Welcome!

Introductions

- Me
- Greg Wilson (next Monday)

- Three TAs:
  - Rose
  - Jason
  - Likit

- Several visiting speakers/profs
  - Ian Dworkin
  - Jeff Barrick (resequencing – next Tues)
  - Mark Robinson (ChiP-seq – next Wed)

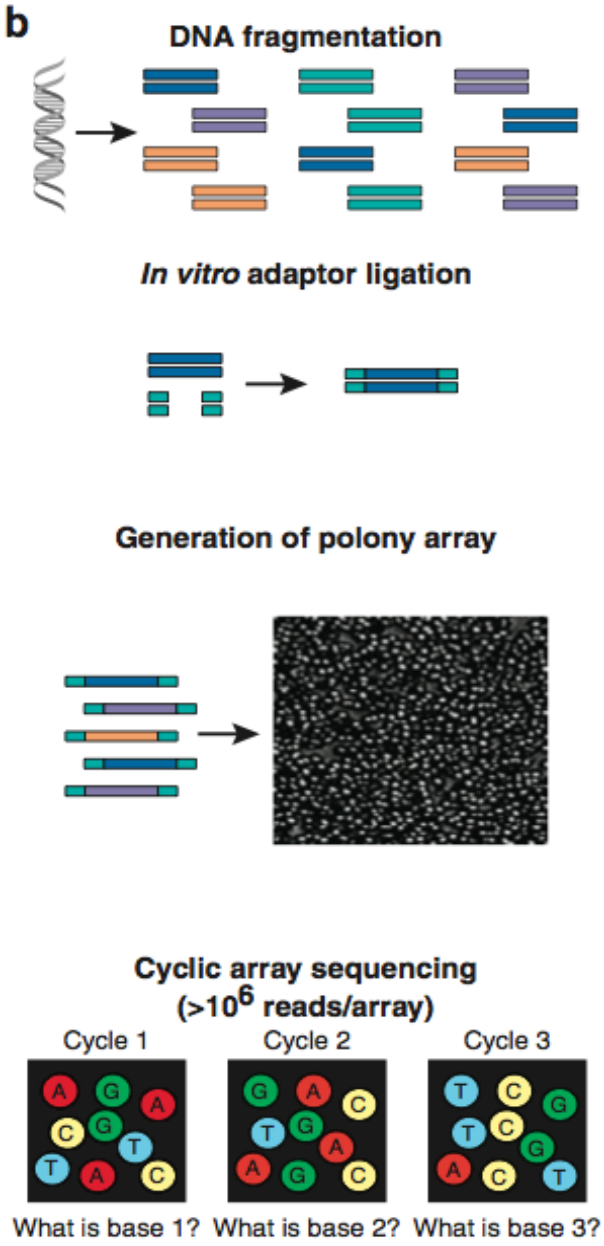
What the heck is BEACON?

- BEACON is a Science & Technology Center that I'm associated with at MSU (to do with Evolution in Action)
- The KBS folk assumed that this course was associated with BEACON.
- It's not, really. Now you know.

OK, …next-gen sequencing.

Basic technology

- 454, Illumina, and SOLiD all use "polonies", PCR colonies.
- By amplifying in isolation, they avoid PCR biases towards one sequence over another.



Shendure and Ji, Nat Biotech, 2008

Basic uses

- Genome (re)sequencing
- mRNA (re)sequencing
- Protein-DNA quantitation

Basic analysis

- Map to known sequence

OR

- Assemble sequence from scratch

Mapping to known genome

- Requires reasonably good reference sequence.
- Pitfalls surrounding computational assumptions and heuristics (shortcuts).
- Even for most sequenced organisms, annotations kind of suck.

Assembly

- Much more challenging than mapping.
- Requires considerably more computational resources…
- Very hard to validate (esp, what are you missing?)

Practical problems dominate!

Data size is a problem

- Just working with large files, transferring them around, and backing them up is annoying.
- Also causes problems for computers with limited memory, etc.
- Particularly a problem for assembly…

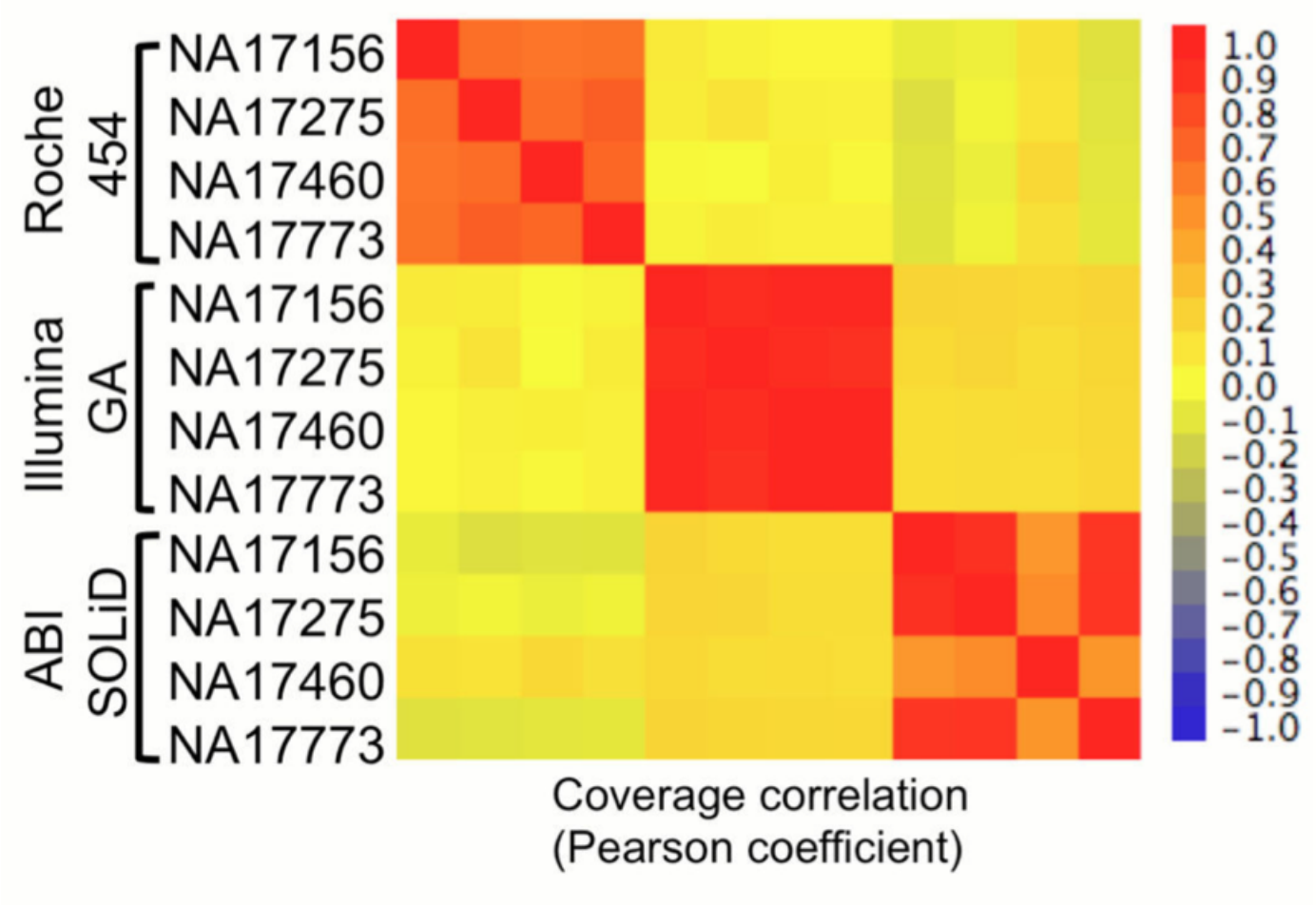
Analysis scaling

- Analysis algorithms rarely scale linearly with the amount of data.
- For example, naïve sequence comparisons scale as N*N: in order to compare N sequences against themselves, you need to do N*N operations.
- N=1 => N=10**3, analysis time increases by 10**6.

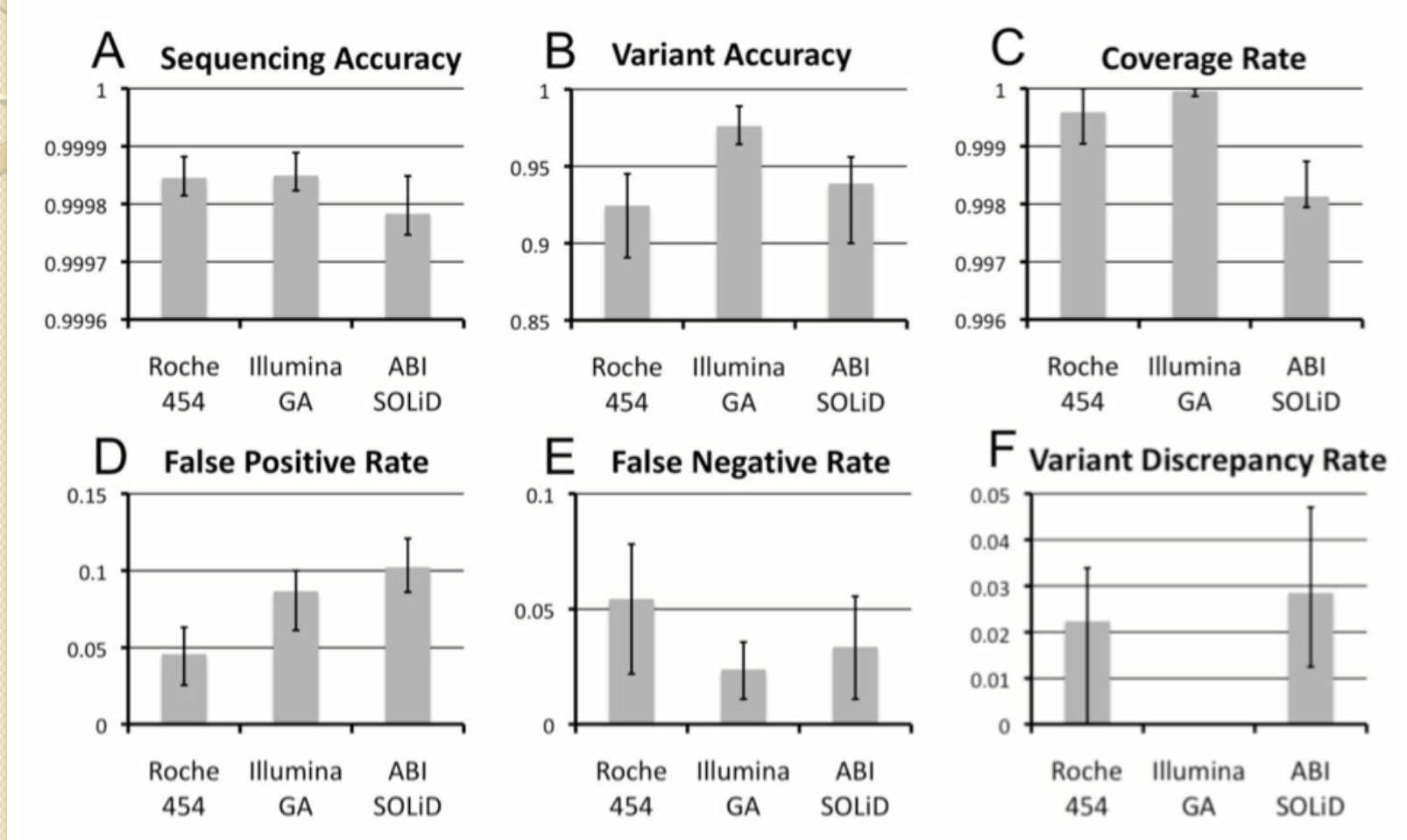
Data characteristics are a problem

- Errors
- Biases
- Software to deal with errors and biases

Platform-specific variation in coverage

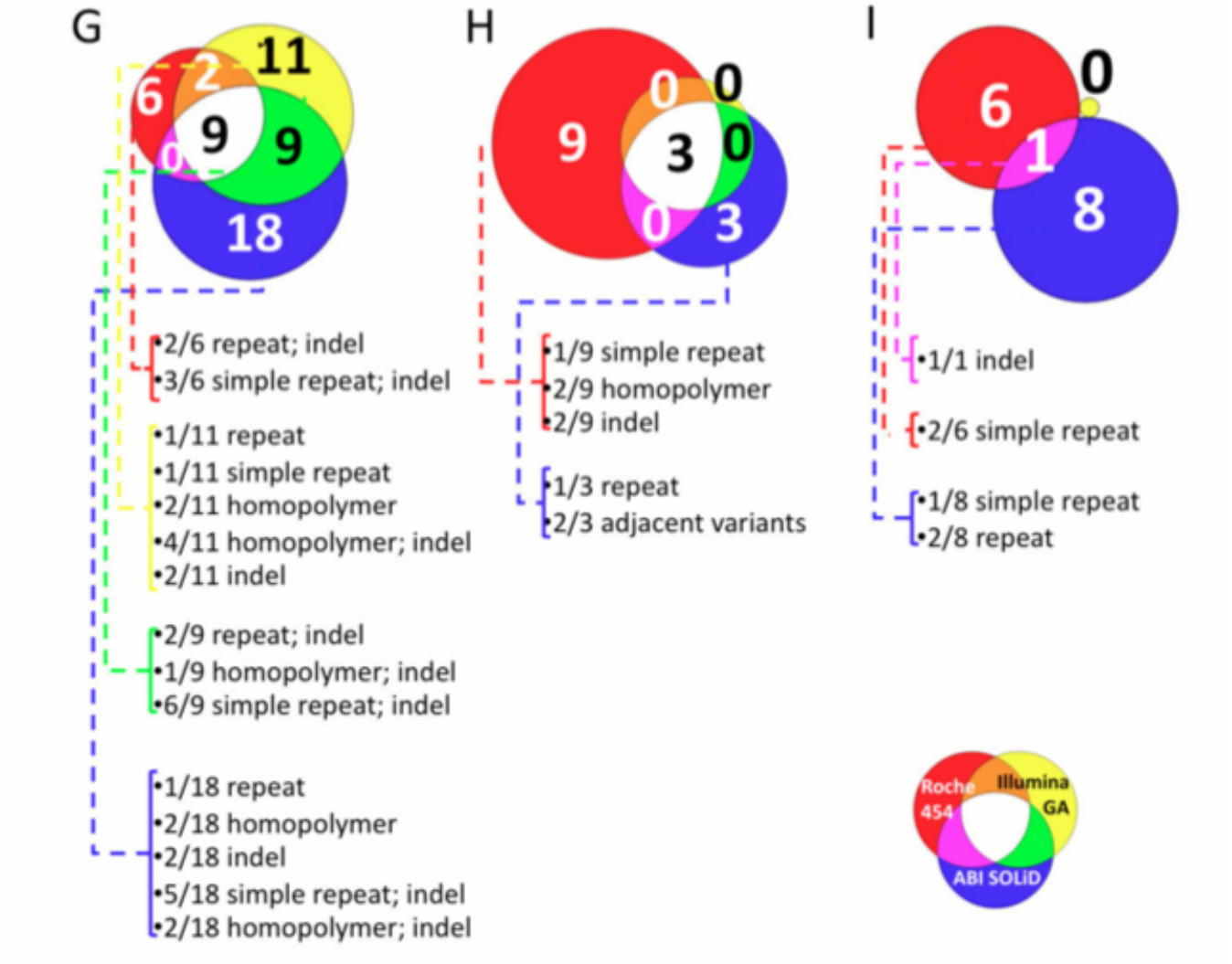


Harismendy et al., Genome Biol. 2009, pmid: 19327155



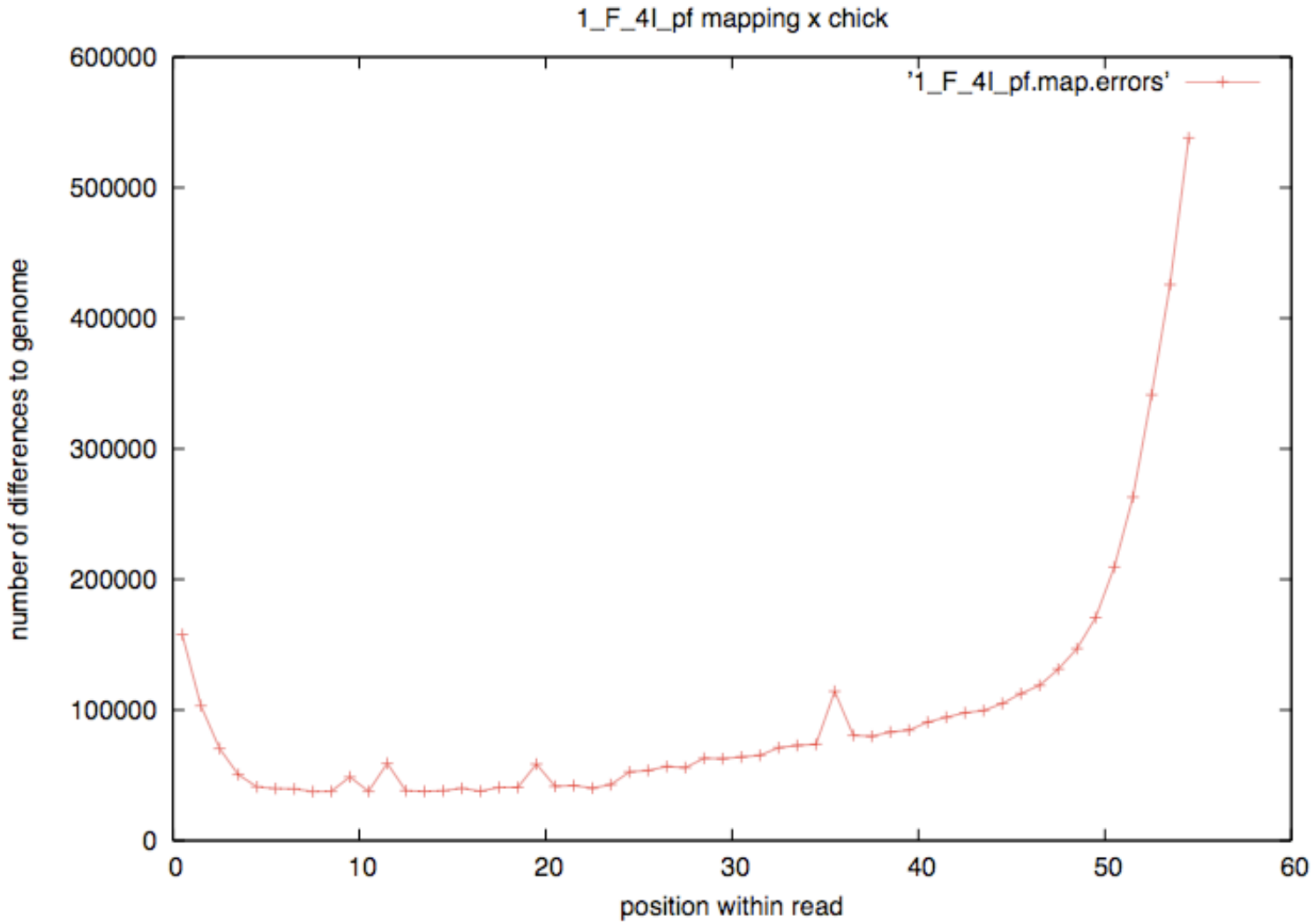
Harismendy et al., Genome Biol. 2009, pmid: 19327155

Mis-called variants associate with annoying sequence features, + technology-specific bias

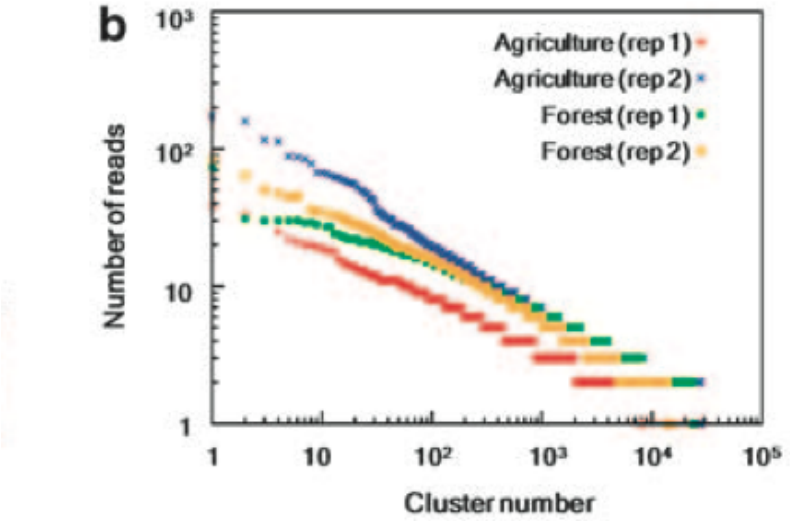


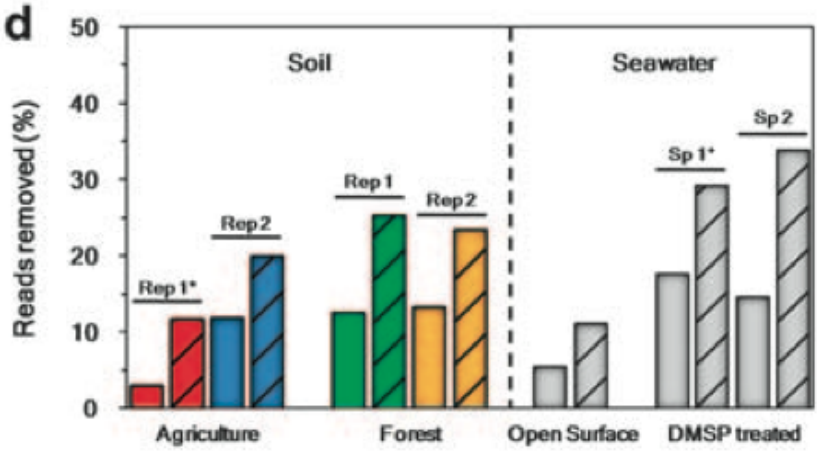
Harismendy et al., Genome Biol. 2009, pmid: 19327155

Error rates rise dramatically with length (Illumina)



Artificial replicates in 454:





11-35% of sequences in a typical metagenome are artifacts of 454.

Gomez-Alvarez et al., ISME, 2009

Skeptical questions

- How much observed variability is due to:
  - Sequencing error/bias
  - Poor trimming of sequence
  - Mis-mapping of reads
  - …?

- Need to evaluate entire pipeline, including software!

Tool usability is a big problem

- Almost all of the tools are "command-line", which means there's no graphical interface.
- This is because building graphical interfaces is a lot of (extra) work!
- In this course you will be learning to use the command-line tools.

Data volume is increasing

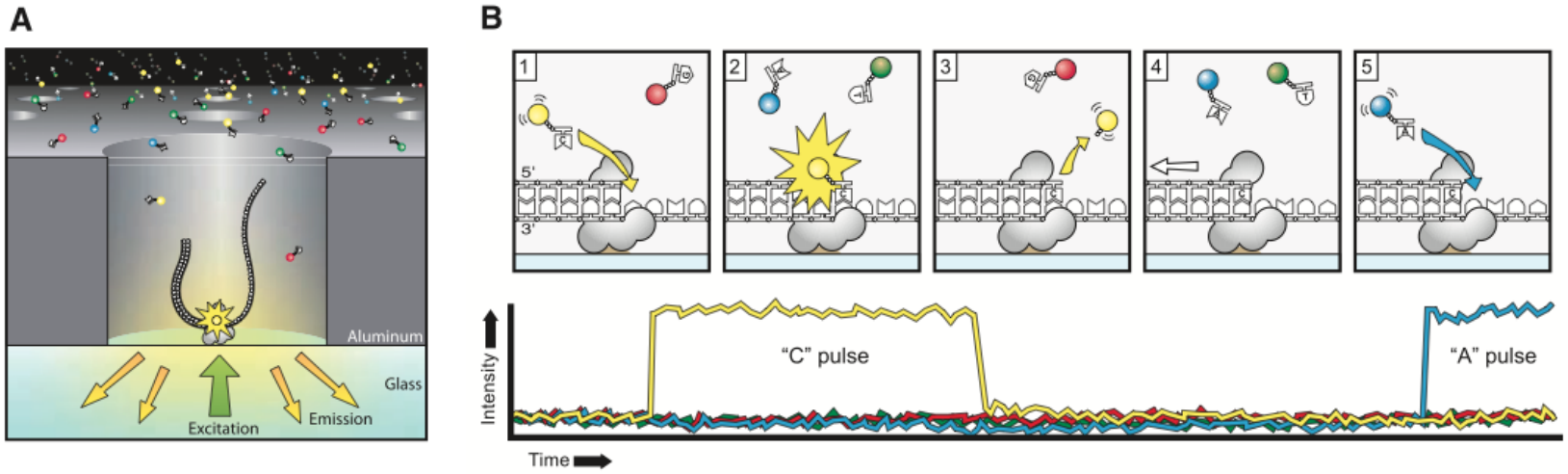
- Number of reads, % mappable, & length of reads has already increased dramatically in the past year for Illumina, 454.

- Now ~200m reads / Illumina flow cell, @ 100bp or more.

(That's 20gb of *sequence* alone.)

New techniques are emerging, too

PacBio promises 1-10kb reads



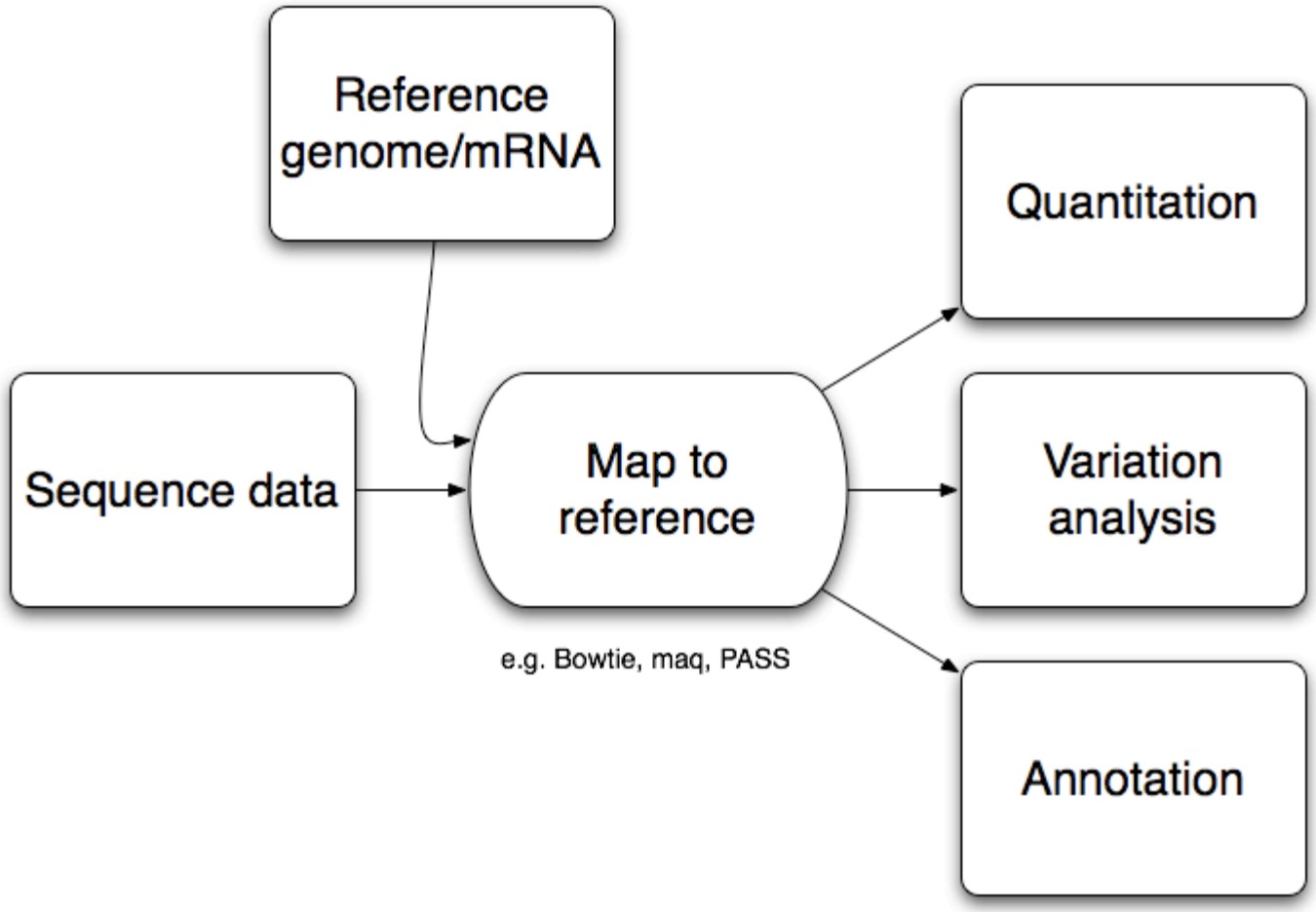
Eid et al., Science, 2009

(and there are others, too)

Purpose of this course

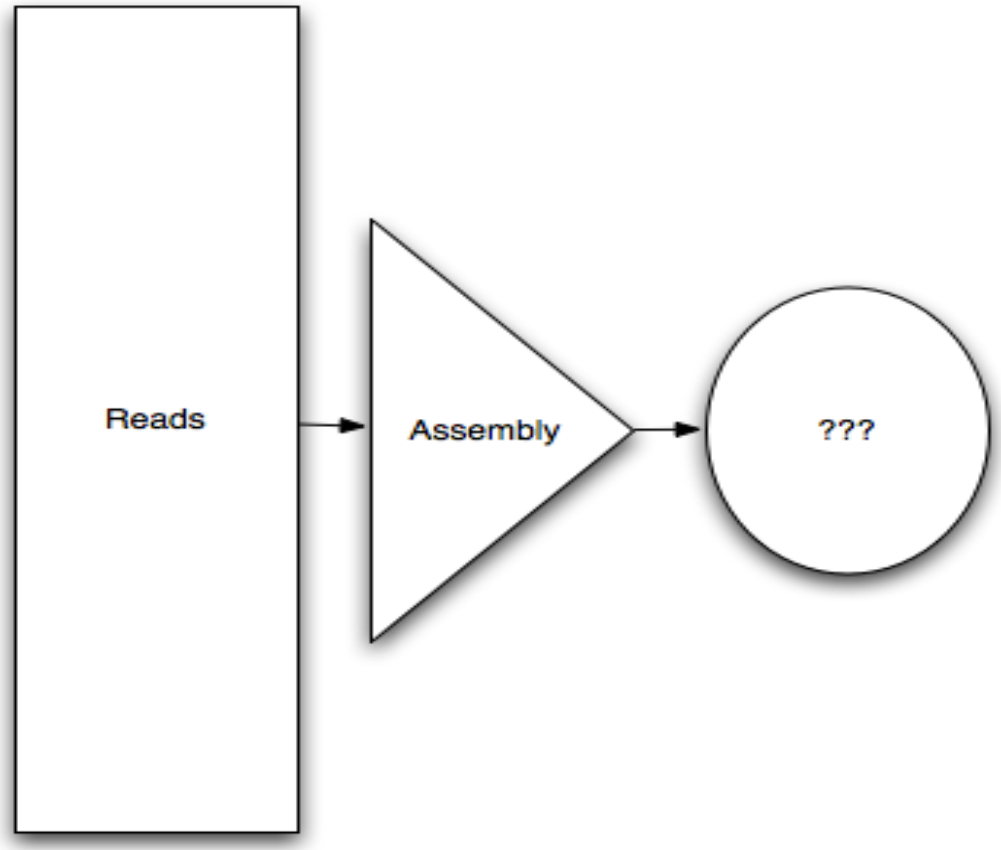
- Get you used to thinking mechanistically about dealing with the data.
- Provide you with practical experience.
- Create a reasonably detailed Web site for you to refer back to, and for others to use.

Main focus: data reduction step



Once you've mapped to a reference, the data becomes much more manageable.

2ndary focus: assembly & eval



Assembly is very technically challenging. Hopefully we can give you the benefit of our experience (=> Jason and Likit, esp.)

3rd focus: post-reduction analysis

What you do after the mapping or assembly?

No general answers… we'll try to help you out, though.

Any questions, comments, thoughts?

- Stuff you want to see?

You Are Not Stupid or Inept.

At the *worst*, you're uninformed.

(That's why you're taking this course, right?)

Don’t Panic.

- All of the instructors in the course were selected:
  - Knowledgeable
  - Friendly (except maybe Rose)
  - Good drinkers
- I personally cannot imagine a nicer, friendlier environment for you to learn stuff in.

Try, try, try.

We have provided protocols.

Like all protocols, we will have left stuff out.

Bang your head against the wall
(figuratively)

Ask for help.

You will become more powerful than you can possibly imagine.

Generations of computational scientists have worked with computers this way.

The tools are getting better, and easier, and more powerful.

(Why, back in my day…)

However: "easier" does not mean "easy."

You cannot break anything.

You do not have the ability to do anything destructive.

(Unless you physically destroy one of the computers.)

The worst that will happen is you will lose your work-in-progress. (We will help show you how to avoid that.)

So, *be adventurous*.

Future-proofing you

You will learn specific tools only as one example of how to do it.

The general approach will translate to the future: new tools, new data types, etc.

Ask questions, go your own way

- The tutorials in the first week will give you a minimum level of "useful" knowledge.
- They are not meant to be exhaustive (hah!) and they are not meant to be 100% of your time.
- Ask, ask, ask.

Course structure

We have you for two weeks.

Week 1: fundamentals.

Week 2: details, and your research.

Week 2 is less structured; pick a problem, or several problems, work on them.

Grading

- Grading is attendance-based: if you show up for 5+ lectures you will get a 4.0.

- I would, however, be disappointed if that's all you wanted from the course ☺

Daily structure - tentative

Lecture:	- 9am
Tutorial 1:	- 11:00am
Lunch	- noon
Tutorial 2:	- 2:30pm
Dinner	- 6pm
Tutorial 3:	- 8pm

Week 1

1. Basic computer stuff (logging in, etc.)
2. Running big jobs (with BLAST)
3. Large, messy data sets, and stats
4. Mapping
5. Assembly

…that brings us through Saturday.

Week 2

- Similar format (lecture, tutorials)
- More one-on-one work with tools, approaches.
- This is your chance to begin to solve the problem that made you come here!

Week 2

1. mRNAseq
2. Resequencing
3. ChIP-seq
4. Still undecided.

…and that concludes the 2nd week.

Questions?

A "reflective exercise"

- I'd like to assess the degree to which thought processes have changed due to this course.
- So, I'd like you to write down (from memory is OK) a paragraph or two about a biology problem you are interested in addressing, and how you propose to tackle it.
- Then, at the end of the course, you can address the same problem and see whether or not the course has changed your thoughts, and how.
- I would be interested in applying to funding bodies for support for this course, so this would help me out, too.
- However, it is entirely voluntary to participate.