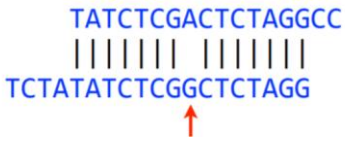
Practice Questions: Key

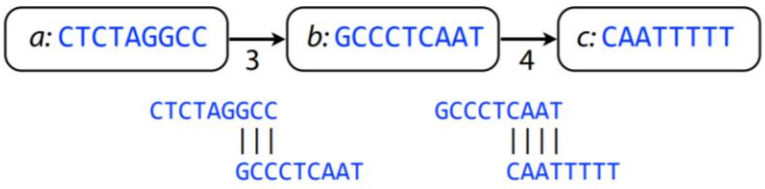
Have a look at these questions and make sure to review the terms in your slides and in the study help guide.

1) Say two reads truly originate from overlapping stretches of the genome like in the figure below. Give one reason why their might there be differences in some of the reads if we are confident they are from the same stretch in the reference genome?



Sequencing error, or from actual allelic differences between loci of a dip- or poly-ploid genome.

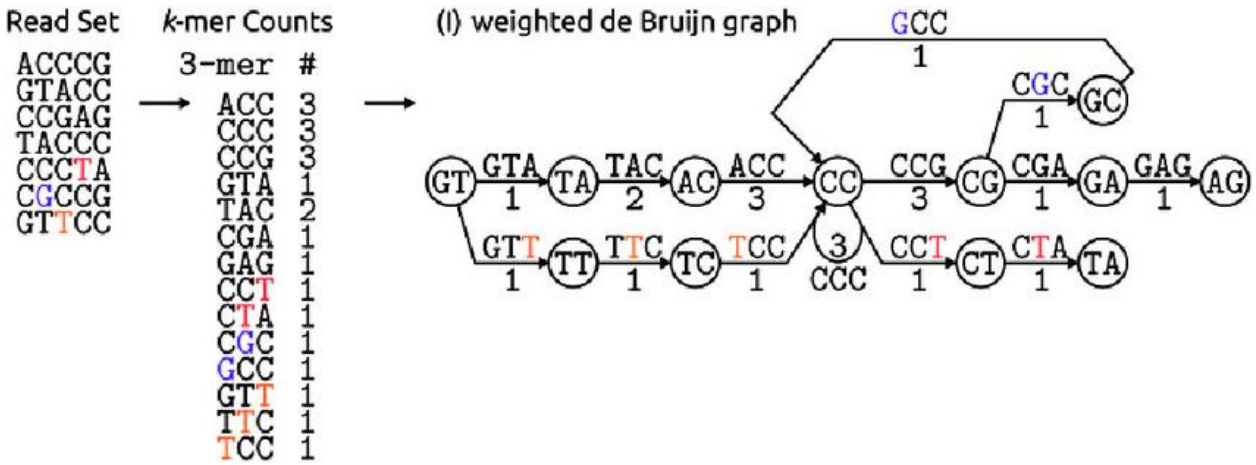
2) Is the graph below an overlap graph or de bruijin graph? Explain why you think its that type of graph?



overlap graph. In a de brujin graph , the direct edges reflect that the suffix of a node (defined by all its nucleotides except the first one) equals/overlaps the prefix of another node (defined by all its nucleotides except the last one).

3) Given a set of reads, length 5, draw a edge centric de brujin graph using kmer length 3, then write out a possible sequence by transversing through the graph. (this has multiple edges for nodes)

ACCCG
GTACC
CCGAG
TACCC
CCCTA
CGCCG
GTTCC



So this question was pretty difficult, there wont be anything this difficult on the exam. But please check the other practice problem regarding de brujin graphs, its good representation of the level we expect you to know.

4) How many 100 bp reads are needed to sequence a 1 Mbp genome (1,000,000 bp) to 5x coverage?

This question depends on the passing rate. Assuming passing rate of 100% youd need about 50000 reads, with an 80 % passing rate youd want 62,500 reads, and then assuming its illumina which from a quick google check have a passing rate of 90% , youd want about 55,556 reads or so. Reads = ((coverage*size of genome) / (read length* passing rate)

90 % passing rate, Number Reads = (5*1000000)/(100*0.9) = 5000000/90 = 55,555 reads

Question somewhat like this, if equation is needed then it will be provided. Take a look at RPKM.

5) You sequence as assembly and the contig sizes are as follows.
35Mb, 40Mb,8Mb,7Mb, 12Mb, 17 Mb and 21 Mb. Calculate the N50.

total assembly size = 140 Mb

rank contigs in size : 40,35,21,12,17,8,7

N50 = min contig length at which at least 50 % of the assembly is contained on

50% of genome = 70 Mb,

Keep adding contigs from largest to smallest until their sum is greater than 50% of the assembly size.

70 Mb > 40 Mb : Not yet

70 Mb < 40 Mb + **35 Mb** : Yes , 40+35 = 75 Mb > 70 Mb

So our N50 is the 35 Mb, 50 % of our assembly is contained on contigs 35 Mb or larger.

DON'T WORRY MUCH ABOUT QUESTIONS 1 and 2.

1. Your team runs a genome sciences centre that has recently received $500,000 USD worth of sequencing funding from a Swaziland start-up to sequence an economically relevant salamander, *Necturus swazilandeus* that produces morphogens with the potential to regenerate human limbs. The Swazi researchers are interested in looking at loci conferring limb-regeneration capabilities. Using cytogenetic information, the genome size has been estimated to contain 20 billion base-pairs (Gbp). The genome is known to be highly repetitive and repeats can be as long as 6 Kbp. What sequencing platform(s) will you choose to use for sequencing this genome assuming you want 50X coverage*? How will this impact your choice of assembly paradigm [de Bruijn | OLC]? (Hint: use an equation to estimate number of sequences and cost per Mbp.)
   * The coverage must sum to 50 for any combination of technologies (e.g. 40X PacBio and 10X Sanger sequencing). At least 10X coverage is required for any one technology used.

10X of PacBio sequencing far exceeds the cost of this project alone ($800,000). The repeats still present a problem and must be sequenced through with a long read sequencing technology to avoid the presence of gaps which arise anytime a repetitive sequence is longer than the fragment that is sequenced.

Start with 10X coverage of Oxford Nanopore long sequencing reads:

$$G = (10x \text{ coverage}) * (20{,}000 \text{ Mbp})$$

$$200{,}000 \text{ Mbp} * \$2/\text{Mbp} = \$400{,}000$$

The remaining 40X coverage can be accomplished using the Illumina HiSeq:

$$\text{HiSeq cost} = \$0.0018/\text{Mbp} * (20{,}000 \text{ Mbp} * 40x \text{ coverage}) = \$1{,}440$$

(Exhausting the remainder of the budget on HiSeq sequencing is another possibility…)

Illumina MiSeq would be slightly too expensive by this nascent start-up company:

$$\text{MiSeq cost} = (\$0.133/\text{Mbp}) * (20{,}000 \text{ Mbp} * 40x \text{ coverage}) = \$106{,}400$$

There are a few ways to combine the short and long sequences to generate a quality genome assembly with few errors and gaps but the best results are obtained when one first corrects the sequencing errors in the long reads by aligning the short reads to these long fragments and finding the consensus sequence (indeed, this involved a multiple-sequence alignment). In the end, the only assembly paradigm that is required is OLC.

A second, slightly less correct but still valid answer is assembling the short reads and long reads separately and merging the assemblies by following the tiling path. (Note: can you see why this may lead to a poorer assembly?)

2. You are working at the Canadian Centre for Disease Control when a viral epidemic breaks out in Vancouver. The virus is a virulent strain of beaver pox that appears to have crossed the species barrier causing infected humans to crave the taste of wood. Your team has isolated the viral particles in a blood sample and needs to sequence it immediately in order to develop an effective vaccination strategy that saves Stanley Park. The viral genome is a circular double stranded DNA molecule approaching 42 Kb in length. Your sample contains approximately one virus nucleotide per 100 thousand human nucleotides. With a sequencing budget of $80,000, what sequencing platform will you use to sequence this "metagenome" assuming you want 100X coverage? How will this impact your choice of assembly paradigm [de Bruijn | OLC]? (Hint: use an equation to estimate number of sequences and cost per Mbp.)

The most cost-effective way to sequence the virus is using the HiSeq but using the MiSeq is would yield a better assembly while still meeting budget restraints. Using anything else would not be reasonable for the amount of sequencing required. The following equation is for the HiSeq and the read length is 100bp. This could be modified for 125bp or 150bp reads but we did not provide information for these run types.

$$G = (42{,}000*10\text{^}5bp)$$

$$c = LN/G$$

$$100 \text{ reads/bp} = 100bp*(N \text{ reads}) / 4{,}200{,}000{,}000bp$$

$$N = 4{,}200{,}000{,}000 \text{ reads}$$

For an Illumina HiSeq 4000, this is equivalent to $42*10^4$ Mbp.

Cost = \$0.0018/Mbp * $42*10^4$Mbp = \$756

Alternative calculation using \$/read (where 0.00000018\$/read on HiSeq 4000) comes to the same answer.

Using a different equation and MiSeq which takes the pass rate into account:

$R_N = CT/rL(P_f)$

$= 100*(42{,}000*10\text{^}5bp)/300(0.8)$

$= 1.4*10^9(0.8)$

$R_N = 1.12*10^9 = 3.36*10^5$ Mbp

336,000 Mbp * ($0.133 / Mbp) = $44,688

The only way to assemble the HiSeq data is by using a de Bruijn graph based assembler, and the recommended method for assembling MiSeq data is also using a DBG-based assembler. Cannot use OLC since there are FAR too many reads!

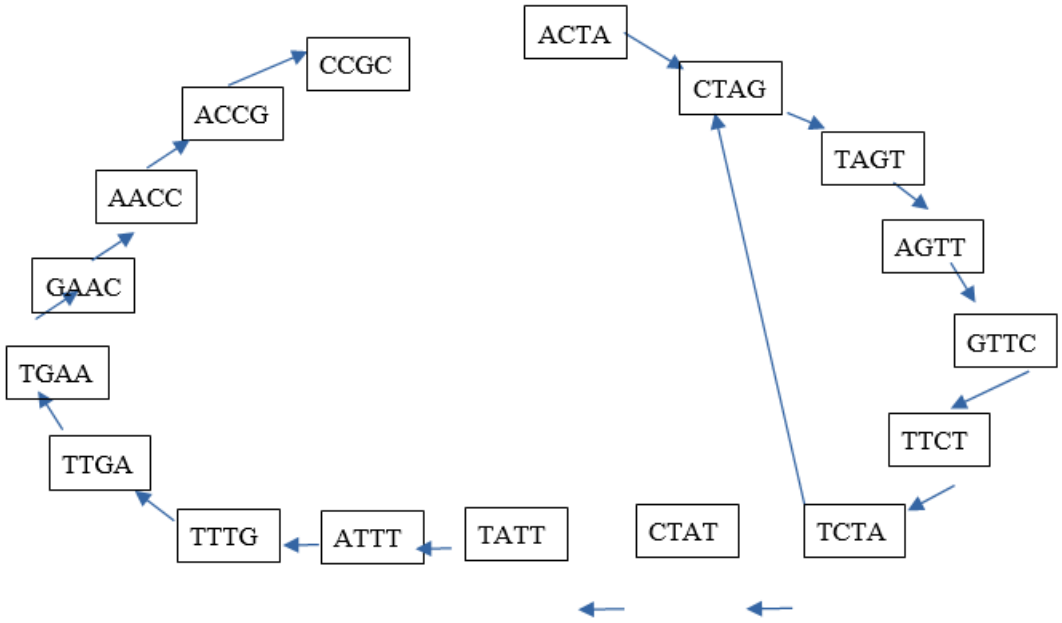
3. As a standard quality control step you are inspecting four reads from a sequencing run conducted at your genome sciences centre:

ACTAGTTCT TTTGAACC TTCTATTTG TGAACCGC

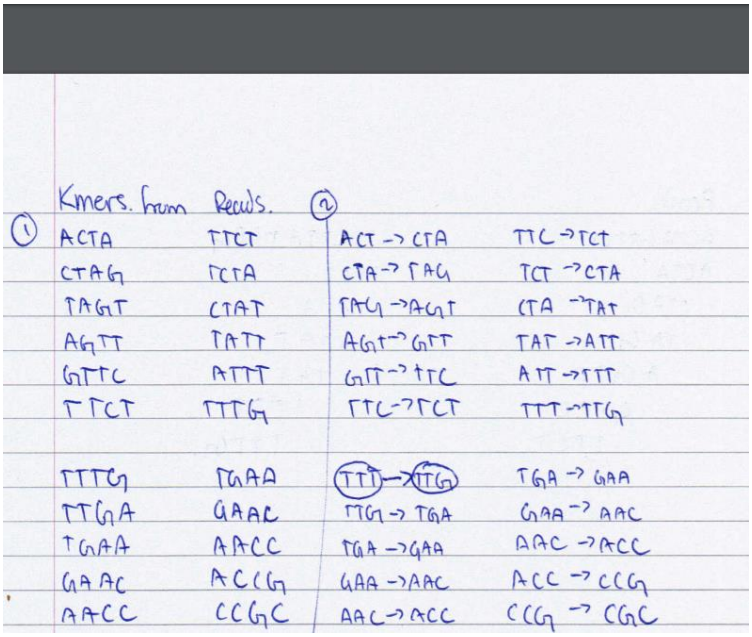
Write out a de Bruijn graph representation of these reads using a k of 4. (Hint: First decompose each read into k-mer space and then find the overlaps between (k-1) mers.)

ACTA AGTT	CTAG CTAT	GTTC **GAAC**	TAGT **TTCT**
AACC	CCGC		TTTG
ATTT ACCG			TTGA **TGAA**
			TCTA
			TATT

Above are the 4-mers that are decomposed from the reads. Bolded 4-mers are duplicates. **There are 2 ways to correctly represent a de Bruijn graph**, actually. Either nodes are k-mers and edges represent an overlap of k-1 OR nodes are k-1 mers and edges are k-mers, representing an overlap of k-2. Here is the de Bruijn graph that would solve this question, using the first representation:



For edge centric, its pretty much the same , expect the edges are the kmers (4mers in this case), and the nodes are k-1. I take my 4 mers , and I can create two nodes, with the kmer at its edge. Then I just match nodes and build the graph. The numbers just mean how many times that edges comes up (don't worry about it), and then I take eulerian walk through the graph to transverse every edge.



You can see the graph is pretty similar to the other graph , but in our case, we want to transverse every edge exactly once, eulerian. In the node centric, you want to find the hamilitonian path which is the one that visits each node etc.