From the Frying Pan to the Fire: Bioinformatics Computing in Two Lectures

Richard Tillett
Nevada Center for Bioinformatics
rltillett@unr.edu
BCH 709: Intro to Bioinformatics Fall 2015

Overview

- Recap of last lecture
- Find out what we assembled
- Dig into the assembly to answer specific questions

Recap of last time

- Learned a lot of unix commands
  - And used many of them
- Learned how to use the UNR grid
  - SGE commands like qsub
- Fed mystery Illumina sequences to the SPAdes assembler with the grid

- Now let's return to the grid

Back to the grid

- Execute Putty.exe
- Type “login.research.unr.edu” into the “Host Name” box
- Type “22” into the “Port” box
- Click “Open” button
- Input NetID as user name
- Input NetID password as password

A new folder?

- $ cd ~/scratch
- $ cd bch-709-intro-bioinformatics-2015f
- $ ls

- Is there a new directory here? Called spades_output? This has your results!
- $ cd spades_output

What's inside?

- $ ls –lh
  - These options show you the file sizes in human readable form
- We see a lot of files, only one of which concerns us
- contigs.fasta
- Check it out
  - $ less contigs.fasta

contigs.fasta

- This contains all of the assembled contiguous sequences generated last week!
- How big is the first sequence?

- How many sequences did you get?
  - Exit less by pressing q
  - $ grep –c '>' contigs.fasta
- So, what did we sequence?

Today's first puzzle

- What did we sequence?
- Can we make an educated guess using any tools we already know?

- By sequence homology to known sequences
  - NCBI's blast website
  - Transfer file over to PC
  - Open in notepad
  - Copy / paste
  - Compare results

Obtaining the file

- Our contigs.fasta is probably too large to safely open in notepad.exe as-is
- Let's cut it down on the command line and then sftp it to ourselves
- Cut it to 10 megabytes like this
  - $ head -c 10MB contigs.fasta > 10meg.contigs.txt
- WinSCP it with the app on your windows desktop

Obtaining the file

- Click WinSCP on your desktop
- Type “login.research.unr.edu” into the “Host Name” box
- Type “22” into the “Port” box
- Click “Open” button
- Input NetID as user name
- Input NetID password as password
- Navigate to the spades_output folder
- Download 10meg.contigs.txt to your Desktop

NCBI's web blast

- Open a web browser and go to www.ncbi.nlm.nih.gov
  - Click on blast, then nucleotide blast
- Open 10meg.contigs.txt with notepad.exe
- We will highlight a sequence, copy it, and paste it into the box at ncbi
- Let's try to use different sequences, each
- And compare results!

What did we sequence?

- Did we all get similar results? For the sequences we tested?

- What was the organism?
- Can we get any more specific?
- How much of it % might we have captured?
  - Napkin math by file size

Today's 2nd puzzle

- Let's suppose the original Illumina sequences came from a medical setting
  - Note: they did not actually. HIPAA & medical ethics discourage this

- Patient is a young boy
- with symptoms matching Sickle Cell Disease
- And family history / ethnicity in which SCD is known to occur
- Does he have sickle cell disease? Can we find evidence within our contigs?
- Let's explore SCD on ncbi to learn more

Ncbi quick researching

- Search “sickle cell disease”
- Look at the OMIM pages
- Identify the gene
- Identify the causal mutation

- Our goal: get a sequence to test vs. our contigs

Obtaining the sequence

- Find the cDNA sequence for the gene in question
- Click Fasta in the upper left, Select it, copy it
- Move back to putty window
- Navigate to the dbs folder
- Type nano
- Paste in the copied sequence
- Ctrl-x and save as a new file ‘normal.q’

Turning contigs into a blast db

- We included the binaries for ncbi standalone blast in the git package we've been using
- The command to turn any fasta format set of sequences into a blast db is `makeblastdb`
- We also premade an sge script that executes `makeblastdb`
- Navigate to day-2/sge
- $ qsub ncbi-makeblastdb.sge
- Db made at new blast_db folder

Blasting our gene vs. our db

- In the day-2/sge dir, we need to edit the other .sge file
- It was pre-built to blast a ‘mutation.q’ file, but we want to blast ‘normal.q’
- $ nano ncbi-blastn.sge
- Scroll all the way down until you see “mutation.q”
- Change it to normal.q, exit and save changes

Blasting our gene vs. our db

- $ qsub ncbi-blastn.sge

- Our results will show up in new dir blast_output
- $ cd ../../blast_output
- $ ls
- $ less blast_out.txt

What do we see?

- This is query anchored blast results
- Dots show identical matches
- Letters indicate differences
- Names indicate names of db sequences corresponding to the matches

- Do we see the mutation where we expected it?
- Does that mean we can positively state we can diagnose? Discuss.

Discuss

- Did we see mutation at the location one expects?
- Can we diagnose?
- Why? Why not?
- What is required for SCD? Have we conclusively proven it? What would?

What have we done in 2 days?

- Assembled, de novo, Chromosome ___ from organism _____
- Used ncbi's web blast to reasonably convince ourselves of that
- Interrogated our assembly to test a medical hypothesis of clinical significance
- Obtained an answer that directs us to the next required experiment for definitive proof