Pathogen identification report

Activity 1.1.3. Can you defend the name your BLAST search gave you?

You already ran the search and you already have the hit table. This is the harder half: writing down why anyone should believe it. A sequence match is evidence, not proof, and the difference between those two words is most of what a diagnostic laboratory does for a living.

HELPS ON THE WEBXAM. A LOT | 60 TO 90 MINUTES | ONE SITTING | NEEDS YOUR BLAST RESULTS FROM 11 SEPTEMBER

What this is

Take the BLAST hit table you saved, the E-value and query coverage you read off it, and the control run you did, and turn all three into a report that names an organism and survives somebody asking you how you know.

This is the write-up that used to sit on 14 September. It came off the calendar so that **How sequencing reads an organism** could have a full class block on 10 September instead of only 25 minutes of homework. That trade was made because you cannot write this report well without understanding what a sequence read actually is, and one class period on that is worth more than one class period of silent typing.

DOES THIS HELP ON THE WEBXAM? YES, AND IT IS THE STRONGEST ITEM IN THE PACK FOR THAT

MI is assessed by WebXam **072130**, which is **57.73 percent Bio-Molecular Technology**. That is the single largest thing you are tested on and sequence identification sits right inside it. This report makes you handle the three numbers the exam cares about, percent identity, E-value and query coverage, and say what each one rules out.

It also lands in **Biotechnology Research and Experiments, 11.34 percent**, because you have to explain what your control run proved and name a limitation of identifying anything by sequence alone. Naming your own limitation is a graded skill, not a confession.

What you need in front of you before you start

GATHER THESE FIRST. ALL OF THEM ARE THINGS YOU ALREADY MADE

- **Your BLAST hit table screenshot** from the computer lab, in your portfolio folder.
- **The top hit's organism name and percent identity.**
- **The E-value and the query coverage** for that same top hit, from your notebook.
- **Your control run result.** The known sequence Mr. Mendoza gave you, and what BLAST returned for it.
- **Your two-hit comparison table**, if you made one, and your revised identification sentence.

Missing any of it? The screenshot and the control can be regenerated in ten minutes on any computer: the case sequence is in Activity 1.1.3 in myPLTW and BLAST is free and public at blast.ncbi.nlm.nih.gov. Ask Mr. Mendoza for the control sequence, then run both again and note in your report that you re-ran them. Re-running and saying so is honest. Quoting numbers you no longer have is not.

What good looks like, before you write anything

Percent identity

Of the bases that lined up, how many matched. High identity over a short stretch is easy to get by chance, which is why this number never travels alone.

Query coverage

How much of *your* sequence took part in the alignment. Ninety nine percent identity across eight percent of your read means almost none of your sequence was involved.

E-value

How many matches this good you would expect to see purely by chance in a database this size. Smaller is stronger. An E-value near zero means chance is a poor explanation.

Control run

A sequence whose right answer is already known, put through the identical workflow. If the control comes back wrong, your unknown result means nothing, however good it looks.

Limitation

A specific, named reason your conclusion could still be wrong. Not 'human error'. Something a reader could go and check.

THE DIFFERENCE BETWEEN A WEAK SENTENCE AND A STRONG ONE

Both of these are about a match. Only one of them is evidence.

> **Weak.** 'The BLAST search showed the pathogen. The percent identity was high so the match is good.'
>
> **Strong.** 'The top hit matched at 99 percent identity across 97 percent query coverage with an E-value of 2e-140. Coverage that high means nearly all of my read took part in the alignment, so the identity figure is describing the whole sequence rather than a short conserved stretch. The control sequence returned its known organism as the top hit under the same settings, so the workflow was behaving.'
>
> The strong version does not use a bigger vocabulary. It names the number, then says what that number rules out. Do that four times and you have the report.

Build it in order

1. **Open your hit table and copy three numbers into your notebook:** percent identity, query coverage, E-value. Write them down before you write a single sentence, so the report gets built on the data rather than the data hunted down to fit the report.

   Done when: you have three numbers written down with their labels

2. **Write the Claim in one sentence.** Name the organism and nothing else. No hedging, no 'it might possibly be'. A claim you have to hedge is a claim you have not finished thinking about.

   Done when: one sentence, one organism name

3. **Write the Reasoning paragraph, one number at a time.** Take percent identity first: say what it is and what it would still allow to be true. Then query coverage: say what proportion of your read aligned and why that matters. Then E-value: say what it means for chance as an explanation. Three sentences minimum, one per number.

   Done when: each of your three numbers appears in the paragraph with a consequence attached

4. **Write the control paragraph.** Name the control sequence, say what organism it should have returned, say what it did return, and state what that tells you about the workflow. If your control came back wrong, say so and say what you would change. A failed control honestly reported scores better than a passed control asserted without evidence.

   Done when: a reader can tell whether the workflow was trustworthy on the day you ran it

5. **Write one limitation sentence.** Pick a real one. Three that are genuinely true here: the reference database only contains organisms somebody has already sequenced and deposited; a short region can be nearly identical across close relatives so the match may be right at genus and wrong at species; and a contaminated sample sequences the contaminant just as happily as the pathogen.

   Done when: your limitation names something specific that could be checked

6. **Write one sentence on what you would do next in a real laboratory** to confirm the identification by a second, independent method. It does not have to be a method we have run.

   Done when: one named next step that is not another BLAST search

7. **Attach the screenshot, put your name and period on every page, and submit.**

   Done when: the hit table is legible in the image you attached

Fill this in as you go

YOUR THREE NUMBERS AND WHAT EACH ONE SETTLES

WHAT THE NUMBER IS	YOUR VALUE	WHAT IT RULES OUT, IN YOUR OWN WORDS
Percent identity of the top hit		
Query coverage of the top hit		
E-value of the top hit		
Second hit, for comparison		

YOUR CLAIM, IN ONE SENTENCE

YOUR LIMITATION SENTENCE

If you get stuck

My coverage is low and my identity is high. Is my match good or bad?

Suspicious, and saying so is the right answer. It means a short piece of your sequence matched very well and the rest did not take part. That happens when a conserved region is shared across many organisms. Report the pair honestly and explain what the low coverage prevents you from concluding. That is a better report than one with prettier numbers.

BLAST gave me several hits with almost identical scores.

Then the honest claim is at whatever level they agree on. If four hits are four species of the same genus, name the genus and say why you cannot go finer. Backing off to the level your data supports is a stronger move than picking one at random.

My control run did not return the organism it should have.

Report exactly that. Then check the obvious things: did you paste the whole sequence, did you choose nucleotide BLAST, did any characters other than A, T, C and G come along in the paste. Write down what you found. A troubleshooting paragraph is worth more than a clean result you cannot explain.

I do not know what counts as a limitation.

Ask yourself one question: what would have to be true about the world for my answer to be wrong even though my numbers are right? Every answer to that question is a limitation.

Can I just write that the pathogen was identified successfully?

No, and not because it is against the rules. It is because that sentence contains no information. A reader cannot check it, cannot disagree with it, and cannot learn anything from it.

What to hand in

THE FINISHED REPORT

- **Where it goes.** The same place as everything else. Your class form on the portal, or handed to Mr. Mendoza in person. Nothing about extra credit needs a different route.
- **What to hand in.** Your written pages, plus photographs of anything the handout asks you to photograph. Photographs of handwritten pages are fine. A photograph you cannot read is not.
- **Put your name and period on every page and every photograph filename.** Unlabelled work cannot be credited and it is the single most common way points get lost.
- **When.** Any day after the pack opens, right up to the last week of the semester. There is no deadline inside that window and no penalty for going slowly.
- **Feedback.** You get it back with a note. If a piece is short of full credit you may fix the named thing and resubmit once.

HOW BIG THIS ITEM IS: ONE WORK SESSION

About 60 to 90 minutes in one sitting. Almost all of that is writing. The searching is already done, so nothing here needs a booking, a bench or a partner.

The point value is printed on the portal the day this item opens, on the item's own card and in the gradebook entry. It is not printed here, because a number on paper and a number in the gradebook that disagree is worse than no number at all. Extra credit in this class is capped per semester, so check the cap on the portal before you start a fourth item.