Kiley Huffman
QBIO 478
Spring 2025

Homework 1, Due February 13 (by midnight)

Exercise 1

1. (True or False) Given the counts of each type of nucleotide in a genome, to calculate the probability that the nucleotides could be generated by a process that is independent and identically distributed, complies with the formulation ”Numerically encoded input with a computable statistical test for significance”.

True. Given the counts of each type of nucleotide in a genome, you can compute their relative frequencies and determine if they follow an independent and identically distributed process. Then, you can use a statistical test for significance to see if the observed nucleotide frequencies deviate from the expected distribution, which complies with ”Numerically encoded input with a computable statistical test for significance”.

2. Provide the length of genomes of the following species and their ploidy with to- tal chromosome count (please refer to the RefSeq database): Homo Sapiens (human), Pan troglodytes (chimpanzee), Oryza sativa Japonica Group (Japanese rice), Canis lupus familiaris (dog), Sus scrofa (pig), Danio rerio (zebrafish), and Triticum aestivum (bread wheat). Note you can explore https://www.ncbi.nlm.nih.gov/gdv/ for genome statistics.

Species	**Genome Length (bp)**	**Ploidy**	**Total Chromosome Count**
Homo Sapiens (human)	3.2 billion	Diploid	46 (23 pairs)
Pan troglodytes (chimpanzee)	3.3 billion	Diploid	48 (24 pairs)
Oryza sativa Japonica Group (Japanese rice)	374 million	Diploid	24 (12 pairs)
Canis lupus familiaris (dog)	2.4 billion	Diploid	78 (39 pairs)
Sus scrofa (pig)	2.5 billion	Diploid	38 (19 pairs)
Danio rerio (zebrafish)	1.5 billion	Diploid	50 (25 pairs)
Triticum aestivum (bread wheat)	16 billion	Haploid	42 (21 pairs)

3. There are various types of repeat in the human genome. Provide any 3 types of repeats. Briefly describe the types of repeat and the fraction of the genome they comprise.

1. Short Interspersed Nuclear Elements (SINEs):

SINEs are a type of transposable element that are relatively short (<500 bp). They can amplify and insert themselves into new locations. These repeats make up about 13% of the human genome.

2. Long Interspersed Nuclear Elements (LINEs):

LINEs are another type of transposable elements, but they are much longer than SINEs. They can be as long as 7000 bp. They make up around 20% of the genome.

3. Tandem Repeats (Microsatelites and minisatellites):

Tandem repeats are short sequences of DNA repeated in tandem. Microsatelites consist of sequences of 1-6 bp repeated units, while minisatellites are longer repeats at around 10-60bp. These repeats account for about 3% of the human genome.

4. (True or False) Event A and B are independent if and only if P(A S B) = P(A) + P(B) − P(A|B)P(B).

False. The correct formula for $P(A \cup B)$ is:

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$ and two events are independent if

$P(A \cap B) = P(A)P(B)$

5. Name one of the genes in the human genome that lies in between chromosome 16: 1,000,000 - 1,300,000 (use human reference genome GRCh38/hg38). Provide its position and length. (use genome.ucsc.edu for searching)

One of the genes in the human genome that lies in between chromosome 16: 1,000,000 - 1,300,000 is METTL26. This gene is located from base pair 1,002,307 to to 1,004,228, and is 1922 bp long.

6. Define a structural variant. What sequencing technology is optimal to discover structural variants? You can reference https://www.nature.com/articles/s41576-020-0236-x

(posted online as Long Read Sequencing And Applications), or https://pmc.ncbi.nlm.nih.gov/articles/PMC10373632/, also posted on brightspace.

A structural variant is a genomic alteration involving segments of DNA typically larger than 1 kb in size. These include deletions, duplications, insertions, inversions, and translocation of DNA segments.

The most optimal sequencing technology to discover structural variants is long reading sequence technology because it can generate longer reads, allowing the identification of structural variants with greater accuracy compared to shorter sequencing technologies. The two primary long-reading sequence technologies are Pacific Biosciences (PacBio) Hifi Sequencing, which provides highly accurate long reads, and Oxford Nanopore Technologies (ONT) which can generate ultra long reads.

Exercise 2

1. Explain a primary source of error in Illumina sequencing.

A primary source of error in Illumina sequencing is phasing and prephasing errors. Phasing is when a nucleotide in a DNA strand fails to incorporate during a sequencing cycle. This causes a delay in the reading. Pre-phasing is when a nucleotide incorporates too early. This leads to an advanced signal. Both of these errors cause a loss of synchronization between DNA strands being sequenced, which can lead to substitution errors and decreased read accuracy.

2. Describe the essence of the clonal template generation step. More specifically, how does this step help the sequencing step?

The clonal template generation step ensures high signal accuracy during the sequencing step. This step amplifies individual DNA fragments to create clusters of identical DNA molecules, which are used as templates for sequencing. This step boosts signal intensity, increases read accuracy, and enables parallel processing.

3. Explain why single-molecule sequencing technologies do not have the same read- length limitations as Illumina.

While Illumina sequencing uses clonal amplification and sequencing-by-synthesis techniques, single-molecule sequencing techniques sequence individual DNA molecules without the need for amplification. Thus, single-molecule sequencing technologies do not experience the read-length limitations Illumina sequencing does. These single-molecule sequencing techniques eliminate the need for amplification, synthesis, and cycle-based imaging, allowing them to sequence longer

DNA fragments (10-100s kb). In contrast, Illumina sequencing can only sequence a few hundred base pairs.

4. There are multiple new companies developing short-read sequencing technologies that will compete with Illumina. Identify them from this article https://frontlinegenomics.com/the-latest-developments-in-sequencing-technologie (also on brightspace), and give the latest statistics for (a) instrument throughput, (b) sequence length, and (c) accuracy.

The new companies developing short short-read sequencing technologies that will compete with Illumina are Ultima Genomics and Element Biosciences.

Company	**Instrument Throughput (a)**	**Sequence Length (b)**	**Accuracy ©**
Ultima Genomics	n/a	n/a	n/a
Element Biosciences	800 million reads per run, outputs from 100-800 Gb	Supports read of 2x150bp and 2x300bp	Exceeds 99%

Note: The article states that Ultima Genomics are operating "relatively in stealth," thus I was unable to find the statistics for them (they are not yet publicly available).

5. Find the latest published estimates of (a) sequence throughput and (b) read accuracy for PacBio HiFi and Oxford Nanopore. Cite your sources.

Company	**Sequence Throughput (a)**	**Read Accuracy (b)**
Pac Bio HiFi	Up to 25 kb in length (PacBio 2024)	99.9% (PacBio 2024)
Oxford Nanopore	10-20 Gb per MinION flow cell; up to 200 Gb per PromethION flow cell (Center for Genetic Medicine)	99% accuracy (Center for Genetic Medicine)

Works Cited:

HiFi reads - highly accurate long-read sequencing. PacBio. (2024, December 2). https://www.pacb.com/technology/hifi-sequencing/

Nanopore long-reads sequencing. Center for Genetic Medicine: Feinberg School of Medicine. (n.d.). https://www.cgm.northwestern.edu/cores/nuseq/services/next-generation-sequencing/nanopore-sequencing.html#:~:text=Oxford%20Nanopore%20currently%20offers%20three,Gb%20(170%20Gb%20maximum).

Exercise 3

The amino acid isoleucine is encoded by ATT, ATC, and ATA. For a genome with base composition defined by p_A,p_C,p_G, and p_T.

1. Give an expression that is the probability of generating a triplet that is an isoleucine.

P(isoleucine) = P(ATT) + P(ATC) + P(ATA)

P(isoleucine) = $(p_A p_T p_T) + (p_A p_T p_C) + (p_A p_T p_A)$

2. For a tiny 14-base genome, what is the maximum number of triplets matching an isoleucine encoding that may exist (note: this is a trick question).

14 / 3 = 4 R2

Since only full, non-overlapping codons are counted, the maximum number of triplets matching an isoleucine encoding that may exist is 4.

Exercise 4

In a DNA sequencing experiment, a machine reads a sequence of nucleotides (A, C, G, T) from a given DNA sample. However, the machine has a small error rate, and each nucleotide may be read incorrectly with a small probability. A common way to cope with sequencing errors is to sequence a base multiple times. Suppose a sequencing technology is claimed to have an error rate of 0.01 on A/T bases and 0.02 on C/G bases.

1. For a base sequence with an error rate of 0.01, which symbol should you use to en- code the phred quality score of this base? (You can refer to information on this website for the encoding rule https://www.drive5.com/usearch/manual/quality_score.html)

You should use the following symbol to encode the phred quality score of this base: ‘5’

2. Suppose the sequencing technology is used to sequence a base A independently by 4 times (i.e., 4 readouts). What is the probability that the base is sequenced correctly for all 4 readouts?

The probability of base A being sequenced correctly in a single read:

P(correct) = 1 - P(error) = 1 - 0.01 = 0.99

The probability of sequencing A correctly is:

P(correct) = 1 - 0.01 = 0.99

The probability of sequencing A correctly four times is:

P(correct 4 times) = 0.99^4 = 0.9606

The probability that the base is sequenced correctly for all 4 readouts is 0.9606.

3. Following (b), what is the probability that at least one of the 4 readouts of this A base is correct?

P(at least one correct) = 1 - P(all incorrect) = 1 - $(0.01)^4$ = 0.00000001

P(at least one correct) = 1 - P(all incorrect) = 1 - $(0.01)^4$ = 1-0.00000001 = 0.99999999

The probability that at least one of the 4 readouts of this A base is correct is 99.99%.

4. A genome of size 10, 000 bases and 60% GC content is sequenced using this sequencing technology. Suppose each base of the genome is sequenced independently by exactly once. What is the expected number of correctly sequenced bases?

n = 10 000 bases, error rate of 0.01 on A/T bases and 0.02 on C/G bases

10000 * 0.6 = 6000 GC Content

10000-6000 = 4000 AT Content

P(correct for AT) = 1 - 0.01 = 0.99

P(correct for GC) = 1 - 0.02 = 0.98

E(correct AT bases) = 10000*0.4*0.99 = 3960

E(correct GC bases) = 10000*0.6*0.98 = 5880

E(total correct bases) = 3960 + 5880 = 9840

The expected number of correct bases is 9840.

Exercise 5

Say one million separate DNA sequences are randomly generated, each of length 6 using the following approach: let X be the random variable that can take on A, C, G, or T with probabilities $p_A = .31, p_C = .20$, $p_G = .17$, and $p_T = .32$. Let w be a random work of length 6 that is the result of 6 samples of X that are joined together.

1. What is the expected number of times the sequence ACCGTA appears among all 1M sequences?

P(ACCGTA at a specific position) = $(p_A p_C p_C p_G p_T p_A)$

P(ACCGTA at a specific position) = 0.31*0.20*0.20*0.17*0.32*0.31 = 0.00021

Total positions possible = 1,000,000 - 6 + 1 = 999,995 positions

E(ACCGTA occurrences) = 999995*0.00021 = 209.11

The expected number of times the sequence ACCGTA appears among 1M sequences is 209.

2. Using the previous definition of purines R = A, G and pyrimidines Y = C, T, what is the expected total number of appearances of the set of sequences that confirm to RCYGTR (for example, GCCGTA is one such appearance)?

P(R) = $p_A + p_G$ = 0.31 + 0.17 = 0.48, P(Y) = $p_C + p_T$ = 0.20 + 0.32 = 0.52

P(RCYGTR) = $p_R p_C p_Y p_G p_T p_R$ = 0.48*0.20*0.52*0.17*0.32*0.48

E(RCYGTR) = 999995* 0.48*0.20*0.52*0.17*0.32*0.48 = 1303

The expected number of appearances of the sequences that confirm to RCYGTR is 1303.

3. What would be the expected number of times the sequence ACCGTA or RCYGTR occurs?

Since ACCGTA confirms to RCYGTR, the expected number of times the sequence ACCGTA occurs is included in the expected number of times the sequence RCYGTR occurs. Thus, the expected number of times the sequence ACCGTA or RCYGTR occurs is the same as the expected number of times the sequence RCYGTR occurs: 1303 occurrences.

P(RCYGTR and ACCGTA) = P(RCYGTR) = 1303

4. Now assume the nucleotide probabilities were updated to P (A) = P (C) = P (G) = P (T) = 0.25. First, count the number of different six-nucleotide sequences in RCYGTR. Then, calculate the expected number of appearances of the RCYGTR in the updated genome simu- lation.

Number of different six nucleotide sequences = 2*1*2*1*1*2 = 8

P(RCYGTR at specific position) = 0.5*0.25*0.5*0.25*0.25*0.5

Total positions = 999995

E(RCYGTR occurrences) = 8*999995*(0.5*0.25*0.5*0.25*0.25*0.5) = 31249

The expected number of appearances of RCYGTR is 31249.

5. Write three different statistical tests to evaluate the probability of observing the sequence ACCGTA exists n times in the genome of size 1M.

#1 Binomial Test: tests number of successes in fixed number of independent trials

Null Hypothesis (h0): the sequence ACCGTA occurs exactly n times in the genome

Alternate Hypothesis (h1): the sequence ACCGTA occurs a different number of times than n

p = P(ACCGTA at a specific position), the probability of observing the sequence at a single position

n = 999995, the total number of possible positions

Then, the binomial distribution gives the probability of observing k successes in n trials:

$P(k) = (n\ choose\ k)p^{k}(1 - p)^{N-k}$

#2 Poisson Test:

Null hypothesis (h0): the number of occurrences of ACCGTA follows a Poisson distribution with a given rate

Alternative Hypothesis (h1): the number of occurrences does not follow the Poisson distribution

Expected number of occurrences: $\lambda = N \times P(ACCGTA\ at\ a\ given\ position)$

Then, the probability mass function of the Poisson distribution is:

$P(k)\ =\ \frac{\lambda^k e^{-\lambda}}{k!}$ where λ is the expected number of occurrences and k is the observed number of occurrences.

<u>#3 Chi-Square Test for Goodness of Fit:</u>

Null hypothesis (h0): the number of occurrences of ACCGTA follows the expected distribution

Alternative Hypothesis (h1): the number of occurrences deviates from the expected distribution

The Chi Square Test is:

$x^2 = \sum \frac{(O_i - E_I)^2}{E_I}$ where O_i is the observed frequency

and E_i is the expected frequency

Exercise 6

Suppose the GC percent of a bacterial genome is 50 and there are equal number of C/G bases as well as equal number of A/T bases on the strand (Note that this equal base assumption does not always hold). Recall that the chi-square statistic is expressed as χ_2=X$_4$ (O$_i$−E$_i$)$_2$ $_{i=1}$ E$_i$ where O$_i$ denotes the number of times base i is observed in the sequence, and E$_i$ denotes the expected number. Suppose a part of the genome of 100 bases has a count of the four bases as A:26, C:24, G:22, T:28.

1. Calculate the χ2 statistic (Show your work).

P(A) = P(T) = 0.25, so E(A) = E(T) = 100*0.25 = 25

P(G) = P(C) = 0.25, so E(G) = E(C) = 100*0.25 = 25

Thus, the expected count for all four nucleotides is 25.

$$x^2 = \sum_{i=1}^{4} \frac{(O_i - E_I)^2}{E_I} = \frac{(O_A - 25)^2}{25} + \frac{(O_C - 25)^2}{25} + \frac{(O_G - 25)^2}{25} + \frac{(O_T - 25)^2}{25}$$

$$x^2 = \frac{(26-25)^2}{25} + \frac{(24-25)^2}{25} + \frac{(22-25)^2}{25} + \frac{(28-25)^2}{25}$$

$$x^2 = \frac{(1)^2}{25} + \frac{(-1)^2}{25} + \frac{(-3)^2}{25} + \frac{(3)^2}{25} = + \frac{1}{25} + \frac{1}{25} + \frac{9}{25} + \frac{9}{25} = \frac{20}{25} = 0.8$$

The x^2 value is 0.8.

2. If the observed number is close to the expected number, $\chi 2$ will be small and vice versa. Using the $\chi 2$ calculator at https://www.graphpad.com/quickcalcs/chisquared1. cfm, you can calculate a p-value. Based on this calculation, comment on this 100-base part of the genome, does it reflect the overall distribution of bases of the entire bacterial genome well?

Using the x^2 calculator, I calculated the p-value to be 0.8495. Since this p-value is > 0.05, there is not significant evidence that this 100-base part of the genome deviates from what is expected. Thus, it does seem to reflect the overall distribution of bases of the entire bacterial genome well.

3. Let X be the random variable of the number of bases A in a region of 100 bases of this bacterial genome. What is the expectation and variance of X (i.e., E(X) and V ar(X)).(Hint: X follows binomial distribution)

n = 100, p = 0.25

E(X) = np = 100(0.25) = 25

Var(X) = np(1-p) = 25*(1-0.25) = 18.75

4. Write an expression for the probability P(X = 30)? (You do not need to expand the combination and power terms)

Where n = 100 and k = 30: P(X=30) $= (n\ choose\ k)p^k(1 - p)^{N-k}$

P(X = 30) = $(100\ choose\ 30)(0.25)^{30}(1 - 0.25)^{100-30}$

P(X = 30) = $(100\ choose\ 30)(0.25)^{30}(0.75)^{70}$

5. Calculating exact values of a binomial distribution can be overwhelming. Therefore, one could approximate a binomial distribution by a normal distribution. Using a normal approximation of X, what is the probability P(X ≥ 30)? (You can use this normal distribution calculator https://onlinestatbook.com/2/calculators/normal_dist.html)

μ = np = 25, σ = $\sqrt{18.75}$ = 4.33, P(X ≥ 30) ≈ P(X ≥ 29.5)

Standardize X:

$$z = \frac{X-\mu}{\sigma} = \frac{29.5-25}{4.33} = 1.0392$$

Using the normal distribution calculator, P(X ≥ 30) = 0.1492

Exercise 7

In PacBio sequencing, the raw-read accuracy is about 87%. The HiFi sequences are built up as a consensus among reads, for which there is a probabilistic formulation for building the consensus. We will explore that here with an over-simplified model of the sequencer output. Assume there is a template of length m bases, for which a read of length n is generated. Without loss of generality, we will assume n = m * P , for an integer value of P . The bases output by the machine are referenced using the notation L_{ji} , where i is the position in the read, and j is the position on the template. Consider the machine has simplistic output: it either outputs the correct base with probability p, and incorrect base with probability 1 − p (we do not yet distinguish between different bases, and there are no insertions or deletions even though in reality there are mostly only insertions or deletions).

1. Write a formal definition of a random variable that can compute the expected number of correctly sequenced bases for a read of length n.

Let X be a random variable that represents the number of correctly sequenced bases in a read of length n. Then, each base is correctly sequenced with the probability p = 0.87 and incorrectly sequenced with the probability p = 1 - 0.87 = 0.13. Since each base is independent of one another, the random variable X follows a binomial distribution, where m*p is the total number of bases in the read, and p = 0.87 is the probability of sequencing each base correctly. So the expected number of correctly sequenced bases is given by: E(X) = np = m*P*0.87.

2. Write a formal definition for the probability of k correctly sequenced bases for a particular position in the template, k ≤ P .

Let Y_J be a random variable that represents the number of correctly sequenced bases at a specific position, J, in the template. Since each position is sequenced P times, and each sequencing attempt is correct with probability p = 0.87, the number of accurate bases at this position follows a binomial distribution. Here, P is the number of times the position is sequenced, and p = 0.87 is the probability that any single sequencing attempt is correct.

Then, the probability that there are k correctly sequenced bases for a particular position in the template is: P(Y_J=k) $= (P\ choose\ k)p^k(1 - p)^{P-k}$

For questions 3 and 4. The consensus can be determined correctly if the number of times a base is read correctly outnumbers the number of times it is read incorrectly. Assume the probability that the base is read correctly is 0.87.

3. Write an expression for the probability that the number of bases correctly read outnumbers the number of bases incorrectly read.

For the number of correct bases to outnumber incorrect bases, Y_J > P/2

Then, the probability that the number of bases correctly read outnumbers the number of bases incorrectly read is: P(Y_J > P/2) $= \sum_{k=(P/2)+1}^{P} (P\ choose\ k)p^{k}(1-p)^{P-k}$

4. What is the number of times a base must be read so that the probability that the number of correctly read bases outnumbers the incorrect bases is at least 0.9999 (QV40)?

Need to find a P value so that P(Y_J > P/2) > 0.9999

Normalize Data:

μ = np = P*p , σ = $\sqrt{P * p * (1-p)}$

$$z = \frac{X-\mu}{\sigma} = \frac{P/2-P*p}{\sqrt{P*p*(1-p)}}$$

The z-score that corresponds to a p value of 0.9999 is 3.8906 so,

$$3.8906 = \frac{P/2-P*(0.87)}{\sqrt{P*(0.87)*(1-0.87)}} = \frac{P/2-P*(0.87)}{\sqrt{P*(0.87)*(0.13)}} = \frac{P(0.5-0.87)}{\sqrt{P*0.1131}} = \frac{P(-0.37)}{\sqrt{P*0.1131}}$$

$$3.8906^2 = \frac{P^2(-0.37)^2}{P*0.1131} = \frac{P(0.1369)}{0.1131} = P(1.2104332)$$

P = 12. 5

Thus, the number of times a base must be read so that the probability that the number of correctly read bases outnumbers the incorrect bases is at least 0.9999 is 13 times (rounded up from 12.5 to get desired accuracy).