Genomics

Data Analysis & Visualization

Camilo Valdes
cvaldes3@miami.edu
https://github.com/camilo-v
Center for Computational Science, University of Miami • ccs.miami.edu

Today

- **Sequencing Technologies**
  - Background & Fundamentals
  - Commercial Platforms
  - Applications
- **RNA-Seq**
  - RNA & the Transcriptome
  - Typical Workflow
  - Data Analysis
- **Challenges**
  - Data Challenges: Statistical & Computational
- **Visualization**

Sequencing Technologies

DNA Sequencing

- In biochemistry, sequencing means to **determine the primary structure** of an unbranched biopolymer.

- Sequencing results in a **symbolic linear depiction** known as a sequence, which succinctly summarizes much of the atomic-level structure of the sequenced molecule.

- **DNA sequencing** is the process of determining the nucleotide order of a given DNA fragment.

- Most DNA sequencing had, up until a few years ago, been performed using the chain termination method developed by Frederick Sanger in 1977.

Genomes

- The genome sequence of an organism includes the **collective DNA sequences** of each chromosome in the organism.
  - For a bacterium containing a single chromosome, a genome project will aim to map the sequence of that chromosome and any plasmids.
  - For the human species, whose genome includes 22 pairs of autosomes, 2 sex chromosomes, and 1 mitochondrial chromosome, a complete genome sequence will involve 25 separate chromosome sequences.
- Genome projects are efforts that ultimately aim to **determine the complete genome sequence of an organism** and to annotate genes and other important genome-encoded features.

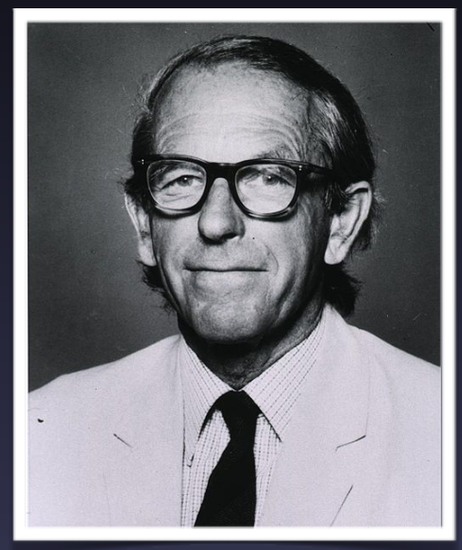
Some History

- 1870, DNA first isolated by the Swiss physician Friedrich Miescher.
- 1953, Watson & Crick describe double-helix model.
- **1977, Frederick Sanger determined the entire genome sequence of the bacteriophage OX174.**
- 1984, The entire sequence of the HIV-1 genome was determined by Chiron Corp.
- 1995, The first genome of a free living organism (H. influenzae) is completely sequenced.
- 1996, The complete genome of the E. coli bacteria was sequenced.
- **2001, First draft of a Human Genome.**
- **2007, High-Throughput Sequencing**

Rosalind Franklin

Sanger Sequencing

- The Sanger sequencing method results in sequence strings (“reads”) that are, on average, ~800 bases long, but may be extended to above 1,000 bases.



- The chief limitation is the small amounts of DNA that can be processed per unit time, referred to as **throughput**, as well as high cost, resulting in it taking roughly 10 years and three billion dollars to sequence the first human genome.

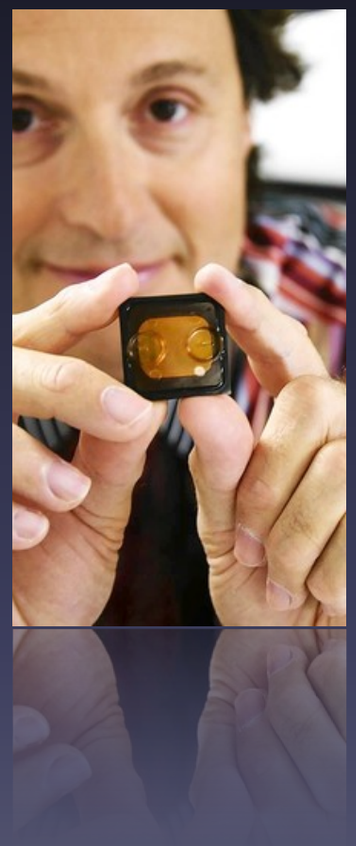
- “Gold Standard” → **high quality** sequence reads

- Considered a ‘first-generation’ technology.

- Newer methods are referred to as “non-Sanger” sequencing, or **Next-Generation Sequencing (NGS)**.

Next-Generation Sequencing, NGS

- Non-Sanger based sequencing.
- Characterized by
  - **Parallel Sequencing**
  - **High Throughput**
  - **Reduced Cost**



- Main difference between Sanger and NGS is **massive parallelization**.
- **Billions** of DNA sequences can be generated in parallel.
- Relatively **Cheap** — to generate data, but analysis…

High-Throughput Sequencing

- NGS expands the realm of experimentation beyond just determining the order of bases.

- These newer technologies constitute various strategies that rely on a combination of **template preparation**, **sequencing**, **imaging**, **genome alignment** and **assembly** methods.

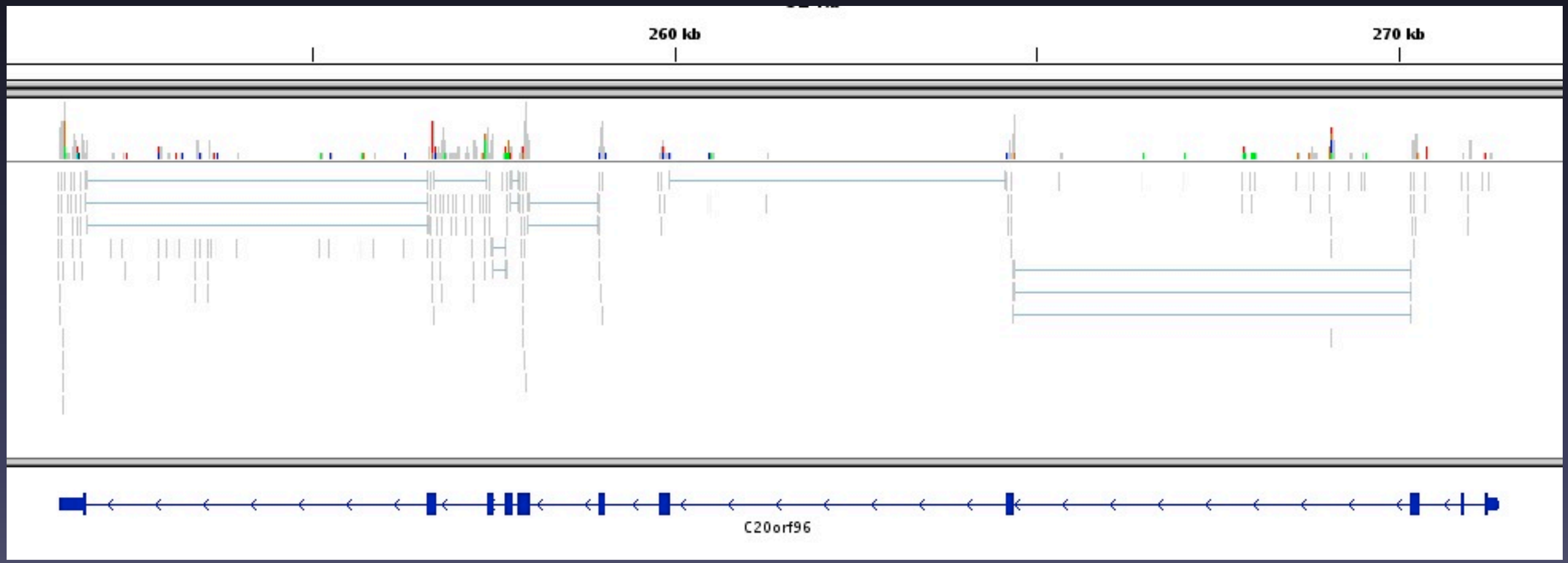
- The arrival of NGS technologies in the marketplace has changed the way we think about scientific approaches in basic, applied and clinical research.

- The major advance offered by NGS is the **ability to produce an enormous volume of data cheaply** — billions of short reads per instrument run.

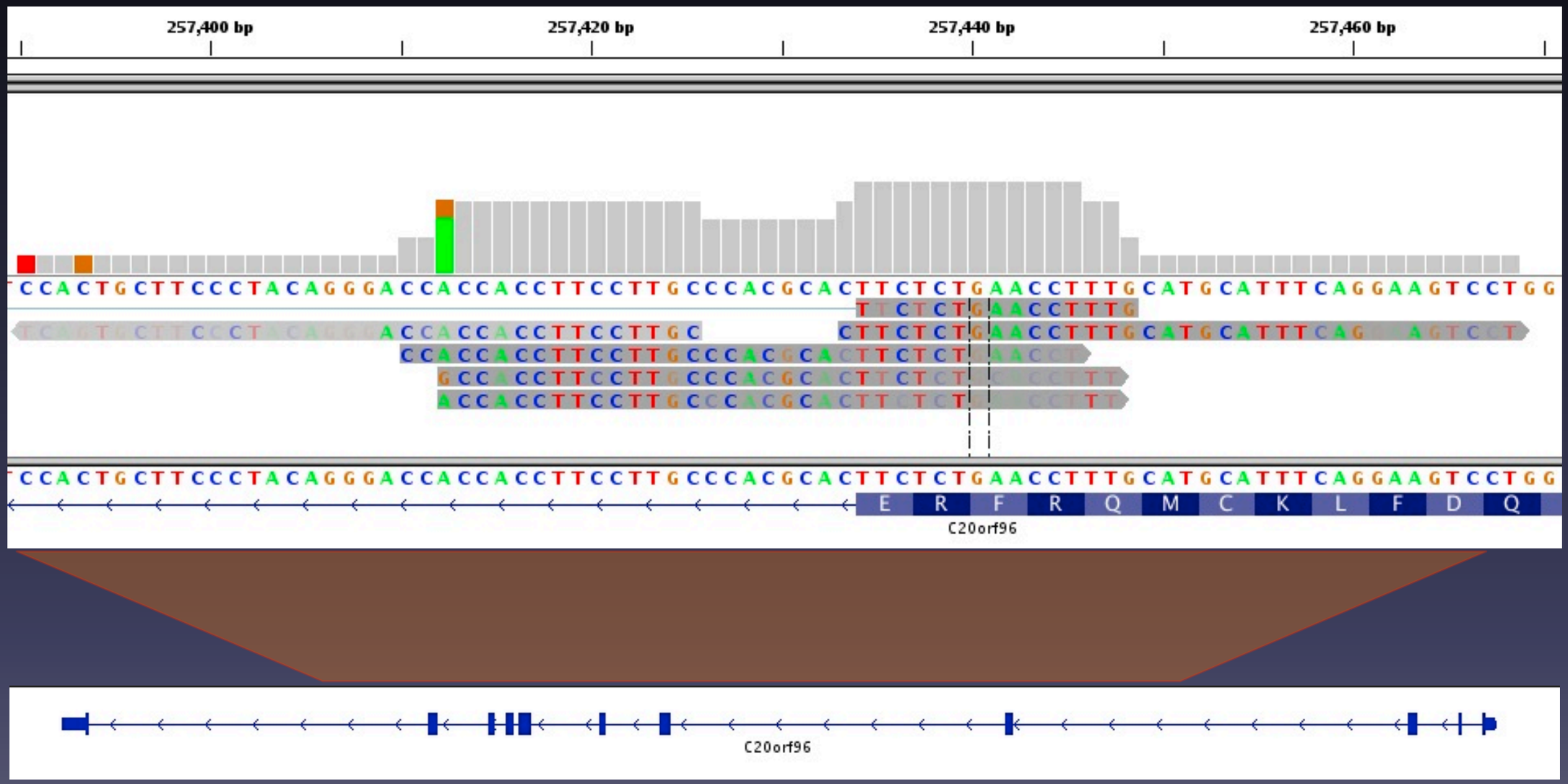
Sequence Coverage

- To sequence a person's genome, many copies of the **DNA are broken into short pieces** and each piece is sequenced.
- The many copies of DNA mean that the DNA pieces are more-or-less randomly distributed across the genome.
- The pieces are then **aligned to the reference sequence** and joined together.
- To find the complete genomic sequence of one person with current sequencing platforms requires **sequencing that person's DNA the equivalent of about 28 times** (called 28X).
- If the amount of sequence done is only an average of once across the genome (1X), then much of the sequence will be missed, because some genomic locations will be covered by several pieces while others will have none.

Genome Coverage



- The **deeper** the sequencing coverage, the more of the genome will be covered at least once.
- Deeper coverage is particularly useful for detecting structural variants, and allows sequencing errors to be corrected.


257,400 bp
257,420 bp
257,440 bp
257,460 bp
E R F R Q M C K L F D Q
C20orf96
C20orf96

Commercial Platforms

illumina®

Oxford NANOPORE Technologies™

Applied Biosystems

PACIFIC BIOSCIENCES™

life technologies™

Helicos BioSciences Corporation

ion torrent

Roche

Commercial Platforms

- NGS platforms differ on their **template preparation**, **sequencing** and **imaging**, and **data analysis** methods.

- The unique combination of specific protocols distinguishes one technology from another, and determines the type of data produced from each platform. Chiefly the **number of sequences generated**, the **length**, and the **quality**.

- Each sequence is a subsequence of the overall genomic sequence — a "**read**".

- These differences in data output present challenges when comparing platforms.

illumina®

PACIFIC BIOSCIENCES™

illumina HiSeq 2500

- 600 GB per run (7-8 days)
  - Paired-End reads @ 2 x 150bp
- 120 GB in 27 hours
  - 1 human genome @ 30X coverage
- 1-2% Error Rate
- On-board Computer
  - 2 Intel XEON Quad-Core Processors
  - 48GB RAM
- 1 Run
  - 2 human genomes @ 30x
- Cost of $10,000 per genome in consumables

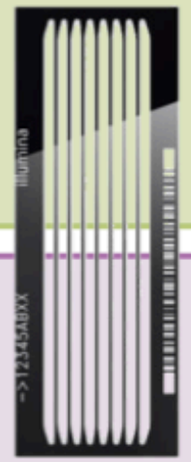
Workflow

1 Library Preparation

Fragment DNA
Repair ends
Add A overhang
Ligate adapters
Purify

2 Cluster Generation

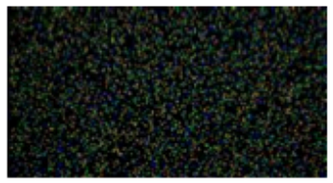
Hybridize to flow cell
Extend hybridized template
Perform bridge amplification
Prepare flow cell for sequencing



3 Sequencing

Perform sequencing
Generate base calls

4 Data Analysis



Images
Intensities
Reads
Alignments

Data Volumes

Data Volume	Total	Final	Comment
HiSeq 2000 200G run			
Image Data	32 TB	0	
Intensity Data	2 TB	0	Optionally transferred
Base Call / Quality Score Data	250 GB	250 GB	1 byte/base (raw) assuming qseq generation offline
Alignment Output	6 TB (3 TB)	1.2 TB	Remove intermediate files
GA_{IIx} 50G run			
Image Data	6.9	0	Optionally transferred
Intensity Data	0.93	0.93	
Base Call / Quality Score Data	0.17	0.17	
Alignment Output	1.2TB	1.2 TB	

1,000,000,000

What do you do with billions of DNA reads?

RNA-Sequencing

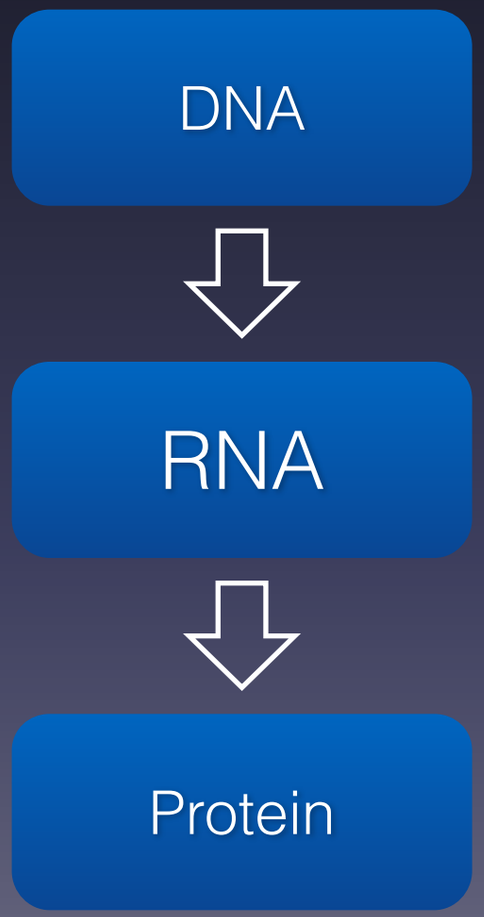
RNA-Sequencing, “RNA-Seq”

- **RNA** is a nucleic acid and, along with proteins and carbohydrates, constitute the three major macromolecules essential for all known forms of life.

- Among other things, RNA **carries the information** from DNA to the ribosomes, the sites of protein synthesis in the cell.

- The **transcriptome** is the set of all RNA molecules in a cell.

- **RNA-Seq** is the application of NGS to reveal a **snapshot** of the **RNA presence** and **quantity** from a genome at a given moment in time.



Transcriptome Analysis

- The main goal is basically to identify all expressed genes & transcripts within a cell, and to determine their structures and measure their abundances

- Some specific aims:

  - Catalogue all species of RNA transcripts

    - mRNAs, noncoding RNAs, long-intergenic RNAs, etc.

  - Accurately quantitate gene expression in a cell.

  - Quantify the changing expression levels of each transcript during development, or under different conditions, i.e., **differential gene expression**.

Gene Expression

The ability to quantify the level at which a particular gene is expressed within a cell, tissue or organism can give a huge amount of information:

- Identify viral infection of a cell
- Determine an individual's susceptibility to cancer
- Find if a bacterium is resistant to penicillin

Ideally measurement of expression is done by detecting the final gene product, however it is often easier to detect one of the precursors, typically mRNA.

Digital Counts

RNA-Seq provides a "**digital measure**" of the presence and prevalence of mRNA transcripts from **known** and previously **unknown** genes.

Mapping and quantifying mammalian transcriptomes by RNA-Seq

Ali Mortazavi[1,2], Brian A Williams[1,2], Kenneth McCue[1], Lorian Schaeffer[1] & Barbara Wold[1]

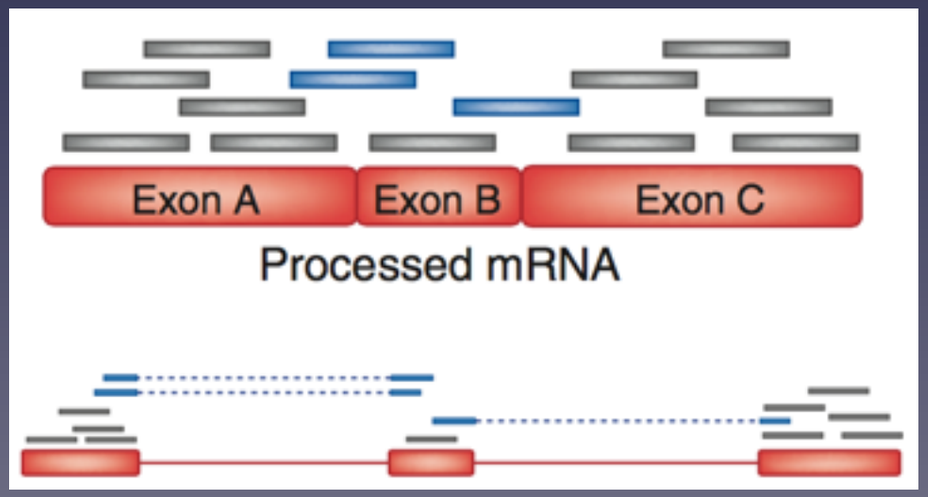
We have mapped and quantified mouse transcriptomes by deeply sequencing them and recording how frequently each gene is represented in the sequence sample (RNA-Seq). This provides a digital measure of the presence and prevalence of transcripts from known and previously unknown genes. We report reference measurements composed of 41–52 million mapped 25-base-pair reads for poly(A)-selected RNA from adult mouse brain, liver and skeletal muscle tissues. We used RNA standards to quantify transcript prevalence and to test the linear range of transcript detection, which spanned five orders of magnitude. Although >90% of uniquely mapped reads fell within known exons, the remaining data suggest new and revised gene models, including changed or additional promoters, exons and 3′ untranscribed regions, as well as new candidate microRNA precursors. RNA splice events, which are not readily measured by standard gene expression microarray or serial analysis of gene expression methods, were detected directly by mapping splice-crossing sequence reads. We observed 1.45×10^5 distinct splices, and alternative splices were prominent, with 3,500 different genes expressing one or more alternate internal splices.

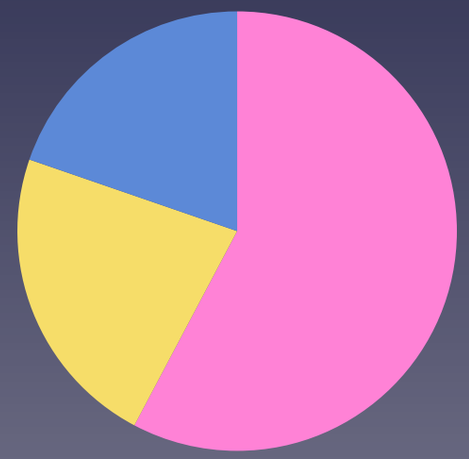
The mRNA population specifies a cell's identity and helps to gov-

approaches to large-scale RNA analysis are serial analysis of gene expression (SAGE)[4,5] and related methods such as massively parallel signature sequencing (MPSS)[6], which use DNA sequencing of previously cloned tags 17–25 base pairs (bp) from terminal 3′ (or 5′) sequence tags. These sequence tags are then identified by informatic mapping to mRNA reference databases or, for longer tag lengths, to the source genome. A strength of SAGE and SAGE-like methods is that they produce digital counts of transcript abundance, in contrast to the analog-style signals obtained from fluorescent dye–based microarrays. However, SAGE-family assays provide no information about splice isoforms or new gene discovery, and fully comprehensive measurements of lower-abundance-class RNAs have not been achieved owing to cost and technology constraints. Expressed sequence tag (EST) sequencing of cloned cDNAs has long been the core method for reference transcript discovery[7–9]. It has both qualitative and quantitative limitations, imposed partly by historic sequencing capacity and cost issues, and more crucially by bacterial cloning constraints that affect which sequences are represented and how sequence-complete each clone is. Recently, dense whole-genome tiling microarrays have been developed and applied to transcriptomes for measuring expression and for transcript discovery[10–14]. In contrast to expression arrays, these tiling arrays can discover new genes and

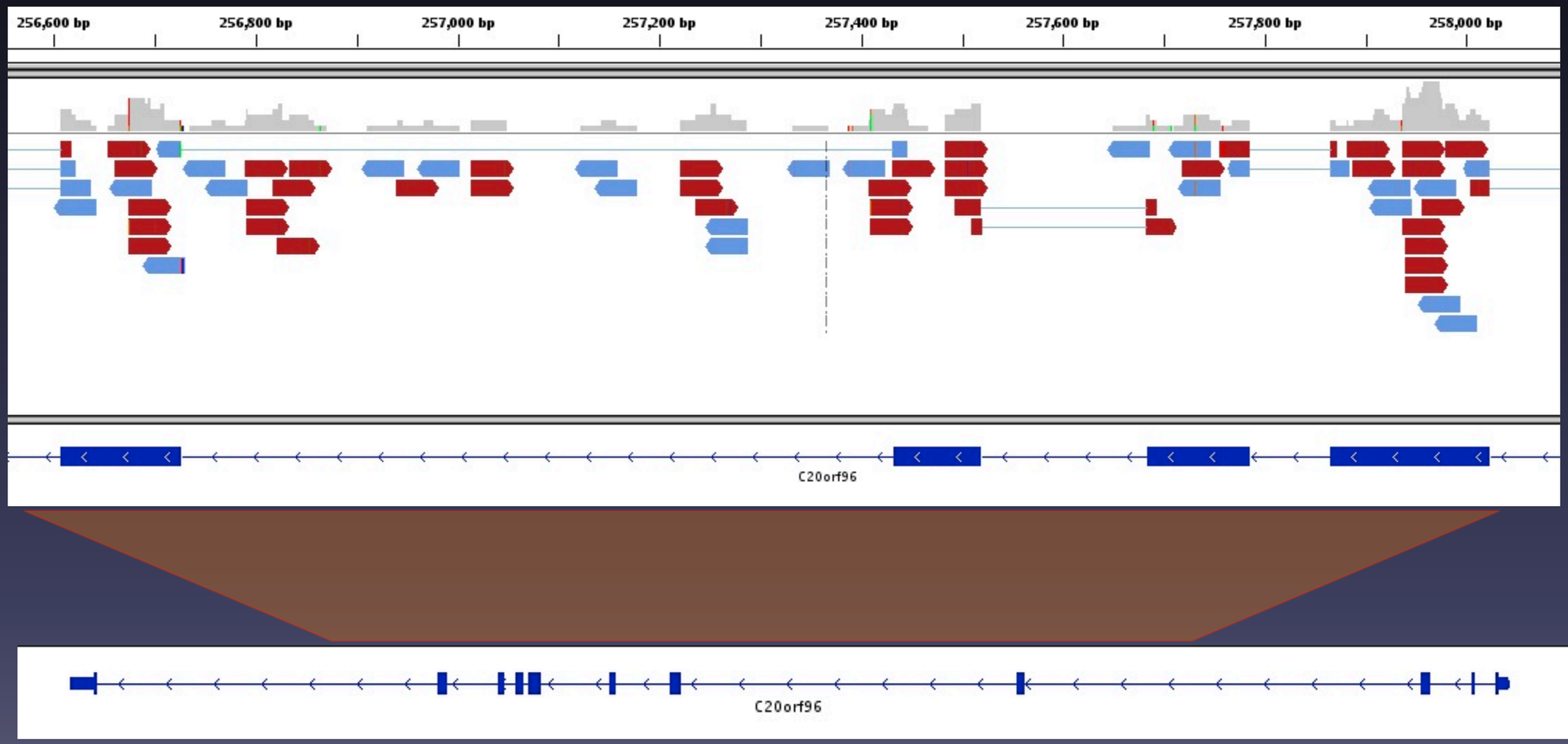
Counting Experiments

- Some RNA-Seq projects begin with an existing backbone “reference” sequence (**mapping**), while others attempt **de-novo** assembly.
- In a **mapping project**, the position of each sequence read within the reference transcriptome must be determined.
- In a **de-novo project**, the architecture of the transcriptome is created without any reference guide.





Trapnell, Cole, and Steven L Salzberg. “How to Map Billions of Short Reads Onto Genomes..” Nature Biotechnology 27, no. 5 (May 2009): 455–57. doi:10.1038/nbt0509-455.

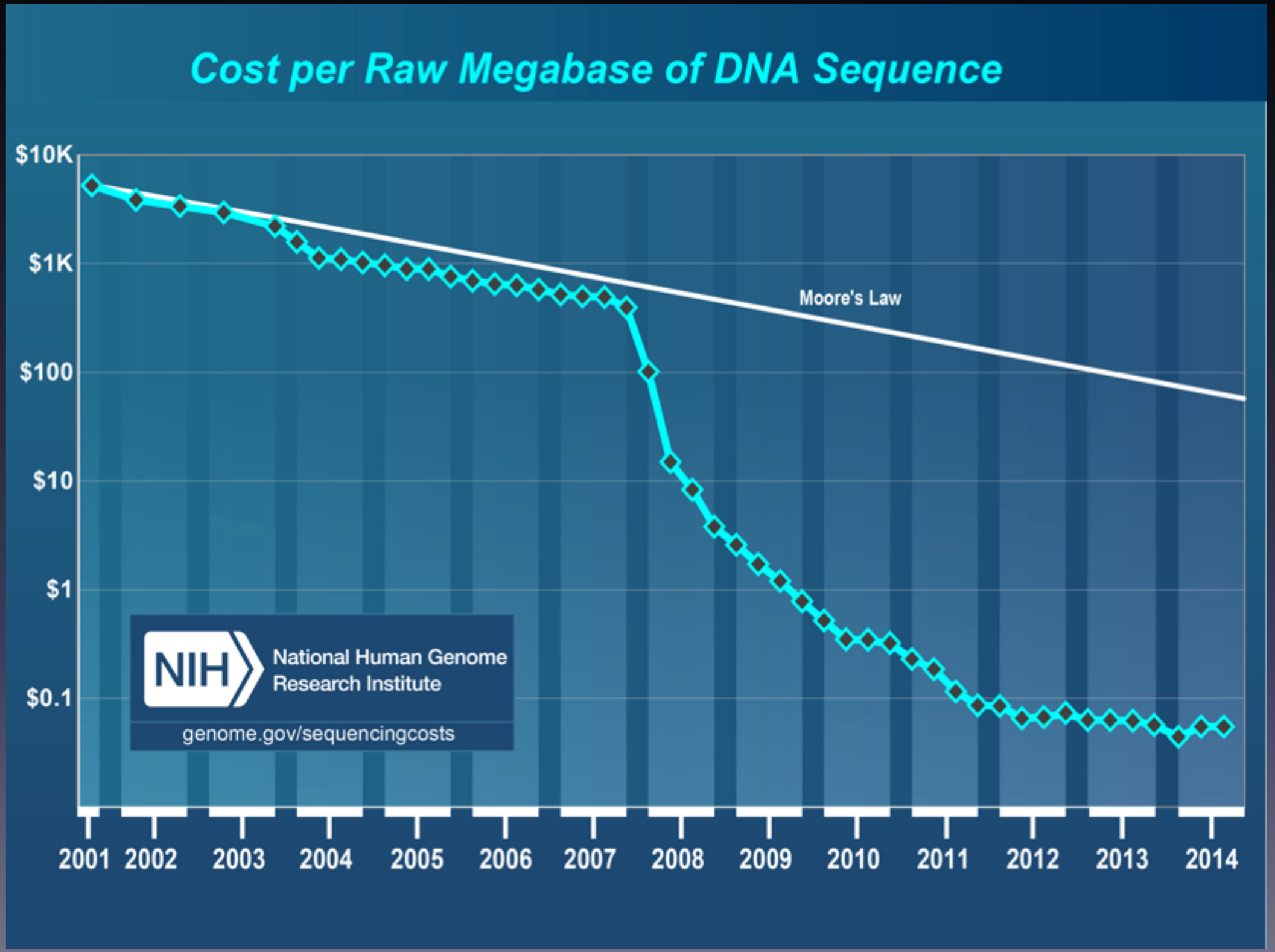

256,600 bp
256,800 bp
257,000 bp
257,200 bp
257,400 bp
257,600 bp
257,800 bp
258,000 bp
C20orf96
C20orf96

Challenges

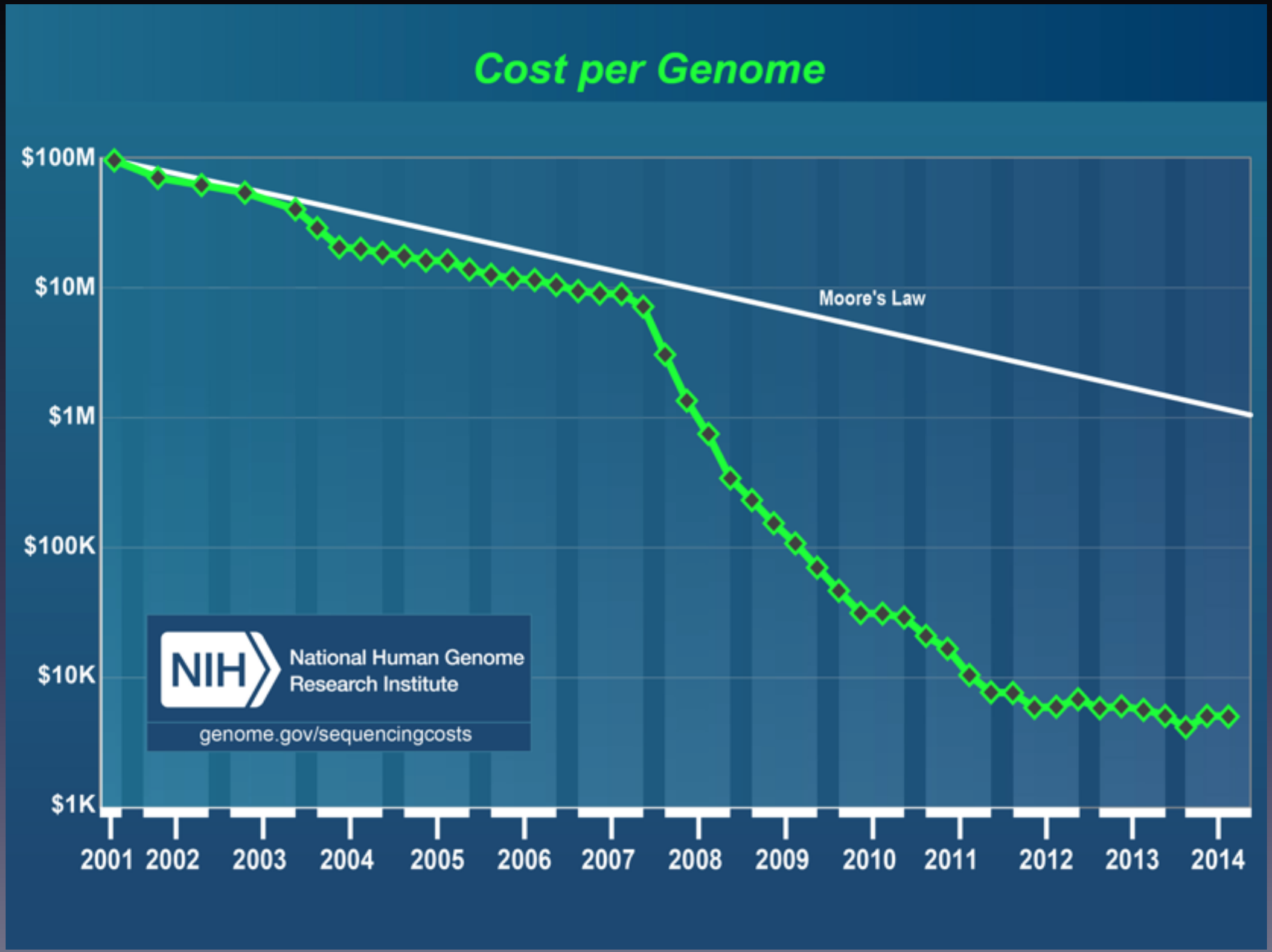
Small Data

- **Data** is “measurements of something”… in our case, its biological measurements.
- But the data has not gotten “Big”, its gotten **smaller**. More **precise**. More **accurate**.
- Sequencing data has allowed us to study things that were impossible just a few years ago.

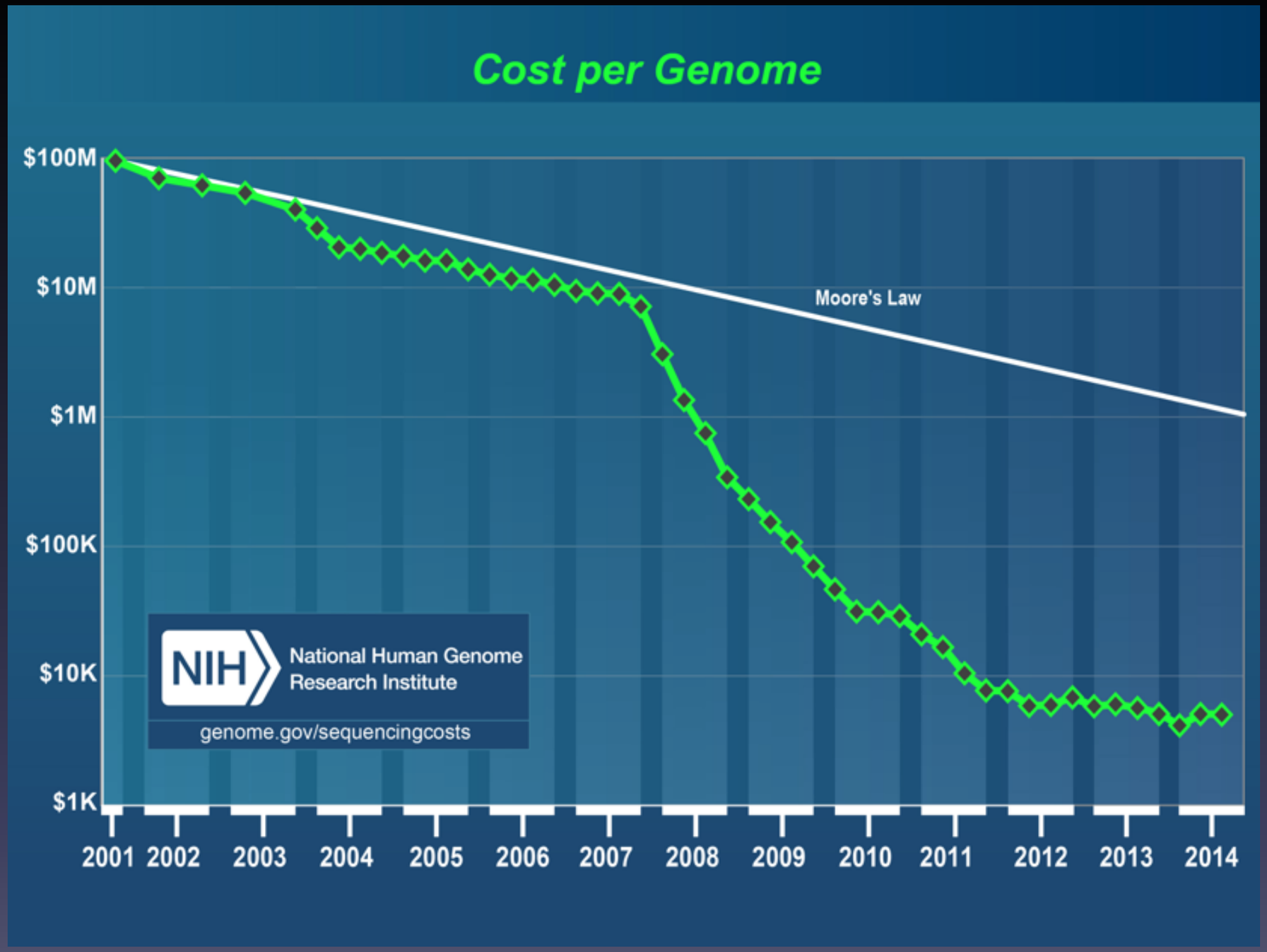
Imielinski, Marcin, Alice H Berger, Peter S Hammerman, Bryan Hernandez, Trevor J Pugh, Eran Hodis, Jeonghee Cho, et al. “**Mapping the Hallmarks of Lung Adenocarcinoma with Massively Parallel Sequencing**..” Cell 150, no. 6 (September 14, 2012): 1107–20. doi:10.1016/j.cell.2012.08.029.



https://www.genome.gov/sequencingcosts/



https://www.genome.gov/sequencingcosts/



- Graph reflects projects involving the '**re-sequencing**' of the human genome, where an available reference human genome sequence is available to serve as a backbone for downstream data analyses.

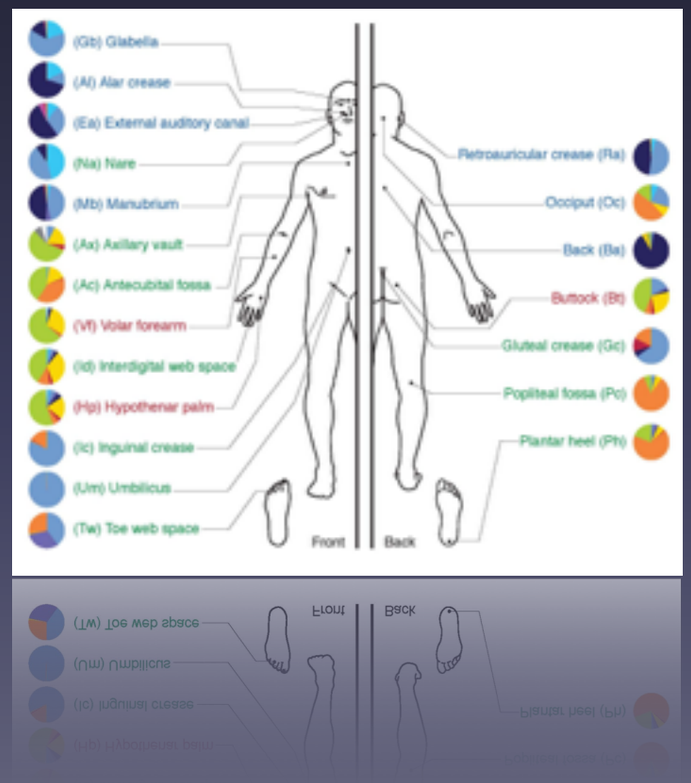
- The required '**sequence coverage**' would be greater for sequencing genomes for which no reference genome sequence is available.

Big Datasets

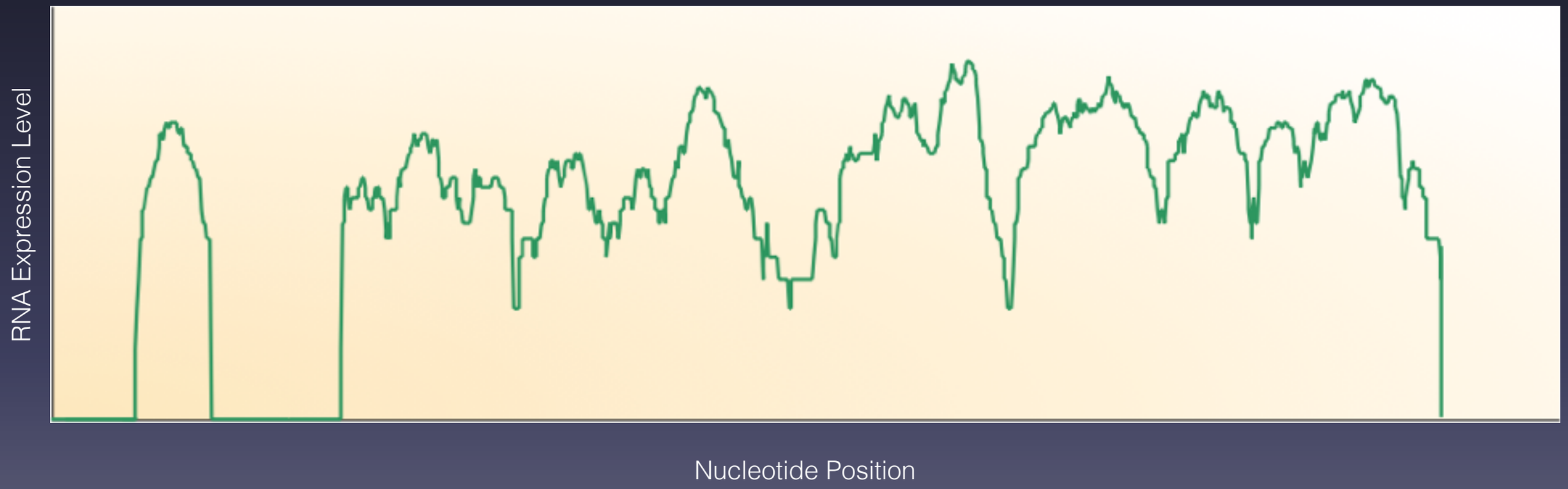
- **The Cancer Genome Atlas (TCGA)**
  - An effort high-throughput genome analysis techniques to improve our ability to diagnose, treat, and prevent cancer through a better understanding of the genetic basis of this disease.

- **Human Microbiome Project (HMP)**
  - Investigate how changes in the human microbiome are associated with human health or disease.

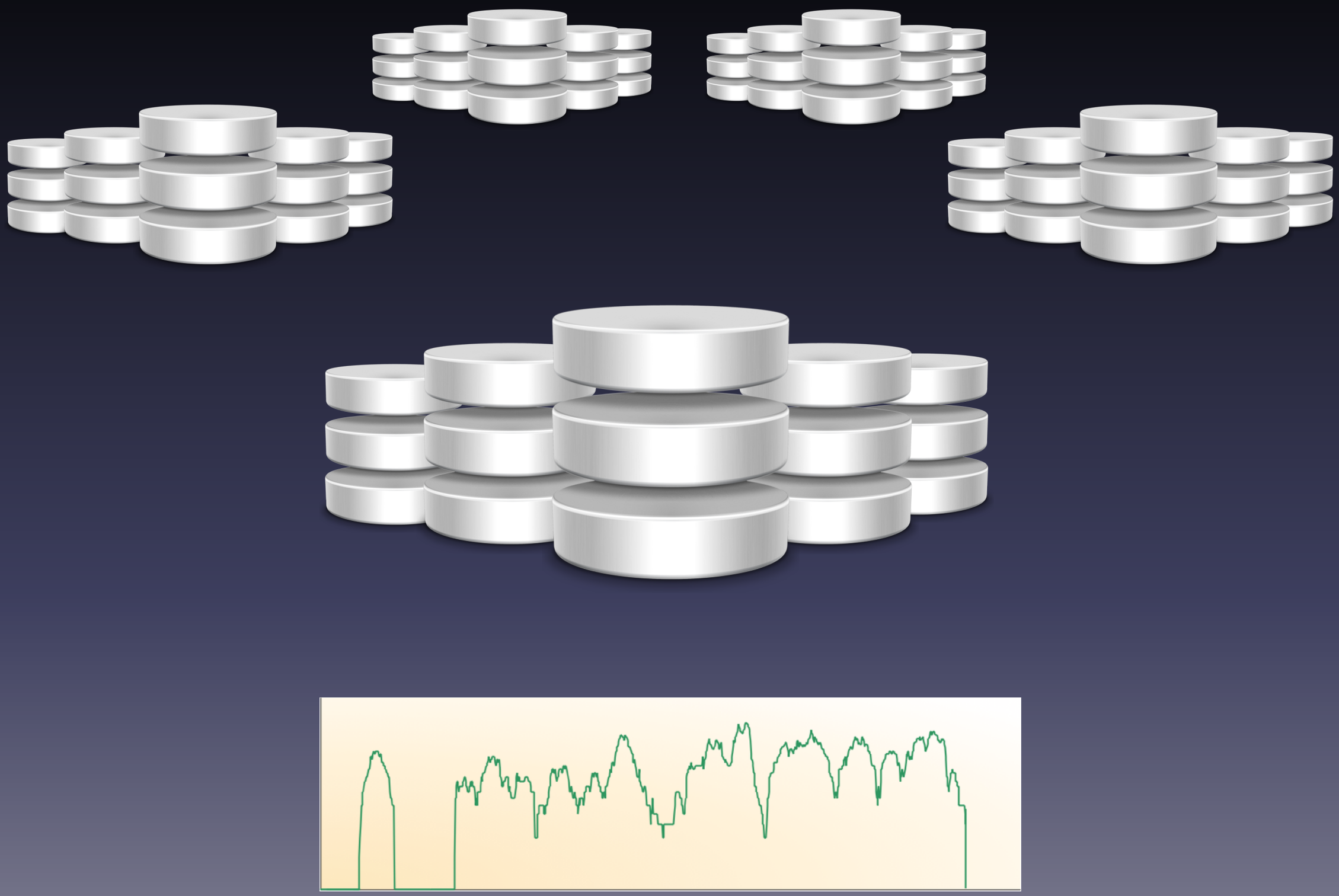
Human Microbiome Project

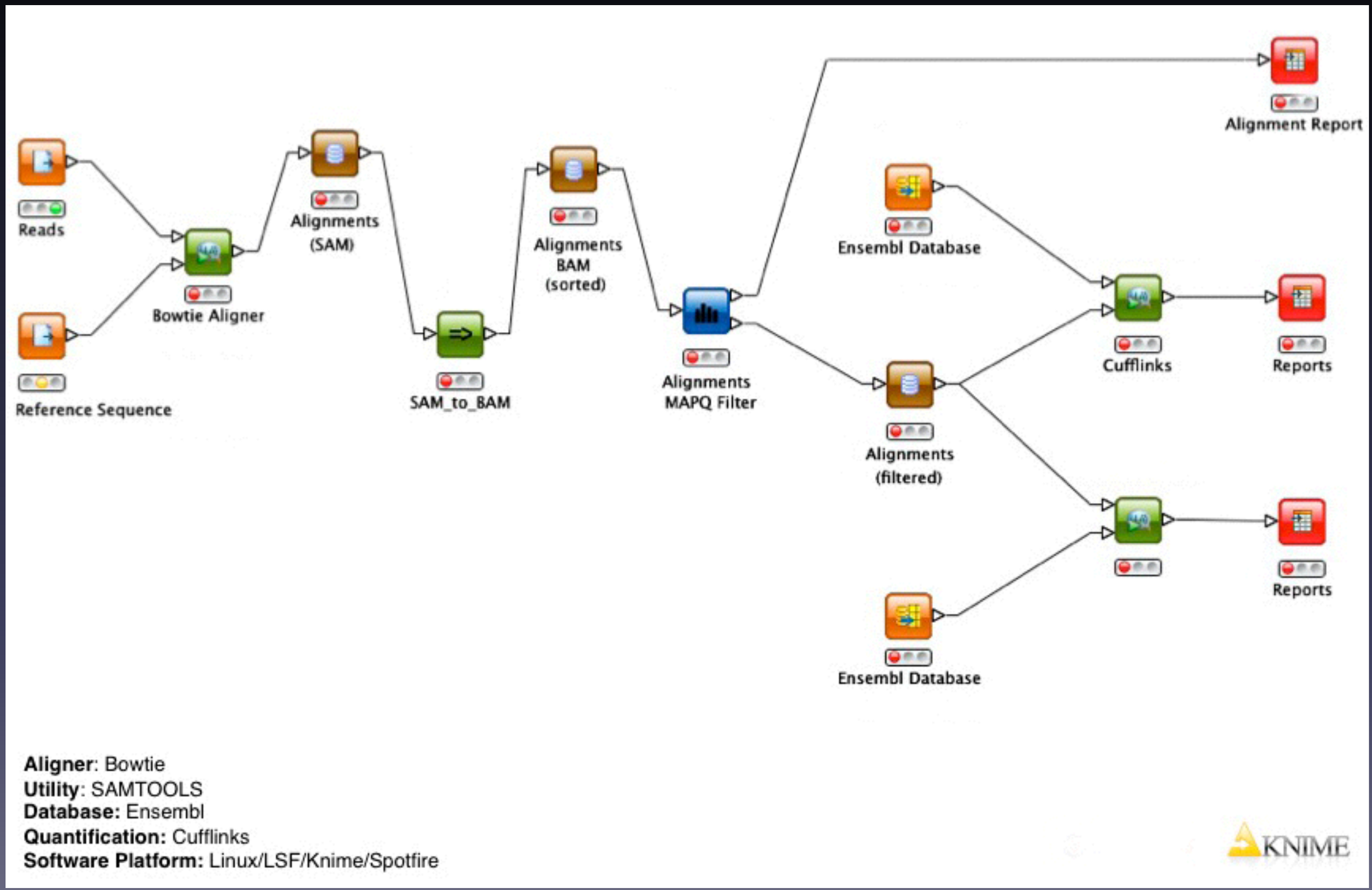


Base-Resolution Expression Profile



Wang, Zhong, Mark Gerstein, and Michael Snyder. “RNA-Seq: a Revolutionary Tool for Transcriptomics..” Nature Reviews Genetics 10, no. 1 (January 2009): 57–63. doi:10.1038/nrg2484.





Aligner: Bowtie
Utility: SAMTOOLS
Database: Ensembl
Quantification: Cufflinks
Software Platform: Linux/LSF/Knime/Spotfire

Data Deluge

- Sequencing technologies generate data at a much faster rate, and cheaper costs, than it is to store, manage, and analyze it.
- How do we turn data into knowledge?

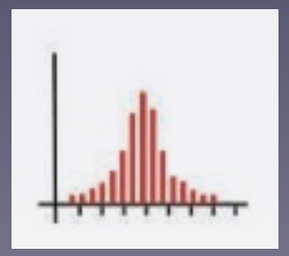
"Computational tools are quickly becoming inadequate for analyzing the amount of genomic data that can now be generated, and this mismatch will worsen."

Eric Green
National Human Genome Research Institute
NHGRI

Visualization

Visualizing Biological Data

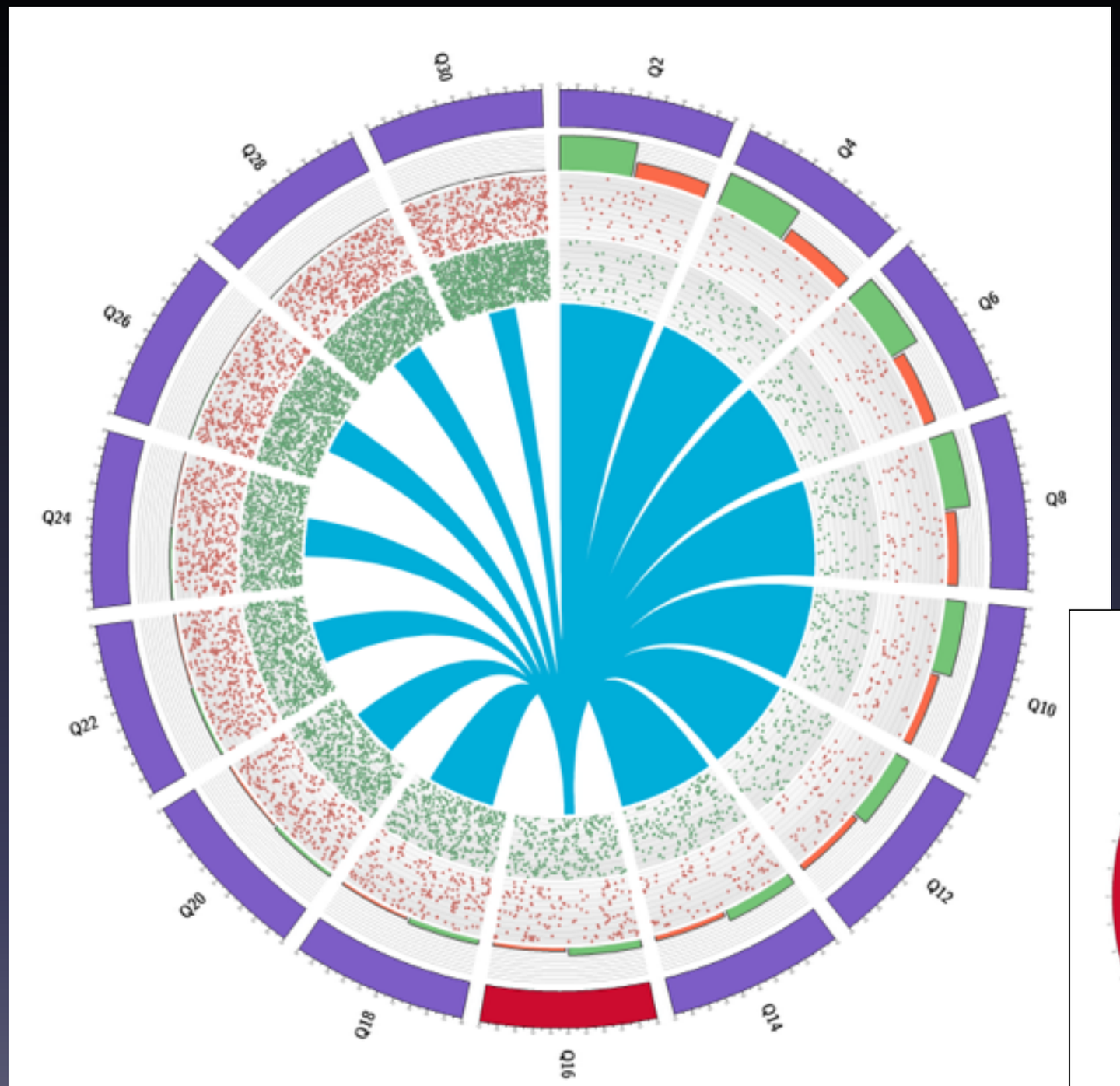
- Visualization is very important in genomics as data grows rapidly in volume and complexity.
- Keep things simple, 7± 2
- What is the *job-to-be-done* for this visualization?



What's the Job to be done?

- What are we trying to show?
- What are the important features and patterns?
- How can I effectively minimize their visual impact?
- What data resolution is required?
- In what media will the figure be shown?

http://circos.ca/intro/data_visualization/

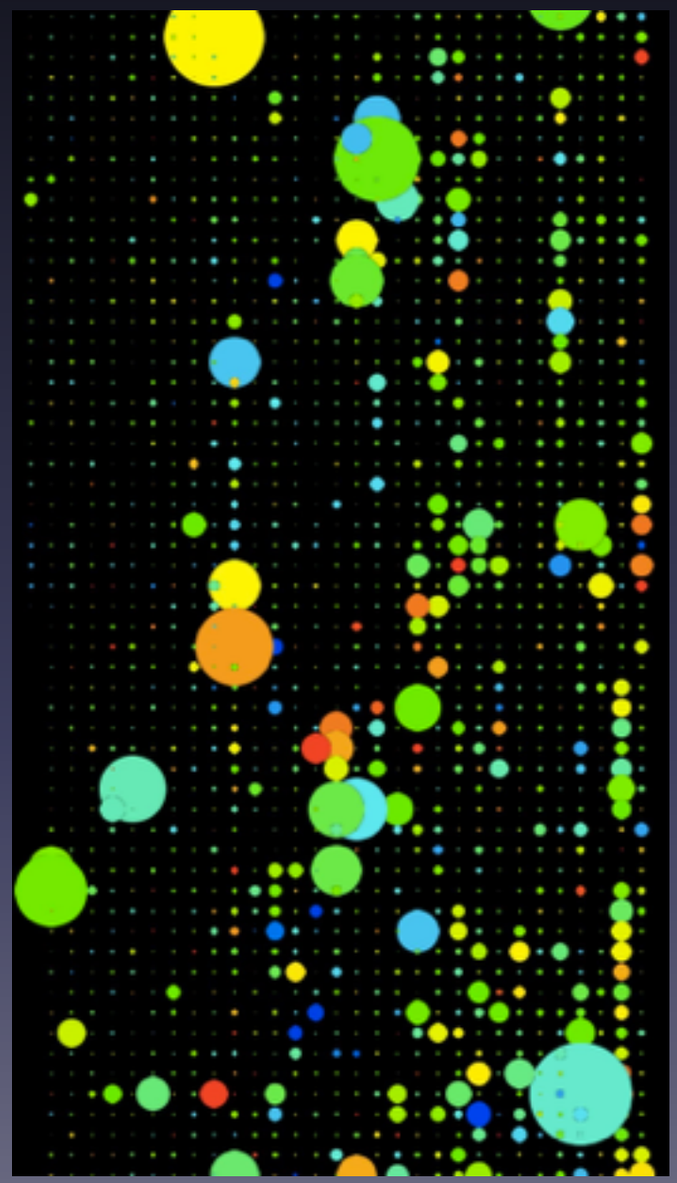

Q30
Q2
Q28
Q4
Q26
Q6
Q24
Q8
Q22
Q10
Q20
Q12
Q18
Q14
Q16

Visualization Tools

- Circos
  - http://circos.ca/
- R & ggplot2
  - http://ggplot2.org/
- Python MatLib
  - http://matplotlib.org/
- Knime
  - https://www.knime.org/

Resources

- Jer Thorp
  - http://blog.blprnt.com/
- Flowing Data
  - http://flowingdata.com/
- Tabletop Whale
  - http://tabletopwhale.com/
- Vizual Statistix
  - http://vizual-statistix.tumblr.com/



credit: Jer Thorp

Thank You!