

BIMM 143

Introduction to Bioinformatics

Barry Grant
UC San Diego

<http://thegrantlab.org/bimm143>

HELLO
my name is

BARRY

bjgrant@ucsd.edu

HELLO
HIS — my name is

ALEX

ajweitze@ucsd.edu

HELLO
HER — my name is

YUSI

cyusi@ucsd.edu



Introduce Yourself!

Your preferred name,
Place you identify with,
Major area of study/research,
Favorite joke (optional)!

Today's Menu

Course Logistics

Website, screencasts, survey, ethics, assessment and grading.

Learning Objectives

What you need to learn to succeed in this course.

Course Structure

Major lecture topics and specific learning goals.

Introduction to Bioinformatics

Introducing the *what, why and how* of bioinformatics?

Bioinformatics Database

Hands-on exploration of several major databases and their associated tools.

<http://thegrantlab.org/bimm143/>

The screenshot shows a web browser window with the URL http://bioboot.github.io/bimm143_W20/ in the address bar. The page content is as follows:

UCSanDiego

BIMM 143

A hands-on introduction to the computer-based analysis of genomic and biomolecular data from the Division of Biological Sciences, UCSD.

Overview

Lectures

Computer Setup

Learning Goals

Assignments & Grading

Ethics Code

Course Director
Prof. Barry J. Grant bjgrant@ucsd.edu

Instructional Assistants
Alex Weitzel ([Email: ajweitze@ucsd.edu](mailto:ajweitze@ucsd.edu))
Yusi Chen ([Email: cysi@ucsd.edu](mailto:cysi@ucsd.edu))

Course Syllabus
[Fall 2019 \(PDF\)](#)

Overview

Bioinformatics - the application of computational and analytical methods to biological problems - is a rapidly maturing field that is driving the collection, analysis, and interpretation of the avalanche of data in modern life sciences and medical research.

This upper division 4-unit course is designed for biology majors and provides an introduction to the principles and practical approaches of bioinformatics as applied to

<http://thegrantlab.org/bimm143/>

UCSanDiego

BIMM 143

A hands-on introduction to the computer-based analysis of genomic and biomolecular data from the Division of Biological Sciences, UCSD.

Overview

Lectures

Computer Setup

Learning Goals

Assignments & Grading

Ethics Code

[Twitter](#) [GitHub](#) [Email](#) [RSS](#)

bioboot.github.io/bimm143_W20/

Home Gmail Gcal GitHub BIMM143 BGGN213 Atmosphere BIMM194 Blink News +

Bioinformatics (BIMM 143, Winter 2020)



Course Director
Prof. Barry J. Grant [✉](#) (Email: bjgrant@ucsd.edu)

Instructional Assistants
Alex Weitzel (Email: ajweitze@ucsd.edu)
Yusi Chen (Email: cysi@ucsd.edu)

Course Syllabus
[Fall 2019 \(PDF\) \[✉\]\(#\)](#)

Overview

Bioinformatics - the application of computational and analytical methods to biological problems - is a rapidly maturing field that is driving the collection, analysis, and interpretation of the avalanche of data in modern life sciences and medical research.

This upper division 4-unit course is designed for biology majors and provides an introduction to the principles and practical approaches of bioinformatics as applied to

What essential concepts and skills should YOU attain from this course?

The screenshot shows a web browser window with the URL bioboot.github.io/bimm143_W18/goals/ in the address bar. The page content is as follows:

Learning Goals

At the end of this course students will:

- Understand the increasing necessity for computation in modern life sciences research.
- Be able to use and evaluate online bioinformatics resources including major biomolecular and genomic databases, search and analysis tools, genome browsers, structure viewers, and select quality control and analysis tools to solve problems in the biological sciences.
- Be able to use the R environment to analyze bioinformatics data at scale.
- Understand the process by which genomes are currently sequenced and the bioinformatics processing and analysis required for their interpretation.
- Be familiar with the research objectives of the bioinformatics related sub-disciplines of Genomics, Transcriptomics and Structural bioinformatics.

In short, students will develop a solid foundational knowledge of bioinformatics and be able to evaluate new biomolecular and genomic information using existing bioinformatic tools and resources.

Specific Learning Goals

The left sidebar of the website includes the following navigation links:

- UCSanDiego
- BIMM 143
- A hands-on introduction to the computer-based analysis of genomic and biomolecular data from the Division of Biological Sciences, UCSD [link]
- Overview
- Lectures
- Computer Setup
- Learning Goals
- Assignments & Grading
- Ethics Code

Below the sidebar are social media sharing icons for Twitter, Facebook, Email, and RSS.

At the end of this course students will:

- Understand the increasing necessity for computation in modern life sciences research.
- Be able to use and evaluate online bioinformatics resources and analysis tools to solve problems in the biological sciences.
- Be able to use the R environment to analyze bioinformatics data at scale.
- Be familiar with the research objectives of the bioinformatics related sub-disciplines of Genome informatics, Transcriptomics and Structural informatics.

In short, you will develop a solid foundational knowledge of **bioinformatics** and be able to evaluate new biomolecular and genomic information using **existing bioinformatic tools and resources.**

Specific Learning Goals....

What I want you to know by course end!

The screenshot shows a web browser window with the URL bioboot.github.io/bimm143_W18/goals/. The page content is as follows:

UCSanDiego

BIMM 143

A hands-on introduction to the computer-based analysis of genomic and biomolecular data from the Division of Biological Sciences, UCSD.

Navigation menu:

- Overview
- Lectures
- Computer Setup
- Learning Goals** (highlighted with a red border)
- Assignments & Grading
- Ethics Code

Specific Learning Goals

Teaching toward the specific learning goals below is expected to occupy 60%-70% of class time. The remaining course content is at the discretion of the instructor with student body input. This includes student selected topics for peer presentation, as well as one student selected guest lecture from an industry based genomic scientist.

All students who receive a passing grade should be able to:

		Lecture(s):
1	Appreciate and describe in general terms the role of computation in hypothesis-driven discovery processes within the life sciences.	1, 2, 20
2	Be able to query, search, compare and contrast the data contained in major bioinformatics databases and describe how these databases intersect (GenBank, GENE, UniProt, PFAM, OMIM, PDB, UCSC, ENSEMBLE).	2, 12, 13
3	Describe how nucleotide and protein sequence and structure data are represented (FASTA, FASTQ, GenBank, UniProt, PDB).	3, 10
4	Be able to describe how dynamic programming works for pairwise sequence alignment and appreciate the differences between global and local alignment along with their major application areas.	4, 5
5	Calculate the alignment score between two nucleotide or protein sequences using a provided scoring matrix and be able to perform BLAST, PSI-BLAST, HMMER and protein structure based database	5, 10

Course Structure

Derived from specific learning goals

The screenshot shows a web browser window with the URL bioboot.github.io/bimm143_S18/lectures/. The page title is "Lectures".
The left sidebar includes links for Overview, Lectures (which is highlighted with a red box), Computer Setup, Learning Goals, Assignments & Grading, and Ethics Code.
The main content area displays a table of lectures for Spring 2018:

#	Date	Topics for Spring 2018
1	Tu, 04/03	Welcome to Bioinformatics Course introduction, Learning goals & expectations, Biology is an information science, History of Bioinformatics, Types of data, Application areas and introduction to upcoming course segments, Hands on with major Bioinformatics databases and key online NCBI and EBI resources
2	Th, 04/05	Sequence alignment fundamentals, algorithms and applications Homology, Sequence similarity, Local and global alignment, classic Needleman-Wunsch, Smith-Waterman and BLAST heuristic approaches, Hands on with dot plots, Needleman-Wunsch and BLAST algorithms highlighting their utility and limitations
		Advanced sequence alignment and database searching

Course Structure

Derived from specific learning goals

The screenshot shows a web browser window with the URL bioboot.github.io/bimm143_S18/lectures/. The page title is "BIMM 143". The left sidebar has links for Overview, Lectures (highlighted with a red box), Computer Setup, Learning Goals, Assignments & Grading, and Ethics Code. The main content area has a heading "Lectures" and a table titled "Topics for Spring 2018". The table has columns for #, Date, and Topics. The first row (Tu, 04/03) has the topic "Welcome to Bioinformatics" highlighted with a red box. The second row (Th, 04/05) has the topic "Sequence alignment fundamentals, algorithms and applications".

#	Date	Topics for Spring 2018
1	Tu, 04/03	Welcome to Bioinformatics Course introduction, Learning goals & expectations, Biology is an information science, History of Bioinformatics, Types of data, Application areas and introduction to upcoming course segments, Hands on with major Bioinformatics databases and key online NCBI and EBI resources
2	Th, 04/05	Sequence alignment fundamentals, algorithms and applications Homology, Sequence similarity, Local and global alignment, classic Needleman-Wunsch, Smith-Waterman and BLAST heuristic approaches, Hands on with dot plots, Needleman-Wunsch and BLAST algorithms highlighting their utility and limitations
		Advanced sequence alignment and database searching

Class Details

Goals, Class material, Screencasts & **Homework**

The screenshot shows a web browser window with the URL bioboot.github.io/bimm143_W18/lectures/#1. The page has a dark blue background with a glowing blue circular graphic on the left.

UCSanDiego

BIMM 143

A hands-on introduction to the computer-based analysis of genomic and biomolecular data from the Division of Biological Sciences, UCSD [\[link\]](#).

Overview

Lectures

Computer Setup

Learning Goals

Assignments & Grading

Ethics Code

1: Welcome to Foundations of Bioinformatics

Topics:

Course introduction, Learning goals & expectations, Biology is an information science, History of Bioinformatics, Types of data, Application areas and introduction to upcoming course segments, Student 30-second introductions, Student computer setup.

Goals:

- Understand course scope, expectations, logistics and [ethics code](#).
- Understand the increasing necessity for computation in modern life sciences research.
- Get introduced to how bioinformatics is practiced.
- Complete the [pre-course questionnaire](#) [\[link\]](#).
- Setup your [laptop computer](#) for this course.

Material:

- Pre class screen casts (also see below):
 - SC1: [Welcome to BIMM-143](#) [\[link\]](#),
 - SC2: [What is Bioinformatics?](#) [\[link\]](#) and
 - SC3: [How do we do Bioinformatics?](#) [\[link\]](#).
- Lecture Slides: Large PDF, Small PDF
- [Handout: Class Syllabus](#) [\[link\]](#)

Homework

Goals, Class material, Screencasts & Homework

The screenshot shows a web browser window with the URL bioboot.github.io/bimm143_W18/lectures/#1. The page content includes:

- UC San Diego BIMM 143**: A hands-on introduction to the computer-based analysis of genomic and biomolecular data from the Division of Biological Sciences, UCSD.
- Homework:**
 - [Questions](#)
 - Readings:
 - PDF1: [What is bioinformatics? An introduction and overview](#),
 - PDF2: [Advancements and Challenges in Computational Biology](#),
 - Other: [For Big-Data Scientists, ‘Janitor Work’ Is Key Hurdle to Insights](#) New York Times, 2014.
- Screen Casts:**

Welcome to “Foundations of Bioinformatics” (BGGN-213)

The video player displays a screen cast of a man speaking in front of a colorful molecular model. The video controls show it's at 2:05 / 4:05. The URL <http://thegrantlab.org/baan213> is visible at the bottom of the video frame.

1 Welcome to BIMM-143: Course introduction and logistics.

Homework

Goals, Class material, Screencasts & Homework

The screenshot shows a web browser window with the URL bioboot.github.io/bimm143_W18/lectures/#1. The page content includes:

- Homework:**
 - [Questions](#) (highlighted with a red box)
 - Readings:
 - PDF1: [What is bioinformatics? An introduction and overview](#),
 - PDF2: [Advancements and Challenges in Computational Biology](#),
 - Other: [For Big-Data Scientists, ‘Janitor Work’ Is Key Hurdle to Insights](#) New York Times, 2014.
- Screen Casts:**
 - Welcome to “Foundations of Bioinformatics” (BGGN-213)**: A video player showing a man speaking in front of a background of colorful 3D molecular models. The video is at 2:05 / 4:05. The URL <http://thegrantlab.org/baan213> is displayed below the video player.

On the left side of the browser, the UC San Diego logo and the course title **BIMM 143** are visible. The course description is: "A hands-on introduction to the computer-based analysis of genomic and biomolecular data from the Division of Biological Sciences, UCSD". Below the course title, there is a sidebar with links to **Overview**, **Lectures**, **Computer Setup**, **Learning Goals**, **Assignments & Grading**, and **Ethics Code**.

Homework

Goals, Class material, Screencasts & **Homework**

BIMM143 Lecture 1 Homework (W19)

Please answer the following questions including your main @ucsd.edu email address and UCSD PID number so you can receive credit for your responses.

* Required

Email address *

Your email

UCSD PID number (exam number)

Your answer

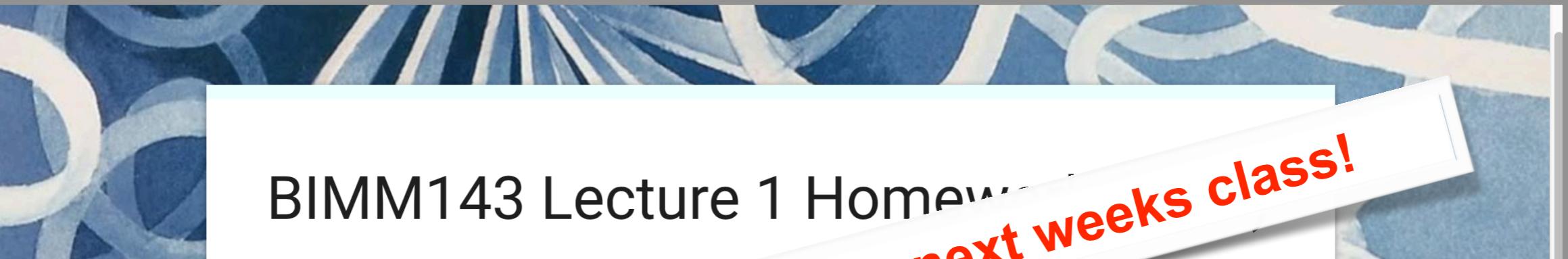
Which of the following operating systems is most frequently used for bioinformatics tool development

1 point

Homework

(35% of course grade)

Goals, Class material, Screencasts & **Homework**



BIMM143 Lecture 1 Homework

Please answer the following questions and include your UCSD email address and UCSD PID number so you can receive feedback.

Homework is due before the next weeks class!

Email address *

Your email

UCSD PID number (exam number)

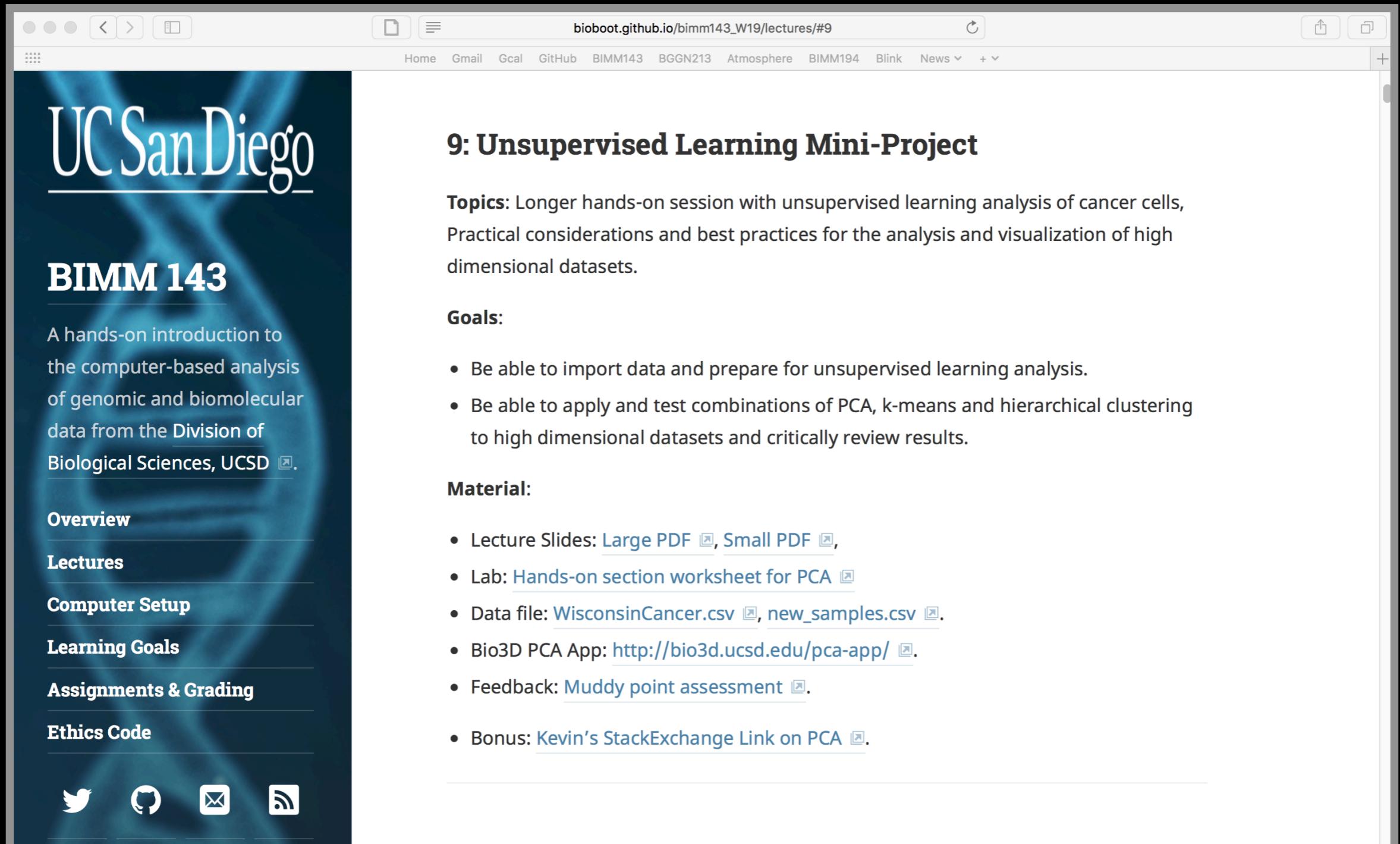
Your answer

Which of the following operating systems is most frequently used for bioinformatics tool development

1 point

Projects

Week long **mini-projects** (x2),
and 1 five week main project



The screenshot shows a web browser window with the URL bioboot.github.io/bimm143_W19/lectures/#9. The page content is as follows:

UCSanDiego

BIMM 143

A hands-on introduction to the computer-based analysis of genomic and biomolecular data from the Division of Biological Sciences, UCSD.

Overview

Lectures

Computer Setup

Learning Goals

Assignments & Grading

Ethics Code

9: Unsupervised Learning Mini-Project

Topics: Longer hands-on session with unsupervised learning analysis of cancer cells, Practical considerations and best practices for the analysis and visualization of high dimensional datasets.

Goals:

- Be able to import data and prepare for unsupervised learning analysis.
- Be able to apply and test combinations of PCA, k-means and hierarchical clustering to high dimensional datasets and critically review results.

Material:

- Lecture Slides: [Large PDF](#), [Small PDF](#),
- Lab: [Hands-on section worksheet for PCA](#)
- Data file: [WisconsinCancer.csv](#), [new_samples.csv](#).
- Bio3D PCA App: <http://bio3d.ucsd.edu/pca-app/>
- Feedback: [Muddy point assessment](#)
- Bonus: [Kevin's StackExchange Link on PCA](#)

Projects

Week long **mini-projects** (x2),
and 1 five week main project

The image shows a Mac OS X desktop with two browser windows open. The top window is the course homepage for BIMM 143 at UC San Diego, featuring the title "BIMM 143" and a description of the course as a hands-on introduction to genomic analysis. The bottom window is a lecture slide titled "Designing a personalized cancer vaccine" from BIMM-143 Lecture 18, dated March 7, 2018. The slide includes notes on somatic mutations and variant calling algorithms.

BIMM 143
A hands-on introduction to the computer-based analysis of genomic and biomolecular data from the Division of Biological Sciences, UCSD.

Designing a personalized cancer vaccine

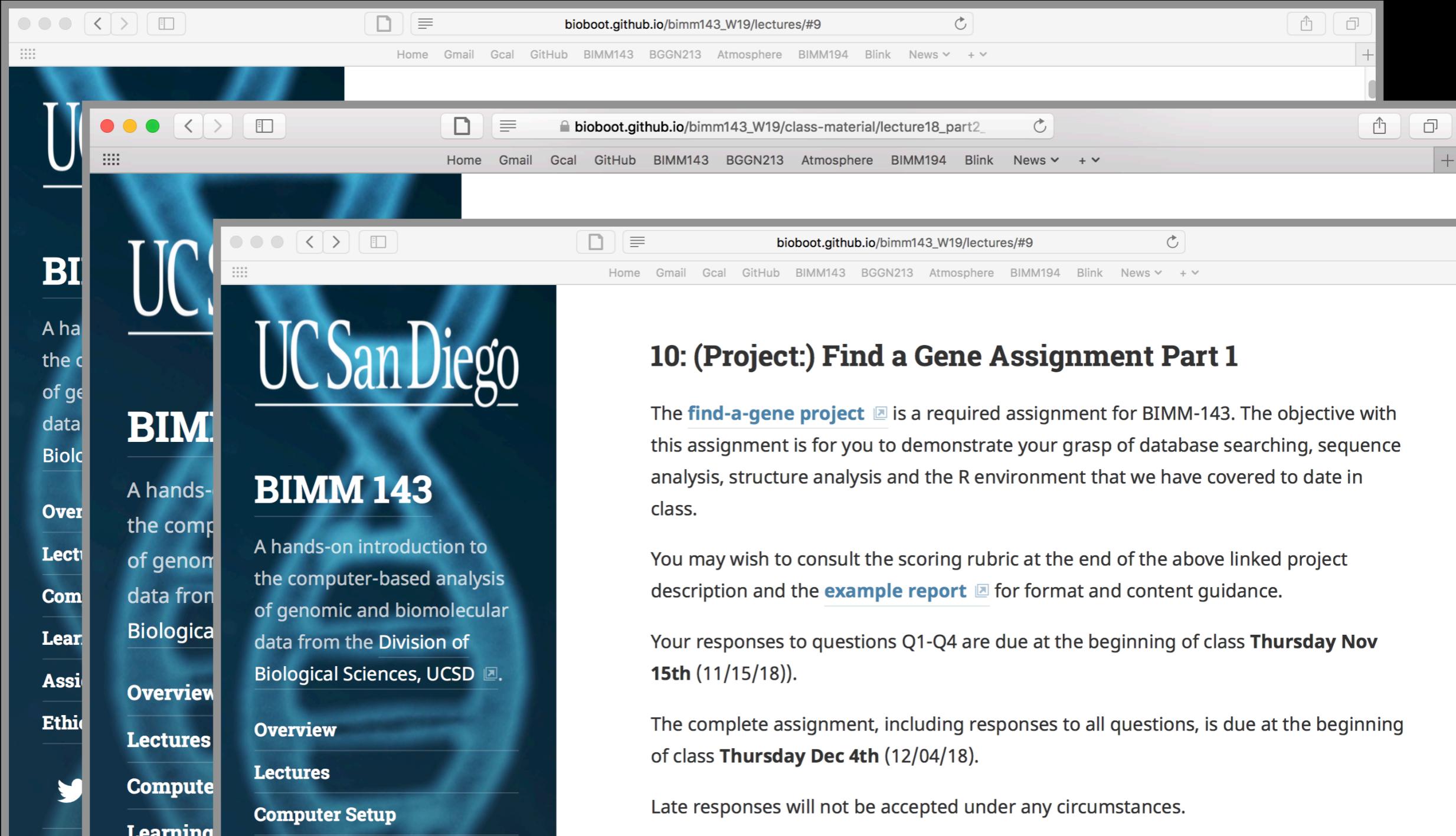
BIMM-143 Lecture 18:
Barry Grant <<http://thegrantlab.org>>
Date: 2018-03-07 (15:24:21 PST on Wed, Mar 07)

Notes: To identify somatic mutations in a tumor, DNA from the tumor is sequenced and compared to DNA from normal tissue in the same individual using *variant calling algorithms*. Comparison of tumor sequences to those from normal tissue (rather than 'the human genome') is important to ensure that the detected differences are not germline mutations.

To identify which of the somatic mutations leads to the production of aberrant proteins, the location of the mutation in the genome is inspected to identify non-

Projects (20% of course grade)

Week long mini-projects (x2),
and 1 five week **main project**



The **find-a-gene project** is a required assignment for BIMM-143. The objective with this assignment is for you to demonstrate your grasp of database searching, sequence analysis, structure analysis and the R environment that we have covered to date in class.

You may wish to consult the scoring rubric at the end of the above linked project description and the [example report](#) for format and content guidance.

Your responses to questions Q1-Q4 are due at the beginning of class **Thursday Nov 15th (11/15/18)**.

The complete assignment, including responses to all questions, is due at the beginning of class **Thursday Dec 4th (12/04/18)**.

Late responses will not be accepted under any circumstances.

Why Projects?

- Projects allow you to practice your new Bioinformatics skills in a less guided environment.
- In Projects, we provide datasets and ask you questions about them; just like a research project.
- Projects help build a personal portfolio and showcase your new skills, as well as help put what we have learned into practice.

Final Exam

Open-book, open-notes 150-minute test
(45% of course grade)

The screenshot shows a web browser window with the URL bioboot.github.io/bimm143_W19/lectures/#18. The page content is as follows:

UCSanDiego
BIMM 143

A hands-on introduction to the computer-based analysis of genomic and biomolecular data from the Division of Biological Sciences, UCSD

Overview
Lectures
Computer Setup
Learning Goals
Assignments & Grading
Ethics Code

20: Final Exam

This open-book, open-notes 150-minute test consists of 35 questions. The number of points for each question is indicated in green font at the beginning of each question. There are 80 total points on offer.

Please remember to:

- Read all questions carefully before starting.
- Put your name, UCSD email and PID number on your test.
- Write all your answers on the space provided in the exam paper.
- Remember that concise answers are preferable to wordy ones.
- Clearly state any simplifying assumptions you make in solving a problem.
- No copies of this exam are to be removed from the class-room.
- No talking or communication (electronic to otherwise) with your fellow students once the exam has begun.
- **Good luck!**

Bonus:

Online portfolio of **your** bioinformatics work!

The screenshot shows a web browser window with the URL jasonpbennett.github.io/bimm143/. The page content is as follows:

Introduction to Bioinformatics Class S18

A repository to store and display my work completed during the Spring 2018 quarter in BIMM-143 at UCSD.

[View the Project on GitHub](#)
jasonPBennett/bimm143

Bioinformatics Class BIMM-143

This is my repository for my Bioinformatics class from UC San Diego in S18.

Index of Material

Introductory Material: Working With R

- [Class 5 - Basic Data Exploration and Visualization in R](#)
- [Class 6 - Creating R Functions](#)
- [Class 7 - R Packages, working with CRAN, and working with Bioconductor](#)

Using R and Other Tools for Bioinformatics Analysis

- [Class 8 - An Introduction to Machine Learning \(Hierarchical Clustering\)](#)
- [Class 9 - Analyzing High Dimensional Datasets and Unsupervised Learning](#)
- [Class 11 - Structural Bioinformatics: Analyzing Protein Structure and Function](#)
- [Class 12 - Drug Discovery: Techniques and Analysis](#)
- [Class 13 - Genome Informatics and High Throughput Sequencing \(NGS, RNA-Seq, and FastQC\)](#)
- [Class 14 - Transcriptomics and RNA-Seq Analysis](#)
- [Class 15 - Genome Annotation and Using Functional Databases \(KEGG and GO - Gene Ontology\)](#)
- [Class 16 - Transposons: A Sample Workflow](#)

This project is maintained by
JasonPBennett

Bonus:

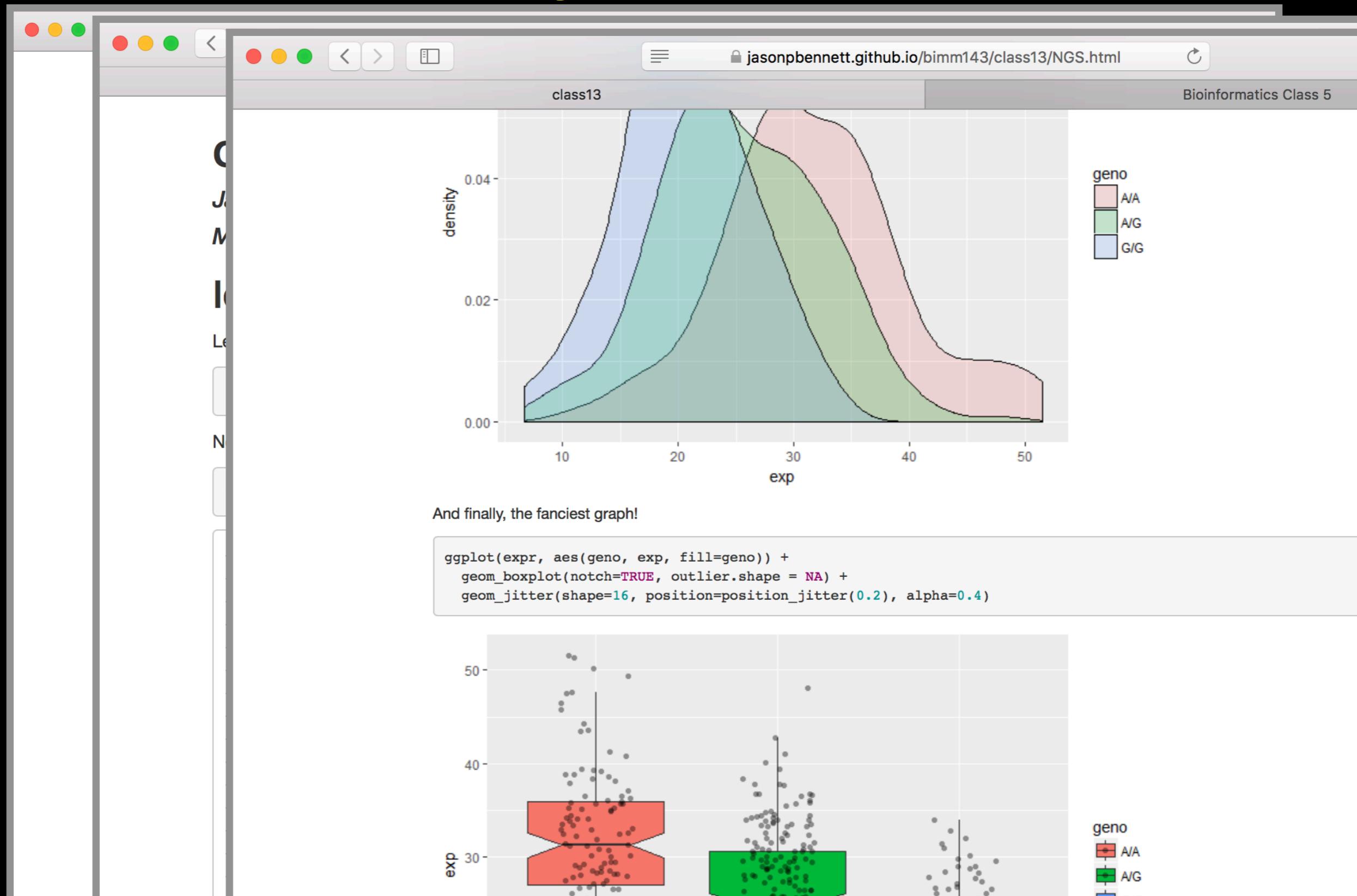
Online portfolio of **your** bioinformatics work!

The screenshot shows a web browser window with the following details:

- Address Bar:** jasonpbennett.github.io/bimm143/class13/NGS.html
- Page Title:** class13
- Page Subtitle:** Bioinformatics Class 5
- Content:**
 - # class13
 - Jason Patrick Bennett*
 - May 15, 2018*
 - ## Identifying SNP's in a Population
 - Lets analyze SNP's from the Mexican-American population in Los Angeles:
 - ```
genotype <- read.csv("373531-SampleGenotypes-Homo_sapiens_Variation_Sample_rs8067378.csv")
```
  - Now lets look at a table of the data:
  - ```
table(genotype)
```
 - ```
, , Population.s. = ALL, AMR, MXL, Father = -, Mother = -
##
Genotype..forward.strand.
Sample..Male.Female.Unknown. A|A A|G G|A G|G
NA19648 (F) 1 0 0 0
NA19649 (M) 0 0 0 1
NA19651 (F) 1 0 0 0
NA19652 (M) 0 0 0 1
NA19654 (F) 0 0 0 1
NA19655 (M) 0 1 0 0
NA19657 (F) 0 1 0 0
NA19658 (M) 1 0 0 0
NA19661 (M) 0 1 0 0
NA19663 (F) 1 0 0 0
NA19664 (M) 0 0 1 0
NA19669 (F) 1 0 0 0
```

# Bonus:

## Online portfolio of **your** bioinformatics work!



## **Side Note: Why stick with this course?**

**Provides a hands-on practical introduction to major bioinformatics concepts and resources.**

Covers modern hot topics and the intimate coupling of informatics with biology - highlighting the impact of computing advances and 'big data' on biology!

Designed for biology majors with no programming experience or high level math skills.

Provides a hook for increasing computational and data science competencies in the biosciences - valuable high demand translational skills!

## **Side Note: Why stick with this course?**

**Provides a hands-on practical introduction to major bioinformatics concepts and resources.**

Covers modern hot topics and the intimate coupling of informatics with biology - highlighting the impact of computing advances and 'big data' on biology!

Designed for biology majors with no programming experience or high level math skills.

Provides a hook for increasing computational and data science competencies in the biosciences - valuable high demand translational skills!

# BIMM-143 Learning Goals....

## Data science R based learning goals

UCSanDiego

BIMM 143

A hands-on introduction to the computer-based analysis of genomic and biomolecular data from the Division of Biological Sciences, UCSD.

Overview

Lectures

Computer Setup

Learning Goals

Assignments & Grading

Ethics Code

bioboot.github.io/bimm143\_W18/goals/

BIMM 143 Home Gmail Gcal Bitbucket GitHub News Disqus BGGN-213 BIMM-143 GDocs

|    |                                                                                                                                                                                                                                                                                         |                          |
|----|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------|
| 5  | Calculate the alignment score between two nucleotide or protein sequences using a provided scoring matrix and be able to perform BLAST, PSI-BLAST, HMMER and protein structure based database searches and interpret the results in terms of the biological significance of an e-value. | 5, 10                    |
| 6  | Use R to read and parse comma-separated (.csv) formatted files ready for subsequent analysis.                                                                                                                                                                                           | 8, 9, 10, 11, 13, 15, 16 |
| 7  | Perform elementary statistical analysis on biomolecular and "omics" datasets with R and produce informative graphical displays and data summaries.                                                                                                                                      | 9, 10, 11, 13, 15, 16    |
| 8  | View and interpret the structural models in the PDB.                                                                                                                                                                                                                                    | 10, 11                   |
| 9  | Explain the outputs from structure prediction algorithms and small molecule docking approaches.                                                                                                                                                                                         | 11                       |
| 10 | Appreciate and describe in general terms the rapid advances in sequencing technologies and the new areas of investigation that these advances have made accessible.                                                                                                                     | 13, 14, 15               |
| 11 | Understand the process by which genomes are currently sequenced and the bioinformatics processing and analysis required for their interpretation.                                                                                                                                       | 13                       |
| 12 | For a genomic region of interest (e.g. the neighborhood of a particular gene), use a genome browser to view nearby genes, transcription factor binding regions, epigenetic information, etc.                                                                                            | 14                       |

# BIMM-143 Learning Goals....

Delve deeper into “real-world” bioinformatics

UC San Diego

**BIMM 143**

A hands-on introduction to the computer-based analysis of genomic and biomolecular data from the Division of Biological Sciences, UCSD.

**Overview**

**Lectures**

**Computer Setup**

**Learning Goals**

**Assignments & Grading**

**Ethics Code**

|    | view and interpret the structural models in the PDB.                                                                                                                                         | 10, 11     |
|----|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------|
| 9  | Explain the outputs from structure prediction algorithms and small molecule docking approaches.                                                                                              | 11         |
| 10 | Appreciate and describe in general terms the rapid advances in sequencing technologies and the new areas of investigation that these advances have made accessible.                          | 13, 14, 15 |
| 11 | Understand the process by which genomes are currently sequenced and the bioinformatics processing and analysis required for their interpretation.                                            | 13         |
| 12 | For a genomic region of interest (e.g. the neighborhood of a particular gene), use a genome browser to view nearby genes, transcription factor binding regions, epigenetic information, etc. | 14         |
| 13 | Given an RNA-Seq data file, find the set of significantly differentially expressed genes and use online tools to interpret gene lists and annotate potential gene functions.                 | 15, 16     |
| 14 | Perform a GO analysis to identify the pathways relevant to a set of genes (e.g. identified by transcriptomic study or a proteomic experiment).                                               | 16         |
| 15 | Use the KEGG pathway database to look up interaction pathways.                                                                                                                               | 17         |
| 16 | Use graph theory to represent biological data networks.                                                                                                                                      | 17, 18     |
| 17 | Understand the challenges in integrating and interpreting large heterogeneous high throughput data sets into their functional                                                                | 19         |

# **These support a major learning objective**

**At the end of this course students will:**

- Understand the increasing necessity for computation in modern life sciences research.
- Be able to use and evaluate online bioinformatics resources and analysis tools to solve problems in the biological sciences.
- Be able to use the R environment to analyze bioinformatics data at scale.
- Be familiar with the research objectives of the bioinformatics related sub-disciplines of Genome informatics, Transcriptomics and Structural informatics.

# Why use R?

Productivity

Flexibility

Genomic data analysis

# IEEE 2016 Top Programming Languages

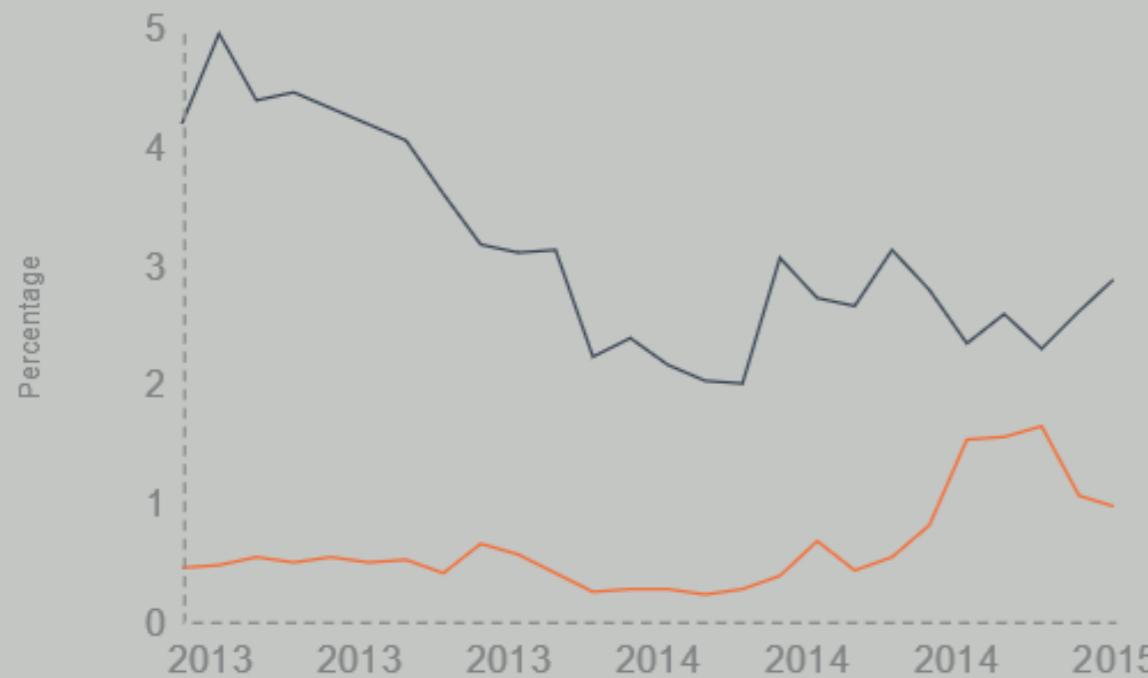
| Language Rank | Types | Spectrum Ranking |
|---------------|-------|------------------|
| 1. C          |       | 100.0            |
| 2. Java       |       | 98.1             |
| 3. Python     |       | 98.0             |
| 4. C++        |       | 95.9             |
| 5. R          |       | 87.9             |
| 6. C#         |       | 86.7             |
| 7. PHP        |       | 82.8             |
| 8. JavaScript |       | 82.2             |
| 9. Ruby       |       | 74.5             |
| 10. Go        |       | 71.9             |

<http://spectrum.ieee.org/computing/software/the-2016-top-programming-languages>

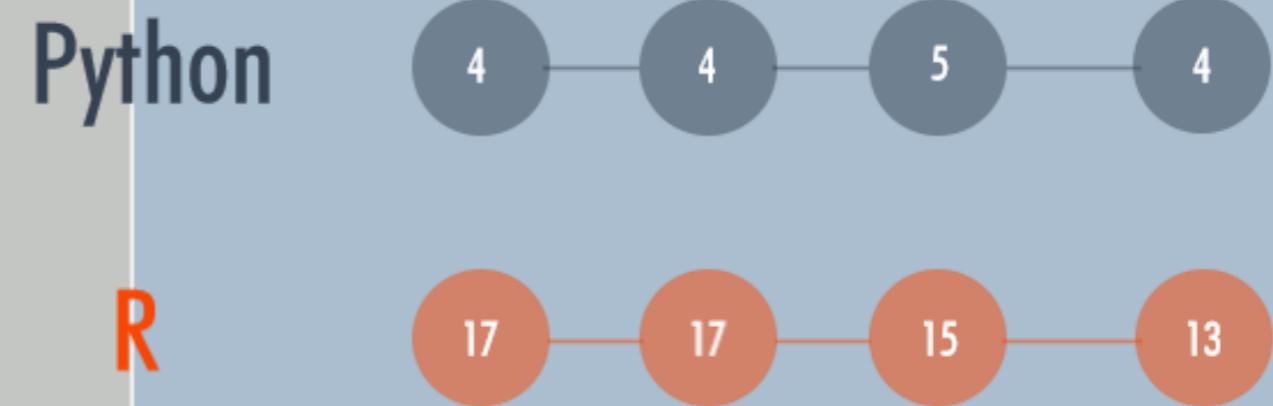
# R and Python: The Numbers

## Popularity Rankings

R and Pythons popularity between 2013 and February 2015 (Tiobe Index)



Redmonk ranking, comparing the relative performance of programming languages on GitHub and Stack Overflow (September 2012 and January 2013, 2014, 2015)



## Jobs And Salary?

2014 Dice Tech Salary Survey:  
Average Salary For High Paying Skills and Experience



\$ 115,531



Python

\$ 94,139

[http://www.kdnuggets.com/2015/05/r-vs-python-data-science.html?  
utm\\_medium=email&utm\\_source=flipboard](http://www.kdnuggets.com/2015/05/r-vs-python-data-science.html?utm_medium=email&utm_source=flipboard)

# R is designed specifically for data analysis

- Large friendly user and developer community.
  - As of Jan 6th 2019 there are 13,645 add on **R packages** on **CRAN** and 1,649 on **Bioconductor** - much more on these later!
- Virtually every statistical technique is either already built into R, or available as a free package.
- Unparalleled data analysis environment for **high-throughput genomic data**.

< https://www.datacamp.com/ >

The screenshot shows the DataCamp homepage with a red circle highlighting the user profile icon in the top right corner. The profile icon has a red notification badge with the number '3'.

**Your Latest Activity**

**Introduction to Spark in R using dplyr**

You are doing awesome barryus! So far you've earned 250 XP!

The last chapter you were working on was [Light My Fire: Starting To Use Spark With dplyr Syntax](#).

**DAILY PRACTICE**

Learning data science requires practice **every day**. Build your data science fluency with DataCamp practice mode.

**Notifications:**

- You have a new assignment: Conditionals and Con... 16 days ago
- You have a new assignment: Working with the RSt... 16 days ago
- You have a new assignment: Introduction to R 16 days ago
- bjgrant invited you to the group 'Foundations o... 16 days ago
- You have a new assignment: Orientation 9 months ago

[See all notifications](#)

< https://www.datacamp.com/ >

The screenshot shows a DataCamp course page titled "What is an IDE anyway?". The page includes a brief description of RStudio, a question about what IDE stands for, and a section titled "Possible Answers" with five options. The option "Integrated Development Environment" is circled in red. Below it is a "Take Hint (-15xp)" button. A large yellow "Submit Answer" button at the bottom is also circled in red. To the right of the course content is an RStudio IDE window displaying R version 3.3.1 output and an empty environment.

What is an IDE anyway? | R

Secure | https://campus.datacamp.com/courses/working-with-the-rstudio-ide-part-1/orientation?ex=2

DataCamp

Course Outline

What is an IDE anyway?

50xp

RStudio is an IDE that makes R easier to use by combining a set of tools into a single environment.

What does IDE stand for?

Possible Answers

- Intensive Design Environment
- Integrated Document Environment
- Independent Developer Ecosystem
- Integrated Development Environment

Take Hint (-15xp)

Submit Answer

R version 3.3.1 (2016-06-21) -- "Bug in Your Hair"  
Copyright (C) 2016 The R Foundation for Statistical Computing  
Platform: x86\_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.  
You are welcome to redistribute it under certain conditions.  
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.  
Type 'contributors()' for more information and  
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help,  
or  
'help.start()' for an HTML browser interface to help.  
Type 'q()' to quit R.

> |

Environment History

Import Dataset List Global Environment

Environment is empty

Files Plots Packages Help Viewer

New Folder Upload Delete Rename

Home

| Name | Size |
|------|------|
|      |      |

< https://www.datacamp.com/ >

The screenshot shows the RStudio IDE interface running within a web browser window. The browser title bar reads "What is an IDE anyway? | R". The address bar shows a secure connection to "https://campus.datacamp.com/courses/working-with-the-rstudio-ide-part-1/orientation?ex=2". The DataCamp logo is in the top left, and a course outline navigation bar is at the top right.

The main content area displays a course exercise titled "What is an IDE anyway?". A prominent message says "Exercise Completed" with a blue button containing "50xp" (experience points). Two specific buttons are circled in red: one labeled "Continue" in a yellow box, and another labeled "Take Hint (-15xp)" in a grey box.

The RStudio interface includes:

- Console**: Shows R version 3.3.1 (2016-06-21) -- "Bug in Your Hair" and the platform as x86\_64-pc-linux-gnu (64-bit).
- Environment**: Shows an empty global environment.
- Files**: Shows a "Home" folder with a single file named "R" (Size 0).

The overall layout is a combination of the DataCamp course UI and the RStudio development environment.

< https://www.datacamp.com/ >

# Homework assignments will be via DataCamp

The screenshot shows a DataCamp exercise interface. On the left, there's a sidebar with a 'DataCamp' logo, a 'Course Outline' button, and a '5+' notifications icon. Below that is a 'Exercise' tab, followed by a section titled 'PCA analysis'. The main area has tabs for 'script.R' and 'RDocumentation'. The 'script.R' tab contains the following code:

```
1 # Transform the normalized counts
2 vsd_smoc2 <- vst(dds_smoc2, blind = TRUE)
3
4 # Plot the PCA of PC1 and PC2
5 ---(_____, intgroup=____)
```

Below the code editor are three buttons: a blue 'Run Code' button, a green 'Submit Answer' button, and a small circular icon with a '5'.

At the bottom, there are tabs for 'R Console' and 'Slides'. The 'R Console' tab shows the following session history:

```
> ?plotPCA
> plotPCA(vsd_smoc2)
Error: object 'vsd_smoc2' not found
> vsd_smoc2 <- vst(dds_smoc2, blind = TRUE)
+
> plotPCA(vsd_smoc2)
>
```

On the far left, under the exercise title, there's a 'Instructions 1/2 50 XP' section with two numbered steps. Step 1 is 'Run the code to transform the normalized counts.' Step 2 is 'Perform PCA by plotting PC1 vs PC2 using the DESeq2 `plotPCA()` function on the DESeq2 transformed counts object, `vsd_smoc2` and specify the `intgroup` argument as the factor to color the plot.'

At the bottom left, there's a 'Take Hint (-15 XP)' button.

< https://www.datacamp.com/ >

The screenshot shows a web browser window for DataCamp. The URL in the address bar is <https://www.datacamp.com/groups-foundations-of-bioinformatics-bggn-213/details>. The DataCamp logo is in the top left. The top navigation bar includes 'Learn', 'Groups' (which is circled in red), 'About', and user stats ('1,550 XP'). Below the navigation is a 'Back to My Dashboard' link. The main title is 'Foundations of Bioinformatics (BGGN-213)'. A 'Leaderboard' section displays a table of users based on their activity over the last 30 days. The table has columns for Member, XP, Courses, and Chapters. The top 8 members are listed:

| Rank | Member            | XP    | Courses | Chapters |
|------|-------------------|-------|---------|----------|
| 1    | Angela Nicholson  | 22450 | 4       | 20       |
| 2    | Ben Song          | 12850 | 2       | 11       |
| 3    | Ana Grant         | 12120 | 2       | 9        |
| 4    | Delaney Pagliuso  | 12085 | 2       | 11       |
| 5    | oehernan          | 11055 | 2       | 10       |
| 6    | Erin Schiksnis    | 10350 | 2       | 9        |
| 7    | Zachary Warburg   | 9110  | 1       | 8        |
| 8    | Alexander Weitzel | 6950  | 1       | 6        |

# Today's Menu

## Course Logistics

Website, screencasts, survey, ethics, assessment and grading.

## Learning Objectives

What you need to learn to succeed in this course.

## Course Structure

Major lecture topics and specific learning goals.

## Introduction to Bioinformatics

Introducing the *what, why and how* of bioinformatics?

## Bioinformatics Database

Hands-on exploration of several major databases and their associated tools.

**“What is Bioinformatics?”**

*“Bioinformatics is the application of computers to the collection, archiving, organization, and analysis of biological data.”*

... A hybrid of biology and computer science

*“Bioinformatics is the application of computers to the collection, archiving, organization, and analysis of biological data.”*

**Bioinformatics is computer aided biology!**

*“Bioinformatics is the application of computers to the collection, archiving, organization, and analysis of biological data.”*

**Bioinformatics is computer aided biology!**

**Goal: Data to Knowledge**

# There are many useful definitions...

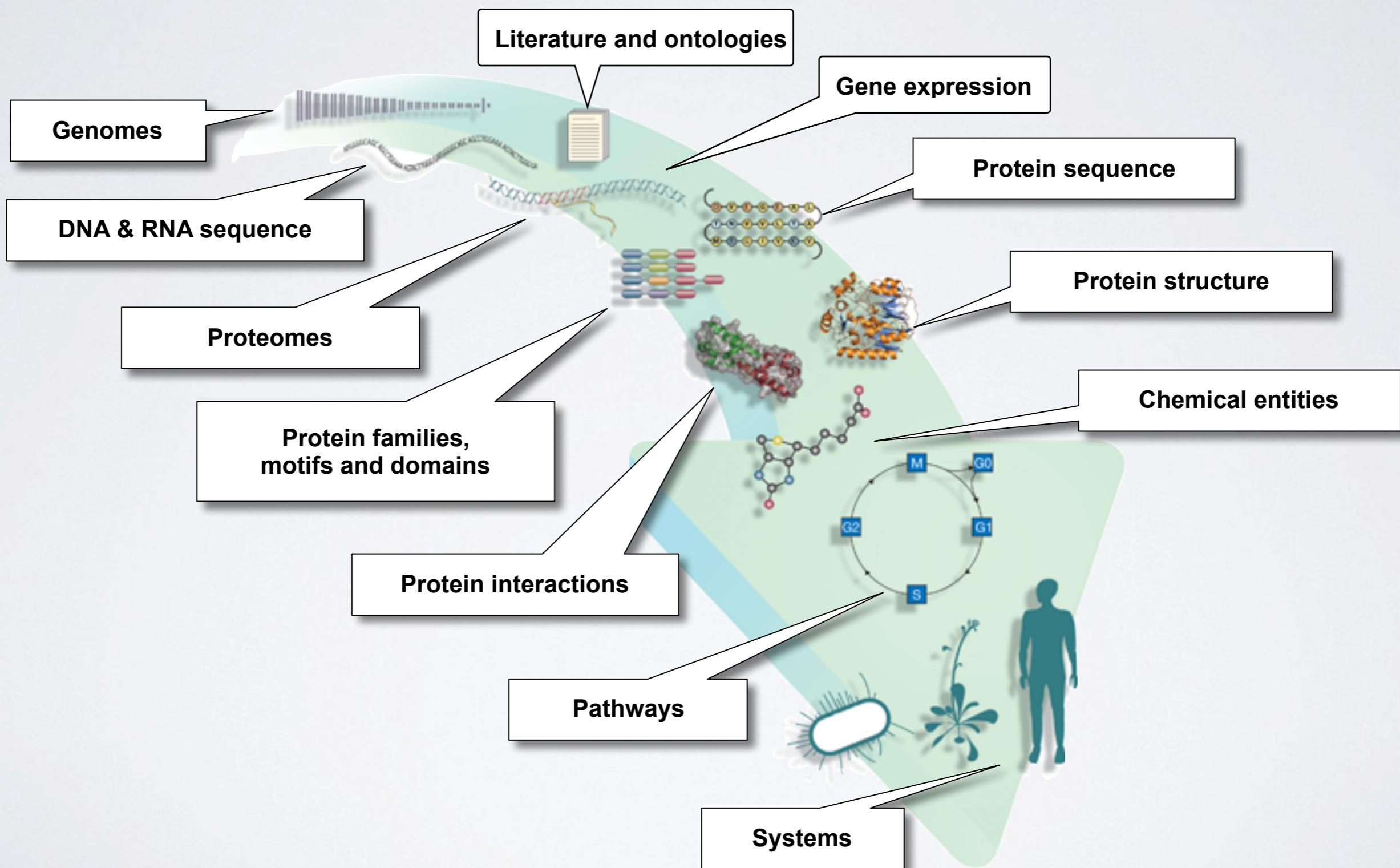
- "Computer based **management** and **analysis** of biological and biomedical data with useful applications in many disciplines, particularly **genomics**, **proteomics**, **metabolomics**, and related fields."  
**(BIMM-143)**
- "Bioinformatics is conceptualizing biology in terms of **macromolecules** and then applying "**informatics**" techniques (derived from disciplines such as applied maths, computer science, and statistics) to **understand** and **organize** the information associated with these molecules, on a **large-scale**."  
**(Luscombe *et al.* 2001)**
- "Bioinformatics is research, development, or application of **computational approaches** for expanding the use of biological, medical, behavioral or health data, including those to acquire, **store**, **organize** and **analyze** such data ...<cut>..."  
**(National Institutes of Health: <http://tinyurl.com/l3gxr6b>)**

# There are many useful definitions...

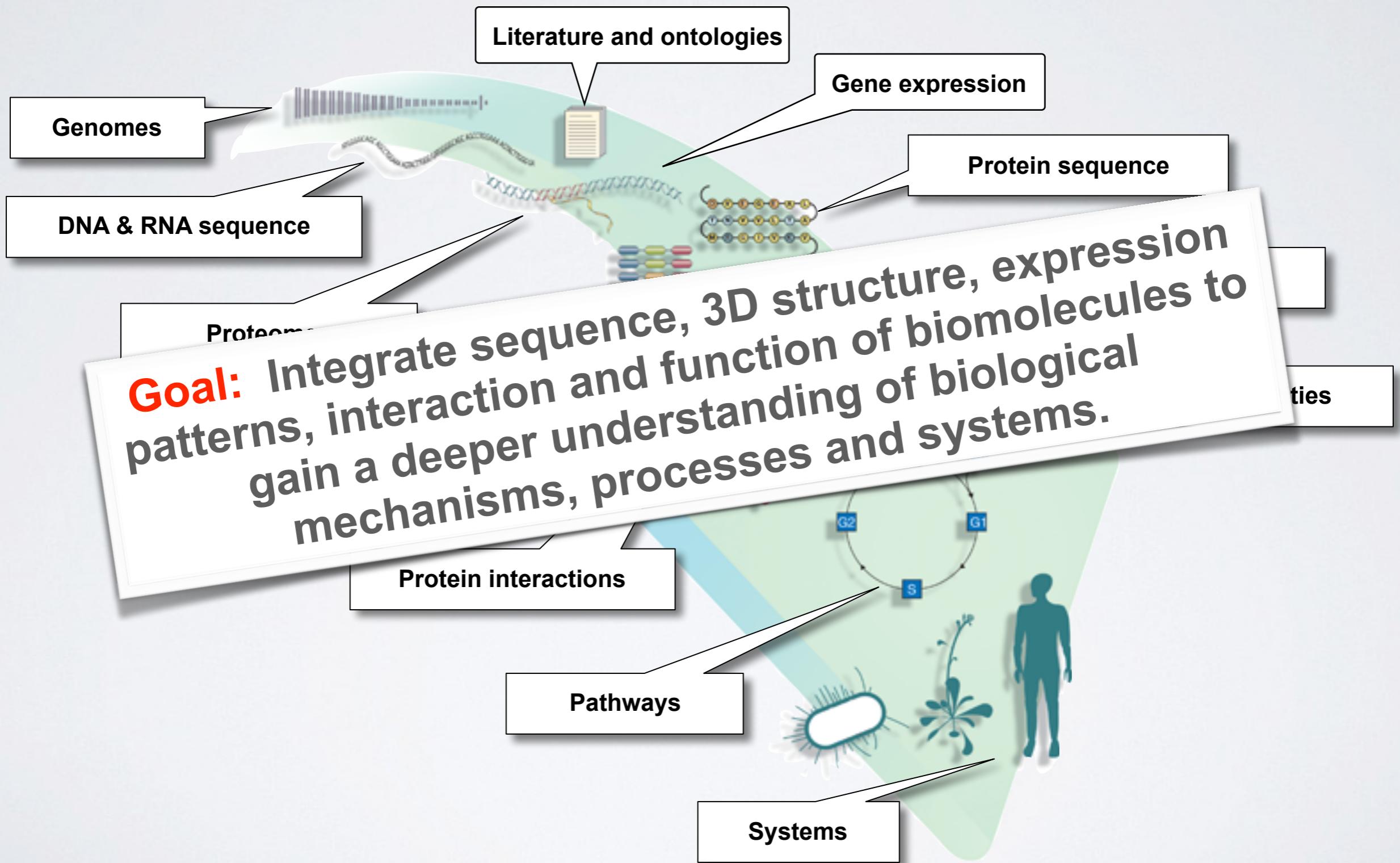
- "Computer based **management** and **analysis** of biological and biomedical data with useful applications in many disciplines, particularly **genomics**, **proteomics**, **metabolomics**, and related fields."  
(BIMM-143)
- "Bioinformatics is conceptualizing biology at the level of **macromolecules** and then applying "informatics" techniques (derived from disciplines such as applied mathematics, computer science, and statistics) to **understand** and **organize** the information associated with these molecules at the **bio-scale**."  
(Luscombe et al., 2001)
- "Bioinformatics is research, development, or application of **computational approaches** for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize and **analyze** such data ...<cut>..."  
(National Institutes of Health: <http://tinyurl.com/l3gxr6b>)

**Key Point:** Bioinformatics is Computer Aided Biology

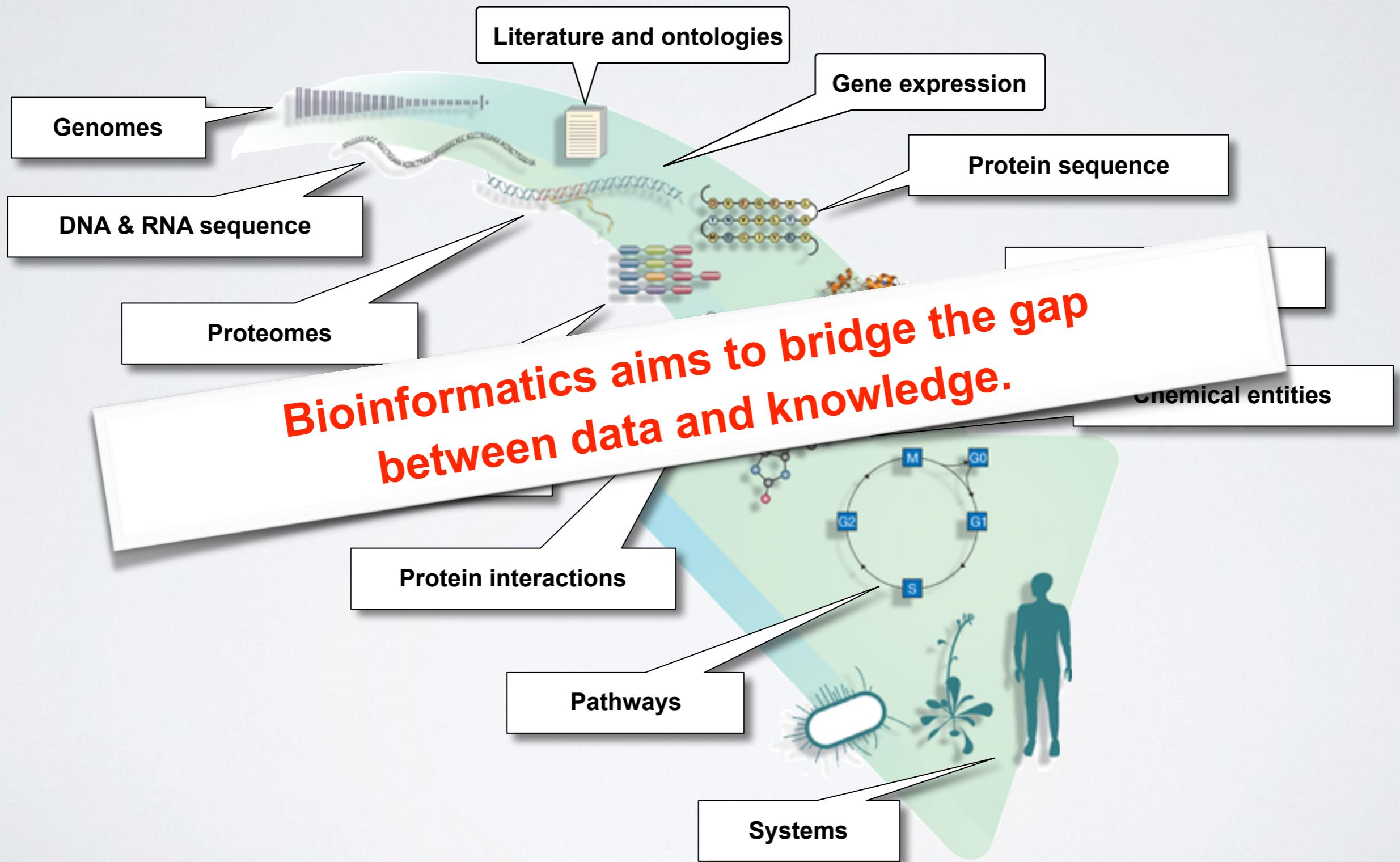
# Major types of Bioinformatics Data



# Major types of Bioinformatics Data

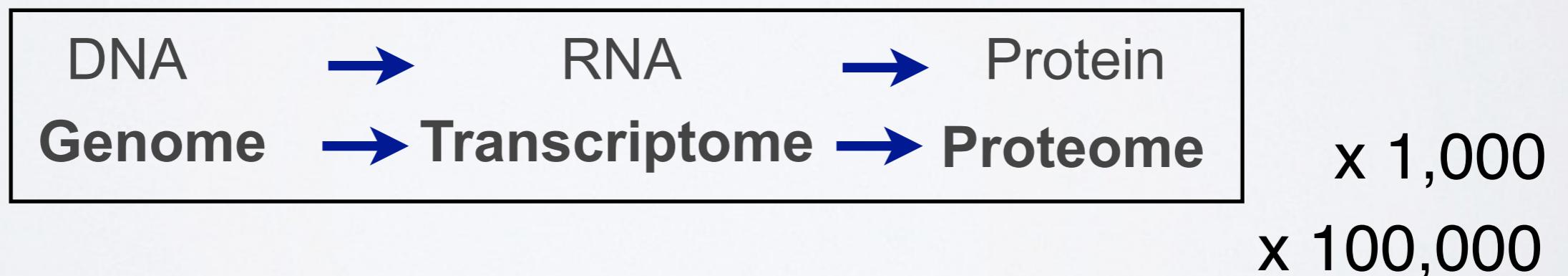


# Major types of Bioinformatics Data



# How do we do Bioinformatics?

- A “*bioinformatics approach*” involves the application of **computer algorithms**, **computer models** and **computer databases** with the broad goal of understanding the action of both individual genes, transcripts, proteins and large collections of these entities.



# How do we *actually* do Bioinformatics?

## Pre-packaged tools and databases

- Many online
- Most are free to use
- Time consuming methods require downloading...

## Advanced tool application & development

- Mostly on a UNIX environment
- Knowledge of programing languages frequently required  
(e.g. R, Python, Perl, C, Java, Fortran)
- May require specialized high performance computing...

# How do we *actually* do Bioinformatics?

## Pre-packaged tools and databases

- Many online
- Most are free to use
- Time consuming methods require downloading...

## Advanced tool application & development

- Mostly on a UNIX environment
- Knowledge of programing languages frequently required  
(e.g. R, Python, Perl, C, Java, Fortran)
- May require specialized high performance computing...

# NSF Extreme Science and Engineering Discovery Environment (XSEDE)

The screenshot shows a web browser window with the URL [www.xsede.org/community-engagement/educator-programs](http://www.xsede.org/community-engagement/educator-programs). The page features a dark blue header with the XSEDE logo and navigation links for Home, Gmail, Gcal, Bitbucket, GitHub, News, and Disqus. A secondary navigation bar includes About, For Users, Ecosystem, Community Engagement (which is currently selected), News, XUP, and a search icon. The main content area has a background image of a star-filled galaxy. A large white title 'Curriculum and Educator Programs' is centered above a black sidebar. The sidebar contains the text 'XSEDE pursues innovation and collaboration in computational science education.' and a section titled 'Campus Visits' with descriptive text. To the right of the sidebar is a 'Key Points' section with a bulleted list and a 'Related Links' section with several blue hyperlinks.

XSEDE pursues innovation and collaboration in computational science education.

## Campus Visits

XSEDE campus visits emphasize the need for computational science education and offer guidance concerning course content.

Campus visits bring together faculty, students, and administrators to discuss the importance of having a workforce that is ready to use modeling and simulation, advanced data analysis, and visualization to explore problems in science and engineering, in both academic and non-academic settings.

A typical campus visit consists of a general presentation affirming the essentiality of computational science education and suggesting approaches to inserting the appropriate content into the curriculum. Discussions are held with faculty and administrators about the current curriculum. Some visits are also combined with a half-day workshop on

*XSEDE campus visits emphasize the need for computational science education and offer guidance concerning course content*

**Key Points**

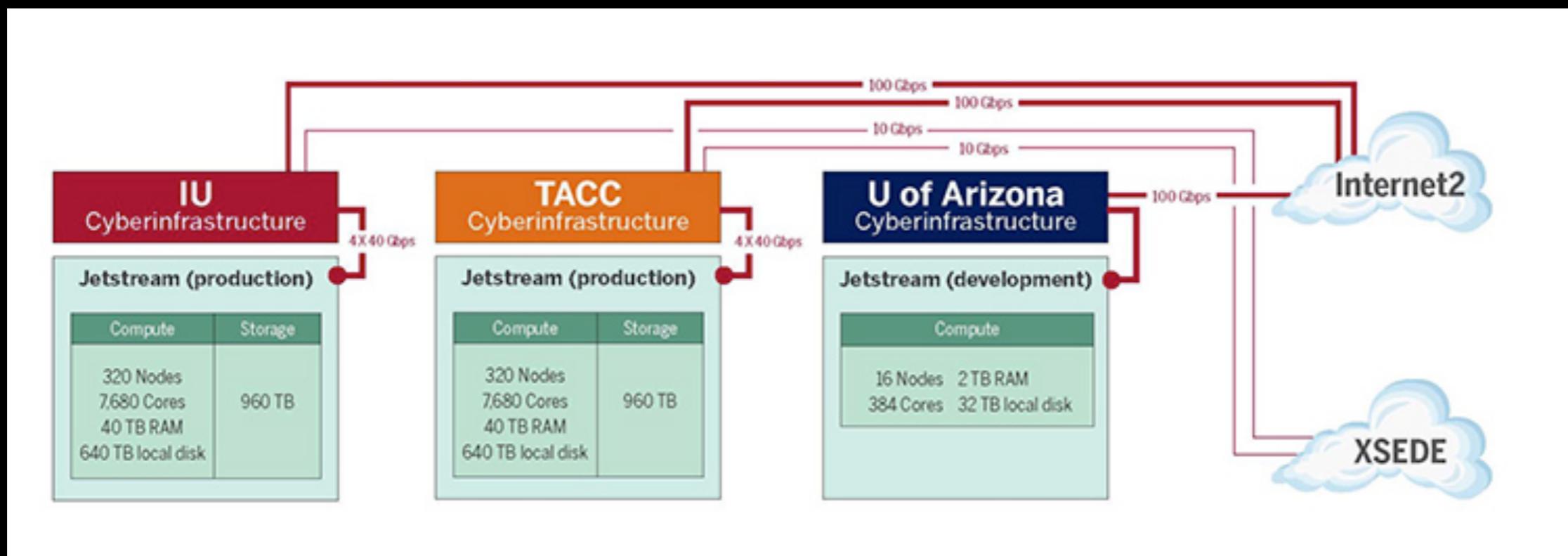
- XSEDE sponsors full-semester online courses
- Collaborations with faculty at participating institutions
- Campus visits offer guidance concerning course content

**Related Links**

[Diversity and Inclusion](#)  
[Student Engagement](#)  
[Campus Champions](#)  
[XSEDE Scholars Program](#)

# What is *Jetstream*?

- A new cloud computing environment based at Indiana University and the Texas Advanced Computing Center (TACC) providing on-demand access to interactive computing and data analysis resources.



# Skepticism & Bioinformatics

We have to approach computational results the same way we do wet-lab results:

- Do they make sense?
- Is it what we expected?
- Do we have adequate controls, and how did they come out?
- Modeling is modeling, but biology is different...

*What does this model actually contribute?*

- Avoid the miss-use of ‘black boxes’

# Skepticism & Bioinformatics

Gunnar von Heijne in “*Sequence Analysis in Molecular Biology*” states:

→“Think about what you’re doing; use your knowledge of the molecular system involved to guide both your interpretation of results and your direction of inquiry; use as much information as possible; and do not blindly accept everything the computer offers you”.

**Key-Point: Avoid the miss-use of ‘black boxes’!**

# Common problems with Bioinformatics

Confusing multitude of tools available

- ▶ Each with many options and settable parameters

Most tools and databases are written by and for nerds

- ▶ Same is true of documentation - if any exists!

Most are developed independently

Notable exceptions are found at the:

- **EBI** (European Bioinformatics Institute) and
- **NCBI** (National Center for Biotechnology Information)

### General Parameters

|                                                                             |                                                                                                                 |
|-----------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------|
| Max target sequences                                                        | 500                                                                                                             |
| Select the maximum number of aligned sequences to display <a href="#">?</a> |                                                                                                                 |
| Short queries                                                               | <input checked="" type="checkbox"/> Automatically adjust parameters for short input sequences <a href="#">?</a> |
| Expect threshold                                                            | 10                                                                                                              |
| Word size                                                                   | 3                                                                                                               |
| Max matches in a query range                                                | 0                                                                                                               |

### Scoring Parameters

|                           |                               |
|---------------------------|-------------------------------|
| Matrix                    | BLOSUM62                      |
| Gap Costs                 | Existence: 11 Extension: 1    |
| Compositional adjustments | Conditional compositional sco |

### Filters and Masking

|        |                                                                                                                           |
|--------|---------------------------------------------------------------------------------------------------------------------------|
| Filter | <input type="checkbox"/> Low complexity regions <a href="#">?</a>                                                         |
| Mask   | <input type="checkbox"/> Mask for lookup table only<br><input type="checkbox"/> Mask lower case letters <a href="#">?</a> |

### PSI/PHI/DELTA BLAST

|                      |                              |
|----------------------|------------------------------|
| Upload PSSM Optional | Choose File no file selected |
| PSI-BLAST Threshold  | 0.005                        |
| Pseudocount          | 0                            |

Even Blast has many settable parameters

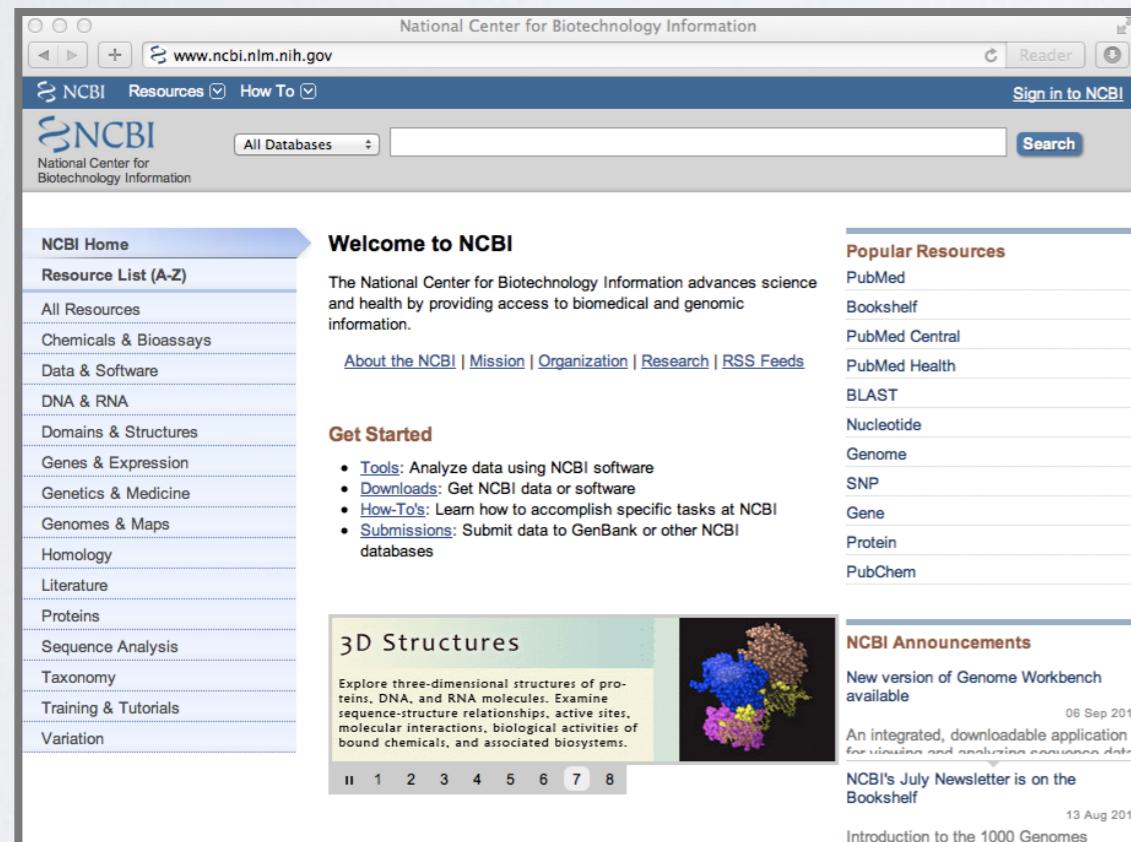
STEP 3 - Set your PROGRAM FASTA

|                      |            |                |                       |                         |                         |
|----------------------|------------|----------------|-----------------------|-------------------------|-------------------------|
| MATRIX               | GAP OPEN   | GAP EXTEND     | KTUP                  | EXPECTATION UPPER VALUE | EXPECTATION LOWER VALUE |
| BLOSUM50             | -10        | -2             | 2                     | 10                      | 0 (default)             |
| DNA STRAND           | HISTOGRAM  | FILTER         | STATISTICAL ESTIMATES |                         |                         |
| N/A                  | no         | none           | Regress               |                         |                         |
| SCORES               | ALIGNMENTS | SEQUENCE RANGE | DATABASE RANGE        |                         | MULTI HSPs              |
| 50                   | 50         | START-END      | START-END             | no                      |                         |
| SCORE FORMAT Default |            |                |                       |                         |                         |

Related tools with different terminology

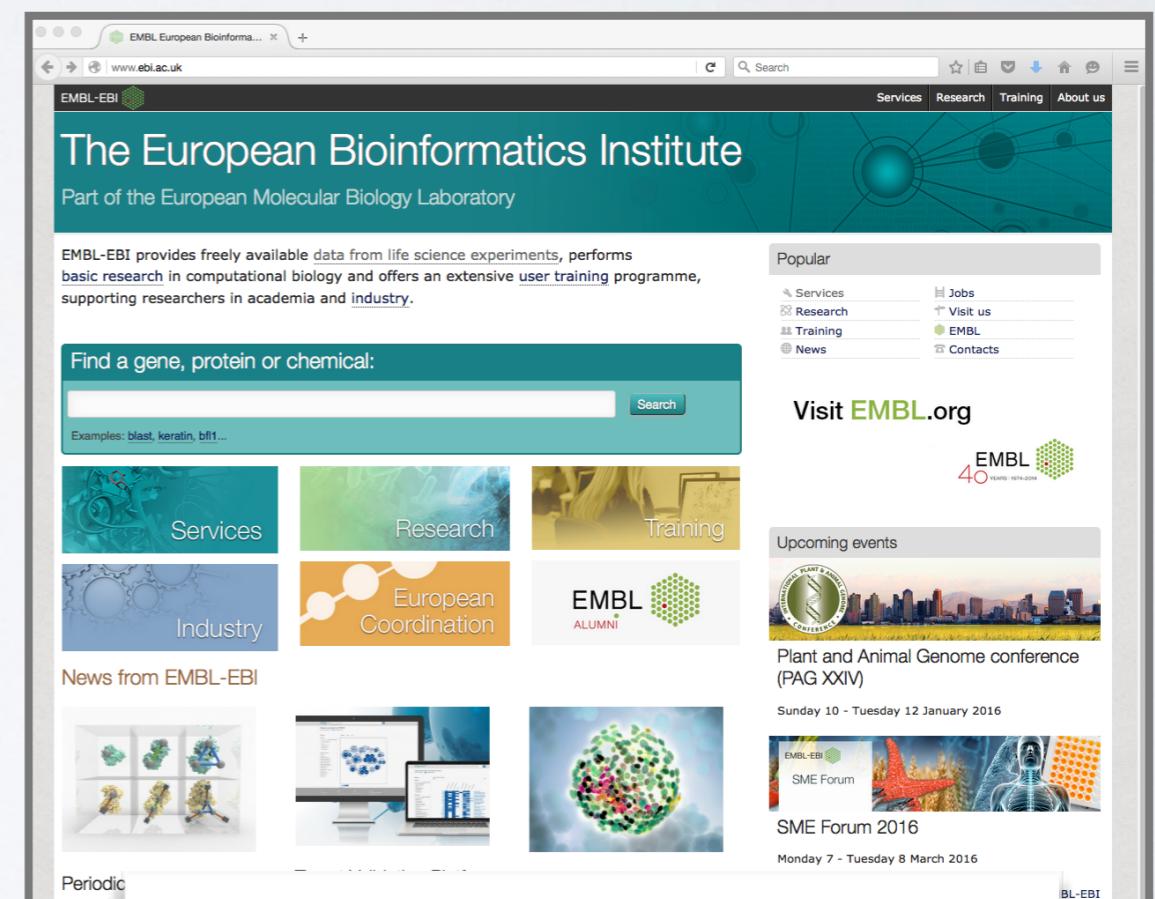
# Key Online Bioinformatics Resources: NCBI & EBI

The NCBI and EBI are invaluable, publicly available resources for biomedical research



The screenshot shows the homepage of the National Center for Biotechnology Information (NCBI). The top navigation bar includes links for 'Resources' (with a dropdown menu), 'How To', 'Sign in to NCBI', and a search bar. The main content area features a 'Welcome to NCBI' section, a 'Get Started' section with links to tools, downloads, and how-to guides, and a '3D Structures' section showing a molecular model. On the right, there's a sidebar titled 'Popular Resources' listing links to PubMed, Bookshelf, PubMed Central, and other databases. A 'NCBI Announcements' section highlights the new version of the Genome Workbench.

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



The screenshot shows the homepage of the European Bioinformatics Institute (EMBL-EBI). The top navigation bar includes links for 'Services', 'Research', 'Training', and 'About us'. The main content area features a large search bar for finding genes, proteins, or chemicals, and sections for 'Services', 'Research', 'Training', 'Industry', 'European Coordination', and 'EMBL ALUMNI'. A 'News from EMBL-EBI' section shows images of scientific work. On the right, there's a sidebar titled 'Popular' with links to 'Services', 'Research', 'Training', and 'News', and a 'Visit EMBL.org' section with the EMBL logo and a '40 years' anniversary graphic. Below this are sections for 'Upcoming events' (Plant and Animal Genome conference) and 'SME Forum 2016'.

<https://www.ebi.ac.uk>

# National Center for Biotechnology Information (NCBI)

- Created in 1988 as a part of the National Library of Medicine (NLM) at the National Institutes of Health
- NCBI's mission includes:
  - ▶ Establish **public databases**
  - ▶ Develop **software tools**
  - ▶ **Education** on and dissemination of biomedical information
- We will cover a number of core NCBI databases and software tools in this class



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

National Center for Biotechnology Information

NCBI Resources How To Sign in to NCBI

All Databases Search

NCBI Home Resource List (A-Z) All Resources Chemicals & Bioassays Data & Software DNA & RNA Domains & Structures Genes & Expression Genetics & Medicine Genomes & Maps Homology Literature Proteins Sequence Analysis Taxonomy Training & Tutorials Variation

## Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

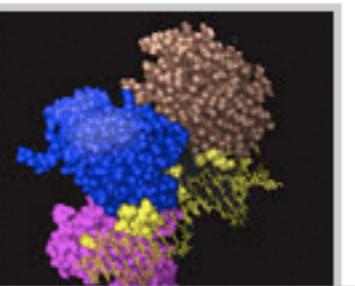
[About the NCBI](#) | [Mission](#) | [Organization](#) | [Research](#) | [RSS Feeds](#)

### Get Started

- [Tools](#): Analyze data using NCBI software
- [Downloads](#): Get NCBI data or software
- [How-To's](#): Learn how to accomplish specific tasks at NCBI
- [Submissions](#): Submit data to GenBank or other NCBI databases

### 3D Structures

Explore three-dimensional structures of proteins, DNA, and RNA molecules. Examine sequence-structure relationships, active sites, molecular interactions, biological activities of bound chemicals, and associated biosystems.



### Popular Resources

PubMed  
Bookshelf  
PubMed Central  
PubMed Health  
BLAST  
Nucleotide  
Genome  
SNP  
Gene  
Protein  
PubChem

### NCBI Announcements

New version of Genome Workbench available 06 Sep  
An integrated, downloadable applicati

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

National Center for Biotechnology Information

NCBI Resources How To Sign in to NCBI

All Databases

Search

Popular Resources

- PubMed
- Bookshelf
- PubMed Central
- PubMed Health
- BLAST
- Nucleotide
- Genome
- SNP
- Gene
- Protein
- PubChem

Welcome to NCBI

The National Center for Biotechnology Information provides access to unique information, tools and resources in molecular biology, genetics and health by providing access to its databases and information.

About the NCBI | Mission | Our History

Get Started

- Tools: Analyze data using NCBI's bioinformatics tools
- Downloads: Get NCBI data files and software
- How-To's: Learn how to access and use NCBI's resources
- Submissions: Submit data to NCBI's databases

3D Structures

Explore three-dimensional structures of proteins, DNA, and RNA molecules. Examine sequence-structure relationships, active sites, molecular interactions, biological activities of bound chemicals and associated biosystems.

New version of Genome Workbench available

06 Sep

An integrated, downloadable application

A screenshot of the NCBI homepage. The sidebar on the left lists various resources like NCBI Home, Resource List (A-Z), and 3D Structures. The main content area features a 'Welcome to NCBI' section and a 'Get Started' list. On the right, there's a 'Popular Resources' sidebar with links to PubMed, Bookshelf, PubMed Central, PubMed Health, BLAST, Nucleotide, Genome, SNP, Gene, Protein, and PubChem. A red bracket on the right side of the sidebar groups the links for BLAST, SNP, Gene, Protein, and PubChem, with red arrows pointing from the bracket to each of these five links.

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

National Center for Biotechnology Information

NCBI Resources How To Sign in to NCBI

All Databases Search

NCBI Home Resource List (A-Z)

Welcome to NCBI  
The National Center for Biotechnology Information advances science

Popular Resources PubMed

Notable NCBI databases include:  
**GenBank**, **RefSeq**, **PubMed**, **dbSNP**

and the search tools **ENTREZ** and **BLAST**

Homology Literature Proteins Sequence Analysis Taxonomy Training & Tutorials Variation

databases

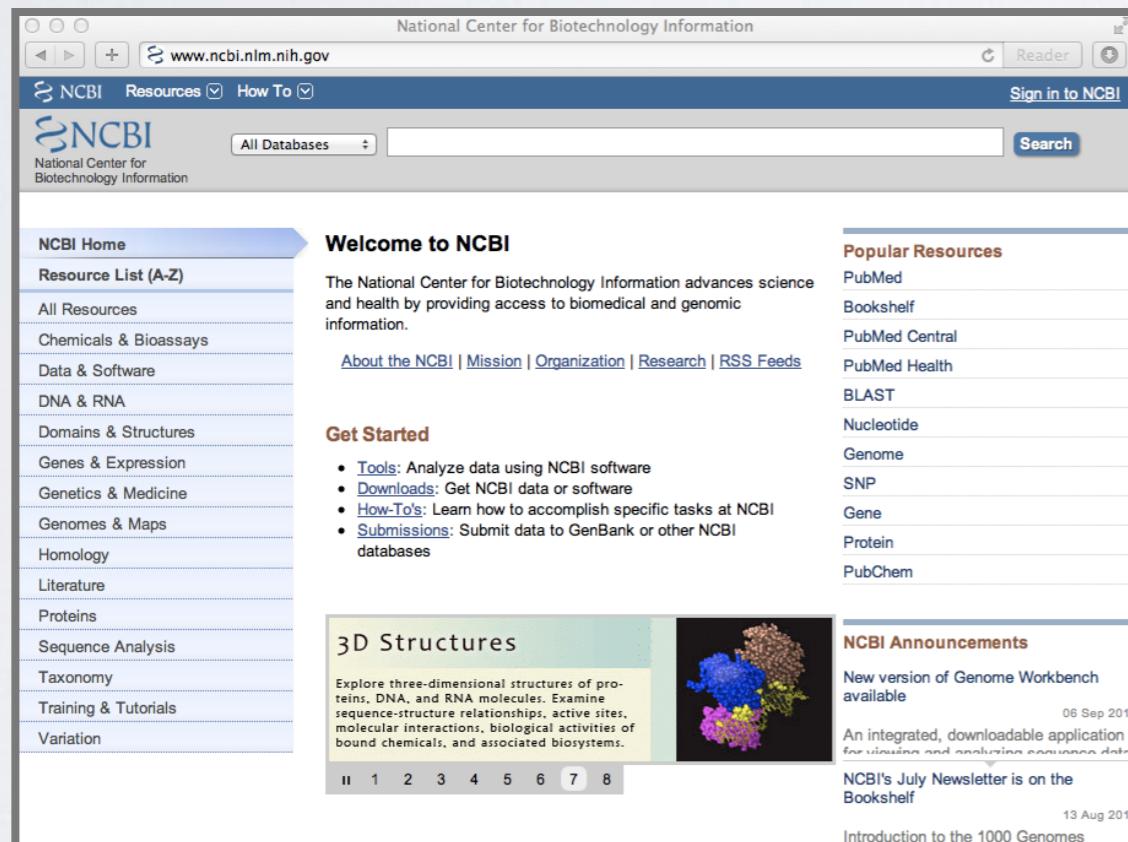
3D Structures  
Explore three-dimensional structures of proteins, DNA, and RNA molecules. Examine sequence-structure relationships, active sites, molecular interactions, biological activities of bound chemicals and associated biosystems

Protein PubChem

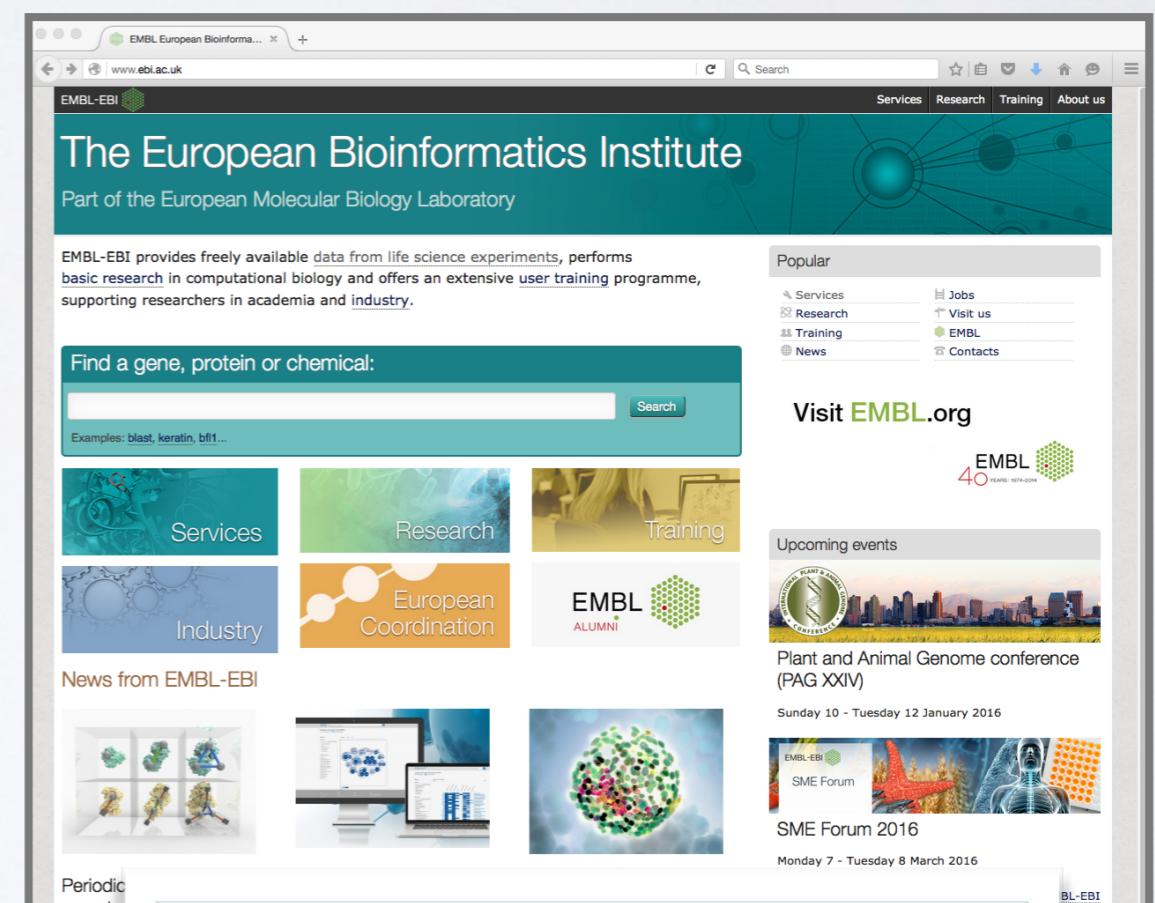
NCBI Announcements  
New version of Genome Workbench available 06 Sep  
An integrated, downloadable applicati

# Key Online Bioinformatics Resources: NCBI & EBI

The NCBI and EBI are invaluable, publicly available resources for biomedical research



The screenshot shows the NCBI homepage with a blue header bar containing the NCBI logo, a search bar, and links for "Resources", "How To", and "Sign in to NCBI". Below the header is a navigation menu with links to "NCBI Home", "Resource List (A-Z)", "All Resources", "Chemicals & Bioassays", "Data & Software", "DNA & RNA", "Domains & Structures", "Genes & Expression", "Genetics & Medicine", "Genomes & Maps", "Homology", "Literature", "Proteins", "Sequence Analysis", "Taxonomy", "Training & Tutorials", and "Variation". A "Popular Resources" sidebar on the right lists "PubMed", "Bookshelf", "PubMed Central", "PubMed Health", "BLAST", "Nucleotide", "Genome", "SNP", "Gene", "Protein", and "PubChem". The main content area features a "Welcome to NCBI" section, a "Get Started" section with links to tools, downloads, how-to's, and submissions, and a "3D Structures" section showing a molecular model.



The screenshot shows the EMBL-EBI homepage with a teal header bar containing the EMBL-EBI logo, a search bar, and links for "Services", "Research", "Training", and "About us". Below the header is a main content area with a teal banner stating "The European Bioinformatics Institute Part of the European Molecular Biology Laboratory". It features a "Find a gene, protein or chemical:" search bar, several colored boxes for "Services", "Research", "Training", "Industry", "European Coordination", and "EMBL ALUMNI", and a "News from EMBL-EBI" section with images of scientific data. On the right, there's a "Popular" sidebar with links to "Services", "Research", "Training", and "News", and a "Visit EMBL.org" section with the EMBL 40th anniversary logo and information about the Plant and Animal Genome conference (PAG XXIV) and the SME Forum 2016.

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

<https://www.ebi.ac.uk>

# European Bioinformatics Institute (EBI)

- Created in 1997 as a part of the European Molecular Biology Laboratory (EMBL)
- EBI's mission includes:
  - ▶ providing freely available **data and bioinformatics services**
  - ▶ and providing advanced **bioinformatics training**
- We will cover a number of EBI databases and tools that have advantages over those offered at NCBI



# The EBI maintains a number of high quality curated **secondary databases** and associated tools

The screenshot shows the homepage of the EMBL European Bioinformatics Institute (EBI) at [www.ebi.ac.uk](http://www.ebi.ac.uk). The page features a dark blue header with the EMBL-EBI logo and navigation links for Services, Research, Training, and About us. Below the header is a teal banner with the text "The European Bioinformatics Institute" and "Part of the European Molecular Biology Laboratory". A search bar is located above a main content area. The content area includes a section about EMBL-EBI's mission, a search bar for finding genes, proteins, or chemicals, and several promotional boxes for Services, Research, Training, Industry, European Coordination, and EMBL ALUMNI. A sidebar on the right lists Popular links: Services, Research, Training, News, Jobs, Visit us, EMBL, and Contacts. At the bottom, there is information about an upcoming event: the Plant and Animal Genome conference (PAG XXIV) from Sunday 10 to Tuesday 12 January 2016.

EMBL European Bioinforma... [www.ebi.ac.uk](#) Search Services Research Training About us

# The European Bioinformatics Institute

Part of the European Molecular Biology Laboratory

EMBL-EBI provides freely available data from life science experiments, performs basic research in computational biology and offers an extensive user training programme, supporting researchers in academia and industry.

Find a gene, protein or chemical:

Examples: blast, keratin, bfl1...

Search

Services

Research

Training

Industry

European Coordination

EMBL ALUMNI

Popular

- Services
- Research
- Training
- News
- Jobs
- Visit us
- EMBL
- Contacts

Visit [EMBL.org](#)

EMBL 40 YEARS 1974-2014

Upcoming events

INTERNATIONAL PLANT & ANIMAL GENOME CONFERENCE

Plant and Animal Genome conference (PAG XXIV)

Sunday 10 - Tuesday 12 January 2016

News from EMBL-EBI

The EBI also provides a growing selection of **online tutorials** on EBI databases and tools

A screenshot of a web browser displaying the EBI Train online website. The title bar reads "Train online | EBI Train online". The address bar shows the URL "www.ebi.ac.uk/training/online/". The page header includes the EMBL-EBI logo, a search bar, and links for "Find", "Help", and "Feedback". A red "Beta" badge is visible in the top right corner. The main menu bar has links for "Databases", "Tools", "Research", "Training", "Industry", "About Us", and "Help". A secondary navigation bar on the left is titled "Navigation" and includes a link to "Train online Home". The main content area features a large heading "Notable EBI databases include:" followed by a list of databases: ENA, UniProt, Ensembl.

Notable EBI databases include:  
**ENA**, **UniProt**, **Ensembl**

and the tools **FASTA**, **BLAST**, **InterProScan**,  
**MUSCLE**, **DALI**, **HMMER**

#### Find a course

##### Browse by subject



[Genes and Genomes](#)



[Gene Expression](#)



[Interactions, Pathways, and Networks](#)

# Bioinformatics Databases

AATDB, AceDb, ACUTS, ADB, AFDB, AGIS, AMSdb, ARR, AsDb, BBDB, BCGD, Beanref, BiolImage, BioMagResBank, BIOMDB, BLOCKS, BovGBASE, BOVMAP, BSORF, BTKbase, CANSITE, CarbBank, CARBHYD, CATH, CAZY, CCDC, CD4OLbase, CGAP, ChickGBASE, Colibri, COPE, CottonDB, CSNDB, CUTG, CyanoBase, dbCFC, dbEST, dbSTS, DDBJ, DGP, DictyDb, Picty\_cDB, DIP, DOGS, DOMO, DPD, DPInteract, ECDC, ECGC, EC02DBASE, EcoCyc, EcoGene, EMBL, EMD db, ENZYME, EPD, EpoDB, ESTHER, FlyBase, FlyView, GCRDB, GDB, GENATLAS, Genbank, GeneCards, Genlilesne, GenLink, GENOTK, GenProtEC, GIFTS, GPCRDB, GRAP, GRBase, gRNAsdb, GRR, GSDB, HAEMB, HAMSTERS, HEART-2DPAGE, HEXAdb, HGMD, HIDB, HIDC, HIVdb, HotMolecBase, HOVERGEN, HPDB, HSC-2DPAGE, ICN, ICTVDB, IL2RGbase, IMGT, Kabat, KDNA, KEGG, KloTho, LGIC, MAD, MaizeDb, MDB, Medline, Mendel, MEROPS, MGDB, MGI, MHCPEP5 Micado, MitoDat, MITOMAP, MJDB, MmtDB, Mol-R-Us, MPDB, MRR, MutBase, MycDB, NDB, NRSub, O-lycBase, OMIA, OMIM, OPD, ORDB, OWL, PAHdb, PatBase, PDB, PDD, Pfam, PhosphoBase, PigBASE, PIR, PKR, PMD, PPDB, PRESAGE, PRINTS, ProDom, Prolysis, PROSITE, PROTOMAP, RatMAP, RDP, REBASE, RGP, SBASE, SCOP, SeqAnaiRef, SGD, SGP, SheepMap, Soybase, SPAD, SRNA db, SRPDB, STACK, StyGene, Sub2D, SubtiList, SWISS-2DPAGE, SWISS-3DIMAGE, SWISS-MODEL Repository, SWISS-PROT, TeIDB, TGN, tmRDB, TOPS, TRANSFAC, TRR, UniGene, URNADB, V BASE, VDRR, VectorDB, WDCM, WIT, WormPep, etc ..!!!!

# Bioinformatics Databases

AATDB, AceDb, ACUTS, ADB, AFDB, AGIS, AMSdb, ARR, AsDb, BBDB, RCCP, Beanref, TKbase, CANSITE, CarbBank, CARBHYD, CATH, CAZy, ChickGBASE, Colibri, COPE, CottonDB, dbSTS, DDBJ, DGP, DictyDb, ECGC, EC02DBASE, FlyBase, GDB, HEPDB, KEGG, MHCDB, MycoDB, PDBe, PDB, Pfam, PhosphoBase, PigBASE, PIR, PKR, PMD, PPDB, PRESAGE, PRINTS, ProDom, Prolysis, PROSITE, PROTOMAP, RatMAP, RDP, REBASE, RGP, SBASE, SCOP, SeqAnaiRef, SGD, SGP, SheepMap, Soybase, SPAD, SRNA db, SRPDB, STACK, StyGene, Sub2D, SubtiList, SWISS-2DPAGE, SWISS-3DIMAGE, SWISS-MODEL Repository, SWISS-PROT, TelDB, TGN, tmRDB, TOPS, TRANSFAC, TRR, UniGene, URNADB, V BASE, VDRR, VectorDB, WDCM, WIT, WormPep, etc ..!!!!

**There are lots of Bioinformatics Databases**

For a annotated listing of major bioinformatics databases please see the online handout

< Major Databases.pdf >

# Side-note: Databases come in all shapes and sizes



Databases can be of variable quality and often there are multiple databases with overlapping content.

# Today's Menu

## Course Logistics

Website, screencasts, survey, ethics, assessment and grading.

## Learning Objectives

What you need to learn to succeed in this course.

## Course Structure

Major lecture topics and specific learning goals.

## Introduction to Bioinformatics

Introducing the *what, why and how* of bioinformatics?

## Bioinformatics Database

**Hands-on** exploration of several major databases and their associated tools.

Your Turn!

# Hands-on section

<http://thegrantlab.org/bimm143/>

The screenshot shows a web browser window with the URL [bioboot.github.io/bimm143\\_W18/lectures/#1](http://bioboot.github.io/bimm143_W18/lectures/#1). The page title is "1: Welcome to Foundations of Bioinformatics". The left sidebar has links for Overview, Lectures, Computer Setup, Learning Goals, Assignments & Grading, and Ethics Code. The "Lectures" link is highlighted with a red box. The main content area includes sections for Topics, Goals, and Material, each with a bulleted list of links.

**Topics:**

Course introduction, Learning goals & expectations, Biology is an information science, History of Bioinformatics, Types of data, Application areas and introduction to upcoming course segments, Student 30-second introductions, Student computer setup.

**Goals:**

- Understand course scope, expectations, logistics and [ethics code](#).
- Understand the increasing necessity for computation in modern life sciences research.
- Get introduced to how bioinformatics is practiced.
- Complete the [pre-course questionnaire](#).
- Setup your [laptop computer](#) for this course.

**Material:**

- Lecture Slides: [Large PDF](#), [Small PDF](#),
- Lab: [Hands-on section worksheet](#)
- Feedback: [Muddy Point Assessment](#)
- Handout: [Class Syllabus](#)
- Computer Setup Instructions.

## BIMM-143: INTRODUCTION TO BIOINFORMATICS (Lecture 1)

### **Bioinformatics Databases and Key Online Resources**

[https://bioboot.github.io/bimm143\\_W18/lectures/#1](https://bioboot.github.io/bimm143_W18/lectures/#1)

Dr. Barry Grant

Jan 2018

**Overview:** The purpose of this lab session is to introduce a range of bioinformatics databases and associated services available on the Web whilst investigating the molecular basis of a common human disease.

Sections 1 and 2 deal with querying and searching GenBank, GENE and OMIM databases at NCBI. Sections 3 and 4 provide exposure to EBI resources for comparing proteins and visualizing protein structures. Finally, section 5 provides an opportunity to explore these and other databases further with additional examples.

**Side-note:** The Web is a dynamic environment, where information is constantly added and removed. Servers "go down", links change without warning, etc. This can lead to "broken" links and results not being returned from services. Don't give up - give it a second go and try a search engine using terms related to the page you are trying to access.

### **Section 1**

The following transcript was found to be abundant in a human patient's blood sample.

```
>example1
ATGGTGCATCTGACTCCTGTGGAGAAGTCTGCCGTTACTGCCCTGTGGGCAAGGTGAACGTGGATGAAG
TTGGTGGTGAGGCCCTGGGCAGGCTGCTGGTGGTCTACCCCTGGACCCAGAGGTTCTTGAGTCCTTGG
GGATCTGTCCACTCCTGATGCAGTTATGGGCAACCCCTAACGGTAAGGCTCATGGCAAGAAAGTGCTCGGT
GCCTTAGTGTGATGGCCTGGCTCACCTGGACAACCTCAAGGGCACCTTGCCACACTGAGTGAGCTGCACT
GTGACAAGCTGCACGTGGATCCTGAGAACTTCAGGCTCCTGGCAACGTGCTGGTCTGTGTGCTGGCCA
TCACTTGGCAAAGAATTCACCCCACCAGTGCAGGCTGCCTATCAGAAAGTGGTGGCTGGTGTGGCTAAT
GCCCTGGCCCACAAGTATCACTAAGCTCGCTTCTTGCTGTCCAATTT
```

The only information you are given is the above sequence so you must begin your investigation with a sequence search - for this example we will use NCBI's **BLAST** service at: <http://blast.ncbi.nlm.nih.gov/>

*Note that there are several different "basic BLAST" programs available at NCBI (including nucleotide BLAST, protein BLAST, and BLASTx).*

# YOUR TURN!

- There are five major hands-on sections including:
  1. BLAST, GenBank and OMIM @ **NCBI** [~35 mins]
  2. GENE database @ **NCBI** [~15 mins]  
— BREAK —
  3. UniProt & Muscle @ **EBI** [~25 mins]
  4. PFAM, PDB & NGL [~30 mins]  
— BREAK —
  5. Extension exercises [~30 mins]

- ▶ Please do answer the last review question (**Q19**).  
▶ We encourage discussion and exploration!

# SUMMARY

- Bioinformatics is computer aided biology.
- Bioinformatics deals with the collection, archiving, organization, and interpretation of a wide range of biological data.
- The NCBI and EBI are major online bioinformatics service providers.
- Introduced Gene, UniProt, PDB databases as well as a number of ‘boutique’ databases including PFAM and OMIM.

# HOMEWORK

<http://thegrantlab.org/bimm143/>

- Complete the initial course questionnaire:
- Check out the “background reading” material online:
- Complete the lecture 1 homework questions:

