Biologists and Graph Interpretation Network

From Nano to Life: Interpreting graphs from Nanopore sequencing of bacterial genomes

A BioGraphI Module Lesson Guide for Instructors

Original Authors
Carlos C. Goller, North Carolina State University, Raleigh, NC
Carly Sjogren, North Carolina State University, Raleigh, NC
Jason Williams, Cold Spring Harbor Laboratories

Course Information
Department/Program: Biotechnology Program (BIT)
Level: Introductory
Course type: Lab and lecture
Delivery mode: Asynchronous online
Students: Mixed majors/nonmajors
Number of students: 10-16
Estimated duration of activity: One hour in-person/synchronous and one-hour asynchronous work.

Expected date or dates of implementation
Spring 2023

Purpose/Background
The sustainable reuse of electronics is a global challenge. Sustainability is truly an interdisciplinary challenge, and the use of biotechnologies and microbial genomics can be leveraged to understand how microbes can help tackle this global challenge. For the 200-level Biotechnology & Sustainability course, we will develop a lesson that includes the following elements:

- Interdisciplinary and diverse researchers
- Analysis of genomic data using the Oxford Nanopore Technologies (ONT) MinION portable sequencing device.
- Exploration of applications of ONT long-read sequencing

List of materials needed for the lesson

- Link to interview on YouTube (total running time, 41 min)

    - https://youtu.be/vx93Re_rP5c (video, 41 min, captioned by rev.com and corrected)
- Link to presentation slides
    - Goller PresentationSlides_FromLifetoNanoSequencing
- 3_Assessment_NanotoLife
- 4_H5P activity file

About BioGraphI modules

This lesson is a BioGraphI module. BioGraphI modules address data literacy while fostering diversity in undergraduate biology classrooms. They are lessons about graph and data interpretation, featuring the scientific contributions of biologists who are members of historically excluded groups (HEGs). They include video interviews with these biologists, allowing students to hear directly from HEGs about their discoveries. For more information about how the BioGraphI project is advancing inclusion in biology and improving data literacy, visit our webpage.

Student Learning Objectives

The BioGraphI Student Learning Outcomes (LOs) describe what students can expect to gain by the end of the BioGraphI lesson. They are written in a format that can be shared directly with students.

Quantitative learning objectives

1. **Identify** specific components of graphs from Oxford Nanopore Technologies sequencing output.
2. **Reflect** on your perceptions about using graphs or figures in biology.

Diversity/equity/inclusion learning objectives

3. **Reflect** on your perceptions of people who do biology.
4. **Compare** your own interests and/or identities to those of people who do biology.

Content learning objectives

5. In 3-5 sentences, including specific details about the graph, **describe** the major findings and limitations of the Nanopore sequencing dataset depicted.
6. **Phrase** a testable hypothesis based on the use of Oxford Nanopore Technologies sequencing for bacterial genome sequencing.
7. Using miro.com, GoogleSlides, or Jamboard, **design** a simple flowchart describing a future experiment.

Assessments

To help the BioGraphI Project to measure the effectiveness of our modules in improving data literacy and fostering diversity in biology classrooms, we invite your students to participate in a voluntary, anonymous pre-/post-lesson survey (Geneseo IRB #202021048). This survey is designed as an opportunity for reflecting on the Quantitative and D/E/I learning objectives above and administered via LimeSurvey. Click Instructions for access to BioGraphI PrePost-Lesson Student Survey to request a survey to be set up for your students, at least 7 to 10 days in advance of your class meeting date.

Objectives	Formative Assessment	Summative Assessments
1	Introductory videos through an H5P activity with with multiple-choice questions with multiple attempts. During the synchronous session, polls to compare responses about the data set.	A post-lesson learning management system (LMS) quiz with an essay and or audio recording option to describe again thoughts on the major findings of a related graph.
2	Reflective questions during H5P activity.	BioGraphI Student Post-Lesson reflection.
3, 4	Reflective questions during H5P activity. Group chat waterfall on Zoom and discussion board.	Post-lesson reflection on Scaffolded Notes for the unit.
5, 6, 7	Brainstorming activity with peers. Miro flowchart design activity in small groups.	Post lesson quiz with an essay question

Lesson context

Learning goals of unit

Module 7. Demystifying Big Data & High-throughput Approaches

- **MO 7.1. Define** 'big data' and provide examples. (CO 3)
- **MO 7.2. Explain** how microbial genomes are sequenced. (CO 3)

- **MO 7.3. Discuss** the role of automation in modern scientific discovery. (CO 3)
- **MO 7.4. Analyze** a large dataset using bioinformatics tools (CO 3)

Prerequisite skills or knowledge

- Skills practiced prior to this lesson such as basic elements of graphs and use of Zoom or similar video conferencing system and Miro digital flowcharts for sharing ideas in the chat and through flowcharts.
- Conceptual background knowledge on the fundamentals of what is a gene, a genome, and basic understanding of DNA sequencing.

Preparation for lesson

- Assign BioGraphI Student Pre-Lesson H5P activity as homework for students to complete before this in-class activity.
- Assign a short video and text to introduce the Oxford Nanopore Technologies sequencing platform and goals of the lesson. (H5P)
- Remind students that we will work in teams on the lesson during the synchronous session.
- Share the link to the Google Slides and videos through a course announcement using the Moodle LMS.

Lesson Sequence

Each row of this table is a step of the activity. Column headings reveal what the instructor, interviewee, and students do at each step.

Information from instructor (live in-class)	**Information from scientist (within pre-recorded video interview)**	**Student follow-up or transition activity**
Prior to class, assign any preparation activities	N/A	Students complete homework/pre-lesson survey for BioGraphI project (combination of reflective writing prompts and closed-ended questions) ● *Purpose*: prepare for class; LO 2 & 3 – Pre-lesson reflection on data literacy and perceptions of scientists
Provide a skeleton graph on a slide and highlight axes	N/A	**Fundamental activity:** Make reasonable guesses to what the graph is about, share out ● *Purpose*: Warm-up and

and components.		icebreaker, rewards creativity, normalizes errors (nearly impossible to be correct) **Advanced activity:** Make the graph using the data ● *Purpose*: Warm-up and icebreaker
Play first video segment	Orient us to this graph we've been working on. (What do these axes mean? How did you measure stuff? What do the units mean? How did you make this graph? Why did you decide to show the data in this way?) ● *Vocabulary to incorporate:* types of graphs, units, independent (explanatory, predictor) variables, dependent (response) variables,	**Fundamental activity:** Students interpret/re-interpret graphs with the correct axis labels. Infer the hypothesis that was studied. Evaluate whether or not their hypothesis about the graph was supported. Discuss with peers. Share out. ● *Purpose*: LO 1A & 2 - Check student understanding of the concepts and vocabulary **Advanced activity**: Students compare the graphs they made to those described by the interviewee ● *Purpose*: LO 1B & 2 - Check student mastery
Plays next video segment	Tell me about the research that these data came from: scientist confirms the hypothesis inferred by & the interpretations made by students. ● *Vocabulary to incorporate:* hypothesis, prediction, (explanatory, predictor) variables, dependent (response) variables, controlled (standardized) variables if applicable	Based on these results, what should the scientist investigate next? Share out. ● *Purpose*: Transition to next steps; could be used to align Content LOs or LO 1A or B
Plays next video segment	Tell me about the research that these data came from: Describe the follow up experiment or potential follow up experiments	What's significant about these findings? Share out. ● *Purpose*: LO 1A or B, and could be used to align Content LOs
Plays next video segment	Tell me about the research that these data came from: Explain the significance of this research	*Purpose*: Closes this portion of the lesson

Adds a little info about the scientist (ppt slide): a picture, affiliation	N/A	What questions do you have about [scientist]'s path to this work/where they are today? Share out. ● *Purpose*: LO 4 - Identifying potential connections between students and scientist
Selects questions that align best to student generated questions; Must show the answers to at least the required interview Qs to students	***Required interview Qs*** **What kinds of scientific questions interest you the most?** **When and how did you know you wanted to be a scientist?** **Tell us about the paths that led you to your current job.** **What is it like being a scientist who identifies as [fill in the blank]?** **Tell me about a moment when you felt like you really belonged in the field of science.** *Strongly suggested Qs* What is a typical day like for you? How do you balance work with other interests? *Optional interview Qs* Where did you grow up, and what was it like growing up there? What keeps you motivated to do research? What is the most important advice you would give me [context/audience decided by interviewee & interviewer]? What do you do if you have questions about how to analyze your data?	*Purpose:* Answers the questions identified by students
Instructor shares contact info (twitter, email, etc) of scientist so that students can ask leftover questions! #BioGraphI…	(braces themselves for media deluge)	Students ask their questions during class (or homework!) ● *Purpose*: LO 4 - Identifying potential connections between students and scientist. Closes the in-class portion of the lesson.
Instructor could provide other	N/A	Students complete homework/post-lesson survey for

resources on scientist (e.g., website, papers, etc.)		BioGraphI project (combination of reflective writing prompts and closed-ended questions) ● *Purpose*: LO 2 & 3 – Post-lesson reflection on data literacy and perceptions of scientists

Alignment to Universal Design for Learning Guidelines

UDL Guideline	**Lesson Alignment**
Multiple means of Engagement	Recruiting interest: We will begin with an intriguing "hook" to capture student interest. Students have the opportunity to generate their own questions for the scientist by posting to a Jamboard or Padlet that we will share with Jason Williams. Self-regulation: Students reflect on their dispositions and knowledge of scientists and graph interpretation before and after the lesson
Multiple means of Representation	Perception: Captions are provided on the interview video. Google Slides offers automatic captioning for presenters. Alt text for images provided in Google Slides. Multiple types of media are used in the lesson (visual, audio, text). I will also create animations to share how the interactivity of graphs is displayed AND verbally describe what is happening. Slide Notes will contain tips, links, and scripts. The graph colors will be selected to avoid combinations that can be difficult for color blindness. We will use a web-based Grackle Slides checker to review slides for accessibility concerns and follow 5 Tips on Designing

	Colorblind-Friendly Visualizations suggestions for color combinations for data visualization.
Multiple means of Action and Expression	Expression & communication: Learners are able to interact with the graphs as images and as dynamic HTML renderings from Nanopore output.

Implementation notes

Notes on our implementation
The asynchronous H5P activity was well-received and helped students watch the videos and learn from the interviewee at their own pace. In our small class, the synchronous session had fifteen attendees who reflected on the activity and analysis of graphs.

Suggested adaptations
The synchronous lab session is another opportunity to include skeleton graphs and HTML reports from Nanopore output to encourage students to think deeply about the content. A rubric would help learners know what is expected of them in their reflections and help instructors grade consistently. The post-activity quiz had several multiple-choice questions that could be revised to include open-ended questions and more feedback to promote deeper thinking and creativity.

References/resources

1. 5 Tips on Designing Colorblind-Friendly Visualizations
2. Nanopore sequencing data analysis
3. Basic QC Tutorial

Funding for BioGraphI

This project is funded by the Directorate for Biological Sciences, Division of Biological Infrastructure as part of efforts to address the challenges posed in Vision and Change in Undergraduate Biology Education: A Call to Action (http://visionandchange/finalreport/). Award number 2120679.

This award reflects NSF's statutory mission and has been deemed worthy of support through evaluation using the Foundation's intellectual merit and broader impacts review criteria.

Transcript of the interview

[00:00:01.06]
- Welcome, Jason Williams.

[00:00:05.00]
How are you today?

[00:00:06.08]
- Good, thank you for
the invitation, Carlos.

[00:00:09.03]
- No, thank you for agreeing.

[00:00:10.09]
We've worked together

[00:00:14.05]
and you've taught me in workshops

[00:00:17.03]
and I thought this was a
really fantastic opportunity

[00:00:20.08]
for you to share with us

[00:00:24.03]
and students how you got
into this latest set of tools

[00:00:31.03]
and use of Oxford Nanopore sequencing.

[00:00:37.01]
So, today we have the opportunity
to chat with you a little

[00:00:41.08]
and learn about your career

[00:00:44.04]
and learn about graphs

[00:00:47.07]
and analyzing this
nanopore sequencing data.

[00:00:52.09]
So, tell us a little
bit about your position

[00:00:57.01]
and we'll get started with that.

[00:01:00.07]
How does that sound?

[00:01:02.02]
- Sure.

[00:01:03.02]
I guess I gave you a couple slides,

[00:01:06.00]
so at least I'm not surprised
what's up on the screen.

[00:01:11.01]
- You tell me when to advance.

[00:01:13.01]
- Yeah, yeah, yeah.

[00:01:14.06]
That also requires me

[00:01:15.05]
to remember whatever the next
slide is, but I'll do my best.

[00:01:18.06]
So yeah, I'm an Assistant Director,

[00:01:21.01]
Diversity Research Readiness
at the DNA Learning Center

[00:01:23.09]
which is at Cold Spring Harbor Laboratory.

[00:01:26.04]
I don't know what that
tells you very much,

[00:01:28.01]
but let me try to break it down.

[00:01:30.01]
I'll start from the bottom up.

[00:01:32.07]
So, Cold Spring Harbor Laboratory
is a research institution

[00:01:36.06]
that focuses on things
like cancer, neuroscience,

[00:01:39.08]
plant biology where I got my start,

[00:01:41.04]
in fact actually happens to be
my background picture today.

[00:01:43.07]
So, that's a little bit about the lab.

[00:01:47.01]
And within the lab is
the DNA Learning Center

[00:01:49.09]
where we focus on hands-on

science education,

[00:01:53.01]
including training for faculty,

[00:01:55.04]
which is how I've often

[00:01:56.07]
gotten a chance to work with you, Carlos.

[00:01:58.09]
And because we both have
the educational aspect

[00:02:03.03]
and we teach about genetics and genomics,

[00:02:06.06]
but also that we work

[00:02:08.01]
really in the context of a research lab,

[00:02:10.04]
we often are really excited

[00:02:12.06]
about bringing new things
into the classroom.

[00:02:15.03]
And nanopore sequencing just
happens to be the latest thing.

[00:02:19.01]
It's a favorite toy of mine,

[00:02:21.09]
but of course we want it
to be more than a toy.

[00:02:23.07]
We want it to be an educational
tool that puts everyone,

[00:02:29.03]
including students in touch with

[00:02:31.06]
the data that they're generating,

[00:02:32.09]
but then also helps us to
answer scientific questions.

[00:02:36.09]
So, that's maybe a little bit about

[00:02:39.00]
how I've gotten into nanopore.

[00:02:42.00]
- [Carlos] You're always
trying to teach students

[00:02:43.08]
about visualizing these big
and often complex datasets.

[00:02:52.00]
Tell us a little bit about what you think

[00:02:54.05]
the power of visualizing datasets,

[00:02:58.03]
including sequencing datasets is.

[00:03:01.03]
- [Jason] Great, yes,
that's a good question.

[00:03:03.08]

You're absolutely right.

[00:03:05.05]
Visualization is extremely
important to doing science.

[00:03:09.09]
I'm gonna broaden the
term in just a moment

[00:03:11.07]
with the next slide,

[00:03:12.05]
but I better stay on
this slide for a moment

[00:03:14.07]
about this idea of communicating patterns.

[00:03:17.01]
I really like this animated slide,

[00:03:20.05]
which shows some,
basically a scatter plot,

[00:03:24.06]
which I think you will have
talked with your students about,

[00:03:27.02]
pointing something on
an X-axis and a Y-axis.

[00:03:30.09]
And the point of this animated slide

[00:03:33.00]
that's on the screen currently

[00:03:35.01]

is that you could have
a group of data points

[00:03:38.07]
and they could all, if you average their,

[00:03:41.02]
the mean on the X-axis or
the mean on the Y-axis,

[00:03:44.00]
their standard deviation,
their correlation,

[00:03:46.07]
you could have different datasets

[00:03:48.02]
which have exactly the same numbers.

[00:03:53.04]
In this one, the mean is 0.5426, right?

[00:03:56.08]
So to a decent amount of
precision, they have the same.

[00:04:01.05]
And then if you go out to the
thousandth or the millionth

[00:04:03.09]
or whatever, they change
the numbers slightly.

[00:04:06.04]
But the point is,

[00:04:07.02]
is that you can get almost
any pattern out of a dataset

[00:04:11.06]
by just looking at summary statistics.

[00:04:13.05]
And in fact, the aim of this visualization

[00:04:15.06]
is to show you can even make a dinosaur,

[00:04:17.08]
or a star, whatever kind of pattern.

[00:04:20.01]
To our eyes, we can quite
clearly see those patterns,

[00:04:24.01]
but looking at the numbers,

[00:04:26.06]
just in a summary, we can't
really see those patterns.

[00:04:29.08]
And the reason why I also,

[00:04:32.08]
I appreciate the word visualization,

[00:04:34.06]
but I also make it to more the idea of

[00:04:36.08]
how do you communicate patterns

[00:04:38.05]
is because visuals are not the
only way that we can do that.

[00:04:42.08]
And so, for people who have
low vision or no vision,

[00:04:47.01]
that doesn't mean that
you can't visualize.

[00:04:49.06]
On my next slide, if you
wouldn't mind advancing, Carlos,

[00:04:52.03]
I also thought of just
a quick analogy to this,

[00:04:55.05]
which it's a bit of a trick,
what I have on the screen.

[00:04:59.02]
What I have on the
screen now is some music.

[00:05:04.03]
I could put you on the
spot and ask you, Carlos,

[00:05:08.01]
what is this piece of music?

[00:05:10.04]
- [Carlos] Oh, you are
bringing back piano lessons.

[00:05:13.09]
- [Jason] Well, I'm glad
that you at least noticed

[00:05:16.06]
that it's piano music

[00:05:17.06]
because we see they're both hands,

[00:05:21.03]
the bass and the treble clef.

[00:05:24.00]
But this is even more tricky
than what I'm asking you to do,

[00:05:26.09]
to say what piece of music,

[00:05:28.00]
because this is actually a piece,

[00:05:32.03]
just a moment of one that I like

[00:05:35.05]
from Beethoven's "Ninth Symphony."

[00:05:37.08]
And it's a bit of a double trick

[00:05:39.09]
because the Beethoven's Ninth piano

[00:05:41.07]
is absolutely not for piano.

[00:05:43.01]
You know, Beethoven's Ninth
Symphony is a symphony.

[00:05:45.00]
It's not a piano piece.

[00:05:46.03]
So this is even more tricky

[00:05:48.00]
because Liszt wrote transcriptions

[00:05:51.01]

of all of the Beethoven
symphonies for piano.

[00:05:55.03]
And so you would not only
have to know the music,

[00:05:58.01]
but you'd have to know
how the music sounds

[00:06:00.05]
taken down from being
a symphony to a piano.

[00:06:03.02]
My point here is,

[00:06:04.00]
is that the information
is actually still there.

[00:06:07.04]
If you were to play it on the piano,

[00:06:09.02]
you would be able to hear
Beethoven's Ninth Symphony

[00:06:12.01]
and spot the patterns from the chords.

[00:06:13.08]
The visual element might be a
little bit of irrelevant here,

[00:06:17.09]
but it may be the take
home message is that I bet

[00:06:21.06]
for non-musicians, if

you play their favorite,

[00:06:24.00]
if you were to show them a copy of their,

[00:06:25.09]
of a song without maybe
the words that they like,

[00:06:29.08]
they might not be able to recognize a song

[00:06:31.08]
that they listen to
really frequently or enjoy

[00:06:34.02]
if they just saw the music

[00:06:35.03]
and they don't know how to read music.

[00:06:36.08]
But hearing the song they
would be able to recognize it

[00:06:39.06]
even if it was played by somebody else,

[00:06:42.03]
or if it was a piece that
was without the words,

[00:06:44.06]
or they changed it in some other way.

[00:06:46.02]
As long as you kept the
key pattern together,

[00:06:50.08]
we'd be able to recognize it.

[00:06:52.01]
And so, no matter what our sensory input,

[00:06:56.07]
there are ways to communicate
patterns that come across,

[00:06:59.07]
whether we're talking literal,

[00:07:00.09]
about literal vision with our eyes

[00:07:02.07]
or with other types of
sensory perceptions.

[00:07:04.06]
So just a thought there

[00:07:06.02]
to expand the idea of
communicating patterns.

[00:07:08.06]
Anything that helps us
communicate patterns

[00:07:11.06]
could fit into that idea of a graph.

[00:07:13.03]
And in the case of music we
actually do have a graph.

[00:07:16.00]
It has a Y-axis, which
represents the time.

[00:07:20.05]
And it has different

values represent pitch.

[00:07:23.07]
So it's not so different
from our scatter plot,

[00:07:26.07]
but it just gives a nod to the fact that

[00:07:28.05]
we can use all of our senses
to communicate patterns

[00:07:31.01]
and to think about [Carlos
speaking over Jason].

[00:07:32.09]
- [Carlos] That's awesome analogy, Jason.

[00:07:36.00]
- [Jason] Very good.

[00:07:36.08]
- [Carlos] I love the
emphasis on communicating

[00:07:39.06]
and not just visual.

[00:07:40.08]
It gets to this multiple
means of representing

[00:07:45.09]
and engaging with information.

[00:07:48.07]
Not just visual, but to
communicate patterns.

[00:07:53.00]

- [Jason] Absolutely.

[00:07:57.00]
- [Carlos] So what is

[00:07:58.09]
this next slide that you shared with us?

[00:08:02.01]
You shared a data set.

[00:08:04.00]
And we often,

[00:08:05.05]
I often emphasize to students
garbage in, garbage out.

[00:08:12.02]
The data we get, we have
to, from sequencing runs,

[00:08:17.04]
we have to assess for quality.

[00:08:20.00]
So what is this quality control
graph you're showing us?

[00:08:24.07]
- [Jason] Great, great,
great point, Carlos.

[00:08:27.04]
Yeah, so,

[00:08:30.09]
I guess before we even
talk about a visualization,

[00:08:35.00]
a summarize of data,

[00:08:36.08]
we definitely need to have expectations,

[00:08:40.05]
things that we're looking for,

[00:08:42.02]
and we have to understand

[00:08:43.08]
where do those expectations come from?

[00:08:46.02]
Are the expectations about
the type of experiment

[00:08:48.05]
that we're running?

[00:08:49.03]
Are the expectations about the biology

[00:08:51.02]
that we think we expect to find
when we do this experiment?

[00:08:55.07]
What you're seeing here is a graph,

[00:08:59.01]
and I'll,

[00:09:00.08]
we'll look at them both quickly

[00:09:02.01]
and then we'll come back and focus on one,

[00:09:04.04]
but what we have on the
screen now is a line graph.

[00:09:09.05]
And on one axis, the X-axis
we have a percentage,

[00:09:15.01]
and on the Y-axis we have a value.

[00:09:18.03]
It's actually kind of bend
and I'll explain this.

[00:09:20.04]
So let me first talk
about what the data are

[00:09:24.03]
and what we're trying to see here.

[00:09:26.03]
So this is one of a
half dozen reports or so

[00:09:30.02]
that you can get inside of a
software tool called FASTQC,

[00:09:33.02]
or Fast Quality Control,

[00:09:34.09]
that we use to look at sequence data.

[00:09:37.03]
And this dataset,

[00:09:40.02]
since I happen to have on the screen

[00:09:41.08]
a section called summary statistics,

[00:09:44.03]
is about I think 250,000
individual DNA reads

[00:09:48.08]
that we are summarizing in this dataset.

[00:09:52.06]
And what this dataset is
trying to ask the question is

[00:09:56.02]
how many of those reads are unique?

[00:09:58.09]
In other words, if I were to look at

[00:10:02.01]
250,000 DNA sequence reads
that I got from some organism,

[00:10:07.08]
would I expect them to
be the same or different?

[00:10:12.00]
If I chop them up,

[00:10:13.01]
in this case it says the
sequence length is 40 base pairs.

[00:10:16.03]
So 40 base pairs, if you
think about a password,

[00:10:20.07]
if you had a password that
was 40 characters long,

[00:10:23.05]
it's probably gonna be

pretty unique, right?

[00:10:25.04]
Hopefully.

[00:10:27.00]
It would differ from another,

[00:10:28.08]
it would be pretty hard to crack

[00:10:30.03]
even with a computer to
try to guess your password

[00:10:33.01]
if you had 40.

[00:10:34.03]
Well, what we hope to see on a
graph, usually like this one,

[00:10:38.01]
is that most of the sequences are unique.

[00:10:41.06]
And if you look at this graph,

[00:10:43.03]
on the X-axis it tells you how many copies

[00:10:46.02]
of the sequence do we have.

[00:10:48.04]
And it's almost 100% of the sequences

[00:10:51.06]
of the total percentage
is only having one,

[00:10:54.09]

somewhere between one and two copies.

[00:10:56.06]
So in other words, most of
the sequences are unique.

[00:11:00.05]
However, there is a little bump.

[00:11:02.08]
As you start going down the Y-axis,

[00:11:04.06]
it starts becoming a
bit, I guess logarithmic

[00:11:07.03]
where you have,

[00:11:08.04]
well how much of the sequence

[00:11:11.04]
do you have greater than 10 copies,

[00:11:13.03]
greater than 50, greater
than 100, greater than 500,

[00:11:15.08]
greater than 1,000, so on and so forth.

[00:11:18.01]
So there is a little bit of a bump.

[00:11:20.07]
Somewhere around some sequences
that have a little bit more,

[00:11:25.03]
there's a little small population just by,

[00:11:27.06]
maybe it's somewhere around 3 to 5%

[00:11:29.09]
of sequences where you
have more than 10 copies.

[00:11:34.00]
So why is that?

[00:11:36.01]
Well, your genome, as you may know, is,

[00:11:39.01]
or any genome usually has some areas of it

[00:11:41.04]
that are pretty repetitive,

[00:11:43.03]
where you have the same thing

[00:11:45.03]
written over and over and over again.

[00:11:47.01]
Transposons do this.

[00:11:49.00]
Areas of the centromere

[00:11:50.08]
or the telomere where you have

[00:11:52.01]
the same sequence motif repeated

[00:11:54.00]
over and over and over again.

[00:11:55.02]
So we'd actually expect that

[00:11:56.08]
the sequence isn't totally unique,

[00:11:58.03]
but that we'd have a lot
of unique sequence data.

[00:12:02.04]
And if we notice,

[00:12:03.03]
those green check marks
are there to say that, ooh,

[00:12:06.01]
this is good data according
to whoever made the rules,

[00:12:08.06]
the people who wrote the
software for what is good.

[00:12:11.01]
If you go to the next slide, however,

[00:12:14.07]
this is data

[00:12:15.08]
and I've just shown you all
of the different reports.

[00:12:18.02]
This is supposed to be
an example of bad data.

[00:12:21.03]
Well, it at least has a warning.

[00:12:23.04]
And so in this case we see a line graph

[00:12:26.04]
and it's talking about
the overall percentage

[00:12:28.06]
of total sequences that have duplications.

[00:12:32.07]
And here, you're somewhere
in the 65% I'd say

[00:12:37.01]
with that blue line are
unique, only one copy.

[00:12:43.01]
But then there are large
number of sequences

[00:12:45.02]
sort of around the category

[00:12:48.00]
that's called greater than 1,000 copies.

[00:12:50.01]
So somewhere around,
I would say 18 to 20%.

[00:12:53.05]
I don't know, it looks
around there just by the eye,

[00:12:56.06]
where there's a large number of sequences

[00:12:58.07]
that are greater than a thousand copies.

[00:13:01.01]
So it's given a warning

here, but how would you know?

[00:13:04.03]
Maybe it's bad sequence,
something went wrong,

[00:13:07.07]
but more likely it's a
different type of experiment.

[00:13:11.01]
So in a sequence experiment

[00:13:13.08]
where we're looking at RNA-seq data,

[00:13:16.04]
we're actually looking
for transcripts of things

[00:13:18.06]
that are being made in the cell.

[00:13:20.01]
And actually the cell needs to make

[00:13:22.02]
lots of the same proteins all the time.

[00:13:25.01]
There's just proteins you're
gonna need to make every day,

[00:13:27.08]
maybe even several times a day

[00:13:29.05]
in order to maintain the basic function.

[00:13:31.05]
Things in the cytoskeleton.

[00:13:33.04]
Other types of proteins.

[00:13:34.08]
So it depends on what type
of experiment you're doing.

[00:13:37.09]
If you were doing an RNA-seq experiment,

[00:13:40.04]
you'd actually like to see

[00:13:41.07]
a distribution where you are getting

[00:13:43.09]
lots and lots of copies of the same thing

[00:13:46.01]
over and over again

[00:13:46.09]
because that's what the cell
is actively transcribing.

[00:13:50.00]
Or it could be that

[00:13:51.08]
you have a really,
really repetitive genome.

[00:13:54.00]
So lots of plant genomes.

[00:13:55.05]
There's some that have,

[00:13:56.03]
there are many plant
genomes that have genomes

[00:13:58.02]
that are much larger than human genomes.

[00:13:59.07]
And a lot of the times
those are transposons

[00:14:02.03]
that are duplicated.

[00:14:03.08]
So it's sort of in the
eye of the beholder.

[00:14:06.09]
But this is a nice clue

[00:14:09.06]
and it's something that would
be really hard, I think.

[00:14:12.02]
I can't think of any other way
other than to do a line graph

[00:14:15.01]
or a bar graph to kind of
communicate to somebody

[00:14:18.04]
how many of your reads in a
sequencing run are the same.

[00:14:23.09]
And to give you an appreciation of whether

[00:14:28.03]
this is what you would expect

[00:14:29.04]
because you know your genome is repetitive

[00:14:31.07]
or this is what you would expect

[00:14:32.09]
because you did an RNA-seq experiment

[00:14:34.08]
and you're expecting to get lots of copies

[00:14:38.05]
of single genes being replicated.

[00:14:41.02]
So it all is kind of not
just taking in the data,

[00:14:47.02]
but really trying to ask the question,

[00:14:49.05]
is it what I would expect?

[00:14:51.06]
Does it reflect the biology that I know?

[00:14:54.04]
- [Carlos] What does this percentage mean?

[00:14:57.08]
- [Jason] So it says,

[00:14:58.08]
so percentage of sequences
remaining if deduplicated.

[00:15:02.08]
So this is the, this is
related to the blue line

[00:15:05.09]
at the very left hand side

where you see that it says

[00:15:09.07]
we've got about 60 some
odd percent of sequences

[00:15:12.03]
that are unique.

[00:15:13.03]
So one of the tools that you could do,

[00:15:15.09]
and there's reasons for
it, is you could say, okay,

[00:15:19.02]
look at all of my duplicated sequences

[00:15:21.03]
and anything that's there more than once,

[00:15:23.00]
just get rid of it.

[00:15:24.06]
So you could normalize the data

[00:15:26.09]
by saying I just wanna get
all of the unique reads.

[00:15:30.02]
And there are reasons why
you might wanna do that.

[00:15:32.09]
If you were doing RNA-seq you
probably don't wanna do that

[00:15:35.04]
because you just gotten

rid all of your data.

[00:15:37.05]
RNA-seq is sometimes a
way of referring to it

[00:15:41.00]
as a digital expression or
digital count of expression.

[00:15:45.00]
In other words,

[00:15:45.09]
you're counting this read
has 20 copies of this RNA,

[00:15:49.03]
therefore probably it's replicating

[00:15:51.07]
or it's being transcribed at that rate.

[00:15:53.09]
But another gene might have
2,000 copies of that transcript,

[00:15:58.02]
therefore probably more
of that is being made

[00:16:00.02]
at any given time

[00:16:01.02]
or the transcript is more
stable or other reasons.

[00:16:03.04]
But that's information.

[00:16:05.07]

On the other hand, recently,

[00:16:07.08]
in fact actually I'm doing this right now,

[00:16:10.02]
I want to know what the
size of my genome is.

[00:16:13.02]
And so in order to know what
the size of the genome is,

[00:16:16.02]
I actually don't wanna get that
repetitive sequence so much.

[00:16:20.03]
I actually wanna get rid of it

[00:16:21.05]
because I'm actually asking the question,

[00:16:25.09]
if I were to look at the genome,
I don't know how big it is,

[00:16:28.05]
and this is the case where
sequencing with some students,

[00:16:31.03]
an endangered species,

[00:16:32.06]
and we don't know what
the size the genome is.

[00:16:34.01]
We have a guess based on other
things, but we don't know.

[00:16:37.02]
So we could ask the question,

[00:16:39.02]
how many unique segments of DNA,

[00:16:44.00]
if we break the DNA into 40 base pairs

[00:16:46.06]
or whatever the length is,
how many of those unique?

[00:16:49.09]
Let me give you an analogy
'cause it's kind of abstract.

[00:16:52.05]
If I were to ask you how big, how many,

[00:16:57.01]
what's the size of the English language?

[00:17:00.06]
That could be a question.

[00:17:02.02]
And let's say I'm not
gonna give you a dictionary

[00:17:04.00]
so you can't use a dictionary.

[00:17:05.04]
I could have you read newspapers

[00:17:07.02]
and then just basically count
every time you find a new word

[00:17:11.05]
and give you lots of newspapers.

[00:17:13.02]
Let's say read every single
newspaper for the next year

[00:17:15.05]
and every time you find a new word.

[00:17:19.01]
Finding the word the is not
going to be really interesting

[00:17:22.06]
for you because it's gonna
be repeated many, many times.

[00:17:25.04]
You do wanna count it,

[00:17:26.04]
but you only wanna count it once

[00:17:27.08]
and then it's no longer
new and you move on.

[00:17:30.04]
Eventually, if you had
a big enough sample,

[00:17:32.04]
you'd be able to guess

[00:17:33.04]
at what the size of
the English language is

[00:17:35.01]
based upon something like a newspaper.

[00:17:37.08]
Obviously there are lots of words

[00:17:39.07]
that probably wouldn't
make it into a newspaper

[00:17:41.03]
'cause it's probably written
at a certain grade level,

[00:17:43.05]
but eventually you'd have
some kind of estimate.

[00:17:45.08]
So if we got rid of all of our duplicates

[00:17:48.06]
it's saying that we would have,

[00:17:50.04]
that would be 70% roughly of the,

[00:17:53.03]
in this case dataset, 395,000.

[00:17:57.06]
That's the percent of that
395,000 total sequences

[00:18:03.01]
that we would have left, roughly,

[00:18:05.07]
if we got rid

[00:18:06.06]
of the duplication.
- Really good point.

[00:18:09.09]
What a really nice way to think about it.

[00:18:15.03]
It's to search for unique words in,

[00:18:21.06]
or in this case, unique
sequences in a dataset.

[00:18:27.06]
You also shared,

[00:18:29.03]
or I shared this dataset
that I often ask students,

[00:18:35.00]
here's that dataset from
nanopore sequencing,

[00:18:38.02]
this technology where we have nanopores

[00:18:41.09]
and we string DNA through
these pore proteins

[00:18:46.08]
and are able to now interpret
those changes in current

[00:18:51.08]
and call them bases.

[00:18:54.07]
- [Jason] Yes.

[00:18:56.03]
- [Carlos] And this is one of the graphs

[00:18:58.05]
that we often get and share with students,

[00:19:04.06]

but it has, it's rich in data.

[00:19:08.09]
What do you think about this one?

[00:19:12.00]
- [Jason] All right, so yes, definitely.

[00:19:15.03]
I love looking at graphs like that.

[00:19:17.09]
And one of the things
that you also I think

[00:19:22.08]
pay attention into a graph
like this is its shape

[00:19:26.06]
that we see this sort of
big curve at the beginning

[00:19:29.09]
and then we see it sort
of ski slope off, right?

[00:19:34.02]
So I guess maybe if I was a statistician,

[00:19:37.05]
I could say this a bit more elegantly,

[00:19:39.06]
but we might be thinking of other things,

[00:19:43.01]
distributions that we might
have seen in basic stats class,

[00:19:46.02]
like the normal distribution

[00:19:49.01]
where there's certain
things that are average

[00:19:52.02]
and they end up in the middle

[00:19:53.07]
and there's certain things
that are really exceptional

[00:19:55.08]
and so they don't happen very often.

[00:19:57.08]
Like if I asked you

[00:19:59.01]
how many times have you
seen a person that is 5'7"?

[00:20:04.05]
You probably see people who
are 5'7" every single day.

[00:20:08.07]
If I ask you how often do you see an adult

[00:20:11.06]
that's around 3'5"?

[00:20:14.04]
Maybe once in your life.

[00:20:16.04]
How often have you seen
an adult that's 8'8"

[00:20:21.03]
or whatever the extraordinary height is.

[00:20:24.03]
And it's also probably pretty rare.

[00:20:27.01]
So average things are average

[00:20:29.05]
and other things are exceptional.

[00:20:31.04]
We don't see them very often.

[00:20:32.07]
And so in nanopore sequencing,

[00:20:34.05]
what I think we're seeing from this

[00:20:36.05]
is that we're asking the question,

[00:20:38.05]
if we ask how long a piece of DNA is

[00:20:40.07]
that we sample with nanopore,

[00:20:42.08]
we're asking how many times do we see

[00:20:45.02]
a piece that's any given length long?

[00:20:49.03]
And so there's definitely
an average somewhere.

[00:20:52.03]
If this is in kilobases, which it is,

[00:20:54.00]
somewhere around four to six kilobases

[00:20:59.05]
is average what we're seeing here.

[00:21:03.01]
The N50, which is a number,

[00:21:06.01]
that's an interesting
number in and of itself.

[00:21:08.04]
If we have time I can explain it

[00:21:09.09]
because it's not an average,
it's a little bit of a trick.

[00:21:14.03]
It's basically saying that

[00:21:16.07]
we have an average number of sequences

[00:21:18.09]
and most of them are around 5 kilobases.

[00:21:21.03]
What does that mean?

[00:21:22.01]
Well, when we prepare DNA
for nanopore sequencing,

[00:21:26.04]
and many cases, not every case,

[00:21:27.09]
but in most cases,

[00:21:29.02]
we are trying to get the

longest possible DNA reads.

[00:21:31.08]
But DNA can actually break
when you are pipetting,

[00:21:36.00]
when you're moving it around.

[00:21:38.01]
And so the more careful you are

[00:21:40.04]
and the type of DNA
extraction that you do,

[00:21:42.09]
the more likely you're to get long reads.

[00:21:45.02]
And oftentimes we want long reads

[00:21:46.08]
because rather imagine a puzzle piece,

[00:21:50.03]
a puzzle that had 20,000 pieces

[00:21:52.06]
versus a puzzle that had six pieces.

[00:21:54.03]
The one with six pieces is
really easy to figure out,

[00:21:56.09]
a jigsaw puzzle.

[00:21:58.01]
The one with 20,000 pieces,

[00:21:59.04]

it's a lot of work to figure it together.

[00:22:02.00]
And so bigger is better in this case,

[00:22:06.03]
but it reflects that
there's some sort of cutoff.

[00:22:09.04]
So if we look at the pieces of DNA there

[00:22:13.03]
that are really short,

[00:22:15.01]
you're getting some that are
less than a thousand like 0.13,

[00:22:21.07]
which is probably as short
as you can get a read

[00:22:24.00]
because any shorter than that,

[00:22:26.04]
then the adapter sequences,

[00:22:27.04]
there's nothing really that the DNA can,

[00:22:29.06]
the adapters that are required

[00:22:32.02]
to get the DNA through the pore.

[00:22:33.06]
And I know you might have
shared some visualizations

[00:22:36.02]
or some explanations.

[00:22:37.07]
It's not just the raw piece of
DNA that we get in our tube.

[00:22:40.04]
It's actually that DNA plus
some other DNA adapters

[00:22:44.00]
that we add to it.

[00:22:45.03]
So there's a lower limit.

[00:22:46.09]
In other words, it can't be
shorter than a certain size.

[00:22:51.05]
So we see a hump there.

[00:22:52.07]
So we see that the graph doesn't go off

[00:22:54.05]
to the left infinitely.

[00:22:56.02]
Then we see the average.

[00:22:58.02]
And then longer pieces
are increasingly rare.

[00:23:02.03]
And oftentimes when you do an experiment,

[00:23:03.09]
one of your goals is to push

this curve off to the right

[00:23:07.04]
as much as you could so
that you can get longer

[00:23:10.03]
and you want the average
to be above 10 kilobases

[00:23:15.06]
or above 20 is really great.

[00:23:17.03]
- [Carlos] Yeah, this
was actually a genomics,

[00:23:23.02]
a meta-genomics sample.

[00:23:24.09]
So a microbial community

[00:23:26.06]
and we got the DNA from this
compost sample by beating it,

[00:23:33.09]
beating the soil sample
with beads for a while.

[00:23:37.00]
So we sheared some of the DNA.

[00:23:42.04]
One other graph that pops up

[00:23:45.00]
and students and researchers in the lab

[00:23:50.02]
often have questions

about is this mystery.

[00:23:53.08]
How reliable are these sequences?

[00:23:56.04]
And we often get stuck on this graph

[00:23:59.03]
from nanopore sequencing.

[00:24:01.05]
What do you think about it?

[00:24:03.05]
- [Jason] Okay, so let's take a look

[00:24:05.04]
and see what we're seeing.

[00:24:07.00]
The first thing that we're seeing

[00:24:08.04]
is this Q or quality score.

[00:24:12.02]
And it does mention that
this is a PHRED scale.

[00:24:14.07]
So you gotta know all of a sudden like,

[00:24:16.02]
what are we talking about here?

[00:24:19.00]
So what we're talking about here,

[00:24:20.07]
there's something called the PHRED scale,

[00:24:22.06]
which is logarithmic.

[00:24:25.01]
And basically it says,

[00:24:26.04]
what's the chance that any given base

[00:24:29.06]
in a DNA read is an error?

[00:24:32.04]
So I don't have the chart in front of me,

[00:24:35.01]
but if you get a score of 10,

[00:24:36.07]
I think it's something like

[00:24:37.07]
you've got a 90% accuracy rating.

[00:24:40.05]
If you get a score of 20,
then you're up to sort of 90,

[00:24:44.00]
99% and then 99.9% and so on and so forth.

[00:24:48.01]
It's an interesting thing

[00:24:49.01]
because sometimes you might think

[00:24:52.02]
95 or 99% accuracy sounds pretty good.

[00:24:55.05]
I'd love a 99 on a test.

[00:24:58.01]
But then when you consider

[00:25:00.04]
how much sequence you're often looking at,

[00:25:02.03]
maybe millions or billions of base pairs,

[00:25:04.09]
I think there's a graph

[00:25:06.01]
or a chart at least out there somewhere

[00:25:07.07]
called if 99% were good enough

[00:25:10.03]
and then it tells how
many babies would go home

[00:25:12.03]
with the wrong parents every single day.

[00:25:14.04]
How many pieces of mail
would be delivered.

[00:25:16.02]
When you start to get
really, really large numbers,

[00:25:20.09]
the decimal points matter
in terms of being 99.999.

[00:25:27.04]
So here we've got a quality score

[00:25:29.04]

and a cutoff

[00:25:32.00]
where I guess they're sort
of making the intimation,

[00:25:34.01]
if I recall, so if I recall,

[00:25:35.06]
when you do base calling on nanopore,

[00:25:38.08]
which is actually getting
better than this graph reflects,

[00:25:42.01]
but when you do a base
quality score on nanopore,

[00:25:44.08]
if you get below an eight,

[00:25:46.07]
so now we're below 90% accuracy,

[00:25:50.09]
it doesn't even register.

[00:25:53.01]
It doesn't,

[00:25:53.09]
it won't really call that
as worth even looking at.

[00:25:57.00]
The way that nanopore often works is that

[00:25:59.06]
if you're getting above an eight

[00:26:00.07]
and you're usually hopefully
somewhere closer to,

[00:26:02.07]
like I said now in the twenties,

[00:26:06.05]
that means that well,
any given base might,

[00:26:10.09]
any given read might
have some errors in it,

[00:26:13.03]
but if you get enough of the reads,

[00:26:14.05]
you can average them out.

[00:26:16.00]
So this is a way of
showing the distribution.

[00:26:18.05]
I guess it's saying that the mode

[00:26:20.03]
which is giving us the most
frequent quality score.

[00:26:22.07]
Here, it looks like it's
a little bit above 12.

[00:26:25.00]
But the other thing that's
kind of interesting here,

[00:26:27.00]
if I see the bottom

right, it's on the time.

[00:26:29.06]
So in other words,

[00:26:31.01]
maybe this is the time of
running the experiment.

[00:26:33.03]
I think it might be in minutes.

[00:26:35.02]
And I think what it's sort of suggesting

[00:26:36.08]
is that as you run the experiment

[00:26:38.09]
longer and longer and longer,
the quality starts to dip off.

[00:26:43.05]
And if that's what I'm seeing,

[00:26:45.09]
part of the explanation for that,

[00:26:47.03]
I don't know the context of
this particular experiment,

[00:26:50.07]
but part of the
explanation of that is that

[00:26:53.00]
as you run longer and longer,

[00:26:55.07]
there's actually enzymes

[00:26:57.01]
and proteins that pull the nanopore,

[00:27:00.05]
the DNA through the nanopore

[00:27:02.07]
and those start to either
need more refueling,

[00:27:05.04]
you have to actually add ATP, add fuel,

[00:27:08.09]
or they can become clogged
or other things to happen.

[00:27:11.07]
They're biological molecules
so they can actually degrade.

[00:27:15.04]
So I think what we're seeing here

[00:27:17.02]
is that we get an average quality score.

[00:27:19.08]
Sometimes it's a little
bit higher or lower,

[00:27:21.06]
which is the blue, which
is what you'd expect.

[00:27:24.05]
And then as you go on in
time at a certain point,

[00:27:28.03]
the quality may degrade.

[00:27:29.07]
And then that's,

[00:27:30.06]
if you were running experiment
at that 60 minute mark,

[00:27:33.01]
probably you would wanna stop
the experiment and refresh it

[00:27:37.00]
or do other things to try to see if

[00:27:38.05]
you can at least keep the average back up.

[00:27:40.08]
- [Carlos] Yeah, this was actually

[00:27:43.08]
a long run part of the previous dataset

[00:27:46.07]
and this was, I think we did three days.

[00:27:51.03]
So this is in the hours
and we noticed this large

[00:27:56.01]
drop at the end.
- Ah, okay.

[00:27:56.09]
That makes a lot of sense.

[00:27:58.08]
- [Carlos] And people were concerned.

[00:28:00.02]
But your point about this

having enzymes and needing fuel

[00:28:09.03]
is really an important one to keep in mind

[00:28:12.00]
because after a while the
system loses energy basically

[00:28:17.01]
to continue sequencing.

[00:28:19.09]
- [Jason] Yeah, you'll lose energy.

[00:28:23.00]
The nanopores themselves can be clogged

[00:28:26.03]
or they can stop functioning

[00:28:27.07]
'cause they are biological molecules.

[00:28:29.03]
So it's,

[00:28:31.08]
like running, at a certain
point you're gonna get tired.

[00:28:34.08]
- [Carlos] You get tired.

[00:28:35.09]
Oh, I like that analogy.

[00:28:39.05]
And similar to that,
we have another graph,

[00:28:44.09]

the same dataset and now
you can see the Y-axis.

[00:28:50.06]
- Hours and minutes.

[00:28:51.05]
I see, yep.
- Hours and minutes.

[00:28:53.02]
This one was similar, but it's the speed.

[00:28:56.08]
- [Jason] Correct.

[00:28:58.04]
And it's quite related to the other graph.

[00:29:02.02]
There's something called
the translocation speed,

[00:29:04.07]
which is how fast, how
many bases per second is,

[00:29:08.04]
are moving through the nanopore.

[00:29:11.07]
And so this again is a reflection of the,

[00:29:16.06]
I mean maybe I,

[00:29:17.08]
it's a little bit tough to call it biology

[00:29:19.08]
because we're not talking
about a living system,

[00:29:21.07]
but we are talking about
biomolecules, proteins.

[00:29:24.03]
So that is what's happening here as well.

[00:29:27.05]
It turns out that in many cases actually

[00:29:30.04]
that one of the technologies,
and I'm not the expert here,

[00:29:32.09]
but one of the things
that the nanopore company

[00:29:35.08]
and the developers of the
technology had to deal with

[00:29:38.03]
was actually making sure

[00:29:39.04]
the DNA doesn't go through too fast.

[00:29:41.09]
Because if the DNA actually goes through

[00:29:43.05]
the nanopore too fast,

[00:29:45.04]
it would make more mistakes

[00:29:48.06]
because it's going through

[00:29:49.08]

faster than it can actually be read.

[00:29:53.00]
Of course if it, the faster
you can get the DNA through,

[00:29:56.00]
the more DNA you could
read in a given time,

[00:29:57.09]
so you get more nucleotides called,

[00:30:01.00]
but you kind of have that trade off.

[00:30:03.00]
And so--
- Yeah, it's balance.

[00:30:05.02]
- [Jason] Yeah, yeah.

[00:30:06.00]
So it's a balance

[00:30:06.08]
and they've gone back and forth

[00:30:08.06]
and they actually, when they design,

[00:30:10.09]
I think there's another graph,

[00:30:12.05]
maybe you'll have to look it
up, but it definitely exists.

[00:30:15.09]
I know I've seen it on the
YouTube nanopore channel

[00:30:19.02]
where they have, they actually perform,

[00:30:22.02]
I guess you would call directed evolution

[00:30:25.02]
combined with machine learning

[00:30:26.06]
to try to come up with new pore shapes

[00:30:31.07]
and ask the question with this pore shape,

[00:30:34.05]
what happens on an XY plot?

[00:30:36.06]
Like one pore shape or one
configuration molecules

[00:30:40.03]
might increase the speed
but decrease the quality.

[00:30:43.02]
Another one might be decreasing the speed

[00:30:45.09]
and therefore the throughput
and the quality goes up.

[00:30:48.09]
And the question is,
where is the sweet spot

[00:30:51.06]
that you can have both
a lot of nucleotides

[00:30:54.09]
and a lot of high quality?

[00:30:57.00]
But this is reflect in the biology.

[00:30:58.07]
Although I think you see that

[00:31:00.07]
the translocation speed
doesn't change as much,

[00:31:03.04]
but the drop-off doesn't
look as significant.

[00:31:05.03]
Although my eyes, you know.

[00:31:07.00]
I don't know if it's really comparable

[00:31:08.08]
because we're not looking at
a log scale on left hand side.

[00:31:12.00]
But I think that that same time point,

[00:31:13.09]
we see that the speed starts to change.

[00:31:16.07]
It starts to, looks like it's going down.

[00:31:19.05]
And then that's associated
with the quality changing.

[00:31:21.08]
Again, probably because

things are breaking down

[00:31:24.03]
and not working.

[00:31:25.06]
- [Carlos] And along those lines,

[00:31:27.04]
I noticed that Nanopore keeps
on improving the chemistry

[00:31:32.07]
and between version 12 and version 14,

[00:31:36.07]
they, I believe they changed
the translocation speed

[00:31:40.05]
and the temperature of the reaction.

[00:31:42.03]
[Jason speaks over Carlos]
- Yes, the temperature,

[00:31:44.04]
things that you all know
from your basic biology,

[00:31:46.09]
temperature, ATP enzymes, right?

[00:31:52.07]
The pH.

[00:31:53.05]
All of those different things
they toy with and tinker with.

[00:31:56.08]
And you could see that

happening in the graphs.

[00:31:58.08]
And actually, although many
people, myself included,

[00:32:01.08]
don't always look at these graphs,

[00:32:03.01]
they're actually things that
you should be looking at

[00:32:06.07]
while you're sequencing

[00:32:07.07]
and you can actually intervene

[00:32:10.01]
to actually recover the sample
and put a new flow cell,

[00:32:13.02]
add some more fuel.

[00:32:14.03]
- Add some fuel maybe
- Et cetera.

[00:32:15.08]
- [Carlos] When you start seeing.

[00:32:17.01]
- So yeah, those are all things,

[00:32:18.01]
or really know when it's time to quit

[00:32:19.07]
because it's not gonna
improve no matter what you do.

[00:32:21.06]
So all of those types of things

[00:32:23.05]
are reasons why we wanna
pay attention to the graphs.

[00:32:26.06]
And I'll comment,

[00:32:27.07]
although we don't have
an animation here that

[00:32:29.06]
many of these are animated in real time.

[00:32:31.07]
So actually as you're doing experiment,

[00:32:33.09]
you can be looking at the graphs

[00:32:35.01]
to make a decision about
do you need to do anything?

[00:32:41.03]
- What kinds of scientific
questions interest you the most,

[00:32:46.04]
Jason?

[00:32:47.02]
- That's a very deep question.

[00:32:49.07]
I don't know if I've thought about it,

[00:32:52.06]

but I just--
- With nanopore maybe?

[00:32:54.04]
- Okay, there we go.

[00:32:55.02]
We can, like we can lower it to, you know,

[00:32:58.01]
narrow it down to something.

[00:33:01.02]
Hmm. Let's see.

[00:33:02.07]
So I think it is just absolutely cool

[00:33:05.05]
to be able to sequence things real time.

[00:33:08.03]
To be able, in a few minutes,

[00:33:11.03]
going from some sample of interest

[00:33:16.04]
to actually seeing the DNA sequence.

[00:33:18.09]
So I'm really interested
in making that process.

[00:33:23.00]
I'm interested a lot
actually in the biochemistry,

[00:33:26.01]
in the lab side of it,
of making that optimal.

[00:33:29.06]
But I also love the data side too,

[00:33:31.07]
of how can you help people
to visualize datasets

[00:33:34.08]
and how can you help people

[00:33:36.05]
to understand what's in the data

[00:33:38.04]
and what the data could be used for.

[00:33:40.01]
Because I may have one question,

[00:33:41.04]
but somebody else may be able
to use that same data set

[00:33:44.03]
to look at other things.

[00:33:46.04]
So yeah, there's a lot that interests me

[00:33:49.07]
and I just love new toys.

[00:33:51.03]
I have a kitchen full of toys.

[00:33:53.08]
Every time something's out, I
have to restrain myself, but.

[00:33:58.01]
- Well, I'm just.
- Yeah, I like,

[00:33:59.04]
I like questions that involve
new ways of looking at things.

[00:34:02.00]
Let's say it that way.

[00:34:03.04]
- Oh, I like that.

[00:34:09.02]
How and when did you know
you wanted to be a scientist?

[00:34:14.01]
- I don't know.

[00:34:15.08]
I know I've always loved science
since sort of earliest age.

[00:34:21.00]
So I guess I've always
been interested in science.

[00:34:26.05]
One of my earliest science memories,

[00:34:28.05]
I think I could say was a pop-up book,

[00:34:32.06]
and it was a pop-up book about weather.

[00:34:35.00]
And I do remember my favorite
page in the pop-up book

[00:34:40.00]
was a 3D pop-up hurricane

[00:34:43.05]
where you would see like
the walls of the hurricane

[00:34:46.02]
and you could like twist it

[00:34:47.05]
and you could see how the eye rotates

[00:34:49.04]
and like you got the
different layers of clouds

[00:34:52.02]
and stuff like that.

[00:34:53.05]
So I've always been interested in science,

[00:34:55.09]
and I guess that's a form
of visualization too, right?

[00:34:58.05]
You know, being in a hurricane,

[00:35:00.02]
it's just gonna be wet
and windy and complex.

[00:35:04.08]
And you're probably
worried about other things,

[00:35:06.05]
but being able to appreciate
it in that case through 3D,

[00:35:12.02]
and something you could see and touch,

[00:35:14.02]
that was always interesting.

[00:35:15.09]
- That's really interesting
that you love the

[00:35:18.06]
3D representation and also tactile.

[00:35:22.03]
And there's some math

[00:35:24.06]
and modeling there too.
- Absolutely.

[00:35:28.01]
Correct.

[00:35:29.00]
- So that may align with what
led you to your current job.

[00:35:33.08]
How did your popup book on hurricane

[00:35:38.03]
lead you to this position?

[00:35:42.00]
- Good question.

[00:35:42.08]
I'm not, I don't know.

[00:35:45.00]
I just obviously took.

[00:35:48.02]
Nothing is planned.

[00:35:49.05]
I took lots of science.

[00:35:51.00]
Anytime I could take science courses

[00:35:52.07]
or be parts of science clubs
or science competitions,

[00:35:55.04]
it was always something
that I was interested in.

[00:35:58.08]
Probably the most influential
thing, I guess I would say,

[00:36:01.05]
and again, not being planned is,

[00:36:03.04]
from the moment I was in
an undergraduate situation

[00:36:08.02]
I worked in a lab.

[00:36:10.08]
I was like, oh, I wanna work in a lab.

[00:36:12.01]
My idea was in the building,

[00:36:14.00]
the science building I went
to at Stony Brook University,

[00:36:17.00]
there were five floors.

[00:36:18.02]

And I thought to myself, well,

[00:36:19.04]
I'll work on one floor every single year.

[00:36:21.02]
Like, start from the top
and work my way down.

[00:36:23.00]
- That's--
- I never left.

[00:36:24.08]
I never left the fifth floor,
which was Ecology & Evolution.

[00:36:28.07]
So I didn't quite follow with that plan,

[00:36:31.02]
but I just always worked in a lab

[00:36:33.06]
and so that made it, when I
came to Cold Spring Harbor

[00:36:37.00]
and naturally like, oh, I
wanna continue working in labs.

[00:36:40.05]
I've been here for a little over 18 years.

[00:36:43.00]
So I also kind of stay in places

[00:36:45.03]
and just sort of sit and do science

[00:36:47.02]
and it's a great place to do science.

[00:36:48.06]
So just being in the lab

[00:36:51.03]
and being in the middle of whatever it is.

[00:36:53.06]
Obviously you take
classes and you do things,

[00:36:56.04]
you have to go on interviews at some point

[00:36:58.06]
or talk to people.

[00:37:00.01]
But I think, you know, I was,

[00:37:02.00]
I didn't need anybody to tell
me this is what I was doing.

[00:37:04.01]
I almost felt it from the beginning.

[00:37:05.05]
Like, oh yeah, of course
I'm gonna be doing this, so.

[00:37:07.09]
- That's awesome.

[00:37:09.01]
I did not realize it was 18 years.

[00:37:13.02]
- Yes. Carlos.

[00:37:14.06]
- Whoa. Cool.

[Carlos speaking over Jason]

[00:37:17.00]
Here's a tough question.

[00:37:18.07]
What is it like being a
scientist who identifies as what,

[00:37:23.09]
how would you identify
yourself as a scientist?

[00:37:27.04]
- Weird.

[00:37:29.04]
So yeah, that was my answer.

[00:37:31.08]
- Weird.

[00:37:32.09]
- Weird.

[00:37:34.03]
Yeah, I definitely always enjoy humor,

[00:37:40.00]
always enjoy a mix of things,

[00:37:41.06]
whether it's music and science and art.

[00:37:47.01]
- I love how--
- What else does it?

[00:37:49.02]
- I love how in your
workshops and webinars,

[00:37:53.01]
you always bring in
really great connections

[00:37:56.02]
like art and food,

[00:37:59.01]
and I've seen you do this with students

[00:38:03.04]
and also with like webinar with faculty

[00:38:10.02]
and instructors or computer scientists.

[00:38:12.02]
So I love how you bring
in different things.

[00:38:16.03]
So what does a typical
day in the life of Jason?

[00:38:19.08]
- I wish, I also wish I knew
the answer to that question.

[00:38:22.06]
It would help me a lot.

[00:38:25.09]
A lot of it is really talking
to people like you, Carlos,

[00:38:28.06]
which is great.

[00:38:31.02]
A lot of it is,

[00:38:32.07]

a lot of the work that
I do is people work.

[00:38:35.05]
Making connections.

[00:38:36.07]
Like you said, whether it's
conceptual connections,

[00:38:39.00]
but also connections with people

[00:38:40.09]
who are doing different projects.

[00:38:43.07]
I've had the privilege
at Cold Spring Harbor

[00:38:45.05]
to really be able to teach

[00:38:47.09]
and work with groups all over the country

[00:38:49.07]
and all over the world.

[00:38:50.09]
So oftentimes you realize

[00:38:52.03]
what somebody's doing somewhere else

[00:38:53.08]
is relevant to what somebody's
doing in another corner.

[00:38:57.03]
And so being able to think that way,

[00:39:01.04]
it's very interesting because
at the learning center,

[00:39:03.09]
although we have lots of visitors,

[00:39:05.01]
we don't have students
like a university does.

[00:39:07.07]
So it almost is a great combination

[00:39:10.01]
because it means that everything we do

[00:39:11.06]
is designed to go outward

[00:39:13.02]
and to be communicated to people and to,

[00:39:15.07]
but the fact that we don't
have our own students,

[00:39:17.07]
that we just focus inwards,
we are all focused outwards.

[00:39:20.07]
So, a lot of my day is about
interacting with groups

[00:39:24.06]
and linking it to what we
do at the learning center

[00:39:26.06]
and hopefully linking them as well.

[00:39:28.04]
- Oh, and I love how you all produce

[00:39:33.04]
and share resources and
also make connections.

[00:39:38.00]
So tell me about the
moment when you felt like

[00:39:41.09]
you really belonged in
the field of science.

[00:39:44.04]
Was it that pop-up book

[00:39:46.00]
or is it that day in Cold Spring Harbor?

[00:39:48.06]
- Interesting.

[00:39:49.04]
Yeah, I'm not sure.

[00:39:50.02]
I know like from a young
age, I had a lab coat.

[00:39:52.03]
I certainly went and just bought lab.

[00:39:55.07]
Of course, I love wearing a
lab coat and I wear it now

[00:39:57.09]
because then you can,
even if your t-shirt,

[00:40:00.08]
[audio warbles]

[00:40:02.00]
you've gotta just throw a
lab coat on and you're fine.

[00:40:04.01]
You could sit in a room.

[00:40:08.00]
That's a question that happens,

[00:40:10.06]
that question or that
identity happens many times,

[00:40:13.06]
I would say, where you realize, ooh,

[00:40:15.08]
I get to do science for a living.

[00:40:17.05]
So it happens, realizations.

[00:40:19.03]
I think maybe one of the ones I remember,

[00:40:20.09]
I don't remember where I was going,

[00:40:21.08]
but sometimes sitting on a plane

[00:40:24.08]
and going somewhere

[00:40:25.06]
and realizing that people
have invested their time

[00:40:28.01]
and energy in having me come visit

[00:40:30.02]
and I'm like, oh, this is cool.

[00:40:32.01]
And like, people here, which
is, you know, rare and unusual.

[00:40:37.09]
So I think that's certainly a moment that,

[00:40:39.08]
and it's many, many moments
of feeling like you belong

[00:40:44.00]
in the sciences.

[00:40:45.04]
- And those people are really lucky

[00:40:47.02]
because you always do a great job

[00:40:51.07]
getting us energized and
interested in new topics.

[00:40:58.00]
Thank you for all you do, Jason.

[00:41:00.07]
- Awesome, thank you
Carlos, for all you do.

[00:41:02.09]