BIOSCIENCES 741: GENOMICS
FALL, 2013

MIDTERM EXAMINATION

Instructions: This is a closed book exam, so the use of cell phones in this room, during the examination (verbal, texting, online web browsing, etc) is not acceptable. If you did bring a cell phone with you, it can be turned off now and stored inside a closed backpack on the floor, or alternatively turned off and stored on the bench at the front of the room. After the exam starts, if you still have a cell phone (that is turned on) in your possession (hands, lap, desk, etc), that will be considered evidence of cheating and will result in a grade of zero.

I understand and will follow these rules:

Signed: _____*answer key*_____

Instructions: Answer the following five questions on the paper provided (front and back if necessary). You will not need a calculator.

Suggestions: Your answers are likely to receive more credit if I can understand what you are saying -- in other words, if your answers are well organized and legible. You have plenty of time, so I suggest that you take some time to outline each answer before starting to write, in order to minimize crossouts and unclear reasoning.

1. Discuss the technical difficulties involved in identifying a complete inventory of the following categories of genes. Each answer should include at least two specific, reasonably likely problems, based on current methods of genome sequencing and gene identification. Your answer should also include an outline of a reasonably practical solution to these problems.

(A) Species-specific genes.

Genes that occur in one species but not in related species may be inadvertantly omitted from assembled sequence contigs, if related species are used to guide genome sequence assembly (contig alignment). Also, a species-specific gene may not be recognized by exon prediction algorithms, if new genes do not pass the criteria for "conserved sequences," which in turn guide exon prediction algorithms. If a species-specific gene originated recently by gene duplication, then its EST (or RNA-seq) data may be grouped with a different gene in the gene family. SOLUTIONS – these problems can be corrected by thorough analysis of EST sequences, particularly more ESTs from subtracted or normalized cDNA libraries, and/or RNA-seq.

(B) Tissue/stage specific genes

Tissue/stage specific transcripts will not be represented in cDNA libraries or EST sequencing projects unless that specific cell type, and/or stage of the life cycle, has been used to build a cDNA library. In some cases, tissue-specific genes are expressed in special cell types that are quite rare, which will make their gene transcripts even less common (and hard to find) in samples of the larger tissue. In other cases, stage-specific genes may be expressed only in very brief stages of the life cycle – 8 cell embryo, for example, but not later embryos. In this case, one sample of each major life stage ("embryo") may not be sufficient, and it may be difficult to predict how many stages need to be sampled. SOLUTIONS – one approach is to sample more tissues, and more stages of the life cycle. A complementary approach is to look for conserved genomic sequences/ORFs, which can allow gene identification even before that mRNA has been found.

(C) Regulatory genes.

In many cases, regulatory genes are expressed at very low levels (on the order of one mRNA molecule per cell) because the cell only needs small amounts of the protein. Some such rare transcripts may not have been detected in EST sequencing projects. A more serious problem arises from the fact that some regulatory genes may themselves be regulated at the transcriptional level, in which case these genes may only be transcribed under specific physiological conditions or brief stages of the life cycle. If so, these transcripts would not be detected unless the right conditions were tested. SOLUTIONS – one approach is to use normalized or subtracted cDNA libraries or RNA-seq, sequence larger numbers of ESTs, and sample a range of life stages and physiological conditions. A complementary approach is to look for conserved genomic sequences, which allows gene identification even before the mRNA has been detected.

(D) Protein-coding genes with poly(A)- transcripts.

Most protein coding genes have poly(A) tails, however some do not (i.e., histone mRNAs). In most cases, oligo(dT) affinity columns are used to purify poly(A)+ mRNA before cDNA synthesis, and/or oligo(dT) oligonucleotide primers are used to prime first-strand cDNA synthesis by reverse transcriptase. These methods of preparing cDNA libraries will not include poly(A)- transcripts such as histone gene transcripts. SOLUTIONS - prepare additional cDNA libraries by other methods (such as random priming), and/or use RNA-seq, so that poly(A)- mRNAs are sampled also.

2. Discuss the major types of Metazoan promoters, and the functional and structural (DNA sequence, chromatin modifications) characteristics of each.

The major types of Metazoan promoters are Type I ('adult), Type II ('ubiquitous'), and Type III ('developmentally regulated'). Although there are additional categories, and although some promoters may be difficult to classify, the majority do fit within these major types, which are discussed in more detail below.

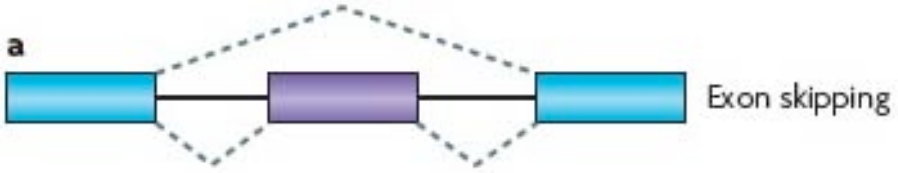
Type I or 'adult' promoters show tissue-specific expression in adult somatic tissues. In other words, these genes are expressed during the adult stage of the life cycle, but are typically not expressed in embryos, and in adults they are expressed in some specific tissues but not others. Type I genes typically have a "sharp" or "focused" transcription start site – in other words, the RNA transcripts start at a single base in the genomic sequence, and that base is located at a fixed distance from the TATA box. Type I promoters generally do have a TATA box, which extends from about 24-30 bp upstream of the RNA start site, but do not have CpG islands. Nucleosomes at Type I promoters are typically disordered (i.e., not phased, no fixed nucleosome locations).

Type II or 'ubiquitous' promoters are characterized by broad expression throughout the life cycle. In other words, these genes are expressed in adults but also in embryos, and even in some cells in the germ cell lineage. They are characterized by a "broad" or "dispersed" transcription start sites. "Broad" promoter does not necessarily mean multiple promoters, rather it means that transcripts from a single promoter have RNA start sites over a range of bases. These genes typically do not have a TATA box, but do have a short CpG island that overlaps the transcription start site. Nucleosome phasing around the transcription start site is typically ordered, presumably due to exclusion of nucleosomes that overlap protein binding sites in the promoter region.

Type III or 'developmentally regulated' promoters are often found in genes that are developmentally regulated (and regulators) during multicellular development and differentiation. In other words, Type III genes are expressed at some points in the life cycle but not others, and they may or may not be expressed in adults. These genes are typically subject to repression by the Polycomb complex, which means that a Type III promoter can be temporarily inactivated (the 'poised' state) when the chromatin in the vicinity of the transcription start site is subjected to histone trimethylation by the Polycomb Complex (which produces a unique chromatin mark, H3K27me3). The transcription start site of a Type III promoter typically does not have a TATA box, but is associated an Inr sequence that overlaps the transcription start site. Without a TATA box, RNA start sites are consequently 'broad' (distributed over a range of bases). Type III promoters contain large CpG islands, which extend into the 'body' (protein coding portion) of the gene. Nucleosome locations around the transcription start site are typically ordered, presumably due to the exclusion of nucleosomes that overlap protein binding sites in the promoter region. Consequently, the nucleosome structure becomes disordered (variable nucleosome locations) after the gene is completely turned off.

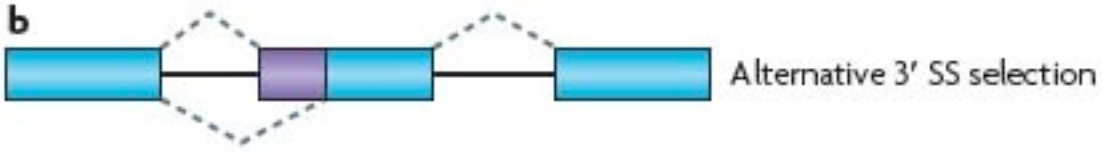
3. In the spaces below, diagram and explain the seven main types of alternative splicing, including an outline of the mechanism(s) by which each type is regulated. Also explain the biological significance of each type, in terms of the specific change(s) in gene function that are usually caused by each splice variant?

(A)



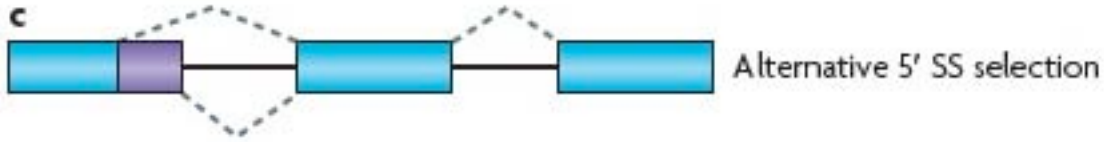
The first (left) splicing event involves the same 5' splice donor in both cases, but different 3' splice acceptors, therefore the 3' splice acceptor is being regulated here. The inclusion/exclusion of an exon will add/subtract a protein module, which may result in adding/subtracting a protein binding or regulatory site.

(B)



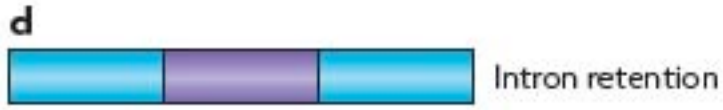
The first (left) splicing event involves the same 5' splice donor in both cases, but different 3' splice acceptors, therefore the 3' splice acceptor is being regulated here. The inclusion/exclusion of the first portion of the exon will add/subtract a protein module, which may result in adding/subtracting a protein binding or regulatory site.

(C)



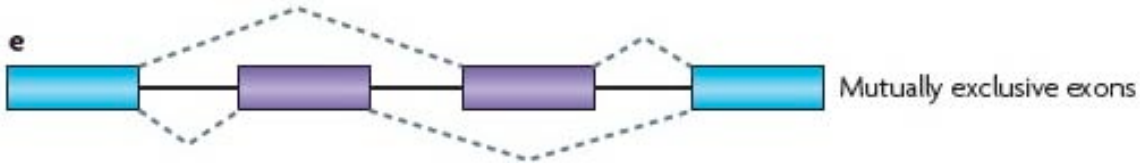
The first (left) splicing event involves different 5' splice donors, but the same 3' splice acceptor, therefore the 5' splice donor is being regulated here. The inclusion/exclusion of the last portion of the exon will add/subtract a protein module, which may result in adding/subtracting a protein binding or regulatory site.

(D)



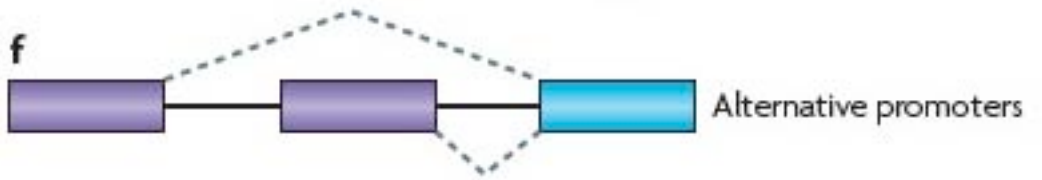
In this case, the intron is either spliced or not, so the 5' splice donor and 3' splice acceptor are either both active or both inactive. However, in most cases only one of them is regulated. If the 5' splice donor were active, it would be paired with another (downstream) splice acceptor and a larger intron spliced out. Therefore, the intron is probably retained due to regulation of the 5' splice donor. Retention of an intron usually results in protein termination (in-frame stop codon) and hence is used to turn genes on and off.

(E)



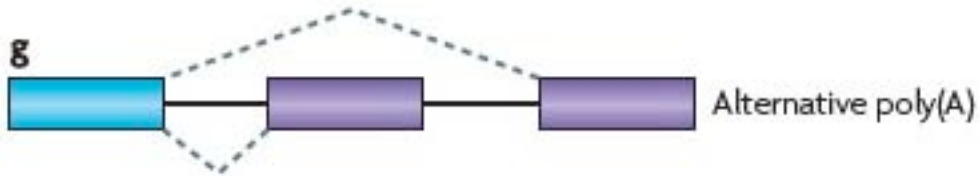
The first (left) splicing event involves the same 5' splice donor in both cases, but different 3' splice acceptors, therefore clearly the 3' splice acceptor is being regulated in that case. A second question is why the 5' splice donor at the end of the second exon always skips the third exon – this could occur by steric hindrance (the third exon may be too close to allow the assembly of a splicing complex with the second exon) or by regulation of the splice 3' acceptors. This spicing pattern – of mutually exclusive exons - is typically used to switch the binding specificities of proteins, for example changing the DNA targets of transcription factors, or the protein partners of cell surface binding proteins.

(F)



In this case both splicing events involve the same 3' splice acceptor, which is therefore not being regulated. On the other hand, whichever 5' splice donor is transcribed is used. In other words, RNA splicing is not being regulated here, rather the regulation is at the transcriptional level (activation of different transcription factors activates different promoters of the same gene). The biological consequences of alternate promoter use may simply allow expression of the gene in different tissues, or different conditions. The alternate exons may also encode N-terminal regulatory domains of the protein, for example that bind to alternate protein partners.

(G)



In this case, the 3' splice acceptor is being regulated. This may result in an alternative C-terminal protein domain, and/or an alternative 3' untranslated region. The 3' untranslated region of the mRNA contains (in some genes) binding sites for mRNA transport/localization, and binding sites for regulation by miRNAs.

4. The following question concerns the assembly of physical genome maps.

(A) Define your terms – what is a BAC clone? How does it differ from a P1 clone? What is FISH? How is it performed?

A BAC clone is a bacterial artificial chromosome. BACs are circular recombinant DNA molecules that include genes derived from the single copy F plasmid that enable DNA synthesis in E. coli and orderly (single copy) chromosome segregation, as well as a selectable marker (such as antibiotic resistance). BACs can accommodate inserts as large as 300 kb. A P1 clone is a circular recombinant DNA molecule that includes a few genes derived from the P1 phage (for phage lytic induction, the initiation of phage packaging, and sites for recombination), as well as selectable markers and single copy F plasmid genes. Because of the requirement for packaging into P1 phage heads, P1 clones can only accommodate inserts up to about 100 kb.

FISH is fluorescent in situ hybridization. "In situ" refers to hybridization to biological tissue ("in site") such as mitotic chromosomes. FISH is performed by fluorescently labeling BAC or P1 clones (by random-primed DNA synthesis), followed by hybridization to mitotic chromosomes that are fixed to glass slides, and fluorescence microscopy to localize the chromosome bands that are fluorescently labeled by the probe.

(B) Briefly describe how BAC (and P1) fingerprinting is performed.

BAC (and P1) fingerprinting is performed by single- and double-restriction digests with restriction enzymes that have six-base recognition sites (such as BamHI, EcoRI, PstI, etc), followed by high performance (agarose and/or acrylamide) electrophoresis with detailed length standards (run in adjacent lanes), digital photography, and the computational assignment of each band to a specific "length bin". Sequence overlaps between BAC and/or P1 clones are then recognized and scored computationally, by matches between their restriction fragment lengths.

(C) Briefly describe how BAC (and P1) fingerprinting can be used together with FISH to build physical genome maps.

BAC (and P1) fingerprinting allows the automated alignment of overlapping clones, before they have been sequenced, into megabase-length contigs. The chromosomal location (and orientation) of these contigs, with respect to chromosomal landmarks such as centromeres, telomeres, and major chromosome bands, is then determined by FISH of clones from each end of the contig. The FISH results are then used to guide further rounds of BAC fingerprinting to close the gaps between adjacent contigs.

(D) How do physical maps differ from cytological or recombinational maps? Which is most relevant to genome projects?

Physical maps are based on molecular experiments (such as restriction mapping and DNA sequencing) and measure genetic distances in quantitative units of kilobases or megabases. Cytological maps are based on microscopy of stained chromosomes, and measure genetic distances in qualitative units of chromosome arms and chromosome bands. Recombinational maps are based on genetic breeding experiments and measure genetic distance in quantitative units of centiMorgans (percent recombination frequency). All of these maps are relevant to genome projects, but the physical maps are arguably the most relevant, because physical maps measure the amount of DNA to be sequenced between any two points on the genome map.

5. The following question concerns studies of single nucleotide polymorphisms (SNPs), haplotypes, and linkage disequilibrium in the human genome.

(A) Define your terms – what is a SNP? What is linkage disequilibrium? What is a haplotype?

A single nucleotide polymorphism (SNP) is a single base location in the genome at which two alternative bases are present in the population. Linkage disequilibrium refers to a non-random association between the specific alleles (bases) of two closely-linked SNPs on the same chromosome. A haplotype is a chromosomal block of these specific alleles (bases) that tend to occur together on the same chromosome.

(B) How is linkage disequilibrium and advantage, and how is it also a disadvantage, in the search for functional SNPs that are associated with human diseases?

Linkage disequilibrium can be an advantage in the initial search for SNPs associated with human diseases, because LD facilitates the initial discovery phase. That is, many SNPs in a haplotype will show significant statistical association with the disease, and this increases the odds that one of the SNPs chosen for testing will show an association.

Conversely, LD can be a disadvantage in the latter efforts to analyze the relative functional contributions of various SNPs within a haplotype to the disease. This is so because the non-random associations of SNPs within a haplotype may tend to obscure which one(s) are most directly responsible for the statistical risk(s) of the disease.

(C) Do haplotypes tend to be larger in African populations or European populations? Why? [Hint: your answer should specify at least two reasons, as well as explaining why those reasons would affect the haplotype size.]

All other things being equal, haplotypes tend to larger in human populations that originated recently, because over time recombination breaks haplotypes into smaller blocks, and a recent population has had less time for such recombination events to occur. Likewise, haplotypes tend to be larger in populations that were founded by a small number of individuals, because these populations have limited genetic diversity, hence also a limited number of haplotypes (but the same genome size) hence larger average haplotype length. Likewise, haplotypes tend to be larger in inbred populations (small isolated populations) because again these populations have limited genetic diversity, hence also a reduced number of haplotypes in the same size genome, hence larger haplotypes.

(D) Does it follow from your answer to part (C) above that the quantitative amount of linkage disequilibrium is also larger in some human populations than others? Why (or why not)?

Yes, linkage disequilibrium varies considerably between human populations, by definition, given that the number of haplotypes and their relative frequencies vary between human populations. This is due to differences in the age of human populations, as well as differences in founder effects, migration, and inbreeding, as explained above in part(C).