Harvard-MIT Division of Health Sciences and Technology
HST.508: Genomics and Computational Biology

Problem Set 3

1. 1. Population Genetics (22 pts)

I) I) **Mutations** (4 pts): define the following terms having to do with mutations in 15 words or less:

- ➢...... Deletion
- ➢...... Nucleotide transversion
- ➢...... Missense mutation
- ➢...... Frameshift mutation

II) II) **Mutagenesis** (2 pts): what is a mutagen (in ten words or less)? Name two types of mutagens and briefly (in 20 words or less) explain how they cause genetic mutations.

III) III) **Gene pool** (6 pts): a transversion in the second codon position for the sixth amino acid in the β-globin chain of hemoglobin is the recessive mutation responsible for sickle cell anemia. When the mutation is homozygous, it is lethal. However, people heterozygous for the sickle cell allele are protected from infection by the protozoan *Plasmodium falciparum*, which causes malaria.

a) a) Define the terms “allele fixation” and “heterozygote superiority” in less than 20 words. Relate them to this case—in areas where malaria is a threat, would either allele become fixed?

b) b) Let HbS^+ denote the dominant allele and HbS^S the sickle cell allele. If in a certain population we find the following genotypic breakdown:

Number of HbS^+/HbS^+ individuals = 3915
Number of HbS^+/HbS^S individuals = 585
Number of HbS^S/HbS^S individuals = 0

What are the genotypic and allelic frequencies of the population in question?

IV) IV) **Hardy-Weinberg** (10 pts): a 32 bp deletion in the gene coding for the human chemokine receptor CCR5, termed Δ32, is found to offer some AIDS resistance. The mutation of the membrane bound receptor hinders HIV infection of T cells. Here let’s denote the normal CCR5 allele as “A” and the Δ32 allele as “a.” Let’s say a 1000 person population was genotyped and we obtain 795 AA, 190 Aa, and 15 aa.

a) a) There are around half a dozens of assumptions upon which the Hardy-Weinberg principle depends in order to have predictive value. Name three of these.

b) b) Use the χ^2 test to determine whether the frequencies provided agree with those predicted by the Hardy-Weinberg law.

2. Monte Carlo Genetic Drift Simulation with Mathematica (10 pts)

Consider a population of sexually-reproducing organisms with an allele A that has an initial frequency of 50%. In the absence of any randomness or other factors, the frequency would remain constant in successive generations. Here, you'll use Mathematica to simulate what happens in a small population of organisms when gene inheritance is random.

In each successive generation, each individual offspring will inherit the A allele if Mathematica's Random[] function returns less than or equal to the frequency of the previous generation. Keep track of how the frequency changes from generation to generation, and output your results with the ListPlot[] function, graphing frequency versus time (in generations). Stop iterating when the frequency of A reaches 0 or 1.

Start with an initial population of 20 organisms. Run your experiment three times, and attach your output. Try modifying the program to use much larger numbers of organisms – what do you observe then?

Provide both your code as well as output with your answers.

3. Genome Sequencing (6 pts)

a) List the major steps involved in modern chain termination sequencing (2 pts)

b) Explain what is meant by shotgun sequence assembly (2 pts)

c) Is shotgun sequencing always the best choice for eukaryotic organisms? Why or why not? (1 pt)

d) As part of an epidemiological study group, you have been asked to sequence the genome of *Salmonella typhi,* the causative agent in typhoid fever. What method would you use, and why? (1 pt)

4. Sequence Analysis (32 pts)

Download the *S. typhi* genome from: ftp://ftp.sanger.ac.uk/pub/pathogens/st/St.dna

Write a perl program which identifies all possible open reading frames (ORFs). For the sake of this exercise, an ORF starts with an ATG (which is the furthest 5', or "left-most", if there are many). The ORF ends when hitting a stop codon (TAA, TGA, TAG). Design your program to search both strands. Include your well-annotated code at the end of your problem set. **Hint:** You may find code from the two previous problem sets useful for this task.

a) Finding small ORFs (8 pts): Most gene-finding programs do not predict genes that are shorter than 100 amino acids. These small, potential genes are not included in most collections of predicted genes. How many such ORFs did your program find that code for proteins between 75 and 100 amino acids (including 75 and 100)? How many were found on each strand of the genome?

b) Checking small ORFs (10 pts): Use BLASTp (http://crobar.med.harvard.edu or http://www2.ebi.ac.uk/blast2/) to compare the first 3 small ORFs (from the first reading frame) that you found to the Swissprot database. For each of these 3 ORFs, list its genome position, amino acid sequence, and the top BLAST match in each case? (Please use single letter annotation for the amino acids.) Are they statistically significant?

c) G+C content and ORFs (8 pts): Download the *E. coli* genome (if you haven't already) from http://www.courses.fas.harvard.edu/~bphys101/problemsets/Ecoli_K12.txt. What are the G+C contents of the *S. typhi* and the *E. coli* genomes? Run your program on the *E.coli* genome. How many small (75-100 aa) ORFs do you find per kb of genome sequence for each of these genomes? How does the average ORF size one expects to find at random change as a function of G+C content? Why?

d) Optimal oligo design (6 pts): You decide to design an oligonucleotide microarray based on the sequences of predicted ORFs greater than 100 amino acids. Name at least 3 criteria that you would use to select sequences to be used as probes on the array.

5. DNA Microarrays (30 pts)

S. cerevisiae, or baker's yeast, is commonly studied using microarrays. Yeast has a powerful genetic system and its 6,220+ ORFs are sequenced and well characterized, hence it is a great candidate for whole-genome profiling using DNA microarrays.

I) I) **Chip construction** (6 pts): after the amplification of DNA from specifically prepared libraries of ORFs, the amplified DNA is arrayed(printed) on slides to create microarray chips used in further experimentation.

a) a) Why are the slides usually coated with compounds like polylysine before DNA is printed onto them? (Hint: polylysine coating gives the slide an overall positive charge)

b) b) Woodie makes microarrays. A microarray containing the entire yeast genome has 6,220 spots of DNA. If the printed area of Woodie's slides are only 20mm by 20mm, and there needs to be around 100μm of space between the edges of each spot, then what's the expected average diameter of each spot?

II) II) **From RNA purification to Hybridization** (6 pts): the next step of the microarray experiment involves harvesting different populations of cells and purifying their RNA, which is then reverse-transcribed into cDNA. The cDNA probes from different populations are then purified and labeled (in various ways) with different fluorochromes (Cy3/Cy5). The probes are then applied to the slide for hybridization to occur.

a) a) While certain researchers extract total RNA from cells to study, others like to extract only messenger RNA. In 20 words or less, why might solely extracting messenger RNA be preferred?

b) b) Why are poly-thymine compounds of 12-17 bp often used to isolate mRNA? (Hint: think of a common feature at the 3' end of eukaryotic mRNA)

c) c) Why is it important to use RT(reverse transcriptase) to convert the RNA to cDNA?

III) III) **Data collection/analysis** (6 pts): after hybridization, the slide is run under an automated scanner which detects the fluorescent intensity of the two channels corresponding to cy3 and cy5.

a) a) Explain what excitation and emission wavelength of fluorochromes are in less than 20 words and find these wavelengths for cy3 and cy5. Why can't rhodamine (another fluorescent dye), with absorption frequency of 570 and emission wavelength of 590, be used as a substitute for Cy5 opposite Cy3?

b) b) Cy5 labeled RNA is purified from a population of *S. cerevisiae* grown a medium where galactose is the only sugar source. Cy3 labeled RNA is purified from the same strain growing in a medium where glucose is the only sugar source. Equal amounts are made into cDNA and competitively hybridized onto a microarray. What color would you expect spots corresponding to genes associated with the glucose metabolism pathway to appear on the computer?

IV) IV) **Error Analysis** (12 pts): because of the numerous steps involved and the high level of automation in microarray experiments, there are numerous possible sources of error, both random and systematic.

a) a) Fluorescence tends to stick non-specifically to the surface of the microarray slides around the DNA spots; this causes a "background" fluorescence that can confound the data. You measure the mean and standard deviation of this background from two arrays:

	Array 1	**Array 2**
Average background	800 units	1000 units
σ of background	40	30

In which case would 60 units above background be more significant?

b) b) Contrast random and systematic errors in 20 words or less.

c) c) For the following errors, indicate whether they are random or systematic and briefly **explain your rationale.**

➢ **➢······ Error in RNA purification.**
➢ **➢······ Cross-hybridization of probes.**
➢ **➢······ Uneven printing, scanning, or hybridization.**
➢ **➢······ Spatial(sector) bias by the scanner.**

d) d) What are *housekeeping genes*? Why would spotting housekeeping genes in every sector of the slide help normalize for spatial bias of the scanner?

e) e) Define cross-hybridization in relation to microarrays.

(2 bonus points) Does the completed sequence of the yeast genome make it possible to limit cross-hybridization at the microarray design step? How can this be done?