

Assembly

Learning objective: Become familiar with some of the basic mechanics underlying *de novo* assembly.

Level of Difficulty: 3/5

Once you've got a filtered set of reads, you may decide you want to piece them together to try to reassemble the genome(s) they came from. This process is kind of like if you were to piece back together a newspaper that had been run through a shredder. What strategies might you use while reassembling? Say you've taped back together as much as you can of the original newspaper, but because the pieces were so small and the shredder ate some of them, you're left over with a bunch of patchy sections, and a bunch of repetitive words and characters like "a", "the", and commas. The sections of assembled pieces are like metagenomic "contigs", and those repetitive leftovers may still be useful down the road in read mapping.

To do: A few copies of two common nursery rhymes representing different microbial genomes have been cut up into pieces, and some of the pieces may be missing. Try to reassemble as much of the rhymes as you can from the following list of "reads". Note that a few areas of the rhymes have repeated words, and both rhymes loop back on themselves at the ends, kind of like a circular genome.

Follow up questions:

- How are ATCGs different from words in a newspaper or song? What context clues in DNA might be used instead of common words, sentence structure, punctuation, and font?
- What other clues in the DNA might guide a metagenomic assembly?
- Why do you suppose some pieces might be missing in metagenomic assembly?
- How might repeated genes affect the assembly process?

-vous? dorm

? Frere J

am row r

atines son

cques Fre

dream r

e is but

eam ro

eam row r

errily merrily

es dorm

es matin

g dong d

g dong Fr

ines ding d

ing din

ing don

m merrily merr

merrily mer

merrily mer

ng ding

ng ding d

nnez les

onnez le

ous? Dor

ous? Fre

ous? sonn

r boat ge

rere Jacq

rilly merrily

rmez-vo

rmez-vou

row ro