

A Perspective: Genomics and Bioinformatics

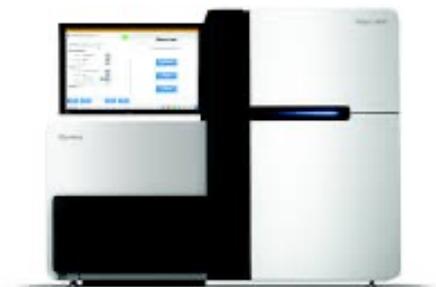
Jie (Jessie) Li, Ph.D.

Genome Center
University of California, Davis
jjqli@ucdavis.edu

Acknowledgement

- Dr. Matthew L. Settles

A Brief History



Sequencing Platforms

- 1986 - Dye terminator Sanger sequencing, peaking at about 900kb/day in early 2000s



'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 (Solexa)

- 2006 – The second ‘Next Generation Sequencing’ platform. Now the dominant platform with 75% market share of sequencer and an estimated >90% of all bases sequenced are from an Illumina machine, Sequencing by Synthesis > 1600Gb/day.

NovaSeq



HiSeq



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 became part of MGI.
- 2020 - Launched DNBSEQ sequencers, which integrate DNA Nanoball technology and PCR-free Rolling Circle Replication. Throughput on par with Novaseq.
- Comparative studies:
 - <https://link.springer.com/article/10.1007%2Fs13258-021-01096-x>
 - <https://academic.oup.com/nargab/article/2/2/lqaa034/5836690>



Bench top Sequencers

❖ Life Technologies

- Ion Torrent
- Ion Proton
- Gene Studio S5



❖ Illumina

- MiSeq
- MiniSeq
- iSeq 100



The 'Next, Next' Generation Sequencers (3rd Generation)

- 2009 – Single Molecule Real Time sequencing by Pacific Biosystems, most successful third generation sequencing platforms, RSII ~2Gb/day, newer Pac Bio Sequel ~14Gb/day, near 100Kb reads
- 2020 - Sequel IIe ~150Gb/flow cell 30 hour run, ~5M HiFi reads.

[SMRT Sequencing](#)



Iso-seq on Pac Bio possible, transcriptome without 'assembly'

Oxford Nanopore



- 2015 – Another 3rd generation sequencer, founded in 2005. The sequencer uses nanopore technology developed in the 90's to sequence single molecules. Throughput is about 500Mb per flowcell, capable of near 200kb reads.
- 2019 – PromethION over 7Tb in a single experiment.
- 2021 – New chemistry promises Q20+ raw sequencing quality.

[Nanopore Sequencing](#)

FYI: 4th generation sequencing is being described as In-situ sequencing



Bioinformatics

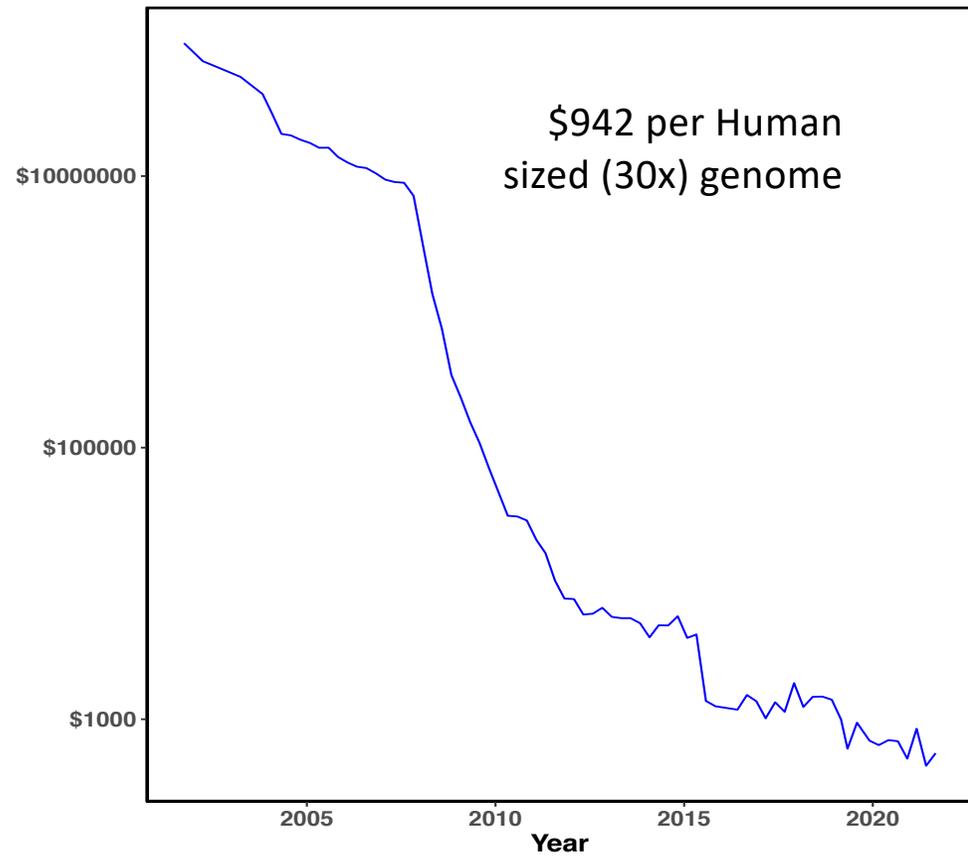
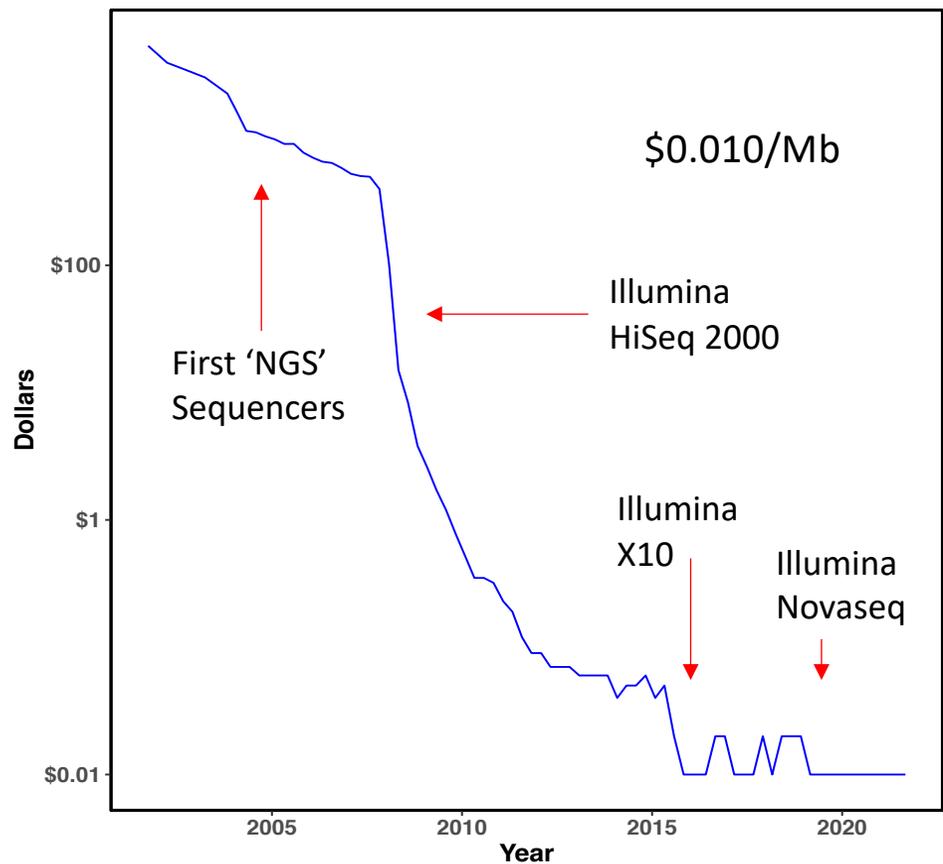
Old Way of
thinking about
Bioinformatics



Appro Cluster

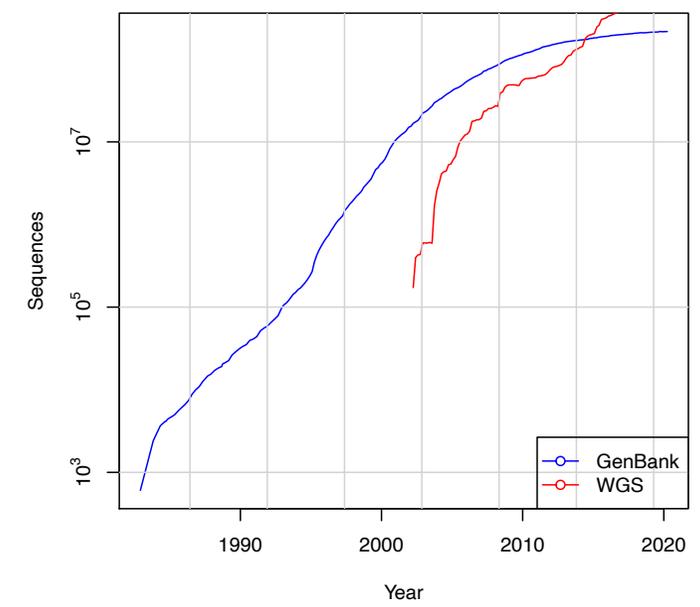
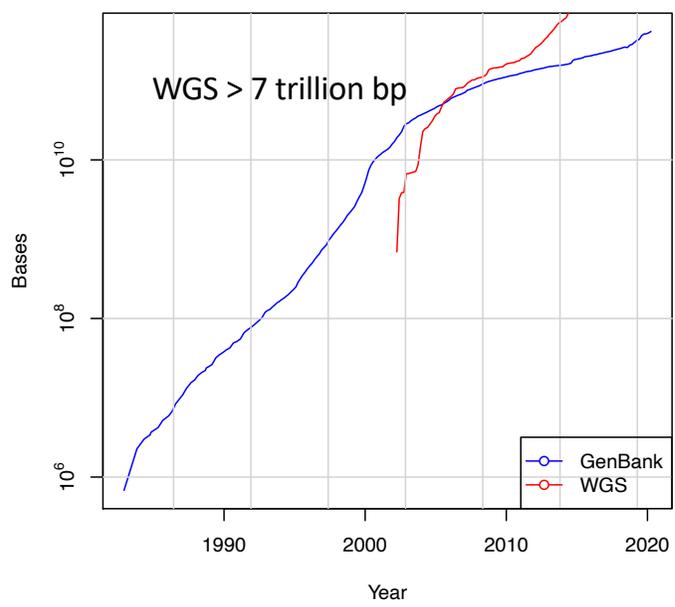
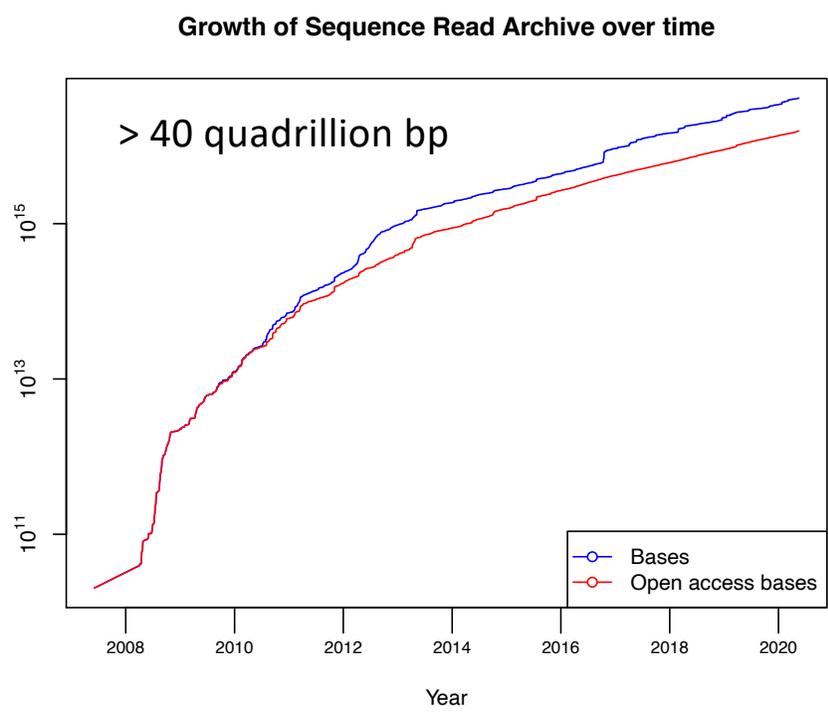
Sequencing Costs

Nov 2021



- 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/>

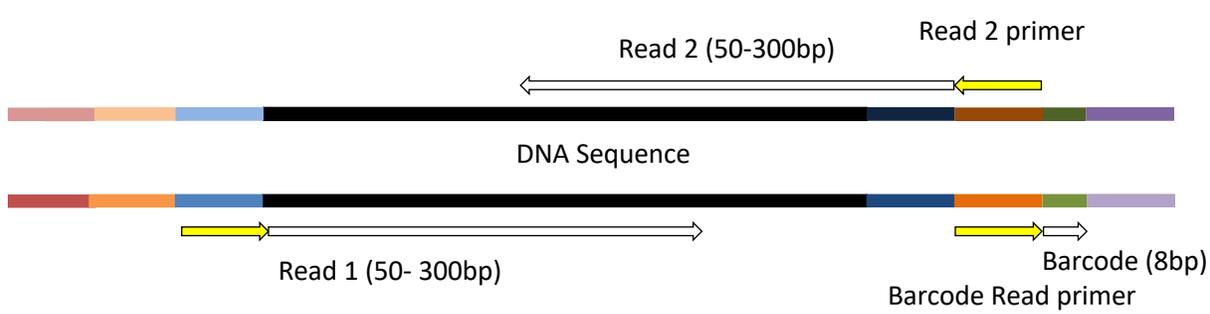
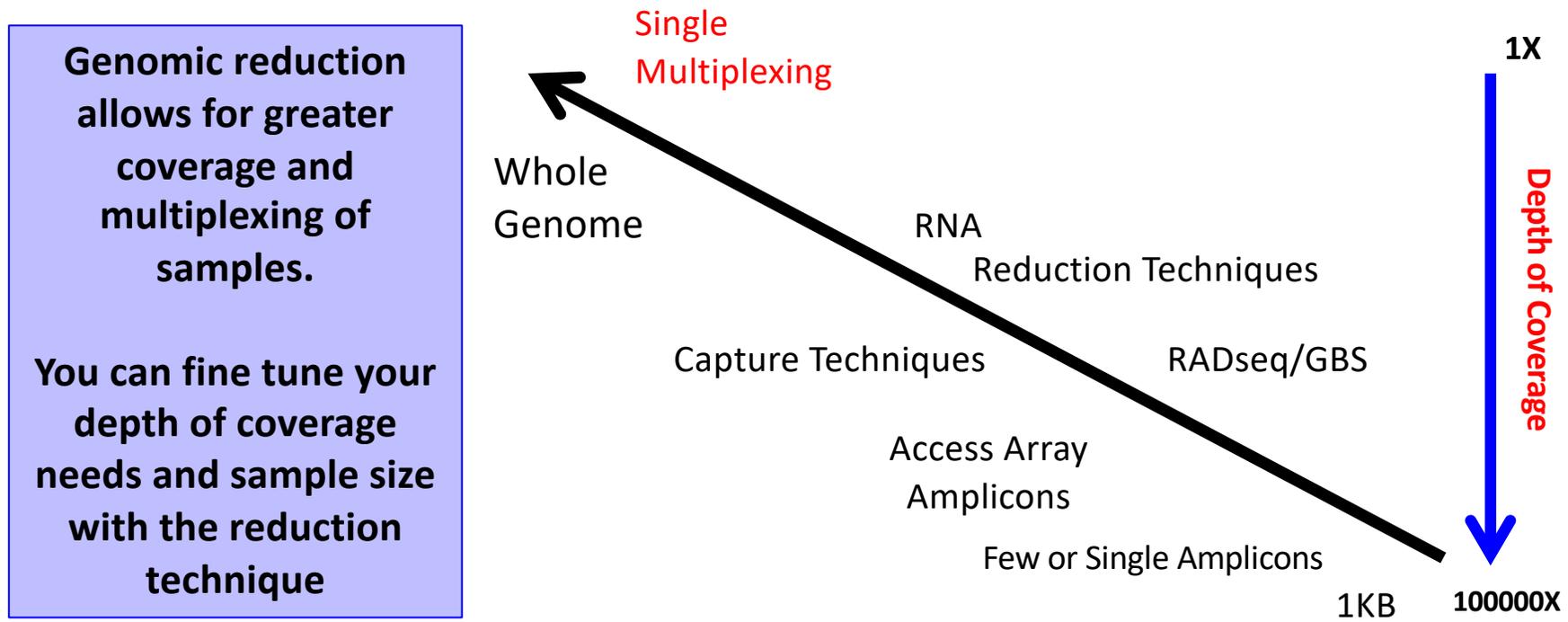
Growth in Public Sequence Database



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

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

Illumina's Flexibility



Greater Multiplexing

Sequencing Libraries:

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

MLA-seq is my favorite seq

The data deluge



- Plucking the biology from the Noise

Reality



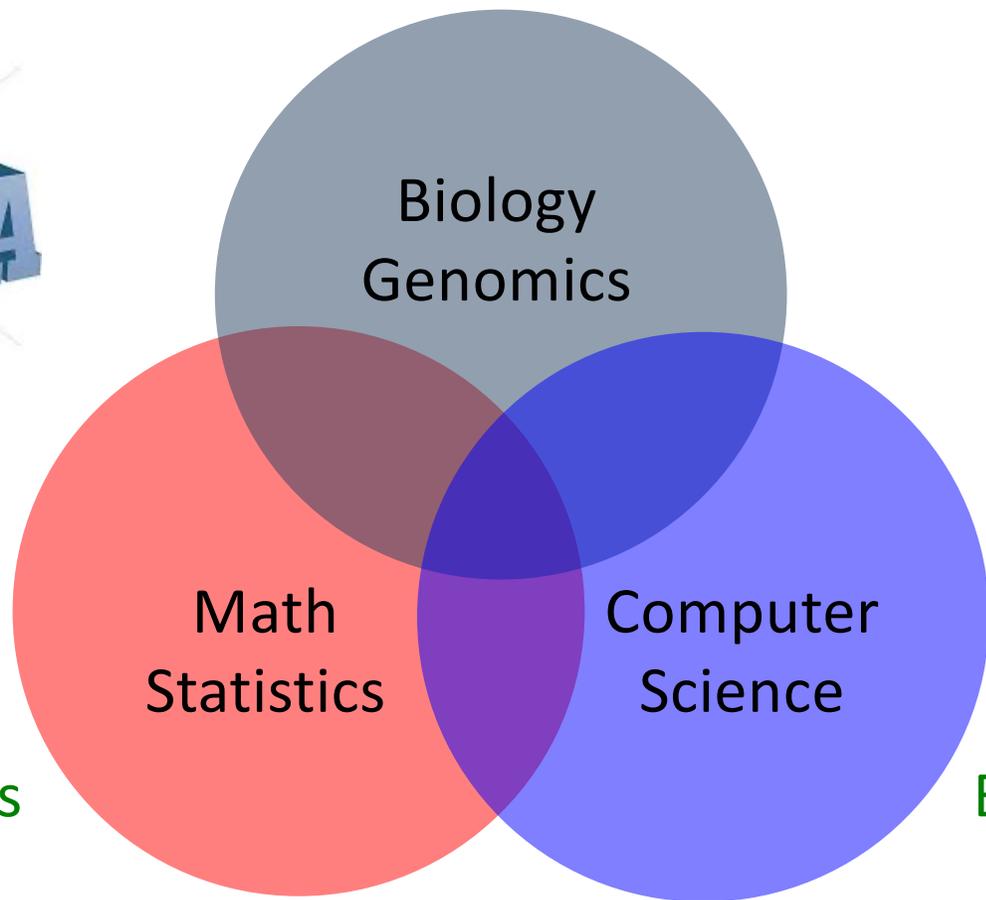
- Its much more difficult than we may first think

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.

Bioinformatician as a Data Scientist

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

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.

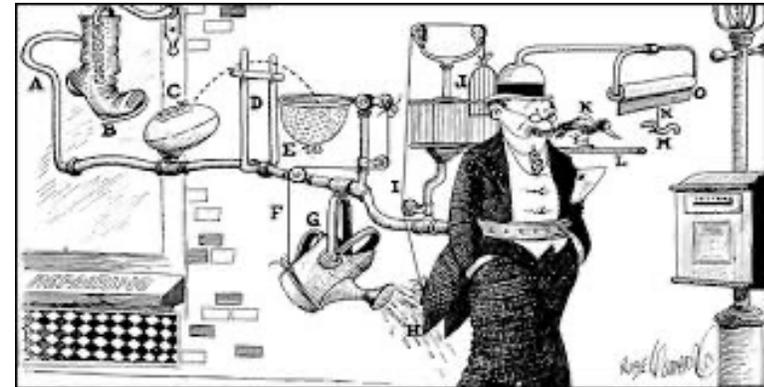


5. Fit statistical models Machine Learning

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.



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

**simplystatistics.org
March 3, 2015 by Jeff Leek**

Bad data science (bioinformatics) also looks easy

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

Genomics and Bioinformatics

Following data science principles, 2 stages in bioinformatics

- **Data reduction**

Sequence data (raw data) to summarized form.

- * Command line, shell scripting, and programming.
- * Requires an understanding of the technology, molecular biology.
- * Removing technical noise from data.

- **Data analysis**

Summarized data to biological interpretation

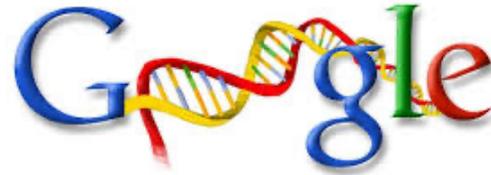
- * R/Python statistical programming
- * Requires an understanding of the biological question, statistics.

Prerequisites for doing Bioinformatics

- 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

Substrate

Cloud Computing



Cluster Computing



BAS™

LINUX

Laptop & Desktop



Environment

“Command Line” and “Programming Languages”



vs

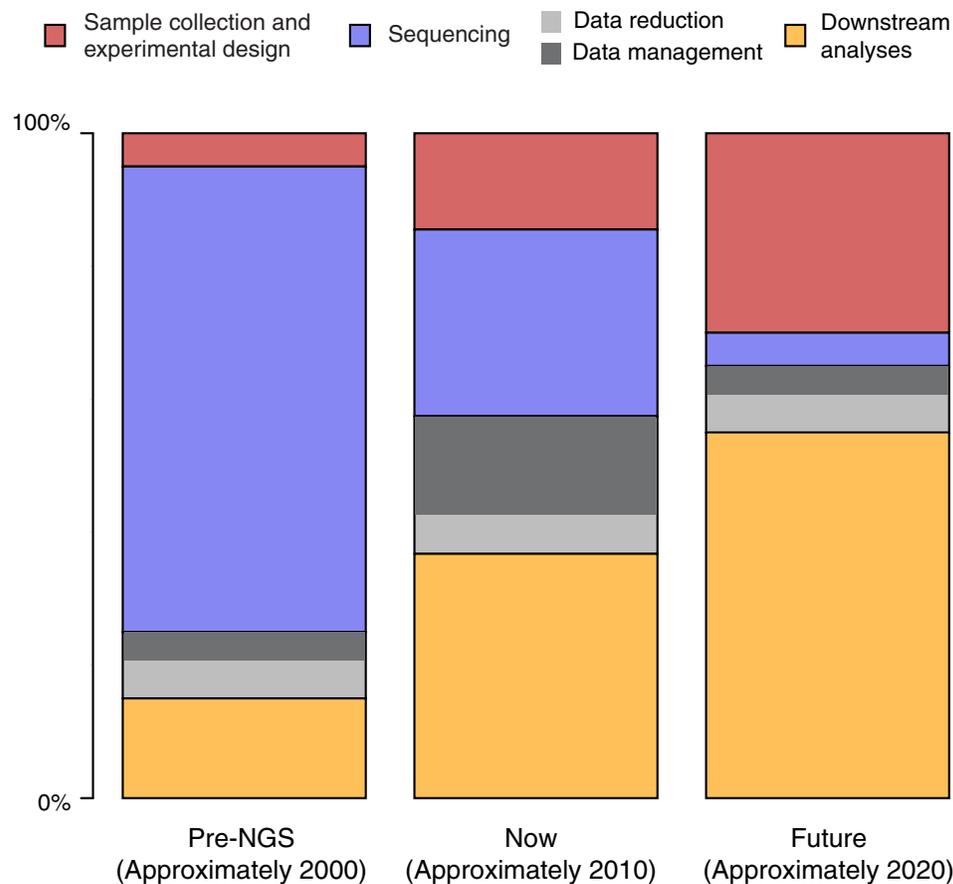
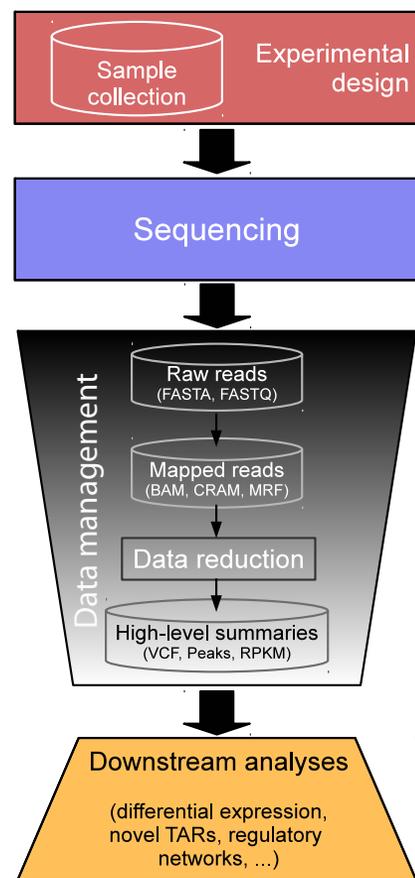
Bioinformatics Software Suite



Training - 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

“The real cost of sequencing”



The last mile



<http://www.bikeblanket.com/blog/suisse>

The Bottom Line: In Genomics

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

Get to know to the data, develop and test expectations, explore and identify patterns.

Result, **spend much less time** (and less money) extracting biological significance and results with fewer failures and reproducible research.