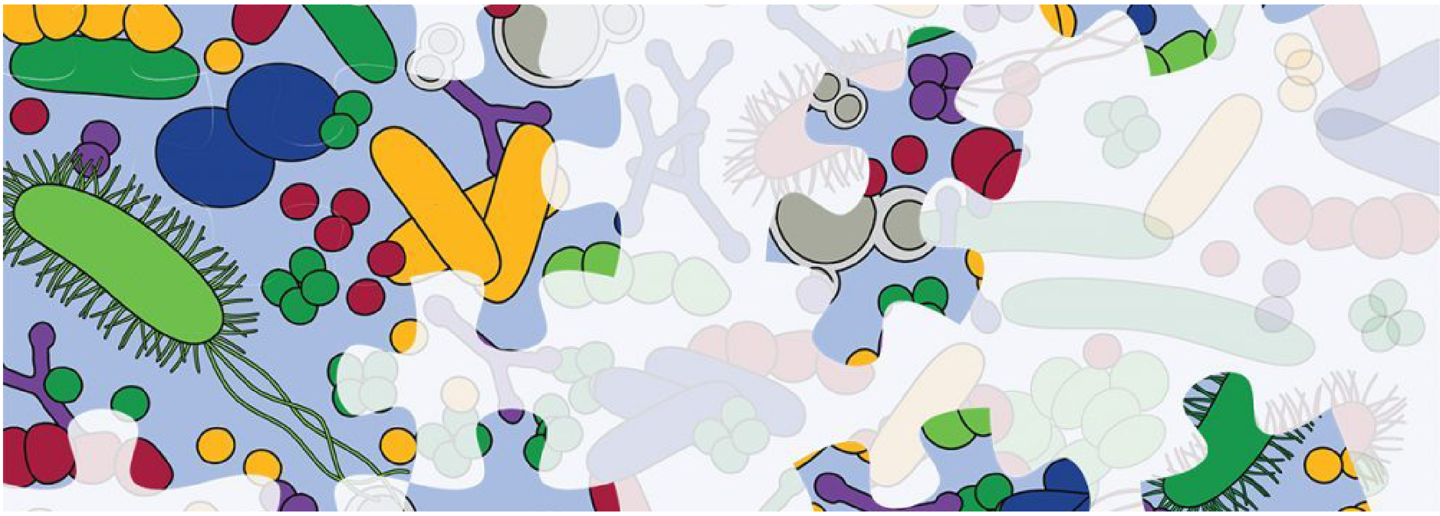
Science in School
The European journal for science teachers

ISSUE 51 – 15/01/2021 Topics Biology | Mathematics



Microbial genome puzzles

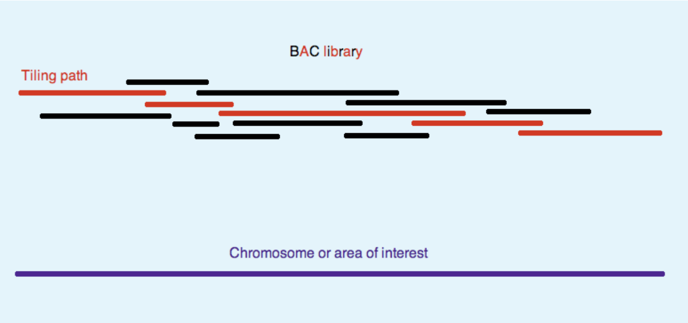
Eleonora Mastrorilli

How do scientists piece together genomic information from sequencing data? Play these two fun online puzzles to find out.

Reconstructing genomes is like solving puzzles

Characterising the makeup of the many microbes that live on and inside us helps us understand the possible roles those microbes play in human health and disease. Scientists use DNA sequencing technologies to analyse the genetic material of microbial samples to identify the microbial species present.

Rather than providing full genome sequences of the organism(s) in a microbial sample, DNA sequencing produces hundreds or thousands of short, linear pieces of microbial DNA. To try identifying the organism(s), scientists often use these DNA fragments and combine them into contiguous fragments of DNA (contigs) using computational approaches.



Since it isn't possible to sequence very long pieces of DNA, isolated DNA is split into fragments to make a collection called a library, in this case using bacterial artificial chromosomes (BACs). In addition to the sequence of interest, the BACs contain elements that allow these fragments to be copied and sequenced. This gives a collection of overlapping sequence fragments that can be aligned using computers to reconstruct the original sequence.
Becchamm – Own work, CC BY-SA 3.0.

This educational resource guides teachers how to use a simple "puzzle" metaphor to introduce students to the concept of genome reconstruction of single bacteria and complex microbial communities.

Scientific introduction

DNA sequencing determines the nucleotide acid sequence of an organism's unique hereditary information. As output, it generates hundreds or thousands of short, linear pieces of microbial DNA, which are fragments of the full DNA genome. The next step after DNA sequencing, therefore, involves combining (assembling) those fragments into contiguous fragments of DNA (contigs) using computational approaches.

Genome reconstruction of a single bacterium

The genome of a single bacterium is generally:

- circular
- double-stranded
- of variable length, but generally in the order of million base pairs in size

Commonly used DNA sequencing technologies (applying so-called 2nd generation sequencing) generate pieces of DNA that are:

- linear
- single stranded
- short (35 to 400 base pairs)

Therefore, you can think of the task of genome reconstruction as a somewhat "hard" puzzle problem: we need to rebuild a whole image from its pieces.

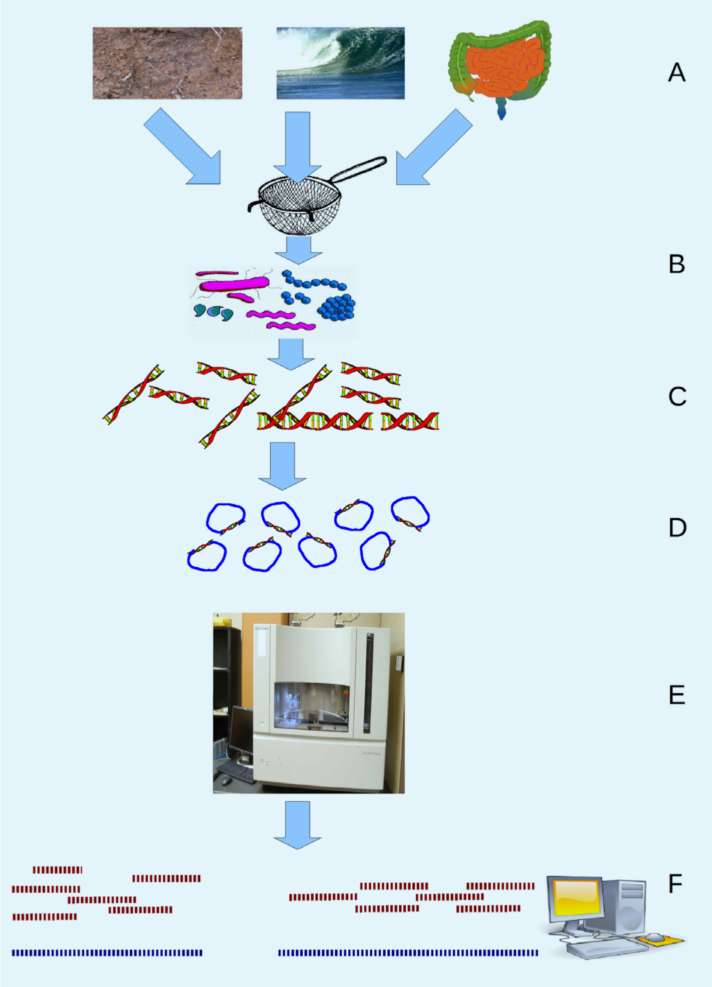
How do we do that exactly when aiming to reconstruct the genome sequence of a single bacterium? In the most straightforward case, our organism has already been sequenced and its genome sequence has been deposited in a public repository (such as the EMBL-EBI's European Nucleotide Archive, ENA). In this case, we can use this sequence to help us rebuild the "puzzle", similarly as you would do by looking at the image on the cover of the puzzle box. This approach is called "mapping" – identifying where a specific piece of DNA comes from by comparing it to a known reference.

Remember that this is obviously a simplistic approach: due to their very high mutation rate, hardly ever is the genome of a sequenced bacterium absolutely identical to that of the reference genome. We must therefore be ready to accept that the mapping will not be perfect, and that the mismatches themselves, if sufficiently proven, might be the most interesting spots in the genome.

Genome reconstruction of complex microbial communities

What are the added challenges when reconstructing the genomes in a complex microbial community such as your gut microbiome?

1. There are multiple genomes mixed together.
2. We do not know which sequence belongs to which genome.
3. We do not have a reference genome to help us rebuild the genome of every single bacterium in the community.
4. Even if the sequences have a certain "depth" (i.e. we collect many pieces of the puzzle), we probably have not collected all sequences (i.e. we might remain with missing pieces of the whole image).



Environmental Shotgun Sequencing (ESS). (A) Sampling from habitat; (B) filtering particles, typically by size; (C) DNA extraction and lysis; (D) cloning and library; (E) sequence the clones; (F) sequence assembly.

John C. Wooley, Adam Godzik, Iddo Friedberg, CC BY 2.5, via Wikimedia Commons

To solve this more complex problem, there are several strategies that, once more, resemble what you would instinctively do with a puzzle:

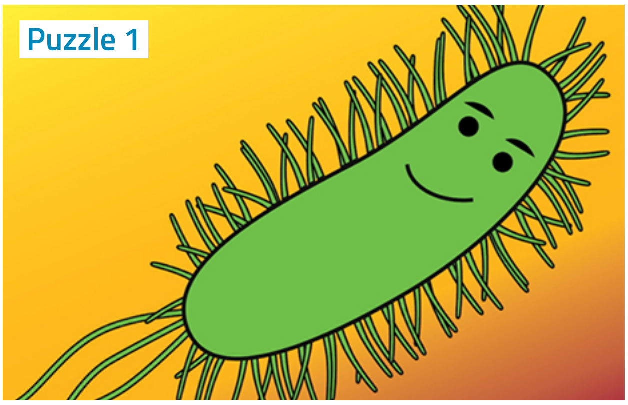
- If there is some piece of the puzzle we have a reference for, we build that first.
- If any of the pieces look "alike" (i.e. they are genetically similar, or, in the puzzle metaphor, they maybe have the same colour or pattern), we group them together.
- If any of the pieces fit together very well (in the bioinformatics jargon, they can be "assembled" together), we assume they belong together.
- If any of the pieces has a known function (in the puzzle metaphor, the corner or border pieces), we try to infer where they belong.

The problem of missing data

As mentioned, however, we might not have all the pieces we need to fully reconstruct the image. Since this image is the starting point to then investigate the bacterial composition in the sample (who is there) and subsequently their possible function (what they might be doing), take a second to think about the impact of the missing part of the data: aside from hampering a complete understanding of the microbial community, we must also understand that we can describe what we see, but we cannot claim any meaning from what we don't see. Simply put, if I grab a few socks from my drawer and none of them is red, I cannot conclude that I have no red socks. Why? The overall complexity of the microbial community is too high for our sampling capacity; therefore, we will end up with missing data.

Instructions

In this educational resource, the microbial genome puzzles are used as metaphors to illustrate how researchers move from raw genetic data produced by DNA sequencing to an overview of the genome of a single bacterium (Puzzle 1) or of a complex community such as a microbial sample of the gut (Puzzle 2).



The puzzle activity consists of two online puzzles: the first one shows a cartoon version of an Escherichia coli bacterium and can be used as a metaphor for single genome reconstruction, the second one is a cartoon representation of multiple microbes and can be used as a metaphor for the reconstruction of the genomes of a microbial community. You can access the puzzles below:

Puzzle 1: single microbe
Puzzle 2: microbial community

Please find below an outline of a possible approach of embedding the microbial genome puzzles in your genomics lessons.

1. Introduce the concepts of the bacterial genome, genome sequencing and genome reconstruction of a single genome.
2. Introduce the puzzle metaphor by asking the students to do Puzzle 1. Students can use the reference image to help them puzzle and can keep track of the time that is needed to complete the puzzle.
3. Introduce the concept of microbial communities and discuss the challenges of reconstructing microbial genomes in this case by asking the students to identify similarities and differences to reconstructing a single genome.
4. Guide students through the challenges described in points 1-4 in the "scientific introduction" section.
5. Discuss with students how they would solve the puzzle without reference image.
6. Ask students to do Puzzle 2. This time, students should complete the puzzle without looking at the reference image and compare the time it takes them to complete the puzzle with their time to solve Puzzle 1.
7. Discuss with the students how solving the puzzle would be complicated further if they were not given all the pieces of the puzzle? Guide the students to mention that a) they might misplace a piece and b) they might not identify one (or more) of the microbes in the community.

8. Start your conclusion by pointing out that the extra time that is needed in reconstructing complex community genomes compared to single bacterial genomes is called "computational complexity". Computational complexity describes the increase in complexity which is due to the increased need in time and computational power (e.g. more server power than manpower) required to retrieve the composition and function of a complex community compared to that of a single bacterium.
9. Finish concluding by pointing out that putting the puzzle together, i.e. reconstructing a somewhat trustworthy image of the microbial community genomes is just the first step in the analysis of the microbiome! Once the genomes are reconstructed, we can identify the bacterial species in the sample. In the next step, those communities can be described and characterised in detail, forming the basis of addressing research questions about the microbial communities and their interaction with us humans.

Acknowledgements

This article was adapted from the one originally published as EMBL-ELLS Teaching Material.

Resources

- Crack the genetic code with this teaching activity.
- Read about a study tracking the evolution of the bacterium *Escherichia coli* over 67 000 generations.
- Learn about the discovery of the DNA structure and how to extract it from onions and peas in your classroom.
- What are cancer mutations and how are they identified? Discover the answers in this teaching activity.
- More resources on the human microbiome.