**Quiz 1 Answer Key**

We are pleased to report that all students got the dot/similarity matrix correct. The range of scores was from 81.7% to 100.8% with a mean of 95.3%. This range is on the high side compared with past years, but we feel it reflects the high quality of student learning and preparation. We ask students who performed lower than they had hoped not to be discouraged: the teaching team is well aware of your diligence, learning, and engagement in class and reading responses, and wish to emphasize that there will be ample opportunities to improve your overall average through exemplary performance on Quiz 2, the upcoming homework assignments, the participation grade, and the final project.

The teaching team believes that this was a fair quiz. Each question was answered completely correctly by a majority of bio-background students, a majority of quantitative-background students, a majority of undergraduate, and of graduate students as best as we could tell from the student course versions provided on the quiz. The teaching team’s assessment that the quiz was fair was supported by the judgment of two former teaching fellows and another two practicing bioinformaticians on whom the quiz was prototyped. Nonetheless, we welcome grading appeals, and do not consider them annoying, especially with grading the Needleman-Wunsch implementation, which one can imagine lends itself to grading errors.

1. D. All of the listed options affect transcription. The key insight is that, of the options listed, only promoter sequences are generally identical from cell type to cell type, so they cannot explain phenotype differences between cells in the same organism.
2. F. The mass to charge ratio of peptides is assayed by MS and not by protein-protein interaction networks. More ambiguously but worth mentioning: the mass-to-charge ratio is only of instrumental interest for peptide identification and is not of primary interest in biology and medicine, unlike the other listed options.
3. C. Simulation. The other options might be ways of doing science, but they were not listed by Jim Grey in his account of data science as a new modern science that uses computers as databases to unify observation, theory, and simulation. (He would consider “deduction” to fall under theoretical knowledge).
4. F. At 3 million, there are a surprisingly large number of SNVs in a typical personal genome with respect to the human reference genome.
5. B. Some students thought that perhaps the excitement about cryo-EM reflects the improved resolution of the technology; however, the truth is that x-ray crystallography currently offers somewhat better resolution, and the reason for the interest in cryo-EM is that cryo-EM does not require crystallization. Crystallization is extremely laborious and challenging, and becomes increasingly challenging for the sort of large, flexible molecules mentioned in the prompt.
6. G1(bridge)-P1(long)-T1(PacBio);

G2(cohort)-P3(cheap)-T2(Illumina);

G3(validation)-P2(accuracy)-T3(Sanger).

PacBio’s long reads are great for spanning repetitive regions when assembling a genome. Illumina’s cheap reads are great for deeply sequencing a substantial research cohort. This first pass using Illumina might discover some key SNVs that are of clinical interest. We really care about getting those key SNVs right because they may affect the clinical implications of the research, and they are few enough in number that we can afford to use an expensive sequencing technology to validate them. Sanger sequencing, despite being an older technology, has the highest per base accuracy of the options listed, so its high expense remains worthwhile when we want to validate just a limited number of important genomic sites.

There were alternative intelligent approaches to this question, which we will consider on a case-by-case basis.

* 1. Error )1 point(: When Carl moved zip codes, he also moved cities, as we can deduce from Spreadsheet 2 (and we can trust Spreadsheet 2 because the prompt makes clear the error is in Spreadsheet 1). The registrar’s error was in failing to update Carl’s city from Washington to New Haven. To get full credit, students must be specific that the city is the wrong one, since only that answer demonstrates complete understanding of the nature of the error as it applies to the data at hand. (Merely stating that the city and zip code are inconsistent with each other gets half credit). The registrar did not make an “error” when choosing to maintain lecturer information as a set of non-normalized excel spreadsheets: whether to store data as a spreadsheet or as a database is a matter of personal style that depends on context. (In this case, the registrar’s data set was small enough that using spreadsheets was a completely reasonable strategy).
  2. Concept )3 points(: Redundant data tables are an invitation to incomplete update errors. We accepted any answer that acknowledged this vulnerability and proposed a way to prevent it such as with automatic updating or reduced redundancy, regardless of whether that proposal involved jargon. Accepted jargon included normalization and no partial key dependencies. We accepted 2nd normal form even though that is technically false since 3rd normal form would have been necessary, so long as students explicitly mentioned a word related to normalization.

1. Integration of genomic data into electronic health records is hoped to enable molecularly-based diagnosis, risk stratification, prevention, and treatment strategies, as well as the discovery of new genotype-phenotype associations (if those records are made available to researchers). One student mentioned use of genomic data as a personal identifier for patients, which was a neat idea that we could discuss in section.
2. ChIP-seq is a valuable technology for identifying binding sites for proteins on DNA, but it has many limitations and complexities. Students correctly cited the need to normalize the ChIP-seq signal by input amount and complexities related to imperfect antibody binding, non-uniform cross-linking, mappabillity issues, interpretive issues regarding causation vs correlation, difficulties in peak calling – which led Dr. Rinehart to conclude that ChIP-seq signals should be best interpreted to offer qualitative rather than quantitative information (although from another perspective, ChIP-seq does offer quantitative information, ). Full credit was also awarded at the graders’ discretion to important limitations that were not explicitly mentioned in class. Partial credit was awarded to answers that did not list or accurately describe an important limitation of ChIP-seq but which did demonstrate some knowledge of what ChIP-seq is used for and how it works. Students generally did very well on this question, and many students cited multiple correct limitations of ChIP-seq (no extra credit there) despite the prompt asking for only one, which is a healthy sign of student preparation and learning because this was not a topic that was advertised to be on the quiz.
3. Similarity matrix

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | A | S | V | A | V | B |
| E |  |  |  |  |  |  |
| A | 1 |  |  | 1 |  |  |
| V |  |  | 1 |  | 1 |  |
| A | 1 |  |  | 1 |  |  |
| B |  |  |  |  |  | 1 |
| D |  |  |  |  |  |  |

Sum matrix

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | A | S | V | A | V | B |
| E | 2.5 | 2 | 2.5 | 1 | 0.5 | 0 |
| A | 3 | 2.5 | 1 | 2.5 | 0.5 | 0 |
| V | 1 | 1 | 2.5 | 1 | 1.5 | 0 |
| A | 1.5 | 0.5 | 0.5 | 1.5 | 1 | 0 |
| B | 0 | 0 | 0 | 0 | 0 | 1 |
| D | 0 | 0 | 0 | 0 | 0 | 0 |

Alignment:

-ASVAVB-

EA-VA-BD

The sum matrix was worth 8 points, divided equally among the 36 boxes, making a correct answer in each box worth 0.22 points each. To avoid double jeopardy, once a student made an error in an early box, no further points were deducted for propagation of that error to subsequent boxes (but conversely, a student box that matches the answer key box could still be marked wrong if the student should have calculated a different score based on his or her earlier incorrect boxes). One student started from the top-left, which was fine. Some students assigned a gap penalty for the leading gap (where D is aligned with gap), placing a 0.5 in row 5, column 6 instead of the answer key’s 1: this was not the method discussed in section but it is an intelligent alternative that did not conflict with the question’s prompt and no points were removed.

We did not penalize students for stylistic departures from class conventions in the traceback (e.g. omitting the connection lines, circling boxes representing gaps) so long as students highlighted the boxes that participate in the optimal alignment and demonstrated in their alignment that they understood the optimal traceback.

Overall, well-done, and congratulations!