



BGGN 213

Foundations of Bioinformatics

Barry Grant
UC San Diego

<http://thegrantlab.org/bggn213>

HELLO
my name is

BARRY

bjgrant@ucsd.edu

HELLO
HIS my name is

KEVIN

kkchau@ucsd.edu

Office Hours:
[SignUp](#)

Location:
TATA, #2501

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/bggn213/>

UC San Diego

BGNN 213

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/bggn213_S18/

**Bioinformatics
(BGNN 213, Spring 2018)**

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

Instructional Assistant
Yuansheng Zhou (Email: [yuz461@ucsd.edu](#))

Course Syllabus
[Spring 2018 \(PDF\)](#)

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 course is designed for bioscience graduate students and provides a hands-on introduction to the computer-based analysis of genomic and biomolecular data.

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

The screenshot shows a web browser window with the URL bioboot.github.io/bggm213_f17/goals/. The page is titled "Learning Goals". On the left, there is a sidebar for the course BGGN 213, which includes links for Overview, Lectures, Computer Setup, Learning Goals (which is highlighted with a red box), Assignments & Grading, Ethics Code, and Screen Cast Videos. Below the sidebar is a footer with social media icons for Twitter, GitHub, Email, and RSS.

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 UNIX command line and 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.

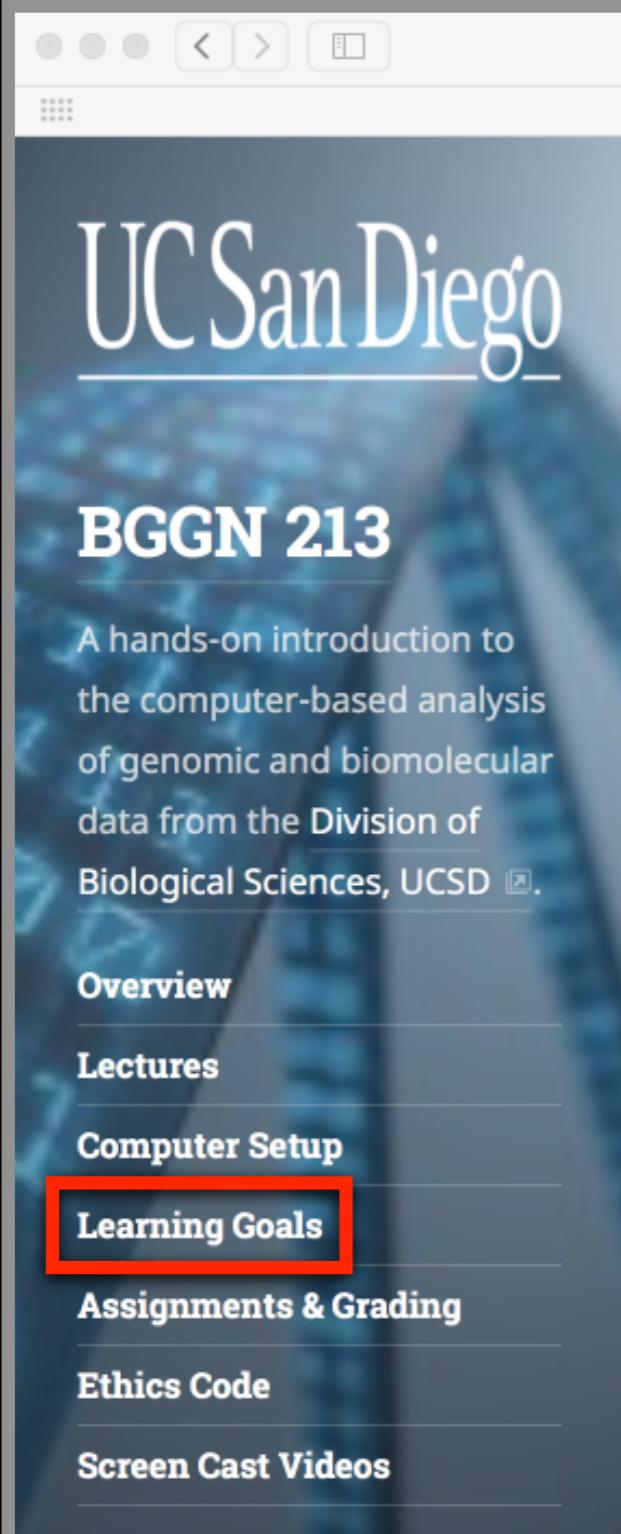
At the end of this course students will:

- 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/bggn213_f17/goals/. The page content is as follows:

UCSanDiego
BGGN 213
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 (highlighted with a red box)
Assignments & Grading
Ethics Code
Screen Cast Videos

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 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	4, 5

Course Structure

Derived from specific learning goals

The screenshot shows a web browser window with the URL bioboot.github.io/bggn213_S18/lectures/. The page title is "UCSanDiego BGGN 213". The left sidebar has links for Overview, Lectures (which is highlighted with a red box), Computer Setup, Learning Goals, Assignments & Grading, and Ethics Code. Social media icons for Twitter, GitHub, and RSS are at the bottom. The main content area has a heading "Lectures" and a table of lectures:

#	Date	Topics for Spring 2018
1	Wed, 04/04	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	Fri, 04/06	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

Course Structure

Derived from specific learning goals

bioboot.github.io/bggn213_S18/lectures/

Lectures

All Lectures are Wed/Fri 1:00-4:00 pm in Warren Lecture Hall 2015 (WLH 2015) ([Map](#)). Clicking on the class topics below will take you to corresponding lecture notes, homework assignments, pre-class video screen-casts and required reading material.

#	Date	Topics for Spring 2018
1	Wed, 04/04	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	Fri, 04/06	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

Class Details

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

The screenshot shows a web browser window with the URL bioboot.github.io/bggn213_f17/lectures/#1. The page content is as follows:

UCSanDiego
BGGN 213
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
Screen Cast Videos

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](#).
- Setup your laptop computer for this course.

Material:

- [Pre class screen cast](#),
- Lecture Slides: Large PDF, [Small PDF](#), (To be updated!)
- [Handout: Class Syllabus](#)
- Computer [Setup Instructions](#).

Homework

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

UCSanDiego

BGGN 213

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

Screen Cast Videos

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-21...)

BGGN 213

Foundations of Bioinformatics

Barry Grant

UC San Diego

<http://thegrantlab.org/bggn213>

1 Welcome to BGGN-213: Course introduction and logistics.

Homework

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

docs.google.com/forms/d/e/1FAIpQLSeN3pg-AaRg5la3PxZuqSj

Home Gmail Gcal Bitbucket GitHub News Disqus

BGGN213 Lecture 1 Homework

Please answer the following questions

* Required

Your name/email address *

Part of your UCSD email address before the '@ucsd.edu' part

Your answer

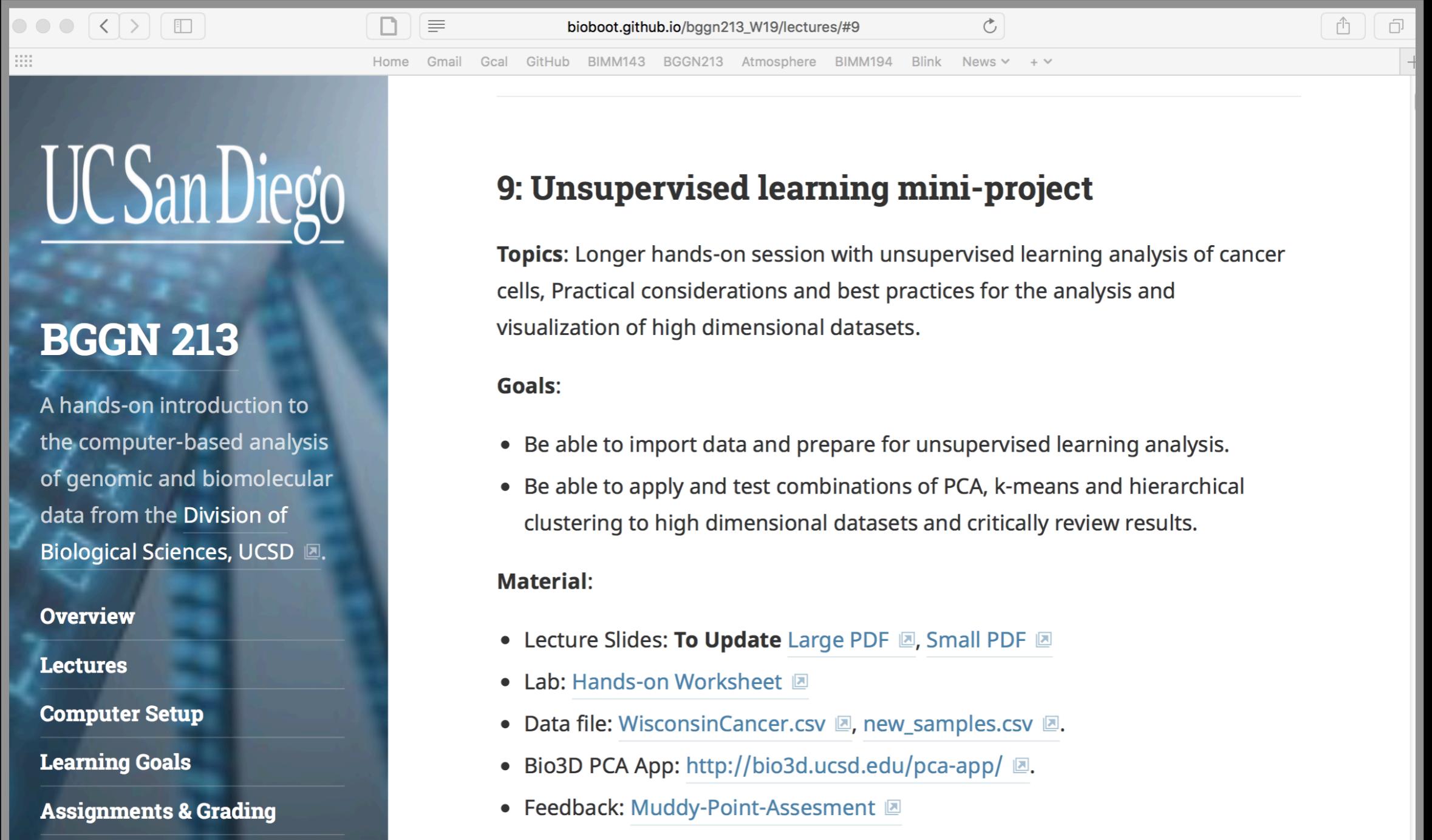
Which of the following operating systems is most frequently used for bioinformatics tool development

- Windows
- iOS
- Unix
- Perl

Which of the following databases contains primarily protein

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/bggm213_W19/lectures/#9. The page content is as follows:

UCSanDiego
BGGN 213
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

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: **To Update** [Large PDF](#) [Small PDF](#)
- Lab: [Hands-on Worksheet](#)
- Data file: [WisconsinCancer.csv](#) [new_samples.csv](#)
- Bio3D PCA App: <http://bio3d.ucsd.edu/pca-app/>
- Feedback: [Muddy-Point-Assesment](#)

Projects

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

The screenshot shows a web browser with two tabs open. The left tab is the course homepage for BGGN 213 at UC San Diego, featuring the title 'UCSanDiego BGGN 213' and a brief description of the course. The right tab is a specific lecture page titled '18: Cancer genomics'.

18: Cancer genomics

Topics: Cancer genomics resources and bioinformatics tools for investigating the molecular basis of cancer. Large scale cancer sequencing projects; NCI Genomic Data Commons; What has been learned from genome sequencing of cancer? **Immunoinformatics, immunotherapy and cancer;** Using genomics and bioinformatics to harness a patient's own immune system to fight cancer. Implications for the development of personalized medicine.

N.B. Find a gene assignment due before next class!

Material:

- Lecture Slides: [Large PDF](#), [Small PDF](#)
- Lab: **T0 UPDATE** [Hands-on Worksheet Part 1.](#)
- Lab: **T0 UPDATE** [Hands-on Worksheet Part 2.](#)
- Data files:
 - [lecture18_sequences.fa](#)

Projects

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 **Fri Feb 22nd (02/22/19)**.

The complete assignment, including responses to all questions, is due at the beginning of class **Wed March 13th (03/13/19)**.

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.

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

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 Content:** Bioinformatics Class 5
- Section Headers:**

class13

,

Jason Patrick Bennett

,

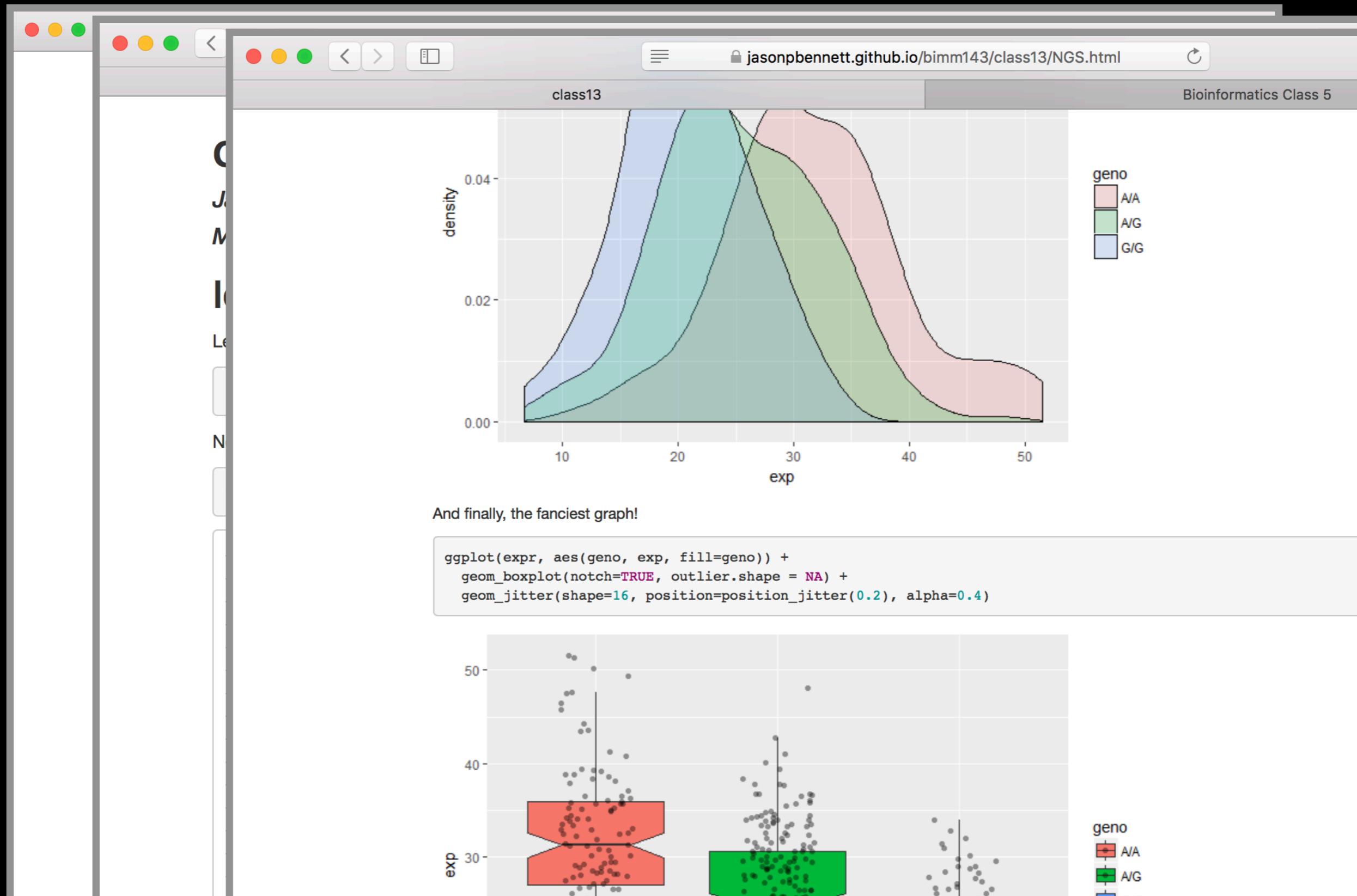
May 15, 2018

,

Identifying SNP's in a Population
- Text:** Lets analyze SNP's from the Mexican-American population in Los Angeles:
- Code Block:** genotype <- read.csv("373531-SampleGenotypes-Homo_sapiens_Variation_Sample_rs8067378.csv")
- Text:** Now lets look at a table of the data:
- Code Block:** table(genotype)
- Table Data:** A table showing SNP genotype data for 14 individuals (NA19648-NA19664) across four positions (A|A, A|G, G|A, G|G). The table includes headers for Population, Sex, and Genotype strand.

Population	Sex	Genotype..forward.strand.	A A	A G	G A	G G
ALL	(F)	NA19648	1	0	0	0
AMR	(M)	NA19649	0	0	0	1
MXL	(F)	NA19651	1	0	0	0
	(M)	NA19652	0	0	0	1
	(F)	NA19654	0	0	0	1
	(M)	NA19655	0	1	0	0
	(F)	NA19657	0	1	0	0
	(M)	NA19658	1	0	0	0
	(M)	NA19661	0	1	0	0
	(F)	NA19663	1	0	0	0
	(M)	NA19664	0	0	1	0
	(F)	NA19666	1	0	0	0

Online portfolio of **your** bioinformatics work!



Bonus:

Bioinformatics & Genomics in industry

The screenshot shows a web browser window with the URL bioboot.github.io/bggn213_W19/lectures/#21 in the address bar. The page content is as follows:

UCSanDiego
BGGN 213
A hands-on introduction to the computer-based analysis of genomic and biomolecular data from the Division of Biological Sciences, UCSD.

21: Bonus: Bioinformatics & Genomics in industry

Friday March 15th at 1pm come and enjoy a set of short open ended guest lectures from leading genomic scientists at Illumina Inc., Synthetic Genomics Inc., Samumed and the La Jolla Institute for Allergy and Immunology. Come prepared for networking and to have your questions about industry careers in Bioinformatics and Genomics answered.

© 2019 Barry J. Grant. All rights reserved. A UCSD Division of Biological Sciences Course

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 graduates in the biosciences with no programing experience or high level math skills.

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

BGGN-213 Learning Goals....

Advanced UNIX and R based learning goals

The screenshot shows a web browser window with the URL bioboot.github.io/bggn213_f17/goals/. The page is titled "BGGN 213". The sidebar on the left includes links for Overview, Lectures, Computer Setup, Learning Goals (which is highlighted with a red box), Assignments & Grading, Ethics Code, and Screen Cast Videos. The main content area displays 12 learning goals, each with a number, a description, and a list of associated numbers. Goals 6, 7, 8, and 9 are grouped together and highlighted with a green border.

Goal	Description	Associated Numbers
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 UNIX command-line tools for file system navigation and text file manipulation.	6, 7, 10, 11, 24, 15
7	Use existing programs at the UNIX command line to analyze bioinformatics data.	7, 10, 11, 13, 14, 15, 16
8	Use R to read and parse comma-separated (.csv) formatted files ready for subsequent analysis.	8, 9, 10, 11, 13, 15, 16
9	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
10	View and interpret the structural models in the PDB.	10, 11
11	Explain the outputs from structure prediction algorithms and small molecule docking approaches.	11
12	Appreciate and describe in general terms the rapid advances in sequencing technologies and the new areas of investigation that have arisen from these advances.	13, 14, 15

BGGN-213 Learning Goals....

Delve deeper into “real-world” bioinformatics

UCSanDiego

BGGN 213

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

Screen Cast Videos

bioboot.github.io/bggn213_f17/goals/

13	sequenced and the bioinformatics processing and analysis required for their interpretation.	13
14	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
15	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
16	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
17	Use the KEGG pathway database to look up interaction pathways.	17
18	Use graph theory to represent biological data networks.	17, 18
19	Understand the challenges in integrating and interpreting large heterogenous high throughput data sets into their functional context.	19
20	Have an appreciation for the social impacts and ethical implications of how genomic sequence information is used in our society	20

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 UNIX and 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

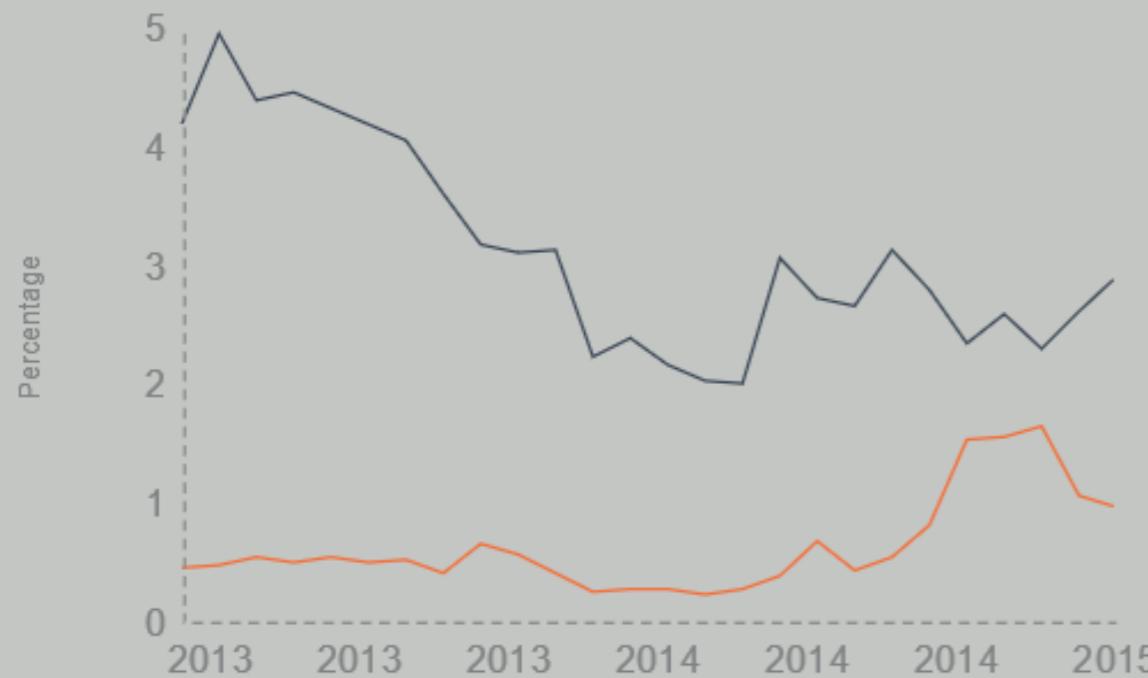
Language Rank	Types	Spectrum Ranking
1. C		100.0
2. Java		98.1
3. Python		98.0
4. C++		95.9
5. R		87.9
6. C#		86.7
7. PHP		82.8
8. JavaScript		82.2
9. Ruby		74.5
10. Go		71.9

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

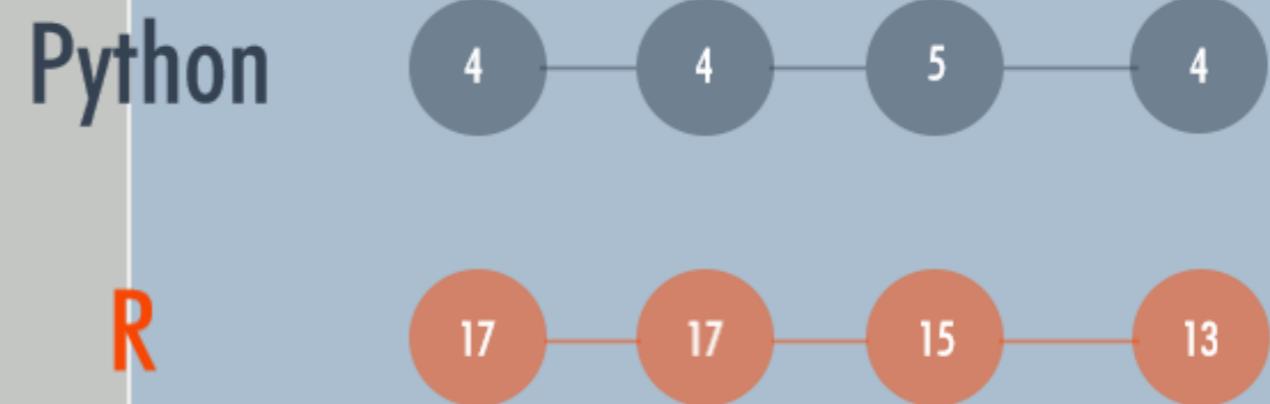
R and Python: The Numbers

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 the “lingua franca” of data science in industry and academia and was 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** - 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. At the bottom is a large red-outlined "Submit Answer" button. To the right is a screenshot of the RStudio IDE interface, showing the console output of an R session, the environment pane, and the files pane.

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

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 URL in the address bar is "https://campus.datacamp.com/courses/working-with-the-rstudio-ide-part-1/orientation?ex=2". The DataCamp logo is visible in the top left of the browser. The main content area displays a course exercise titled "What is an IDE anyway?". The exercise has been completed, indicated by the text "Exercise Completed" and a blue button with "50xp" (experience points). A red circle highlights this button. Below the completion message, there is a note: "Nice job! Move onto the next video to start learning more about the RStudio IDE!". A "Continue" button is also highlighted with a red circle. On the right side of the screen, the RStudio interface is visible, showing the console output of R version 3.3.1, the environment pane, and the files pane. The files pane shows a single item named "Home".

< 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'. Inside this section, there's a text block about performing PCA on normalized counts and a code editor window.

Exercise Content:

To continue with the quality assessment of our samples, in the first part of this exercise, we will perform PCA to look how our samples cluster and whether our condition of interest corresponds with the principal components explaining the most variation in the data. In the second part, we will answer questions about the PCA plot.

To assess the similarity of the `smoc2` samples using PCA, we need to transform the normalized counts then perform the PCA analysis. Assume all libraries have been loaded, the DESeq2 object created, and the size factors have been stored in the DESeq2 object,

```
dds_smoc2 .
```

Instructions: 1/2 50 XP

- Run the code to transform the normalized counts.
- 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.

Code Editor:

script.R RDocumentation

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

Run Code Submit Answer

R Console:

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

< 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 page title is "Foundations of Bioinformatics (BGGN-213)". The top navigation bar includes links for "Learn", "Groups" (which is highlighted with a red circle), and "About". A user profile icon indicates 1,550 XP. Below the title, there are two tabs: "Leaderboard" and "My Assignments", with "Leaderboard" selected. Under the "Leaderboard" tab, there are three filter options: "30 Days", "90 Days", and "Last Year". A table displays the top 8 members of the group, showing their rank, name, XP, Courses completed, and Chapters completed.

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?

Computer Setup

Ensuring your laptop is all set for future sections of this course.

Q. What is Bioinformatics?

Q. What is Bioinformatics?

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

... Bioinformatics is a hybrid of biology and computer science

Q. What is Bioinformatics?

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

- ... Bioinformatics is a hybrid of biology and computer science
- ... **Bioinformatics is computer aided biology!**

Q. What is Bioinformatics?

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

- ... Bioinformatics is a hybrid of biology and computer science
- ... **Bioinformatics is computer aided biology!**

Computer based management and analysis of biological and biomedical data with useful applications in many disciplines, particularly genomics, proteomics, metabolomics, etc...

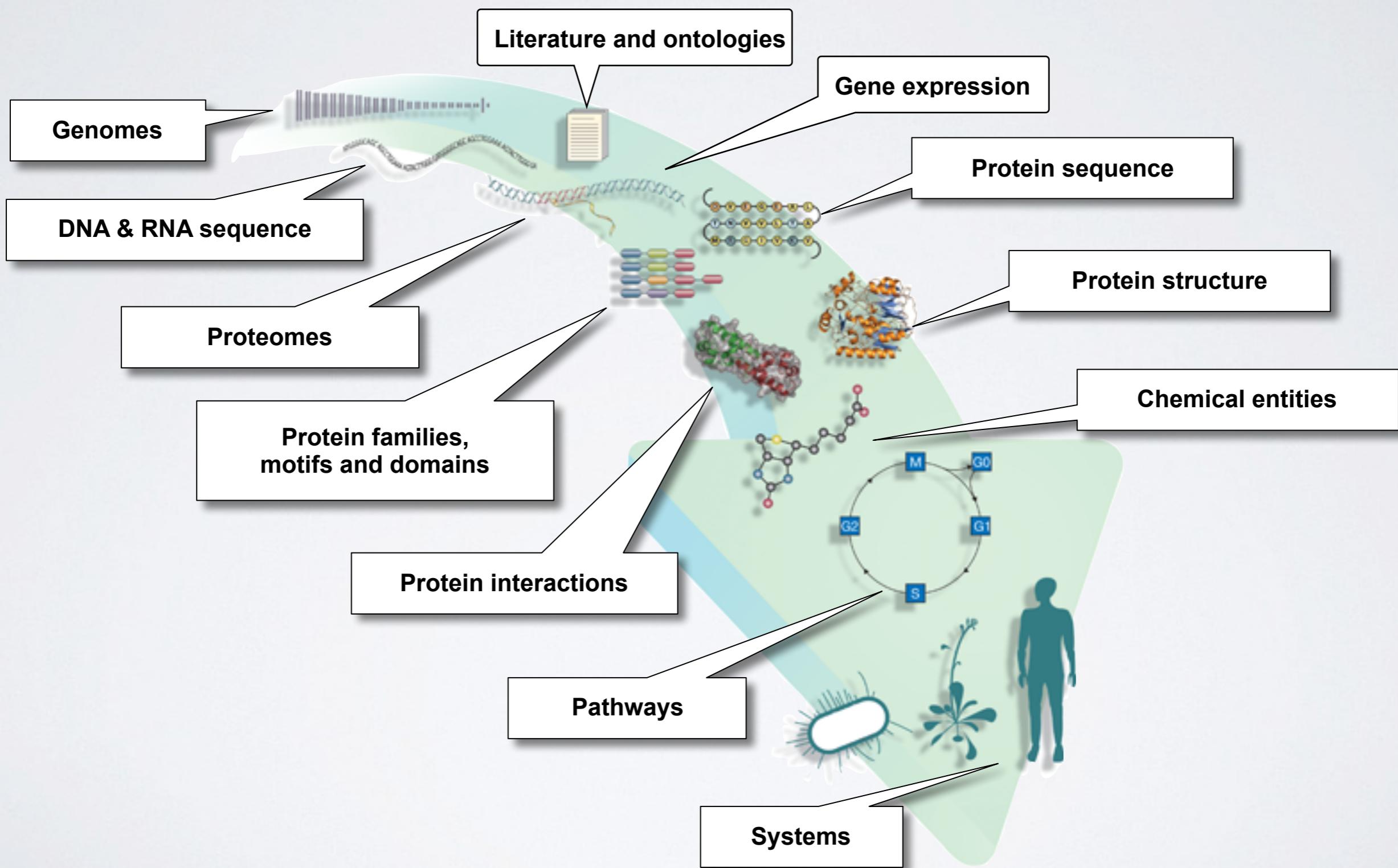
MORE DEFINITIONS

- ▶ “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 NM, et al. Methods Inf Med. 2001;40:346.
- ▶ “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.”
National Institutes of Health (NIH) (<http://tinyurl.com/l3gxr6b>)

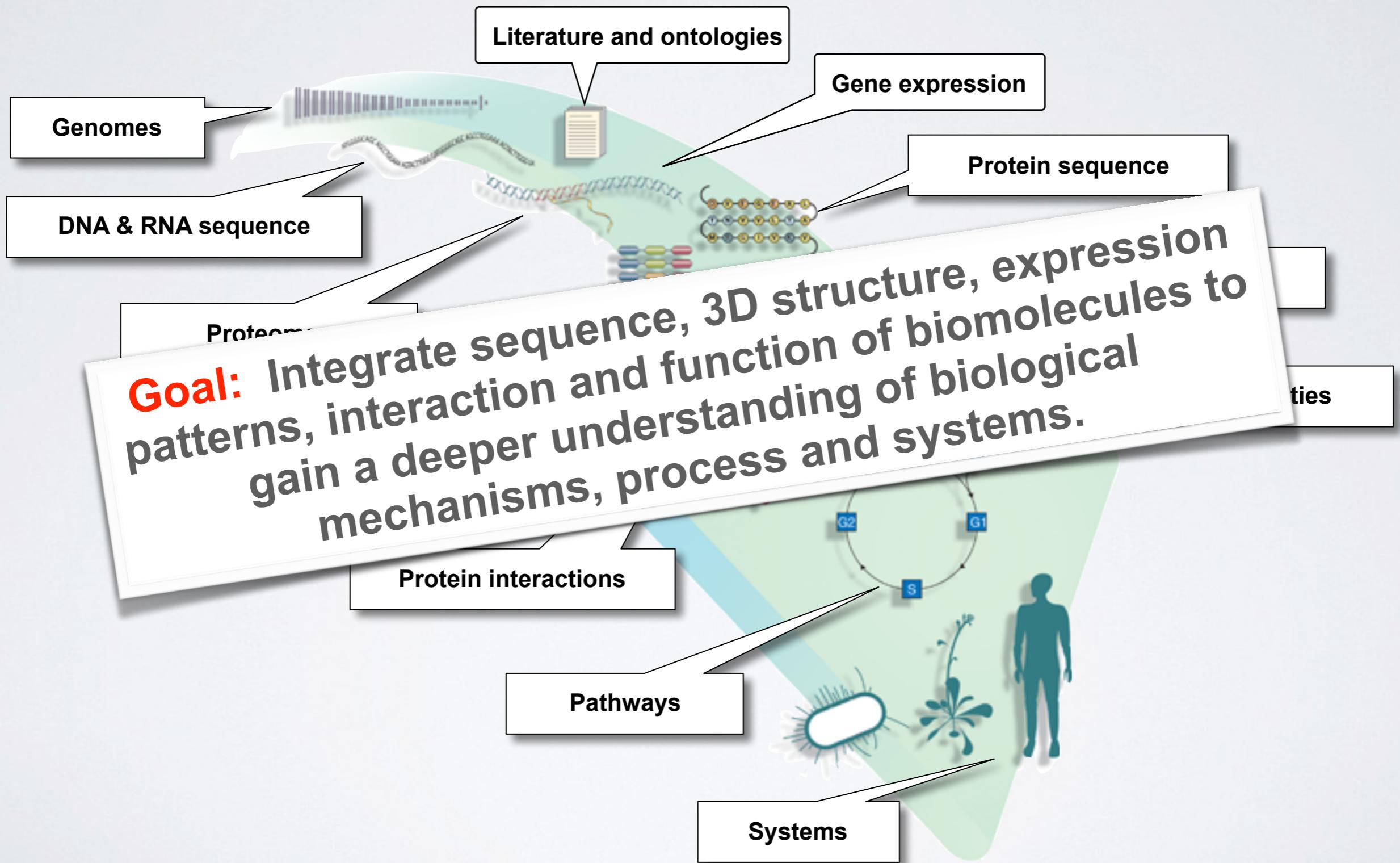
MORE DEFINITIONS

- ▶ “Bioinformatics is conceptualizing biology in terms of **macromolecules** and then applying “**informatics**” (derived from disciplines such as applied mathematics, science, and statistics) to **understand** and **analyze** the information associated with these molecules, on a **large-scale**.
Luscombe NM, et al. Methods 2001;40:346.
 - ▶ “Bioinformatics is the search, development, or application of computer approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize and analyze such data.”
National Institutes of Health (NIH) (<http://tinyurl.com/l3gxr6b>)
- Key Point:** Bioinformatics is Computer Aided Biology*

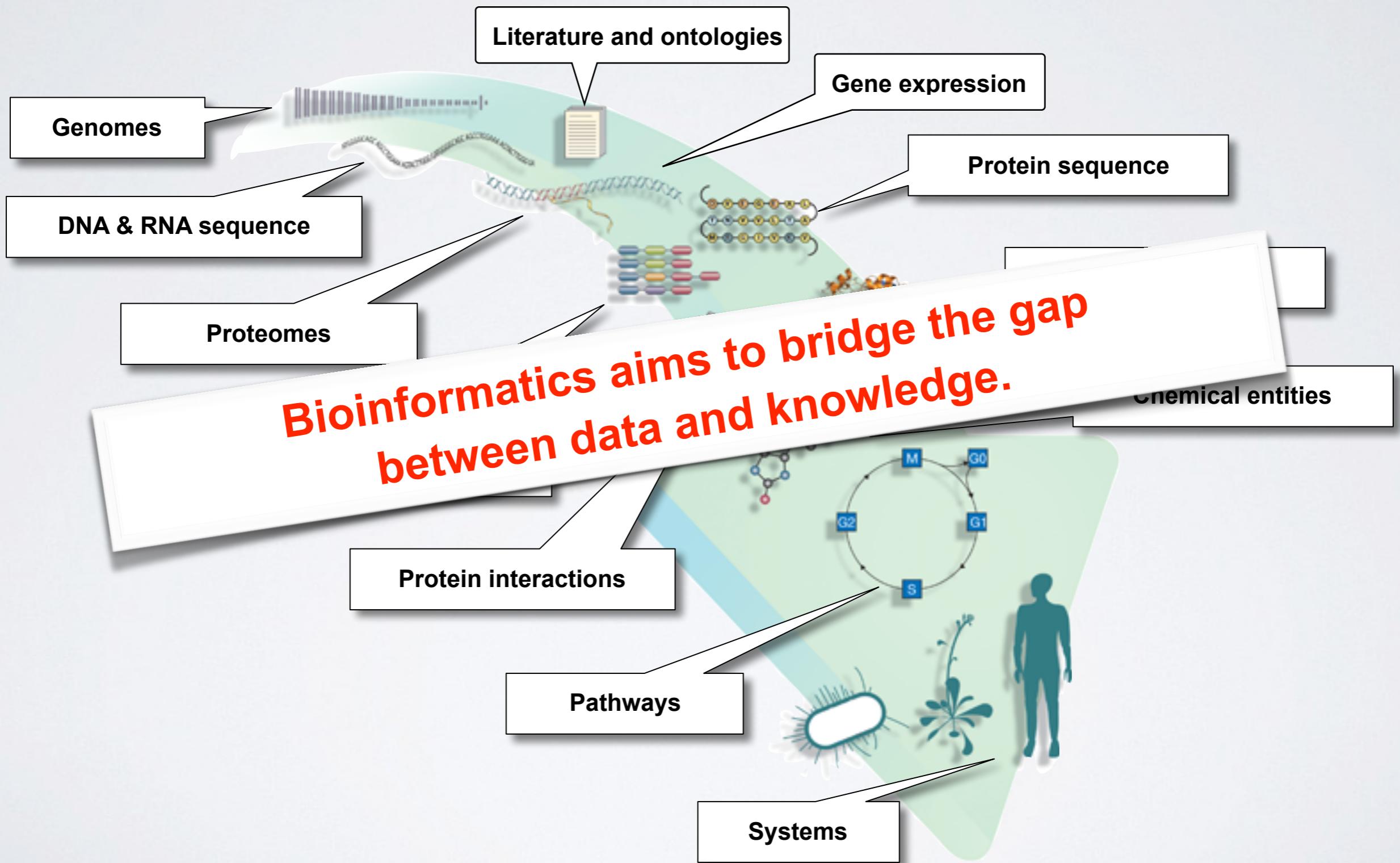
Major types of Bioinformatics Data



Major types of Bioinformatics Data



Major types of Bioinformatics Data



BIOINFORMATICS RESEARCH AREAS

Include but are not limited to:

- Organization, classification, dissemination and analysis of biological and biomedical data (particularly '-omics' data).
- Biological sequence analysis and phylogenetics.
- Genome organization and evolution.
- Regulation of gene expression and epigenetics.
- Biological pathways and networks in healthy & disease states.
- Protein structure prediction from sequence.
- Modeling and prediction of the biophysical properties of biomolecules for binding prediction and drug design.
- Design of biomolecular structure and function.

With applications to Biology, Medicine, Agriculture and Industry

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 programming languages frequently required
(e.g. R, Python, Perl, C, Java, Fortran)
- May require specialized high performance computing...

How do we *actually* do Bioinformatics?

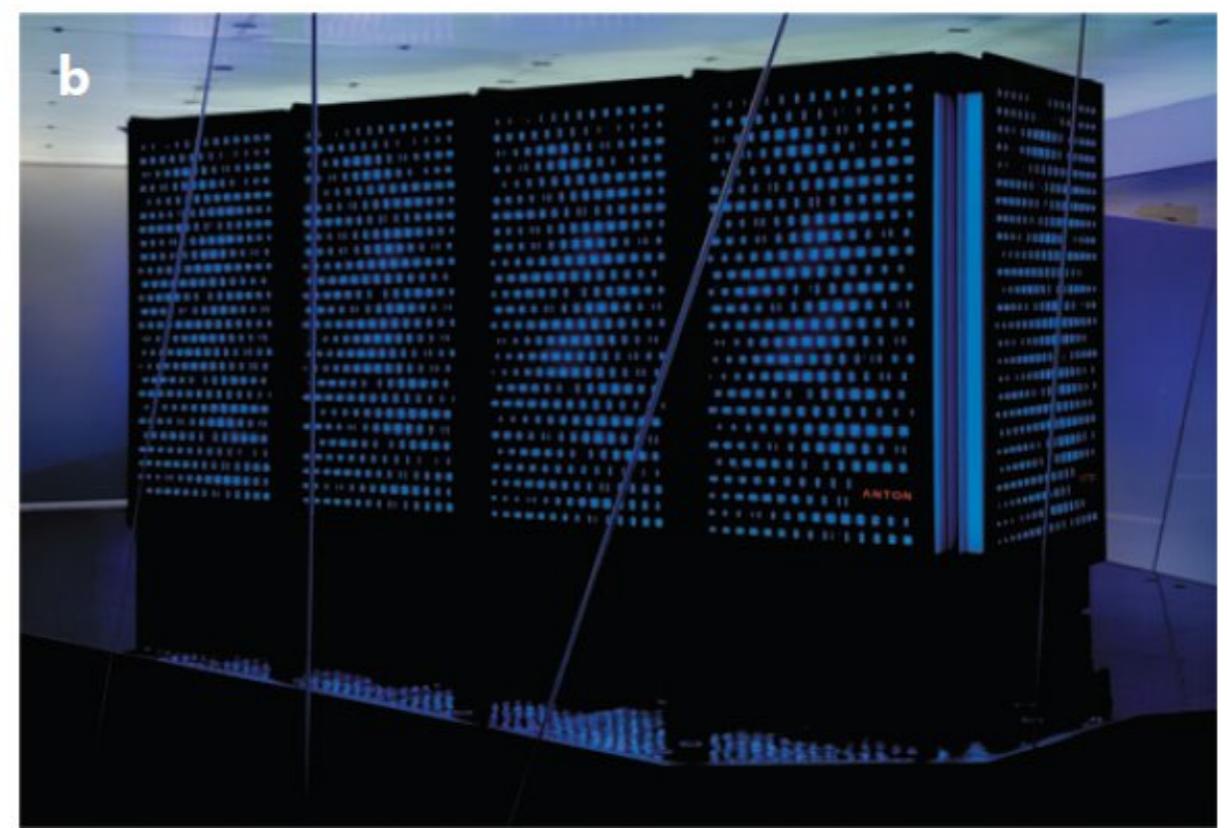
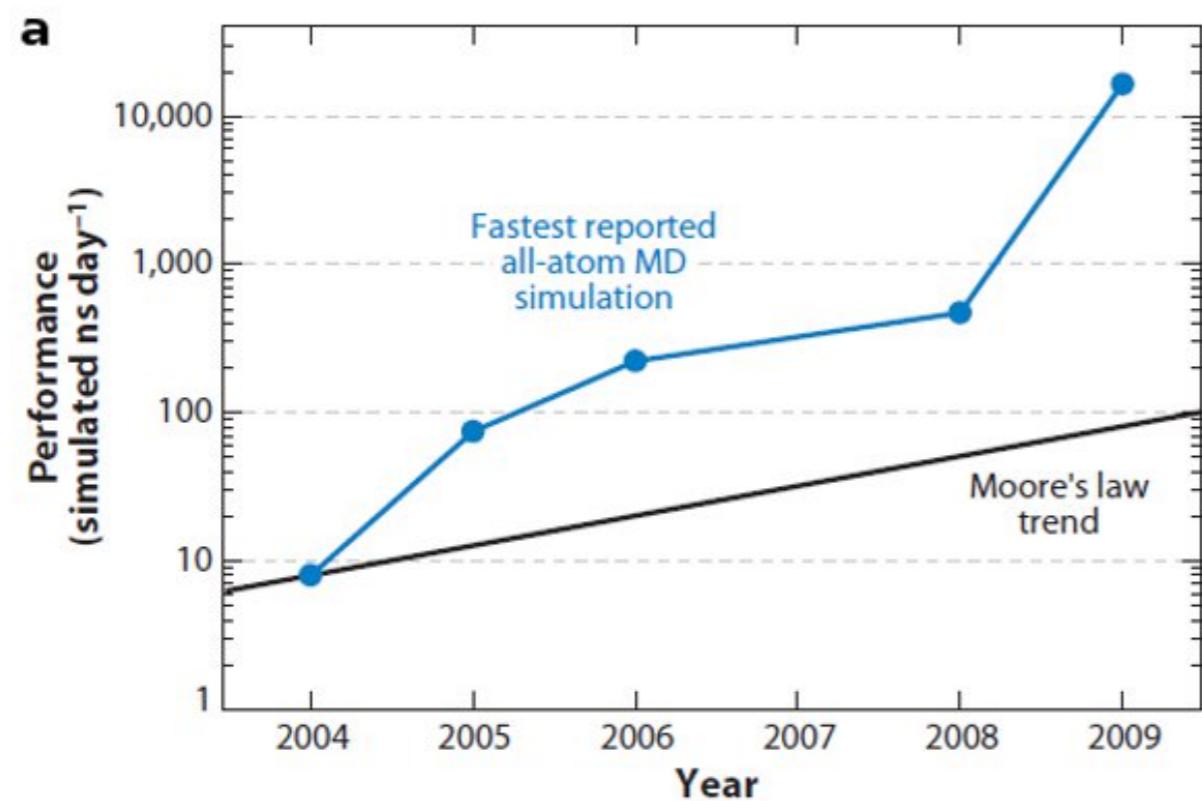
Pre-packaged tools and databases

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

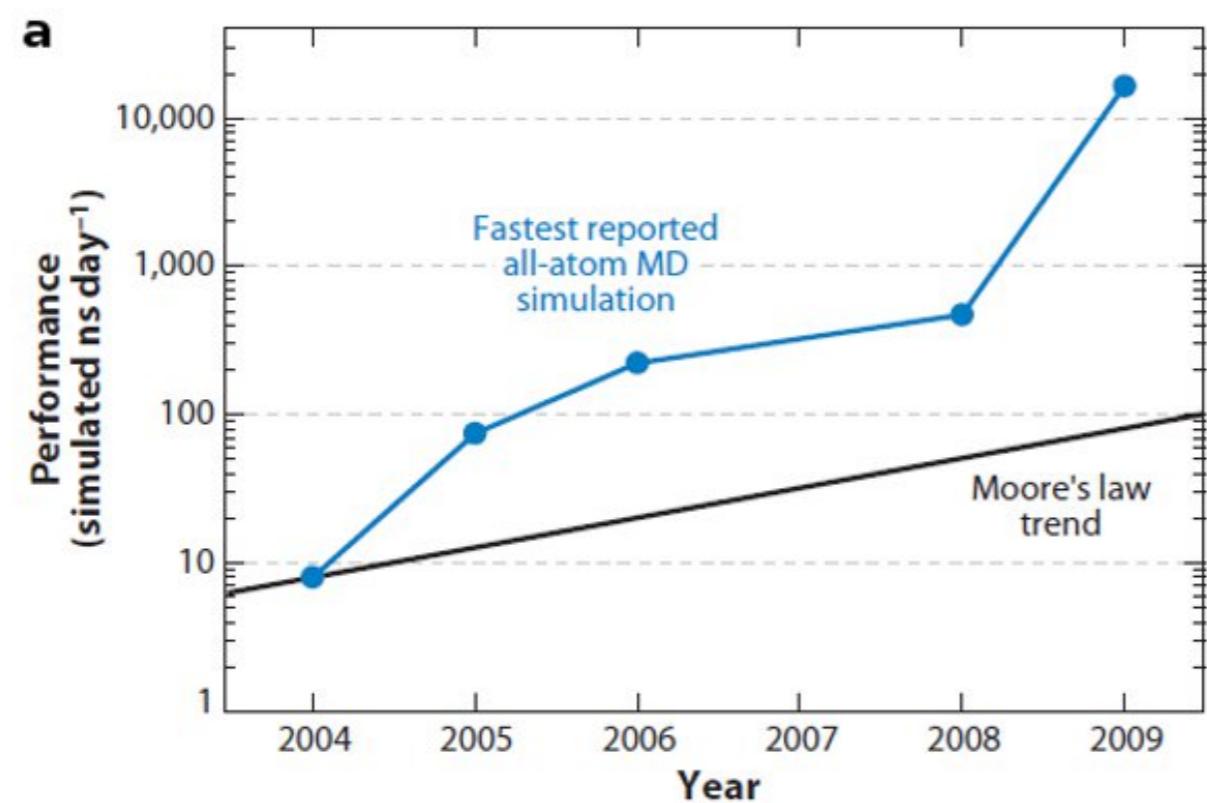
Advanced tool application & development

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

SIDE-NOTE: SUPERCOMPUTERS ANDGPUS



SIDE-NOTE: SUPERCOMPUTERS AND GPUS



HOW COMPUTERS HAVE CHANGED

DATE	COST	SPEED	MEMORY	SIZE
1967	\$40M	0.1 MHz	1 MB	WALL
2013	\$4,000	1 GHz	10 GB	LAPTOP
CHANGE	10,000	10,000	10,000	10,000

If cars were like computers then a new Volvo would cost \$3, would have a top speed of 1,000,000 Km/hr, would carry 50,000 adults and would park in a shadow.



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. 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

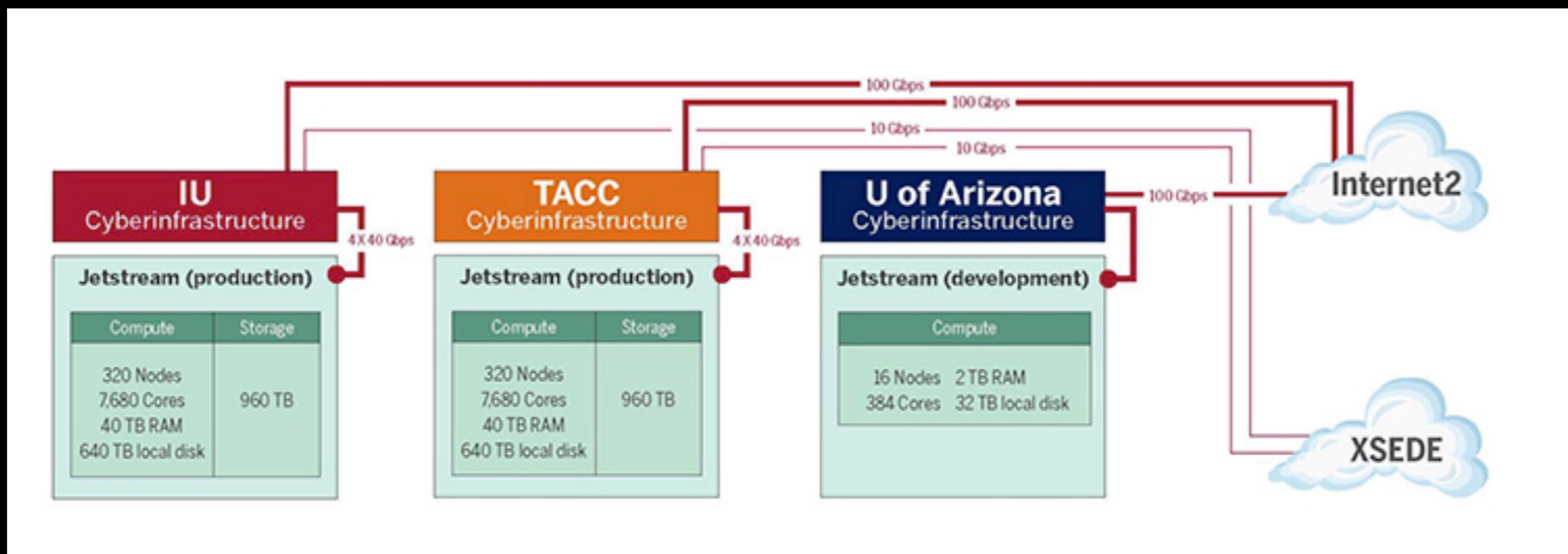
- XSEDE sponsors full-semester online courses
- Collaborations with faculty at participating institutions
- Campus visits offer guidance concerning course content

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.



Jetstream tutorials

Developed *user friendly* labs for Jetstream basics

The screenshot shows a web browser window with two main pages visible.

Left Side (Course Navigation):

- UC San Diego logo
- BGGN 213
- A hands-on introduction to the computer-based analysis of genomic and biomolecular data from the Division of Biological Sciences, UCSD
- Overview
- Lectures (highlighted with a red border)
- Computer Setup
- Learning Goals
- Assignments & Grading
- Ethics Code
- Screen Cast Videos

Right Side (Tutorial Content):

Starting a Jetstream Computer Instance!

Here we describe the process of starting up and managing a [Jetstream](#) service virtual machine instance.

Note: Jetstream is a cloud-based on-demand virtual machine system funded by the National Science Foundation. It will provide us with computers (what we call “virtual machine instances”) that look and feel just like a regular Linux workstation but with thousands of times the computing power!

What we’re going to do here is walk through starting up and running computer (an “instance”) on the Jetstream service.

Below we walk through the process of starting up and accessing one of these instances. To begin with, just think of it like requesting and logging-in to a brand new remote computer. We have provided screenshots of the whole process that you can click on to see a larger version. The important areas to fill in are circled in red.

Note Some of the details may vary – for example, if you have your own XSEDE account, you may want to log in with that – and the name of the operating system or “image” may also vary from “Ubuntu 16.04” depending

Jetstream tutorials

Developed *user friendly* labs for Jetstream basics

The image shows a web browser window with two tabs open. The top tab is titled "bioboot.github.io/bggn213_f17/jetstream/boot/" and shows the UC San Diego BGGN 213 course website. The bottom tab is also titled "bioboot.github.io/bggn213_f17/jetstream/boot/" and shows the Jetstream application login page.

Request to log in to the Jetstream Portal

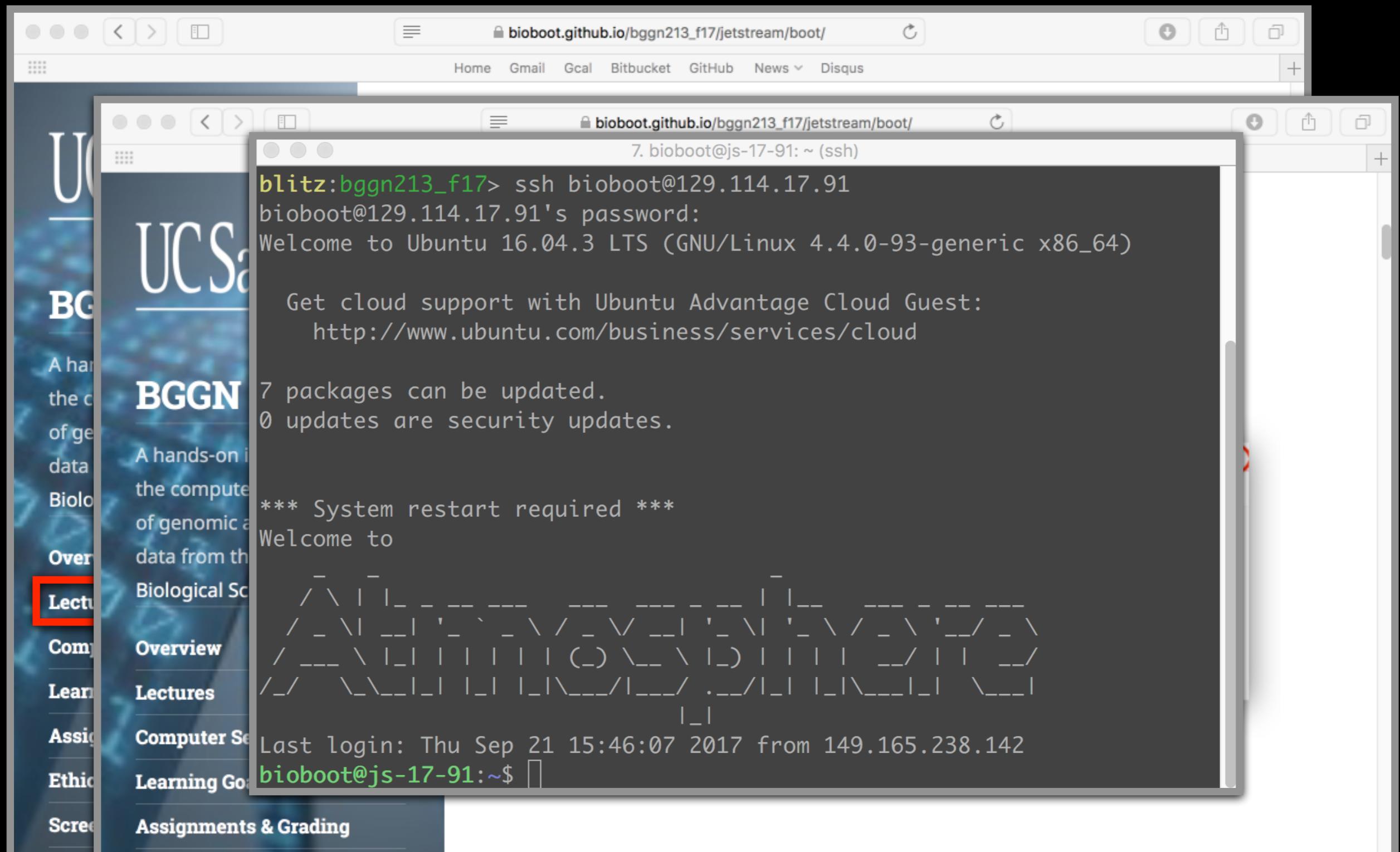
First, go to the Jetstream application at:
<https://use.jetstream-cloud.org/application>.

Now click the **Login** link in the upper right.

The Jetstream application login page is shown. The "Login" button in the top right corner is circled in red. The page features a search bar and a section for "Image Search".

Jetstream tutorials

Developed *user friendly* labs for Jetstream basics



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; Treasure Trove or Trivial Pursuit*” 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 ?	
Short queries	<input checked="" type="checkbox"/> Automatically adjust parameters for short input sequences ?
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 ?
Mask	<input type="checkbox"/> Mask for lookup table only <input type="checkbox"/> Mask lower case letters ?

PSI/PHI/DELTA BLAST

Upload PSSM Optional	<input type="button" value="Choose File"/> no file selected
PSI-BLAST Threshold	0.005
Pseudocount	0

Even Blast has many settable parameters

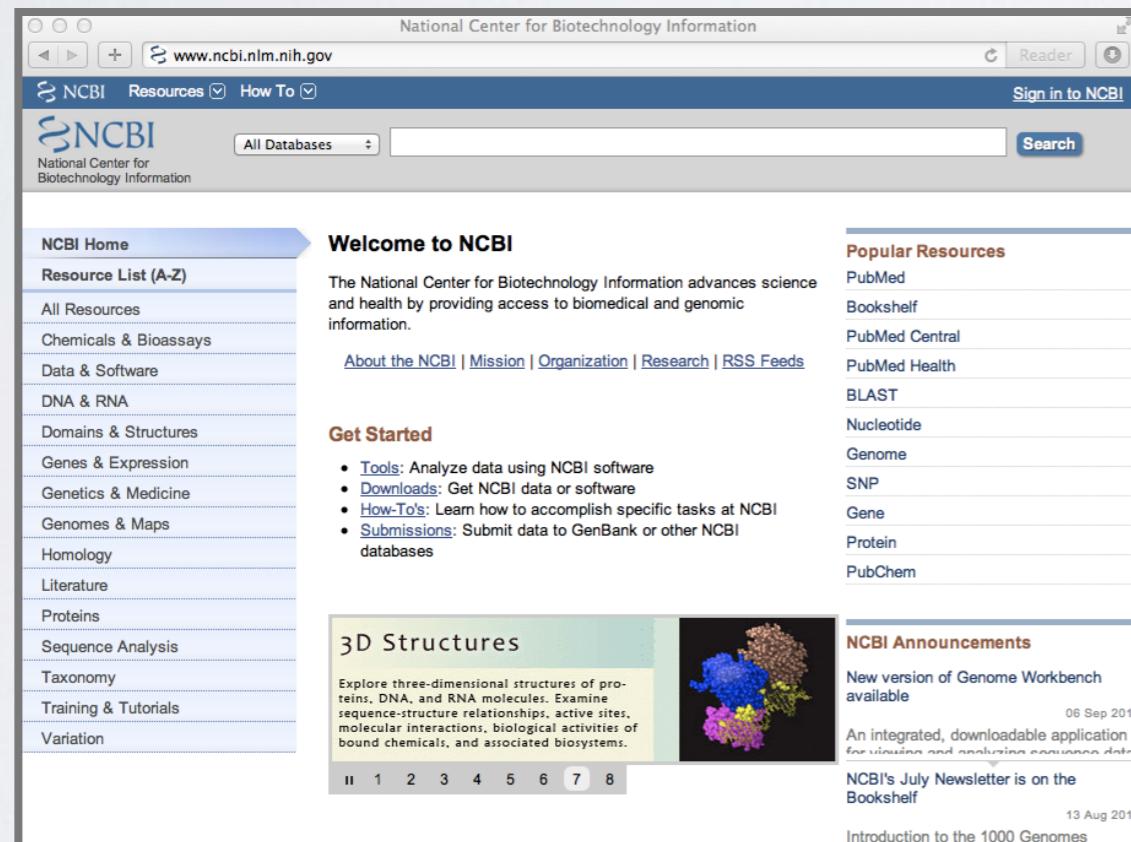
Related tools with different terminology

STEP 3 - Set your PROGRAM FASTA

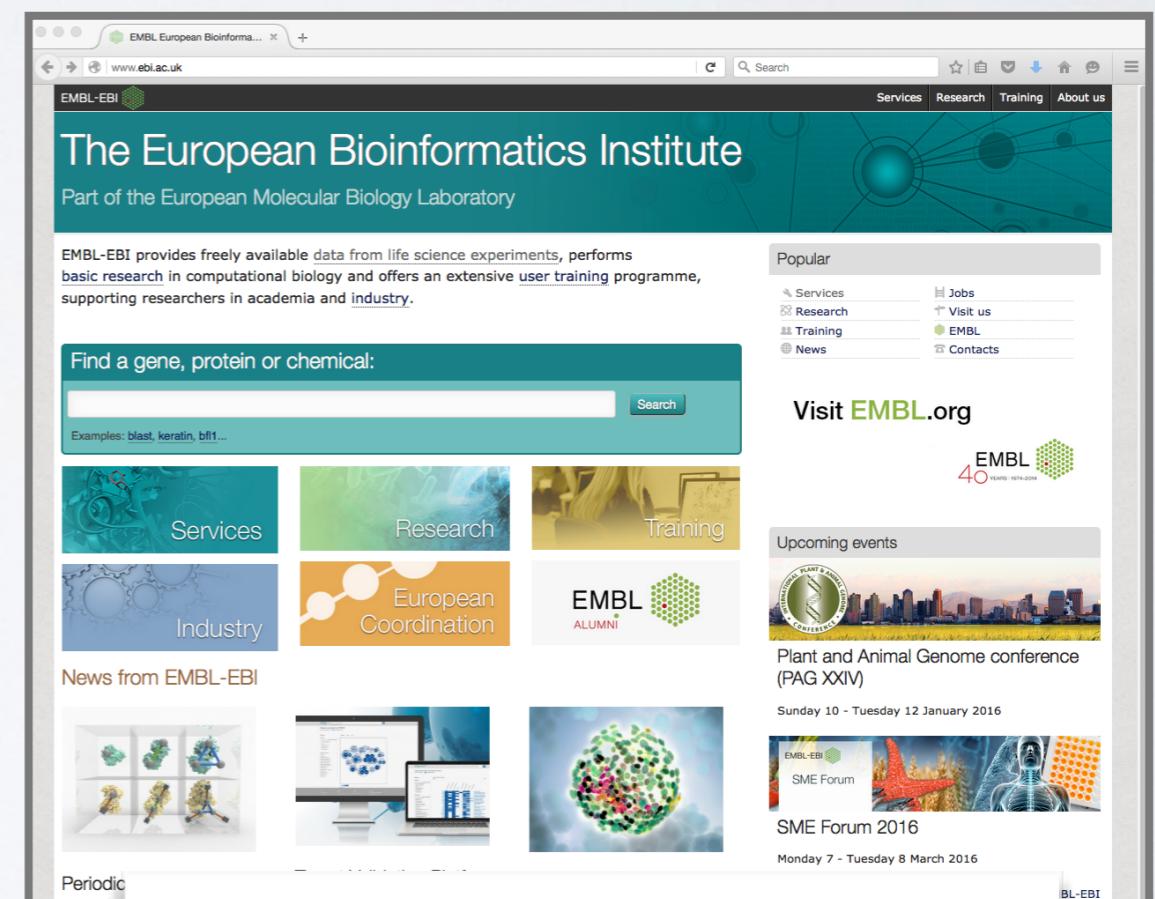
MATRIX	GAP OPEN	GAP EXTEND	KTUP	EXPECTATION UPPER VALUE	EXPECTATION LOWER VALUE
BLOSUM50	-10	-2	2	10	0 (default)
DNA STRAND	HISTOGRAM	FILTER	STATISTICAL ESTIMATES		
N/A	no	none	Regress		
SCORES	ALIGNMENTS	SEQUENCE RANGE	DATABASE RANGE		MULTI HSPs
50	50	START-END	START-END	no	
SCORE FORMAT					
Default					

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 "NCBI Resources", "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. A sidebar on the left lists various NCBI databases and services, and a right sidebar lists "Popular Resources" such as PubMed, Bookshelf, and BLAST.



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 search bar for "Find a gene, protein or chemical", sections for "Services", "Research", "Training", "Industry", "European Coordination", and "EMBL ALUMNI". A sidebar on the right lists "Popular" links for services, research, training, and news. Below the main content, there are sections for "Upcoming events" (Plant and Animal Genome conference) and "SME Forum 2016".

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

<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 the lecture



<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

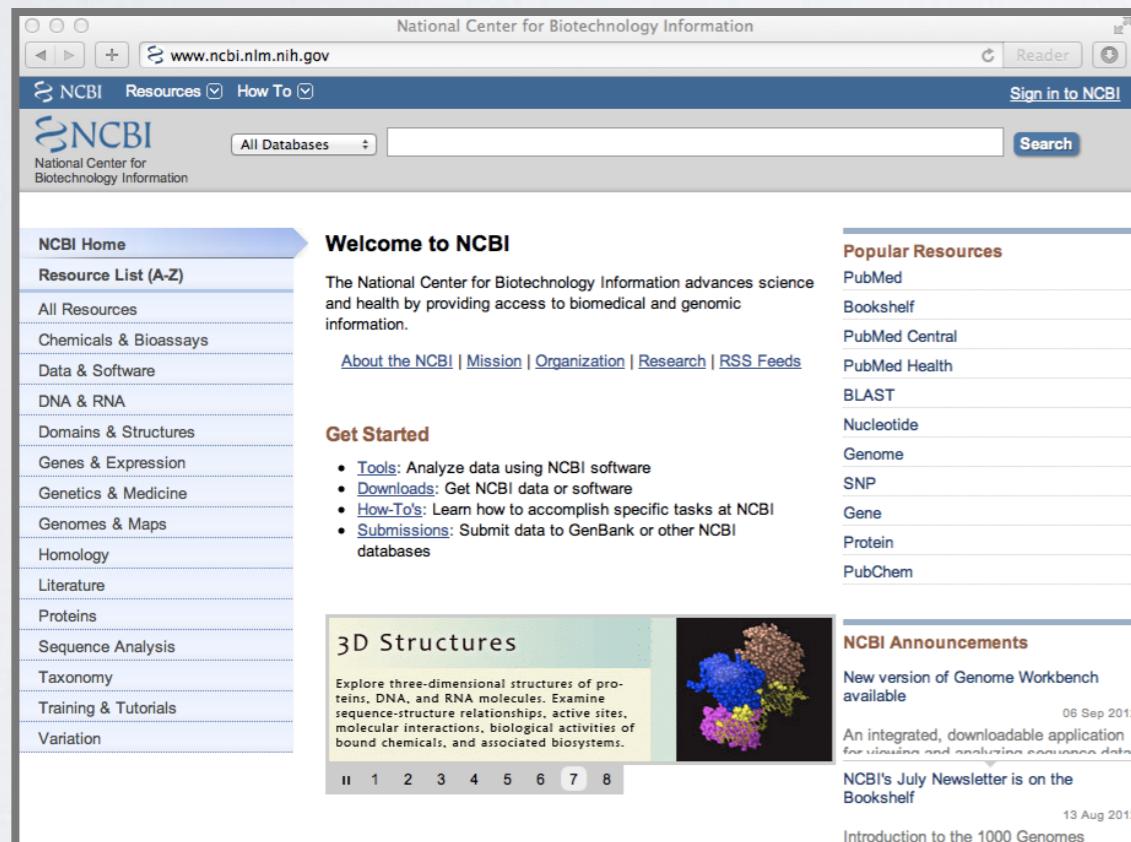
3D Structures
Explore three-dimensional structures of proteins, DNA, and RNA molecules. Examine sequence-structure relationships, active sites, molecular interactions, biological activities of bound chemicals and associated biosystems

Protein PubChem

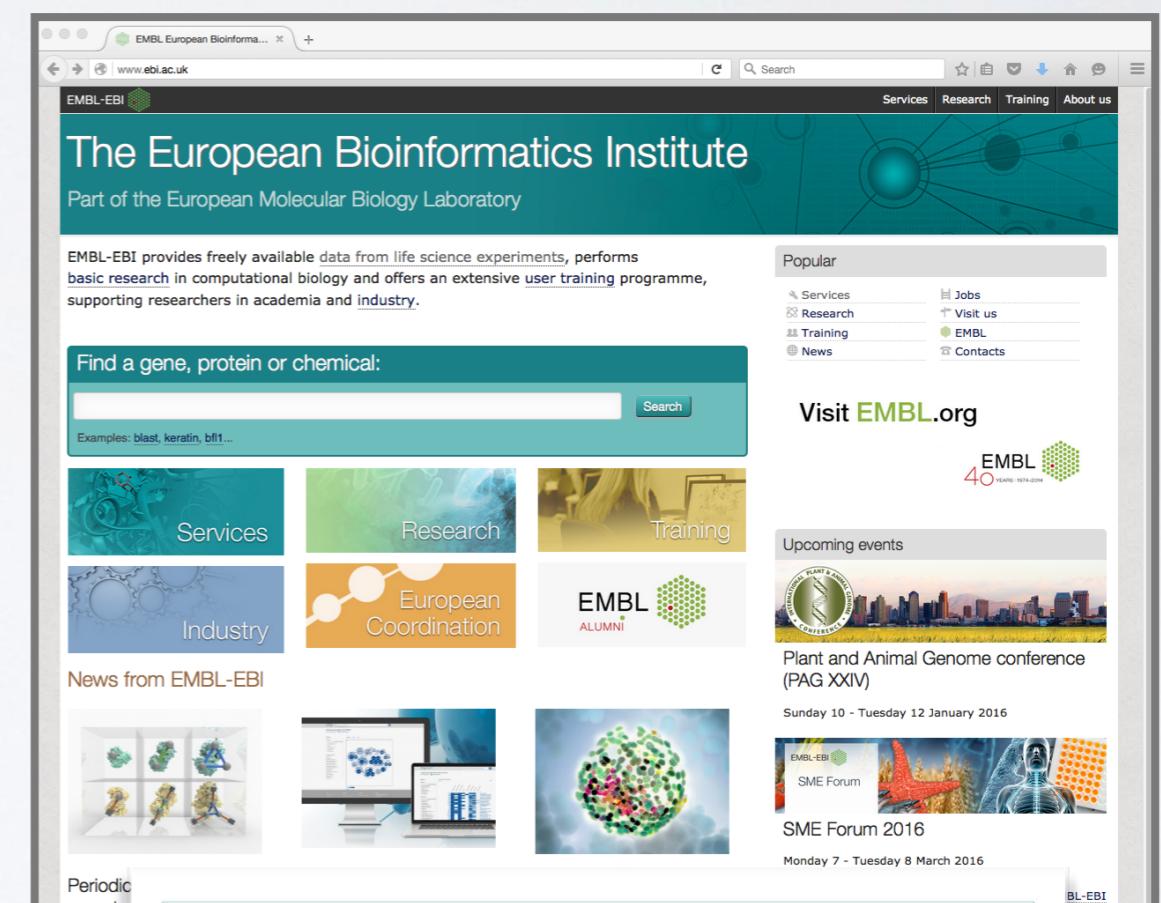
NCBI Announcements
New version of Genome Workbench available 06 Sep
An integrated, downloadable applicati

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 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. On the right, there's a sidebar titled "Popular Resources" listing PubMed, Bookshelf, PubMed Central, PubMed Health, BLAST, Nucleotide, Genome, SNP, Gene, Protein, and PubChem. A "NCBI Announcements" section highlights a new version of the Genome Workbench and the July Newsletter.



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 featuring a "Find a gene, protein or chemical:" search bar, a "Popular" sidebar with links to Services, Research, Training, and News, and sections for "Visit EMBL.org", "Upcoming events", and "SME Forum 2016". The main content area also includes sections for "Services", "Research", "Training", "Industry", "European Coordination", and "EMBL ALUMNI".

<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 briefly cover several 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. 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, and EMBL ALUMNI. On the right side, there are sections for Popular links (Services, Research, Training, News, Jobs, Visit us, EMBL, Contacts), a "Visit EMBL.org" link with the EMBL 40th anniversary logo, an "Upcoming events" section featuring the Plant and Animal Genome conference (PAG XXIV), and a "News from EMBL-EBI" section with three thumbnail images.

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

EMBL ALUMNI

Upcoming events

Plant and Animal Genome conference (PAG XXIV)

Sunday 10 - Tuesday 12 January 2016

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

The screenshot shows the EBI Services website (www.ebi.ac.uk/services) with a teal header and a banner featuring a molecular structure. The main content area displays nine service categories in a grid:

- DNA & RNA**: genes, genomes & variation
- Gene expression**: RNA, protein & metabolite expression
- Proteins**: sequences, families & motifs
- Structures**: Molecular & cellular structures
- Systems**: reactions, interactions & pathways
- Chemical biology**: chemogenomics & metabolomics
- Ontologies**: taxonomies & controlled vocabularies
- Literature**: Scientific publications & patents
- Cross domain**: cross-domain tools & resources

On the right side, there are two sections: "Popular" (listing Ensembl, UniProt, PDB, ArrayExpress, ChEMBL, BLAST, Europe PMC, Reactome, Train online, and Support) and "Service news" (showing a monarch butterfly on a stack of books). A large "Training" section is also visible at the bottom right.

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

The screenshot shows the EBI Services website (www.ebi.ac.uk/services) with a teal header and a sidebar on the left listing services like DNA & RNA, Gene expression, Proteins, Systems, Chemical biology, Ontologies, Literature, and Cross domain. A red box highlights the 'Proteins' service. On the right, a 'Popular' section lists Ensembl, UniProt, PDB, ArrayExpress, and ChEMBL, with a red box around Ensembl. Below this is a banner featuring a monarch butterfly and the word 'Training'.

Services < EMBL-EBI

www.ebi.ac.uk/services

EMBL-EBI

Services | Research | Training | About us

Services

Overview | A to Z | Data submission | Support

Bioinformatics services

We maintain the world's most comprehensive range of **freely available** and up-to-date molecular databases. Developed in collaboration with our colleagues worldwide, our services let you share data, perform complex queries and analyse the results in different ways. You can work locally by downloading our data and software, or use our [web services](#) to access our resources programmatically. You can read more about our services in the journal *Nucleic Acids Research*.

DNA & RNA
genes, genomes & variation

Gene expression
RNA, protein & metabolite expression

Proteins
sequences, families & motifs

Structures
Molecular & cellular structures

Systems
reactions, interactions & pathways

Chemical biology
chemogenomics & metabolomics

Ontologies
taxonomies & controlled vocabularies

Literature
Scientific publications & patents

Cross domain
cross-domain tools & resources

Popular

Ensembl

UniProt

PDB

ArrayExpress

ChEMBL

Training

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

The EBI makes available a wider variety of **online tools** than NCBI

Proteins

Popular services



UniProt: The Universal Protein Resource

The gold-standard, comprehensive resource for protein sequence and functional annotation data.



InterPro

A database for the classification of proteins into families, domains and conserved sites.



PRIDE: The Proteomics Identifications Database

An archive of protein expression data determined by mass spectrometry.



Pfam

A database of hidden Markov models and alignments to describe conserved protein families and domains.



Clustal Omega

Multiple sequence alignment of DNA or protein sequences. Clustal Omega replaces the older ClustalW alignment tools.



HMMER - protein homology search

Fast sensitive protein homology searches using profile hidden Markov models (HMMs). Variety of different search methods for querying against both sequence and HMM target databases.



InterProScan 5

InterProScan 5 searches sequences against InterPro's predictive protein signatures. Please note that [InterProScan 4.8 has been retired](#).

Quick links

- o Popular services in this category
- o All services in this category
- o Project websites in this category

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

The screenshot shows the homepage of the EMBL European Bioinformatics Institute (EBI) at www.ebi.ac.uk. The page features a dark teal header with the EMBL-EBI logo and navigation links for Services, Research, Training, and About us. Below the header is a large 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 "Find a gene, protein or chemical:" search bar with examples like "blast, keratin, bfl1...". To the right is a "Popular" sidebar with links to Services, Research, Training, News, Jobs, Visit us, EMBL, and Contacts. A red box highlights the "Training" link. Other sections include "Services", "Research", "European Coordination", "Industry", "EMBL ALUMNI", and "News from EMBL-EBI". On the right side, there's a "Visit EMBL.org" section with the EMBL 40th anniversary logo, an "Upcoming events" section featuring the "Plant and Animal Genome conference (PAG XXIV)", and a "Sunday 10 - Tuesday 12 January 2016" event date.

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

EMBL ALUMNI

European Coordination

Industry

News from EMBL-EBI

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

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

The screenshot shows a web browser displaying the EBI Training online course page. The URL in the address bar is www.ebi.ac.uk/training/online/course/using-sequence-similarity-searching-tools-embl-ebi. The page title is "Using sequence similarity searching tools at EMBL-EBI: webinar". The main content area shows a thumbnail of the webinar video, which features a blue background with white text and a portrait of a man. Below the thumbnail, a video player interface shows the start time as 0:00 / 37:42. To the left of the video, there's a sidebar with "Course content" sections for "Using sequence similarity searching tools at EMBL-EBI: webinar" and "Contributors". A "Print Course" button is also present. On the right side, there are "Popular" links for "Train online", "Find us", and "Funding", and a "Find us at..." section with links for "Open days and career days", "Conference exhibitions", "EMBL courses and events", "Genome campus events", and "Science for schools". The top navigation bar includes links for "Services", "Research", "Training" (which is highlighted), and "About us".

Using sequence similarity searching tools at EMBL-EBI: webinar

Using sequence similarity searching tools at EMBL-EBI: webinar

Using sequence similarity search tools at EMBL-EBI

Finding homologous sequences with BLAST, FASTA, PSI-Search etc.

Andrew Cowley
andrew.cowley@ebi.ac.uk
support@ebi.ac.uk

EMBL-EBI

0:00 / 37:42

This webinar focuses on how to use tools like **BLAST** and PSI-Search to find homologous sequences in EMBL-EBI databases, including tips on which tool and database to use, input formats, how to change parameters and how to interpret the results pages.

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](#)

Next Class...

**MAJOR BIOINFORMATICS
DATABASES AND ASSOCIATED
ONLINE TOOLS**

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

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.

Primary, secondary & composite databases

Bioinformatics databases can be usefully classified into *primary*, *secondary* and *composite* according to their data source.

- **Primary databases** (or archival databases) consist of data derived experimentally.
 - **GenBank**: NCBI's primary nucleotide sequence database.
 - **PDB**: Protein X-ray crystal and NMR structures.
- **Secondary databases** (or derived databases) contain information derived from a primary database.
 - **RefSeq**: non redundant set of curated reference sequences primarily from GenBank
 - **PFAM**: protein sequence families primarily from UniProt and PDB
- **Composite databases** (or *metadatabases*) join a variety of different primary and secondary database sources.
 - **OMIM**: catalog of human genes, genetic disorders and related literature
 - **GENE**: molecular data and literature related to genes with extensive links to other databases.

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!

https://bioboot.github.io/bggn213_S18/lectures/#1

The screenshot shows a web browser window with the URL https://bioboot.github.io/bggn213_S18/lectures/#1 in the address bar. The page itself is a course landing page for BGNN 213 at UC San Diego. The left sidebar has a blue background with the UC San Diego logo and the course title "BGNN 213". Below the title, it says "A hands-on introduction to the computer-based analysis of genomic and biomolecular data from the Division of Biological Sciences, UCSD". The sidebar menu includes "Overview", "Lectures" (which is highlighted with a red border), "Computer Setup", "Learning Goals", "Assignments & Grading", and "Ethics Code". At the bottom of the sidebar are social media icons for Twitter, GitHub, Email, and RSS. The main content area starts with a section titled "Goals:" followed by a bulleted list of goals. Then there's a section titled "Material:" with a list of resources, where the "Lab" and "Feedback" items are highlighted with a red box. Finally, there's a section titled "Homework:" with a list of resources.

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.
- The goals of the hands-on session is to introduce a range of core bioinformatics databases and associated online services whilst actively investigating the molecular basis of several common human disease.

Material:

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

Homework:

- [Questions](#),
- Readings:
 - PDF1: [What is bioinformatics? An introduction and overview](#),
 - PDF2: [Advancements and Challenges in Computational Biology](#),

BGGN-213: FOUNDATIONS OF BIOINFORMATICS (Lecture 1)

Bioinformatics Databases and Key Online Resources

https://bioboot.github.io/bggn213_S18/lectures/#1

Dr. Barry Grant

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
TTGGTGGTGAGGCCCTGGCAGGCTGCTGGTGTACCCCTGGACCCAGAGGTTCTTGAGTCCTTGG
GGATCTGTCCACTCCTGATGCAGTTATGGCAACCTAAGGTGAAGGCTCATGGCAAGAAAGTGCTCGGT
GCCTTAGTGTAGGCCTGGCTCACCTGGACAACCTCAAGGGCACCTTGCACACTGAGTGAGCTGCACT
GTGACAAGCTGCACGTGGATCCTGAGAACCTCAGGCTCCTGGCAACGTGCTGGTCTGTGTGCTGGCCA
TCACCTTGGCAAAGAACATTCCCCACCAGTGCAGGCTGCCTATCAGAAAGTGGTGGCTGGTGTGGCTAAT
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!

YOUR TURN!

- There are five major hands-on sections including:

1. BLAST, GenBank and OMIM @ **NCBI**

End times:

[2:35 pm]

2. GENE database @ **NCBI**

[2:55 pm]

— BREAK —

— 3:10 pm —

3. UniProt & Muscle @ **EBI**

[3:30 pm]

4. PFAM, PDB & NGL

[4:00 pm]

— BREAK —

— 4:10 pm —

5. Extension exercises

[4:40 pm]

- ▶ 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/bggn213/>

- Complete the **initial course questionnaire**:
- Check out the “**Background Reading**” material online:
- Complete the **lecture 1 homework questions**:

