Week 3 Assembly Issues

Monday, 31 January 2022 8:10 AM

ISSUES

1. we 'can' generate all possible k-mers
2. all k-mers are error free
3. all k-mers only appear once
4. genome is in 1 piece

---

1. how do you generate all possible k-mers

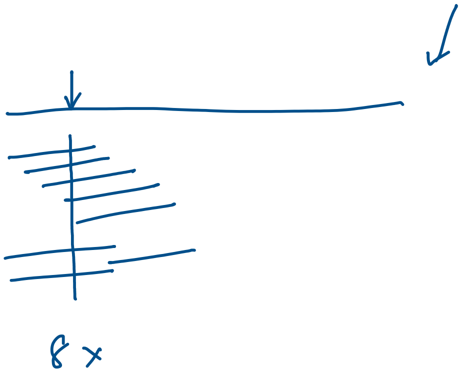
COVERAGE = C

G = haploid genome length

L = read length

N = # of reads

$$C = \frac{LN}{G}$$



~ # of reads that include a particular nucleotide

eg. read length = 150 bp

k = 150] k-mers of length = 40

k = 40 150 - 40 + 1 = 111 - k-mers

↳ smaller the k-mer size ; more likely to capture that across the genome

ISSUE 2

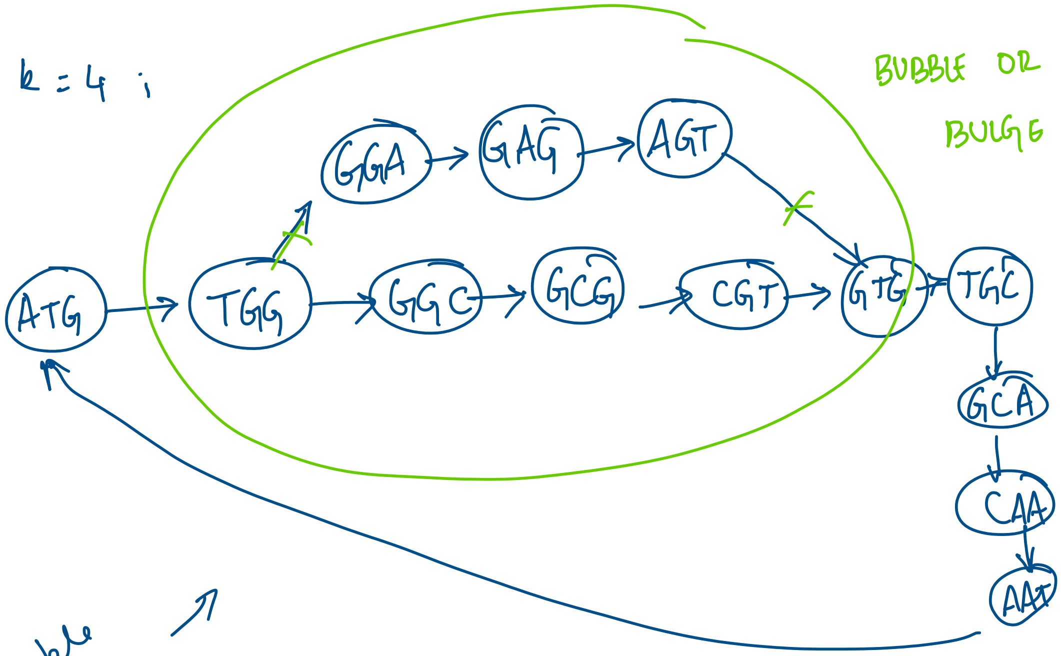
all reads are free of errors

CGTGCAA , TGCAATG , GGCGTGC , ATGGCGT , CAATGGC

ERRONEOUS READ : TGGAGTG

CORRECT READ : TGGCGTG

k = 4 ;



Bubble

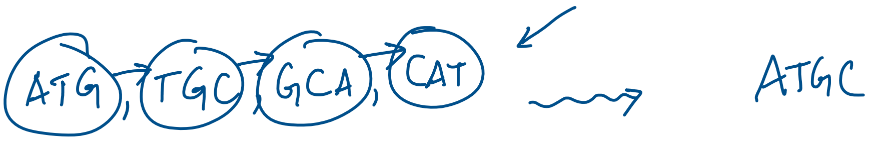
oval

geno

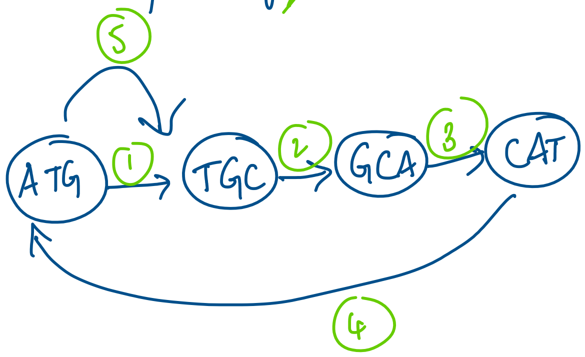
ISSUE 3

Repeats.

eg. genome : ATGC ATGC   k = 4



k-mer multiplicity



ATGCATGC

ISSUE 4

multiple, linear chromosomes?

↳ guided assembly

↳ gaps

VELVET, ALLPATHS-LG, SOAPDENOVO2, UNICYCLER

ASSEMBLY QC

① CONTIGUITY ② COMPLETENESS

↓

N50 → sum all sequence lengths; start at longest contig; observe the length that takes the sum past 50% of total length

eg. 9 contigs

2 Mbp, 3 Mbp, 4, 5, 6, 7, $\boxed{8, 9, 10}$

sum = 54 Mbp ; $\frac{54}{2}$ = 27 Mbp

N50 = 8 Mbp [LENGTH]

L50 - smallest # of contigs that make up 50% of genome

[# of CONTIGS]

eg. L50 = 3

N90, L90, N75 ....

NG50 -> 50% of the actual (known) genome size

-> greater N50 / N90 -> better the assembly; more contiguous

-> smaller L50/L90 ; better the assembly

2 COMPLETENESS

BUSCO , CEGMA ↙]

LTR assembly index -> % of intact]

LTR

greater the completeness, better assembly

3 CONTAMINATION assessment