

# Bioinformatics: A perspective

Dr. Matthew L. Settles

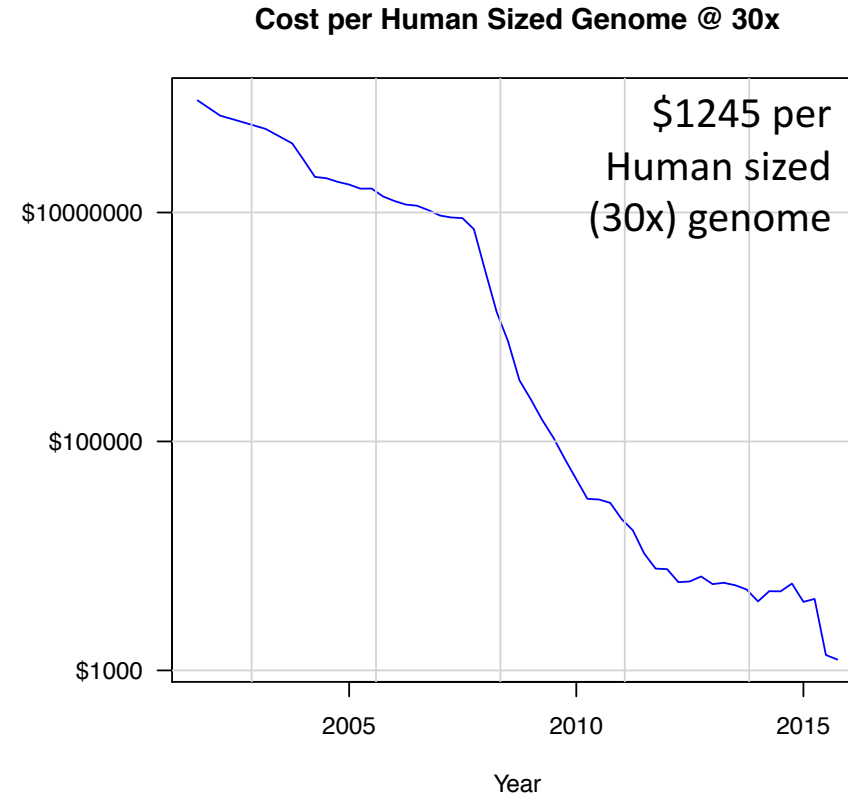
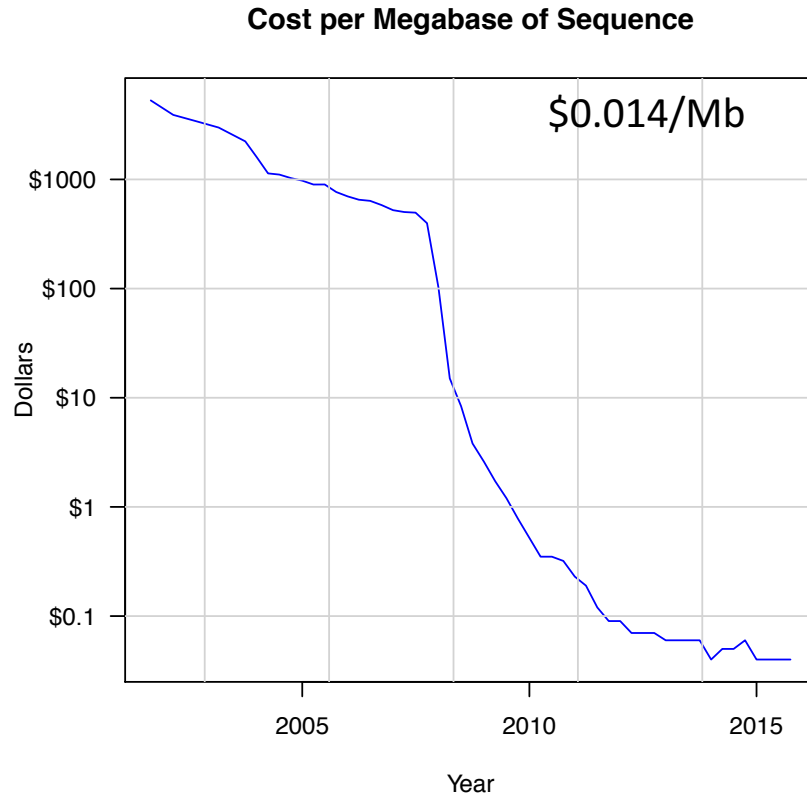
Genome Center  
University of California, Davis  
[settles@ucdavis.edu](mailto:settles@ucdavis.edu)

# Outline

- The World we are presented with
- Advances in DNA Sequencing
- Bioinformatics as Data Science
- Viewport into bioinformatics
- Training
- The Bottom Line

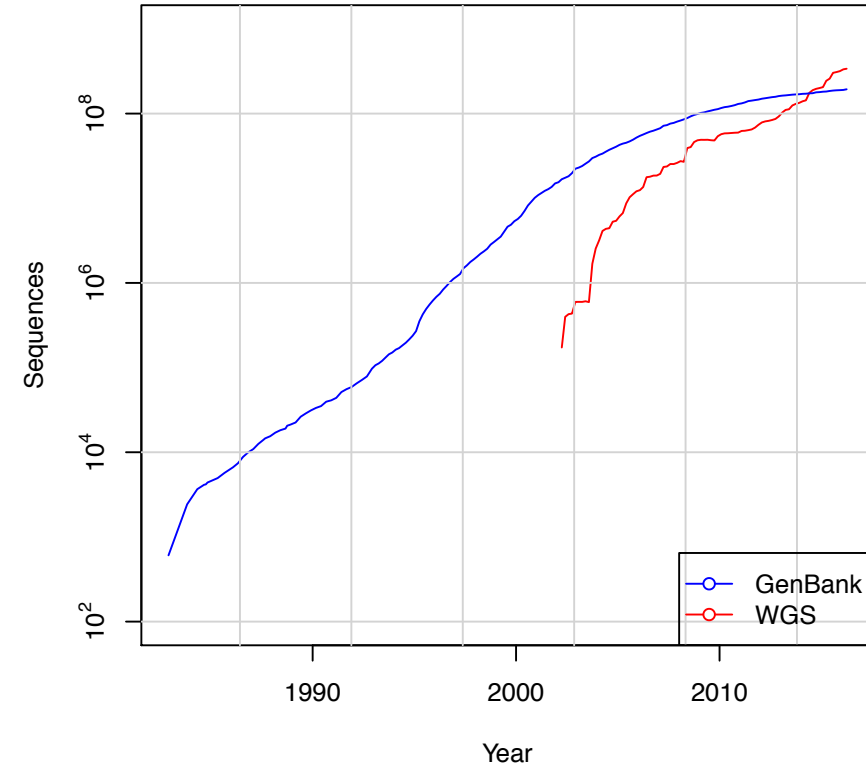
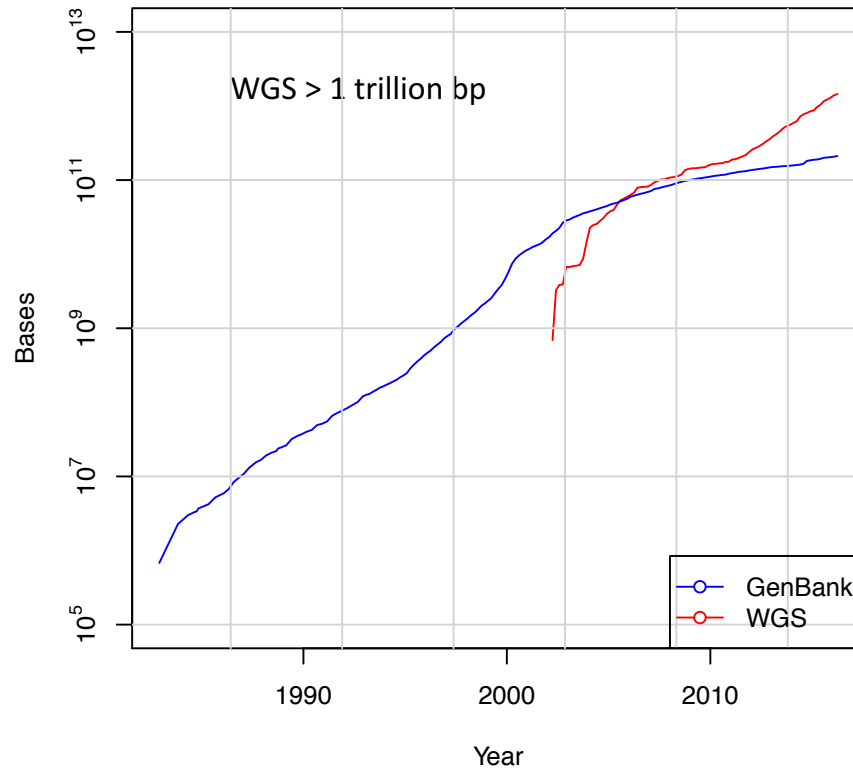
# Sequencing Costs

October 2016



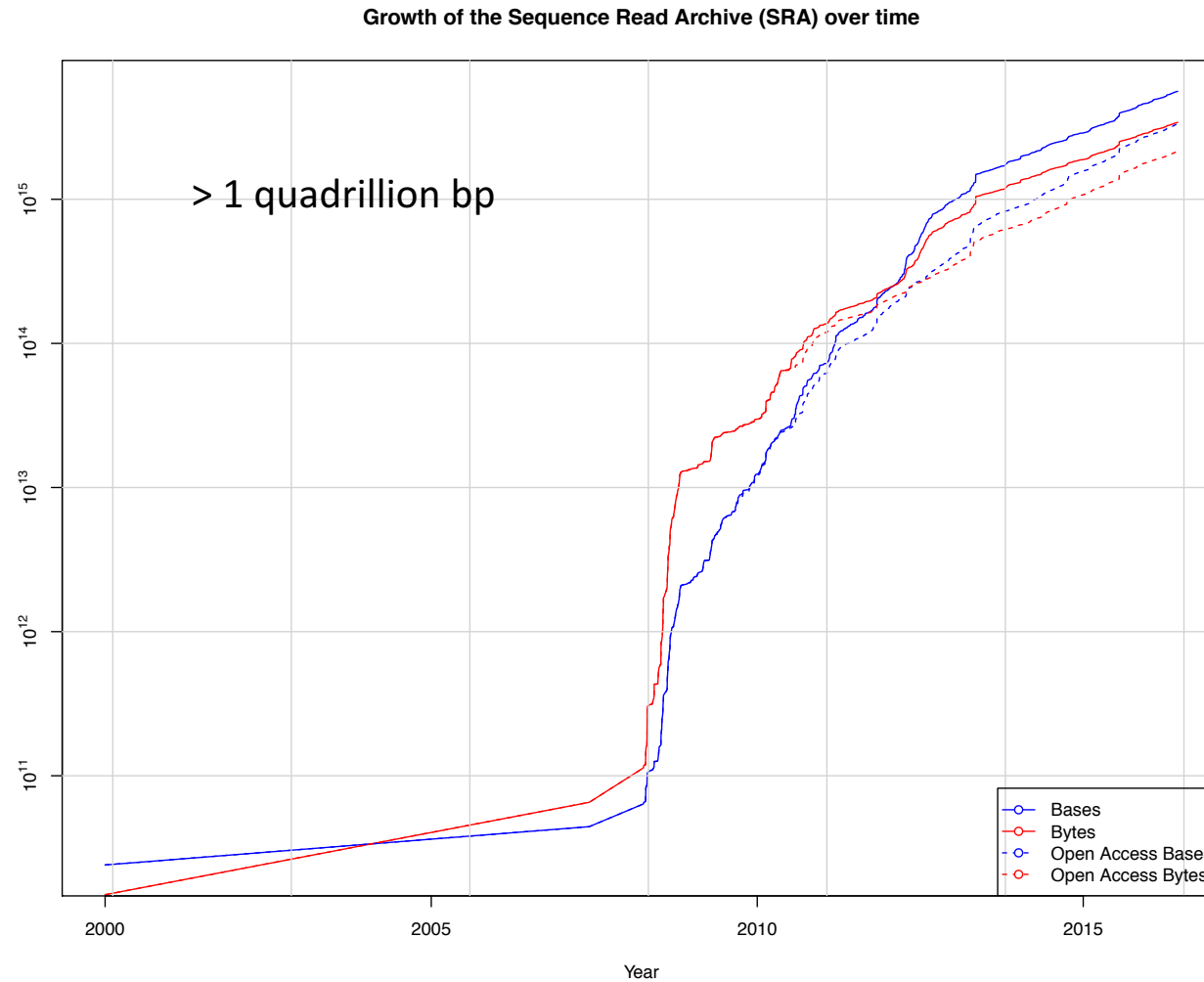
- Includes: labor, administration, management, utilities, reagents, consumables, instruments (amortized over 3 years), informatics related to sequence productions, submission, indirect costs.
- <http://www.genome.gov/sequencingcosts/>

# Growth in Public Sequence Database



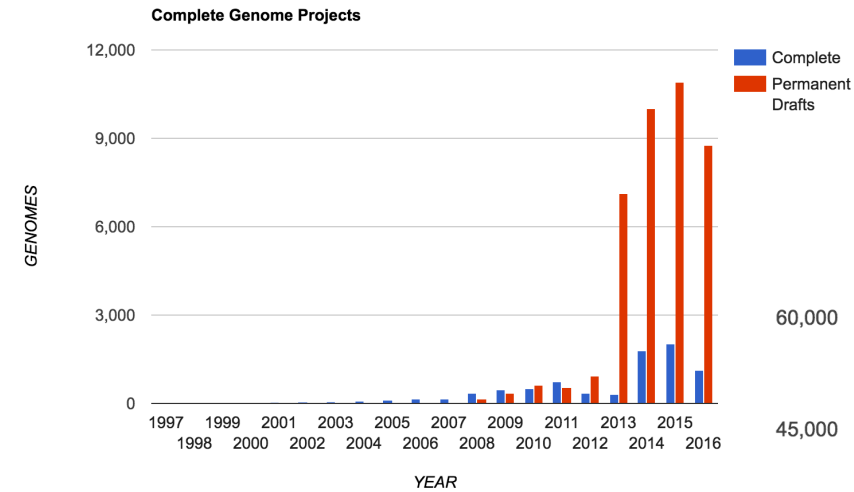
- <http://www.ncbi.nlm.nih.gov/genbank/statistics>

# Short Read Archive (SRA)

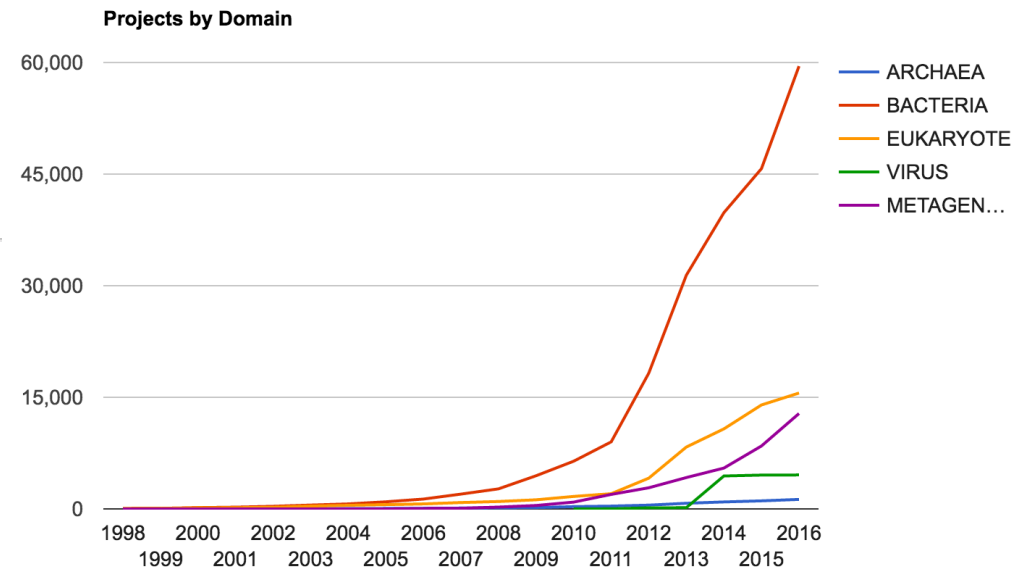


<http://www.ncbi.nlm.nih.gov/Traces/sra/>

# Increase in Genome Sequencing Projects



Lists > 3700 unique genus



- JGI – Genomes Online Database (GOLD)
- 67,822 genome sequencing projects

# Brief History

# Sequencing Platforms

- 1986 - Dye terminator Sanger sequencing, technology dominated until 2005 until “next generation sequencers”, peaking at about 900kb/day





# 'Next' Generation

- 2005 – 'Next Generation Sequencing' as Massively parallel sequencing, both throughput and speed advances. The first was the Genome Sequencer (GS) instrument developed by 454 life Sciences (later acquired by Roche), Pyrosequencing 1.5Gb/day

**Discontinued**



# Illumina

- 2006 – The second ‘Next Generation Sequencing’ platform was Solexa (later acquired by Illumina). Now the dominant platform with 75% market share of sequencer and and estimated >90% of all bases sequenced are from an Illumina machine, Sequencing by Synthesis > 200Gb/day.



# Complete Genomics

- 2006 – Using DNA nanoball sequencing, has been a leader in Human genome resequencing, having sequenced over 20,000 genomes to date. In 2013 purchased by BGI and is now set to release their first commercial sequencer, the Revolocity. Throughput on par with HiSeq

NOW DEFUNCT

Human genome/exomes only.

10,000 Human Genomes per year



# Bench top Sequencers

- Roche 454 Junior
- Life Technologies
  - Ion Torrent
  - Ion Proton
- Illumina MiSeq



# The 'Next Next' Generation

- 2009 – Single Molecule Read Time sequencing by Pacific Biosystems, most successful third generation sequencing platforms, ~2Gb/day, New Pac Bio Sequel ~14Gb/day.



# Oxford Nanopore

- 2015 – Another 3<sup>rd</sup> generation sequencer, founded in 2005 and currently in beta testing. The sequencer uses nanopore technology developed in the 90's to sequence single molecules. Throughput is about 500Mb per flowcell.

**Fun to play with but  
results are highly variable**

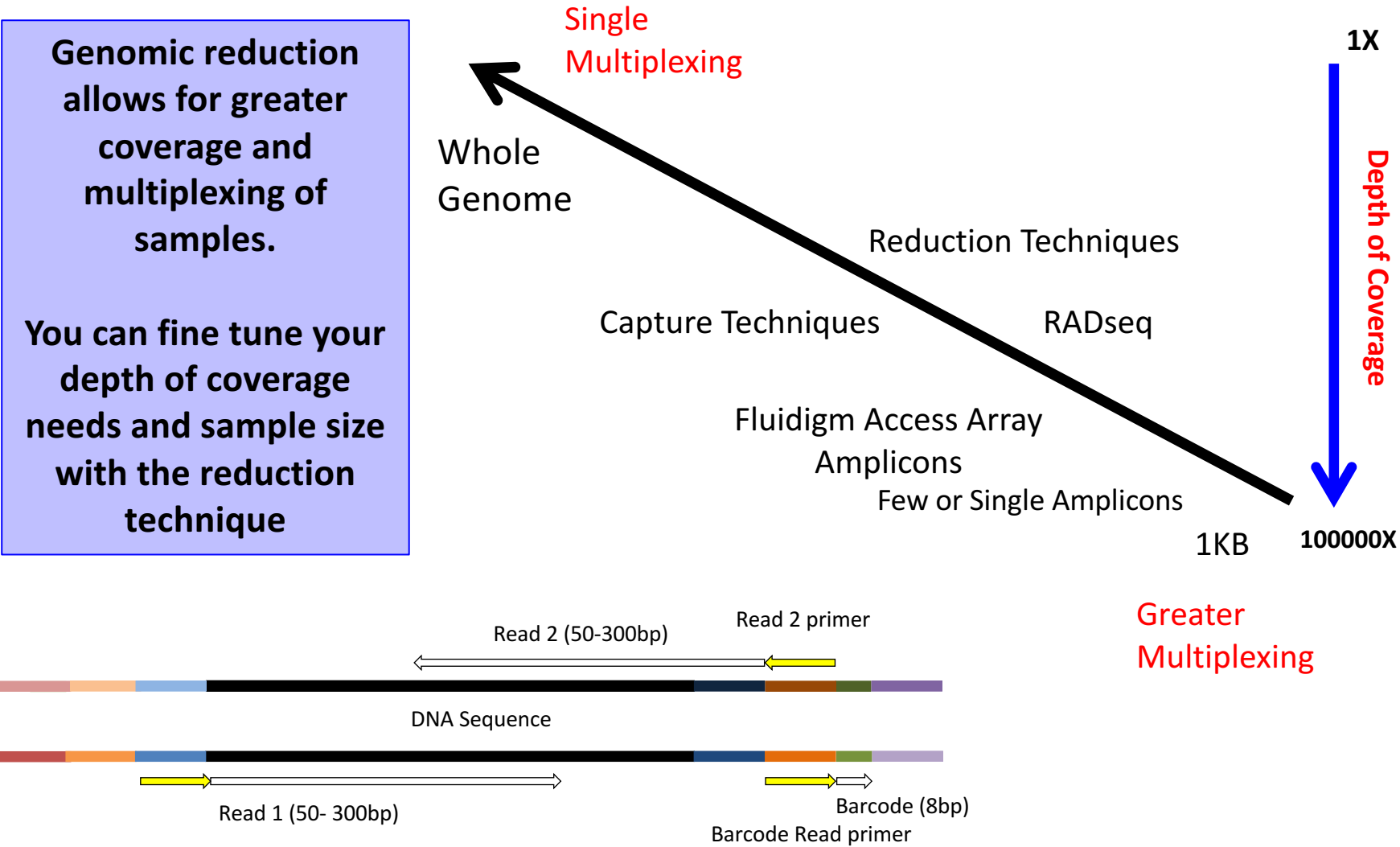


FYI: 4<sup>th</sup> generation sequencing is being described as In-situ sequencing

# Flexibility

Genomic reduction allows for greater coverage and multiplexing of samples.

You can fine tune your depth of coverage needs and sample size with the reduction technique

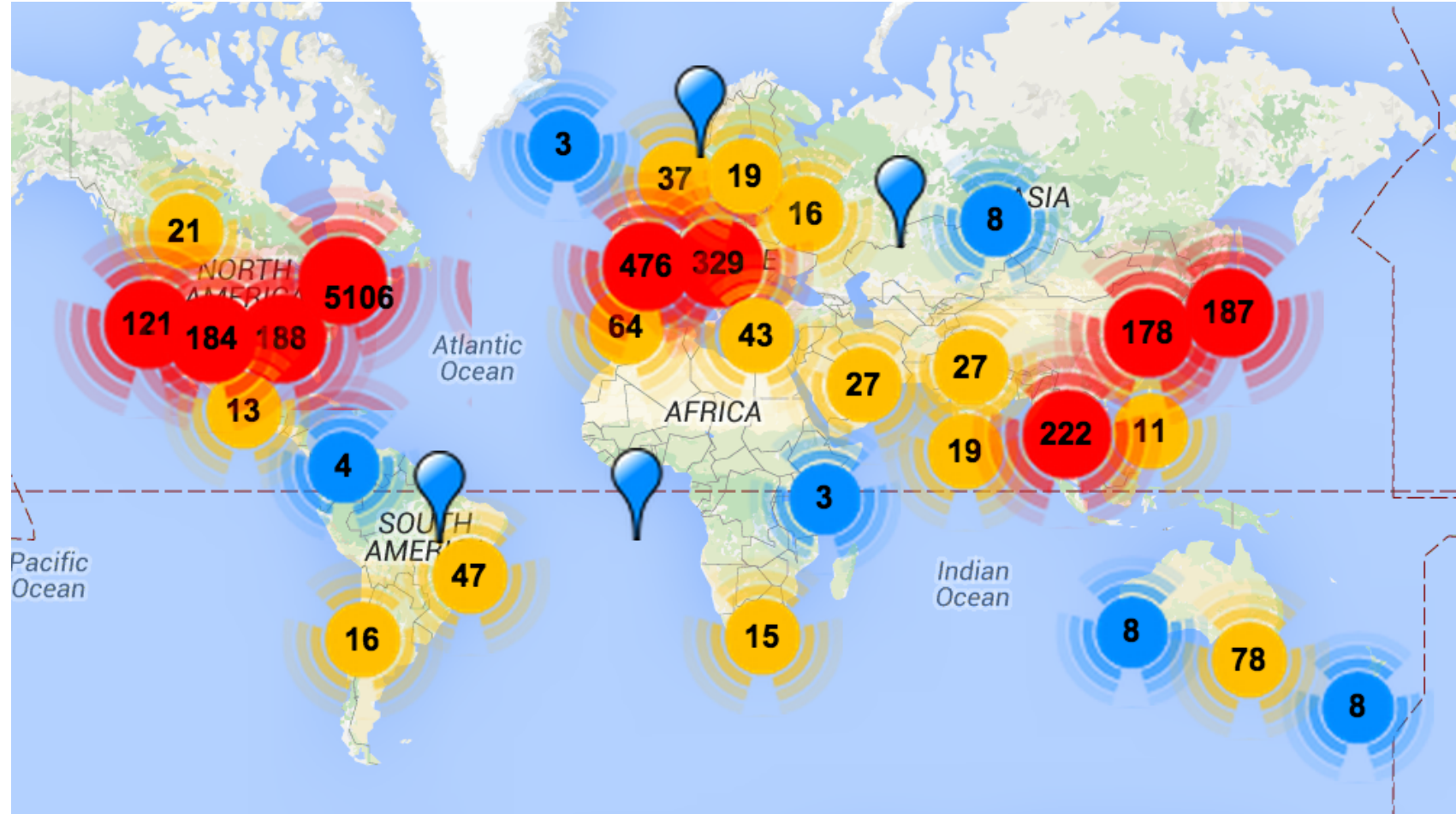


# Sequencing Libraries

• DNA-seq	DNase-seq	tagRNA-seq
• RNA-seq	ATAC-seq	PAT-seq
• Amplicons	MNase-seq	Structure-seq
• ChIP-seq	FAIRE-seq	MPE-seq
• MeDiP-seq	Ribose-seq	STARR-seq
• RAD-seq	smRNA-seq	Mod-seq
• ddRAD-seq	mRNA-seq	BrAD-seq
• Pool-seq	Tn-seq	SLAF-seq
• EnD-seq	QTL-seq	G&T-seq



omicsmaps.com



# The data deluge



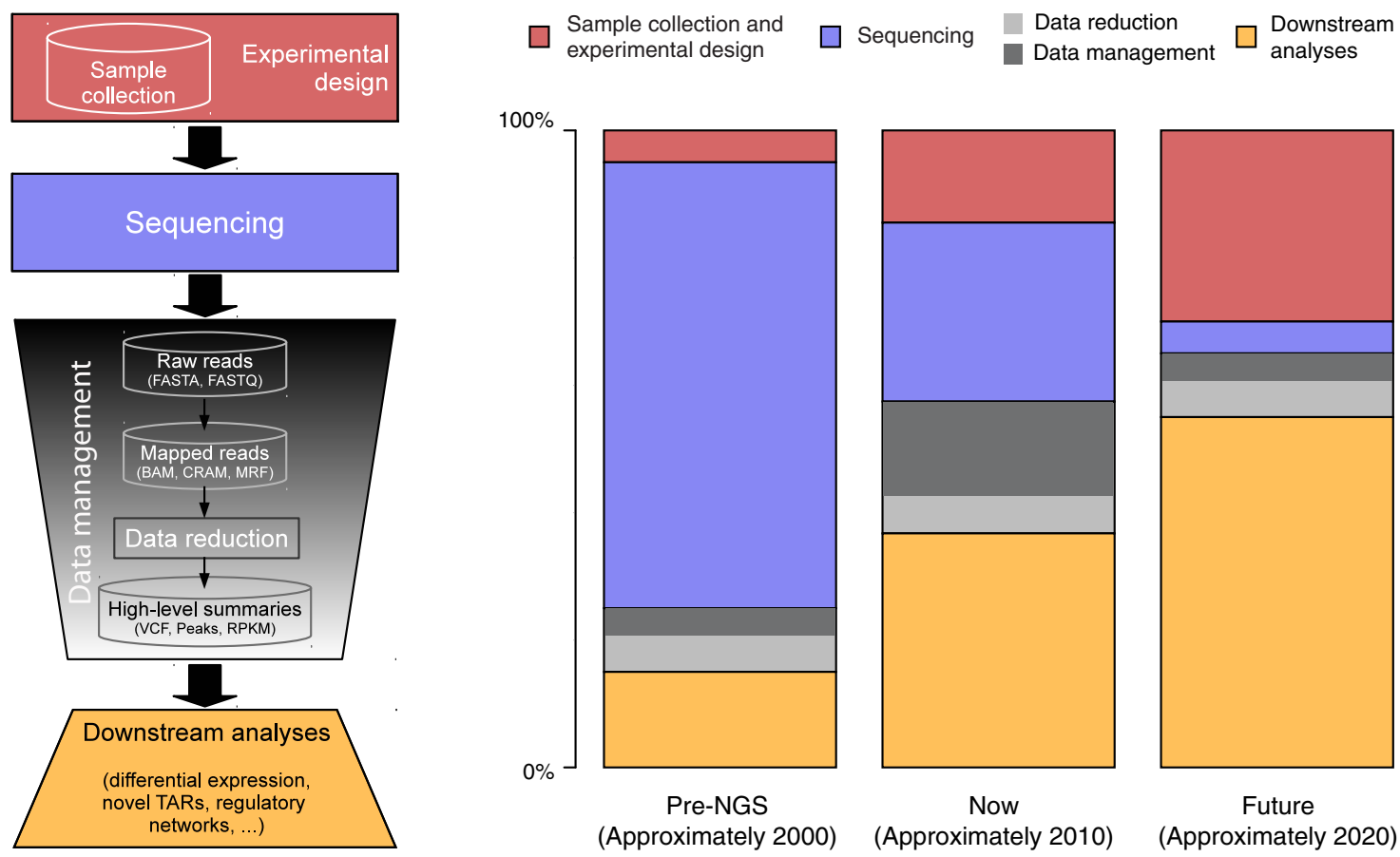
- Plucking the biology from the Noise

# Reality



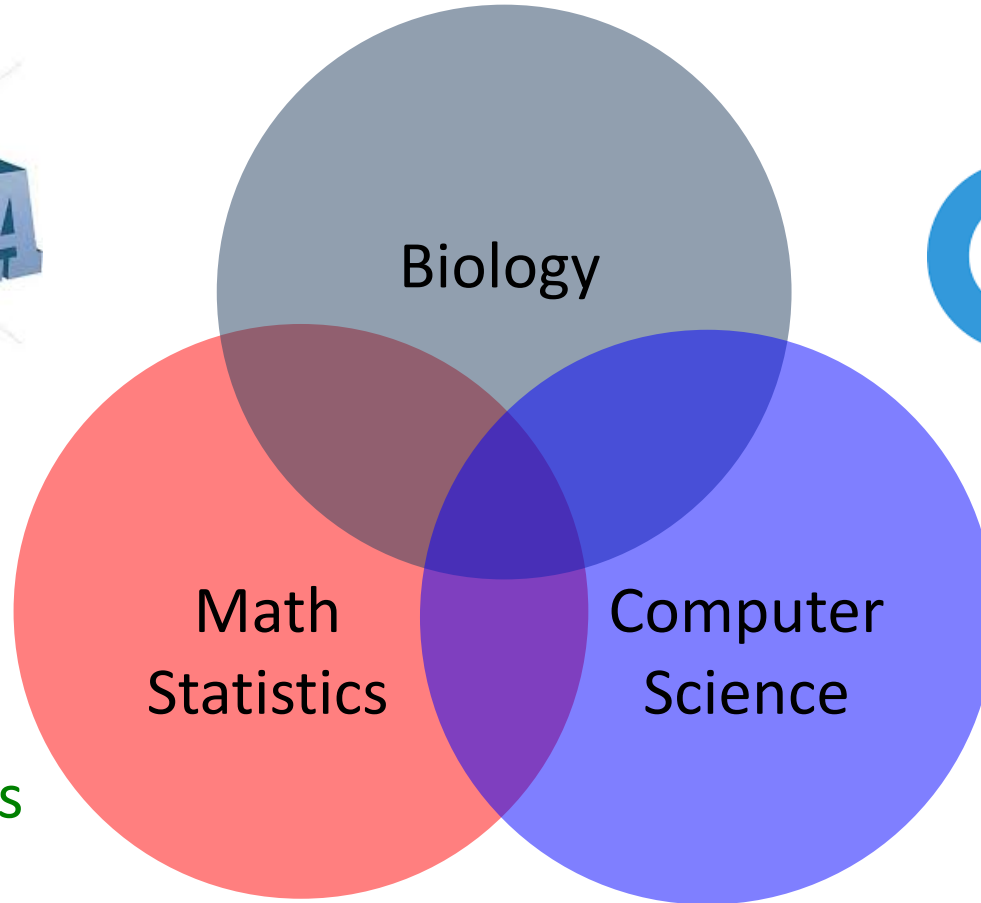
- Its much more difficult than we may first think

# The real cost of sequencing



# Bioinformatics is Data Science

Computational Biology



Biostatistics

Bioinformatics

‘The data scientist role has been described as “part analyst, part artist.”’  
Anjul Bhambhri, vice president of big data products at IBM

# Data Science

Data science is the process of formulating a quantitative question that can be answered with data, collecting and cleaning the data, analyzing the data, and communicating the answer to the question to a relevant audience.



# 7 Stages to Data Science

1. Define the question of interest
2. Get the data
3. Clean the data
4. Explore the data
5. Fit statistical models
6. Communicate the results
7. Make your analysis reproducible

## 1. Define the question of interest

### **Begin with the end in mind!**

what is the question

how are we to know we are successful

what are our expectations

### **dictates**

the data that should be collected

the features being analyzed

which algorithms should be use



2. Get the data
3. Clean the data
4. Explore the data

### **Know your data!**

know what the source was  
technical processing in producing  
data (bias, artifacts, etc.)  
“Data Profiling”

### **Data are never perfect but love your data anyway!**

the collection of massive data sets often leads to unusual ,  
surprising, unexpected and even outrageous.

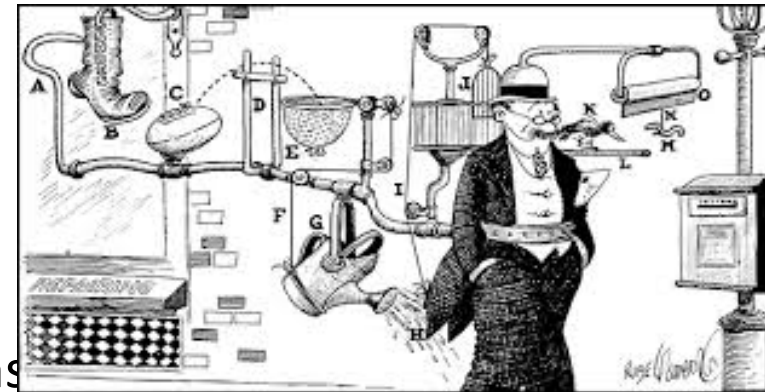


## 5. Fit statistical models

**Over fitting is a sin against data science!**

Model's should not be over-complicated

- If the data scientist has done their job correctly the statistical models don't need to be incredibly complicated to identify important relationships
- In fact, if a complicated statistical model seems necessary, it often means that you don't have the right data to answer the question you really want to answer.



6. Communicate the results
7. Make your analysis reproducible

**Remember that this is 'science'!**

We are experimenting with data selections, processing, algorithms, ensembles of algorithms, measurements, models. At some point these ***must all be tested for validity and applicability*** to the problem you are trying to solve.



**Data science done well looks easy – and  
that's a big problem for data scientists**

**[simplystatistics.org](http://simplystatistics.org)  
March 3, 2015 by Jeff Leek**

# Training: Data Science Bias

Data Science (data analysis, bioinformatics) is most often taught through an apprentice model

Different disciplines/regions develop their own subcultures, and decisions are based on cultural conventions rather than empirical evidence.

- Programming languages
- Statistical models (Bayes vs. Frequentist)
- Multiple testing correction
- Application choice, etc.

These (and others) decisions matter **a lot** in data analysis

*"I saw it in a widely-cited paper in journal XX from my field"*

# The Data Science in Bioinformatics

Bioinformatics is not something you are taught, ***it's a way of life***

*“The best bioinformaticians I know are **problem solvers** – they start the day not knowing something, and they enjoy finding out (themselves) how to do it. It's a great skill to have, but for most, it's not even a skill – it's a passion, it's a way of life, it's a thrill. It's what these people would do at the weekend (if their families let them).”*

Mick Watson – Rosland Institute

# Models

- Workshops
  - Often enrolled too late
- Collaborations
  - More experience persons
- Apprenticeships
  - Previous lab personnel to young personnel
- Formal Education
  - Most programs are graduate level
  - Few Undergraduate



# Bioinformatics

- Know and Understand the experiment
  - “The Question of Interest”
- Build a set of assumptions/expectations
  - Mix of technical and biological
  - Spend your time testing your assumptions/expectations
  - Don’t spend your time finding the “best” software
- Don’t under-estimate the time Bioinformatics may take
- Be prepared to accept ‘failed’ experiments



# Bottom Line

## The Bottom Line:

Spend the time (and money) planning and producing **good quality, accurate and sufficient data** for your experiment.

Get to know to your data, develop and test expectations

Result, you'll **spend much less time** (and less money) extracting biological significance and results during analysis.

# Substrate

Cloud  
Computing



Cluster  
Computing



BAS™

# LINUX

Laptop & Desktop



# Environment

“Command Line” and “Programming Languages”



vs

Bioinformatics Software Suite



# Prerequisites

- Access to a multi-core (24 cpu or greater), 'high' memory 64Gb or greater Linux server.
- Familiarity with the 'command line' and at least one programming language.
- Basic knowledge of how to install software
- Basic knowledge of R (or equivalent) and statistical programming
- Basic knowledge of Statistics and model building