

# Questions

Use the next 15+ minutes to answer and send me your responses: [guertin@uchc.edu](mailto:guertin@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

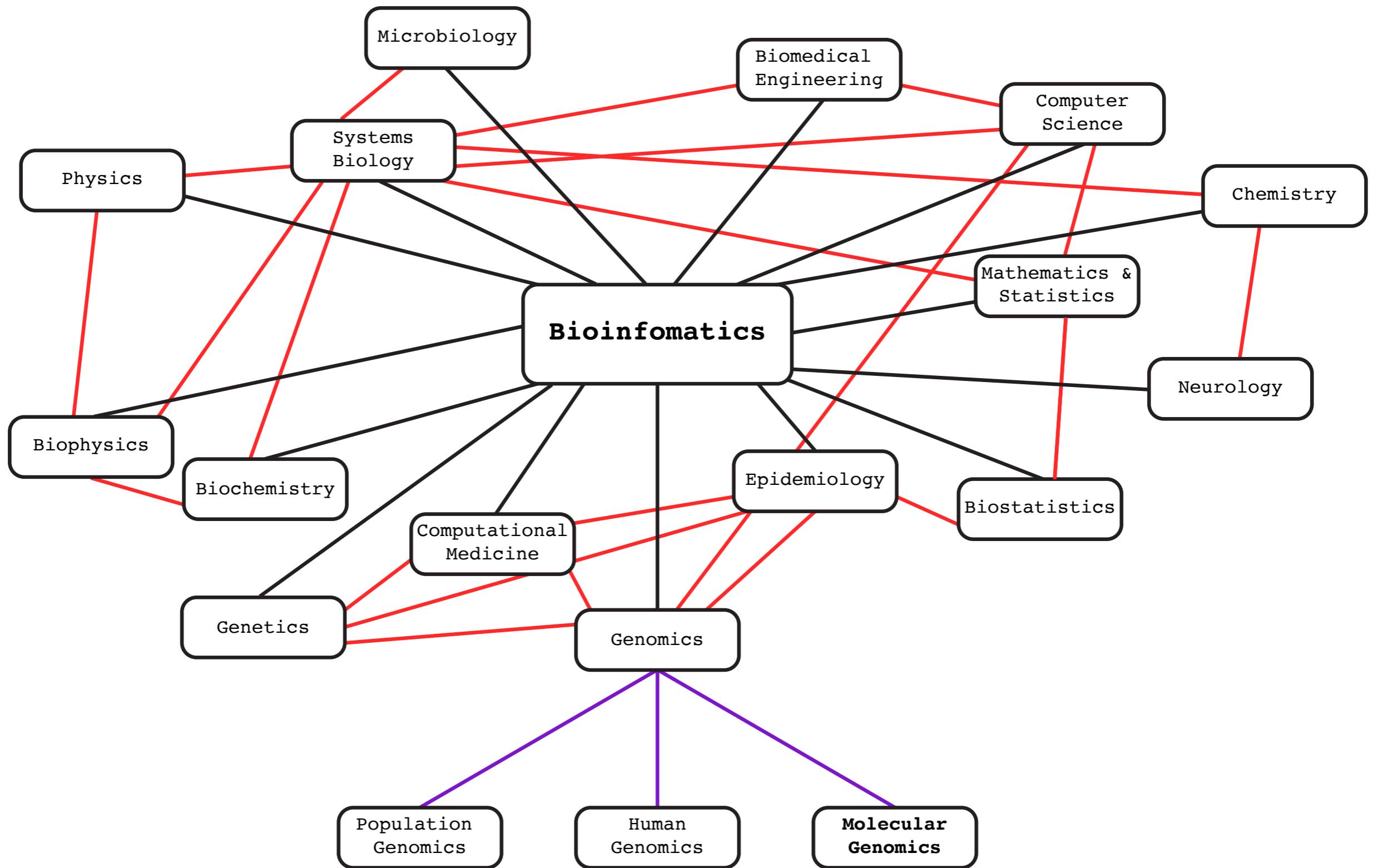


**Spring 2024**  
**Mike Guertin**  
**Featuring Dr. Pedro Miura**

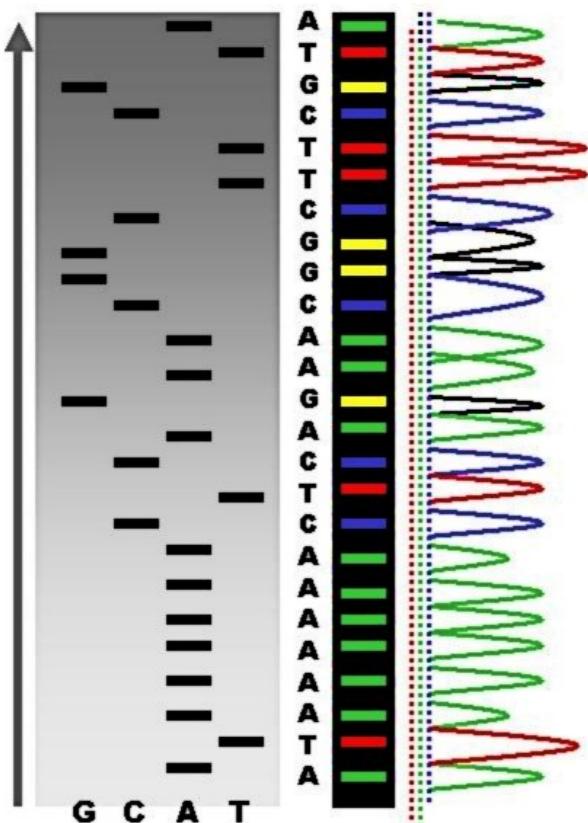
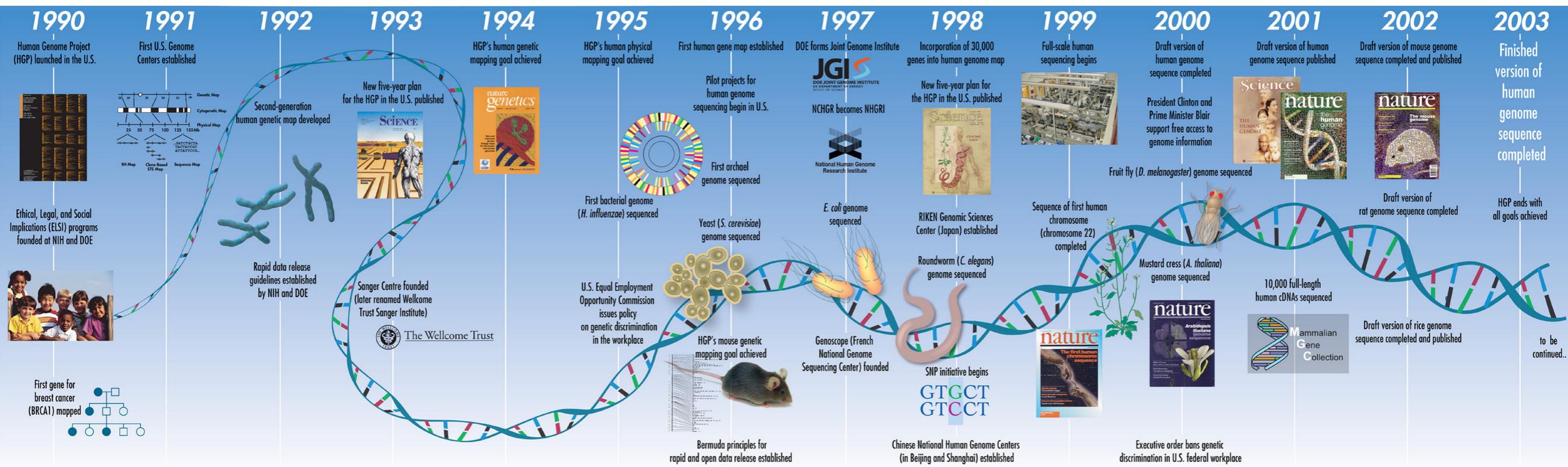


This course is adapted from UConn Professor Leighton Core's course MCB 5430

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



# Human Genome Project



## Sanger sequencing

Cost: 2.7 billion

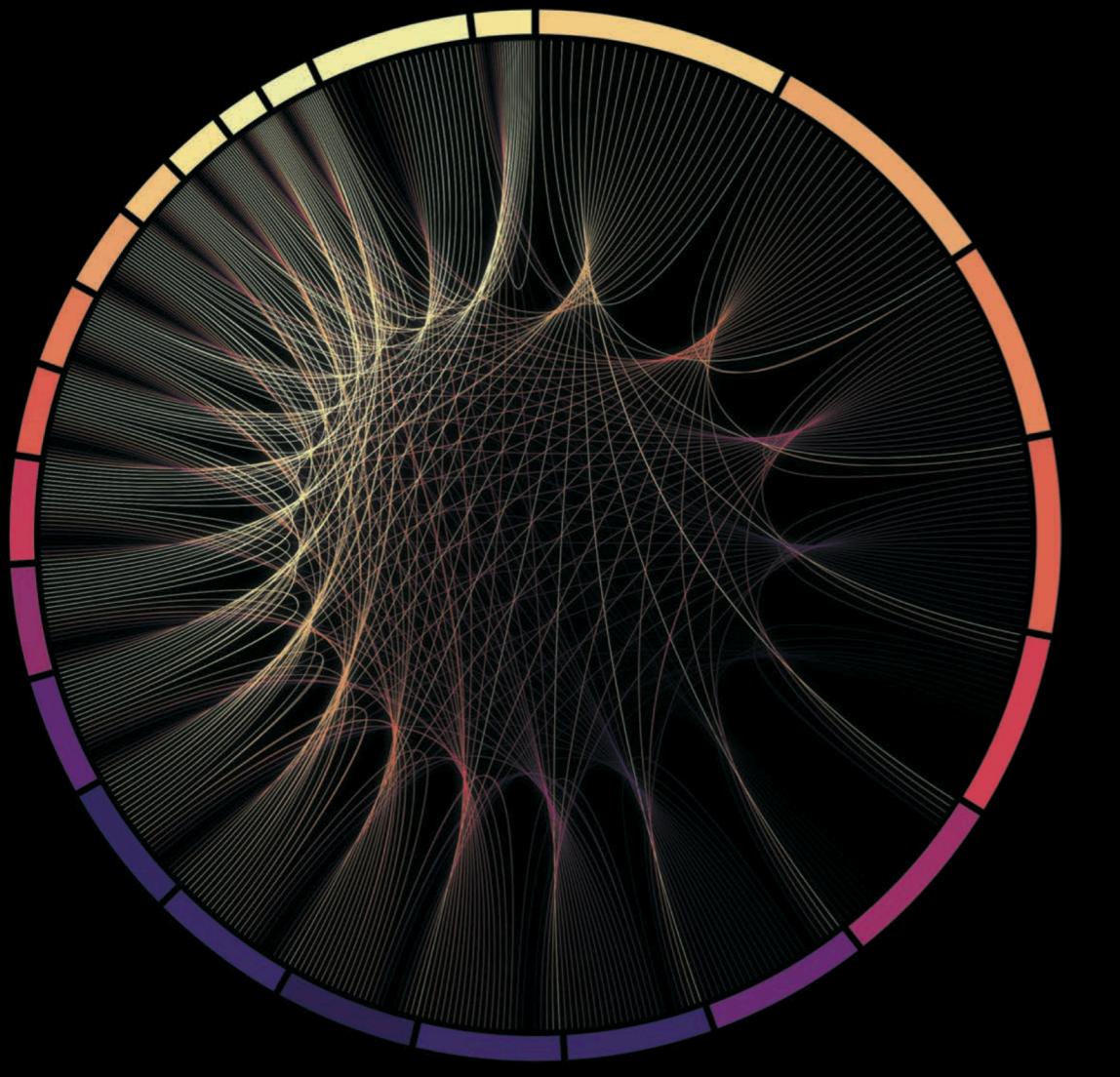
# Contemporary milestones in genomics

[www.nature.com/collections/genomic-sequencing-milestones](https://www.nature.com/collections/genomic-sequencing-milestones)

February 2021

## nature milestones

Genomic sequencing



Produced by:

Nature, Nature Genetics and  
Nature Reviews Genetics

With support from:

illumina®

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

*Genomics of human variation*

*Epigenomics*

*Population Genomics*

*Functional Genomics*

# Genomics?!

*Microbiome Genomics*

*Metagenomics*

*Medical Genomics*

*Structural Genomics*

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

# How do we begin to understand the genome?

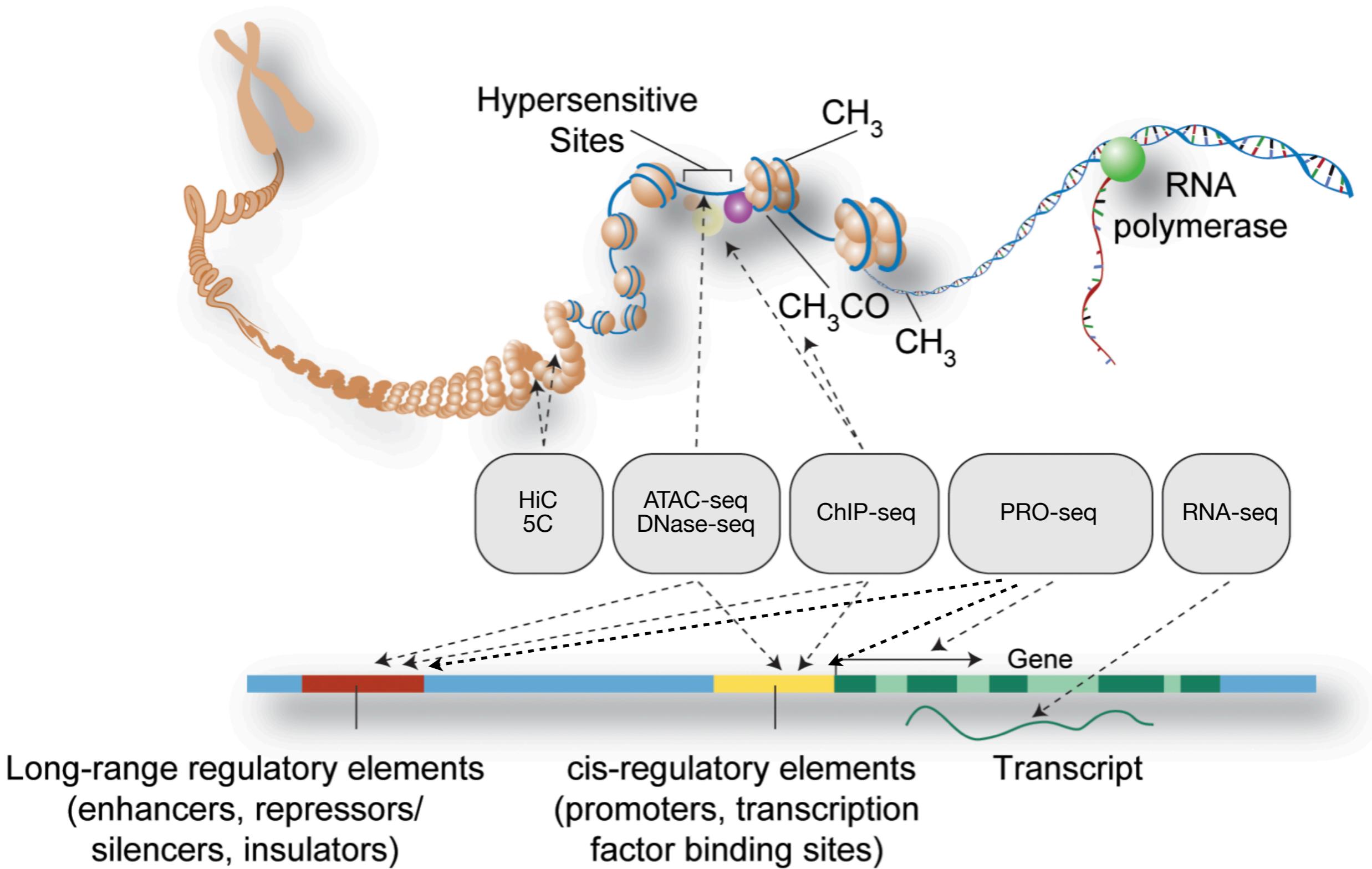


GTCCGCCTAGCGACTGCGTGTACGACGTTACGACTACTGCATGACGCGTACTAGCTAGCATCG  
ACAGTCATCGACTCGCCTCTGCCGTATATATAGCGCTCTCTCTTTTTATATAGAGAGCT  
TCGTGTGGGGTATCAGATCGCATACTGATCGTTGTACGCGATGCAACGCTGCATTGATGAAAAA  
ATCAGACTGCTACGTACGACGATCGATTCTCTGACATGTGAATATGGTCGCGCGCTATGCTA  
CCCGCATATAACGTATCGACATGTCTGCCCGCGATATAATATCCAGACTTGCTGACATAACG  
ATATACTACGATGACCGAT~~GAT~~TAGACTAGCTACAGACGCACTGAAGAGCGCGCTCTATACG  
ATCTATATCTGCATGCTACGACACGTACCGCTATATGCTGCTATGCAGCCGTCACTAGCGCAA  
CGCACTGATGACTAACGCGCTACTGCCCTACTGACTCACTATGCCGCCGCCGTGGGGATA  
TACGCTGATCGTACGCCCTATATCGGATGATCGCGCTCATATCGCATCGCTATCTACGCATA  
TACCAGATCATGCCCTACTATGATTATATCGCTACAGCTAAAGCTCGATCAGATC  
GATAAGACTTATTAGCGTAATATCGTAGCAACTCTATGATTAGCAGGGTCGATAT  
ACGATCAATGAATCTTAACTAACTCTGATATCGATCCGCGCTACAGTTA  
CGCCACGTATCGGATATTTCGATGAAAGTCAGTAGCGCGTATCGGGATT  
ACACGTAATGCGACTACTGACCTACTAGCTAGCACTATT  
TATCATGACGACATCATTCTAGTGTGTGATGATATGCTATA  
GCTACGCGATCGCTAGCTACGTCGTTATGCTACTCTCGTTTTACTA  
ACTGCCGTAATGCGACTACTGACCTACTAGCTAGCACTATT  
CTGACGATGAAATTGGGGGTGTATCATGATGATATGAAATATGACTACTGA  
ACAATCGAAGACTAGCTAGCATGACGCGCTAGCGATGCGCATGCCGATA  
GTCCACAGCTACTATCATGATCGTACGCCCGCGTTCGCCGATGATGC  
ATGCATGCGATGCATACTGCATGACGGGGTAGCATGATCGATCATCAT  
GCAGTACATTGGCATGCTGACTGCATGCATGACTGCATGCATGATGCA  
TACGTCTACAAGGTGCATGCCCACTGACTGACTACTGATGATGAGAGGGGA  
TCGATTGTCGATGCATTGCAGTACTTCATACTAAAGCGCGTGCATA  
CTGACTGCGTACGGGATCGTGTAGCTAGATATGCTAGCTACGGCGATCGATC  
AATATATCGATATAACGCCATAACAGCGGGCTCTCTCGAGAGAGCTCTT

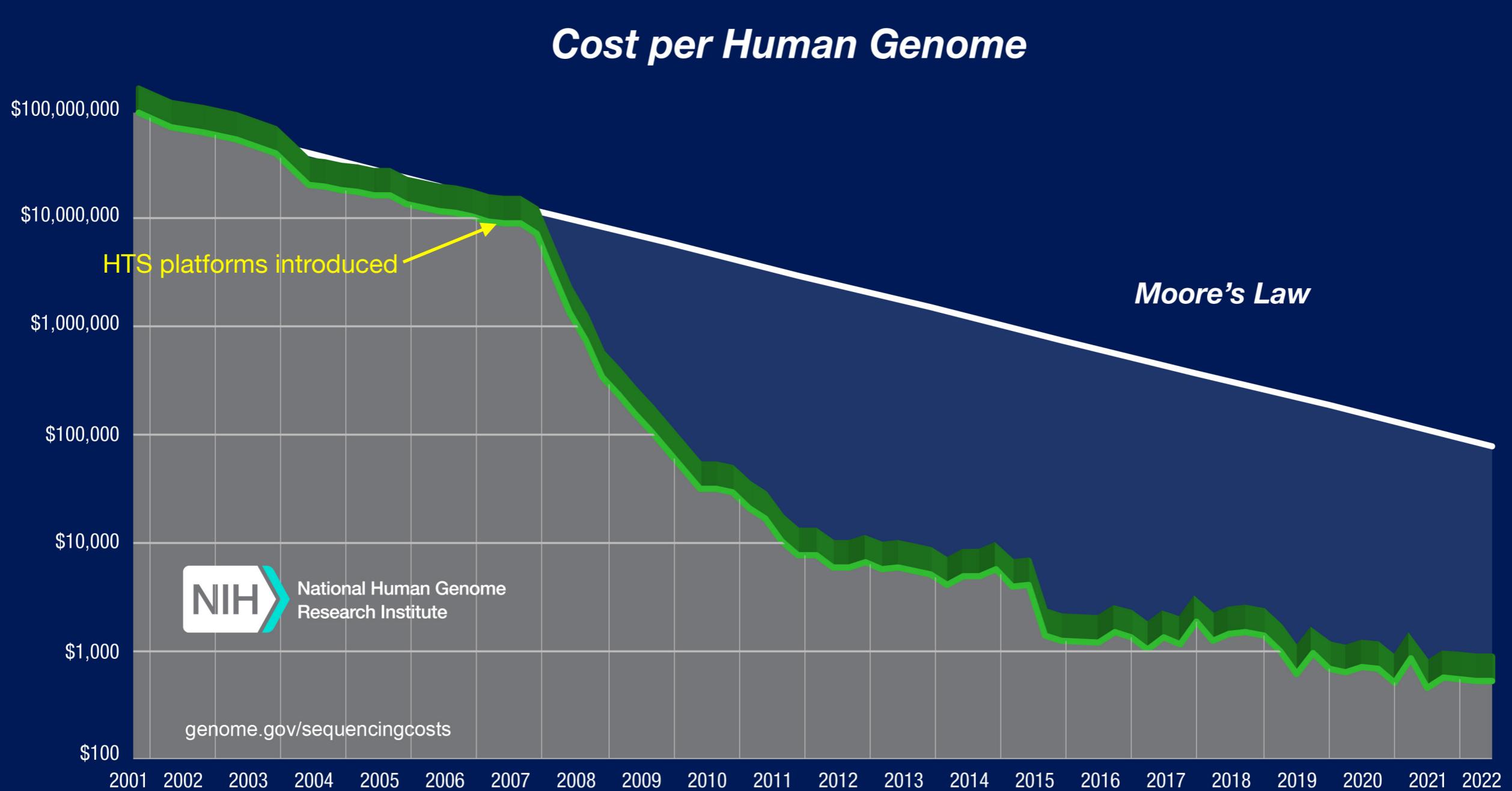
# Questions that can begin to be addressed with Molecular Genomics

- 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



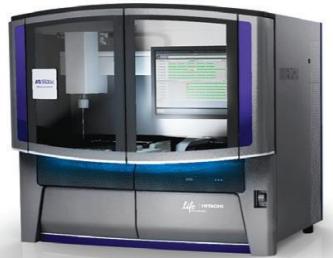
# High throughput sequencing technologies



**Roche 454**



**Ion Torrent**



**ABI Solid**

## Long Read (> 1kb)



**Oxford Nanopore**



**Pacific Biosciences**

## Illumina



**GALIx**



**HiSeq 2500**



**iSeq 100**



**MiniSeq**



**NextSeq 550**



**NextSeq 1000 & 2000**



**NovaSeq 6000**

# 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	GALx	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

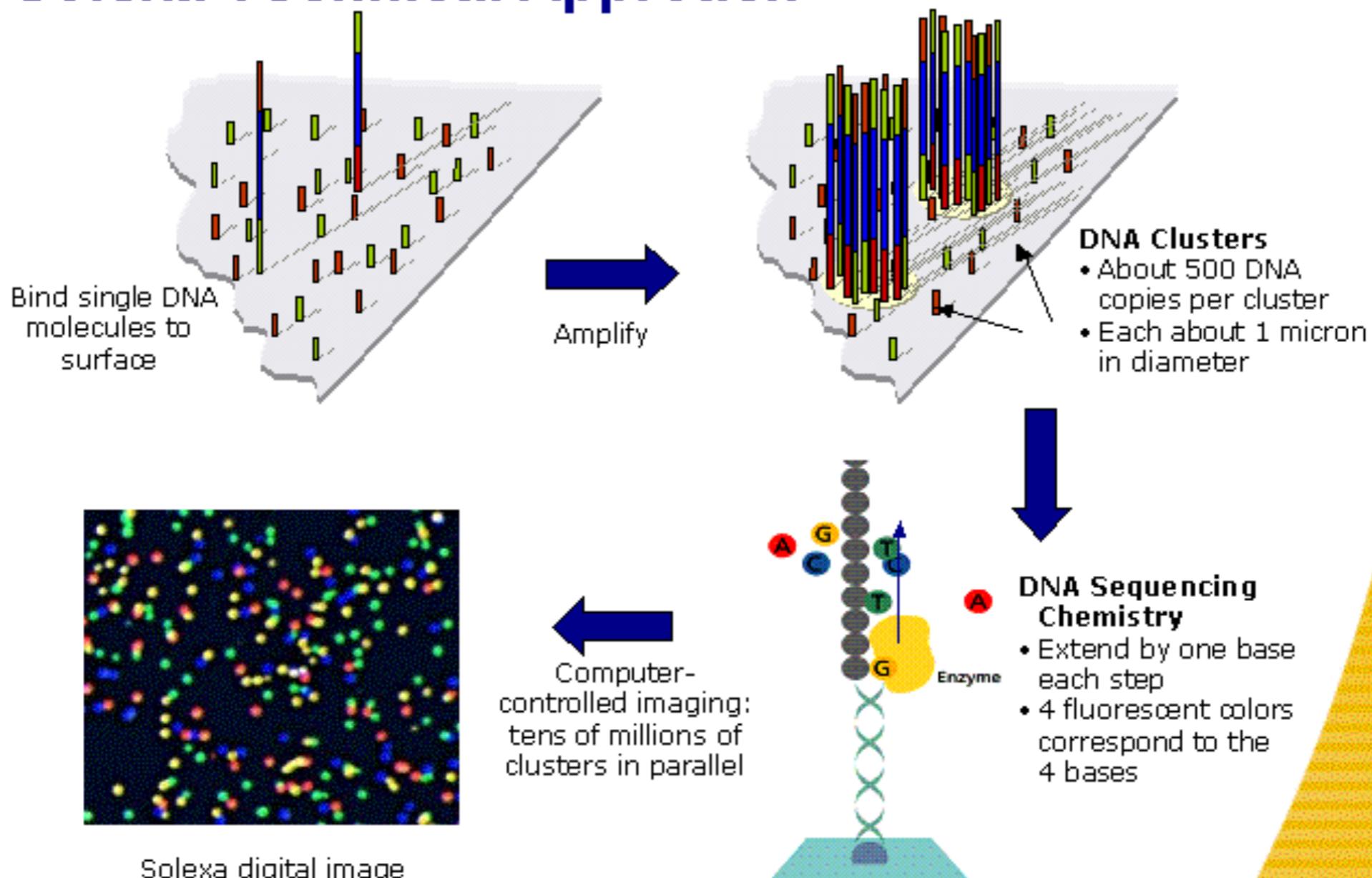


NovaSeq 6000 can sequence the equivalent of 48 human genomes per run at 30x coverage!

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



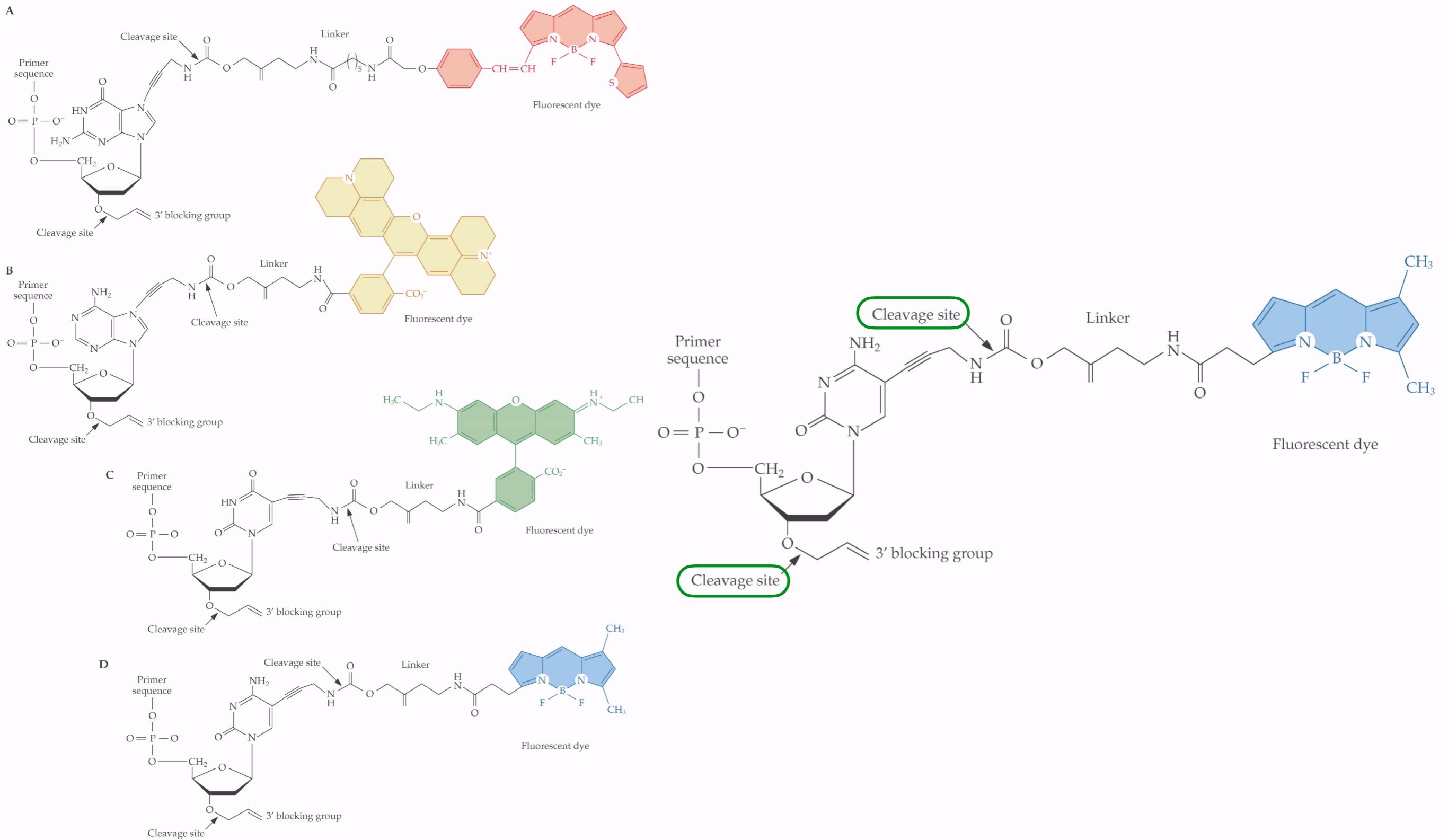
## Solexa Technical Approach



November 2006

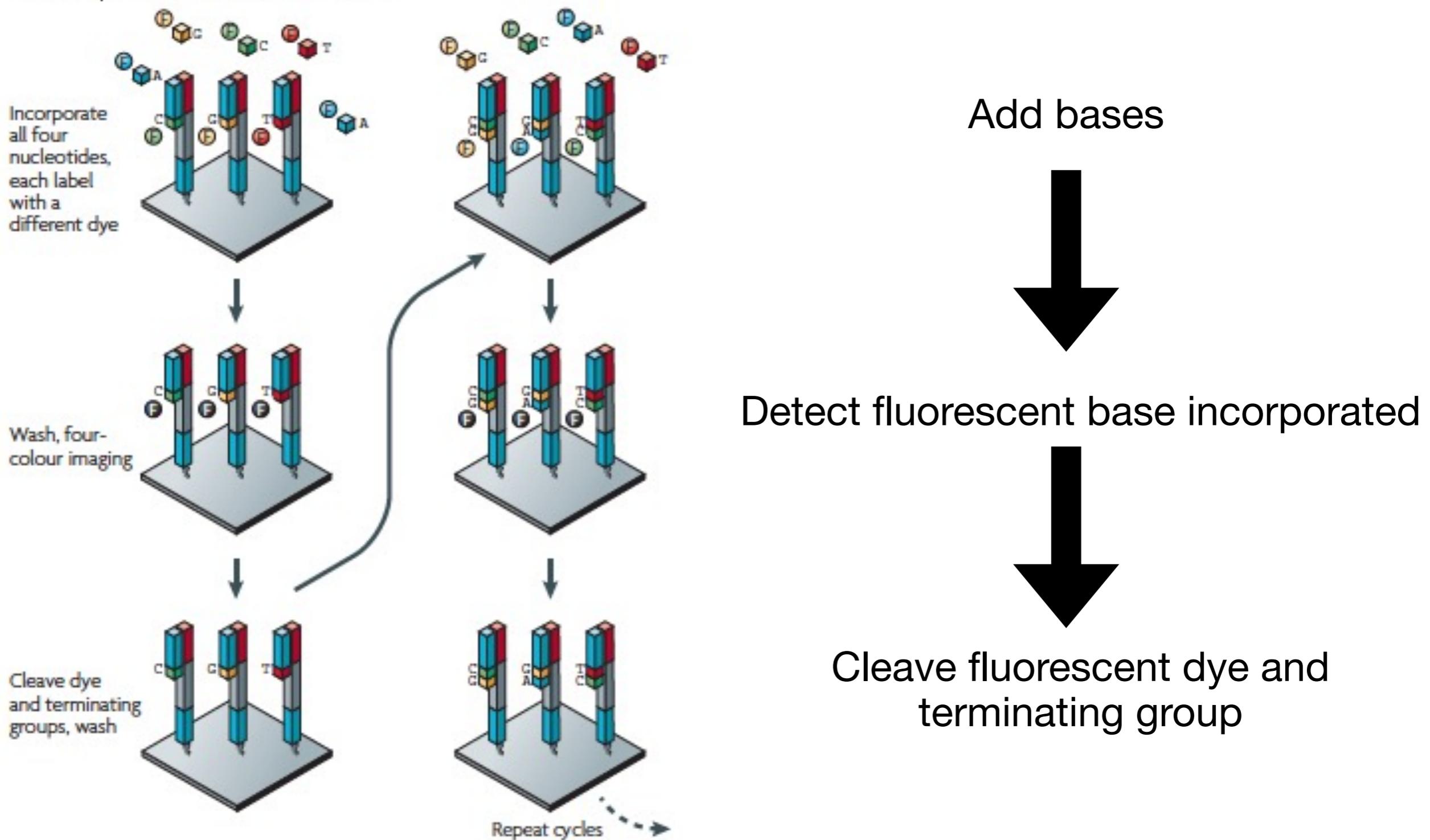
9

# Illumina Sequencing Technology: Dye and Reversible Terminators



# Illumina Sequencing Technology: Dye and Reversible Terminators

a Illumina/Solexa — Reversible terminators



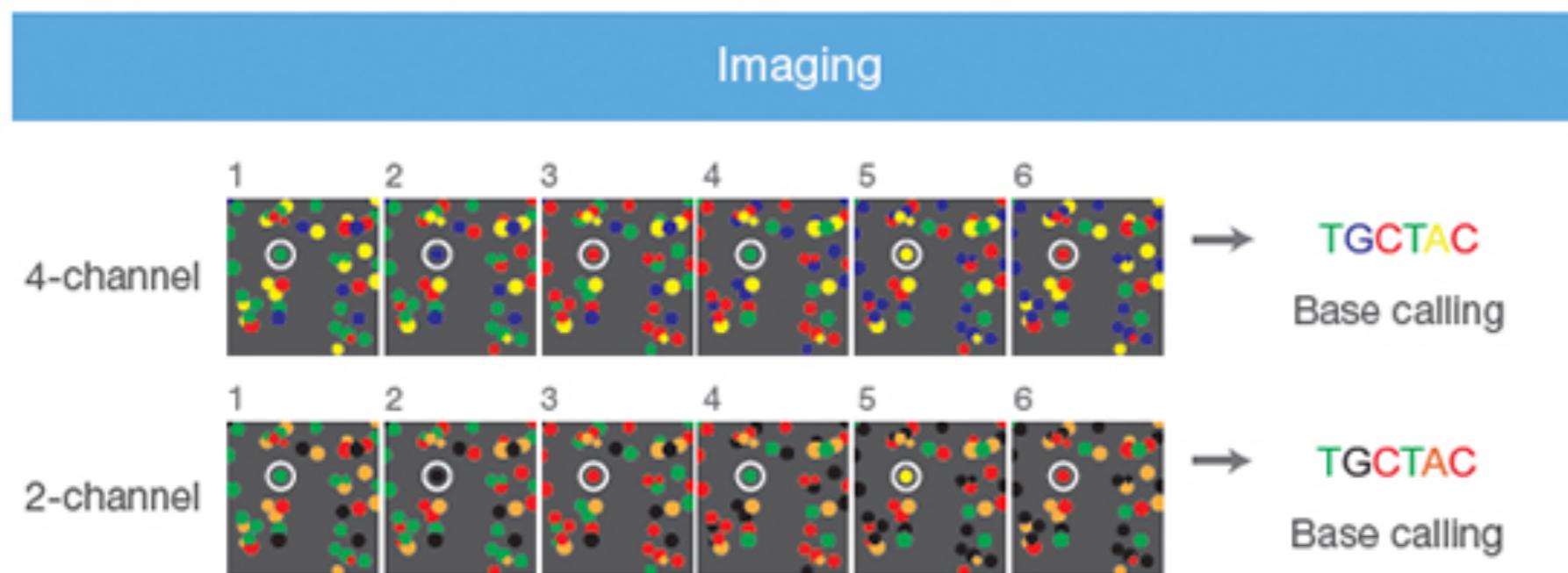
# Illumina sequencing by synthesis (SBS) updates: 2 color imaging



## Benefits:

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

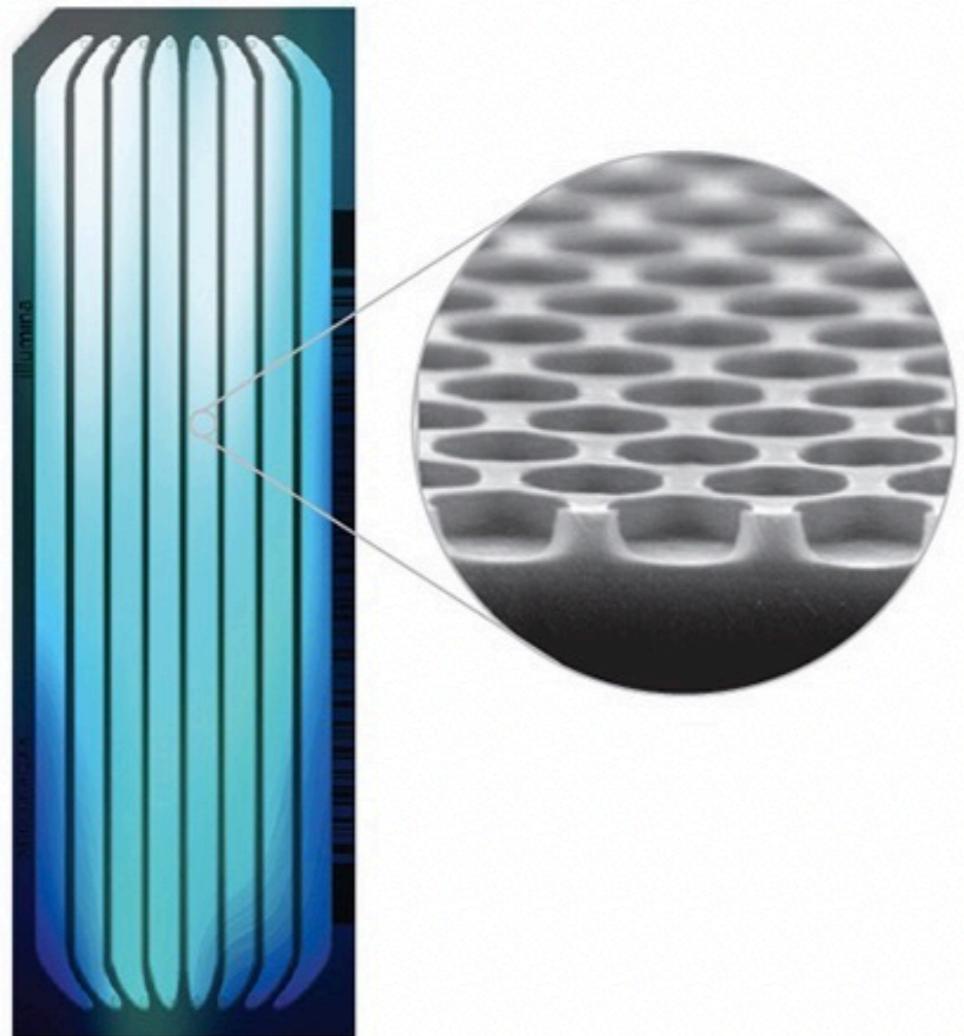
User experience unaffected



Current acquisition method for all Illumina devices

# Illumina SBS updates: patterned flow cell

Distinct, Ordered Nanowell Design



**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>

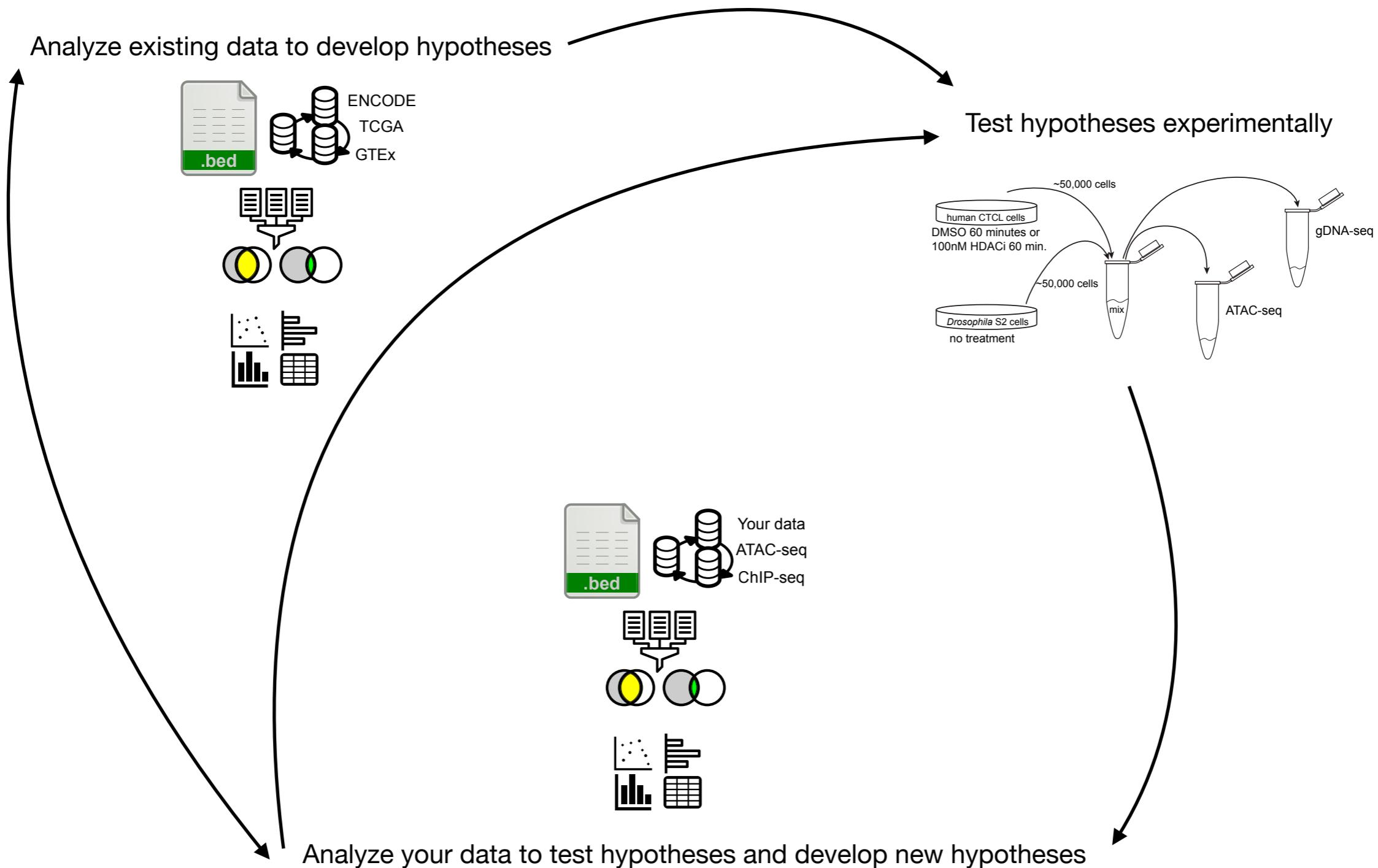
ABI solid: <https://www.youtube.com/watch?v=nIvyF8bFDwM>

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

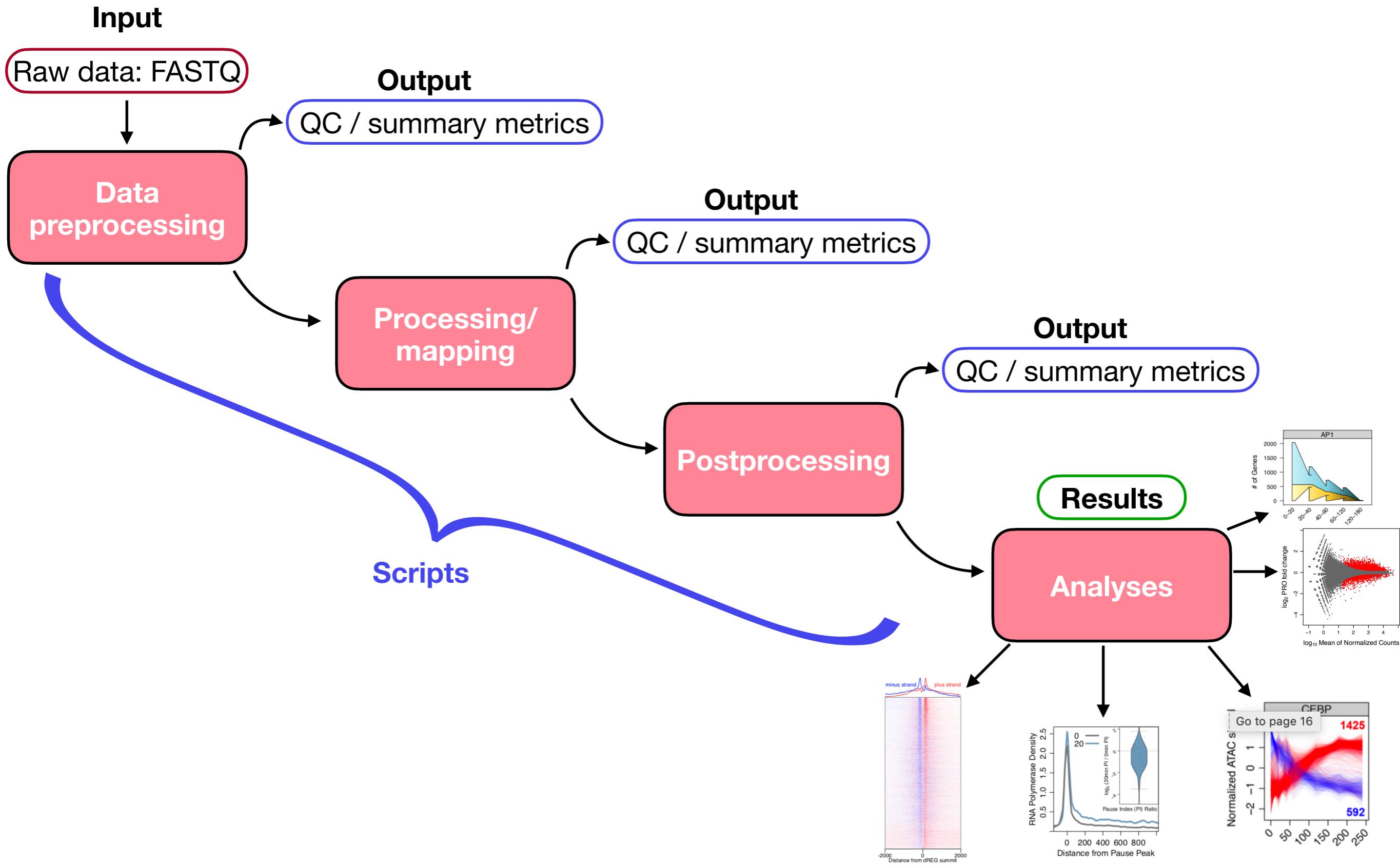


Scientists need to be able to move between the bench and bioinformatics

# Terminology

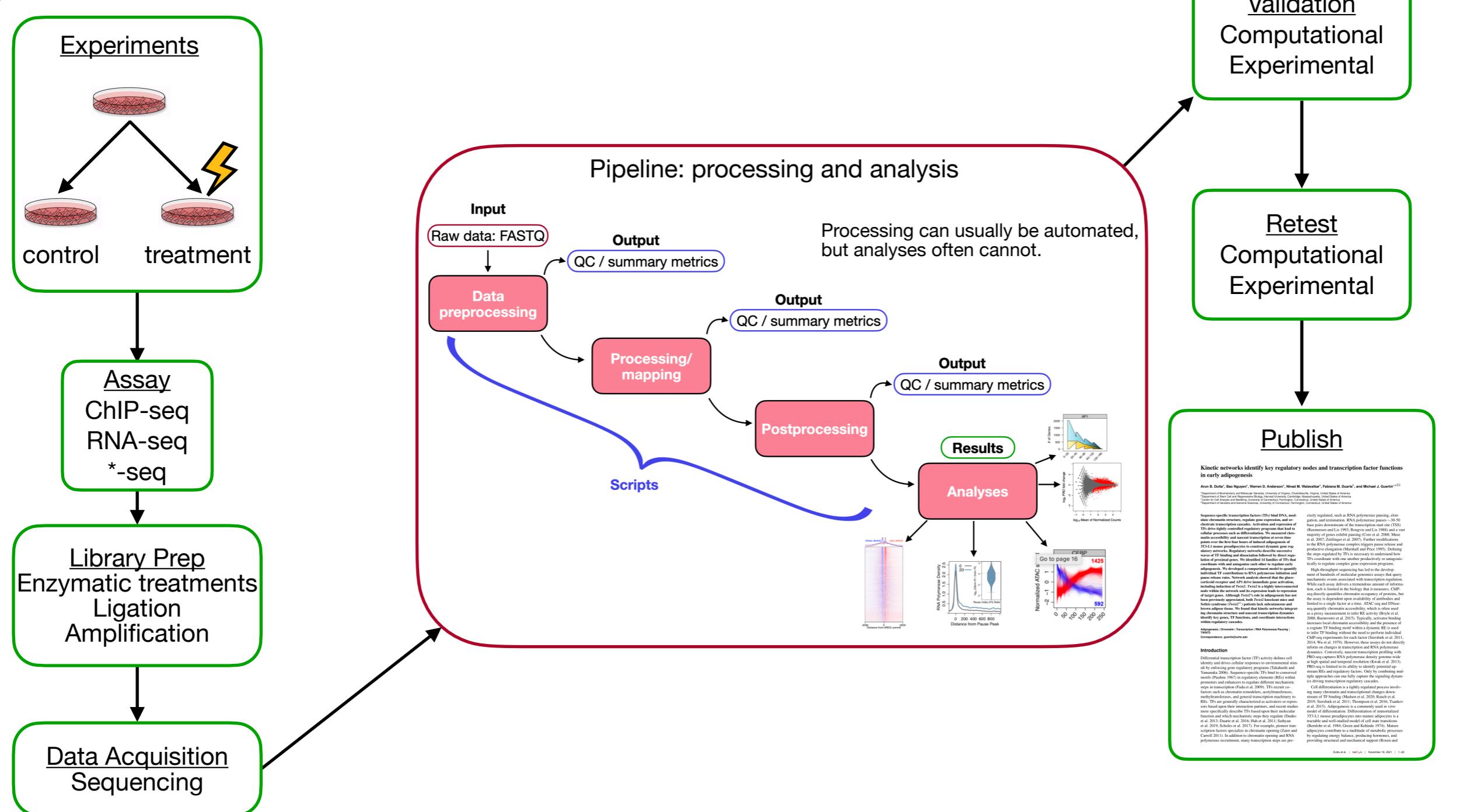
- **Script**
  - Executable document or program listing computer interpretable commands to be executed in sequence.
- **Pipeline**
  - Often a series of independent scripts
    - Output from one script becomes input for next until desired result is achieved
    - Once defined requires limited user effort
    - 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

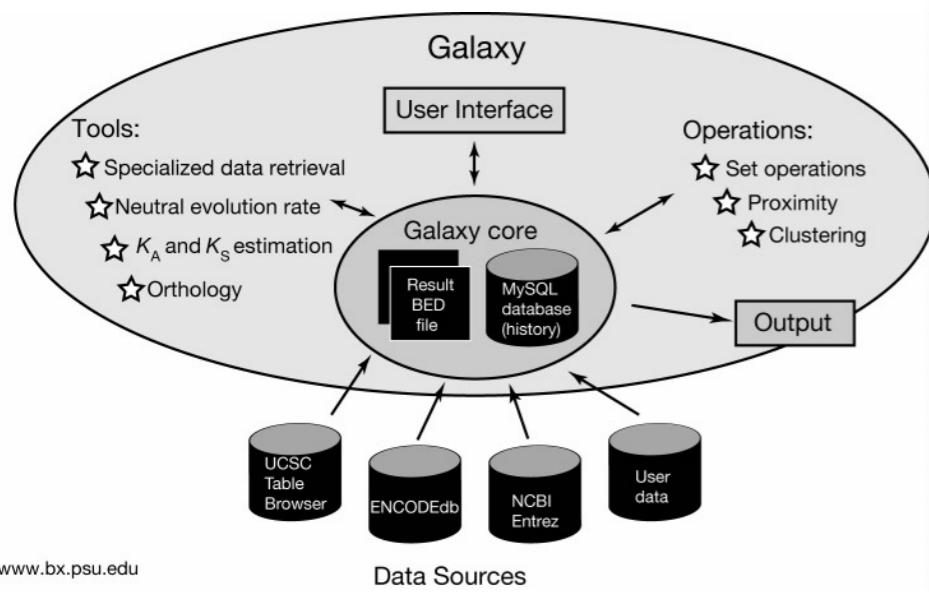


# Workflow can encompass projects and analysis pipelines

## Typical genomics project workflow

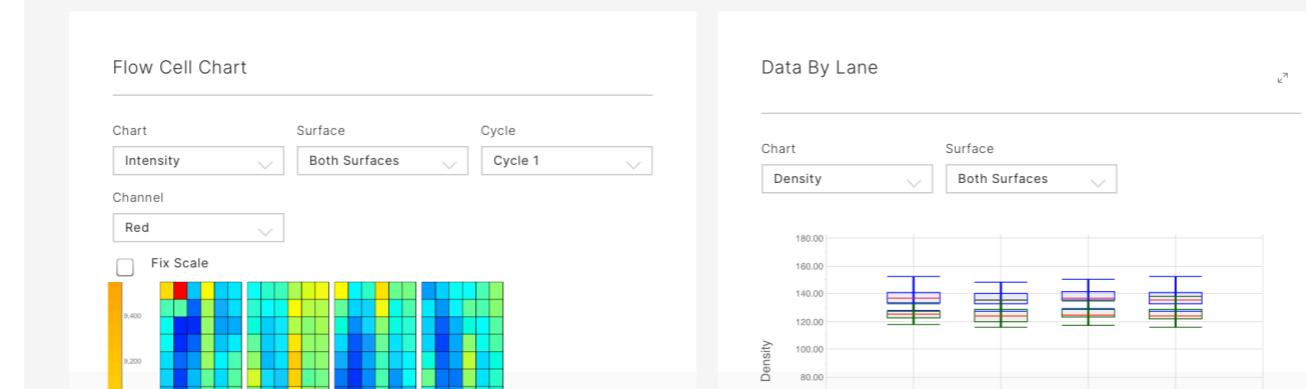
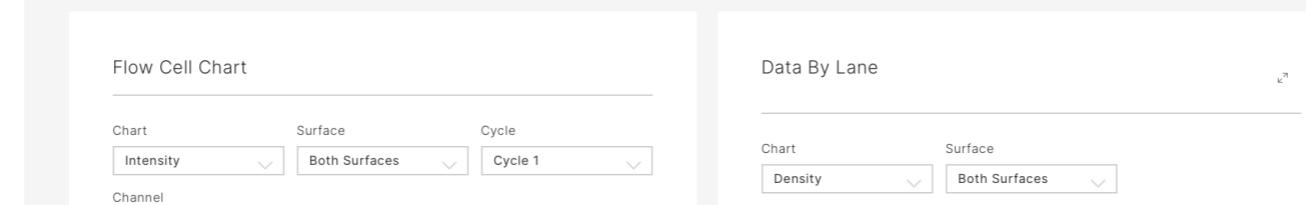
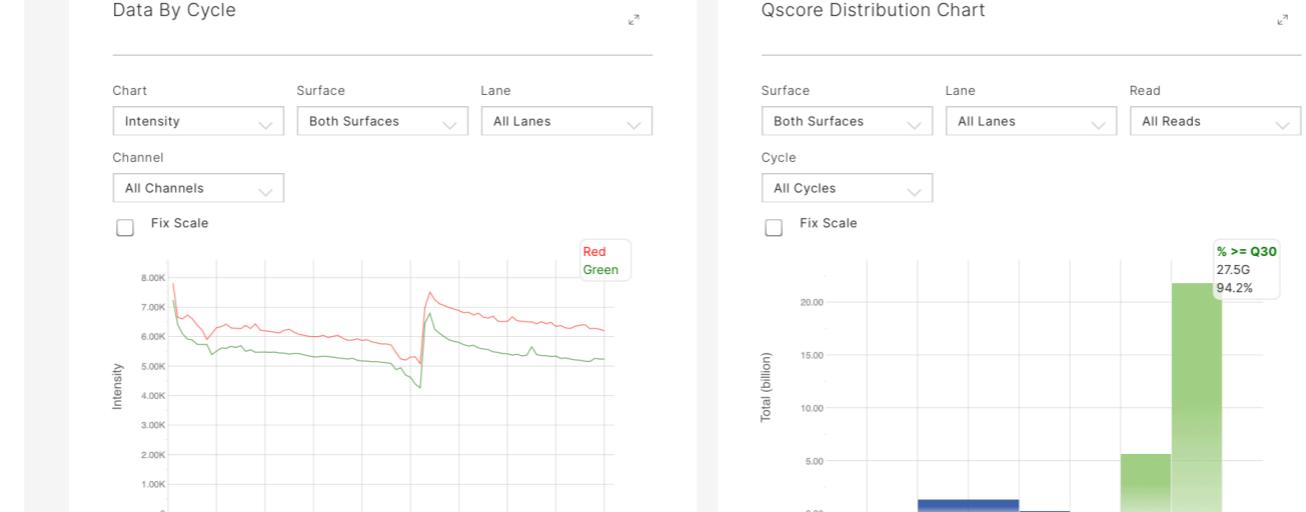
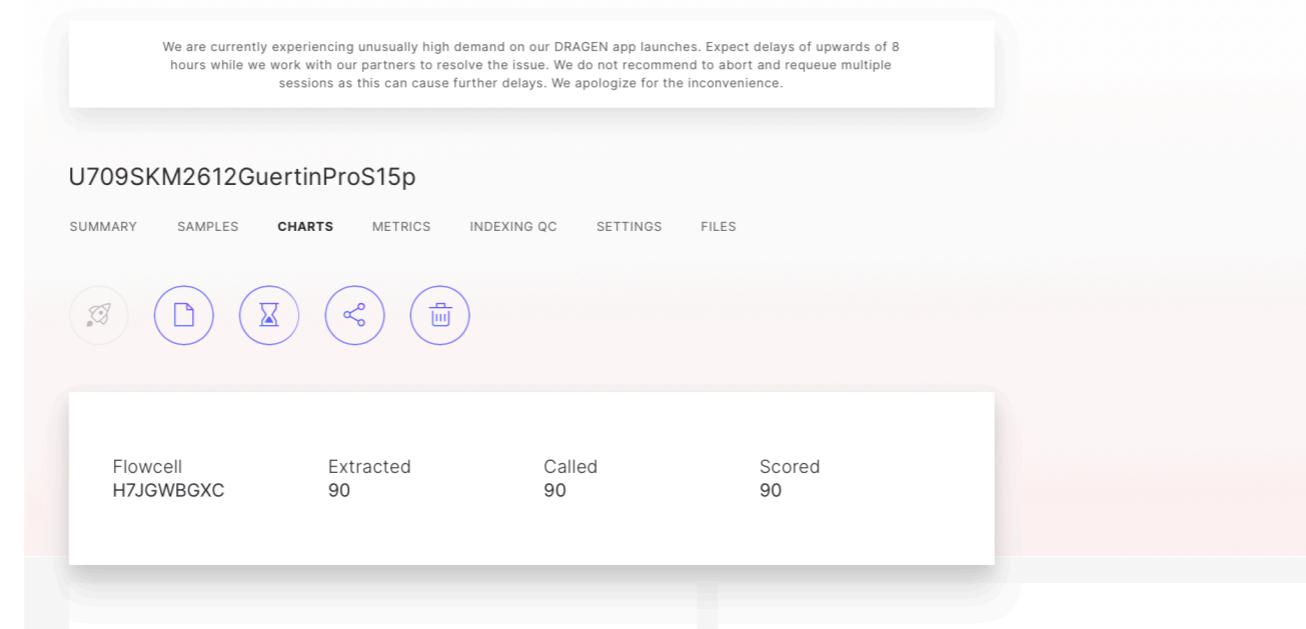
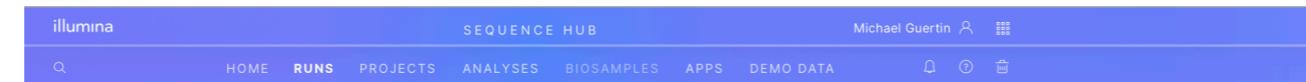


# Web-based solutions for building pipelines



The screenshot shows the Galaxy Workflow Canvas interface. On the left, a sidebar lists various tools under 'Tools'. A workflow history titled 'Workflow constructed from history 'Unnamed history'' is displayed in the center. The workflow consists of three main steps: 1. An 'Input dataset' step with an output labeled 'output'. 2. A 'Map with BWA' step with a reference selected from history, producing a 'FASTQ file' output labeled 'output (sam)'. 3. A 'SAM Filter' step with an 'Edit Distance' field set to '1', producing an output labeled 'output1 (sam)'. The right side of the interface shows the 'Details' panel for the 'SAM Filter' step, where annotations can be added.

# Web-based solutions for building pipelines



# Web-based solutions for building pipelines

The screenshot shows the homepage of the GenomeSpace website. At the top, there is a navigation bar with links to "What is GenomeSpace?", "Tools", "Recipes", "Documentation", "Developers", "Support", and "About". The main header features the "GENOMESPACE" logo with a cloud icon containing DNA helixes and the tagline "Frictionless connection of bioinformatics tools". Below the header, there is a large network graph visualization with blue nodes and green connecting lines. To the right of the graph is a small image of a computer interface showing a pipeline or workflow. On the left side of the page, there is a "STATUS" section indicating "11.18.19 06:02PM" with a red signal icon. A prominent red message states: "With the discontinuation of NHGRI funding for GenomeSpace we have shut down the servers." Below this, text explains that GenomeSpace Recipes can be found at <http://recipes.genomespace.org/>, but data transfer through GenomeSpace will not be available. It also provides a link to more details at <http://www.genomespace.org/news/>. On the right side, there is a "Calendar of Upcoming Events" section with social media icons for Twitter, Facebook, and email. Below that is a "Tweets" section by the @genomespace account, which posts a message about the shutdown due to funding expiration on November 15, 2019. The "WHAT'S NEW" section at the bottom includes "News Highlights" and "GenomeSpace Blog" tabs, with a link to "The GenomeSpace project is ending". The "News Highlights" section contains a summary of the shutdown announcement and a link to "More >".

What is GenomeSpace? Tools Recipes Documentation Developers Support About

# GENOMESPACE

Frictionless connection of bioinformatics tools

STATUS 11.18.19 06:02PM

**With the discontinuation of NHGRI funding for GenomeSpace we have shut down the servers.**

GenomeSpace Recipes can be found at <http://recipes.genomespace.org/> however data transfer through GenomeSpace will not be available.

More details can be found at <http://www.genomespace.org/news/>

### Citing GenomeSpace

To cite your use of GenomeSpace, please reference Qu K, Garamszegi S, Wu F, et al. [Nature Methods](#). 2016 Jan 18. doi: 10.1038/nmeth.3732.

**F1000 Research** Check out our [F1000 GenomeSpace Channel](#) for published, community-contributed [recipes](#).

### WHAT'S NEW

**News Highlights** **GenomeSpace Blog**

**The GenomeSpace project is ending**

**The GenomeSpace project servers are shutting down on November 15, 2019** due to expiration of its NHGRI funding. We would like to thank all GenomeSpace users for their support and for all the important science they have done on the platform over the last nine years. [More >](#)

[See All News Highlights](#)

Calendar of Upcoming Events

Tweets by @genomespace

GenomeSpace Team @genomespace

The GenomeSpace project ends \*tomorrow\* November 15, 2019 due to expiration of its NHGRI funding. Please save any data from your GenomeSpace account by transferring it to your own storage before that date. More details at [genomespace.org/news/the-genom...](#)

Thank you!

Nov 14, 2019

GenomeSpace Team @genomespace

The GenomeSpace project ends on November 15, 2019 due to expiration of its NHGRI funding. Please save any data from your GenomeSpace account by transferring it to your own storage before that date. More details at [genomespace.org/news/the-genom...](#)

Sep 23, 2019

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 18	Overview of Molecular Genomics and High Throughput Sequencing Technology
22	Introduction to the Command Line: directories, head, wc, etc.
25	Introduction to the Command Line: pipes, compression, and grep
29	Introduction to the Command Line: find, cut, variables, scripting, and permissions (Homework 1 assigned)
February 1	Introduction to the Command Line: awk
5	Introduction to the Command Line: Logical operators, loops, and Xanadu
8	Introduction to the Command Line: Batch scripts, interactive sessions, and software installs
12	Illumina data format, QC, and preprocessing (HW1 due)
15	Illumina preprocessing: fastx tools
19	Aligning Illumina data (Homework 2 assigned)
22	Transcription Factors and ChIP-seq lecture
26	Post-mapping processing with samtools and bedtools
29	The UCSC genome browser (Dr. Guertin may pre-record or lecture remotely)
March 4	ChIP-seq Analysis: ChIP-seq peak calling
7	ChIP-seq Analysis: Analyze ChIP-seq peaks with bedtools and awk (HW 2 due) (Midterm assigned)
18	ChIP-seq Analysis: de novo motif analysis (pre-recorded) Dr. Guertin is in Utah
21	ChIP-seq Analysis: Motif queries to genomes and databases
25	Introduction to R
28	RNA-seq lecture

## R

April 1	R functions and plotting (Midterm due) (Homework 3 assigned)
4	RNA-seq Analysis: RNA-seq mapping
8	RNA-seq Analysis: visualization
11	RNA-seq Analysis: transcript assembly (recommended CCAM seminar at 4pm in Grossman Auditorium at 400 Farmington: Dr. Robin Dowell)
15	RNA-seq Analysis: Differential Expression (HW3 due) (Final Assigned)
18	RNA-seq Analysis: Gene Set Enrichment analysis
22	ATAC-seq Analysis
25	ATAC-seq Analysis

# Course goals: Molecular Genomics assays and analysis

January 18	Overview of Molecular Genomics and High Throughput Sequencing Technology
22	Introduction to the Command Line: directories, head, wc, etc.
25	Introduction to the Command Line: pipes, compression, and grep
29	Introduction to the Command Line: find, cut, variables, scripting, and permissions (Homework 1 assigned)
February 1	Introduction to the Command Line: awk
5	Introduction to the Command Line: Logical operators, loops, and Xanadu
8	Introduction to the Command Line: Batch scripts, interactive sessions, and software installs
12	Illumina data format, QC, and preprocessing (HW1 due)
15	Illumina preprocessing: fastx tools
19	Aligning Illumina data (Homework 2 assigned)
22	Transcription Factors and ChIP-seq lecture
26	Post-mapping processing with samtools and bedtools
29	The UCSC genome browser (Dr. Guertin may pre-record or lecture remotely)
March 4	ChIP-seq Analysis: ChIP-seq peak calling
7	ChIP-seq Analysis: Analyze ChIP-seq peaks with bedtools and awk (HW 2 due) (Midterm assigned)
18	ChIP-seq Analysis: de novo motif analysis (pre-recorded) Dr. Guertin is in Utah
21	ChIP-seq Analysis: Motif queries to genomes and databases
25	Introduction to R
28	RNA-seq lecture

## RNA-seq

April 1	R functions and plotting (Midterm due) (Homework 3 assigned)
4	RNA-seq Analysis: RNA-seq mapping
8	RNA-seq Analysis: visualization
11	RNA-seq Analysis: transcript assembly (recommended CCAM seminar at 4pm in Grossman Auditorium at 400 Farmington: Dr. Robin Dowell)
15	RNA-seq Analysis: Differential Expression (HW3 due) (Final Assigned)
18	RNA-seq Analysis: Gene Set Enrichment analysis
22	ATAC-seq Analysis
25	ATAC-seq Analysis

## ATAC-seq

## ChIP-seq

# Course goals: Molecular Genomics assays and analysis

January 18	Overview of Molecular Genomics and High Throughput Sequencing Technology
22	Introduction to the Command Line: directories, head, wc, etc.
25	Introduction to the Command Line: pipes, compression, and grep
29	Introduction to the Command Line: find, cut, variables, scripting, and permissions (Homework 1 assigned)
February 1	Introduction to the Command Line: awk
5	Introduction to the Command Line: Logical operators, loops, and Xanadu
8	Introduction to the Command Line: Batch scripts, interactive sessions, and software installs
12	Illumina data format, QC, and preprocessing (HW1 due)
15	Illumina preprocessing: fastx tools
19	Aligning Illumina data (Homework 2 assigned)
22	Transcription Factors and ChIP-seq lecture
26	Post-mapping processing with samtools and bedtools
29	The UCSC genome browser (Dr. Guertin may pre-record or lecture remotely)
March 4	ChIP-seq Analysis: ChIP-seq peak calling
7	ChIP-seq Analysis: Analyze ChIP-seq peaks with bedtools and awk (HW 2 due) (Midterm assigned)
18	ChIP-seq Analysis: de novo motif analysis (pre-recorded) Dr. Guertin is in Utah
21	ChIP-seq Analysis: Motif queries to genomes and databases
25	Introduction to R
28	RNA-seq lecture

## RNA-seq

April 1	R functions and plotting (Midterm due) (Homework 3 assigned)
4	RNA-seq Analysis: RNA-seq mapping
8	RNA-seq Analysis: visualization
11	RNA-seq Analysis: transcript assembly (recommended CCAM seminar at 4pm in Grossman Auditorium at 400 Farmington: Dr. Robin Dowell)
15	RNA-seq Analysis: Differential Expression (HW3 due) (Final Assigned)
18	RNA-seq Analysis: Gene Set Enrichment analysis
22	ATAC-seq Analysis
25	ATAC-seq Analysis

## ATAC-seq

## ChIP-seq

\*ATAC-seq analysis is not tested in this course.  
2023 ATAC Lectures were not well-attended.  
2023 feedback did not address the usefulness of the ATAC lectures.

We can omit the ATAC and spend more time with other material or keep this as “optional” for the class.

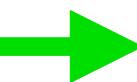
# Course goals: Creating processing and analysis pipelines

January 18	Overview of Molecular Genomics and High Throughput Sequencing Technology
22	Introduction to the Command Line: directories, head, wc, etc.
25	Introduction to the Command Line: pipes, compression, and grep
29	Introduction to the Command Line: find, cut, variables, scripting, and permissions (Homework 1 assigned)
February 1	Introduction to the Command Line: awk
5	Introduction to the Command Line: Logical operators, loops, and Xanadu
8	Introduction to the Command Line: Batch scripts, interactive sessions, and software installs
12	Illumina data format, QC, and preprocessing (HW1 due)
15	Illumina preprocessing: fastx tools
19	Aligning Illumina data (Homework 2 assigned)
22	Transcription Factors and ChIP-seq lecture
26	Post-mapping processing with samtools and bedtools
29	The UCSC genome browser (Dr. Guertin may pre-record or lecture remotely)
March 4	ChIP-seq Analysis: ChIP-seq peak calling
7	ChIP-seq Analysis: Analyze ChIP-seq peaks with bedtools and awk (HW 2 due) (Midterm assigned)
18	ChIP-seq Analysis: de novo motif analysis (pre-recorded) Dr. Guertin is in Utah
21	ChIP-seq Analysis: Motif queries to genomes and databases
25	Introduction to R
28	RNA-seq lecture

April 1	R functions and plotting (Midterm due) (Homework 3 assigned)
4	RNA-seq Analysis: RNA-seq mapping
8	RNA-seq Analysis: visualization
11	RNA-seq Analysis: transcript assembly (recommended CCAM seminar at 4pm in Grossman Auditorium at 400 Farmington: Dr. Robin Dowell)
15	RNA-seq Analysis: Differential Expression (HW3 due) (Final Assigned)
18	RNA-seq Analysis: Gene Set Enrichment analysis
22	ATAC-seq Analysis
25	ATAC-seq Analysis

**Analysis and interpretation  
processing and QC**

# Course goals: important dates

January 18	Overview of Molecular Genomics and High Throughput Sequencing Technology		
22	Introduction to the Command Line: directories, head, wc, etc.		
25	Introduction to the Command Line: pipes, compression, and grep		
29	Introduction to the Command Line: find, cut, variables, scripting, and permissions (Homework 1 assigned)		
February 1	Introduction to the Command Line: awk		
5	Introduction to the Command Line: Logical operators, loops, and Xanadu		
8	Introduction to the Command Line: Batch scripts, interactive sessions, and software installs		
12	Illumina data format, QC, and preprocessing (HW1 due)		
15	Illumina preprocessing: fastx tools		
19	Aligning Illumina data (Homework 2 assigned)		
22	Transcription Factors and ChIP-seq lecture		
26	Post-mapping processing with samtools and bedtools		
29	The UCSC genome browser (Dr. Guertin may pre-record or lecture remotely)		
March 4	ChIP-seq Analysis: ChIP-seq peak calling		
7	ChIP-seq Analysis: Analyze ChIP-seq peaks with bedtools and awk (HW 2 due) (Midterm assigned)		
18	ChIP-seq Analysis: de novo motif analysis (pre-recorded) Dr. Guertin is in Utah		
21	ChIP-seq Analysis: Motif queries to genomes and databases		
25	Introduction to R		
28	RNA-seq lecture		
April 1	R functions and plotting (Midterm due) (Homework 3 assigned)		
4	RNA-seq Analysis: RNA-seq mapping		
8	RNA-seq Analysis: visualization		
11	RNA-seq Analysis: transcript assembly (recommended CCAM seminar at 4pm in Grossman Auditorium at 400 Farmington: Dr. Robin Dowell)		
15	RNA-seq Analysis: Differential Expression (HW3 due) (Final Assigned)		
18	RNA-seq Analysis: Gene Set Enrichment analysis		
22	ATAC-seq Analysis		
25	ATAC-seq Analysis		

**\*\*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
- Plotting data (base R and lattice)
- 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, Thurs. 1:00-3:00pm

Zoom or in-person

(in-person recommended if you are struggling)

400 Farmington Ave.

Room: R 1401

Instructor: Michael Guertin; [guertin@uchc.edu](mailto:guertin@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](mailto: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 your laptop is >3 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!**