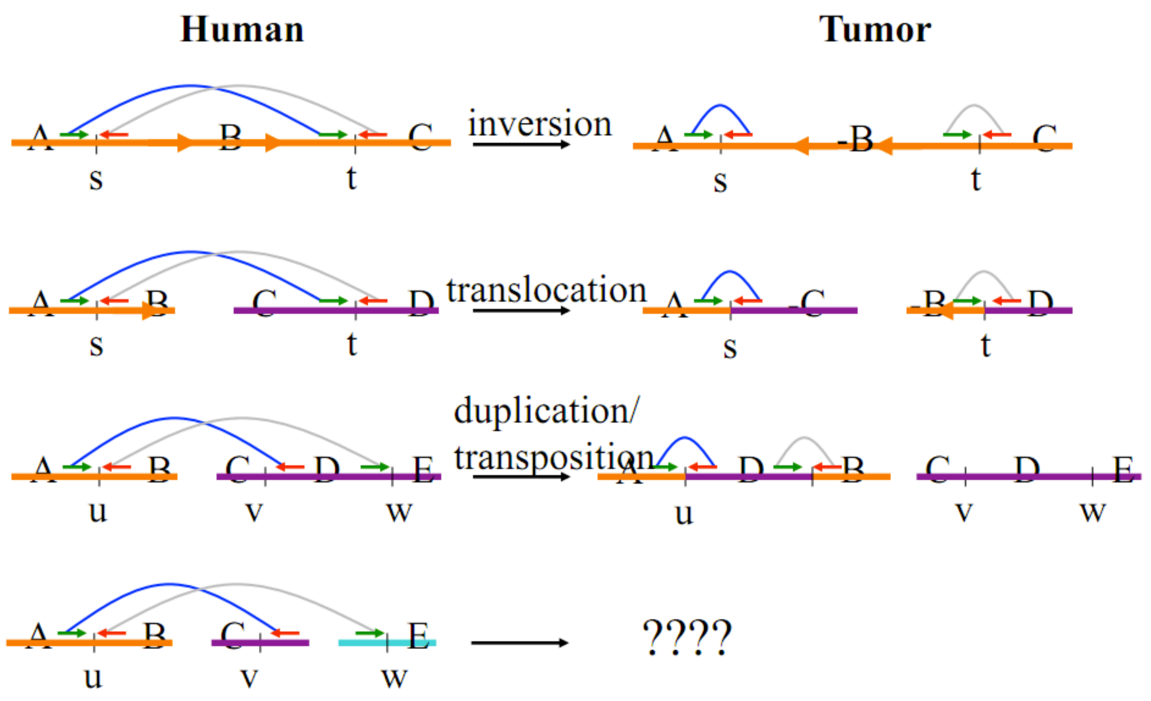
Map PE pairs to human genome.

Reconstruct tumor genome

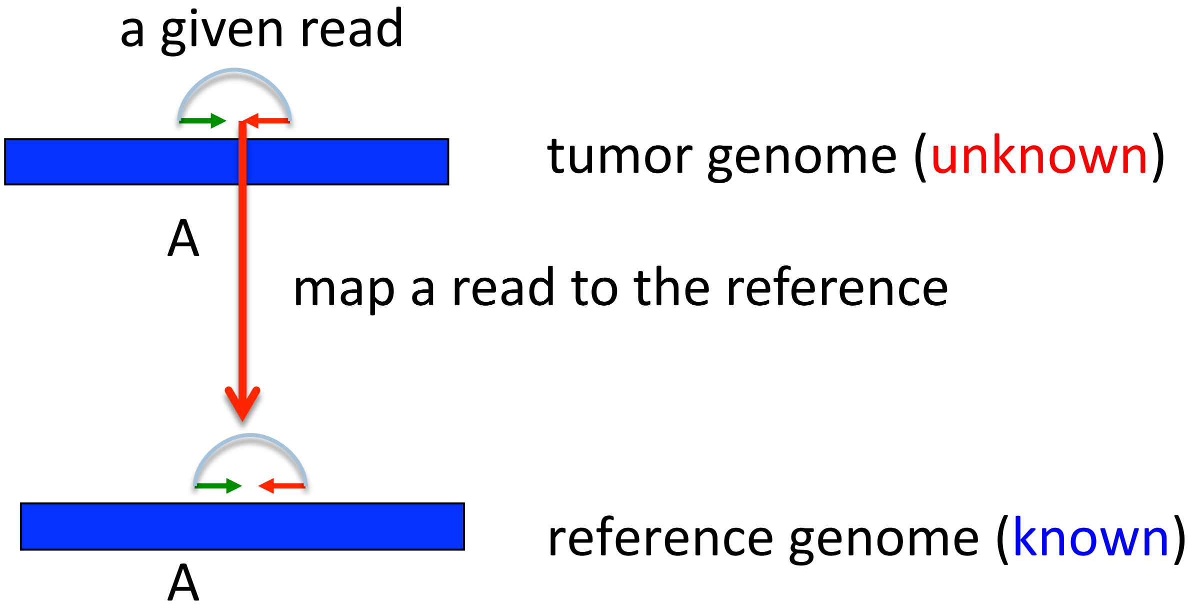
Location of PE pairs in human genome. (known)

# Possible rearrangements

## Rearrangement Signatures



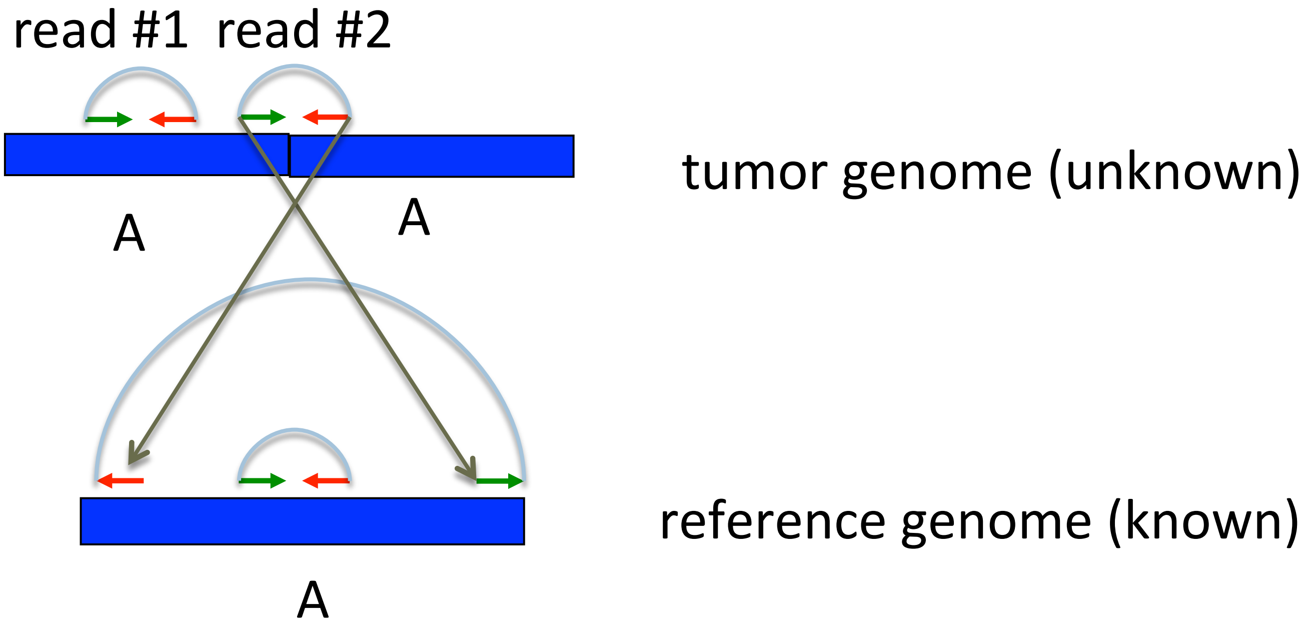
# Tumor Genome – same as reference



- The read from the tumor genome normally maps to the reference genome
  - “Normally” means the gap between two ends of paired-end read is the same (or similar) for both tumor and reference genomes
- Hence, the region contained by this paired-end read is the same for both tumor and reference genomes

# Tumor Genome – Duplication

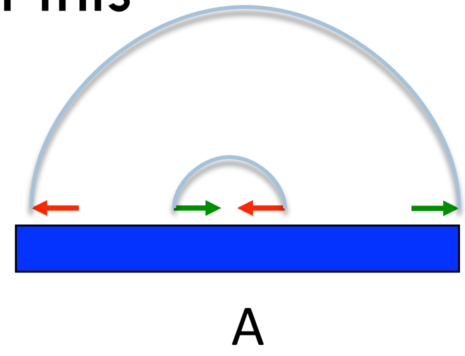
- Assume a region of the reference genome is duplicated in tumor genome



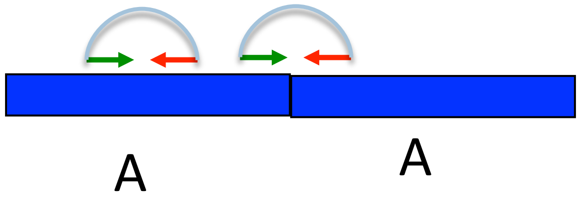
- Read #1 maps normally to the reference genome
- However, when we map read #2, it does not map normally
  - There is a big space between two ends of the paired-end read
  - The order of paired-end read is also different (read #1: green is on left side of region, read #2: red is on left side of region)
- Hence, we can conclude that region A is duplicated in the tumor genome

# Tumor Genome – Clarification

- We are given this

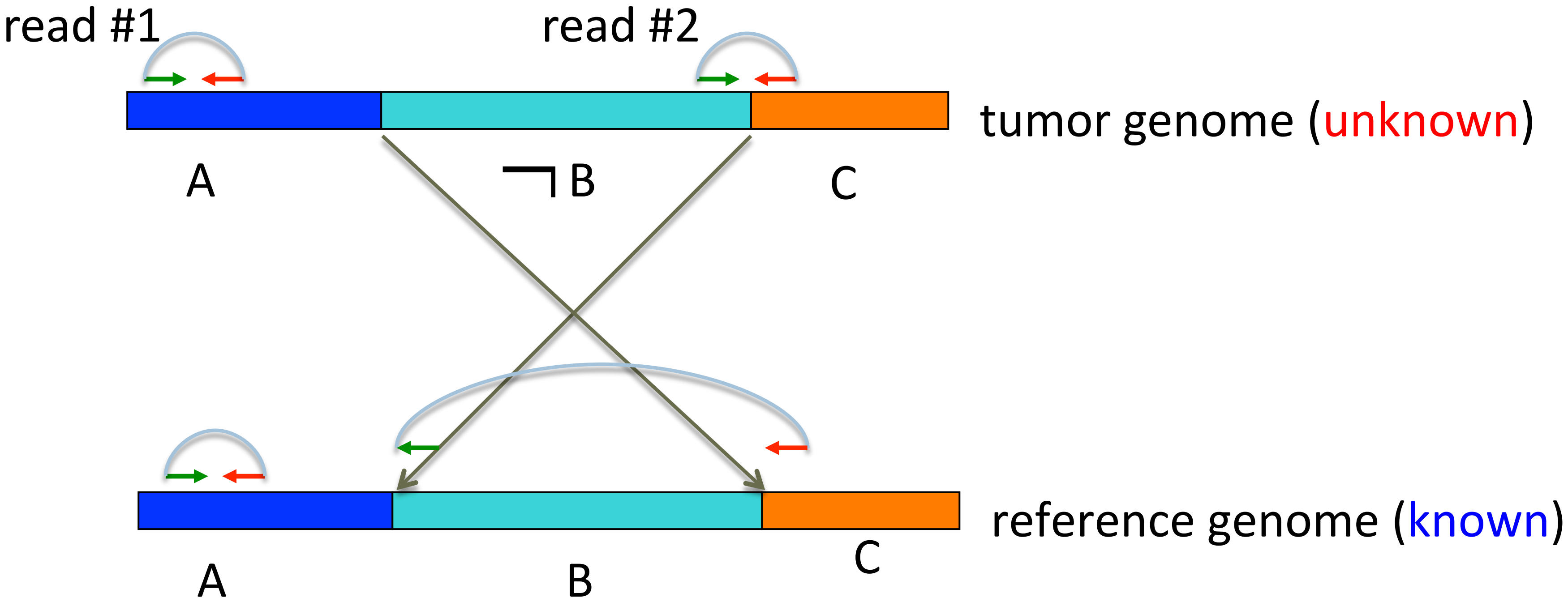


- From this information, we want to reconstruct this



# Tumor Genome – Inversion

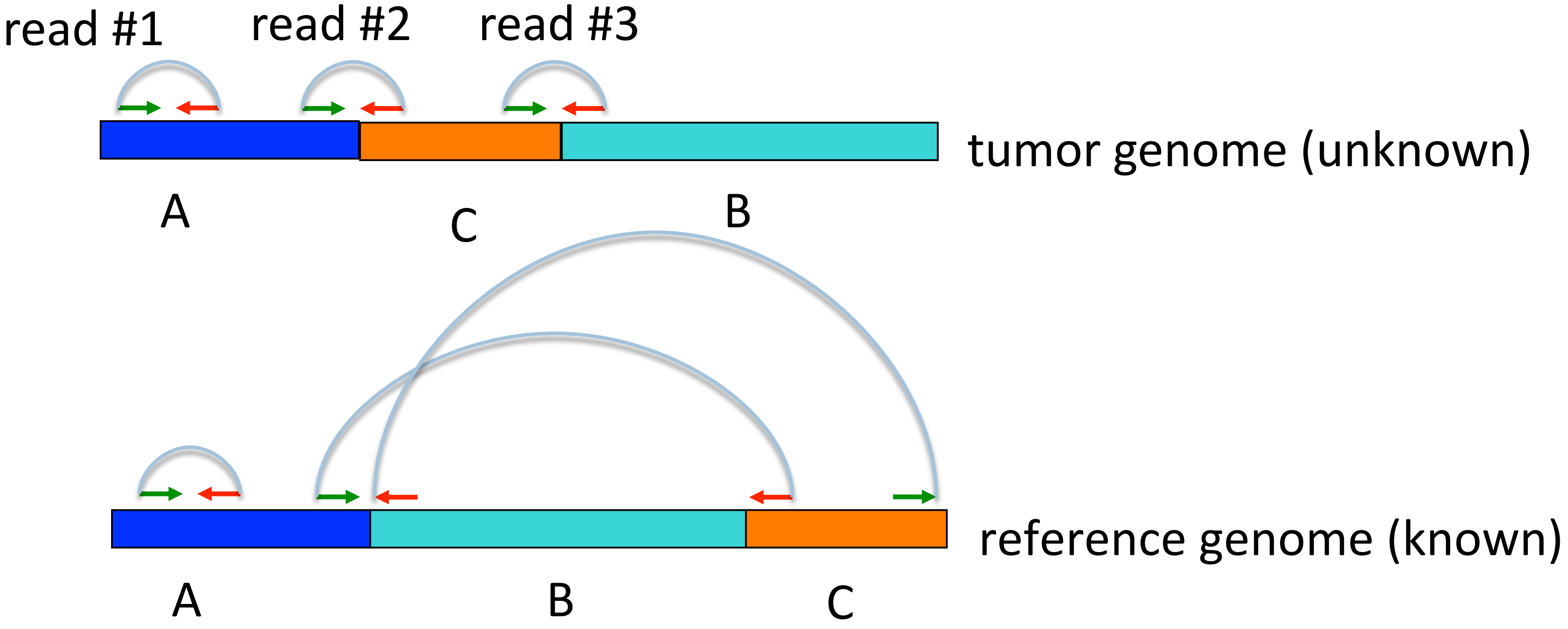
- Assume a region of the reference genome is inverted in tumor genome



- Read #2 does not map normally
  - There is a big space between two ends of the paired-end read
  - The direction of the green arrow is opposite when read #2 maps to the reference genome
- We can conclude that region B is inverted in the tumor genome

# Tumor Genome – Translocation

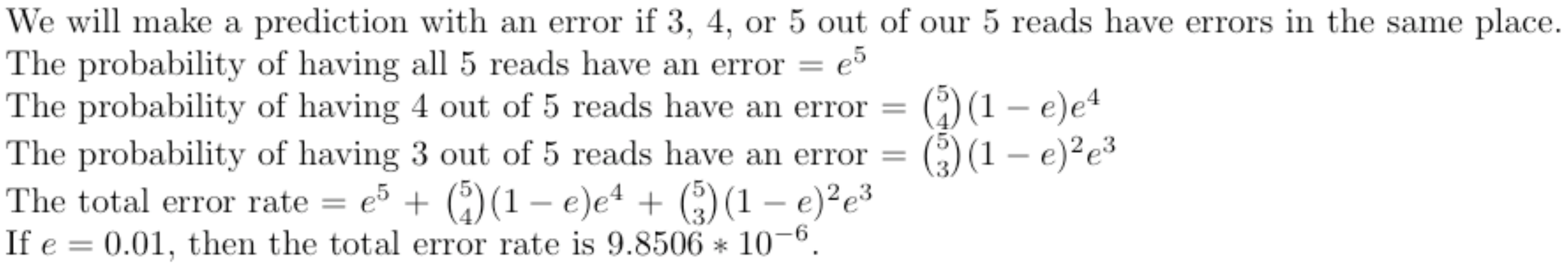
- Assume a region of the reference genome is translocated in tumor genome



- Read #2 and #3 do not map normally
  - There is a big space between two ends of the paired-end read
- We can conclude that region B and C are translocated in the tumor genome

# HW3

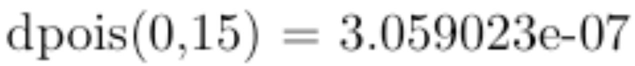
- □ 7. To solve sequencing errors problem, we collect abundant reads (coverage) and take consensus among them. Let's say error rate of reads is 1%, and we are going to predict the consensus sequence. What is the total error rate if the coverage is 5?



We will make a prediction with an error if 3, 4, or 5 out of our 5 reads have errors in the same place.
The probability of having all 5 reads have an error $= e^5$
The probability of having 4 out of 5 reads have an error $= \binom{5}{4}(1-e)e^4$
The probability of having 3 out of 5 reads have an error $= \binom{5}{3}(1-e)^2e^3$
The total error rate $= e^5 + \binom{5}{4}(1-e)e^4 + \binom{5}{3}(1-e)^2e^3$
If $e = 0.01$, then the total error rate is $9.8506 * 10^{-6}$.

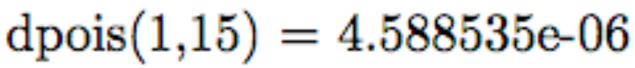
# HW3

- 1. Let’s assume that the coverage is 15 (15x coverage). What percentages of genome have coverage less than 1 (< 1)?



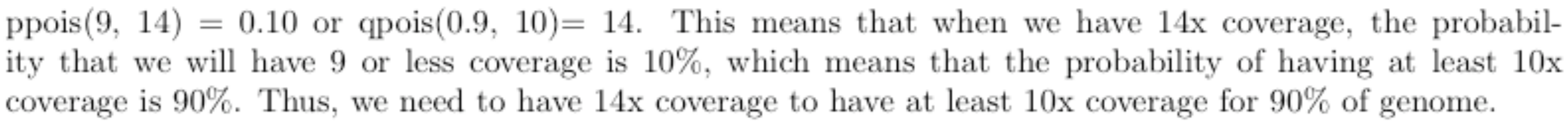
dpois(0,15) = 3.059023e-07

- 2. Again, let’s say the coverage is 15. What percentages of genome have coverage exactly 1?



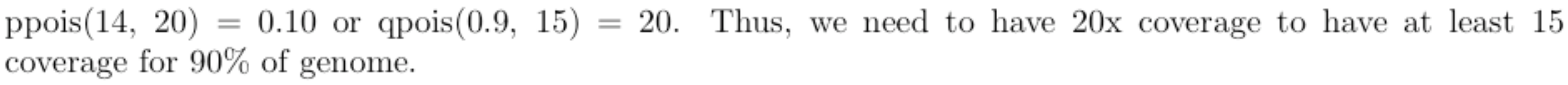
dpois(1,15) = 4.588535e-06

- 3. If we want at least 10 coverage for 90% of genome, what is the minimum overall coverage do we need?



ppois(9, 14) = 0.10 or qpois(0.9, 10)= 14. This means that when we have 14x coverage, the probability that we will have 9 or less coverage is 10%, which means that the probability of having at least 10x coverage is 90%. Thus, we need to have 14x coverage to have at least 10x coverage for 90% of genome.

- 4. What if we want at least 15 coverage for 90% of genome?



ppois(14, 20) = 0.10 or qpois(0.9, 15) = 20. Thus, we need to have 20x coverage to have at least 15 coverage for 90% of genome.

# HW3

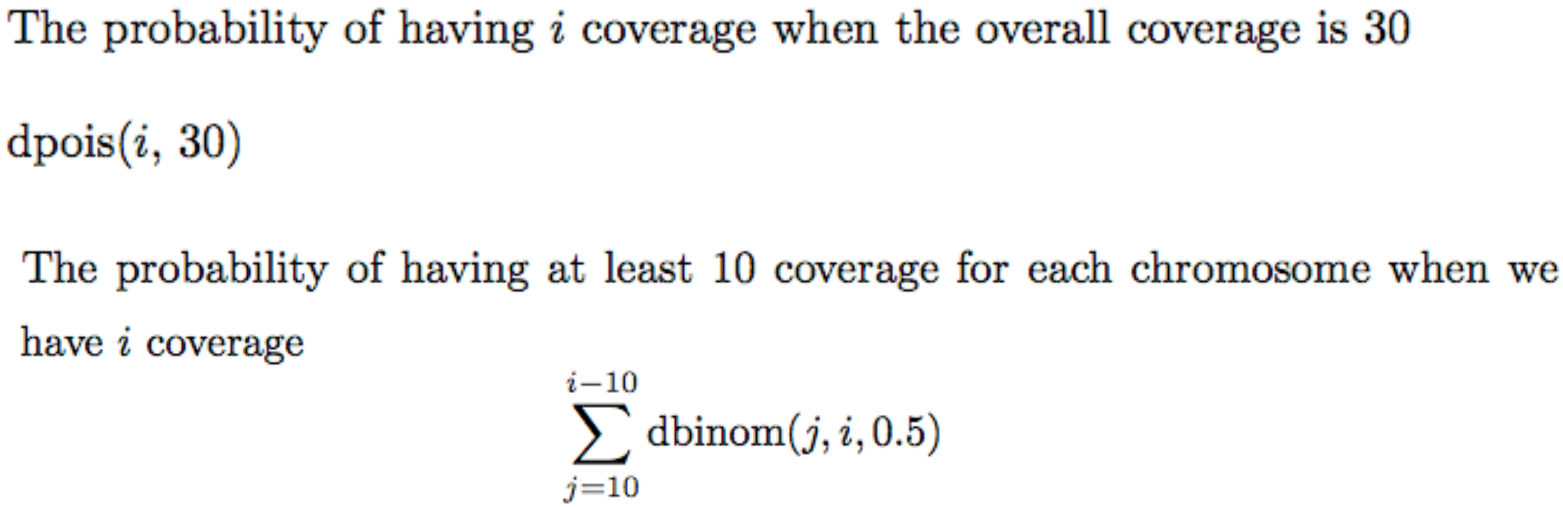
- □ Following problems are related to diplod case. Remember that humans have 2 chromosomes.
- □ 5. Let's assume that for specific positions of genome, we have 30 short reads. What is the probability of having at least 10 coverage for each chromosome? (Hint: Use binomial distribution)

$$\sum_{i=X}^{Y-X} \text{dbinom}(i, Y, 0.5)$$

sum(dbinom(10:20,30,0.5)) = 0.957226. Thus, the probability is about 96%.

# HW3

- 6. Let's assume that the overall coverage is 30. What is the probability of having at least 10 coverage for each chromosome over the whole genome?



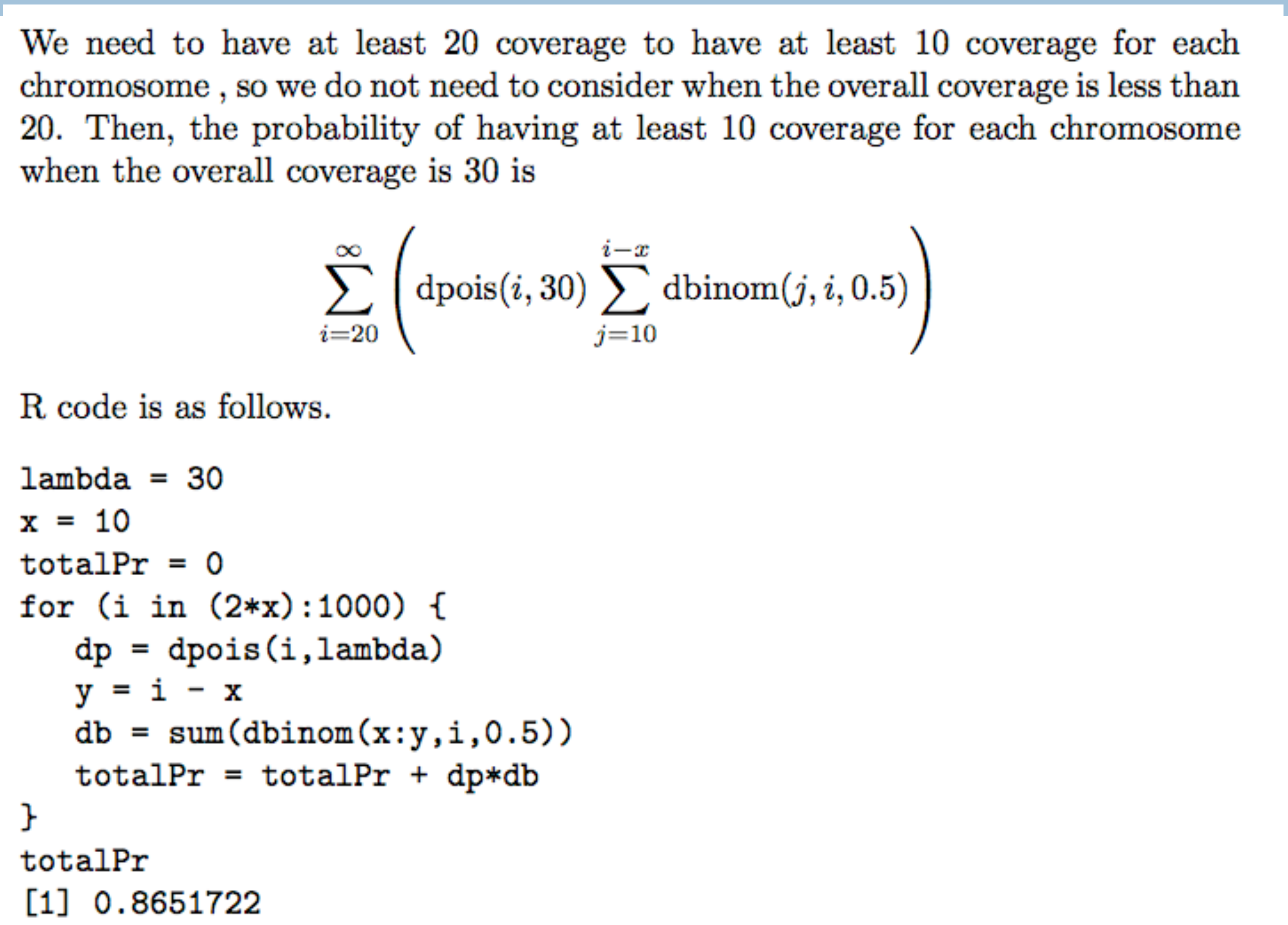
The probability of having $i$ coverage when the overall coverage is 30

dpois($i$, 30)

The probability of having at least 10 coverage for each chromosome when we have $i$ coverage

$$\sum_{j=10}^{i-10} \mathrm{dbinom}(j, i, 0.5)$$

# HW3



We need to have at least 20 coverage to have at least 10 coverage for each chromosome , so we do not need to consider when the overall coverage is less than 20. Then, the probability of having at least 10 coverage for each chromosome when the overall coverage is 30 is

$$\sum_{i=20}^{\infty}\left(\mathrm{dpois}(i,30)\sum_{j=10}^{i-x}\mathrm{dbinom}(j,i,0.5)\right)$$

R code is as follows.

```
lambda = 30
x = 10
totalPr = 0
for (i in (2*x):1000) {
   dp = dpois(i,lambda)
   y = i - x
   db = sum(dbinom(x:y,i,0.5))
   totalPr = totalPr + dp*db
}
totalPr
[1] 0.8651722
```