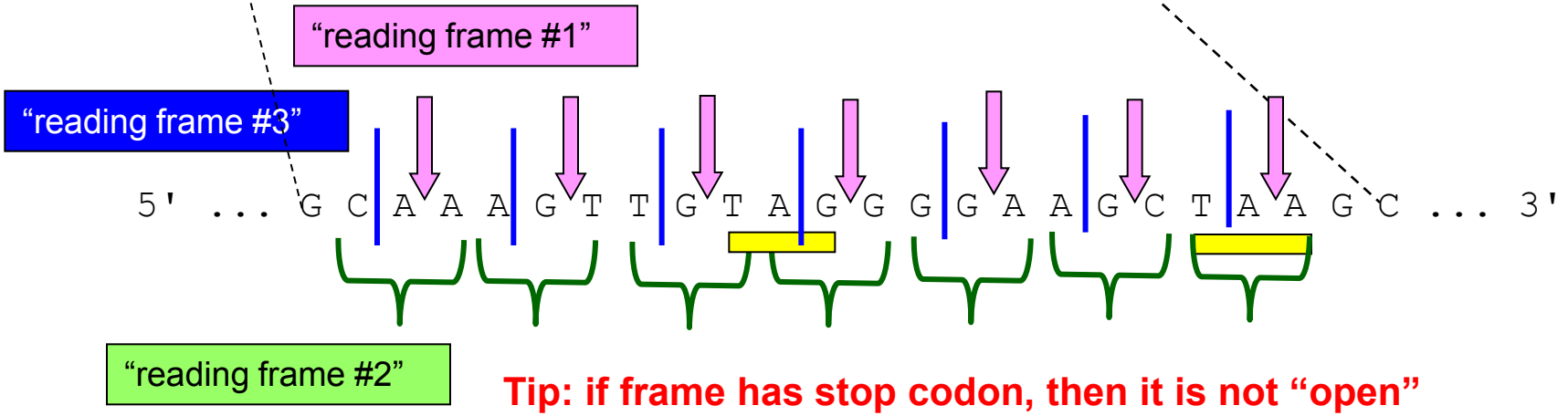
Practice set #1, question 1

Does this sequence have the potential of encoding part of a protein?

```
5′...  GCAAAGTTGTAGGGGAAGCTAAGCTCGAAATAAGGTGTGCCTATT...  3′
3′...  CGTTTCAACATCCCCTTCGATTCGAGCTTTATTCCACACGGATAA...  5′
```

For a segment of sequenced DNA that is suspected to contain an ORF,

… divide into triplets starting at positions 1, 2, and 3



Tip: if frame has stop codon, then it is not "open"

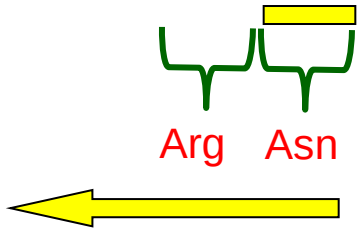
Reading frame # 3 is "open" over the whole region shown...

and its initiation codon would be located upstream of the given sequence

… so it is a better candidate than the other two (shorter) ones

However, gene might be in opposite orientation, so examine 3 reading frames on complementary strand too

```
5′... GCAAAGTTGTAGGGGAAGCTAAGCTCGAAATAAGGTGTGCCTATT... 3′
3′... CGTTTCAACATCCCCTTCGATTCGAGCTTTATTCCACACGGATAA... 5′
```



Second letter

First letter	U	C	A	G	Third letter
U	UUU, UUC: Phe UUA, UUG: Leu	UCU, UCC, UCA, UCG: Ser	UAU, UAC: Tyr UAA: Stop UAG: Stop	UGU, UGC: Cys UGA: Stop UGG: Trp	U C A G
C	CUU, CUC, CUA, CUG: Leu	CCU, CCC, CCA, CCG: Pro	CAU, CAC: His CAA, CAG: Gln	CGU, CGC, CGA, CGG: Arg	U C A G
A	AUU, AUC, AUA: Ile AUG: Met	ACU, ACC, ACA, ACG: Thr	AAU, AAC: Asn AAA, AAG: Lys	AGU, AGC: Ser AGA, AGG: Arg	U C A G
G	GUU, GUC, GUA, GUG: Val	GCU, GCC, GCA, GCG: Ala	GAU, GAC: Asp GAA, GAG: Glu	GGU, GGC, GGA, GGG: Gly	U C A G

Example of computer output for virtual translation (for all 6 frames) shown in one-letter amino acid code

Upper strand

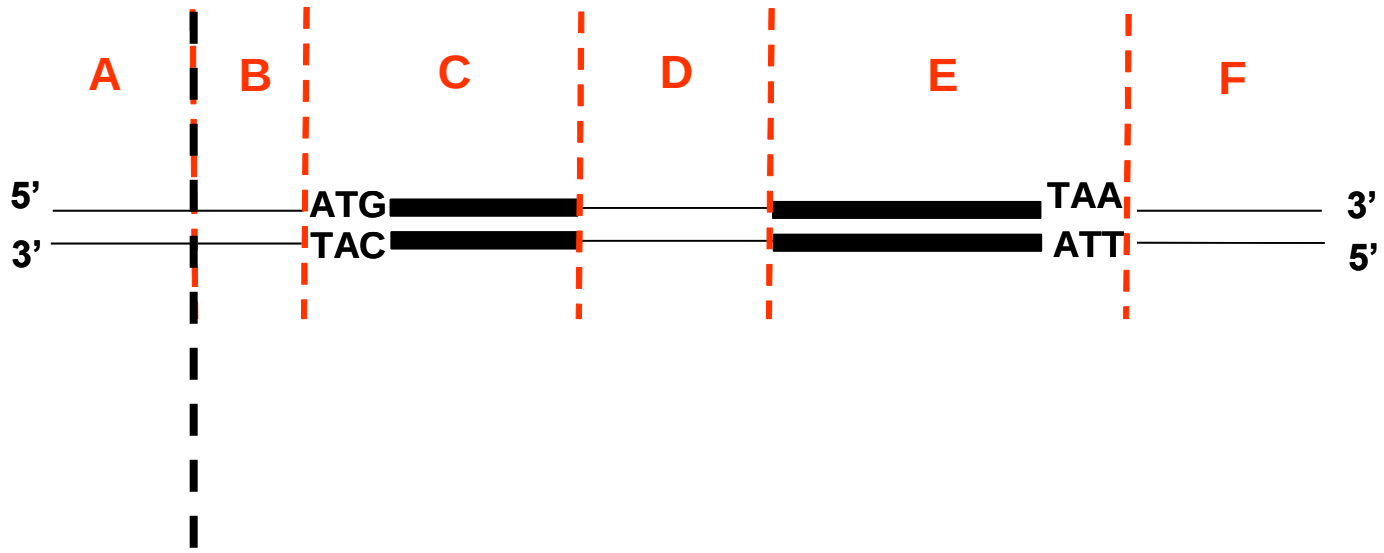
```
>myseq
DNA: GCAAAGTTGTAGGGGAAGCTAAGCTCGAAATAAGGTGTGCCTATT
 +3:   K  V  V  G  E  A  K  L  E  I  R  C  A  Y
 +2:  Q  S  C  R  G  S  *  A  R  N  K  V  C  L
 +1: A  K  L  *  G  K  L  S  S  K  *  G  V  P  I
```

Indicates stop codon

"Lower strand" (shown 5' to 3')

```
>myseq
DNA: AATAGGCACACCTTATTTCGAGCTTAGCTTCCCCTACAACTTTGC
 -1: N  R  H  T  L  F  R  A  *  L  P  L  Q  L  C
 -2:  I  G  T  P  Y  F  E  L  S  F  P  Y  N  F
 -3:    *  A  H  L  I  S  S  L  A  S  P  T  T  L
```

Practice set #1, question 2



For the DNA region shown above, give the locations (using symbols A-F) of:

Gene C-D-E (Structural gene)

Exon C and E (as often shown in texts & papers, but in reality first & last exons also contain UTRs – RNA info is needed to know their lengths)

Intron D

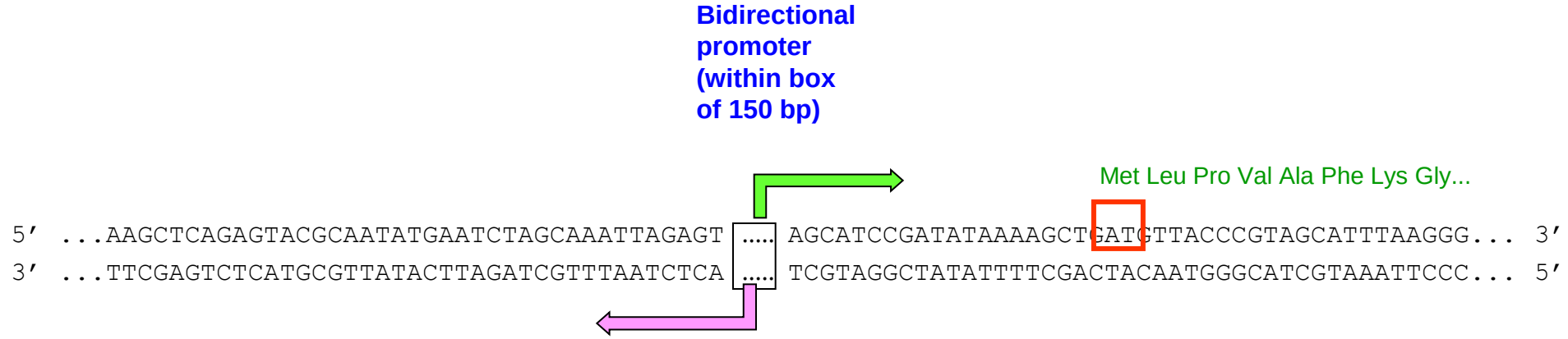
Translation initiation site B/C junction

Translation termination site E/F junction

Transcription initiation site A/B junction

Promoter Slightly upstream of A/B junction

Practice set #1, question #3



Upper strand: look for ATG (other 2 frames have stop codons)

Lower strand: for protein-coding gene in opposite orientation, look for 3'... GTA... 5'

But there aren't any!!!

& all three frames have stop codons

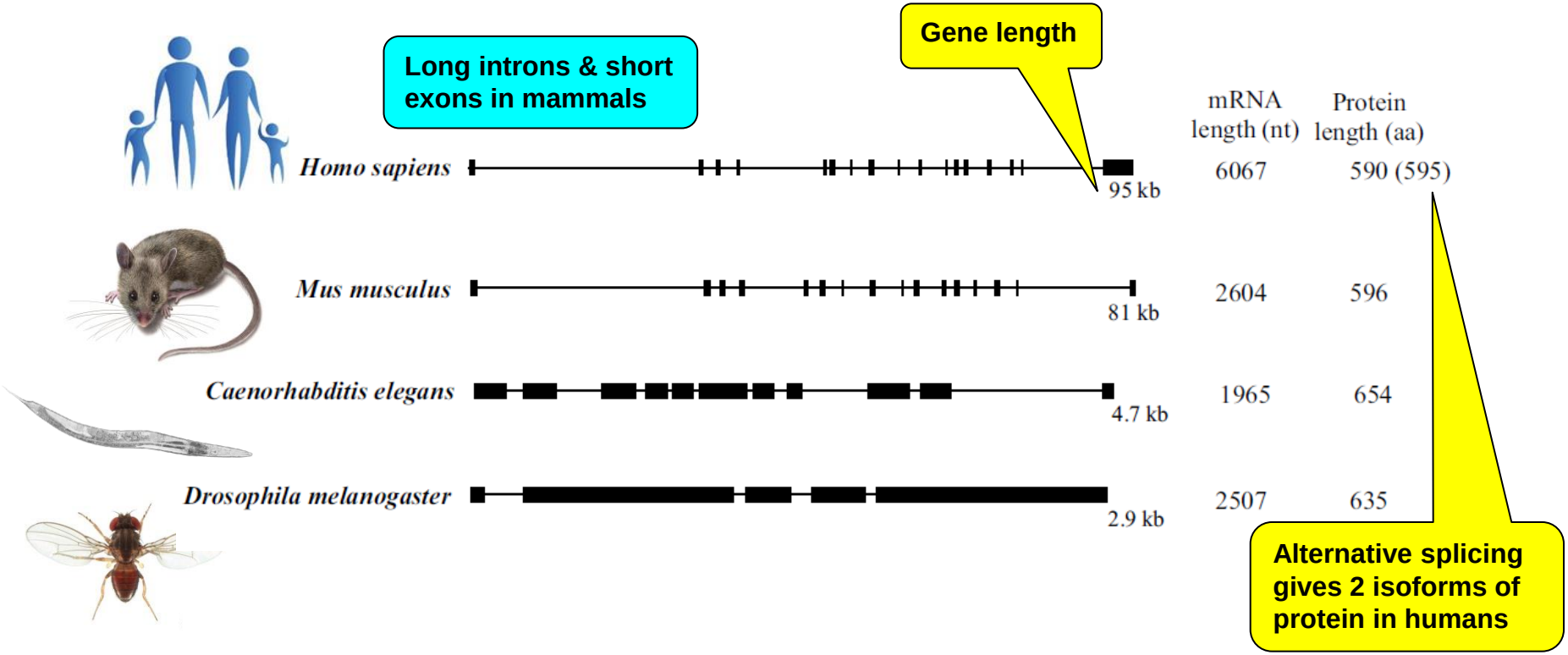
- so maybe long 5' UTR for this gene... ie. start codon is further to left...
- or maybe very short 5' UTR (ie. within box of 150 bp)

But if so, must be short ORF since all 3 frames have stop codons for region shown

- or maybe it's a structural RNA gene (eg. tRNA...)

Practice set #1, question #4

Structure of NF2 (neurofibromatosis type II) gene in various animals



What features of this gene are different among these animals?

Gene is much longer in mouse & human than in nematode or Drosophila

… because more introns and longer introns — 4 introns in Drosophila gene vs. 10 in nematode & 17 in mouse/human

But NF2 protein is longer in invertebrates (~ 650 aa) than in mammals (~ 595 aa)

Human mRNA longer than mouse mRNA…

because 3' UTR is longer

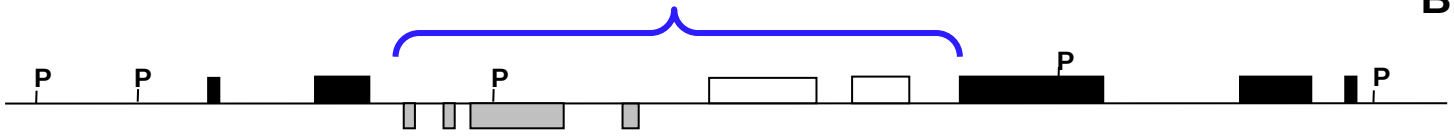
Maybe extra internal domain or extension at N- or C-terminus?

Golovnina et al. *BMC Evol Biol* 2005

Practice set #1, question 5

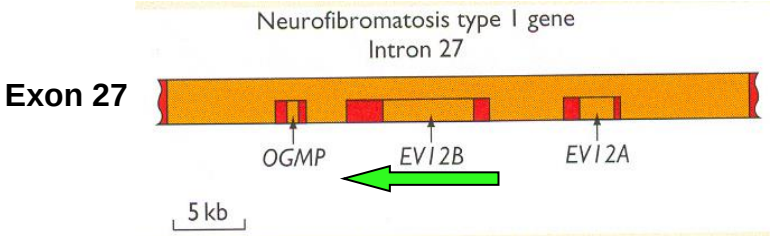
"Genes-within-genes" (Topic 1)

Blocks = exons



Grey gene is in opposite orientation

Aside: This question was modeled on the human neurofibromatosis type I gene, with 3 genes inside intron 27 & one [EV12B] in opposite orientation.

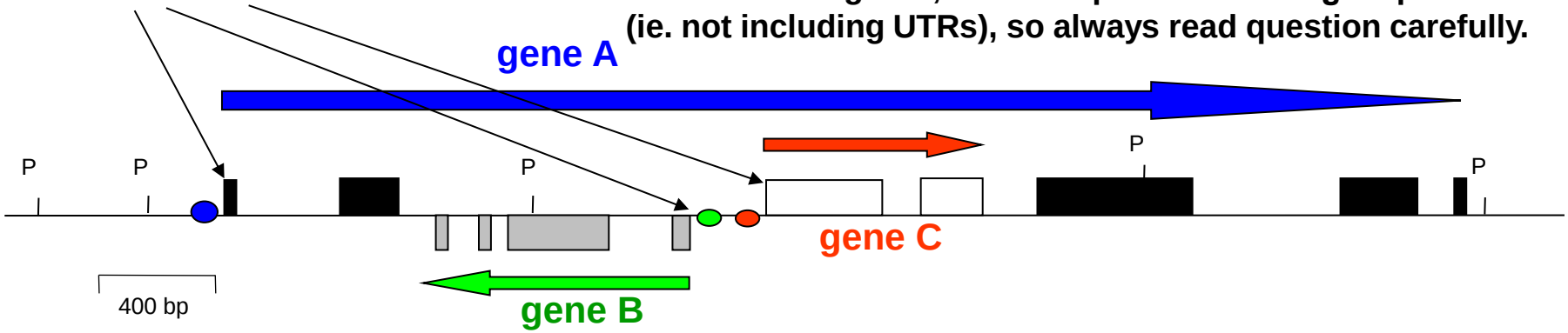


Brown 2d ed. Box 2.2 p.44

Transcription initiation sites slightly downstream from promoters

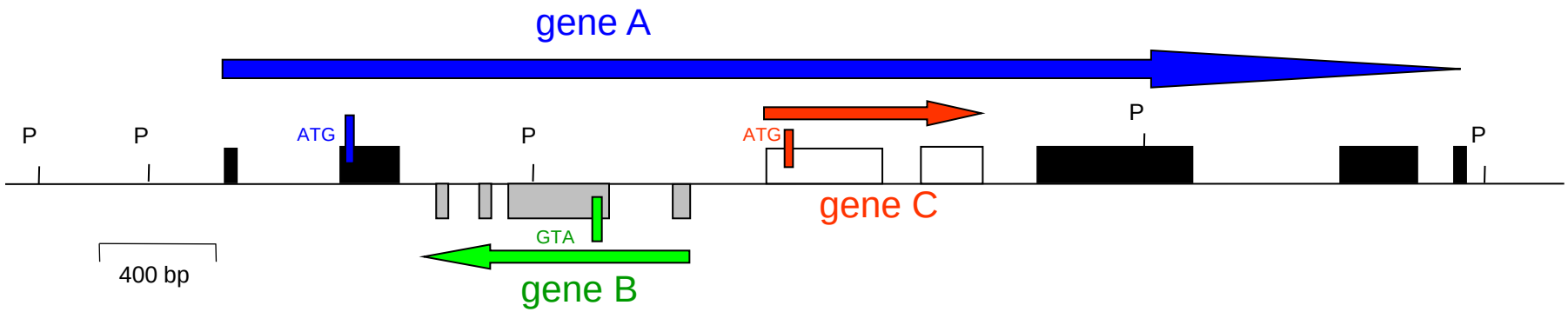
Fundamental definition: "introns" spliced out & "exons" remain in mRNA

NB: In some figures, blocks represent "coding sequences" (ie. not including UTRs), so always read question carefully.



gene A (exons = black blocks), gene B (exons = grey blocks), gene C (exons = white blocks)

(ii) Where are the translation initiation sites for genes A, B & C?



In this question, told that UTRs ~ 100 nt long

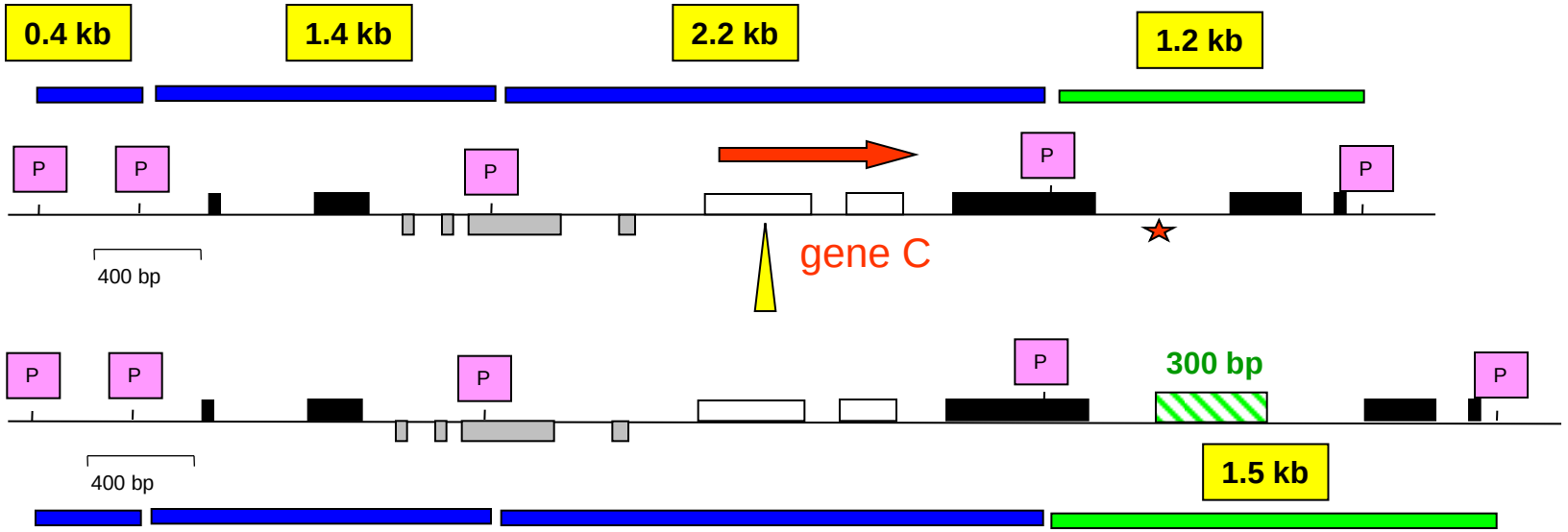
… so ATG will be ~ 100 nt downstream of transcription initiation site (using scale on left, can measure with ruler where start codon will be)

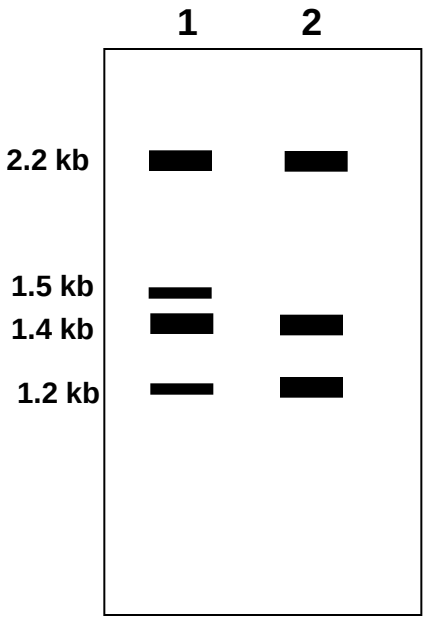
For gene A and gene B, there is an intron located within region corresponding to the 5' UTR

... as is the actual case for EV12B in human neurofibromatosis gene

(iii) Using ruler on figure… protein C is estimated to be ~ 133 amino acids long

Practice set #1, question 5b



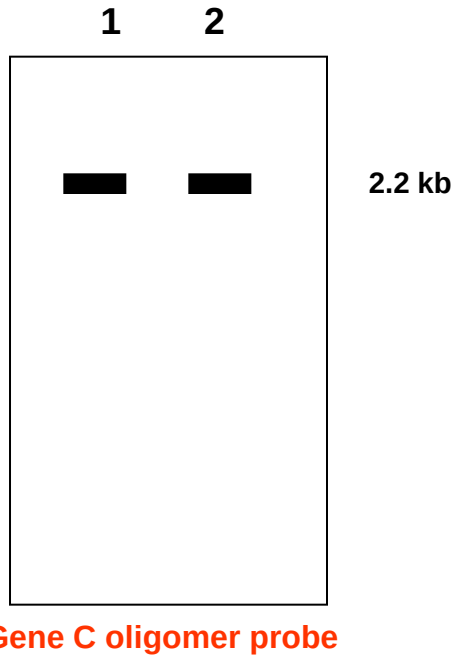


"rare" allele, so father is almost certainly heterozygous

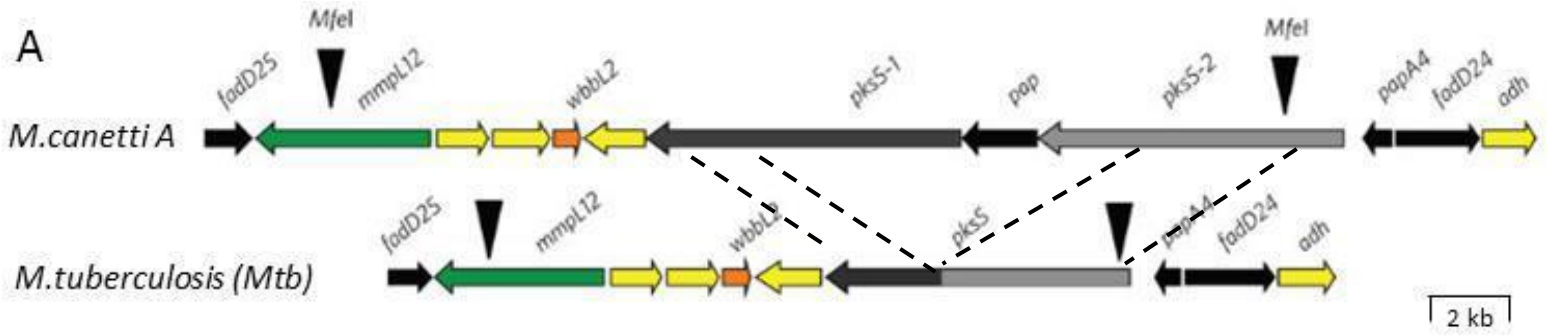
50% probability that mouse X inherits "rare" allele

if yes, profile same as father

If not …

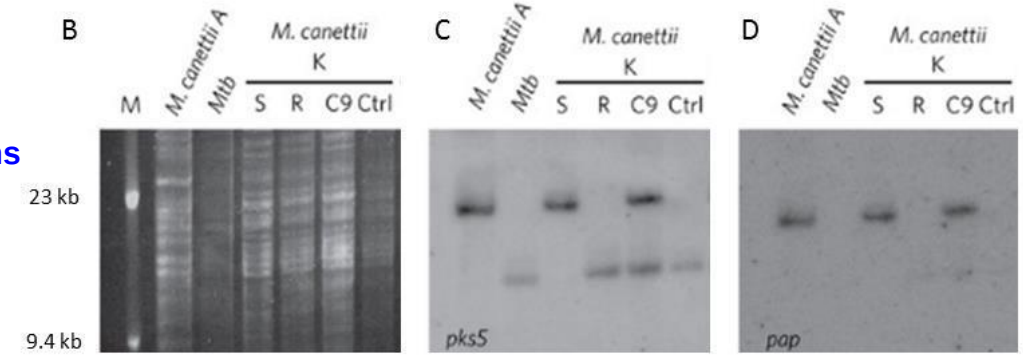


Practice set #1, question 6



Maps for M.cannetti & Mtb look same, except for pks5-1/pap/pks5-2 region

Southerns



***MfeI* restricted DNAs pulsed-field gel**

Probe pks5

Probe: pap

Unstable genome? Homologous recombination across similar sequences in pks5-1 and pks5-2?

Lanes 1 & 2 agree with map : hybridization signals of ~ 20kb and ~ 12 kb with psks5 probe and no signal with pap gene probe for Mtb…

M.canetti strain K variants (S,R,C9,Ctrl) differ in their profiles

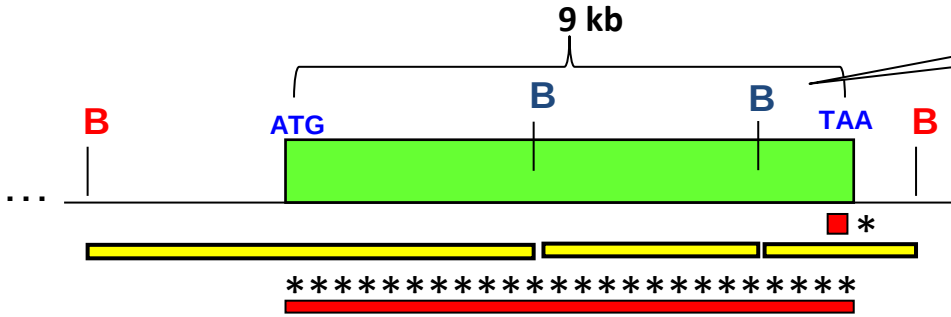
C9 has both "long" and "short" forms

Title of paper: "pks5-recombination-mediated surface remodelling in Mycobacterium"

Boritsch *Nature Microbiol.* 1:15019, 2016

Practice set #1, question 7a

Mouse gene Y = 9kb long, with 2 BamHI sites



But not told exactly where the BamHI sites are within gene Y, so wherever you choose is ok. (It will affect the size of the internal BamHI fragment.)

And we also don't know where <u>flanking</u> BamHI sites are located…

… but we do know 6-bp cutter sites are <u>on average</u> ~4^6 bp apart (~4 kb)

Southern blot analysis

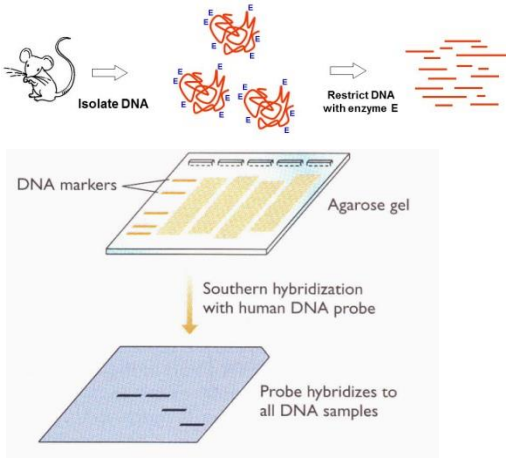
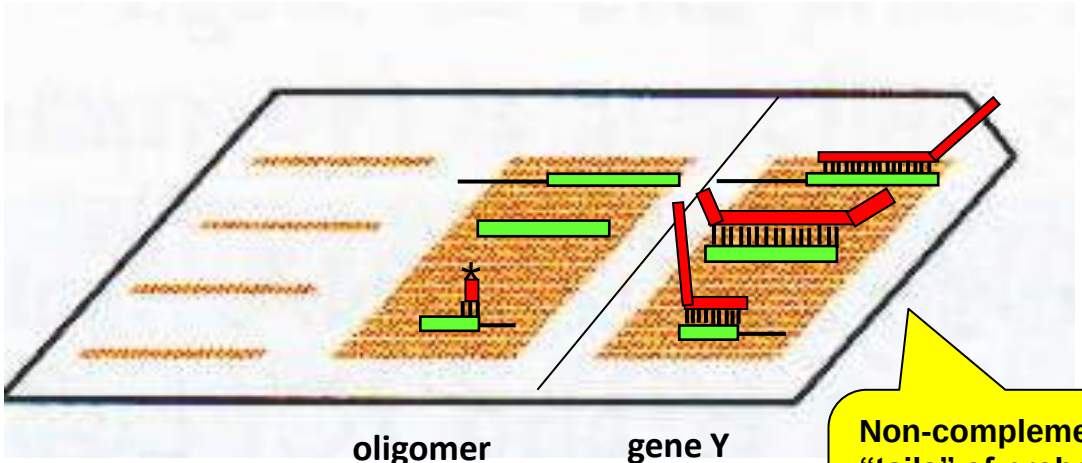


Fig.2.11 (Topic 2)



oligomer probe

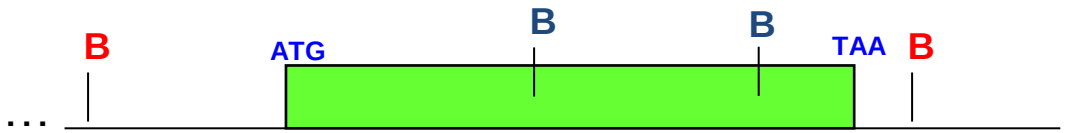
gene Y probe

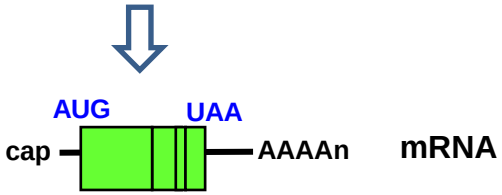
Non-complementary "tails" of probe will not destabilize hybrid, if hybrid is > 50 bp or so (Topic2 , slide 8)

Practice set #1, question 7b

Protein Y = 300 aa, so coding sequence: 300 x 3 = 900 nt

and intron(s) make up 8.1 kb of gene





Length of mRNA Y =

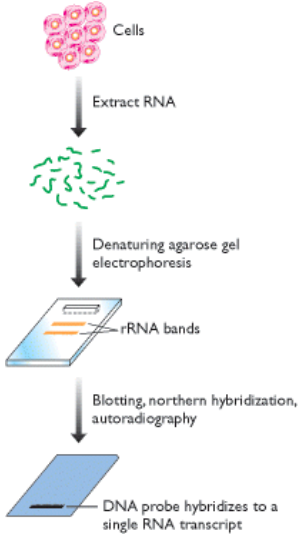
300 x 3 = 900 nt

+ UTRs (~ 100 nt each)

+ polyA tail (~ 200 nt)

= 1.3 kb (or so)

Northern blot analysis



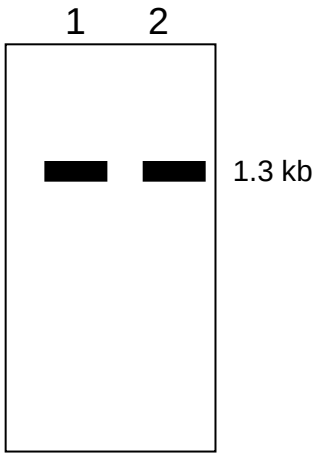
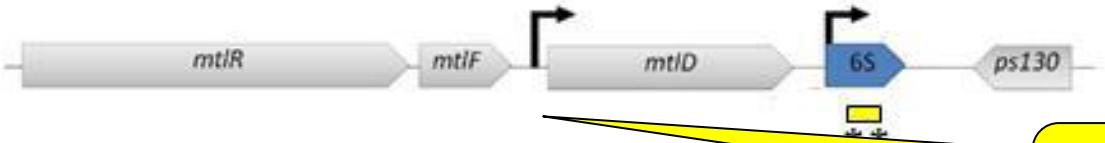


Fig.5.11 (Topic 2)

Practice set #1, question 8



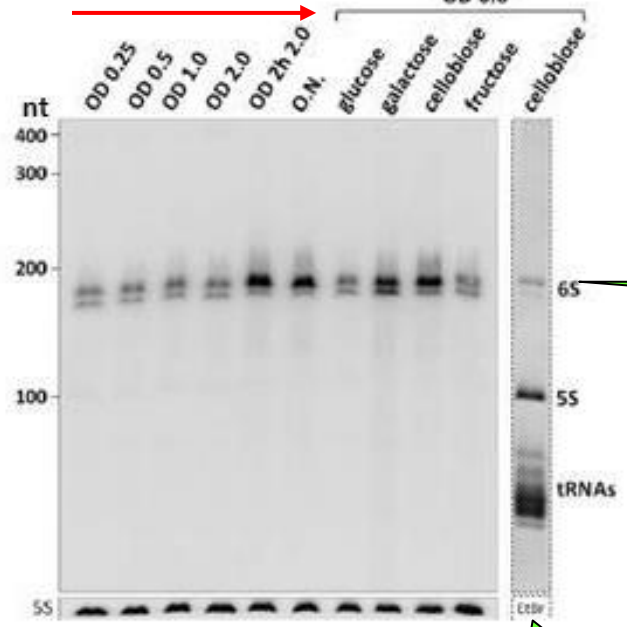
Based on names looks like operon for "mtl", but separate promoter for mtlD

Relative amounts of the two RNAs changed over time (more of larger one)

Grew on different nutrients (carbon source)

Longer growth time

Saw differences in stochiometry of the two ncRNAs



Northerns

Observation: two small RNAs (~ 180 nt & 170 nt) hybridized with "6S" probe

Under these conditions ,180 nt RNA species is so abundant that could it detect by Et Br staining

Northern blot analysis

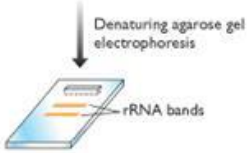


Fig.5.11 (Topic 2)

5S rRNA probe – "control" showing that same amount of RNA was loaded in each lane

Ethidium bromide stained lane

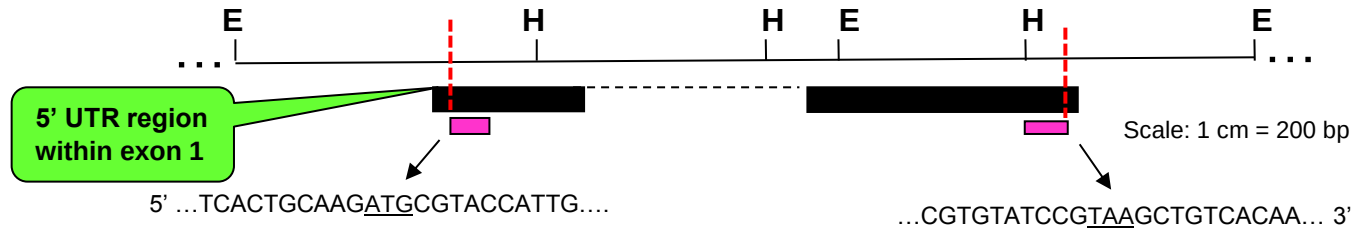
"novel RNAs including a small regulatory RNA involved in carbon uptake and metabolism"

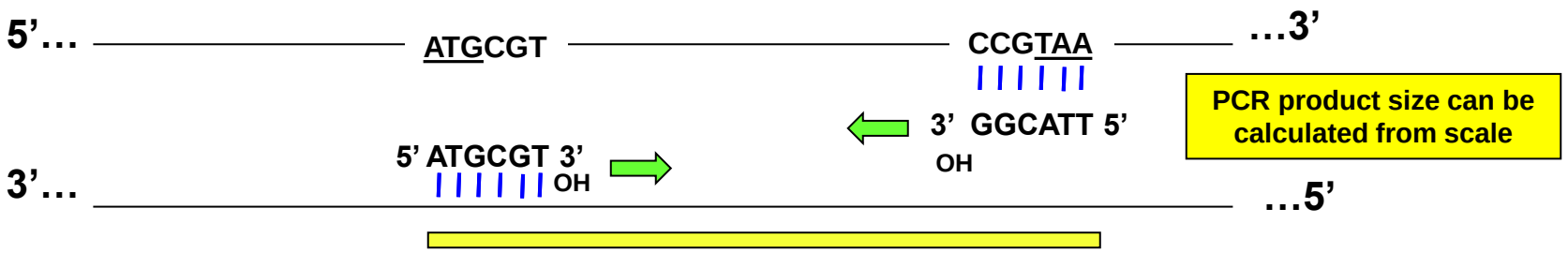
van der Meulen *RNA Biol.* 13: 353, 2016

Practice set #1, question 9

Exons = black bars, intron = dotted line

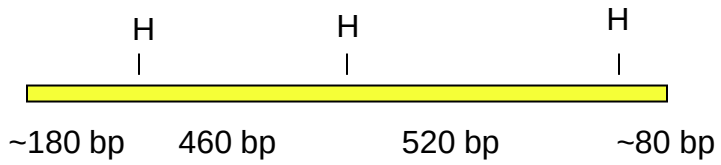
first exon contains 5' UTR…

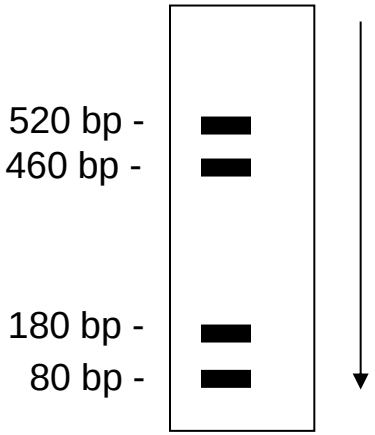




Aside: in "real life", we might choose oligomers slightly upstream & downstream of start/stop codons to ensure that the full coding sequence reflects the template sequence. Especially if primers were designed from a different organism (in case primers have a mis-match with template…).

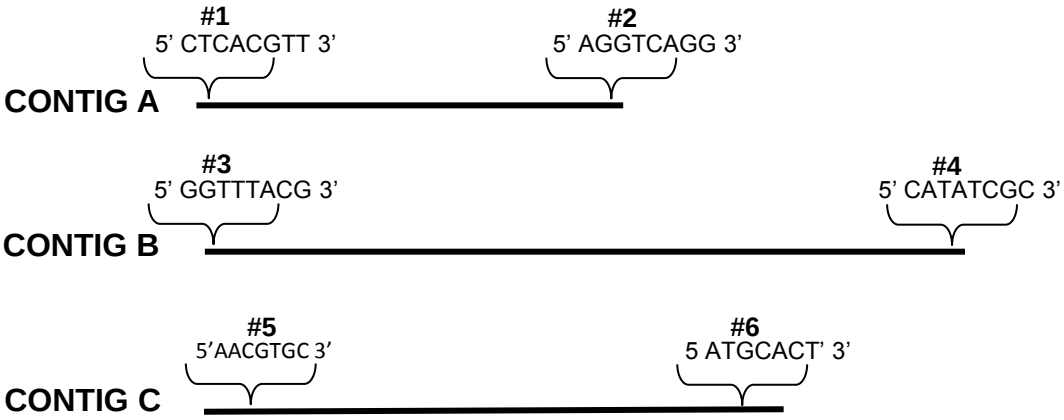
(ii) If PCR product is restricted with HindIII



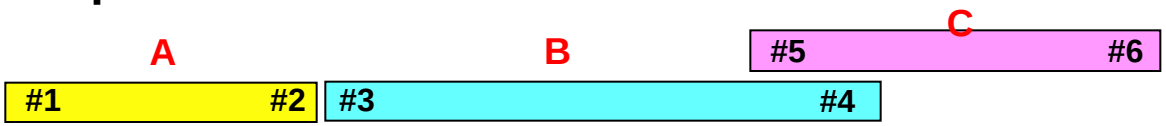


Aside: Your actual numerical values will depend on the magnification of your copy of the figure.

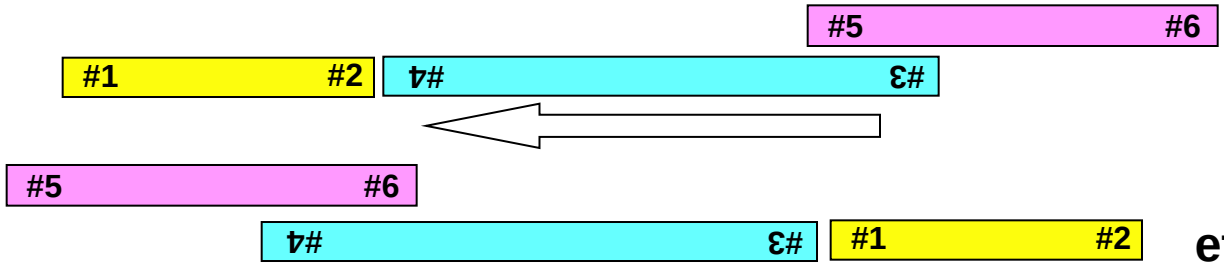
Practice set #1, question 10



For example:

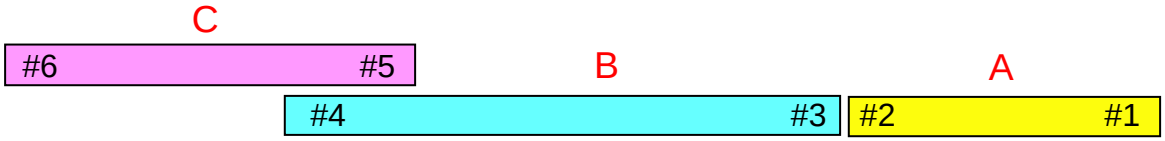


But other possible arrangements too…

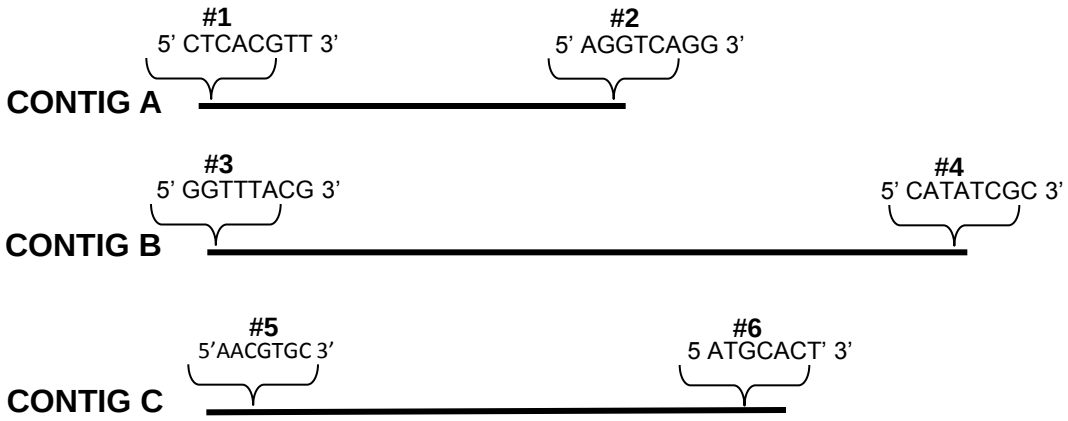


etc.

(It's important to be aware of this, but on a test you wouldn't have time to work through them all, so would just choose one as an example…)

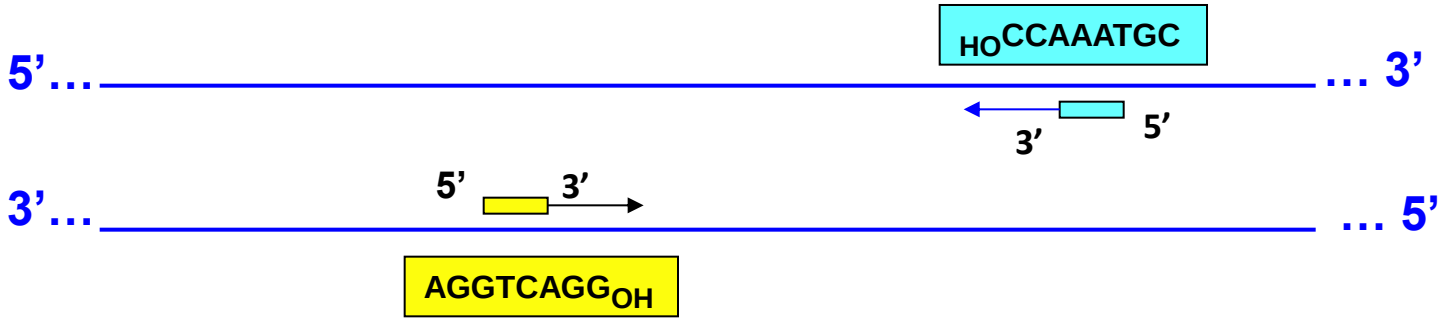


NB: Mirror image (whole thing flipped over) is <u>not</u> a different arrangement

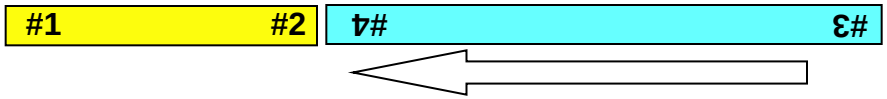


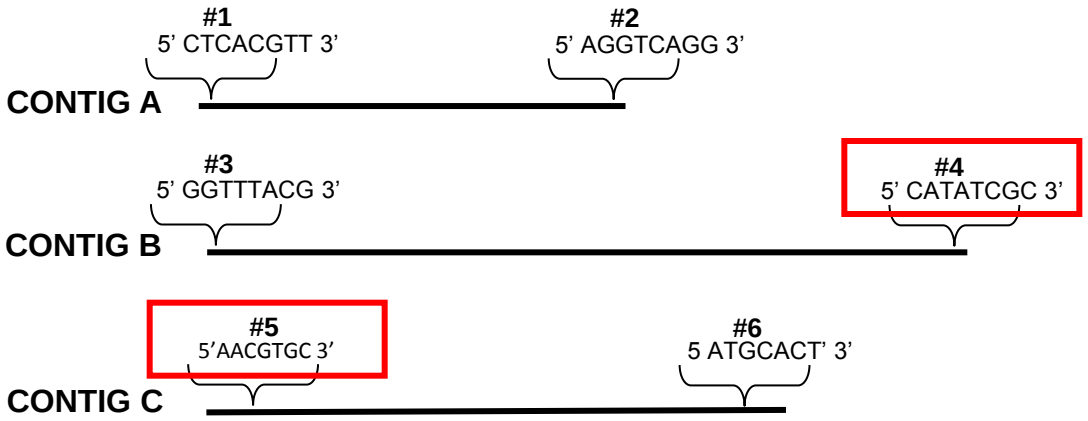
Contig A adjacent to B, so no overlap…

PCR approach "to close the gap" using BAC clone X as template DNA

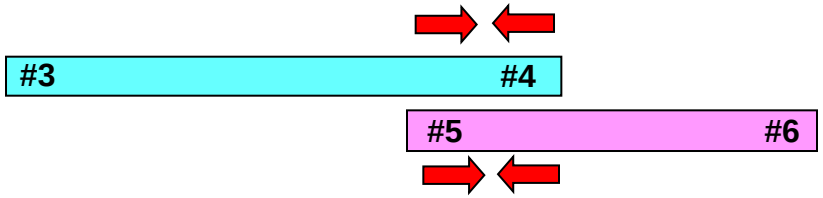


If this is correct arrangement, primer pairs 2 & 4 or 1 & 3 etc. will not give PCR products…



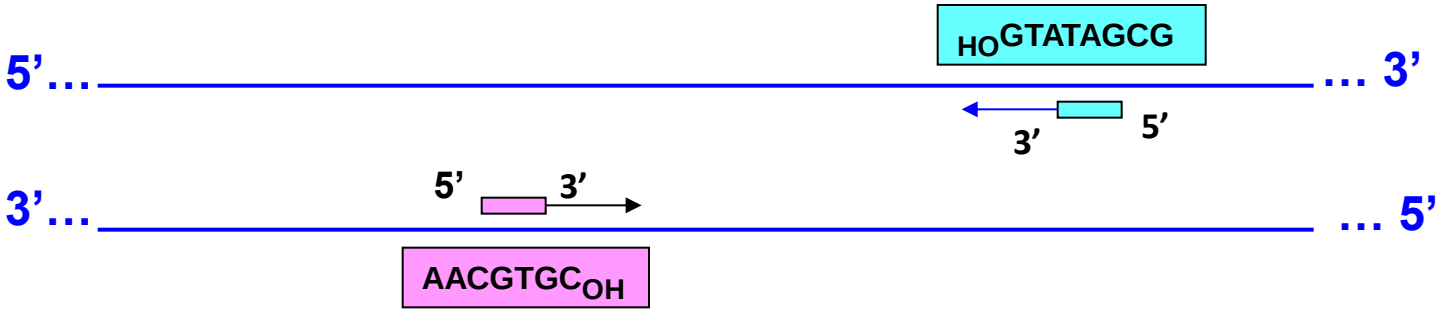


Contig B overlaps C… can use "STS-specific PCR" strategy

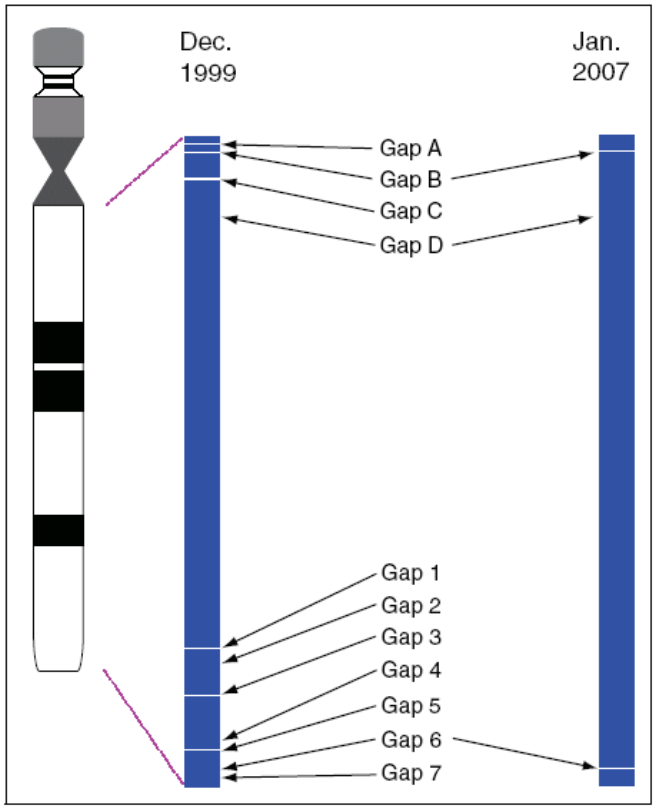


This primer pair will give PCR product using clone mapping near the end of contig B as template

… and also for clone from near end of contig C as template



Practice set #1, question 11a



Question deals with closing "physical gaps"

1. Generate new clone library using <u>different vector</u>

then chromosome walking

using hybridization or PCR or fingerprinting… approaches

(vs. shotgun sequencing of clones from new library since missing only small regions)

2. To close <u>very</u> small gaps, maybe use PCR strategy (uncloned DNA as template & design primers from contig info on each side of gap)

Results: We have used a combination of conventional chromosome walking (aided by the availability of end sequences) in fosmid and bacterial artificial chromosome (BAC) libraries, whole chromosome shotgun sequencing, comparative genome analysis and long PCR to finish 8 of the 11 gaps in the initial chromosome 22 sequence.

Cole et al. *Genome Biol* 9:R78, 2008

Practice set #1, question 11b

In 1999, 545 genes (protein-coding) identified… whereas September 2016, NCBI website says chromosome 22 has … 1129 genes

- gaps short & likely repetitive DNA (difficult to clone/sequence), so probably not very many new genes expected by closing gaps

- maybe some "genes-within-genes" (within introns)

- maybe some genes missed (if very short exons & long introns & no RNA info)

- maybe some very short protein-coding genes missed (if below default length in computer searches)

- probably found new regulatory RNA genes