Introduction to Methods in Computational Biology and Genomics

C. Ryan Campbell

Duke University

c.ryan.campbell@duke.edu

August 29, 2017

Overview

Today's Goals

- Get familiarized with course format
- Meet each other and myself
- Understand course expectations
- Know my teaching methods and reasons for offering this course

Course Objectives

Objective 1

Learning programming basics and best practices within a biology framework (i.e. basic building blocks for students to use biological software)

Course Objectives

Objective 1

Learning programming basics and best practices within a biology framework (i.e. basic building blocks for students to use biological software)

Objective 2

Learning the statistics that underlie these tools and the rapidly evolving suite of measures used in ”genomics”

Course Objectives

Objective 1

Learning programming basics and best practices within a biology framework (i.e. basic building blocks for students to use biological software)

Objective 2

Learning the statistics that underlie these tools and the rapidly evolving suite of measures used in ”genomics”

Objective 3

Combine 1 and 2 to test biological hypotheses with RNAseq data and present those results in an ”IMRD” format (pronounced EM-rod)

Course Syllabus

Course Grade

Course Policies

- Honor Code

Course Policies

- Honor Code
- Group Work vs. Own Work

Course Policies

- Honor Code
- Group Work vs. Own Work
- No Direct Attendence Policy

Course Policies

- Honor Code
- Group Work vs. Own Work
- No Direct Attendence Policy
- Interact and Contribute in Class

Potential Course Policies - TBD

- Latest pre-class email time (5pm? 9pm?)

Potential Course Policies - TBD

- Latest pre-class email time (5pm? 9pm?)
- Slack?

Potential Course Policies - TBD

- Latest pre-class email time (5pm? 9pm?)
- Slack?
- Office Hours?

Day-to-day Course

- Bring a laptop

Day-to-day Course

- Bring a laptop
  - Yes, Every Day

Day-to-day Course

- Bring a laptop
  - Yes, Every Day
- Tuesdays = Lecture

Day-to-day Course

- Bring a laptop
  - Yes, Every Day
- Tuesdays = Lecture
- Thursdays = Lab

Project

Teaching Philosophy

- Clear Goals
- Active Learning
- Student Driven

Clear Goals

- Presented before each class
- Concepts or skills to focus on
- Call me on it if I forget them

Active Learning

- Student Participation
- A different style than lecture-based courses (flipped courses fall into this category)
- Natural fit for smaller class size, advanced material, and learning skills

Student Driven

- Student Participation Required
- Work through examples in class and apply to your own question
- Many skills need to be practiced, not taught
- Grand Bargain

In-Class Activity

1. Pair up *randomly*

In-Class Activity

1. Pair up *randomly*
2. Fill out this Google form

In-Class Activity

1. Pair up *randomly*
2. Fill out this Google form
3. Make a slide in the Google Slideshow for your partner with their answers, including (at minimum):

In-Class Activity

1. Pair up *randomly*
2. Fill out this Google form
3. Make a slide in the Google Slideshow for your partner with their answers, including (at minimum):
    - Name
    - Picture
    - ‘Three words’ response
    - Two answers from ‘’Whimsy’
    - Feel free to expand (see my slide for reference)

In-Class Activity

1. Pair up *randomly*
2. Fill out this Google form
3. Make a slide in the Google Slideshow for your partner with their answers, including (at minimum):
   - Name
   - Picture
   - 'Three words' response
   - Two answers from ''Whimsy'
   - Feel free to expand (see my slide for reference)
4. Introduce partner to class

Computer Requirements

- Bring your computer to class
- Run some software locally
  - How much storage do your computers have?
- Connect to cluster

Research Tools

- R
  - Statistical software, data visualization and analysis

Research Tools

- R
  - Statistical software, data visualization and analysis
- kallisto
  - Fast RNAseq analysis with a command line interface

Research Tools

- R
  - Statistical software, data visualization and analysis
- kallisto
  - Fast RNAseq analysis with a command line interface
- github
  - Website for version control as well as script and code storage

Programming Languages

- R
  - free statistical software
- bin/bash
  - unix language, automate many tasks, interact with cluster computers

Software

- R -¿ R-Studio

Software

- R -¿ R-Studio
- github -¿ SourceTree

Software

- R -¿ R-Studio
- github -¿ SourceTree
- data analysis software -¿ fastqc/trimmomatic/kallisto

Cluster Computing

- Duke Computing Cluster
- SLURM workload manager

Hand Raising Request

- Student Driven - Active Learning

Hand Raising Request

- Student Driven - Active Learning
- Raise your hand if you're confused

Hand Raising Request

- Student Driven - Active Learning
- Raise your hand if you're confused
- Provides me with helpful feedback

What is/are Genomics?

What is/are Genomics?

- Within the field of molecular biology, genomics is the study of genomes, the complete set of genetic material within an organism.

What is/are Genomics?

- Within the field of molecular biology, genomics is the study of genomes, the complete set of genetic material within an organism.
- Genomics involves the sequencing and analysis of genomes.

What is/are Genomics?

- Within the field of molecular biology, genomics is the study of genomes, the complete set of genetic material within an organism.
- Genomics involves the sequencing and analysis of genomes.
- Genomics is also concerned with the structure, function, comparison, and evolution of genomes.

What is/are Genomics?

- Within the field of molecular biology, genomics is the study of genomes, the complete set of genetic material within an organism.
- Genomics involves the sequencing and analysis of genomes.
- Genomics is also concerned with the structure, function, comparison, and evolution of genomes.
- The field also includes studies of intragenomic (within the genome) phenomena such as heterosis (hybrid vigour), epistasis (effect of one gene on another), pleiotropy (one gene affecting more than one trait) and other interactions between loci and alleles within the genome.

What is/are Genomics?

- Within the field of molecular biology, genomics is the study of genomes, the complete set of genetic material within an organism.
- Genomics involves the sequencing and analysis of genomes.
- Genomics is also concerned with the structure, function, comparison, and evolution of genomes.
- The field also includes studies of intragenomic (within the genome) phenomena such as heterosis (hybrid vigour), epistasis (effect of one gene on another), pleiotropy (one gene affecting more than one trait) and other interactions between loci and alleles within the genome.
- In contrast to genetics, which refers to the study of individual genes and their roles in inheritance, genomics uses high throughput DNA sequencing and bioinformatics to assemble, and analyze the function and structure of entire genomes.

Brief History of Sequencing

- Allozymes
  - Electrophoresis separates different proteins by amino acid makeup
- Sanger Sequencing
  - Determines the sequences a single piece of DNA up to 500bp
- NGS - Next Generation Sequencing
  - Reads sequence of many pieces of DNA many billions of times

Brief History of Sequencing

- Allozymes
  - 1960's
- Sanger Sequencing
  - 1977
- NGS - Next Generation Sequencing
  - 2000

Next Generation Sequencing vs. Sanger

- Output for Quality Tradeoff
  - NGS = High Output / Lower Quality
  - Sanger = Low Output / High Quality

Next Generation Sequencing vs. Sanger

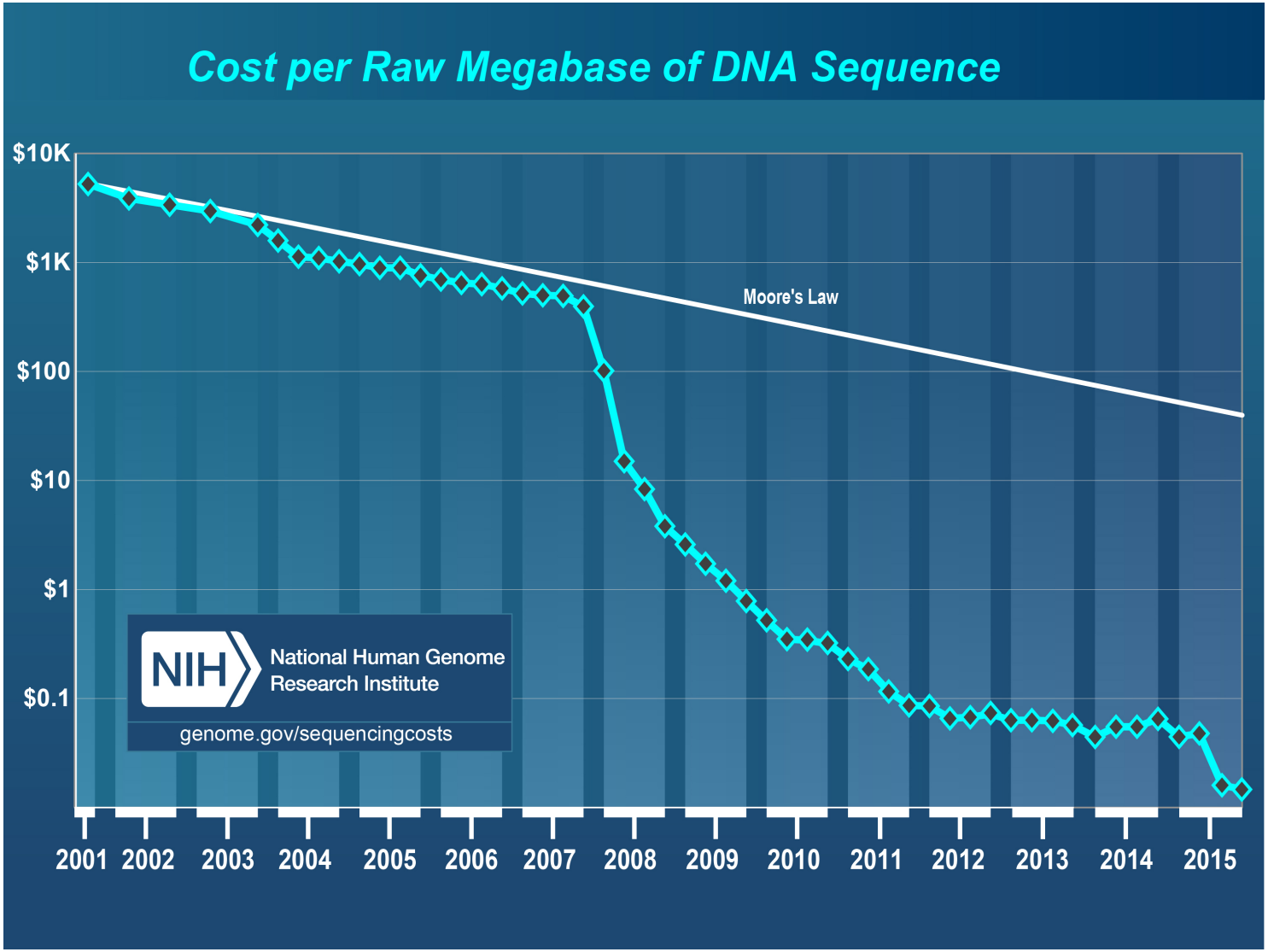
- Output for Quality Tradeoff
  - NGS = High Output / Lower Quality
  - Sanger = Low Output / High Quality
- NGS Methods and Machines
  - IonTorrent
  - Illumina
  - PacBio

Next Generation Sequencing vs. Sanger

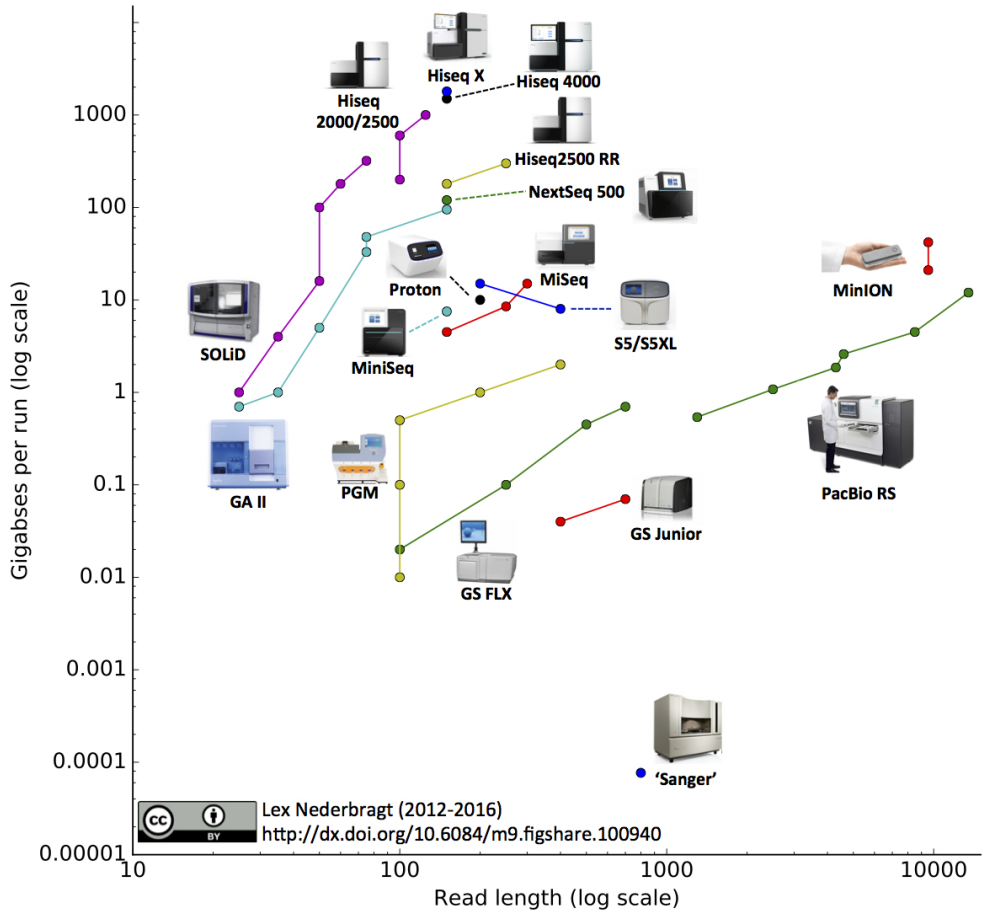
- Output for Quality Tradeoff
  - NGS = High Output / Lower Quality
  - Sanger = Low Output / High Quality
- NGS Methods and Machines
  - PyroSeq
  - IonTorrent
  - Illumina
  - PacBio


Miles Zhang
@ymilesz
Following
This meme is everywhere, so I thought I'd add a biology twist to it.
BIOLOGISTS
NGS
SANGER SEQUENCING
7:11 PM - 25 Aug 2017

Sequencing Cost



Sequencing Output



Generational Shift

- More and more data can be generated cheaper

Generational Shift

- More and more data can be generated cheaper
- Data length and quality are both improving

Generational Shift

- More and more data can be generated cheaper
- Data length and quality are both improving
- How does this change the scope of research?

Generational Shift

- More and more data can be generated cheaper
- Data length and quality are both improving
- How does this change the scope of research?
  - (hint: Sanger is good for studying what?)

Course Motivations

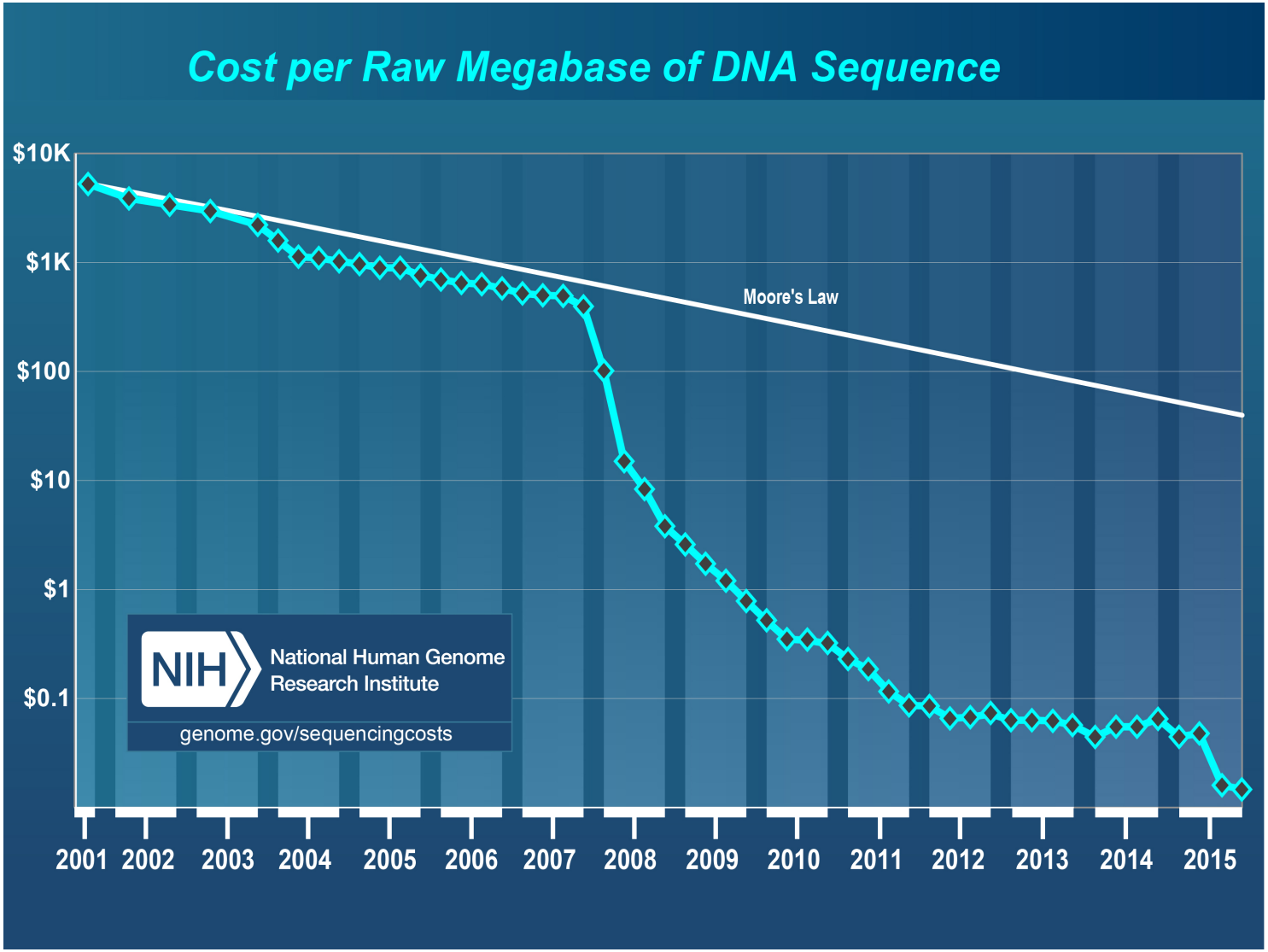
- Biologist increasingly need to be programmers
- Topics often aren't introduced to undergraduates
- Cheap and Free Data

Computational Need

- Biologist increasingly need to be programmers
- Many hypotheses are tested by generating piles of data
- Regardless of your future plans, programming and hypothesis testing are skills that all STEM students should have

Cheap/Free Data

Cheap/Free Data



Cheap/Free Data

- Data is cheap and often free
- Computation is getting faster
- Combined, this means student projects are feasible

Personal Experience

- Worked in Duke IGSP/CHGV Sequencing Core for 4 years
  - Ran Illumina GA, GA2, HiSeq2000
- PhD Thesis on Mouse Lemur Genomics
  - Whole Genome Sequencing and Mutation Rate
  - Rates of Sperm Gene Evolution

The End