# Questions

**Use the next 15+ minutes to answer and send us your responses: guertin@uchc.edu & miura@uchc.edu**

1. Have you ever made high throughput sequencing (HTS) libraries?
2. Does your thesis project involve HTS experiments or analysis? —if so, please describe
3. Have you previously analyzed HTS data?
   - If so, did you use the command line or web-based tools?
4. How would you rate your abilities in the terminal/command line?
   - Rate 1 to 5: 1 = *The Terminal…the 2004 movie starring "America's dad" Tom Hanks?;* 5 = *my stack overflow name is shellHacker1976*
5. How would you rate your abilities in R?
   - Rate 1 to 5: 1 = *My last lecture on "R" was in kindergarten;* 5 = *statistics, parsing, figures…all the things.*
6. Are you familiar with any programming languages? —if so, please list them
7. What type of computer and operating system will you be using for this course?

Bookmark this page:
https://github.com/guertinlab/meds5420

# MEDS 5420:
# Molecular Genomics Practicum

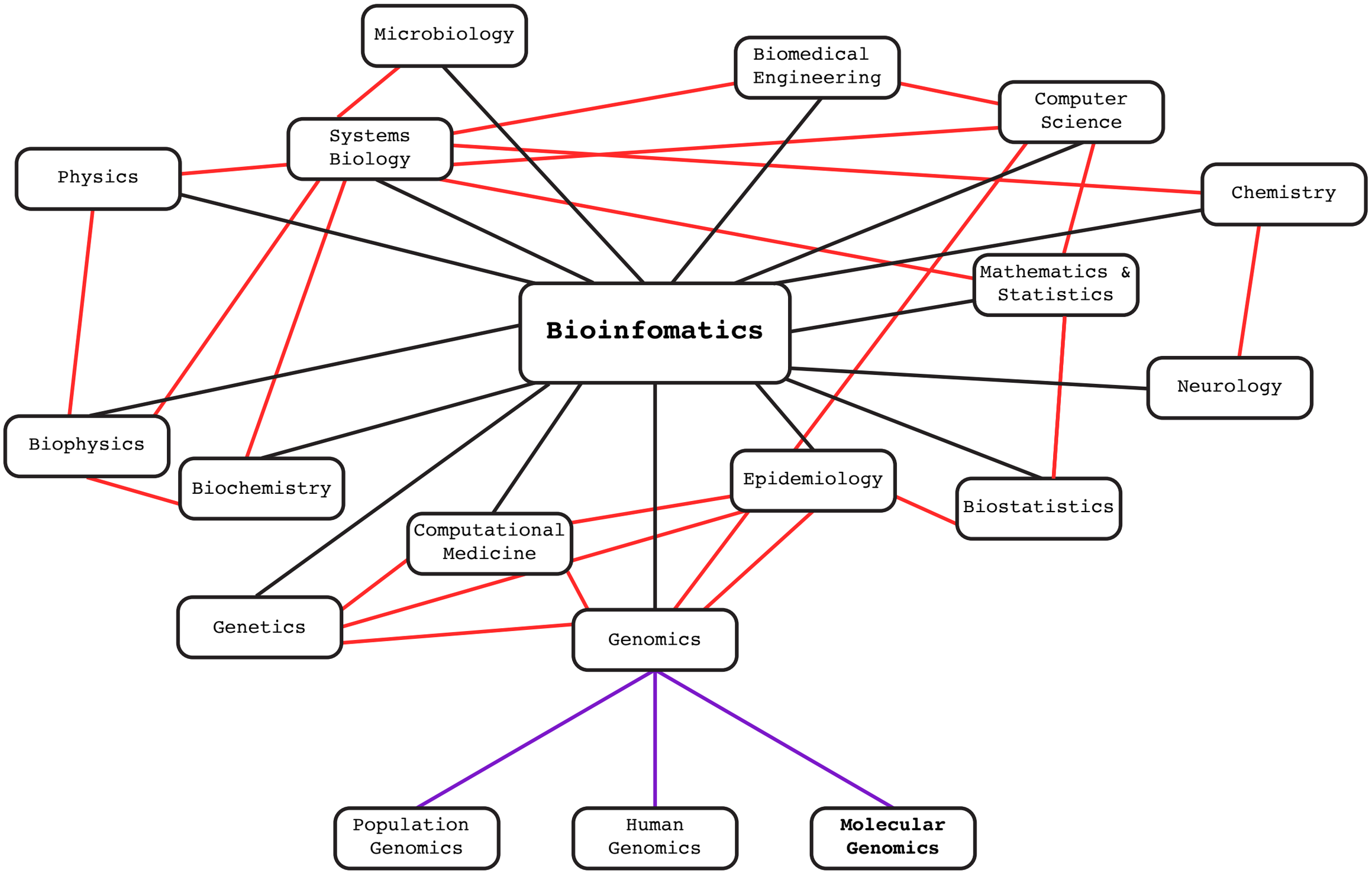
UCONN HEALTH

## Spring 2025

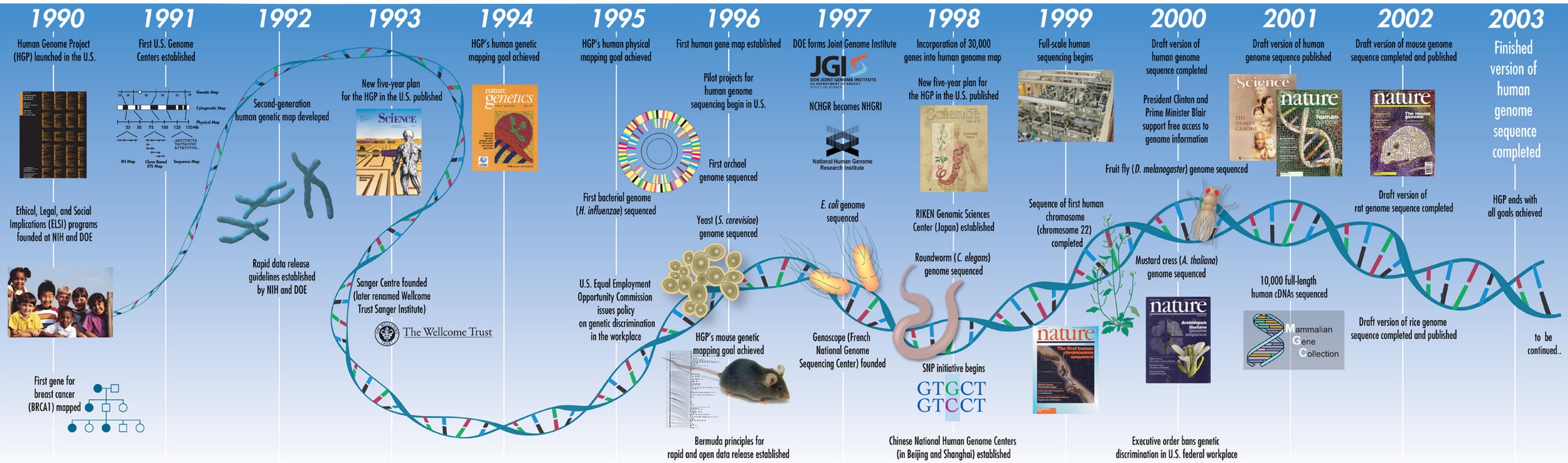
**Mike Guertin & Pedro Miura**

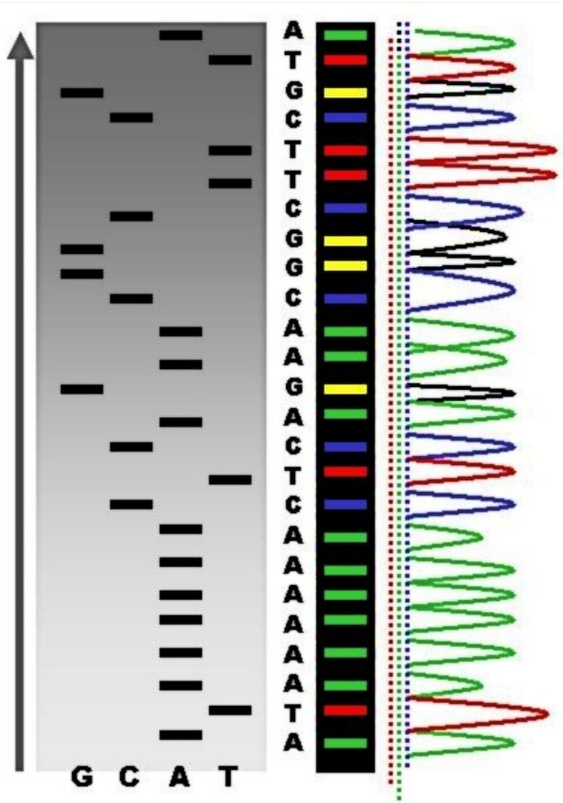
This course is adapted from UConn Professor Leighton Core's course MCB 5430

# Why *molecular genomics* and not **bioinformatics**?



# Human Genome Project

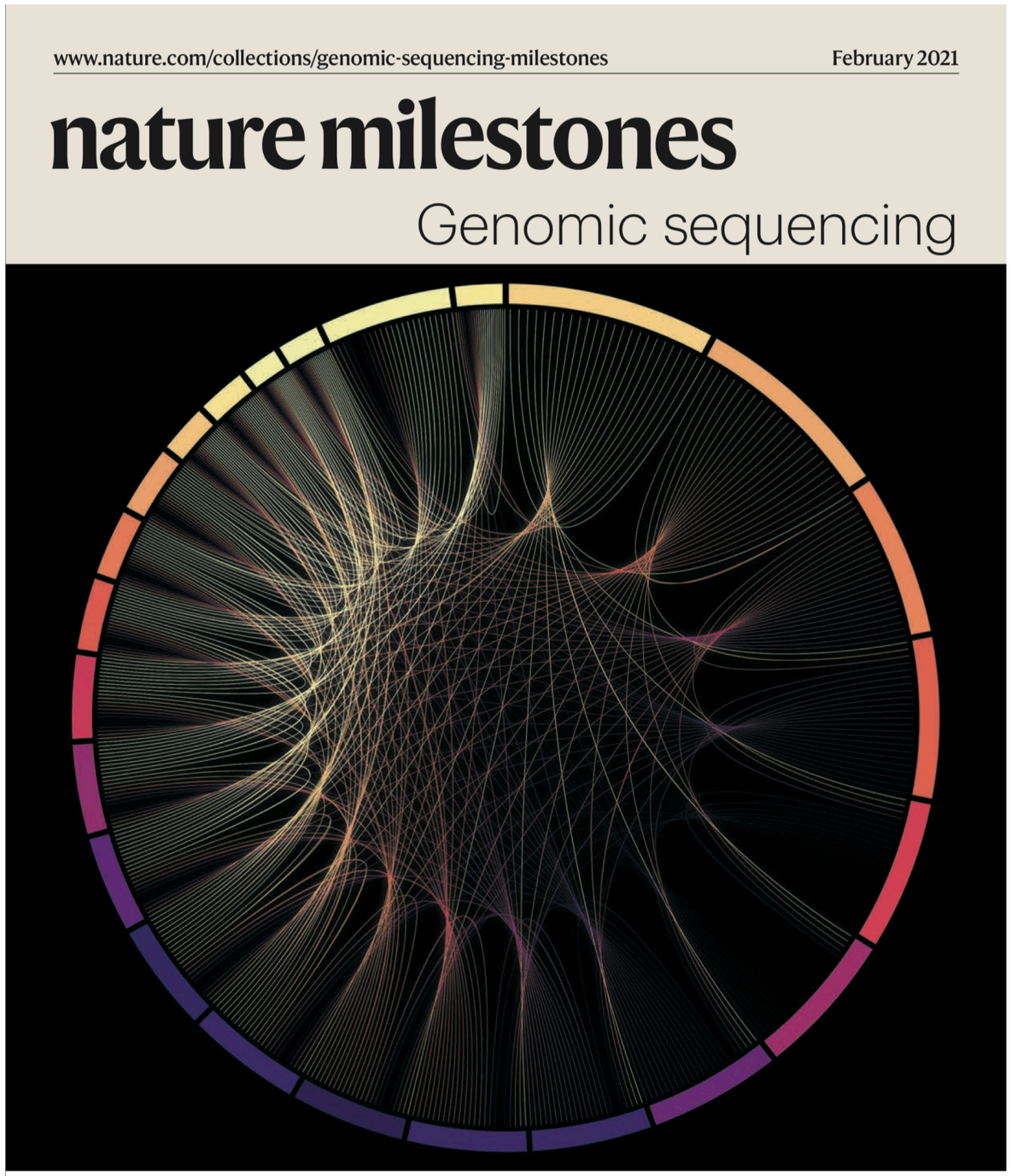



Sanger sequencing

Cost: 2.7 billion

# Contemporary milestones in genomics




https://www.nature.com/immersive/d42859-020-00099-0/pdf/d42859-020-00099-0.pdf

Genomics of human variation

Population Genomics

Epigenomics

Functional Genomics

# Genomics?!

Microbiome Genomics

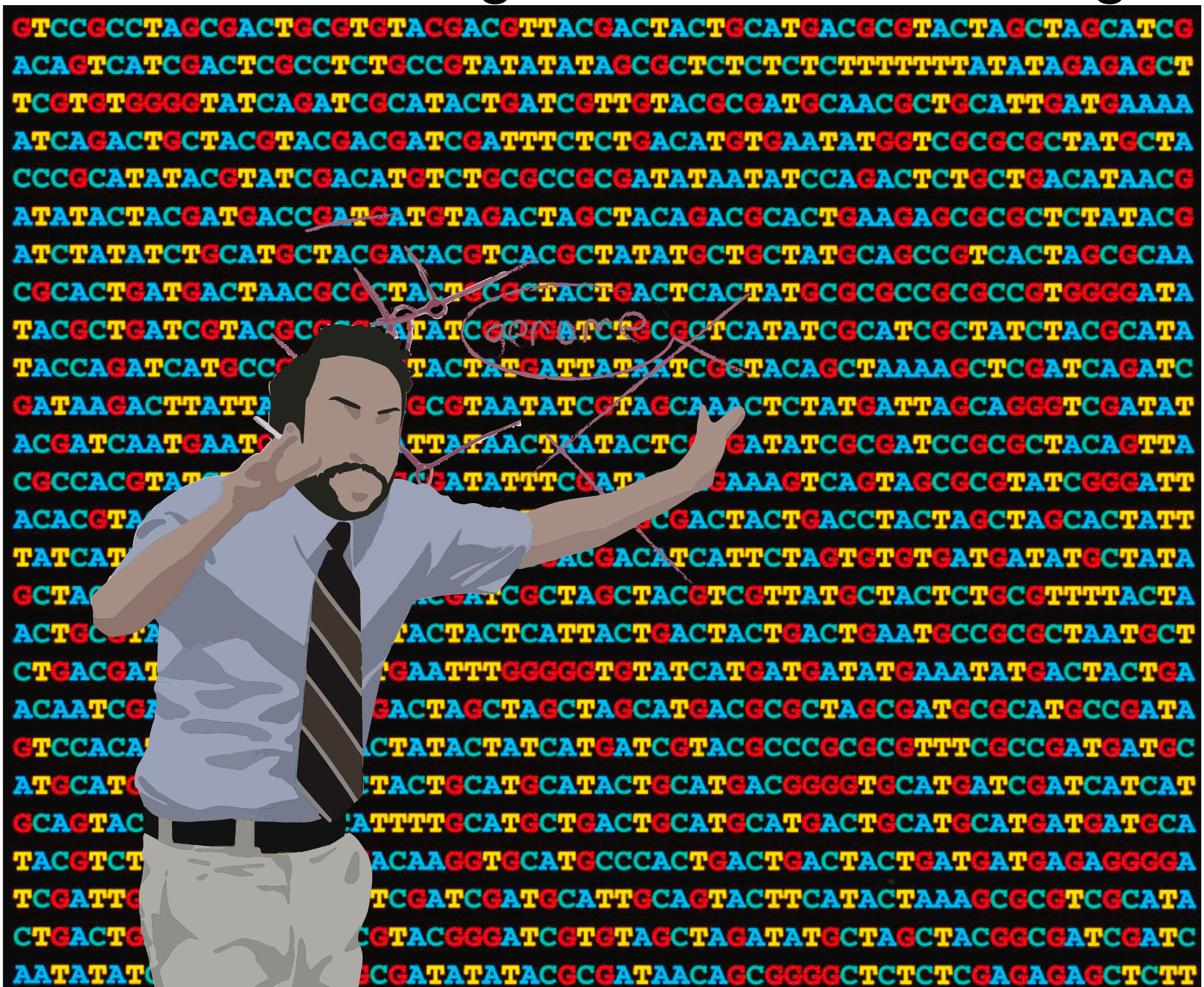
Structural Genomics

Medical Genomics

Metagenomics

**Molecular Genomics:  coupling classic molecular biology techniques to HTS for nucleic acid quantification**

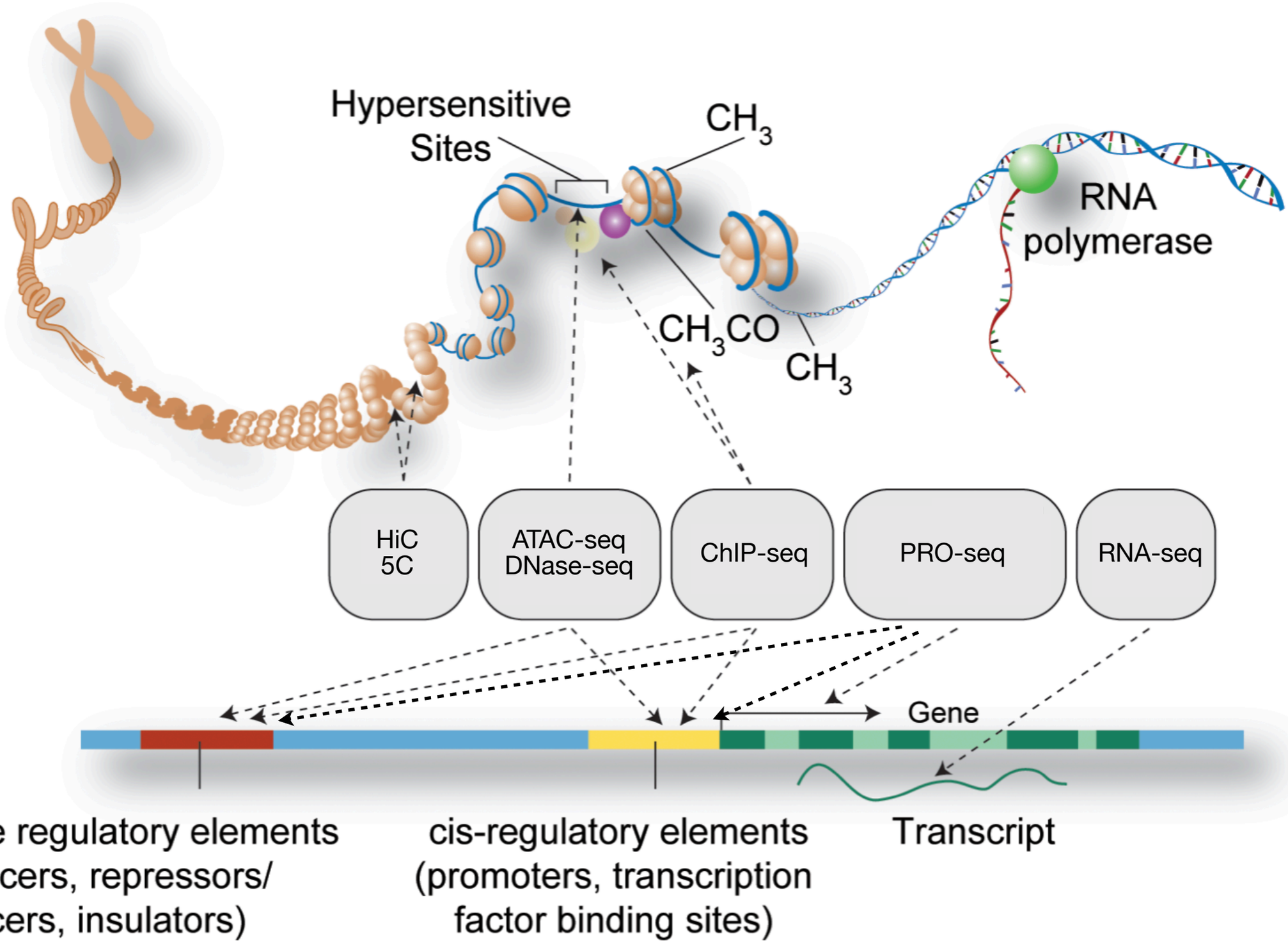
# How do we begin to understand the genome?



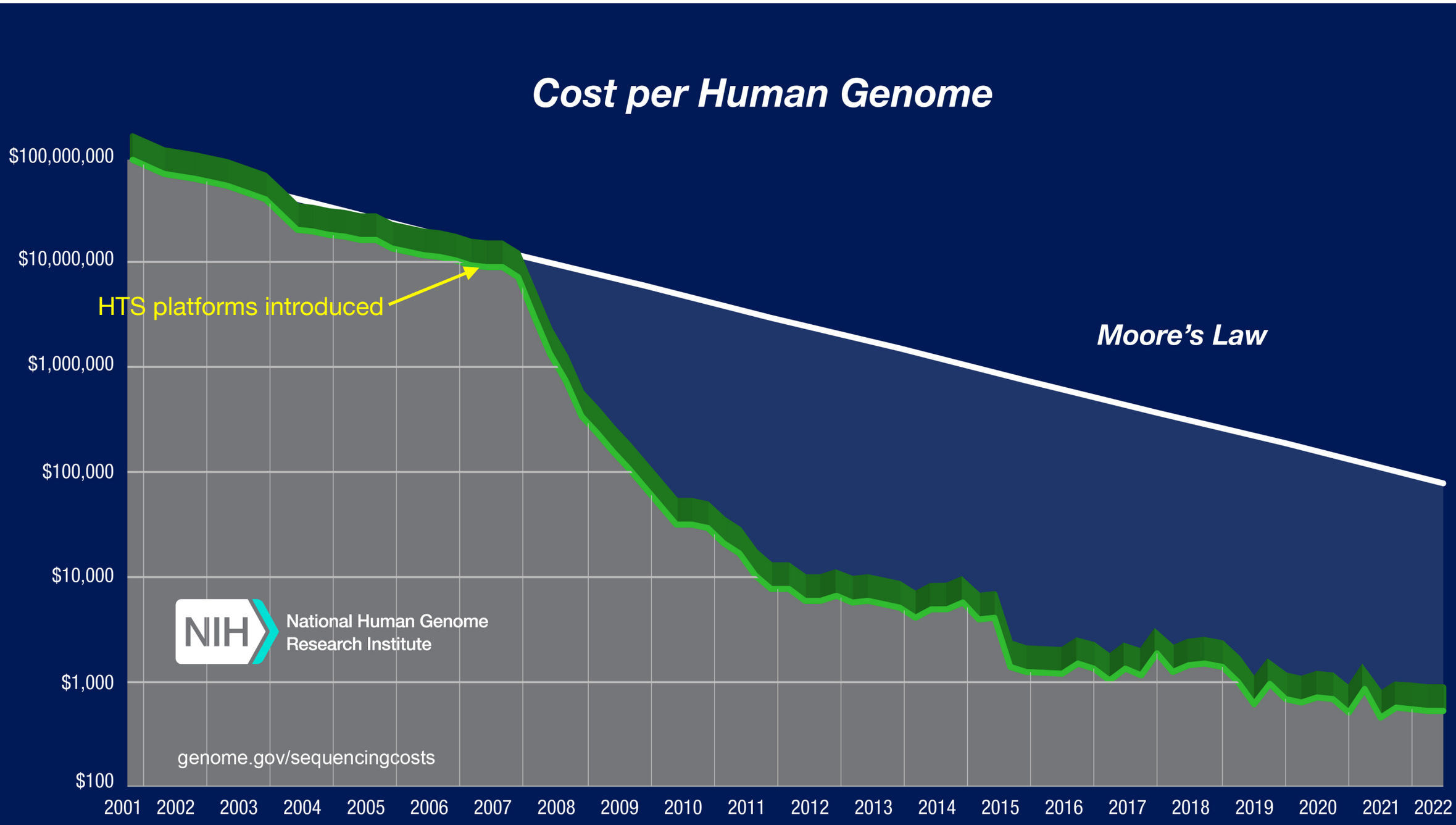
# Questions that can begin to be addressed with Molecular Genomics

- How much of the genome is functional?
- Where are the functional elements?
- How are elements organized 3 dimensionally?
- What constitutes the molecular makeup of regulatory regions?
- How do regulatory regions change throughout development, upon environmental perturbation, or in the presence of mutations?
- What is the identity of proteins bind to regulatory regions to regulate gene expression?

# Molecular genomics assays



# High throughput sequencing costs drove the genomics revolution



# High throughput sequencing technologies

Roche 454

Ion Torrent

ABI Solid

## Long Read (> 1kb)

Oxford Nanopore

Pacific Biosciences

## Illumina

GAIIx

HiSeq 2500

iSeq 100

MiniSeq

NextSeq 550

NextSeq 1000 & 2000

NovaSeq 6000

https://genohub.com/ngs-instrument-guide/

# High throughput sequencing technology

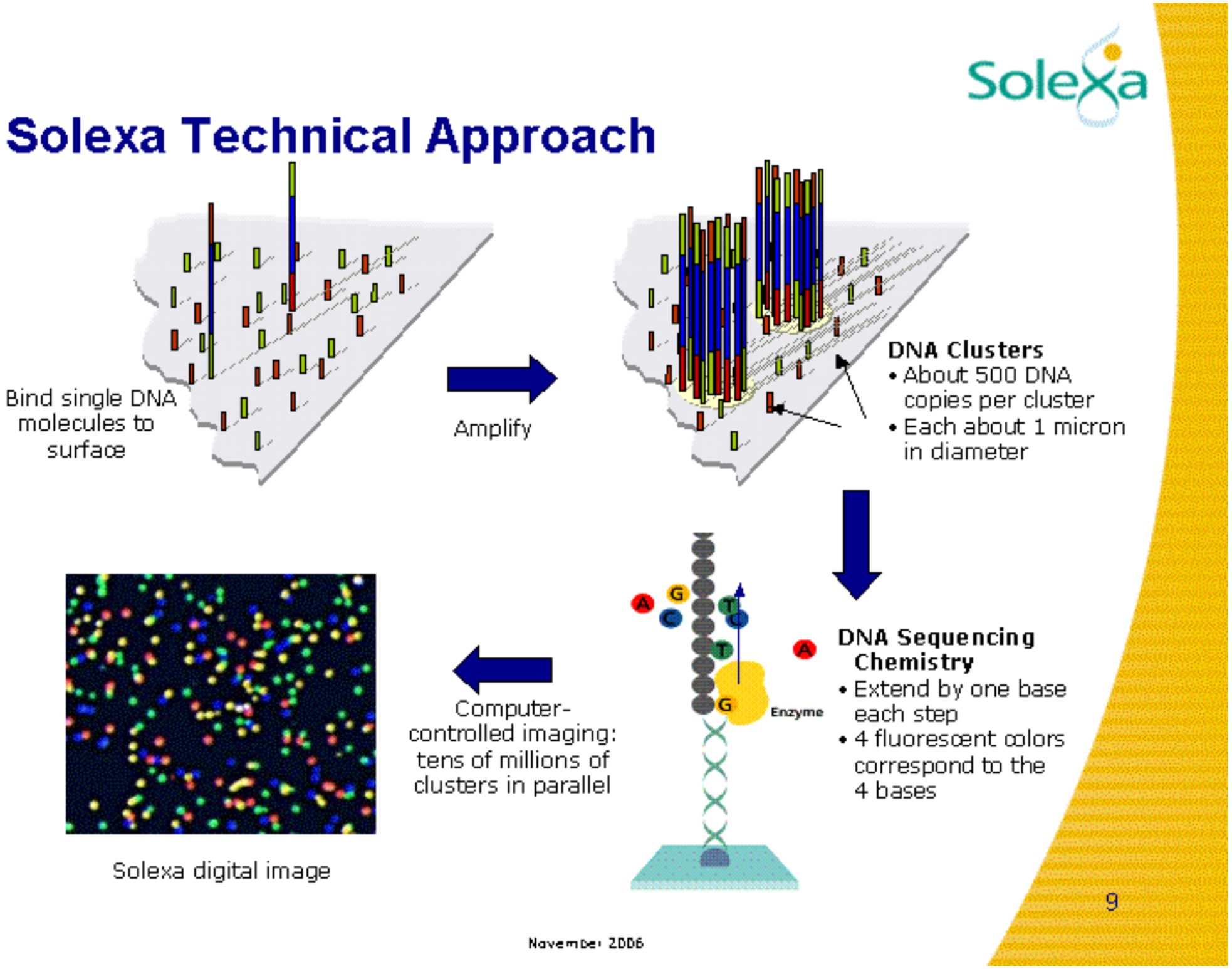
| Platform | Instrument | Reads/unit | Read Length (bp) | Read Type | Error Type |
|---|---|---|---|---|---|
| Illumina | NovaSeq 6000 S4 | 10,000,000,000 | 300 | SR & PE | substitution |
| Illumina | NextSeq 550 High-Output | 1,200,000,000 | 300 | SR & PE | substitution |
| Illumina | HiSeq High-Output v4 | 250,000,000 | 250 | SR & PE | substitution |
| Illumina | GAIIx | 42,075,000 | 300 | SR & PE | substitution |
| Illumina | MiSeq v3 | 25,000,000 | 600 | SR & PE | substitution |
| Illumina | MiniSeq High-Output | 25,000,000 | 300 | SR & PE | substitution |
| Ion | Proton I | 60,000,000 | 200 | SR | indel |
| Ion | PGM 314 | 400,000 | 400 | SR | indel |
| PacBio | PacBio Sequel | 370,000 | 20,000 | NA | indel |
| PacBio | PacBio RS II (P6) | 55,000 | 15,000 | NA | indel |
| Roche 454 | GS FLX+ / FLX | 700,000 | 700 | NA | indel |
| SOLiD | 5500xl W | 266,666,667 | 100 | SR & PE | A/T Bias |
| SOLiD | 5500xl | 81,500,000 | 100 | SR & PE | A/T Bias |
| Oxford Nanopore | PromethION 48 | depends on size (300 Gb total) | length of molecule up to 4,000,000 | NA | sub/indel |

# Genomics at UConn


NovaSeq 6000 can sequence the equivalent of 48 human genomes per run at 30x coverage!

The Center for Genome Innovation with the Institute for Systems Genomics acquired a NovaSeq in 2020

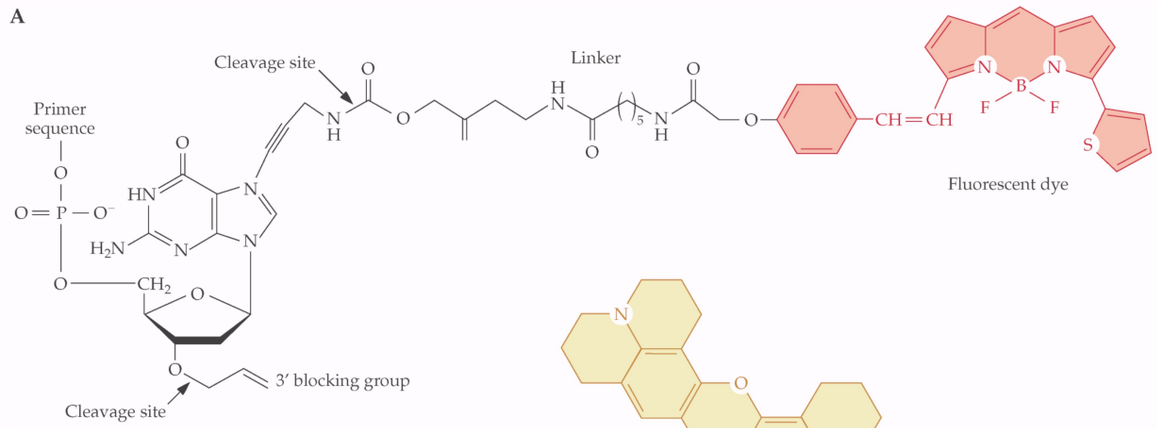
# Illumina (formerly Solexa) Sequencing Technology: Clonal PCR colonies and Reversible Terminators

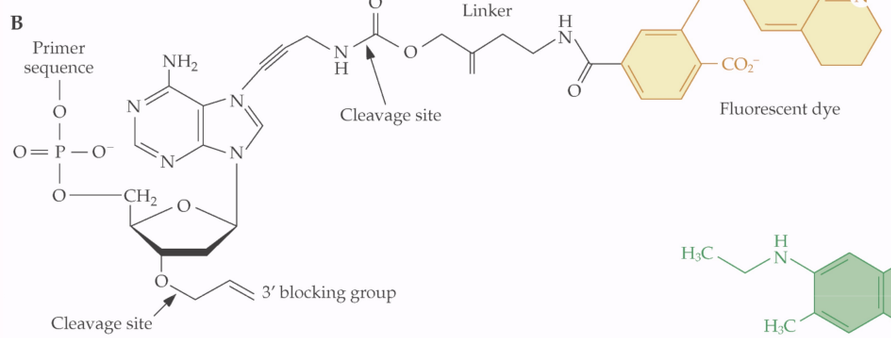


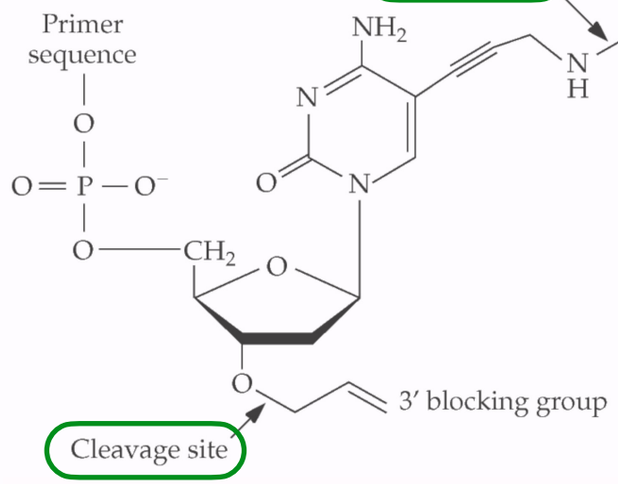

Balance of economy, read length, throughput, versatility, and reliability

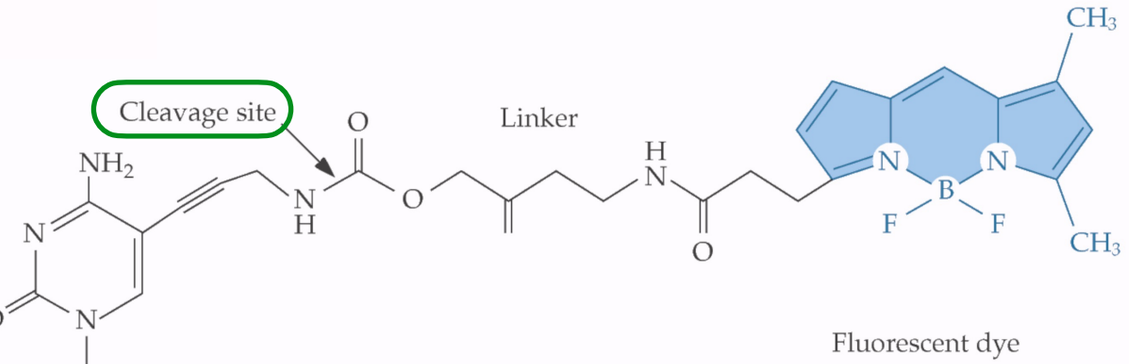
2006 SEC filing for Solexa Technology

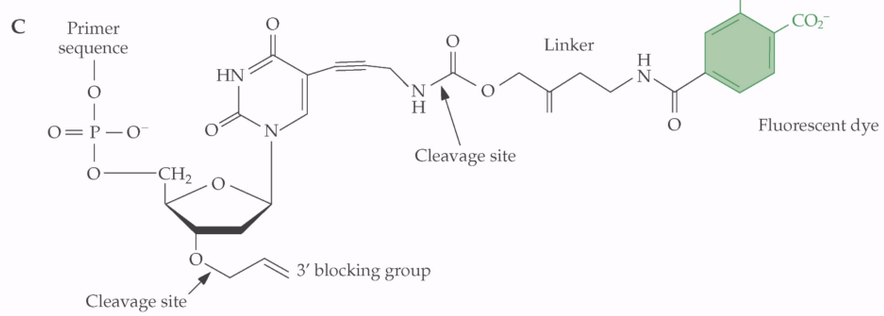
# Illumina Sequencing Technology: Dye and Reversible Terminators

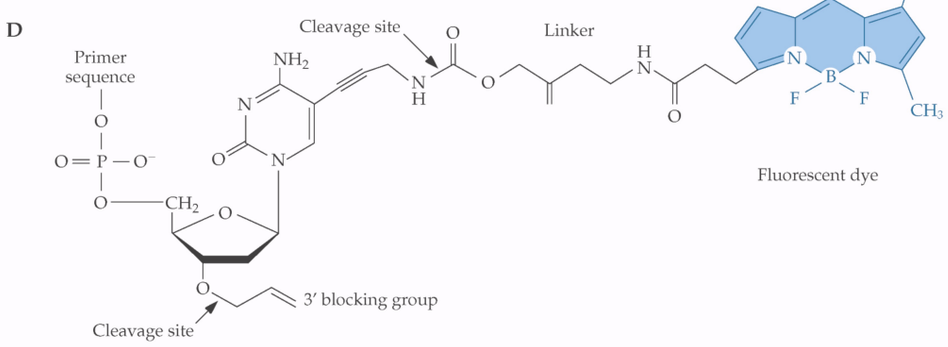


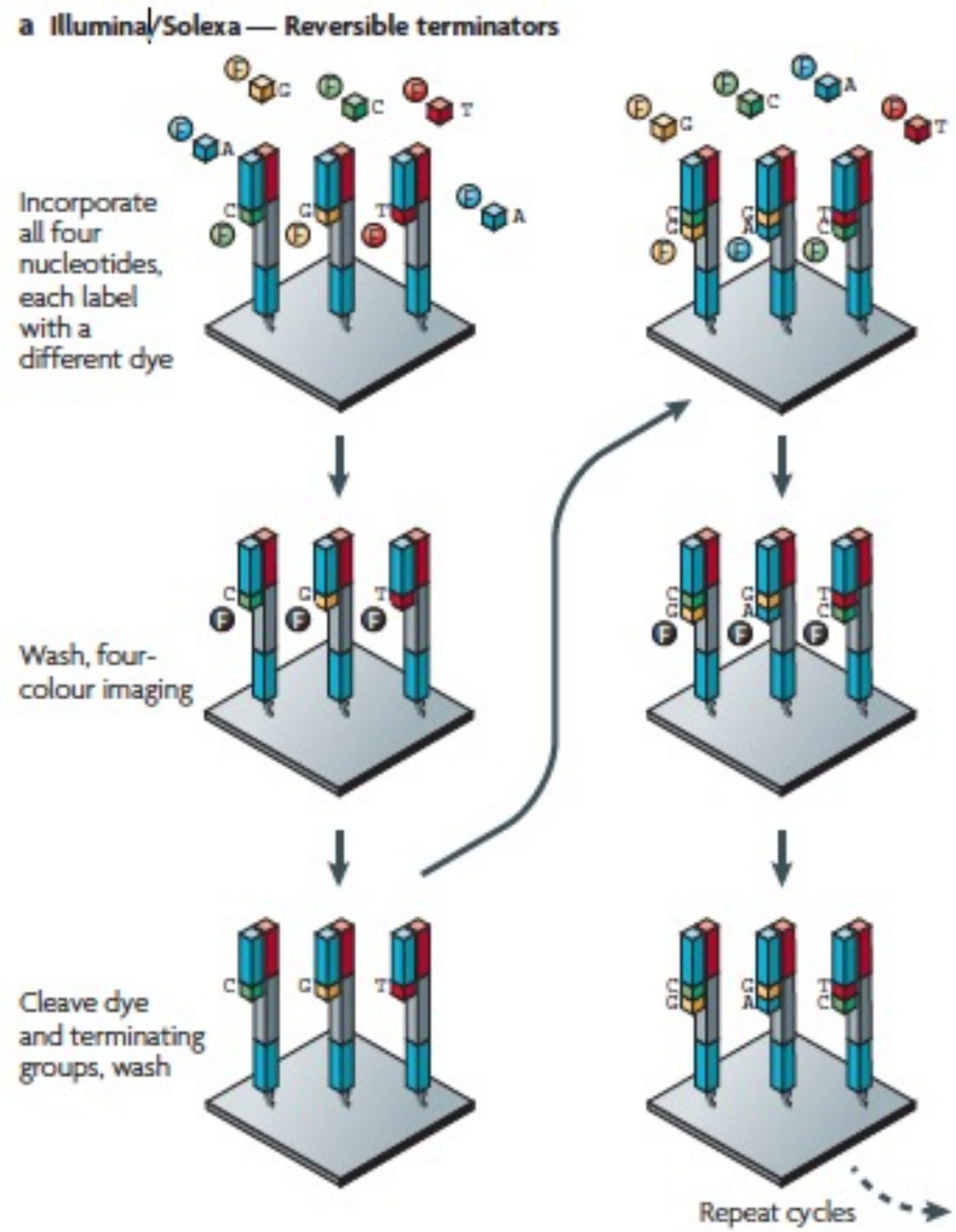



# Illumina Sequencing Technology: Dye and Reversible Terminators



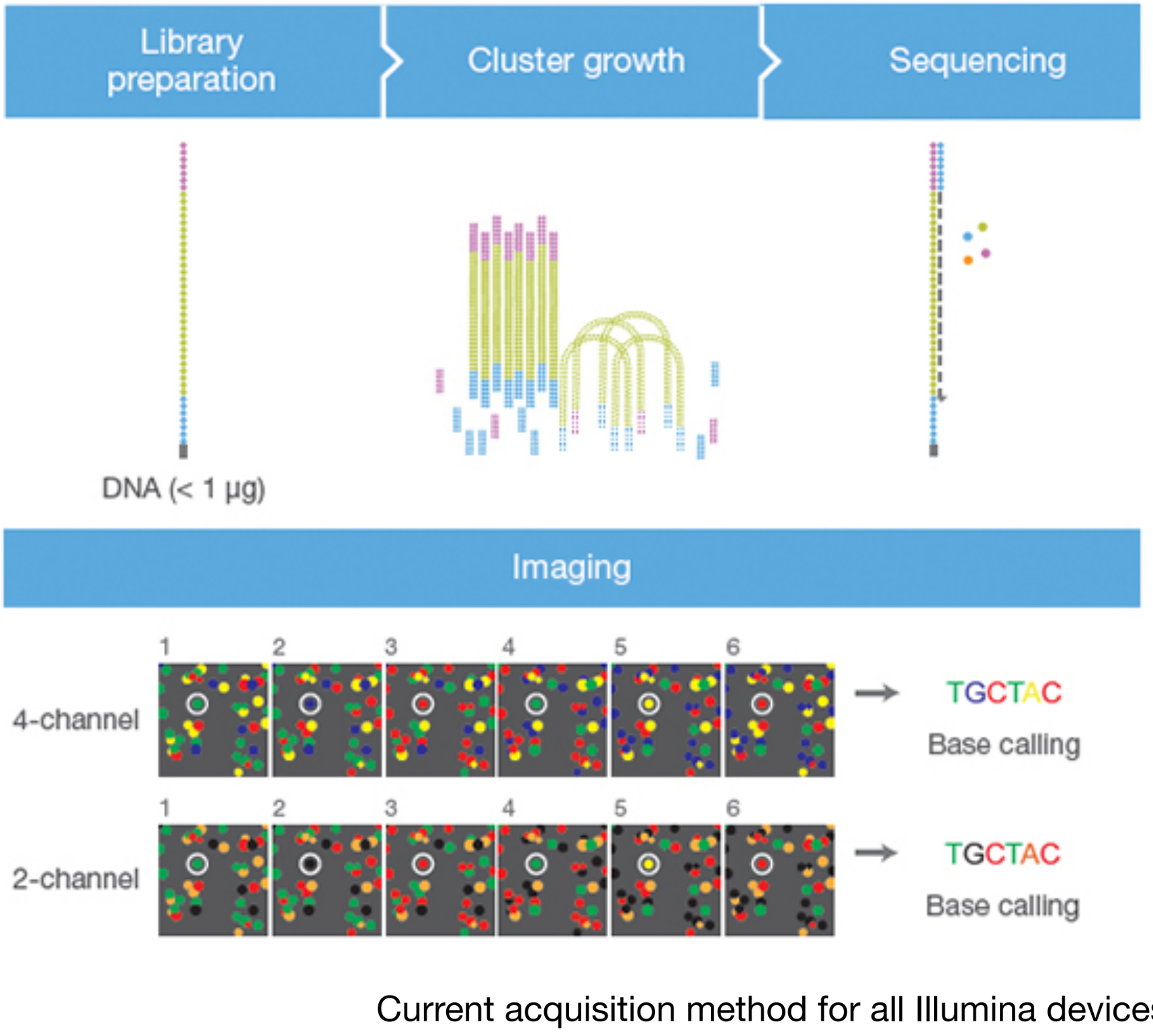

Add bases

Detect fluorescent base incorporated

Cleave fluorescent dye and terminating group

A clear explanation of the entire process from Illumina: https://www.youtube.com/watch?v=womKfikWlxM

# Illumina sequencing by synthesis (SBS) updates: 2 color imaging




**Benefits**:

Fewer images (2 vs 4):
= less data acquisition and processing time

= faster sequencing.

User experience unaffected

T = 🟢

C = 🔴

A = 🟢 + 🔴 = 🟡

G = X

Current acquisition method for all Illumina devices

# Illumina SBS updates: patterned flow cell

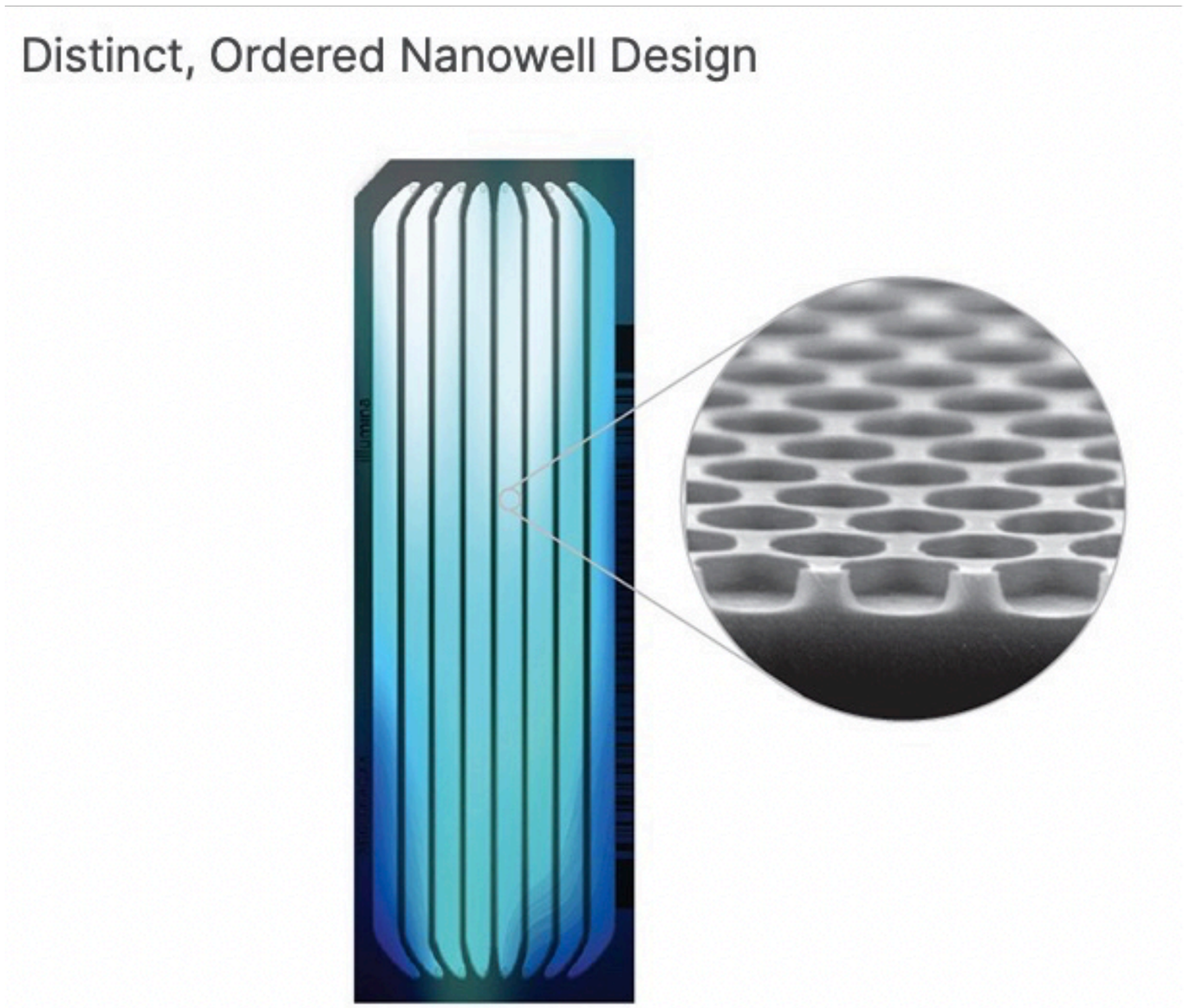



**Figure 1. Advanced Patterned Flow Cell Design Enables Maximum Throughput.**
Patterned flow cells contain billions of nanowells at fixed locations, providing even cluster spacing and uniform density.

**Benefits**:

- Location of clusters known
- Less cluster overlap
- Exclusion Amplification (ExAmp) allows multiple clusters from a single molecule

**Pitfalls**:

- Nanowells favor clustering small adapter/adapter products
- ExAmp creates PCR duplicates—good for genome coverage; bad for quantification of molecular genomics experiments

Currently used on NextSeq2000, NovaSeq

(page 7 of the patent provides an explanation of the technology)
https://patentimages.storage.googleapis.com/f5/8f/f7/a0c052678df60e/WO2013188582A1.pdf

# Videos of HTS technologies

Roche 454: https://www.youtube.com/watch?v=rsJoG-AuINE

Ion Torrent: https://www.youtube.com/watch?v=zBPKj0mMcDg

Pac bio: https://www.youtube.com/watch?v=v8p4ph2MAvI

Illumina: https://www.youtube.com/watch?v=HMyCqWhwB8E

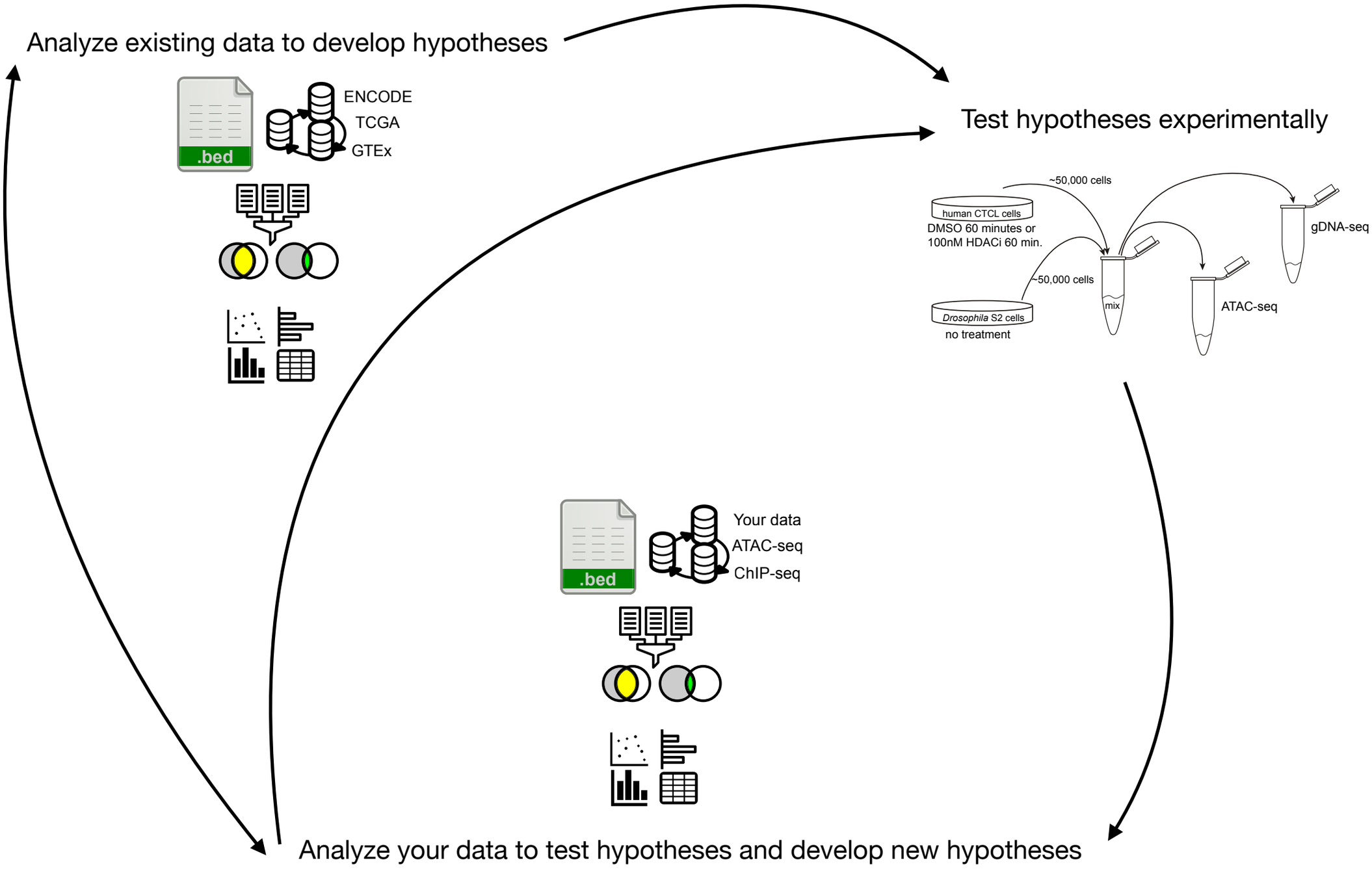
ABI solid: https://www.youtube.com/watch?v=nIvyF8bFDwM

Nanopore: https://www.youtube.com/watch?v=3UHw22hBpAk

# Challenges that arise when working with big datasets

- Computational resources
  - Data storage
  - Processing power
    - RAM
    - CPUs
- Computational competency
  - Adept in a command line environment
  - Knowledge about available utilities
  - Programming languages
  - Pipeline development

# A need for versatile scientists




Scientists need to be able to move between the bench and bioinformatics

# Terminology

- Script
  - Executable document or program listing computer interpretable commands to be executed in sequence.
- Pipeline
  - Often a series of independent scripts
    - Output from one script becomes input for next until desired result is achieved
    - Once defined requires limited user effort
    - Most processes that are routine enough to be automated in a pipeline are limited in the biological insights they can provide. Exploratory analyses are not usually pipelined.
- Workflow
  - A series of steps to be followed in sequence with varying levels of effort
    - May involve one or more pipelines
    - Can encompass entire project starting from experiments at the bench and ending with detailed analyses

# General analysis pipeline for genomics

**Input**

Raw data: FASTQ

**Data preprocessing**

**Output**

QC / summary metrics

**Processing/ mapping**

**Output**

QC / summary metrics

**Postprocessing**

**Output**

QC / summary metrics

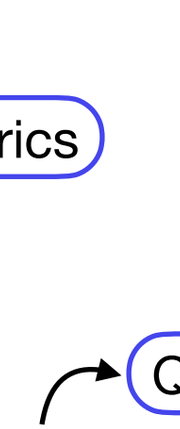
**Scripts**

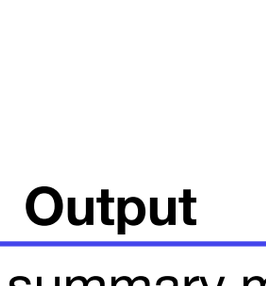
**Results**

**Analyses**







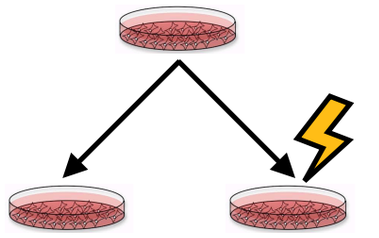


Go to page 16

# Workflow can encompass projects and analysis pipelines

## Typical genomics project workflow

**Experiments**




**Assay**
ChIP-seq
RNA-seq
*-seq

**Library Prep**
Enzymatic treatments
Ligation
Amplification

**Data Acquisition**
Sequencing

### Pipeline: processing and analysis

Input: Raw data: FASTQ

Data preprocessing → Output: QC / summary metrics

Processing/ mapping → Output: QC / summary metrics

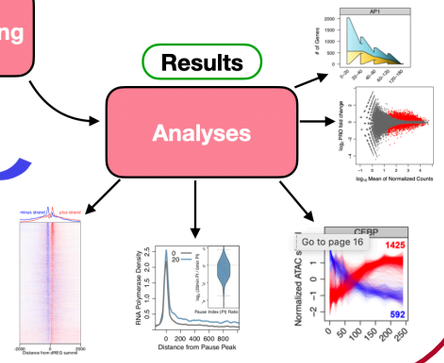
Postprocessing → Output: QC / summary metrics

Scripts

Processing can usually be automated, but analyses often cannot.

**Results**

Analyses




**Validation**
Computational
Experimental

**Retest**
Computational
Experimental

**Publish**

Kinetic networks identify key regulatory nodes and transcription factor functions in early adipogenesis

Arun B. Dutta[1], Bao Nguyen[1], Warren D. Anderson[1], Ninad M. Walavalkar[2], Fabiana M. Duarte[3], and Michael J. Guertin[1,4,5]

Sequence-specific transcription factors (TFs) bind DNA, mediate chromatin structure, regulate gene expression, and orchestrate transcription cascades. Activation and repression of TFs drive tightly controlled regulatory programs that lead to cellular processes such as differentiation. We measured chromatin accessibility and nascent transcription at seven time points over the first four hours of induced adipogenesis of 3T3-L1 mouse preadipocytes to construct dynamic gene regulatory networks. Regulatory networks describe successive waves of TF binding and dissociation followed by direct regulation of proximal genes. We identified 14 families of TFs that coordinate with and antagonize each other to regulate early adipogenesis. We developed a compartment model to quantify individual TF contributions to RNA polymerase initiation and pause release rates. Network analysis showed that the glucocorticoid receptor and AP1 drive immediate gene activation, including induction of *Zfp423*. *Zfp423* is a highly interconnected node within the network and its expression leads to repression of target genes. Although *Zfp423*'s role in adipogenesis has not been previously appreciated, both *Zfp423* knockout mice and Sorlien syndrome (*Zfp423*[+/-]) patients lack subcutaneous and brown adipose tissue. We found that kinetic networks integrating chromatin structure and nascent transcription dynamics identify key genes, TF functions, and coordinate interactions within regulatory cascades.

Adipogenesis | Chromatin | Transcription | RNA Polymerase Pausing | TWIST2

Correspondence: guertin@uchc.edu

Introduction

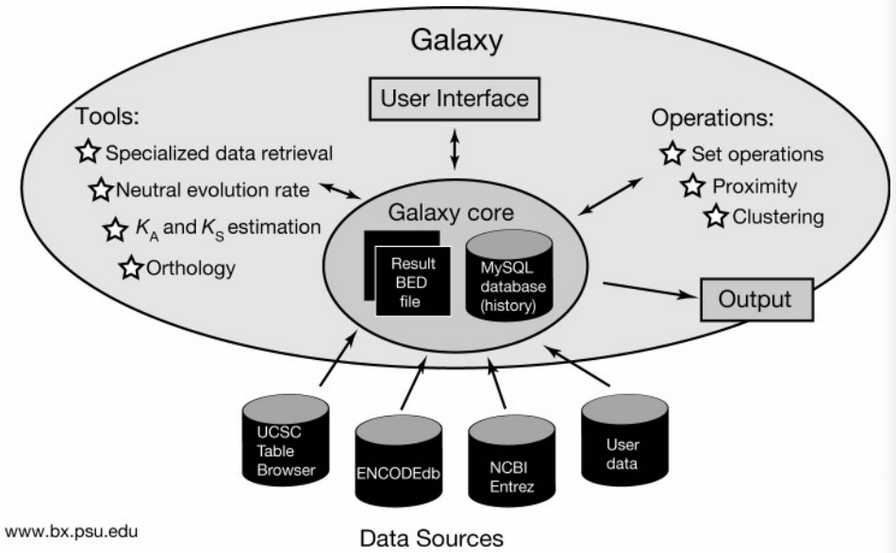
Differential transcription factor (TF) activity defines cell identity and drives cellular responses to environmental stimuli by enforcing gene regulatory programs (Takahashi and Yamanaka 2006). Sequence-specific TFs bind to conserved motifs (Pushne 1967) in regulatory elements (REs) within promoters and enhancers to regulate different mechanistic steps in transcription (Fuda et al. 2009). TFs recruit cofactors such as chromatin remodelers, acetyltransferases, methyltransferases, and general transcription machinery to REs. TFs are generally characterized as activators or repressors based upon their interaction partners, and recent studies more specifically describe TFs based upon their molecular function and which mechanistic steps they regulate (Danko et al. 2013; Duarte et al. 2016; Hah et al. 2011; Sathyan et al. 2019; Scholes et al. 2017). For example, pioneer transcription factors specialize in chromatin opening (Zaret and Carroll 2011). In addition to chromatin opening and RNA polymerase recruitment, many transcription steps are precisely regulated, such as RNA polymerase pausing, elongation, and termination. RNA polymerase pauses ~30-50 base pairs downstream of the transcription start site (TSS) (Rasmussen and Lis 1993; Rougvie and Lis 1988) and a vast majority of genes exhibit pausing (Core et al. 2008; Muse et al. 2007; Zeitlinger et al. 2007). Further modifications to the RNA polymerase complex triggers pause release and productive elongation (Marshall and Price 1995). Defining the steps regulated by TFs is necessary to understand how TFs coordinate with one another productively or antagonistically to regulate complex gene expression programs.

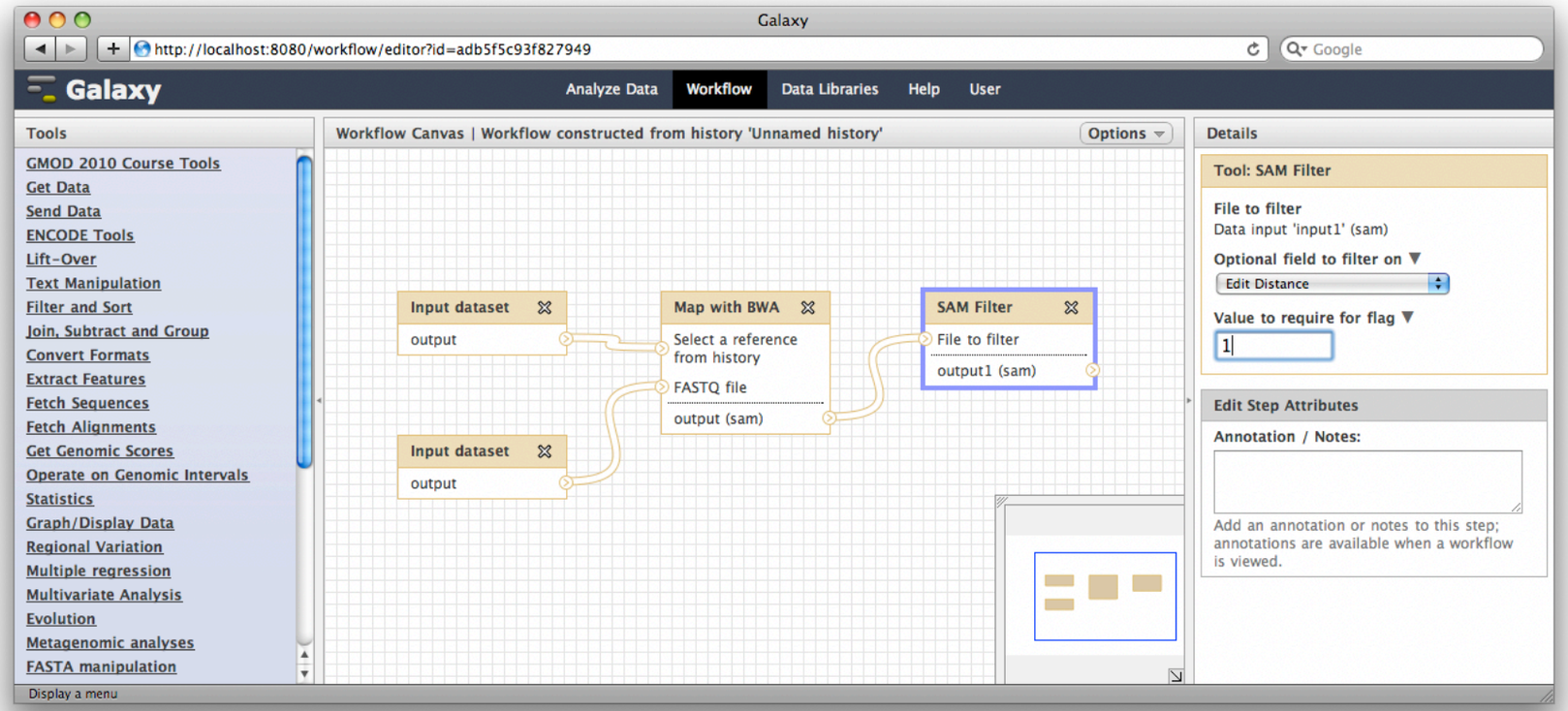
High-throughput sequencing has led to the development of hundreds of molecular genomics assays that query mechanistic events associated with transcription regulation. While each assay delivers a tremendous amount of information, each is limited in the biology that it measures. ChIP-seq directly quantifies chromatin occupancy of proteins, but the assay is dependent upon availability of antibodies and limited to a single factor at a time. ATAC-seq and DNase-seq quantify chromatin accessibility, which is often used as a proxy measurement to infer RE activity (Boyle et al. 2008; Buenrostro et al. 2015). Typically, activator binding increases local chromatin accessibility and the presence of a cognate TF binding motif within a dynamic RE is used to infer TF binding without the need to perform individual ChIP-seq experiments for each factor (Siersbæk et al. 2011, 2014; Wu et al. 1979). However, these assays do not directly inform on changes in transcription and RNA polymerase dynamics. Conversely, nascent transcription profiling with PRO-seq captures RNA polymerase density genome-wide at high spatial and temporal resolution (Kwak et al. 2013). PRO-seq is limited in its ability to identify potential upstream REs and regulatory factors. Only by combining multiple approaches can one fully capture the signaling dynamics driving transcription regulatory cascades.

Cell differentiation is a tightly regulated process involving many chromatin and transcriptional changes downstream of TF binding (Madsen et al. 2020; Rauch et al. 2019; Siersbæk et al. 2011; Thompson et al. 2016; Tsankov et al. 2015). Adipogenesis is a commonly used *in vitro* model of differentiation. Differentiation of immortalized 3T3-L1 mouse preadipocytes into mature adipocytes is a tractable and well-studied model of cell state transitions (Bernlohr et al. 1984; Green and Kehinde 1974). Mature adipocytes contribute to a multitude of metabolic processes by regulating energy balance, producing hormones, and providing structural and mechanical support (Rosen and

Dutta et al. | G3 | November 16, 2021 | 1–32

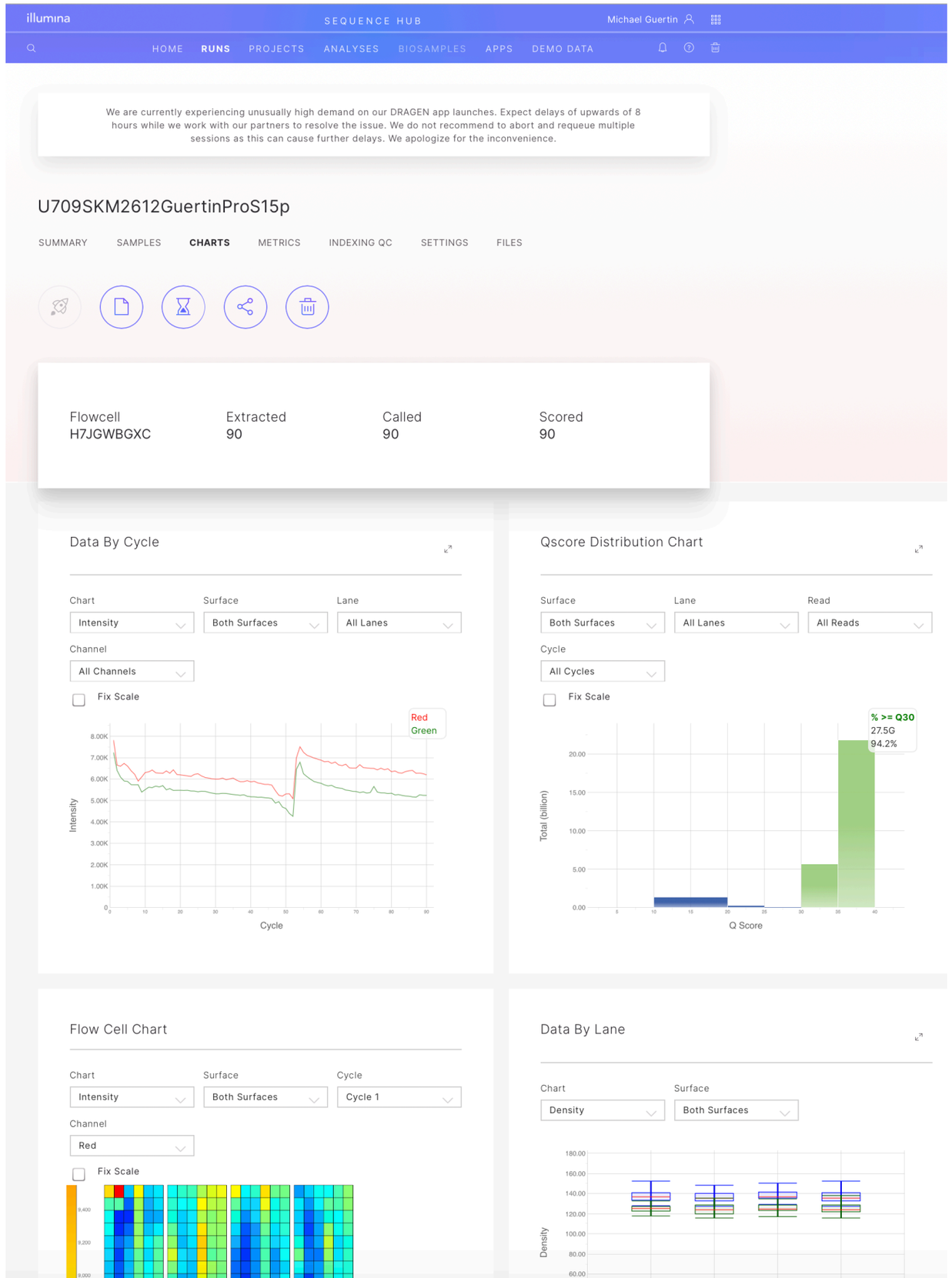
# Web-based solutions for building pipelines




https://usegalaxy.org

# Web-based solutions for building pipelines




https://basespace.illumina.com/

# Web-based solutions for building pipelines


https://genomespace.org

# MEDS 5420 is a GUI-free zone

# Why go GUI-free?

- Less use of system resources
- Generally better for large data
- Remote access to servers
- Easier creation of pipelines and automation
- Flexibility with diverse software
- Customization of parameters and pipelines

# MEDS 5420: what will you gain?

- Learn how to access and navigate your computer via the command line for simple and moderately complex tasks.
- Learn programming strategies useful for processing, parsing, and analysis of data.
- Basic script construction and execution.
- Ability to string together commands ( and / or scripts) and bioinformatics tools into processing pipelines and analysis scripts.
- Visualize data – figure making in R.
- **Google strategies and key words**
- **How to articulate questions and prompts for GPT4**
- **Confidence to analyze genomic data and tackle more complex analyses**

# Course goals: Programming languages

## Command line

| | |
|---|---|
| January 22 | Overview of Molecular Genomics and High Throughput Sequencing Technology |
| 27 | Introduction to the Command Line: directories, head, wc, etc. |
| 29 | Introduction to the Command Line: pipes, compression, and grep |
| February 3 | Introduction to the Command Line: find, cut, variables, scripting, and permissions (Homework 1 assigned) |
| 5 | Introduction to the Command Line: awk |
| 10 | Introduction to the Command Line: Logical operators, loops, and Xanad[illegible] (Miura) |
| 12 | Introduction to the Command Line: Batch scripts, interactive sessions, and software installs |
| 17 | Illumina data format, QC, and preprocessing (HW1 due) |
| 19 | Illumina preprocessing: fastx tools |
| 24 | Aligning Illumina data (Homework 2 assigned) |
| 26 | Transcription Factors and ChIP-seq lecture |
| March 3 | Post-mapping processing with samtools and bedtools |
| 5 | The UCSC genome browser |
| 10 | ChIP-seq Analysis: ChIP-seq peak calling |
| 12 | ChIP-seq Analysis: Analyze ChIP-seq peaks with bedtools and awk (HW 2 due) (Midterm assigned) |
| 24 | ChIP-seq Analysis: de novo motif analysis |
| 26 | ChIP-seq Analysis: Motif queries to genomes and databases |
| 31 | Introduction to R and R studio (Miura Starts) |

## Professor Miura: R

# Course goals: Molecular Genomics assays and analysis

| | |
|---|---|
| January 22 | Overview of Molecular Genomics and High Throughput Sequencing Technology |
| 27 | Introduction to the Command Line: directories, head, wc, etc. |
| 29 | Introduction to the Command Line: pipes, compression, and grep |
| February 3 | Introduction to the Command Line: find, cut, variables, scripting, and permissions (Homework 1 assigned) |
| 5 | Introduction to the Command Line: awk |
| 10 | Introduction to the Command Line: Logical operators, loops, and Xanadu (Miura) |
| 12 | Introduction to the Command Line: Batch scripts, interactive sessions, and software installs |
| 17 | Illumina data format, QC, and preprocessing (HW1 due) |
| 19 | Illumina preprocessing: fastx tools |
| 24 | Aligning Illumina data (Homework 2 assigned) |
| 26 | Transcription Factors and ChIP-seq lecture |
| March 3 | Post-mapping processing with samtools and bedtools |
| 5 | The UCSC genome browser |
| 10 | ChIP-seq Analysis: ChIP-seq peak calling |
| 12 | ChIP-seq Analysis: Analyze ChIP-seq peaks with bedtools and awk (HW 2 due) (Midterm assigned) |
| 24 | ChIP-seq Analysis: de novo motif analysis |
| 26 | ChIP-seq Analysis: Motif queries to genomes and databases |
| 31 | Introduction to R and R studio (Miura Starts) |

**Professor Miura: RNA-seq**

**ChIP-seq**

# Course goals: Creating processing and analysis pipelines

| | |
|---|---|
| January 22 | Overview of Molecular Genomics and High Throughput Sequencing Technology |
| 27 | Introduction to the Command Line: directories, head, wc, etc. |
| 29 | Introduction to the Command Line: pipes, compression, and grep |
| February 3 | Introduction to the Command Line: find, cut, variables, scripting, and permissions (Homework 1 assigned) |
| 5 | Introduction to the Command Line: awk |
| 10 | Introduction to the Command Line: Logical operators, loops, and Xanad[illegible] (Miura) |
| 12 | Introduction to the Command Line: Batch scripts, interactive sessions, and software installs |
| 17 | Illumina data format, QC, and preprocessing (HW1 due) |
| 19 | Illumina preprocessing: fastx tools |
| 24 | Aligning Illumina data (Homework 2 assigned) |
| 26 | Transcription Factors and ChIP-seq lecture |
| March 3 | Post-mapping processing with samtools and bedtools |
| 5 | The UCSC genome browser |
| 10 | ChIP-seq Analysis: ChIP-seq peak calling |
| 12 | ChIP-seq Analysis: Analyze ChIP-seq peaks with bedtools and awk (HW 2 due) (Midterm assigned) |
| 24 | ChIP-seq Analysis: de novo motif analysis |
| 26 | ChIP-seq Analysis: Motif queries to genomes and databases |
| 31 | Introduction to R and R studio (Miura Starts) |

**processing and QC**

**Analysis and interpretation**

# Course goals: important dates

| | |
|---|---|
| January 22 | Overview of Molecular Genomics and High Throughput Sequencing Technology |
| 27 | Introduction to the Command Line: directories, head, wc, etc. |
| 29 | Introduction to the Command Line: pipes, compression, and grep |
| February 3 | Introduction to the Command Line: find, cut, variables, scripting, and permissions (Homework 1 assigned) |
| 5 | Introduction to the Command Line: awk |
| 10 | Introduction to the Command Line: Logical operators, loops, and Xanadu (Miura) |
| 12 | Introduction to the Command Line: Batch scripts, interactive sessions, and software installs |
| 17 | Illumina data format, QC, and preprocessing (HW1 due) |
| 19 | Illumina preprocessing: fastx tools |
| 24 | Aligning Illumina data (Homework 2 assigned) |
| 26 | Transcription Factors and ChIP-seq lecture |
| March 3 | Post-mapping processing with samtools and bedtools |
| 5 | The UCSC genome browser |
| 10 | ChIP-seq Analysis: ChIP-seq peak calling |
| 12 | ChIP-seq Analysis: Analyze ChIP-seq peaks with bedtools and awk (HW 2 due) (Midterm assigned) |
| 24 | ChIP-seq Analysis: de novo motif analysis |
| 26 | ChIP-seq Analysis: Motif queries to genomes and databases |
| 31 | Introduction to R and R studio (Miura Starts) |

**Important**

If “in class” exercises are not completed during class, they must be finished at home between classes.

Come to class with questions about previous lecture!

Dates are subject to change based on our progress. Due dates will be officially determined when assigned.

# Up to the midterm

- Command line usage
- Basic shell scripting
- Server access, usage, etiquette—Xanadu
- QC and preprocessing of Illumina data (ChIP-seq)
- Mapping (alignment to a genome)
- Additional QC and converting files
- Genome browsers
- ChIP-seq analyses:
  - Peak calling
  - Quantification of reads in genomic intervals / windows
  - Sequence motif discovery
  - Transcription factor database queries

# midterm to final

- R language syntax, data types, and resources
- RNA-seq
  - Experimental design
  - Preprocessing, mapping
  - Paired-end vs. single-end processing and visualization in browsers
  - Differential gene expression analysis (DESeq2)
  - Gene set enrichment analysis

# Syllabus: contact and references

MEDS 5420: Molecular Genomic Practicum

Mon, Wed. 1:15-3:15pm
Zoom or in-person
(in-person recommended if you are struggling)

400 Farmington Ave.
Room: R 1401

Instructor: Michael Guertin; guertin@uchc.edu & Pedro Miura; miura@uchc.edu

Office hours: by appointment

Text references:

**Practical Computing for Biologists.** Steven H. D. Haddock & Casey Dunn (2018).

**Getting started with R: an Introduction for Biologists.** Andrew P. Beckerman & Owen L. Petchey (2012)

**R in Action: Data Analysis and Graphics with R.** Robert I. Kabacoff (2011).

**R Graphics 3rd Edition.** Paul Murrell (2018)—https://www.stat.auckland.ac.nz/~paul/RG3e/

Although not necessary for this class, these books can be helpful. Ask your PI to purchase these books.

# Syllabus: assignments and grading

**Homework:** Homework assignments will be announced in class. All assignments will be posted on GitHub and announced in class. Homework will be submitted via email to guertin@uchc.edu. **Assignments should be named with the NetID and assignment number (e.g. xyx15002_HW1).** Assignments are due by 5pm on the scheduled due date. Late assignments will lose 5% of total points per day, including weekends.

| Course Components | Weight |
|---|---|
| In class exercises | 20% |
| Homework | 30% |
| Midterm project | 25% |
| Final project | 25% |

**Grading Scale for MED 5420:**

| Grade | Letter Grade | GPA |
|---|---|---|
| 180-200 | A | 4.0 |
| 165-179 | A- | 3.7 |
| 130-164 | B+ | 3.3 |
| 120-129 | B | 3.0 |
| 110-119 | B- | 2.7 |
| 105-109 | C+ | 2.3 |
| 100-104 | C | 2.0 |
| 95-99 | C- | 1.7 |
| 92-94 | D+ | 1.3 |
| 90-91 | D | 1.0 |
| 88-89 | D- | 0.7 |
| <88 | F | 0.0 |

# Server access at UConn Health

We have access to a special queue on the Xanadu server for this course. I will distribute usernames and passwords during the second week of classes. I recommend using this user account even if you have your own already. This will avoid confusion with directory tree structure and problems with access when the server gets busy. **You will need to transfer your data to your own account before the end of the semester.** To request a personal account fill out the form here: https://bioinformatics.uconn.edu/contact-us/

**Useful links from UConn Computational Biology Core**

Understanding the UConn Xanadu cluster:

https://bioinformatics.uconn.edu/resources-and-events/tutorials-2/xanadu/

Unix basics:

http://bioinformatics.uconn.edu/unix-basics

Other CBC tutorials:

http://bioinformatics.uconn.edu/resources-and-events/tutorials/

# First task: identify / install shell terminal

1. If you're laptop is >4 years old check with me about what type and OS.

2. Mac users will use built in Unix shell called 'Terminal' located in: Applications > Utilities > Terminal.app

3. *PC user resources (posted in syllabus):

   Ubuntu (Linux) is available at Microsoft Store, instruction here:

   https://tutorials.ubuntu.com/tutorial/tutorial-ubuntu-on-windows#0

*I have never owned a PC and haven't used a PC in 15+ years. However, I am confident that we will figure it all out! Any PC experts with command line or remote ssh experience please help out

# Let’s get started!