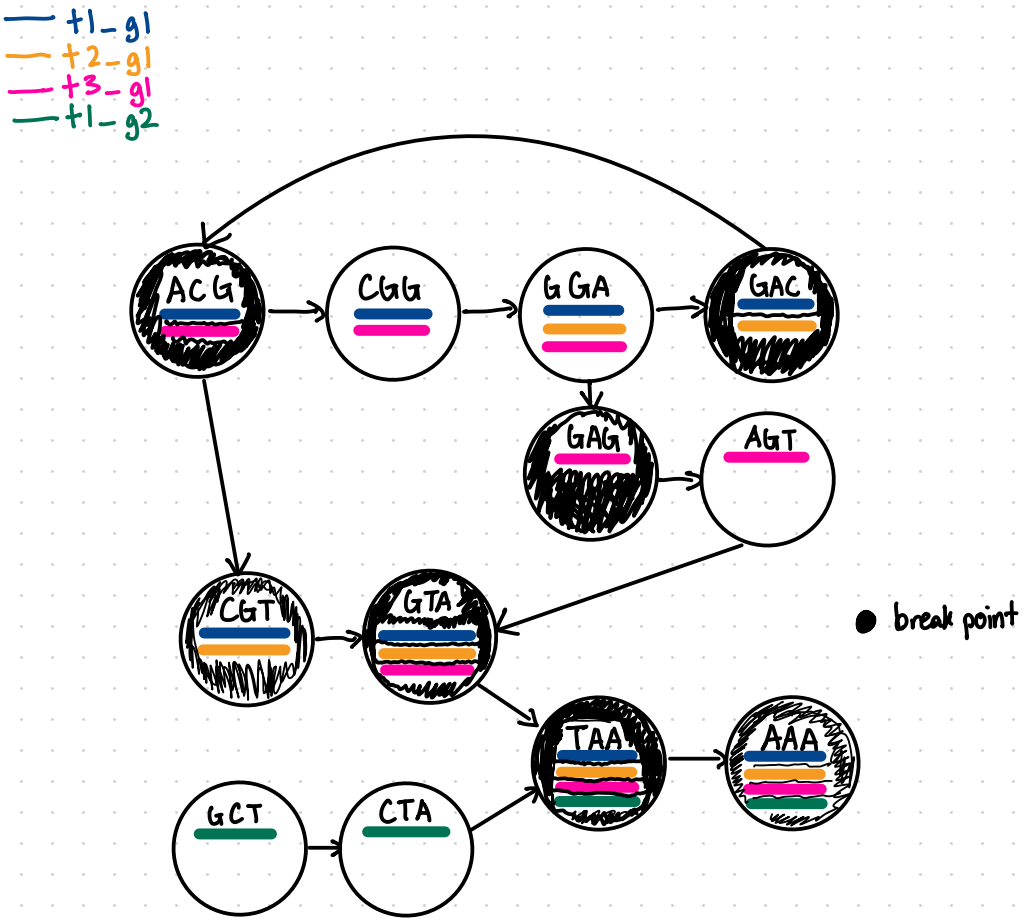
a)


+1_g1
+2_g1
+3_g1
+1_g2
ACG
CGG
GGA
GAC
GAG
AGT
CGT
GTA
TAA
AAA
GCT
CTA
break point

b) First, you start at the 1st k-mer on the constructed de Bruijn, which is GGA, and note the equivalence classes. From there, you can jump to the next breakpoint, which is GAC and get the equivalence classes at that node. The breakpoints are nodes that give you essential distinguishing information. Then, continue to the next breakpoint at the ACG node, then continue to the last breakpoint node at CGT. Finally, get the intersection of all the equivalence classes for the breakpoint nodes traversed through an Eulerian walk.

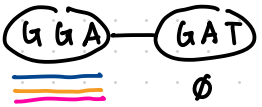
GGACGT

$\cap \cap \cap$ = Gene 1, T_1 and Gene 1, T_2

Read aligned to gene 1, transcript 1 and gene 1, transcript 2

c) First, start at the 1st k-mer of your sequence, which is GGA and note the equivalence classes. Then, go to the next breakpoint, but in this case there are no breakpoints in the constructed de Bruijn graph, which contain any of the remaining k-mers, so the equivalence class would be the null set. Thus, the intersection of your equivalence classes is the null set meaning there is no alignment of this read that is valid.

GGATGT



$\cap \emptyset = \emptyset$

No isoforms of any gene mapped to blc intersection is $\emptyset$.

d) My algorithm will involve replacing the errored based with all other possible bases. Meaning if the sequence is GGATGT and the error was at position 4, alternative sequences will be tested as well including GGAAGT, GGAGGT, and GGACGT. All possible nucleotides will be substituted at the error positions to create new sequences, which will then be pseudoaligned using the constructed de Bruijn graph. Then, each generated sequence will be compared to the reference sequence and the alternative sequence that matches the most with the reference transcriptome will be the aligned sequence. So the alternative sequence with valid intersection equivalence classes will be the one aligned to the transcriptome instead. To account for possibly aligning to multiple positions, you could incorporate information about expression levels to filter out less likely alignments. Some benefits of this would be that there will be reduced false alignments since there's multiple potential corrections, but it doesn't eliminate all false alignments. Some drawbacks include computational efficiency where evaluating all possible corrections can be computationally intensive, especially when you don't know where your error is located in the sequence and also when you have longer sequences. It may not work when you don't have a reference transcriptome or if your reference transcriptome has errors. It also may not work if you have multiple errors because this approach only works if you assume there's a single error in your sequence, not multiple.

e) Given the sequence TTTACG, I notice that it doesn't pseudoalign to any sequences directly but it's reverse complement, CCCTGC does pseudoalign directly. RNA-seq reads can be generated from either the 5' to 3' direction or 3' to 5' direction because RNA molecules can be transcribed in either direction and adapters can be annealed to both ends of complementary DNA. So, to account for this it's important to account for both orientations of a read. So to pseudoalign TTTACG, we should also attempt to pseudoalign its reverse complement as well. This allows us to capture all possible alignments regardles of the read's original orientation and improve the efficiency and accuracy of the alignment process.