The welcome lecture

Challenge #1:

Most biologists still don't know much about computational science.

- Among many biologists, there is a general fear or skepticism of computers.

- This leads to shallow thinking about computational science.

Challenge #2:

Most computational scientists still don't know much about biology.

- Extant computational solutions may not use appropriate heuristics, or default parameters.
- "It works on my data…", but their data != yours!
- Solutions/programs may not be couched in the right terms for the biology, or with proper appreciation for biological complexity.

Challenge #3:

Both biology and computational science are deep, complex fields of study, inhabited by extremely smart people!

- None of this is easy, on any side of things.
- If it were easy, they wouldn't need people as smart as all of us to do it, right??
- A two week course can't possible teach you everything.

Challenge #4:

Sequencing technology is changing very fast.

- We don't understand its limitations or biases very well.
- The software and compute infrastructure lags behind volume of data & type of data.

This is not the #1 problem you will face with bioinformatics.

Here is the #1 problem:

How do you know if your computational answer is right or wrong?

This is not the #1 problem you will face with bioinformatics.

Here is the #1 problem:

How do you know if your computational answer is right or wrong?

If you can't answer this question, then what's the point of doing the computation?

Controls

- Just as with experiments, you can put negative and positive controls in your bioinformatics.
- e.g. with BLAST,
  - Do you see expected matches with the parameters and database you're using?
- Positive controls are often easier than negative, in "discovery" science…

Internal controls

- Use molecules and sequences for which you have expectations.
- “I know this gene comes up, based on qPCR. I expect to see it in my mRNAseq.”
- Or, “human? I didn’t expect to see human!”

External controls & replication

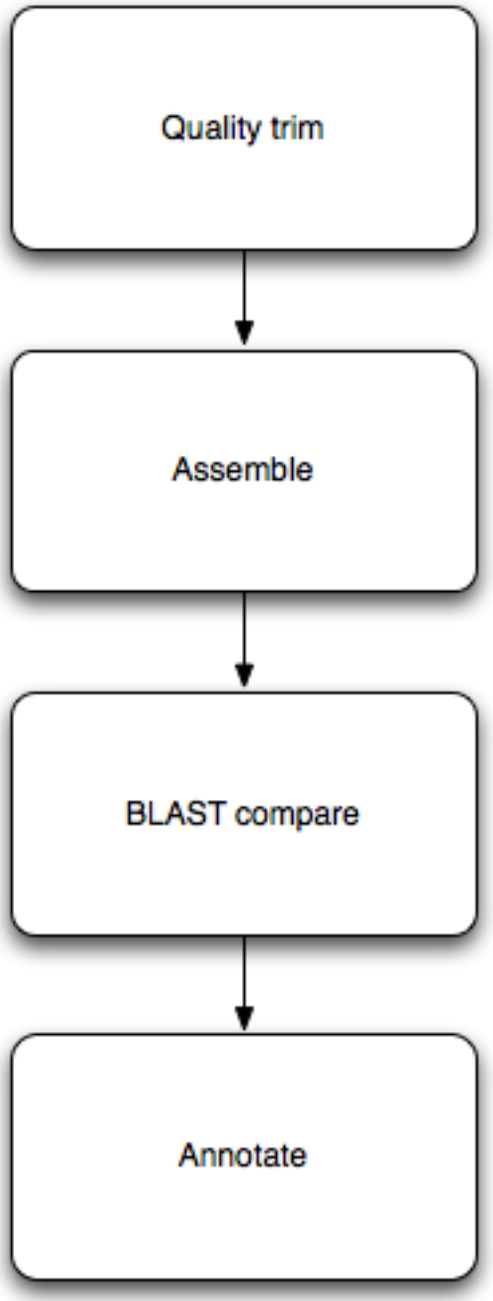
- Does the whole process work?
- "I can reproduce what this other person/lab did, with their data, when I use my own software."
- This is much more rarely done…

Black box nature of algorithms

- When you listen to a computational biologist explain their clever algorithm…
- …it's a **big** mistake to think that they necessarily know what's going on.
- Software is full of bugs and unintended consequences.

Pipelines

- Each step can be understood, and tested/controlled individually.
- Each step is re-usable! Just need to figure out input/ output formats.
- Automate, automate, automate.



The opportunity:

- The sequence is here! As you know!
- "In the land of the blind, the one eyed is king." -- those prepared to *think* about how to use sequencing technology to answer their question will have a substantial leg up.
- Who knows? Some of you might even like this mix!

Our goals

- Provide a safe & welcoming place to experiment.
- Lots and lots of help (in the form of Tas)
- Provide lots of data sets, tools, scripts.
- Research specific help as possible.

Our requirements of you

- Nothing.
- This is a requirements free zone.
- You can safely skip the entire course…

Our expectations

- Questions!
- Ask for help when you need it!
- Tolerance (in both directions)

Our hopes

- Enthusiasm!
- Engagement!

Daily schedule (tentative)

- 7-8: breakfast. They mean it. /cc Frona's Bakery
- 9:15am – lecture
- 10:30am – tutorial 1
- 12-1pm - lunch
- 1:15pm – tutorial 2
- 3pm – free time!
- 5-6:30 - dinner
- 7pm – tutorial/lecture/etc

Weekly schedule – tentative wk1

- Tuesday – BLAST, sequence quality foo
- Wed – mapping & assembly; genomic visualization
- Thursday – Genomic intervals & bioinfo survival
- Friday – SNP calling, experimental design
- Saturday – pipelines & protocols for mRNAseq

Dramatis personae

- Titus Brown (that's me)
- Ian Dworkin -- co-instructor
- Istvan Albert – co-instructor

- Cody Nicks – go-fer and aide-de-camp

Dramatis personae

- Amanda Charbonneau– TA and cruise director
- Elijah Lowe– TA
- Will Pitchers – TA
- Aswathy Sebastian - TA
- Qingpeng Zhang – TA.

Dramatis personae

Other instructors:

Daniel Standage, Meg Staton, Chris Chandler, Adina Howe, Aaron Darling, Matt MacManes.

Written rules

- No night-swimming without a buddy.
- I mean it.

Code of Conduct

http://angus.readthedocs.org/en/2014/code-of-conduct.html

tl;dr? **Don't be a jerk.**

I will post Judi Brown Clark's contact information on the wall shortly.

Note: this is not because of known prior problems, ICYW.

Food and drink

- Anything group-intended can be purchased by Cody. Please write it down on the list in the back.
- Cody can also drive you to the market; he'll probably go every two or three days.
- Please don't ask Cody to spot you $$; ask me.

Games and location.

- We have volleyball, frisbee, frisbee golf, boche ball…?
- Also cards. Other board games needed?
- There's good places to run, to swim, to hike, to bike, and to fish.
- We also have laundry and weight room (?)

Unwritten rules

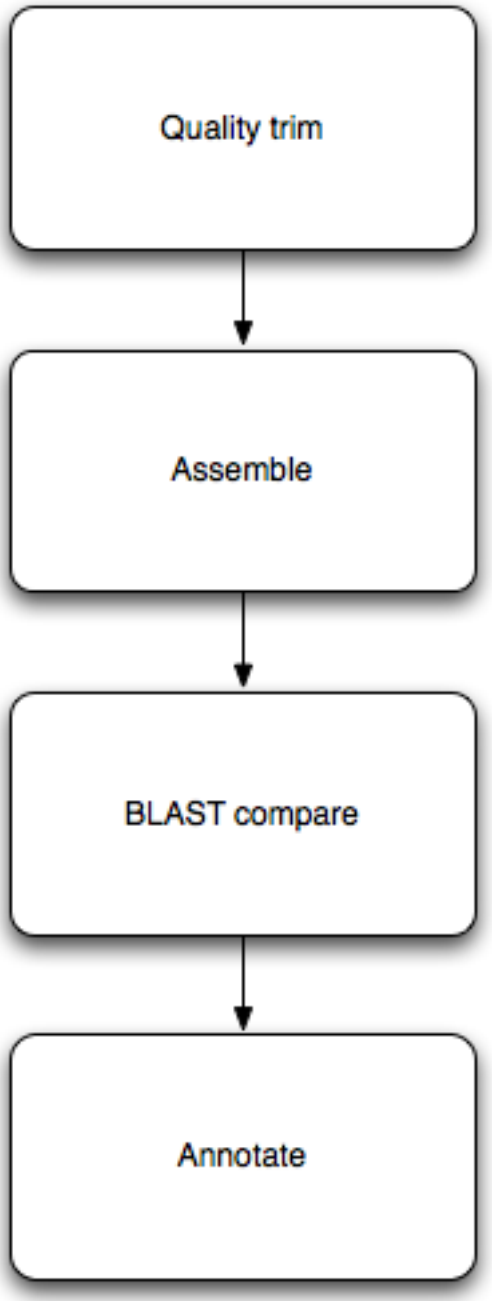
Framing the approach

1. How does all this stuff work, generally?

2. Can we automate things and/or do them more efficiently?

How does this stuff work

- Typically, you need to run multiple different programs in sequence.
- Each program takes in data, in files; and outputs data, in files.
- (Some programs also produce pretty pictures via the Web.)



Automation & computational efficiency matter

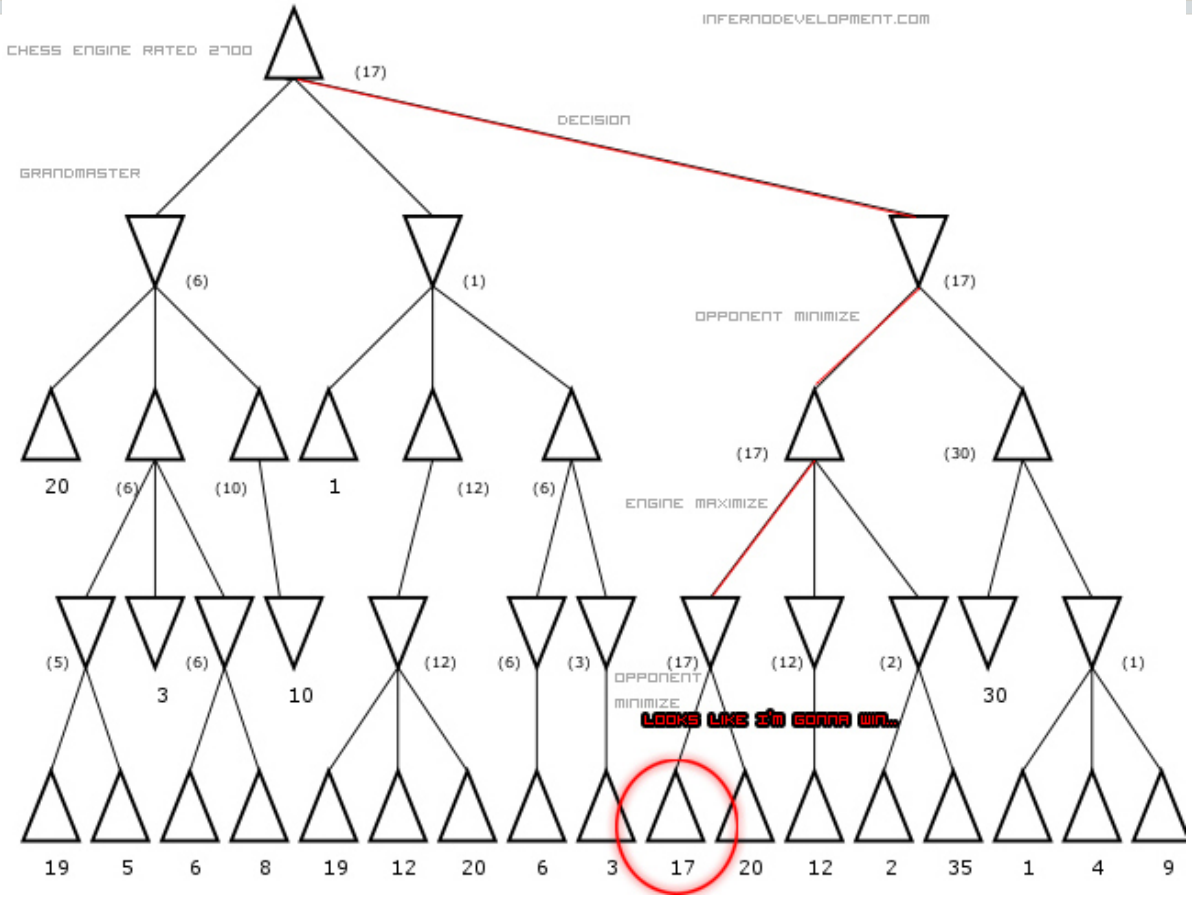
- You'll learn to run lots of different programs here.
- We'll run into some practical problems:
  - Some programs take a long time to run.
  - Some programs take many different parameters; which are best?
  - Some programs don't finish on "cheap" hardware.

How do we run many long-running programs? How do we remember what we did? How do we get our programs to finish?

“Heuristics”

- What do computers do when the answer is either really, really hard to compute exactly, or actually impossible?
- They approximate! Or guess!
- The term “heuristic” refers to a guess, or shortcut procedure, that usually returns a pretty good answer.

Often explicit or implicit tradeoffs between compute “amount” and quality of result



http://www.infernodevelopment.com/how-computer-chess-engines-think-minimax-tree

This kind of issue comes up a lot.

- Mapping.
- Assembly.
- Statistics (Monte Carlo and resampling methods).
- Simulations.

- More generally, most "interesting" algorithms involve approximations and shortcuts. When are they (in)appropriate for your task?

What are the limits of data + compute?

Mappers will ignore some fraction of reads due to errors.

ıapping errors for single (SE) and paired end (PE) reads real transcriptomes

Num Trans	Error	TP (d)	FP (d)	TP (u)	FP (u)	TP
5000	1%	92%	0%	92%	0%	92
5000	1%	87%	5%	81%	0%	92
5000	1%	85%	0%	85%	0%	85
5000	1%	81%	4%	77%	0%	85

ers are default (d), unique (u), and multimap (m). True positives a ed to their originating transcript. False positives are reads that we f the read was an exact match to the alternate transcript).

Pyrkosz et al., unpub.

Does choice of mapper matter?

3. Comparison of Three Common Mapping Programs on the Same Chicken Data

ism	Num Trans	Bowtie TP (d)	FP (d)	BWA TP (d)	FP (d)	SOAP2 TP (d)	
en	100%	78%	22%	78%	20%	78%	
en	90%	72%	21%	72%	20%	72%	
en	80%	65%	22%	65%	21%	65%	
en	70%	58%	22%	58%	21%	58%	
en	60%	51%	20%	50%	19%	51%	
en	50%	44%	19%	44%	18%	44%	
en	40%	36%	16%	37%	16%	36%	
en	30%	27%	13%	27%	13%	27%	
en	20%	19%	11%	19%	11%	19%	
en	10%	9%	5%	9%	6%	9%	

ison of Bowtie, BWA, and SOAP2 mapping programs on the same simulated reads for error-
read sets (triplicate and averaged) with decreasing completeness of the reference transcripto
equivalent results.

Reference completeness matters more!

Pyrkosz et al., unpub.

Real problem? Our data can't uniquely specify solution!

Here, there is no direct way to know if last exon is connected to first exon.

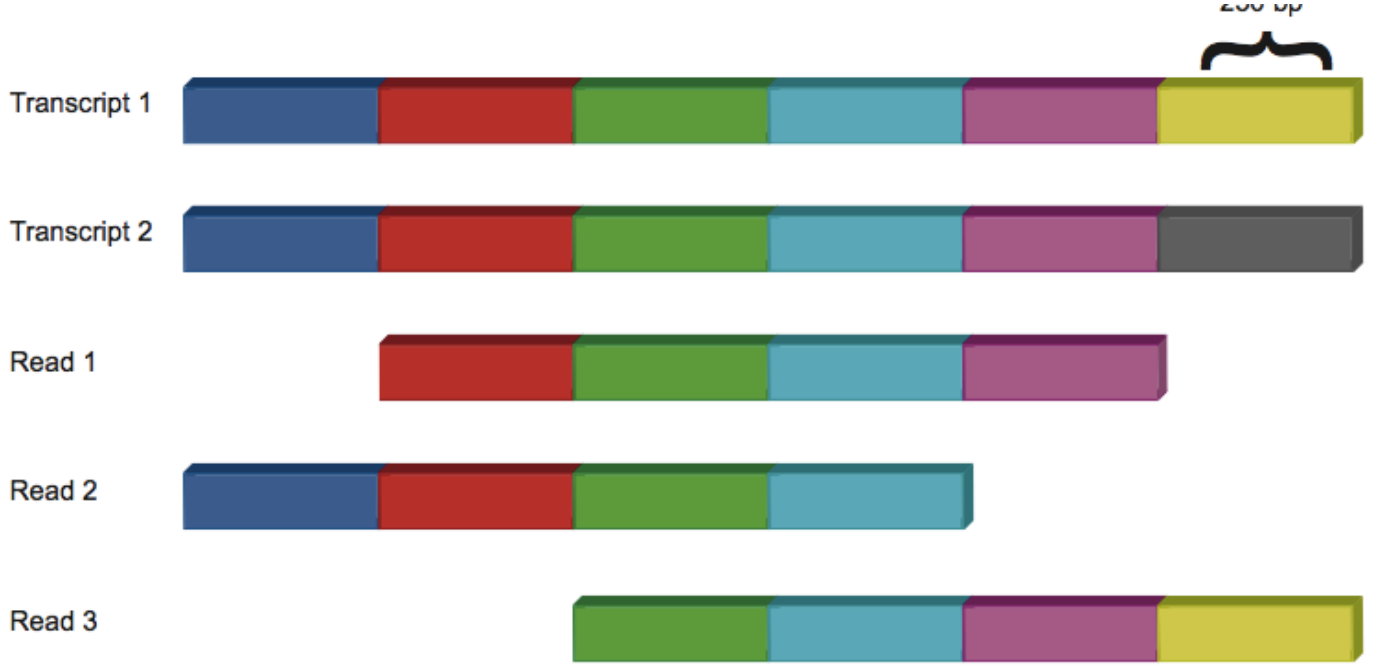


Figure 6. Hypothetical example of 1 kb multimap reads. Only Read 3 can be uniquely

Pyrkosz et al., unpub.

Concluding thoughts

- There's what you can do today, computationally, with existing programs. This is often limited by our time, experience, etc.
- There's what you could, in theory, do with the data you had. This is the upper limit on your accuracy.
- Figuring out the difference is one of the main reasons you're here :)

Any questions or comments?

Process and materials!

- Use the ngs-2014@lists.idyll.org list to organize things!
- Twitter: #ngs2014; I'm @ctitusbrown
- Facebook group?

Use the stickies, Luke…