CS 291 - PERSONAL GENOMICS FOR BIOINFORMATICIANS

Problem Set 1 - Introduction to human genomes

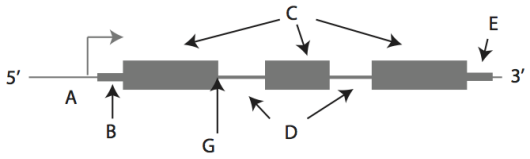
All homework should be sent to mgymrek@ucsd.edu with subject line [CSE291 PS1:LASTNAME] by the beginning of class on **Tuesday, January 10**.

Objectives

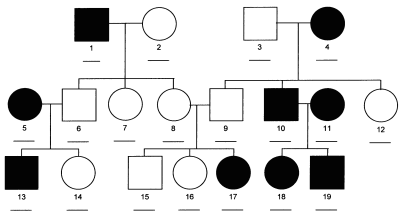
- Review human genetics concepts.
- Get comfortable with standard file formats and tools for analyzing genomes.

Exercises

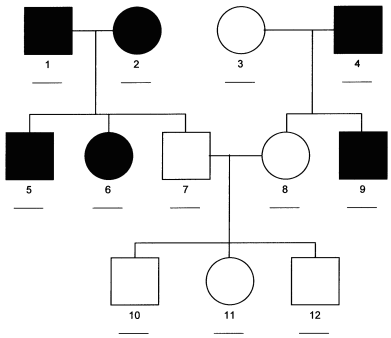
1. How many chromosomes does a typical human have? Remember to consider copies from both parents.
2. Approximately what percent of the genome consists of protein coding genes?
3. Identify the labeled components of a typical gene depicted below. Options are: exon, intron, 5'UTR, 3'UTR, splice site, promoter/regulatory region.



4. You see a gene with the following coding sequence: ATGCCGAATCGATCGTAA. Give (1) the sequence of the transcribed mRNA and (2) the amino acid sequence of the translated protein.
5. A mutation occurs in the gene from the previous question, resulting in a sequence ATGCCGAATCGATCTTAA. What is the amino acid sequence of the new protein? This is known as a *synonymous mutation*.
6. A different mutation occurs in the same gene, now resulting in a sequence ATGCCGAATCGACCGTAA. What is the amino acid sequence of the new protein? This is known as a *missense mutation*.
7. What modes of inheritance are consistent with the following pedigree? Name an example human condition that follows this mode of inheritance.



8. What modes of inheritance are consistent with the following pedigree? Name an example human condition that follows this mode of inheritance.



9. Download PS1 data from the course website. First, we'll get familiar with the VCF file format. For full details of VCF format, see here. Look at the file pset1_1000Genomes_chr16.vcf.

   - Lines at the top of the file marked with "##" indicate the header. What build of the human reference genome is used in this file?
   - Look at the first entry (16:31002227). What is the reference allele? What is the alternate allele? What is the rsid?
   - What is the HG00099's genotype at this position (CC, CA, or AA?).

10. Individual HG00097 comes to the hospital with a blood clot. The doctor would like to treat the patient with warfarin, but is trying to figure out the correct dose. Studies have shown that SNP rs8050894 is associated with warfarin dosage. Individuals with genotypes CC, CG, and GG are recommended to get 2.5, 5, and 6.5 mg/day respectively. We would like to determine the correct dose for our patient.

    VCF genomes can be quite large (the VCF file for the ExAC dataset is more than 1TB!). If we want to query a specific location of the genome, it would be inefficient to scan the entire file. tabix offers utilities to build an index for VCF (and other) files to enable quick access to a specific location. Use the tabix utility to index the VCF file with the following command:

    ```
    bgzip pset1_1000Genomes_chr16.vcf
    tabix -p vcf pset1_1000Genomes_chr16.vcf.gz
    ```

    Now, use the tabix utility to pull out our SNP of interest:

    ```
    tabix --print-header pset1_1000Genomes_chr16.vcf.gz 16:31104509-31104510
    ```

    What dose should the doctor give to HG00097? What about HG00099 and HG00103?

11. The doctor now wonders whether HG00097 has other variants in VKOR1C. We would like to determine whether the individual has any mutations in the protein-coding sequence of this gene. Luckily, we have a file VKORC1_exons_NM_001311311.bed listing the genomic coordinates of each exon. This is an example of a BED file, which gives the chromosome, start coordinate, and end coordinate of each interval.

    Intersecting genomic intervals is an extremely common task. bedtools contains many utilities for working with BED files. Use the intersectBed utility to look at all positions falling within the given intervals.

    ```
    intersectBed -a pset1_1000Genomes_chr16.vcf.gz -b VKORC1_exons_NM_001311311.bed
    ```

    Does HG00097 have any variants from the reference genome in exons of VKOR1C?