Teaching Computational Genomics:

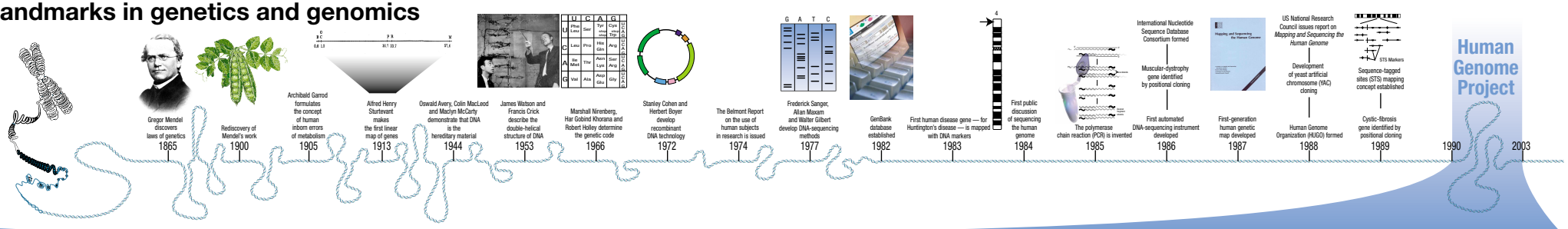
A Tale of Tables and Tests

Texas A&M Research Computing Symposium
2021-05-24

Rodolfo Aramayo - Department of Biology - Texas A&M University

The Genomic Revolution

Landmarks in genetics and genomics



1990
The Human Genome Project (HGP) launched in the United States

Ethical, legal and social implications (ELSI) programmes founded at the US National Institutes of Health (NIH) and Department of Energy (DOE)

First gene for breast cancer (*BRCA1*) mapped

1991
First US genome centres established

1992
Second-generation human genetic map developed

Rapid-data-release guidelines established by the NIH and DOE

1993
New five-year plan for the HGP in the United States published

The Sanger Centre founded near Cambridge, UK, (later renamed the Wellcome Trust Sanger Institute)

1994
The HGP's human genetic mapping goal achieved

1995
The HGP's human physical mapping goal achieved

First bacterial genome (*Haemophilus influenzae*) sequenced

US Equal Employment Opportunity Commission issues policy on genetic discrimination in the workplace

1996
First human gene map established

Pilot projects for human genome sequencing begin in the United States

First archaeal genome sequenced

Yeast (*Saccharomyces cerevisiae*) genome sequenced

The HGP's mouse genetic mapping goal achieved

Bermuda principles for rapid and open data release established

1997
The DOE forms the Joint Genome Institute

US National Center for Human Genome Research becomes the National Human Genome Research Institute (NHGRI)

Escherichia coli genome sequenced

Genoscope (French National Genome Sequencing Center) founded near Paris

1998
Incorporation of 30,000 genes into human genome map

New five-year plan for the HGP in the United States published

RIKEN Genomic Sciences Center established in Japan

Roundworm (*Caenorhabditis elegans*) genome sequenced

Single-nucleotide polymorphism (SNP) initiative begins

Chinese National Human Genome Centers established in Beijing and Shanghai

1999
Full-scale human genome sequencing begins

Sequence of first human chromosome (chromosome 22) completed

2000
Draft version of human genome sequence completed

US President Bill Clinton and UK Prime Minister Tony Blair support free access to genome information

Fruitfly (*Drosophila melanogaster*) genome sequenced

Mustard cress (*Arabidopsis thaliana*) genome sequenced

Executive order bans genetic discrimination in US federal workplace

2001
Draft version of human genome sequence published

10,000 full-length human cDNAs sequenced

Mammalian Gene Collection

2002
Draft version of mouse genome sequence completed and published

Draft version of rice genome sequence completed and published

Draft version of rat genome sequence completed

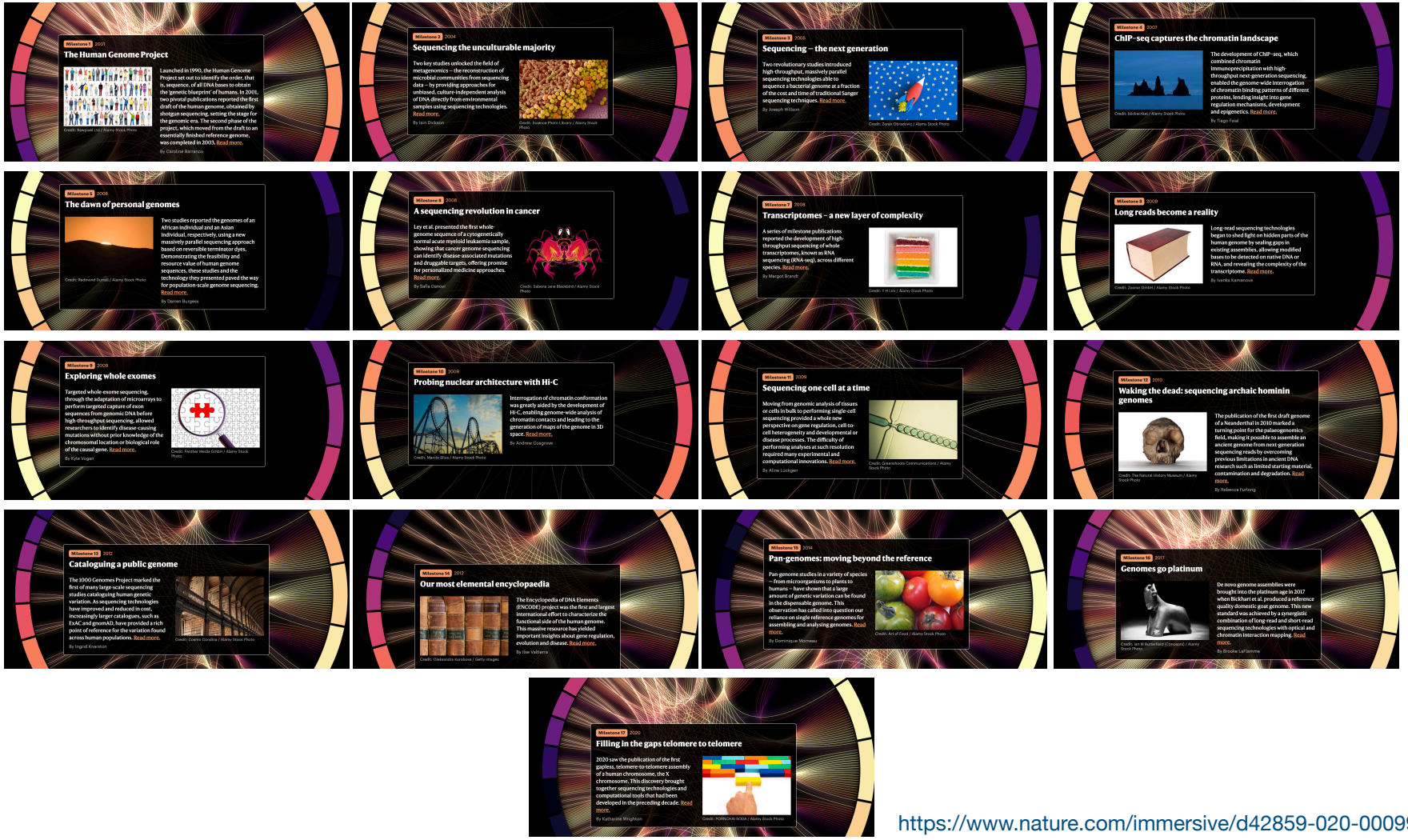
2003
Finished version of human genome sequence completed

The HGP ends with all goals achieved

To be continued...

PEAS COURTESY J. BLAMIRE, CITY UNIV. NEW YORK; WATSON & CRICK COURTESY A. BARRINGTON BROWN/SPL; *SCIENCE* COVERS COURTESY AAAS

Milestones In Genomic Sequencing



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

Milestones In Genomic Sequencing

We Sequenced our own Genome!

Milestone 1 2001

The Human Genome Project

Credit: Rawpixel Ltd / Alamy Stock Photo

Launched in 1990, the Human Genome Project set out to identify the order, that is, sequence, of all DNA bases to obtain the 'genetic blueprint' of humans. In 2001, two pivotal publications reported the first draft of the human genome, obtained by shotgun sequencing, setting the stage for the genomic era. The second phase of the project, which moved from the draft to an essentially finished reference genome, was completed in 2003. Read more.

By Caroline Barranco

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

Milestones In Genomic Sequencing

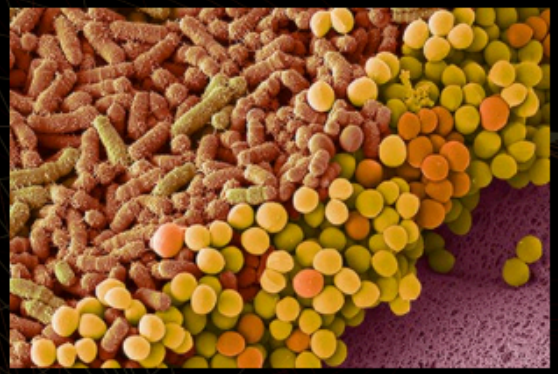
We sequenced what was previously impossible!

Milestone 2 2004

Sequencing the unculturable majority

Two key studies unlocked the field of metagenomics – the reconstruction of microbial communities from sequencing data – by providing approaches for unbiased, culture-independent analysis of DNA directly from environmental samples using sequencing technologies. Read more.

By Iain Dickson



Credit: Science Photo Library / Alamy Stock Photo

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

Milestones In Genomic Sequencing

We invented new sequencing technologies!

Milestone 3 2005

Sequencing – the next generation

Two revolutionary studies introduced high-throughput, massively parallel sequencing technologies able to sequence a bacterial genome at a fraction of the cost and time of traditional Sanger sequencing techniques. Read more.

By Joseph Willson

Credit: Zoran Obradovic / Alamy Stock Photo

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

Milestones In Genomic Sequencing

We understood the need and value of being inclusive!

Milestone 5 2008

The dawn of personal genomes

Credit: Redmond Durrell / Alamy Stock Photo

Two studies reported the genomes of an African individual and an Asian individual, respectively, using a new massively parallel sequencing approach based on reversible terminator dyes. Demonstrating the feasibility and resource value of human genome sequences, these studies and the technology they presented paved the way for population-scale genome sequencing. Read more.

By Darren Burgess

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

Milestones In Genomic Sequencing

Sequencing changed our understanding of disease

Milestone 6 2008

A sequencing revolution in cancer

Ley et al. presented the first whole-genome sequence of a cytogenetically normal acute myeloid leukaemia sample, showing that cancer genome sequencing can identify disease-associated mutations and druggable targets, offering promise for personalized medicine approaches. Read more.

By Safia Danovi

Credit: Sabena Jane Blackbird / Alamy Stock Photo

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

Milestones In Genomic Sequencing

Sequencing changed our understanding of Gene expression and Gene models

Milestone 7 2008

Transcriptomes – a new layer of complexity

A series of milestone publications reported the development of high-throughput sequencing of whole transcriptomes, known as RNA sequencing (RNA-seq), across different species. Read more.

By Margot Brandt

Credit: Y H Lim / Alamy Stock Photo

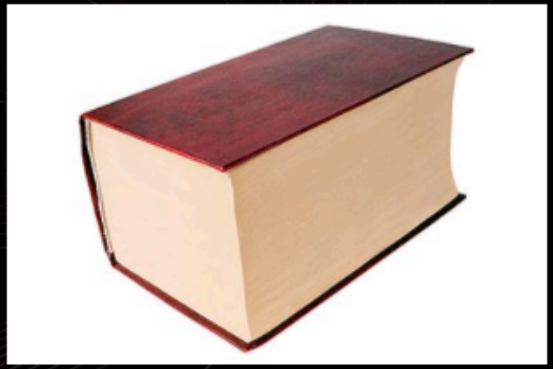
https://www.nature.com/immersive/d42859-020-00099-0/index.html

Milestones In Genomic Sequencing

We invented even better sequencing technologies!

Milestone 8 2009

Long reads become a reality



Credit: Zoonar GmbH / Alamy Stock Photo

Long-read sequencing technologies began to shed light on hidden parts of the human genome by sealing gaps in existing assemblies, allowing modified bases to be detected on native DNA or RNA, and revealing the complexity of the transcriptome. Read more.

By Ivanka Kamenova

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

Milestones In Genomic Sequencing

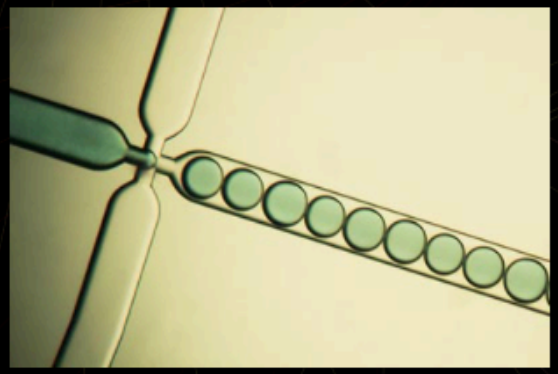
We developed single cell Genomics!

Milestone 11 2009

Sequencing one cell at a time

Moving from genomic analysis of tissues or cells in bulk to performing single-cell sequencing provided a whole new perspective on gene regulation, cell-to-cell heterogeneity and developmental or disease processes. The difficulty of performing analyses at such resolution required many experimental and computational innovations. Read more.

By Aline Lückgen



Credit: Greenshoots Communications / Alamy Stock Photo

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

Milestones In Genomic Sequencing

We started sequencing the genomes of our ancestors!

Milestone 12 2010

Waking the dead: sequencing archaic hominin genomes

Credit: The Natural History Museum / Alamy Stock Photo

The publication of the first draft genome of a Neanderthal in 2010 marked a turning point for the palaeogenomics field, making it possible to assemble an ancient genome from next-generation sequencing reads by overcoming previous limitations in ancient DNA research such as limited starting material, contamination and degradation. Read more.

By Rebecca Furlong

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

Milestones In Genomic Sequencing

We re-defined the meaning of reference genomes!

Milestone 15 2014

Pan-genomes: moving beyond the reference

Pan-genome studies in a variety of species – from microorganisms to plants to humans – have shown that a large amount of genetic variation can be found in the dispensable genome. This observation has called into question our reliance on single reference genomes for assembling and analysing genomes. Read more.

By Dominique Morneau

Credit: Art of Food / Alamy Stock Photo

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

Milestones In Genomic Sequencing

We learned how to sequence a chromosome with no gaps!

Milestone 17 2020

Filling in the gaps telomere to telomere

2020 saw the publication of the first gapless, telomere-to-telomere assembly of a human chromosome, the X chromosome. This discovery brought together sequencing technologies and computational tools that had been developed in the preceding decade. Read more.

By Katharine Wrighton

Credit: PORNCHAI SODA / Alamy Stock Photo

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

The Result was…

2009

An explosion of computational tools

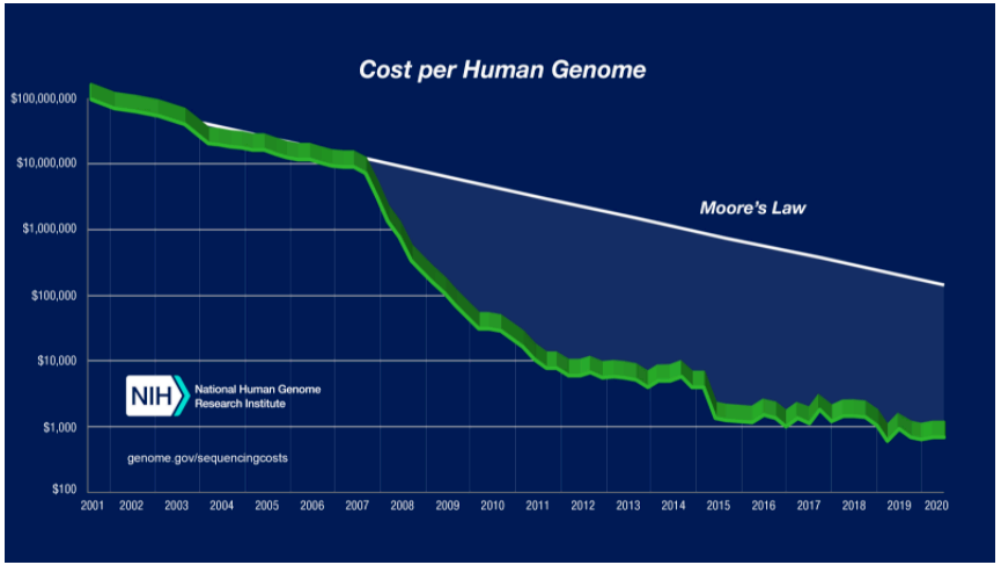
As genome sequencing became more affordable and widespread, its applications rapidly expanded, driving the development of new computational tools to accommodate the requirements of transcriptomics, metagenomics or genetic variant discovery. Read mapping tools such as Bowtie and BWA or the splice-aware aligner TopHat were able to align millions of short reads to the reference genome, and downstream analysis software, such as SAMtools and BreakDancer, facilitated the detection of genetic variants.

Related article: Repetitive DNA and next-generation sequencing: computational challenges and solutions

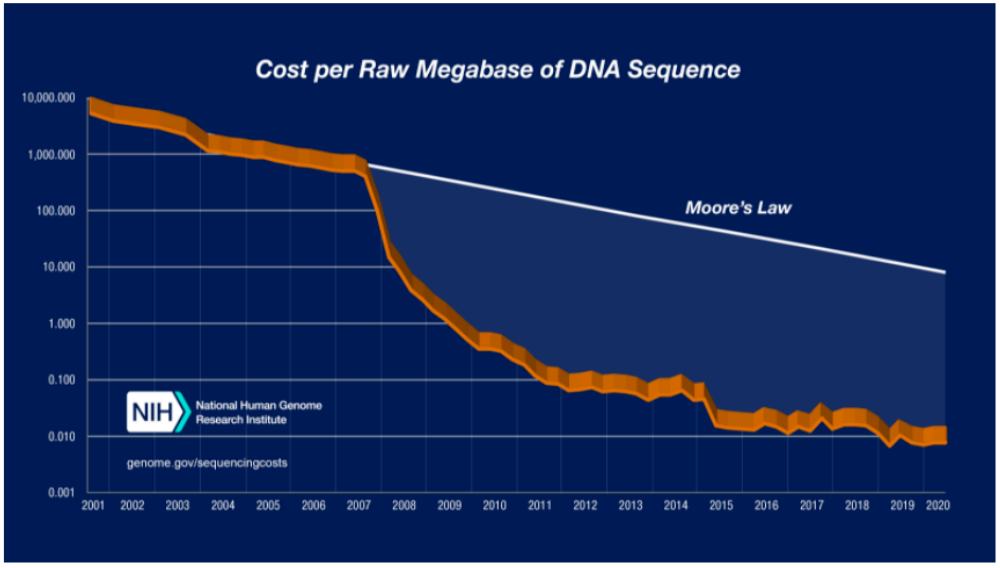
Credit: michalz86 / Alamy Stock Photo

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

The Price of Genomic Sequencing is Dropping…



Cost per genome data - August 2020



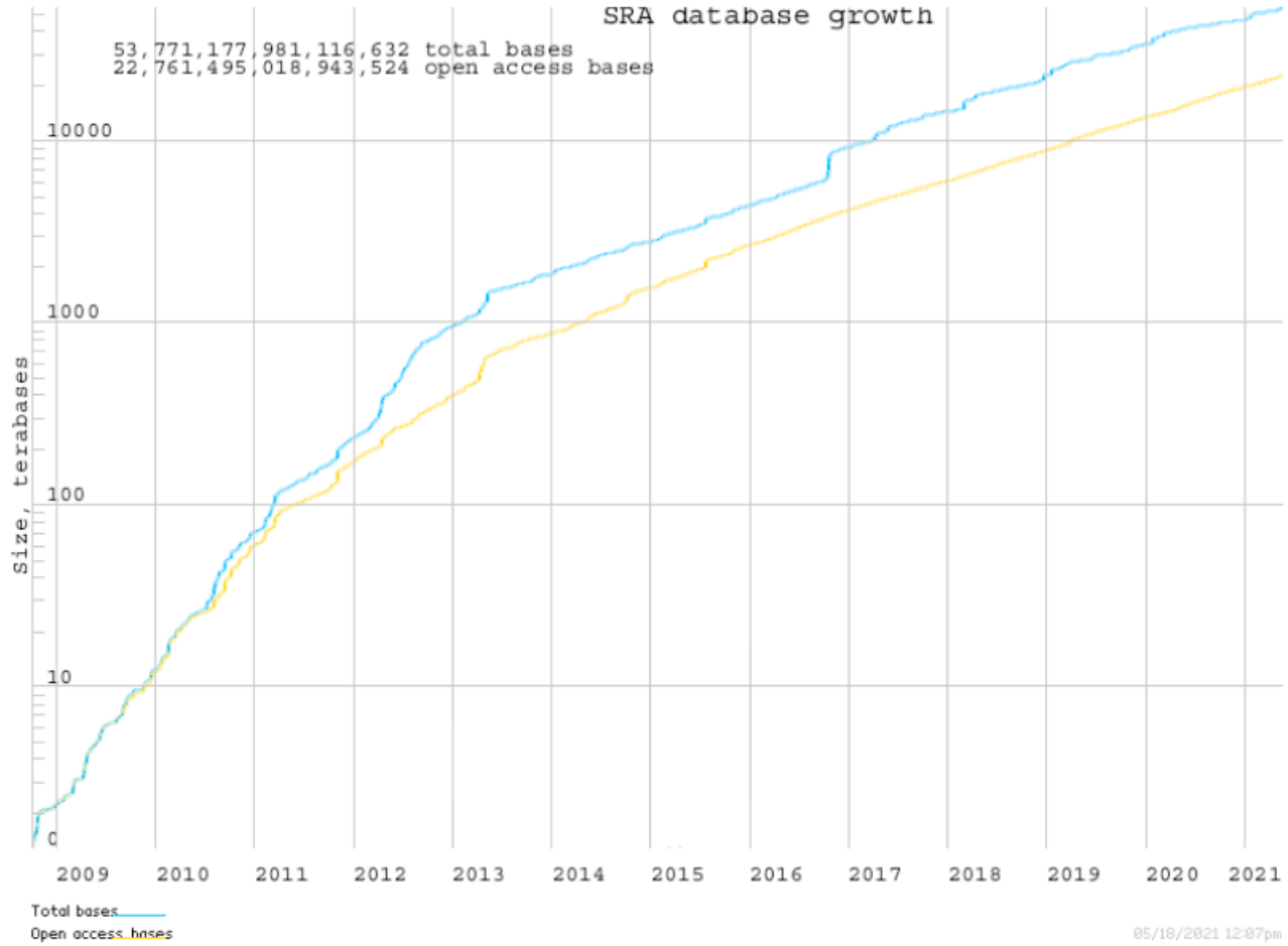
Sequencing cost per megabase - August 2020

Genomic Sequencing Is Generating a Problem

The Volume of Incoming Data is Overwhelming

The Problem

Databases are Growing at an Amazing Rate



The Problem

Databases are Growing at an Amazing Rate

Homo sapiens

Database	Access		all
	public	controlled	
BioSample	3,091,653	651,138	3,742,791
BioProject	32,639	825	33,464
dbGaP		18	18
GEO Datasets	806,746		806,746

As of: 2021-05-18

https://www.ncbi.nlm.nih.gov/sra/?term=Homo+sapiens

The Problem

The Reproducibility Versus Replicability Crisis

Reproducibility of Scientific Results

First published Mon Dec 3, 2018

The terms "reproducibility crisis" and "replication crisis" gained currency in conversation and in print over the last decade (e.g., Pashler & Wagenmakers 2012), as disappointing results emerged from large scale reproducibility projects in various medical, life and behavioural sciences (e.g., Open Science Collaboration, OSC 2015). In 2016, a poll conducted by the journal *Nature* reported that more than half (52%) of scientists surveyed believed science was facing a "replication crisis" (Baker 2016). More recently, some authors have moved to more positive terms for describing this episode in science; for example, Vazire (2018) refers instead to a "credibility revolution" highlighting the improved methods and open science practices it has motivated.

The crisis often refers collectively to at least the following things:

a. the virtual absence of replication studies in the published literature in many scientific fields (e.g., Makel, Plucker, & Hegarty 2012),

b. widespread failure to reproduce results of published studies in large systematic replication projects (e.g., OSC 2015; Begley & Ellis 2012),

c. evidence of publication bias (Fanelli 2010a),

d. a high prevalence of "questionable research practices", which inflate the rate of false positives in the literature (Simmons, Nelson, & Simonsohn 2011; John, Loewenstein, & Prelec 2012; Agnoli et al. 2017; Fraser et al. 2018), and

e. the documented lack of transparency and completeness in the reporting of methods, data and analysis in scientific publication (Bakker & Wicherts 2011; Nuijten et al. 2016).

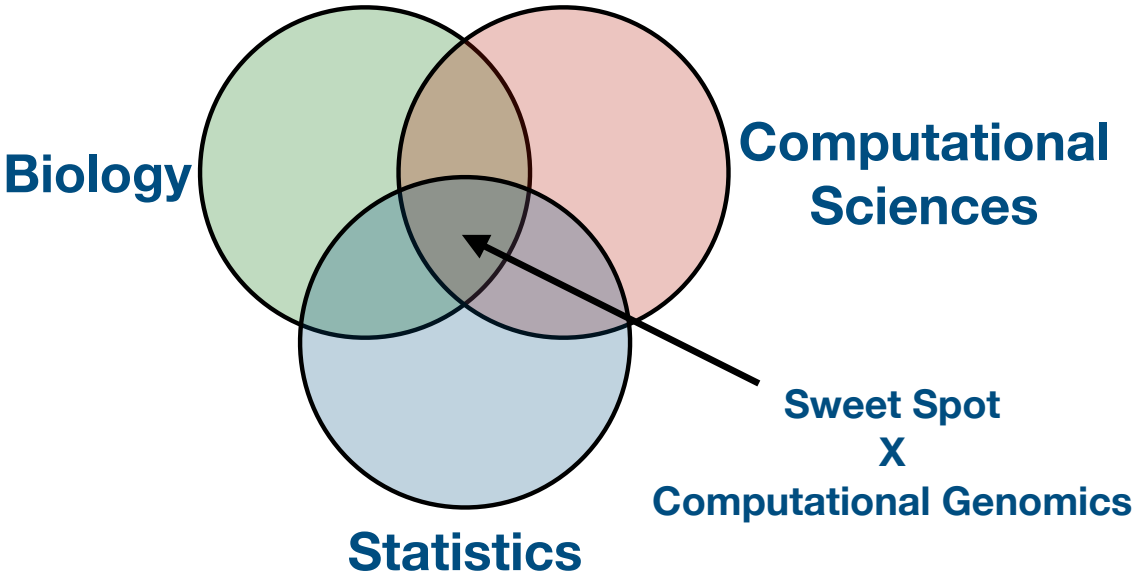
https://plato.stanford.edu/entries/scientific-reproducibility/

The Problem

There is a Divide Between Biology, Computational Sciences and Statistics

And

Biology is developing at an amazing rate, becoming increasingly interdisciplinary



Question

How do you train students for the future, knowing what we know about the past?

One Answer…

You teach them to Think!…
Inter-disciplinarily

How do we do that?

Where do we start?

A Solution…

To Gap The Divide Between

Biology, Computer Sciences and Statistics

We should start by Teaching

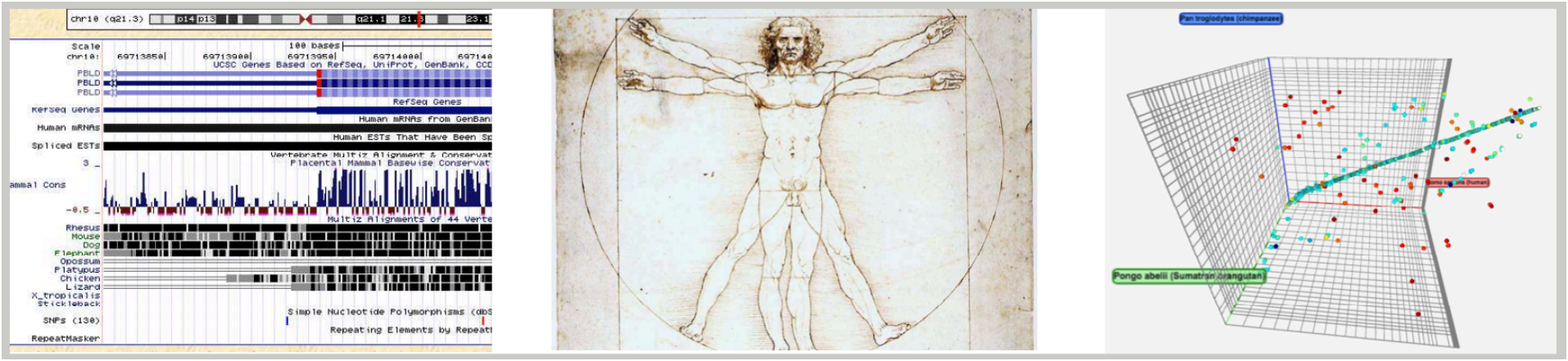
Hypothesis-driven

Computational Genomics Courses

A Solution…

- Computational Genomics gaps the divide between Biology, Statistics and Computational Sciences
- Computational experiments can and should be designed like any other wet-lab experiment
- The belief that data processing is separated, or different from, wet-lab experiments is biased, damaging and wrong
- The rapid evolution and availability of new algorithms, and computational tools make previous data analyses potentially obsolete
- This generates the need to re-test and re-analyze already produced genome data, and provide us with a unique opportunity for training

A Solution…



BIOL350
COMPUTATIONAL GENOMICS

← **Undergraduate Level**
Started 2010

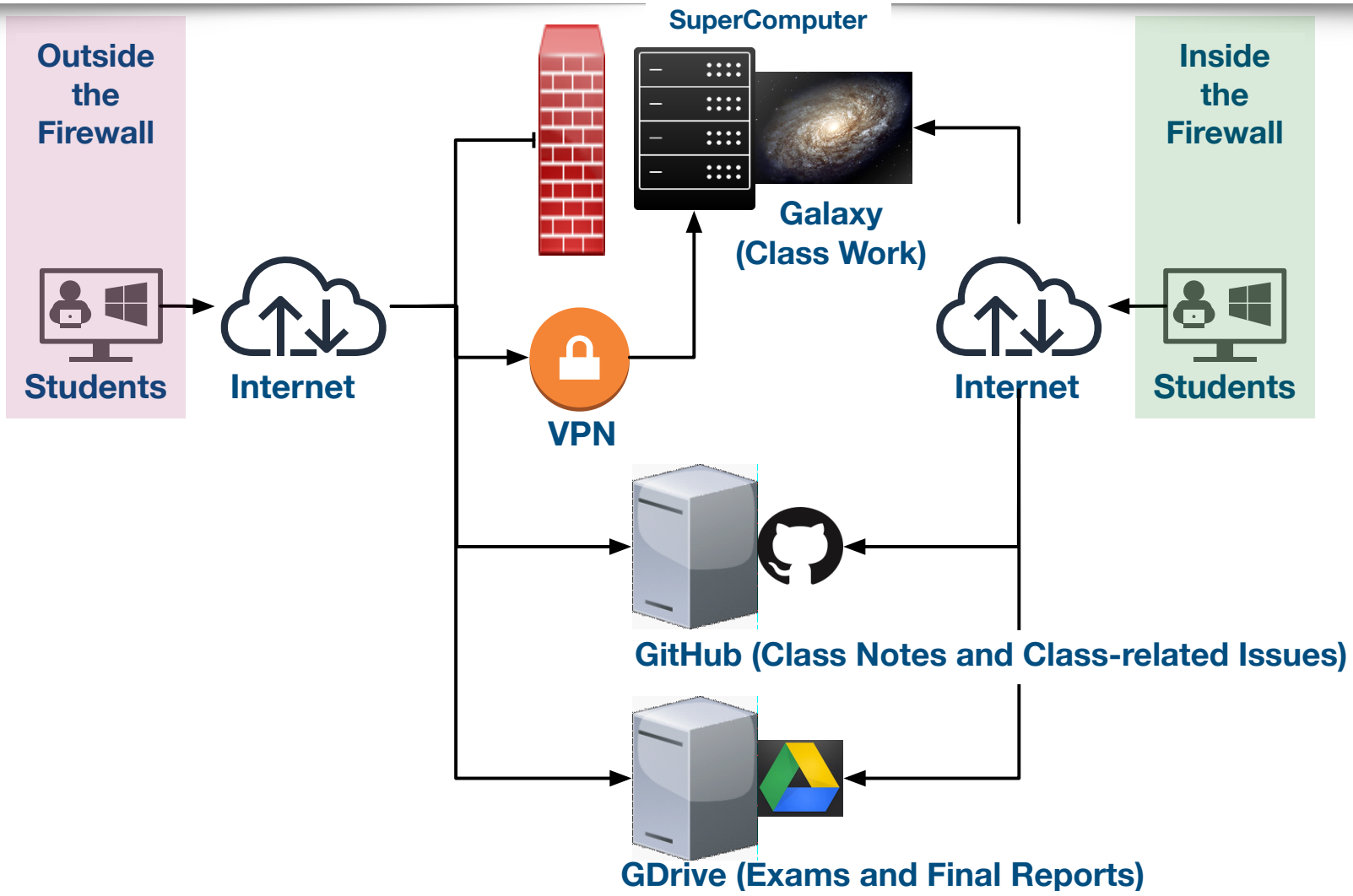
BIOL650
GENOMICS

← **Graduate Level**
Started 2000

BIOL647
DIGITAL BIOLOGY

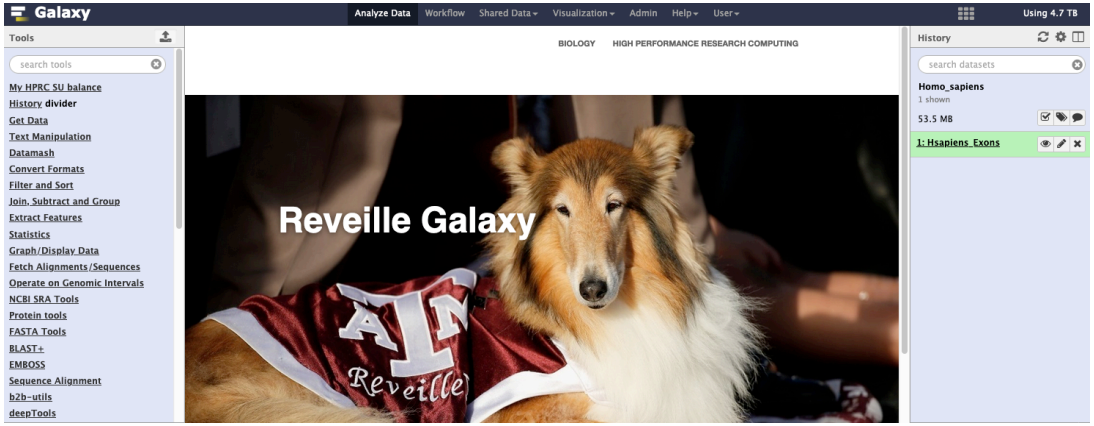
← **Graduate Level**
Started 2012

About Computational Genomics (BIOL350) and Genomics (BIOL650)

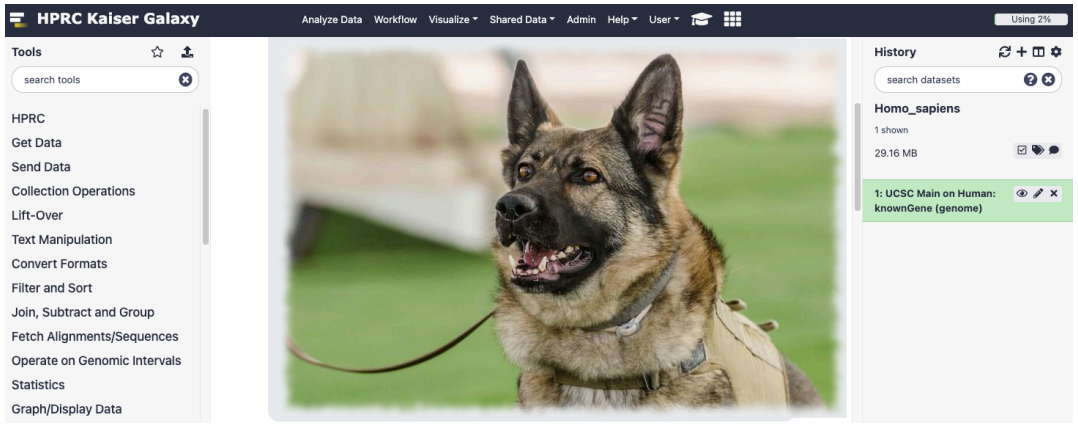


About Computational Genomics (BIOL350) and Genomics (BIOL650)

Galaxy → Web-Based Computational Tools



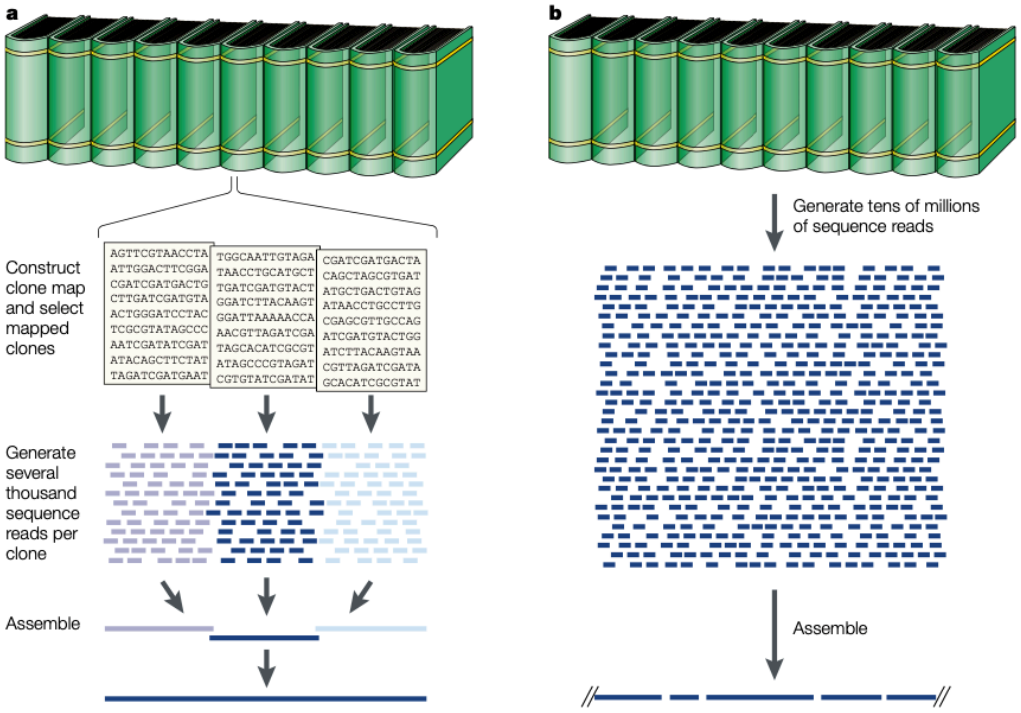
← Reveille Deployed 2016 Ada

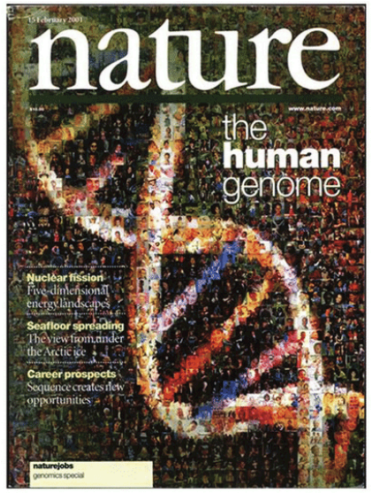


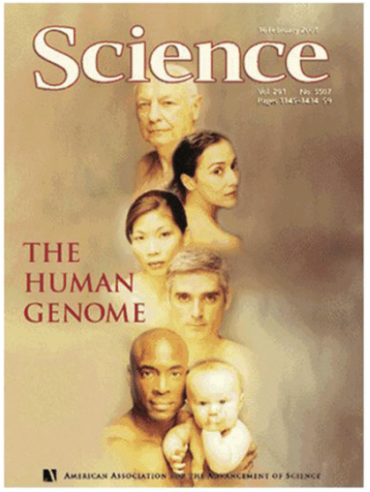
← Kaiser Deployed 2020 Terra

About Computational Genomics (BIOL350) and Genomics (BIOL650)

History of Genomics



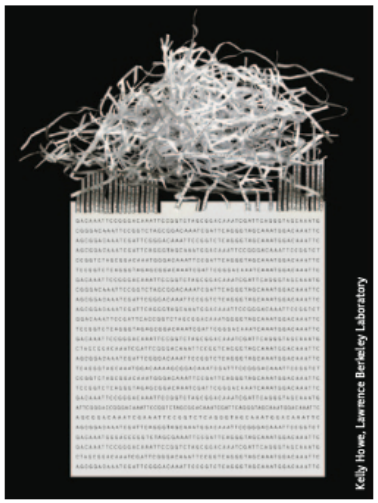


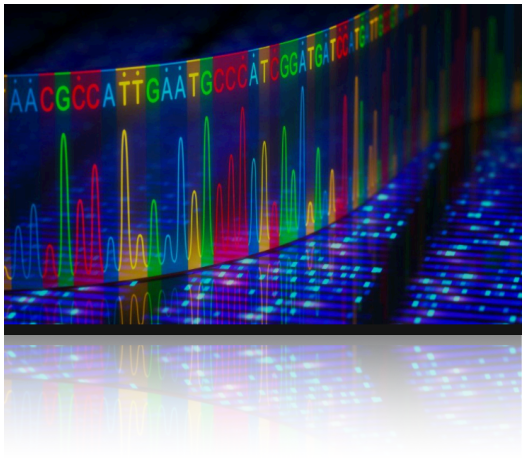


https://www.nature.com/articles/35084503.pdf

About Computational Genomics (BIOL350) and Genomics (BIOL650)

Library Construction → **Sequencing** → **Assembly**





- Sequence reads:

```
CAGCTGTCCCAGATGAC  AACTTCCCTCCCAGCT  TCCGCCTTCAGCTCAAGACTTAACTTC
CGGCCTTTGGGCTCC  ACTTAACTTCCCTCCCAGCTGTCC  TCCCAGCTGTCCCAGATGACGCCATC
 CAGATGACGCC  CGGCCTTTGGGCTCCGCCTTCAGCTCAAGA  GGGCTCCGCCTTCAGCTC
```

- Match up overlaps:

```
CGGCCTTTGGGCTCCGCCTTCAGCTCAAGA    AACTTCCCTCCCAGCT    CAGATGACGCC
          TCCGCCTTCAGCTCAAGACTTAACTTC  TCCCAGCTGTCCCAGATGACGCCATC
       GGGCTCCGCCTTCAGCTC   ACTTAACTTCCCTCCCAGCTGTCC
CGGCCTTTGGGCTCC                          CAGCTGTCCCAGATGAC
```

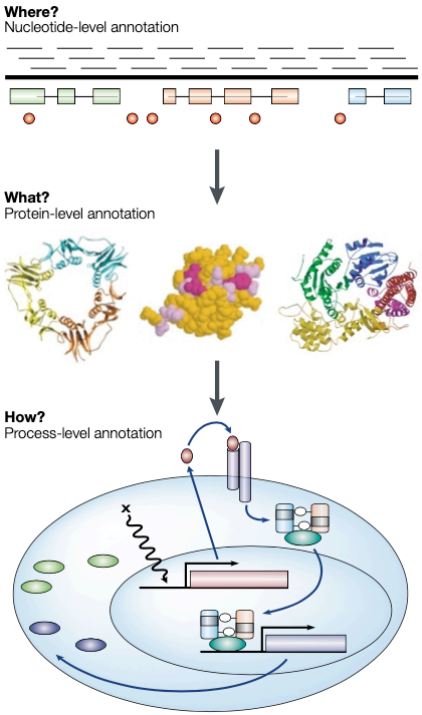
- Contig:

```
CGGCCTTTGGGCTCCGCCTTCAGCTCAAGACTTAACTTCCCTCCCAGCTGTCCCAGATGACGCCATC
```

http://training.ensembl.org/events/2021/2021-05-18-OpenVirtualBrowser_May

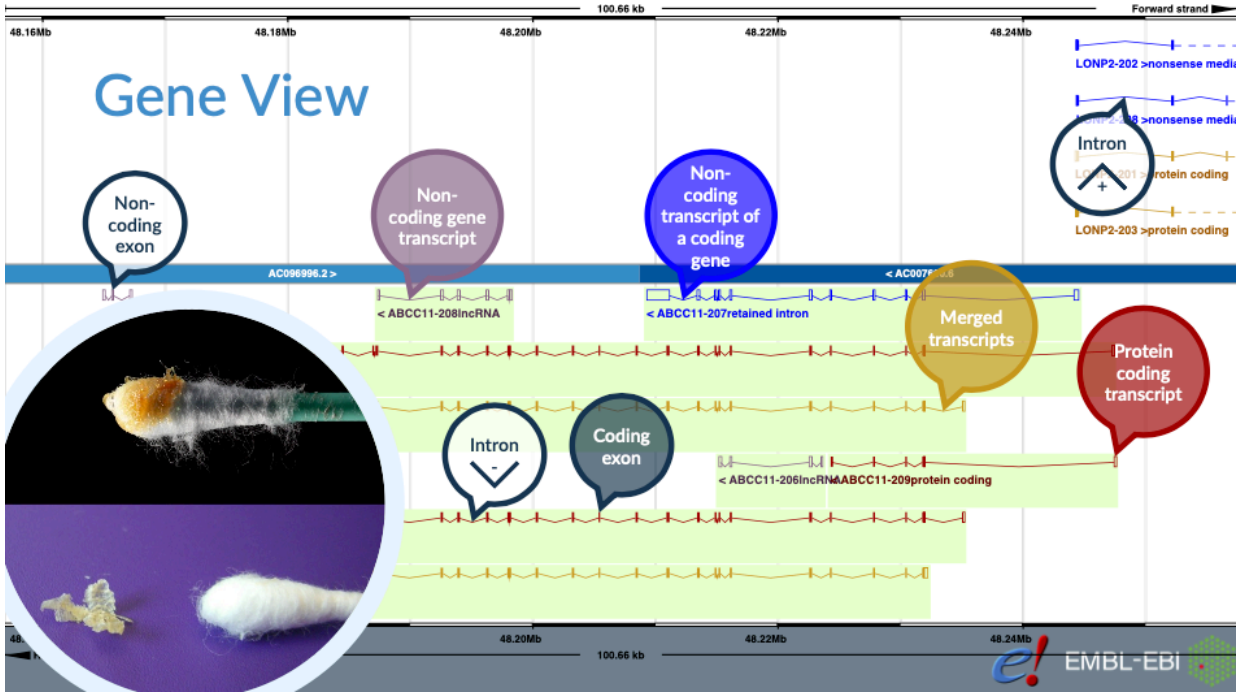
About Computational Genomics (BIOL350) and Genomics (BIOL650)

Genome Annotation



→

Gene Models



http://training.ensembl.org/events/2021/2021-05-18-OpenVirtualBrowser_May

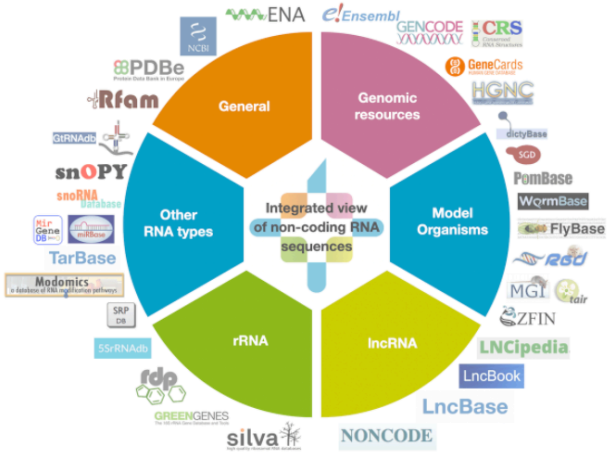
About Computational Genomics (BIOL350) and Genomics (BIOL650)

Databases

UniProt Pfam Rfam

InterPro

Classification of protein families



Data Retrieval and Analysis

How do I use BioMart?

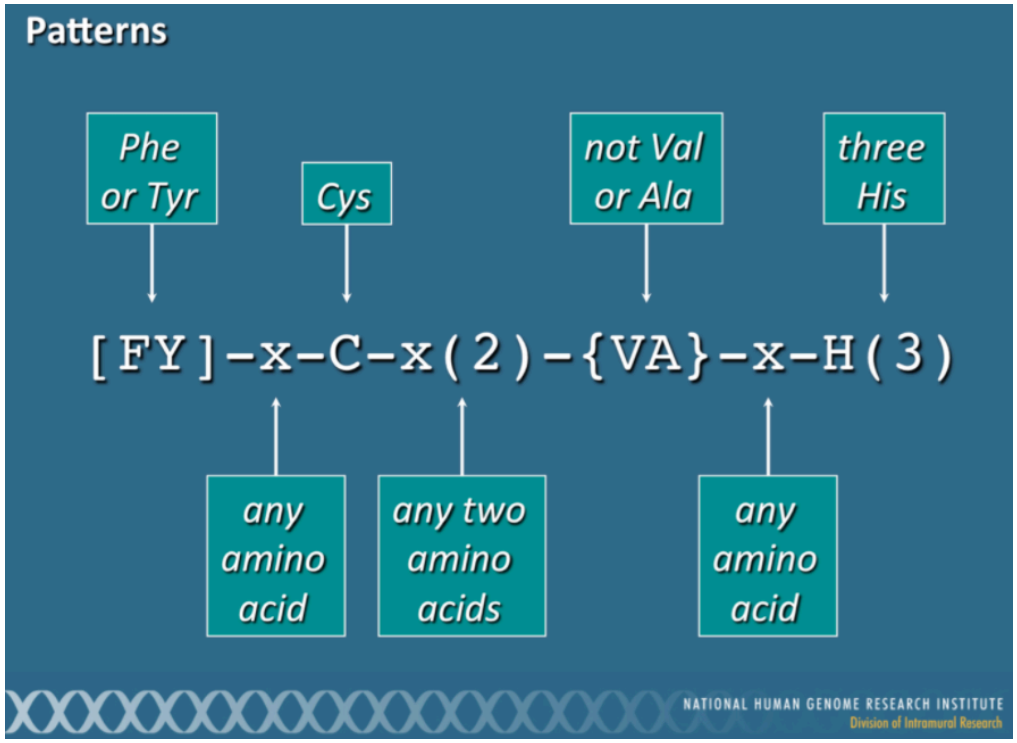
The four steps

Step 1 Dataset	Step 2 Filters	Step 3 Attributes	Step 4 Results
Choose database and species	Narrow down the dataset	Specify your output	View and export

http://training.ensembl.org/events/2021/2021-05-18-OpenVirtualBrowser_May

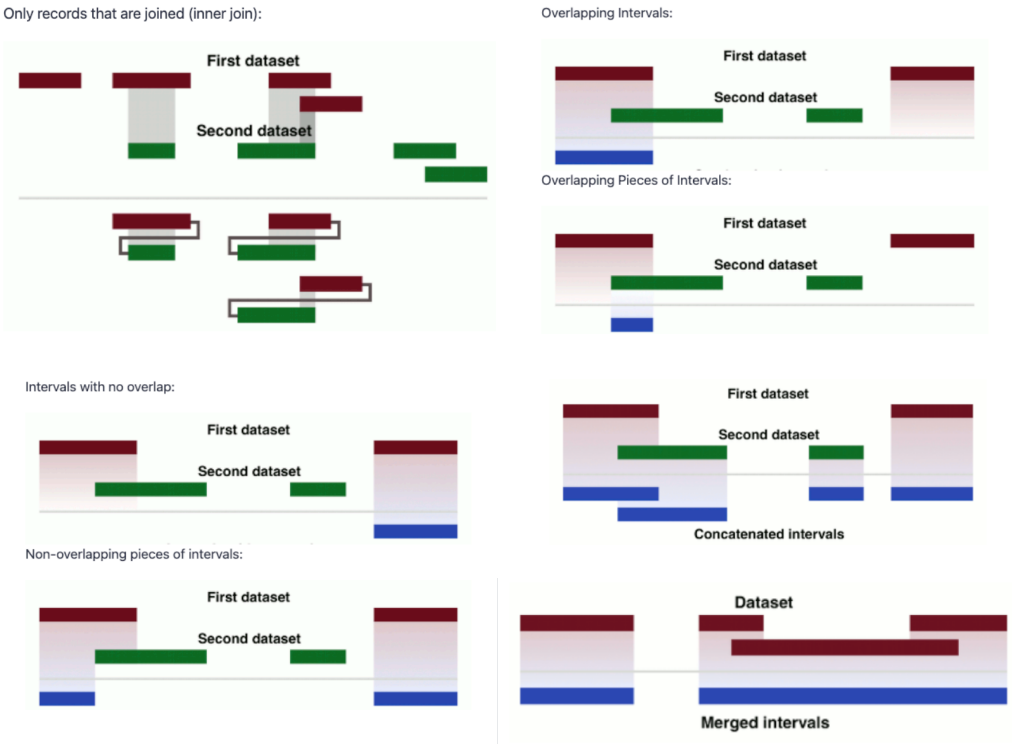
About Computational Genomics (BIOL350) and Genomics (BIOL650)

Patterns as Formulas



→

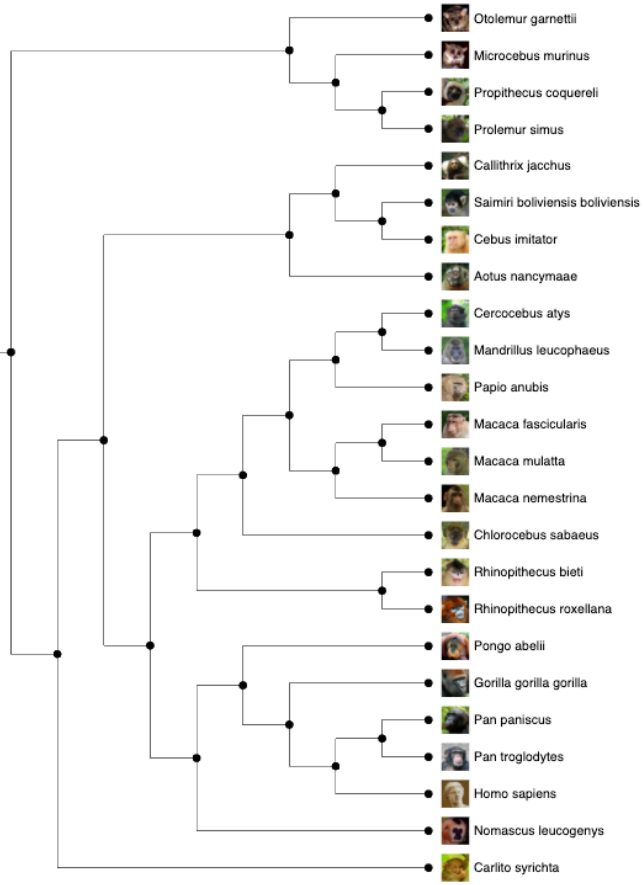
Computational Arithmetics



About Computational Genomics (BIOL350) and Genomics (BIOL650)

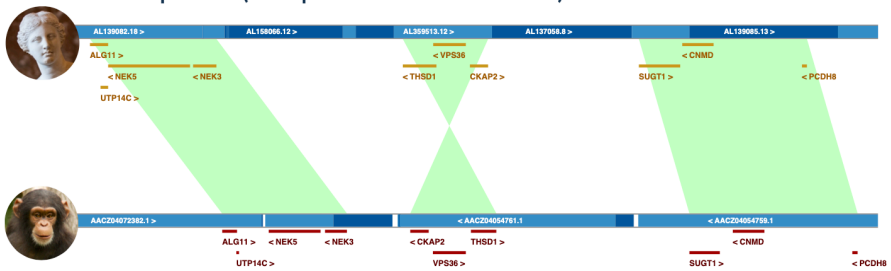
Comparative Genomics

By Taxa



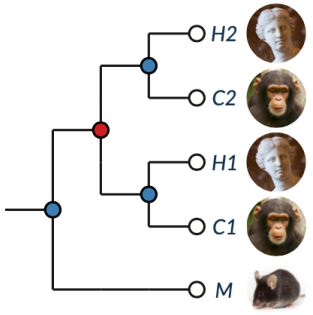
Shared Syntheny

Conserved order of aligned homologoues genomic blocks between species (irrespective of orientation):

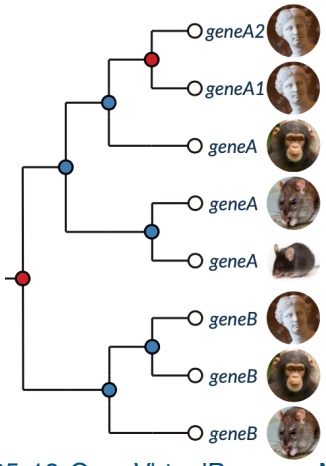


Homology Relationships

- Orthologues
  - Genes emerged through a speciation event, e.g. *C1* and *H1*; *C2* and *M*; *H2* and *M*
  - 1-to-1: *C1* and *H1*
  - 1-to-many: *M* and *H1*, *H2*
- Paralogues
  - Genes emerged through a duplication event, e.g. *C1* and *C2*, *H1* and *H2*



Gene Gain/Loss Tree

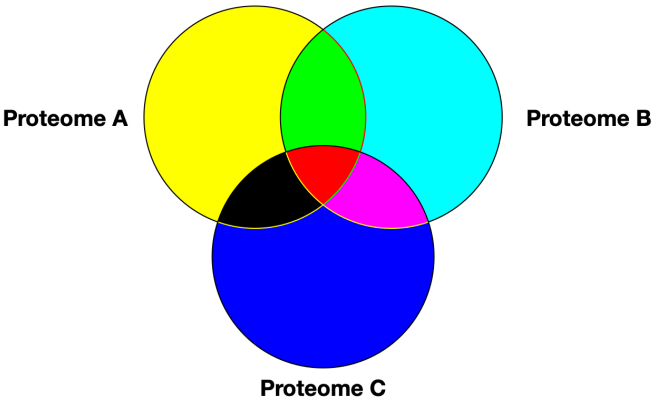


http://training.ensembl.org/events/2021/2021-05-18-OpenVirtualBrowser_May

About Computational Genomics (BIOL350) and Genomics (BIOL650)

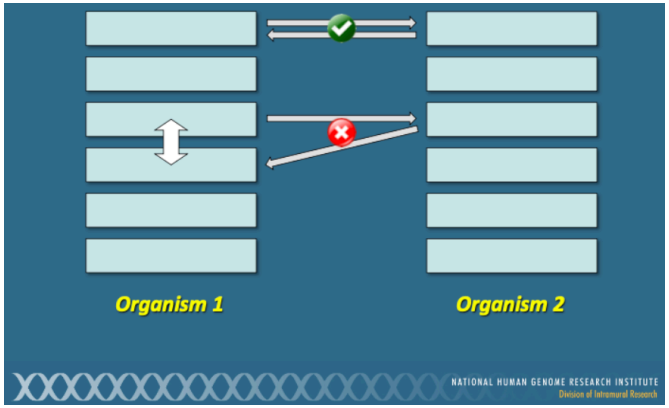
Final Project Hypothesis:
The number of conserved proteins is directly proportional to the evolutionary distance between the proteomes tested

Calculate Intersections Between three Different Proteomes

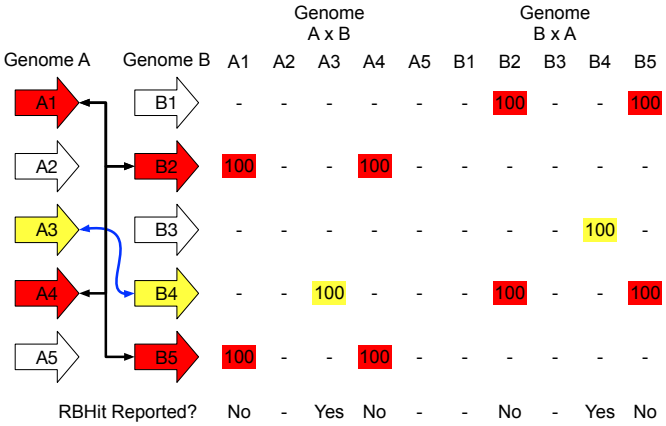


The Final Project workhorse is the Reverse-Blast-Hit (RBH) Algorithm

Basic Idea



Advanced Algorithm



Undergraduate Level Understanding

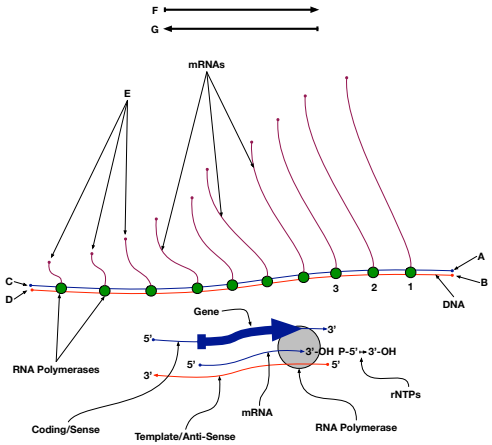
Graduate Level Understanding

All students get the same (control) proteome and then each student gets two different unique proteomes

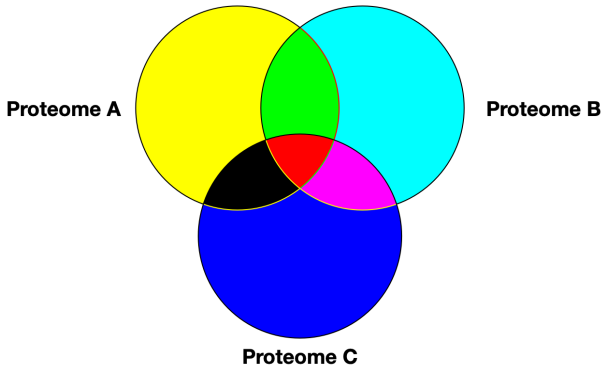
About Computational Genomics (BIOL350) and Genomics (BIOL650)

Key Concepts

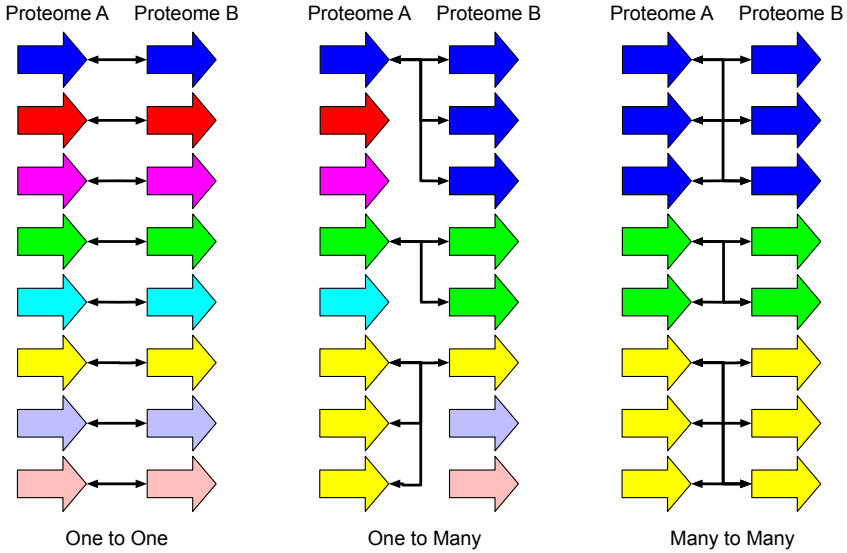
Gene Polarity



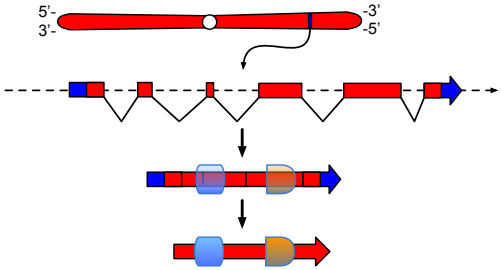
Intersections AND/OR/NOT



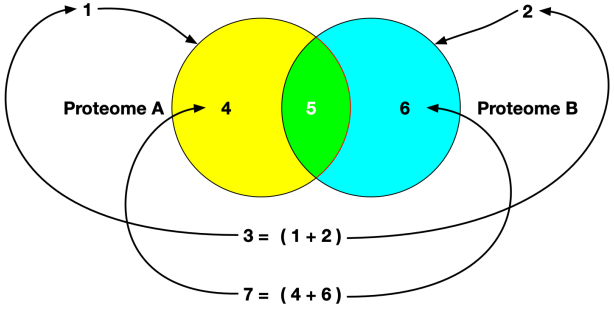
Databases Relationships



Gene Models



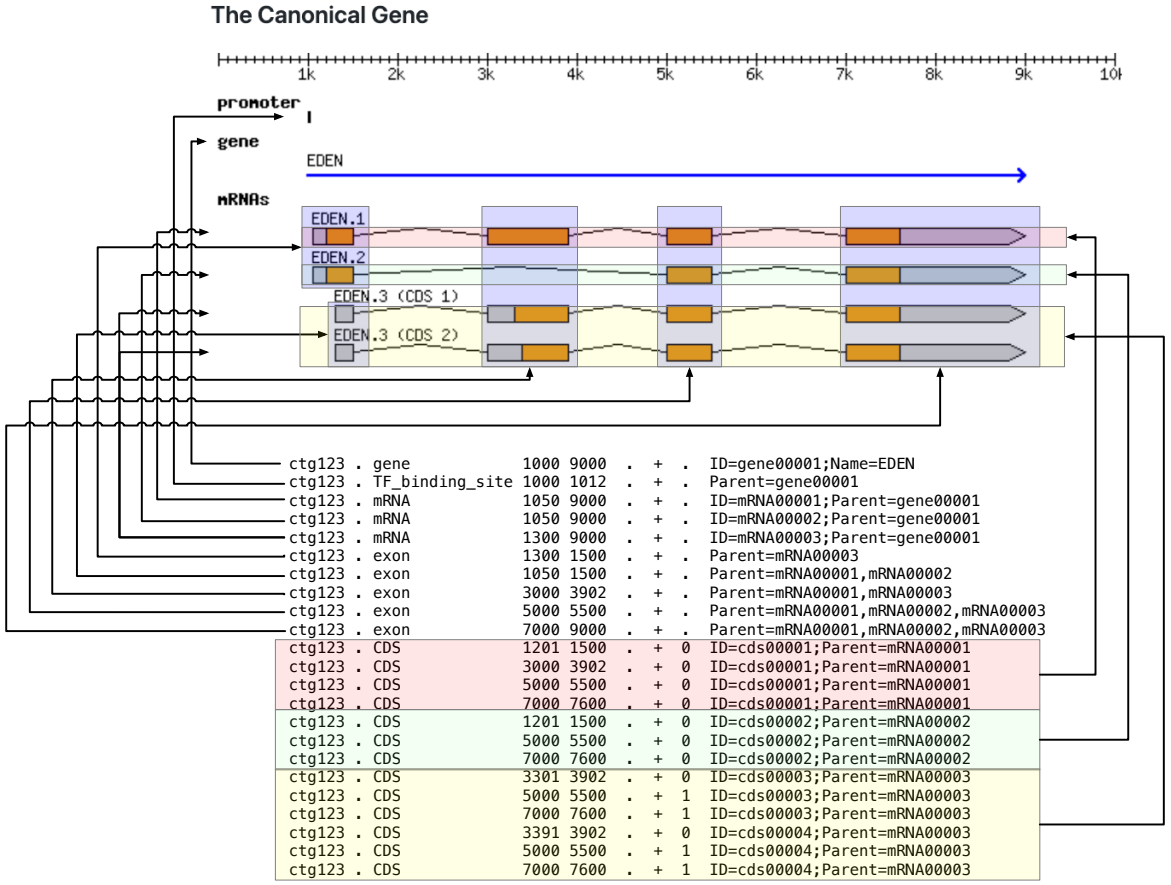
Datasets Combinations



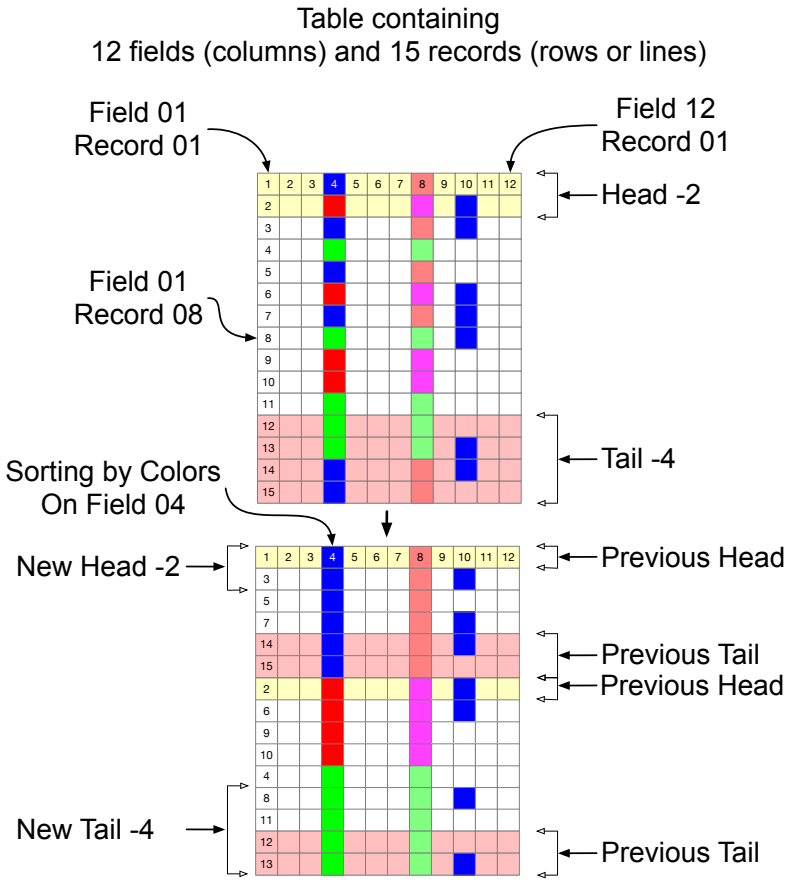
About Computational Genomics (BIOL350) and Genomics (BIOL650)

Key Concepts

Relationships: Tables to Graphs



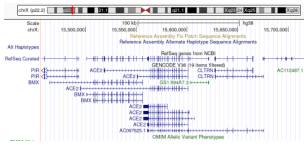
Tables and Text Manipulations



About Computational Genomics (BIOL350) and Genomics (BIOL650)

Google Drive
Or
GITHUB

Databases



Download Question(s)

Upload Answers(s)

Send Queries

Receive Data

Process Data

Send

Receive

Student A

Send

Receive

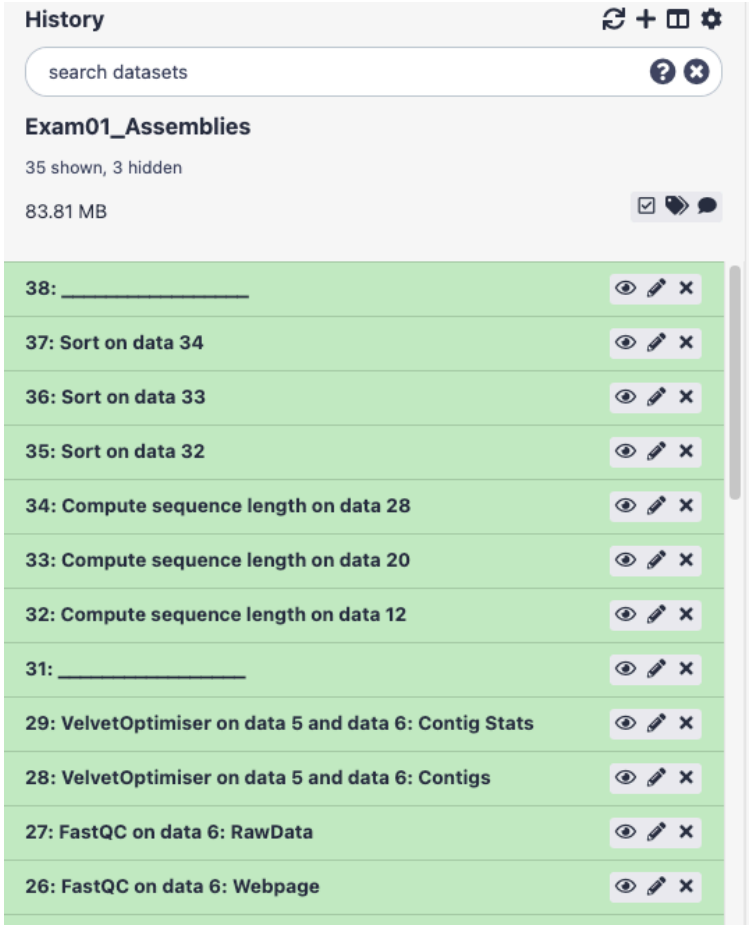
SuperComputer
Galaxy

Text Messaging

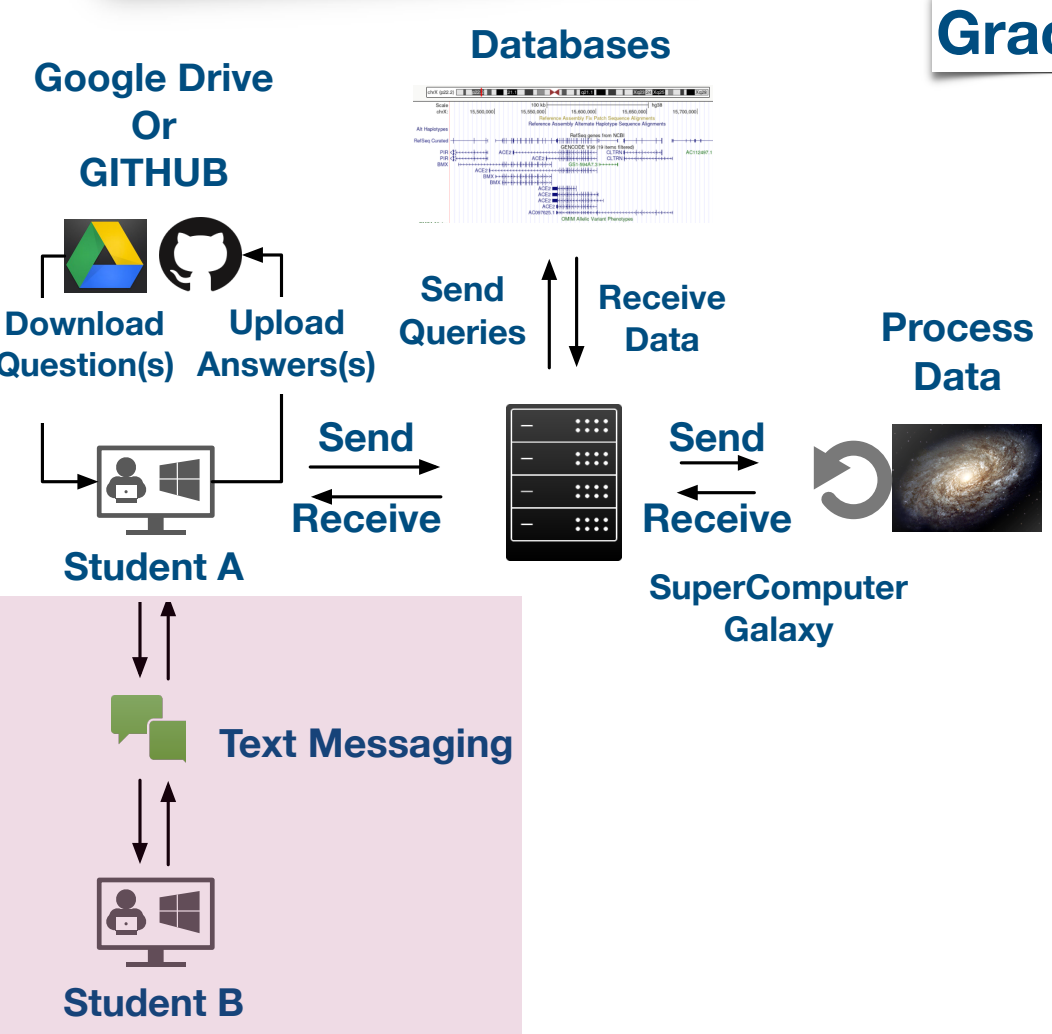
Student B

Grading

Individual Processes



About Computational Genomics (BIOL350) and Genomics (BIOL650)



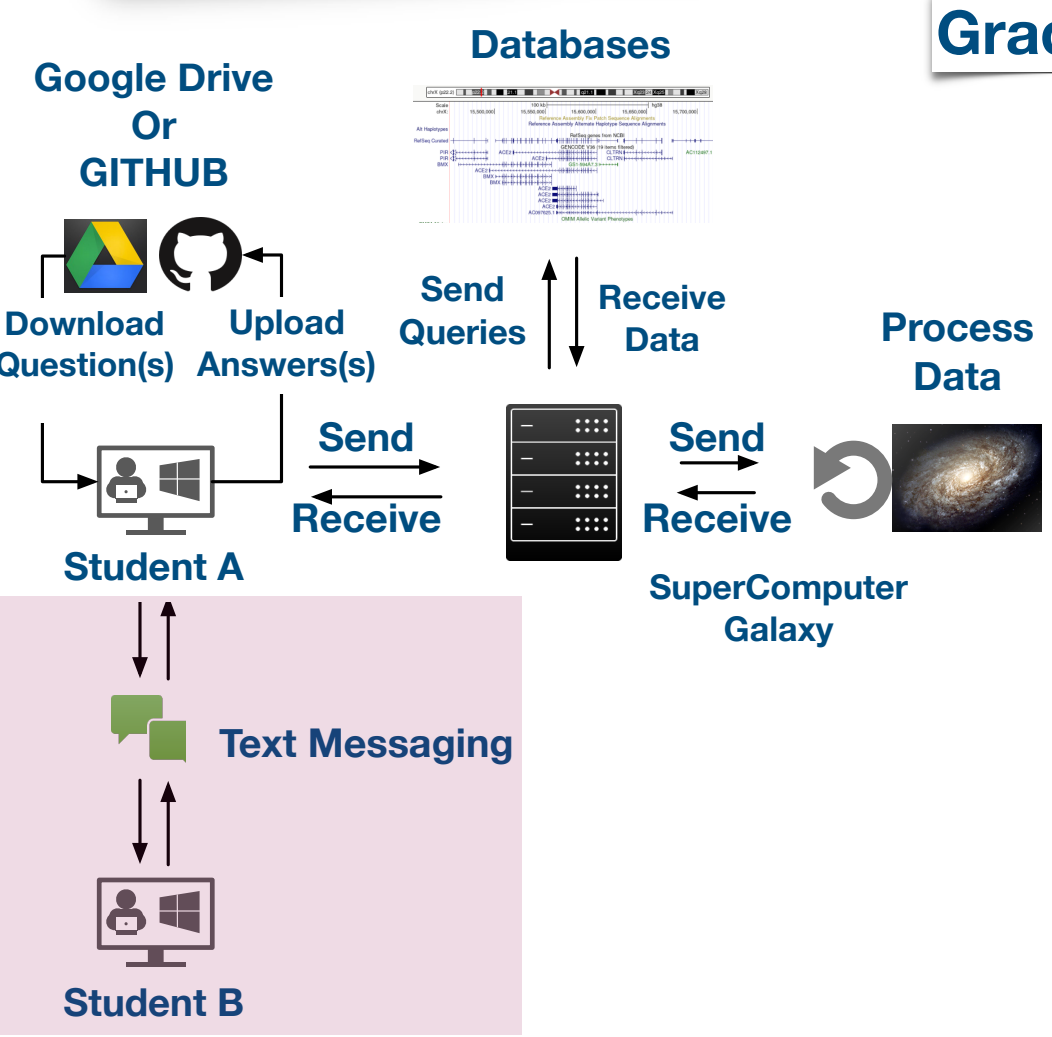
Grading

Grading Galaxy Work

For a given question:

- Is the answer reported correct?
- Does the corresponding Galaxy history exists?
- Does the corresponding Galaxy history contains the corresponding individual processes?
- Are those processes correct?

About Computational Genomics (BIOL350) and Genomics (BIOL650)



Grading

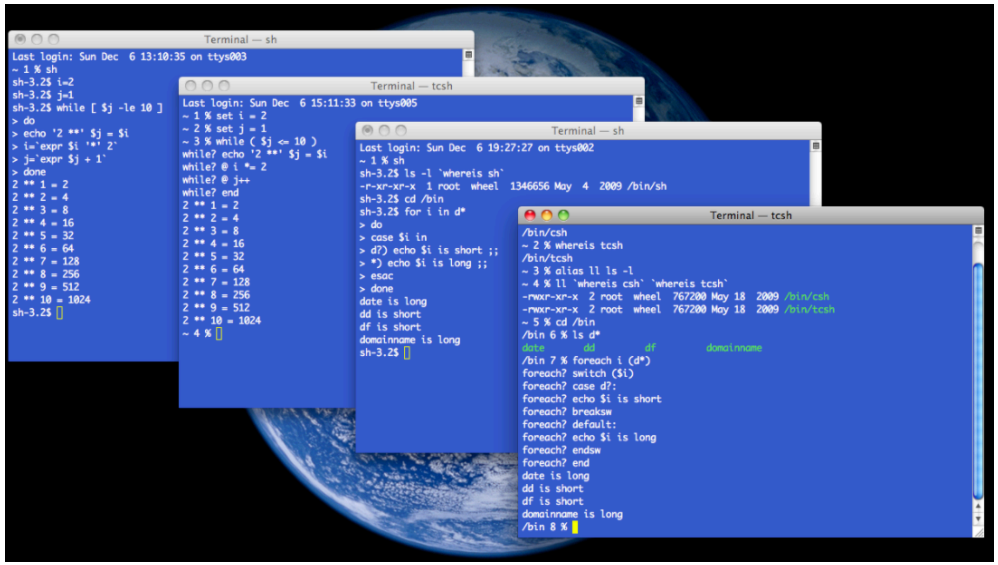
Grading Galaxy Work

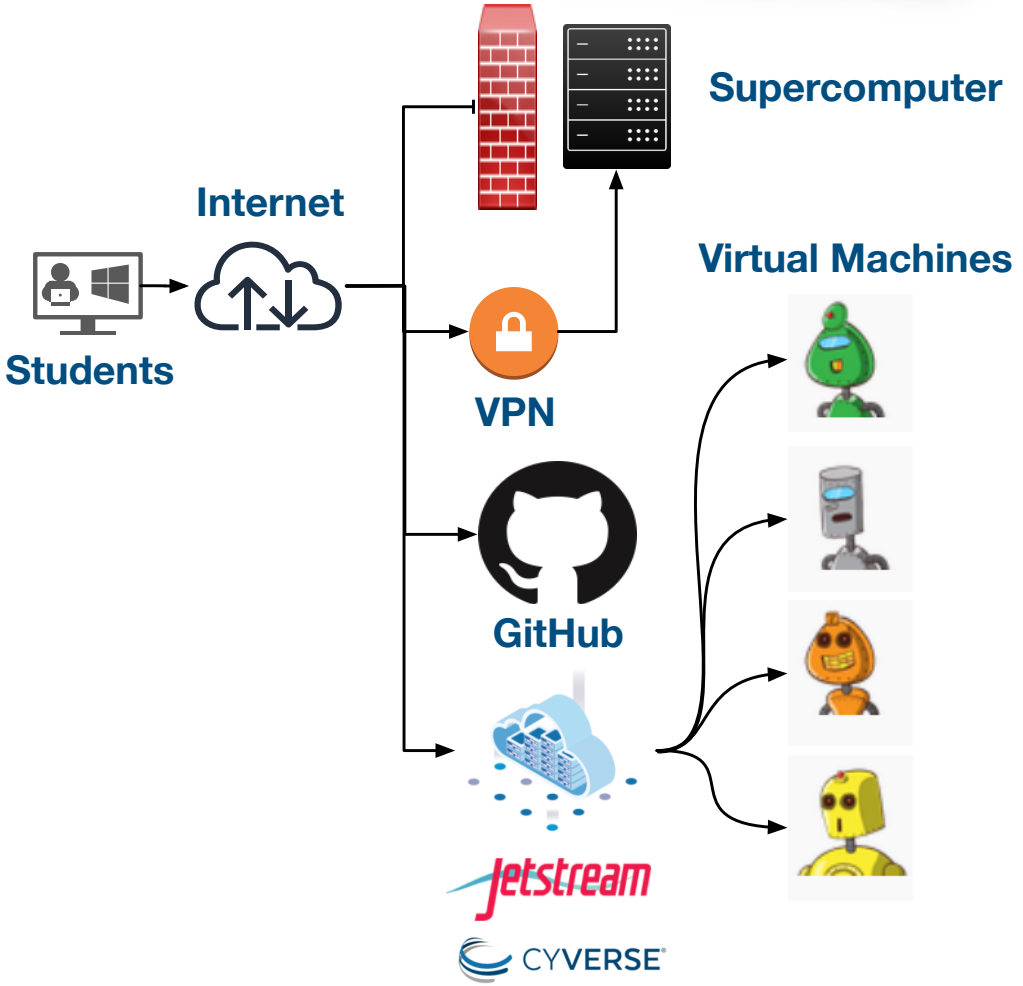
Number of Students	Number of Questions	Number of Processes/ Question	Number of Processes to Grade	Times Six Exams/ Semester
1	1	5	5	30
50	1	5	250	1500
50	10	5	2500	15000
65	10	5	3250	19500

- Grading is Hard because there are different ways to solve the same problem
- Scripting grading is possible but requires files to have consistent names! (…which requires students to name files correctly, which assumes that all students will follow instructions…)

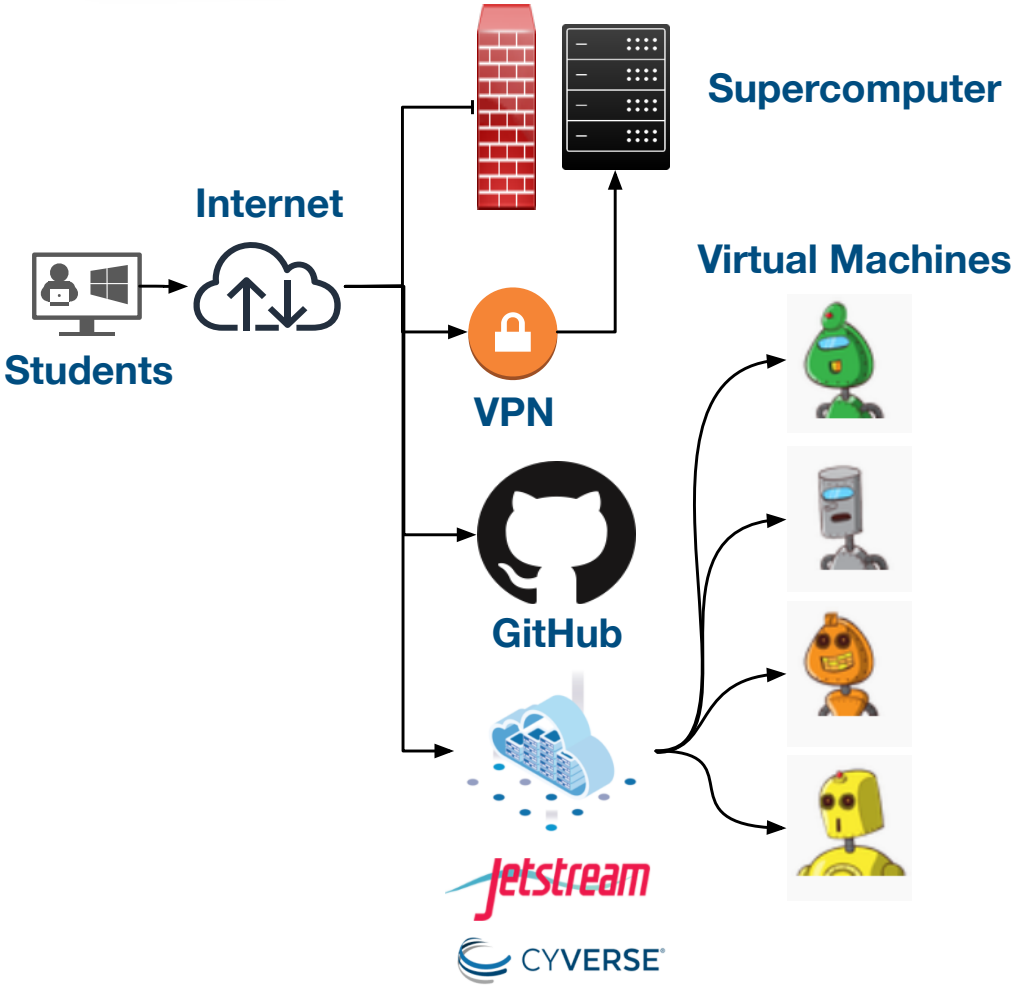
About Digital Biology (BIOL647)

Command-line Driven





About Digital Biology (BIOL647)



Module 01	Module 02	Module 03
Virtual Machines (Jetstream)		Supercomputer (Ada / Grace)
Master the Command-line		Introduction to Supercomputers
Basic GNU/Linux/Unix Introduction	Basic Scripting	Advanced Scripting

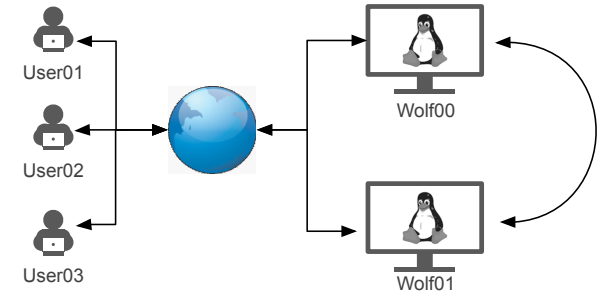
Jetstream work is supported by XSEDE
Grant Nº BIO210035
Funded by NSF ACI-1445604, Awarded to XSEDE

About Digital Biology (BIOL647)

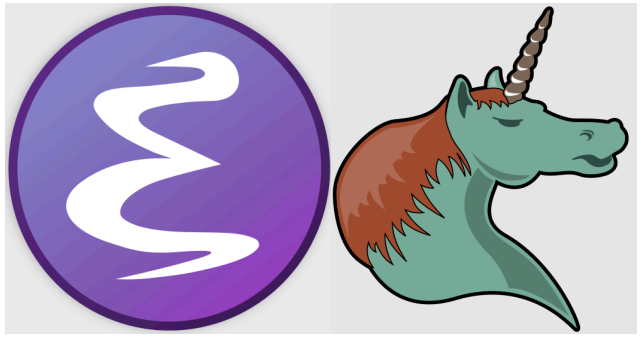
Module 01

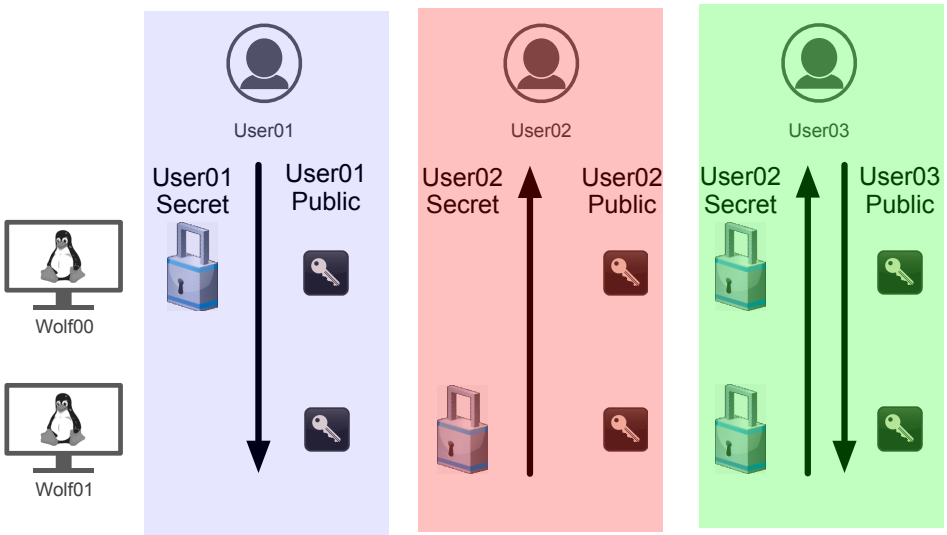
GIT/GITHUB and Version Control

SSH and Data Transfer



Emacs/Org-Mode





About Digital Biology (BIOL647)

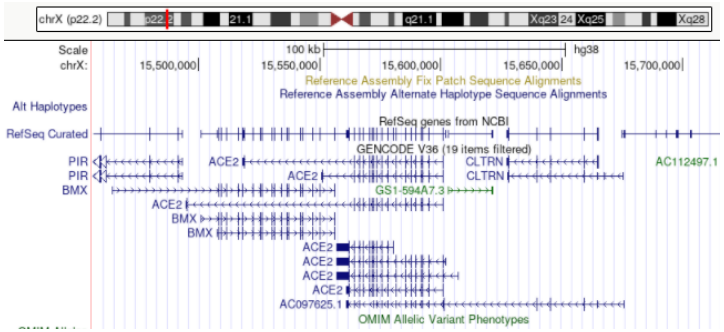
Module 02

Unix101	Comprehensive Introduction to GNU/Linux/Unix
Unix102	Advanced Text manipulations, Software Compilation, and Conda/Bioconda installation and use
Unix103	Introduction to Regular Expressions, GREP, SED, and AWK
Unix104	Fundamentals of Scripting

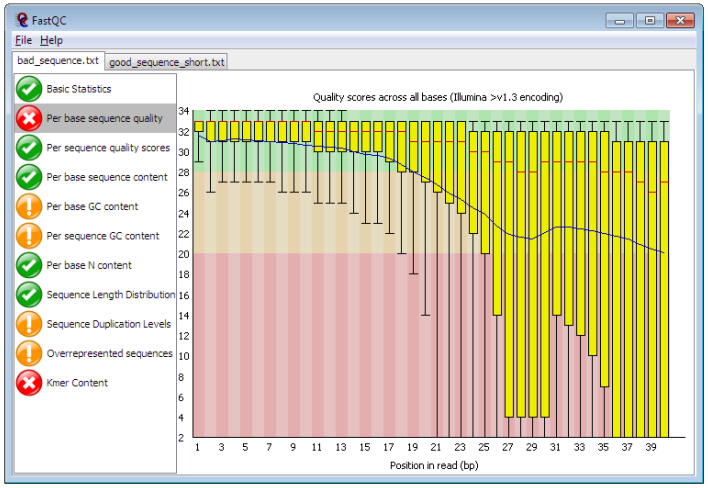
About Digital Biology (BIOL647)

Module 03

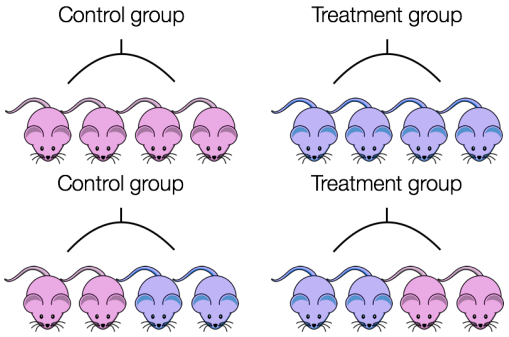
Genome Files, Genes and Genome Browsers



NCBI Databases and NGS Quality Control



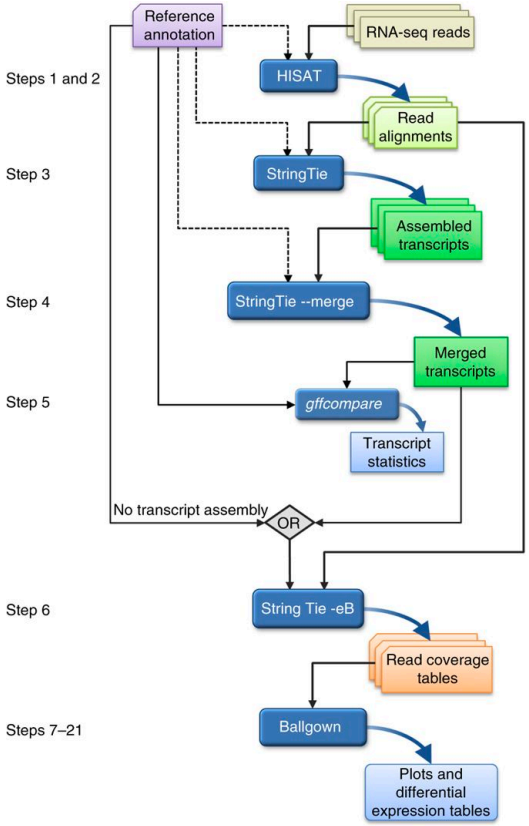
Experimental Design



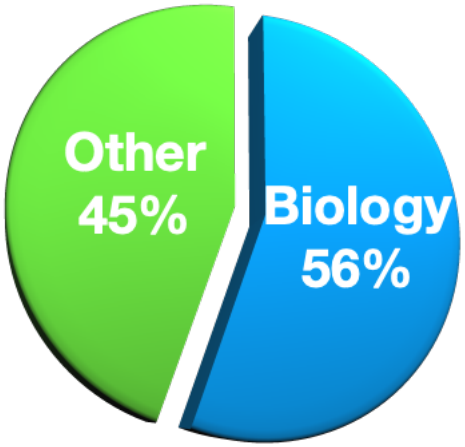
Introduction to and Scripting of Mapping and Alignment

```
#!/usr/bin/env/bash
while getopts :i:n:f:r:c: option;do
    case $option in
        i) index=$OPTARG;;
        n) dName=$OPTARG;;
        f) d1=$OPTARG;;
        r) d2=$OPTARG;;
        c) cores=$OPTARG;;
        ?) echo "I do not know what $OPTARG is!";;
    esac
done
funBWA(){
    bwa \
        mem \
        -t $1 \
        $2 \
        $3 \
        $4 \
        > $5;
}
```

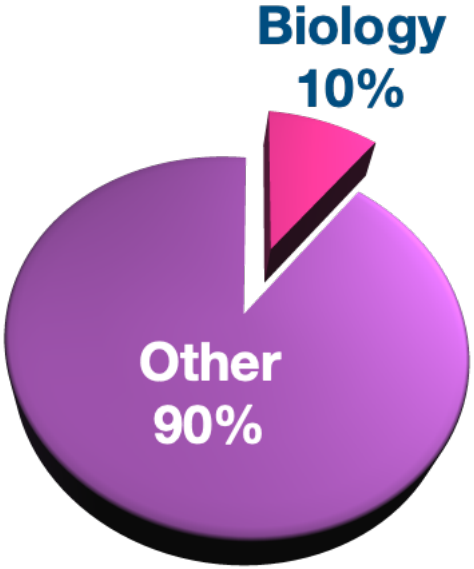
Introduction to and Scripting of Transcriptome Mapping, Assembling and Quantification (Supercomputer)



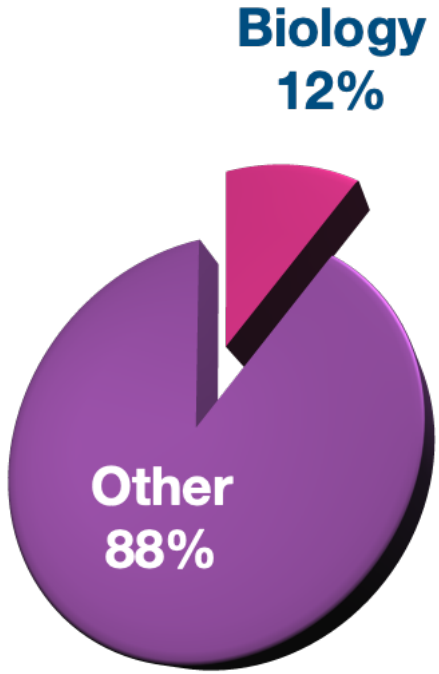
Who Takes BIOL350, BIOL650, and BIOL647?



BIOL350



BIOL650



BIOL647

About Computational Genomics (BIOL350) and Genomics (BIOL650)

Problems Encountered

- Galaxy is intrinsically collaborative, which makes it hard to “isolate” students contributions and/or interactions
- Because we are using the same Galaxy instantiation for Exams and Teaching, we cannot stop students from continuing processing Exam-related data, once a given exam is completed
- Currently, we cannot unmistakably document and isolate a given student’s work on a given exam
- Other Supercomputer users sometimes submit unreal amount of jobs to Ada/Terra while an exam is in progress

Potential Solutions

- To have the ability to float one Galaxy instantiation/student/exam
- Dedicated nodes for Galaxy during exams

Other Observations/Questions

- To access the Supercomputer students need to apply for a Supercomputer account.

  After the semester is over, the majority of students will not be using their Supercomputer account again

  Is this an unnecessary overhead on the Supercomputer facility?

  Should we have a temporary Supercomputer accounts for teaching?

  We urgently need to implement the local deployment of virtual machines and, ideally, to have those machines outside the firewall

Acknowledgements

This talk is dedicated to the memory of

Dr. Jim Hu

My friend and colleague

With whom I developed and taught Genomics for many, many years

Jim's constant sense of humor, and his fearless attitude towards implementing new ideas was the driving force behind these courses

Acknowledgements

I cannot thank enough the Texas A&M High Performance Research Computing facility for their unwavering, constant, support and continuous help

Texas A&M High Performance Research Computing
is one of the jewels of this institution

Special Thanks go to:

Michael Dickens
Lisa Perez
Francis Dang
Mark Huang
Keith E. Jackson
Robin Burns
Honggao Liu