Welcome!

Everything's going to be OK.

This is the 6th year we've run this course (!!)

You're in experienced hands.

(I won't say "good", necessarily. But experienced.)

Everything's going to be OK.

Number of emergency room trips: 1

Let's avoid incrementing that number…

Introducing myself

- Until January, MSU asst professor in Microbiology and Computer Science.
- As of January, assoc professor in School of Veterinary Medicine at UC Davis
- Background: programming, evolutionary biology, math, climatology & astronomy, dev bio, regulatory genomics, genomics, gene regulatory networks, bioinformatics.

Family!
(Not here ☹)

Others in the room:

- Matt MacManes (UNH)
- Amanda Charbonneau (MSU)
- Lisa Cohen (UC Davis)
- Phil Brooks (MSU)
- Jessica Mizzi (MSU / UC Davis)

What are the goals of this workshop?

1. Expose you all to a bunch of approaches and ideas and details.
2. Expose you to a particular way of *thinking.*
3. Train you in a particular way of *learning* and *doing*.
4. Networking!

At the *worst*, you will find out you know more (or less) than you thought.

Our approach.

- 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.

Daily schedule (generally)

- 7-9: breakfast
- 9:15am – lecture
- 10:30am – tutorial 1
- 12-1pm - lunch
- 1:15pm – tutorial 2
- 3pm – free time!
- 6-7- dinner (unless noted otherwise)
- 7pm – tutorial/lecture and/or fun

Something that we don't talk about enough in this workshop:

Most / best bioinformatics information is online:

- Blogs
- Twitter
- SeqAnswers
- Biostars

Spend some time poking around while you're here. We have some suggestions on the course web site.

Written rules

No night-swimming without a buddy.

I mean it.

Code of Conduct

http://angus.readthedocs.org/en/2015/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, ICYAW.

Food and drink

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

Free time, more generally.

Please be sure to take time off as you need it.

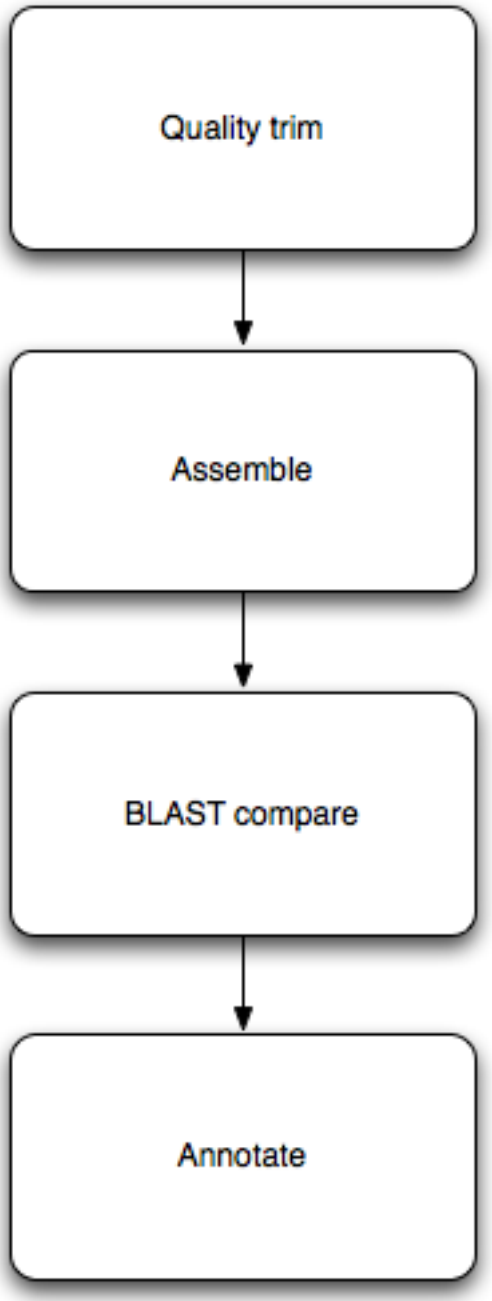
Not only is relaxation important, given the next two weeks, but the networking is also important.

Plus, the TAs and profs need the time off too :>

Unwritten rules

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

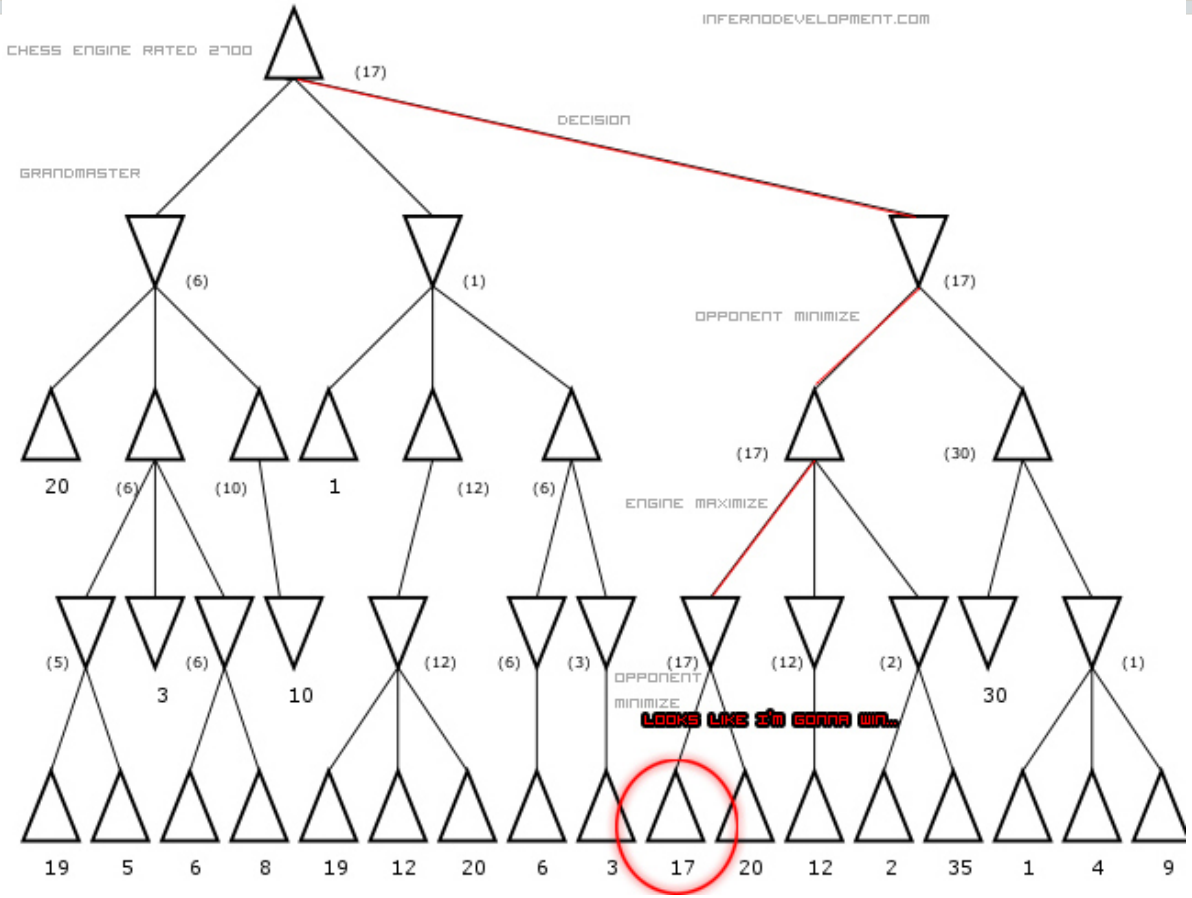
- You'll learn to run lots of different programs here.
- We'll run into some practical problems:
  - Some programs take a long time to run.
  - Some programs take many different parameters; which are best?
  - Some programs don't finish on "cheap" hardware.

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.

Table 1. Read mapping errors for single (SE) and paired end (PE) reads from random (simulated) and real transcriptomes

Organism	Num Trans	Error	TP (d)	FP (d)	TP (u)	FP (u)	TP (m)	FP (m)
Random (SE)	5000	1%	92%	0%	92%	0%	92%	0%
Mouse (SE)	5000	1%	87%	5%	81%	0%	92%	12%
Random (PE)	5000	1%	85%	0%	85%	0%	85%	0%
Mouse (PE)	5000	1%	81%	4%	77%	0%	85%	9%

Mapping parameters are default (d), unique (u), and multimap (m). True positives are reads that were successfully mapped to their originating transcript. False positives are reads that were mapped to other transcripts (even if 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.

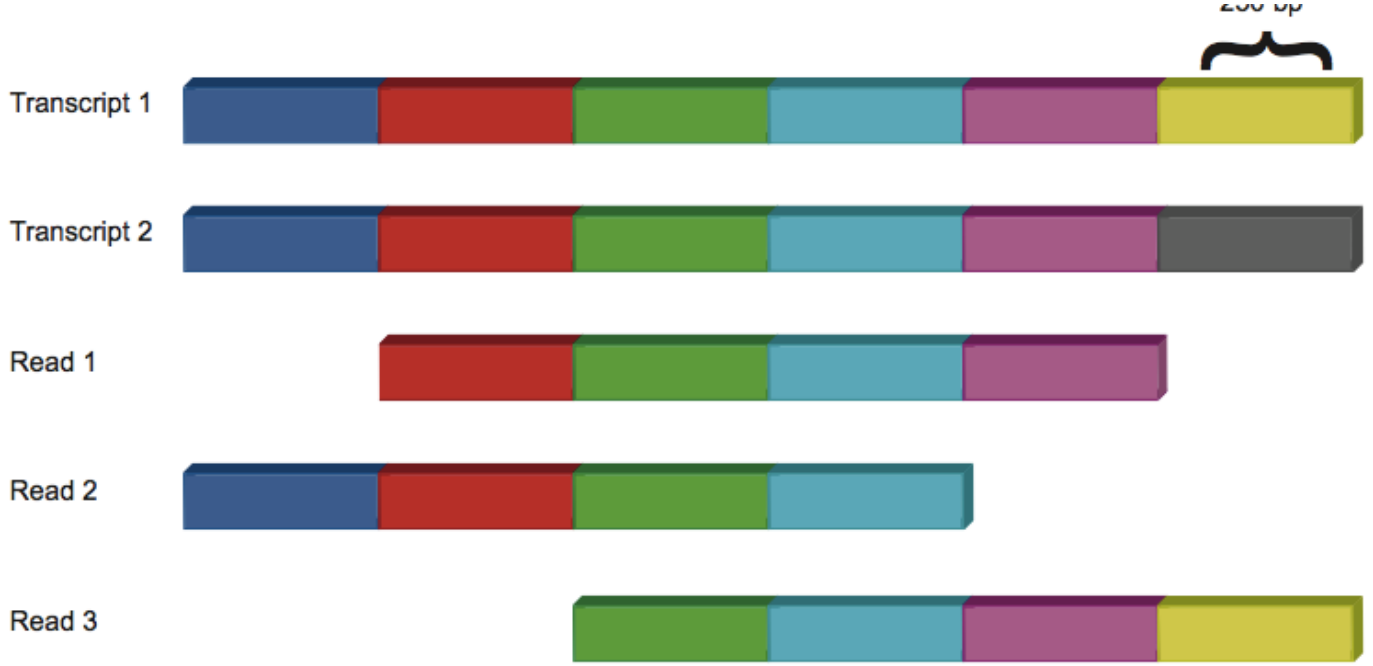


Figure 6. Hypothetical example of 1 kb multimap reads. Only Read 3 can be uniquely

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 :)

Process and materials!

- Use the ngs-2015@lists.idyll.org list to organize things!
- Twitter: #ngs2015; I'm @ctitusbrown
- Facebook group.

Use the stickies, Luke…

Any questions or comments?