



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

05:00

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 leaning goals.
Introduction to Bioinformatics	Introducing the <i>what, why</i> and <i>how</i> of bioinformatics?
Bioinformatics Database	Hands-on exploration of several major databases and their associated tools.

http://thegrantlab.org/bimm143/

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

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

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: cyusi@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/

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

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: cyusi@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?

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

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

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 browser's address bar and navigation buttons are visible at the top. Below the browser window, the page content is displayed. On the left side, there is a dark blue sidebar with the UC San Diego logo and a list of navigation links: **BIMM 143**, Overview, Lectures, Computer Setup, **Learning Goals** (highlighted with a red box), Assignments & Grading, and Ethics Code. The main content area has the heading **Specific Learning Goals** and a paragraph explaining that 60%-70% of class time is dedicated to these goals. Below this is a table with 5 rows, each containing a goal number, a description, and the lecture(s) it covers. A red arrow on the right side of the page points downwards.

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

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

Lectures

All Lectures are Tu/Th 9:00-12:00 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	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

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

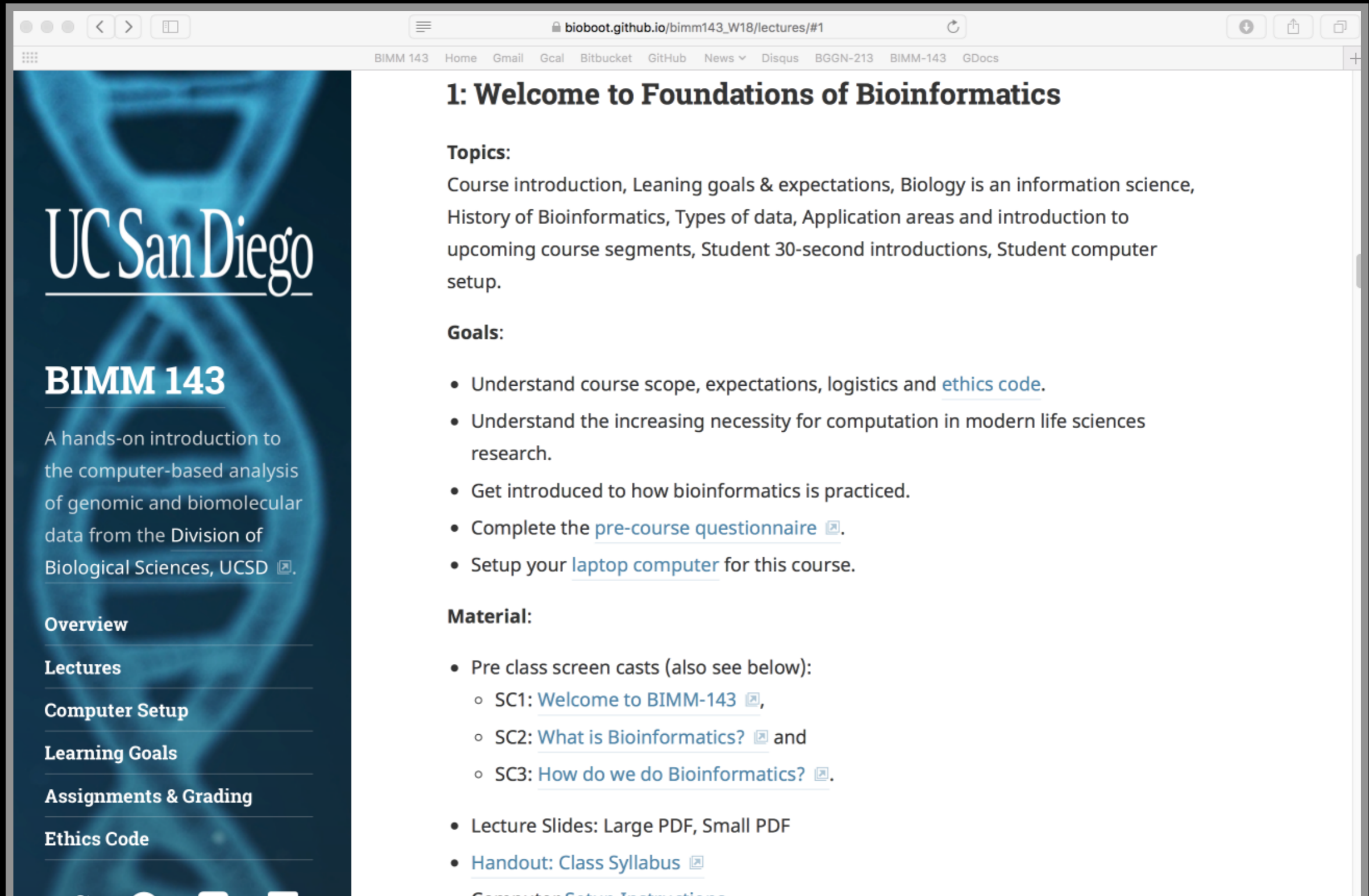
Lectures

All Lectures are Tu/Th 9:00-12:00 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	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 browser's address bar and navigation buttons are visible at the top. Below the browser window, the page content is displayed. On the left side, there is a dark blue sidebar with the UC San Diego logo and the course title **BIMM 143**. The sidebar also contains a list of navigation links: Overview, Lectures, Computer Setup, Learning Goals, Assignments & Grading, and Ethics Code. The main content area on the right is white and features the heading **1: Welcome to Foundations of Bioinformatics**. Below this heading, there are sections for Topics, Goals, and Material, each containing descriptive text and a list of bullet points with links to external resources.

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:

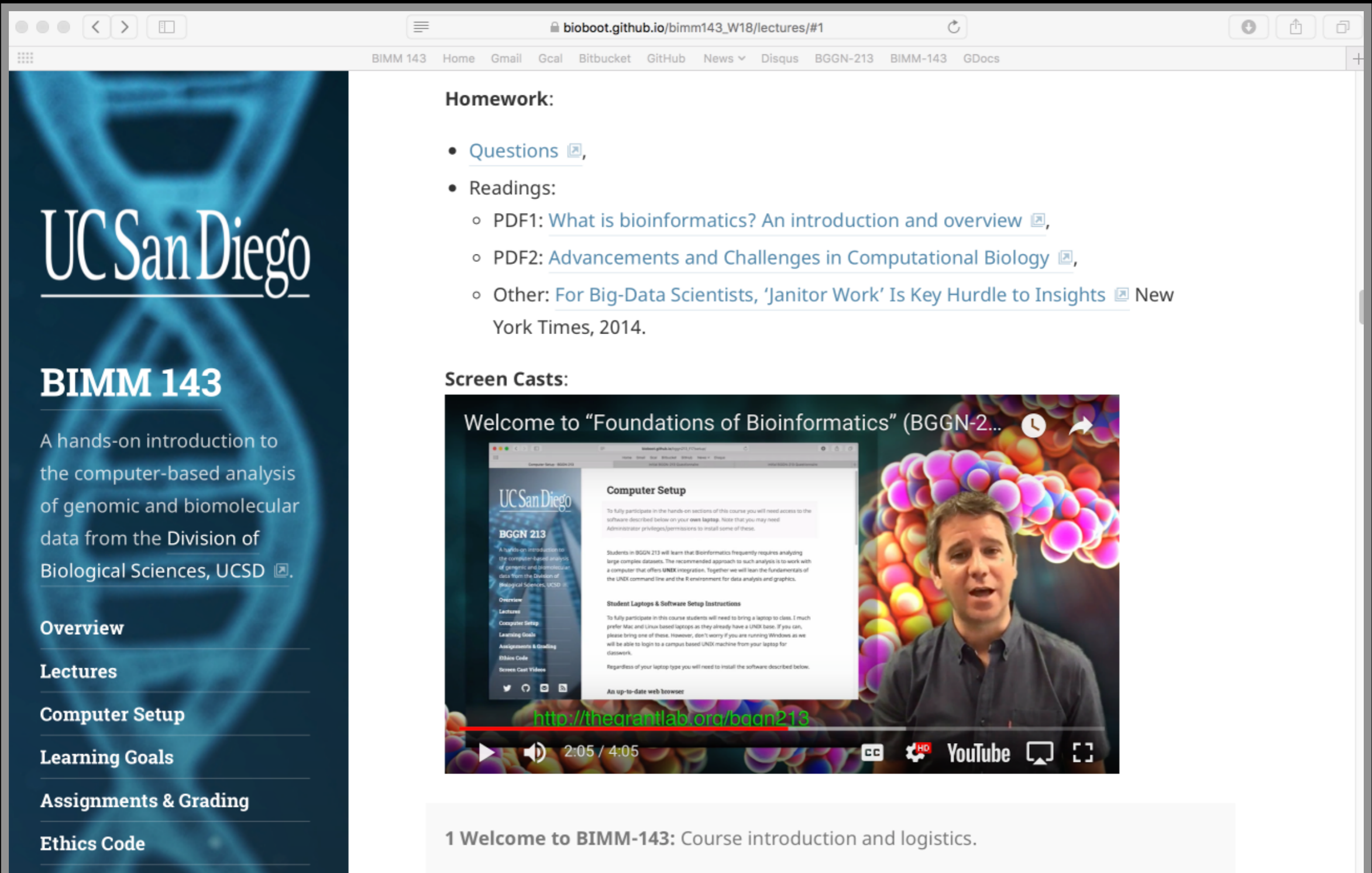
- Understand course scope, expectations, logistics and [ethics code](#).
- Understand the increasing necessity for computation in modern life sciences research.
- Get introduced to how bioinformatics is practiced.
- Complete the [pre-course questionnaire](#).
- Setup your [laptop computer](#) for this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

Material:

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

Homework

Goals, Class material, Screencasts & Homework



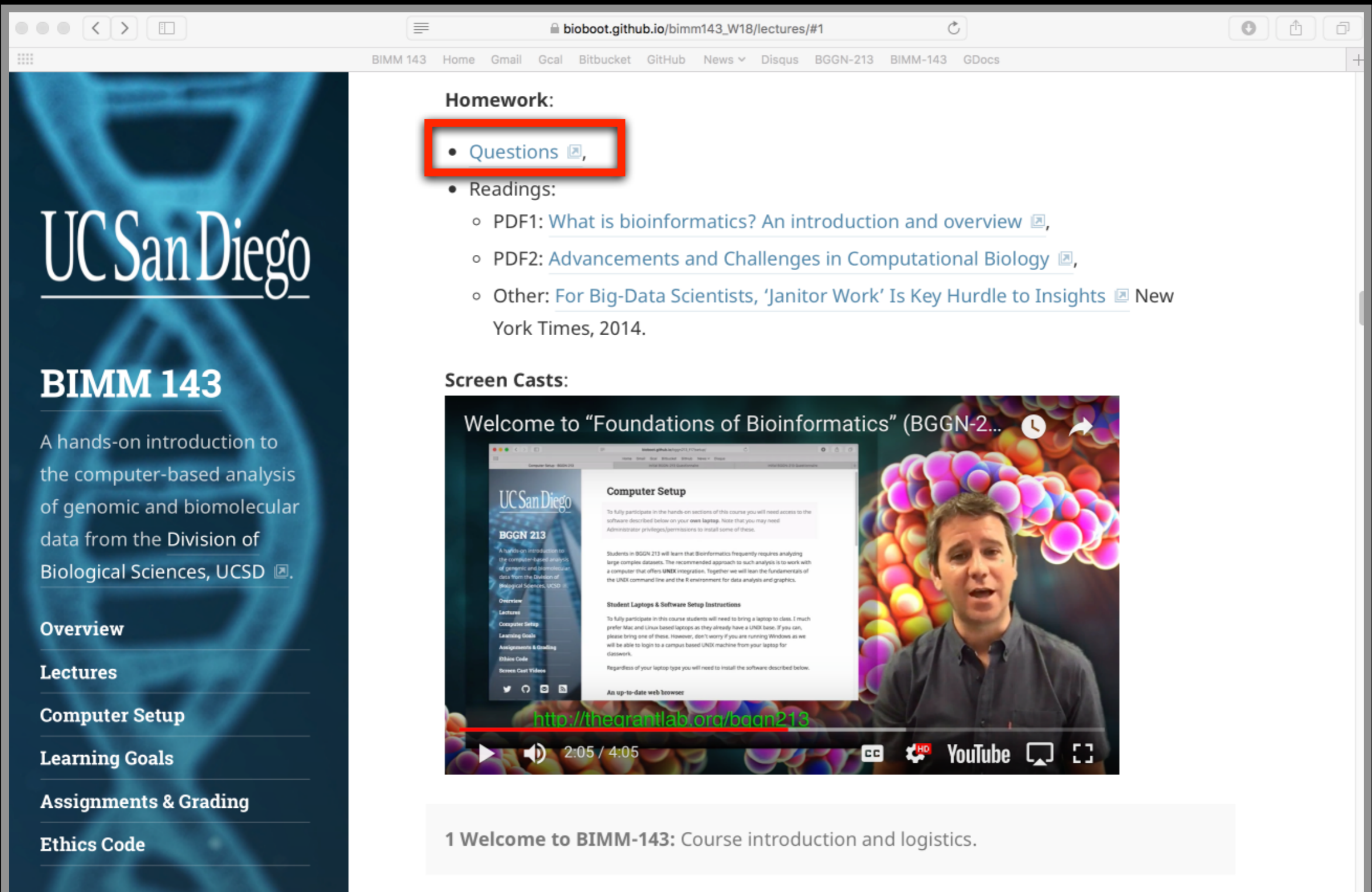
The screenshot shows a web browser window with the URL `bioboot.github.io/bimm143_W18/lectures/#1`. The browser's address bar and tabs are visible at the top. The page content is divided into several sections:

- UC San Diego** logo and **BIMM 143** title.
- A brief description: "A hands-on introduction to the computer-based analysis of genomic and biomolecular data from the Division of Biological Sciences, UCSD."
- A navigation menu on the left with links for Overview, Lectures, Computer Setup, Learning Goals, Assignments & Grading, and Ethics Code.
- 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:**
 - A video player showing a screencast titled "Welcome to 'Foundations of Bioinformatics' (BGGN-2...". The video content includes a "Computer Setup" page with instructions for students, overlaid on a background of colorful molecular models. The video player shows a timestamp of 2:05 / 4:05 and a URL `http://theorantlab.org/bqon213` at the bottom.

At the bottom of the page, a grey box contains the text: "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 is as follows:

- UC San Diego** logo and **BIMM 143** title.
- Text: "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, Assignments & Grading, Ethics Code.
- 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:**
 - Video player: "Welcome to 'Foundations of Bioinformatics' (BGGN-2...)"
 - Video content: A man speaking in front of a background of colorful spheres. A browser window is overlaid showing the "Computer Setup" page for BGGN 213.
 - Video URL: <http://theorantlab.org/bqon213>
 - Video player controls: 2:05 / 4:05, YouTube logo, and other controls.
- Footer: 1 Welcome to BIMM-143: Course introduction and logistics.

Homework

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

BIMM143 Lecture 1 Homework (W19)

Please answer the following questions including your main [@ucsd.edu](mailto: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. Please provide your email address and UCSD
PID number so you can receive your score.

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

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

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

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 browser window with two overlapping pages. The background page is the UC San Diego BIMM 143 course page, which includes the university logo and a navigation menu with links for Overview, Lectures, Computer Setup, Learning Goals, and Ethics. The foreground page is a lecture slide titled "Designing a personalized cancer vaccine" from BIMM-143 Lecture 18, presented by Barry Grant on March 7, 2018. The slide content includes notes on somatic mutations and variant calling algorithms.

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

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 image displays three overlapping browser windows from the website `bioboot.github.io`. The top window shows a navigation menu with links for Home, Gmail, Gcal, GitHub, BIMM143, BGGN213, Atmosphere, BIMM194, Blink, and News. The middle window shows the course page for BIMM 143, featuring the UC San Diego logo and a sidebar menu with links for Overview, Lectures, Computer, and Learning. The bottom window shows a lecture page titled "10: (Project:) Find a Gene Assignment Part 1".

10: (Project:) Find a Gene Assignment Part 1

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).

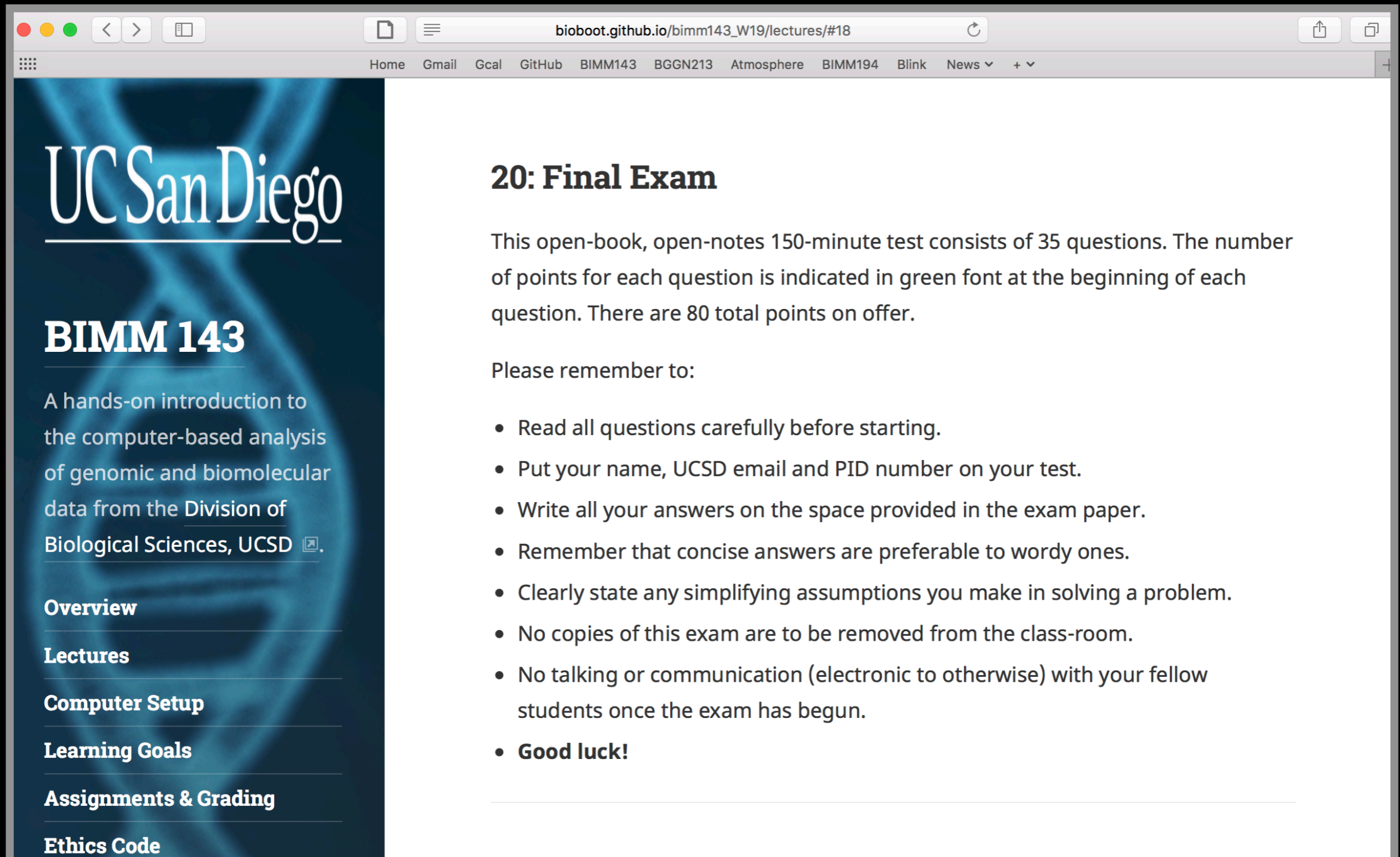
Late responses will not be accepted under any circumstances.

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 browser's address bar and tabs are visible at the top. The page content is divided into a left sidebar and a main content area. The sidebar, on a dark blue background, features the UC San Diego logo and a list of navigation links: **BIMM 143**, Overview, Lectures, Computer Setup, Learning Goals, Assignments & Grading, and Ethics Code. The main content area is white and contains the following text:

20: Final Exam

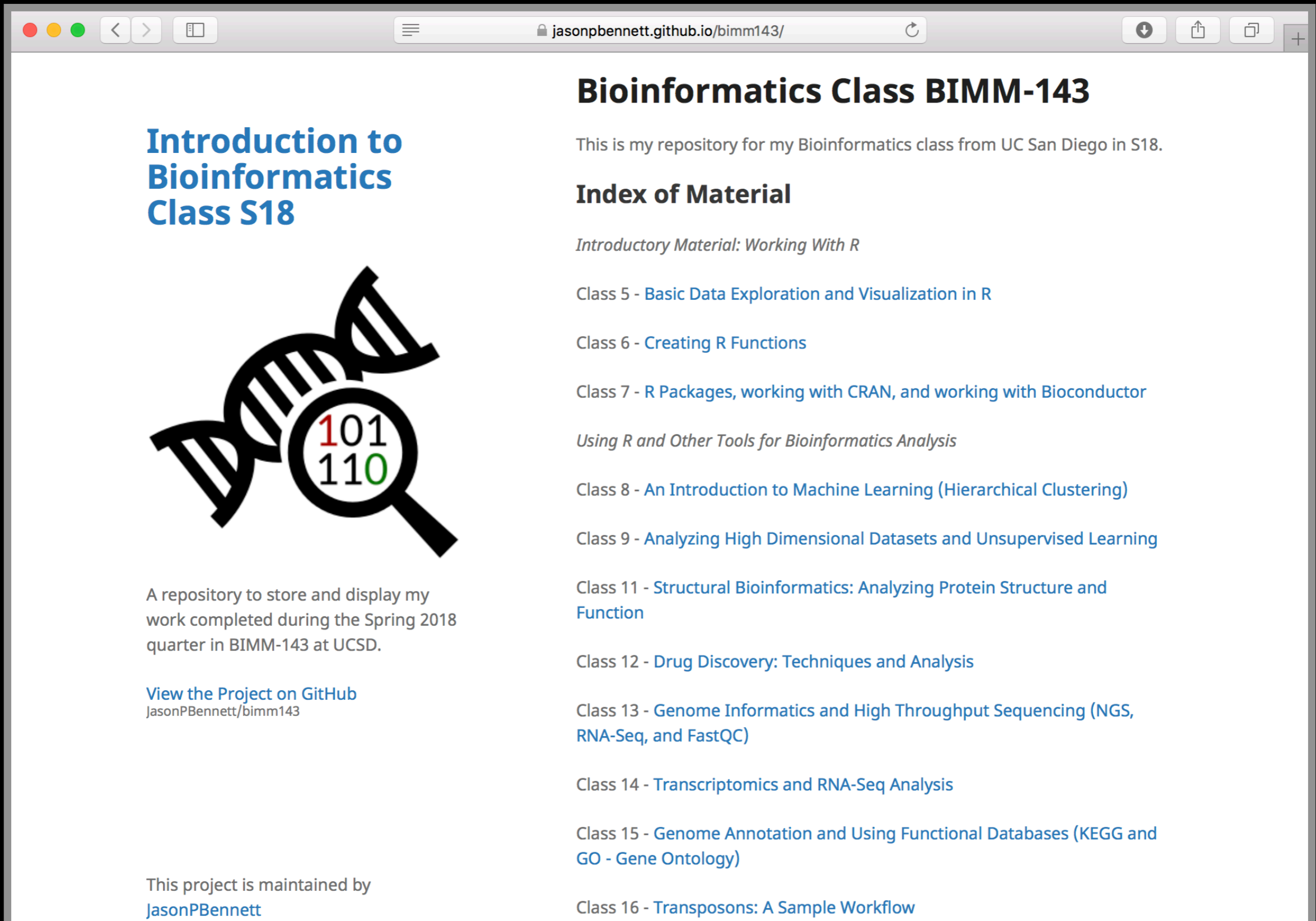
This open-book, open-notes 150-minute test consists of 35 questions. The number of points for each question is indicated in green font at the beginning of each question. There are 80 total points on offer.

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!



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

This project is maintained by
[JasonPBennett](#)

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](#)

Bonus:

Online portfolio of **your** bioinformatics work!

The screenshot shows a web browser window with the address bar containing `vector jasonpbennett.github.io/bimm143/class13/NGS.html`. The browser tabs include `class13` and `Bioinformatics Class 5`. The page content is as follows:

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")
```

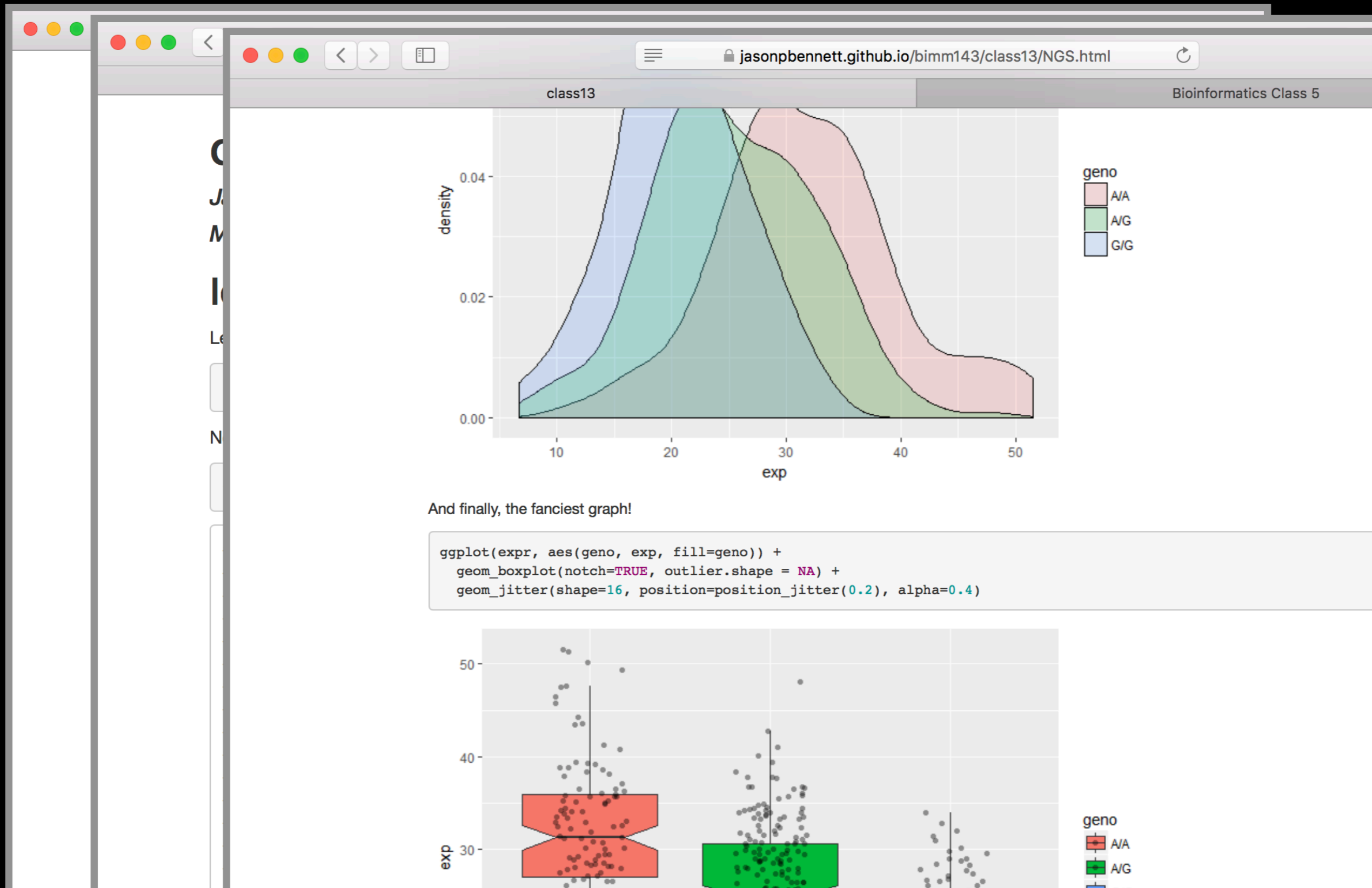
Now lets look at a table of the data:

```
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  
##                               NA19666 (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

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

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
	Given an RNA-Seq data file. find the set of significantly differentially	

BIMM-143 Learning Goals....

Delve deeper into “real-world” bioinformatics

The screenshot shows a web browser window with the URL `bioboot.github.io/bimm143_W18/goals/`. The page content includes a sidebar on the left with navigation links: **BIMM 143**, Overview, Lectures, Computer Setup, **Learning Goals** (highlighted with a red box), Assignments & Grading, and Ethics Code. The main content area displays a table of learning goals:

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
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 heterogenous high throughput data sets into their functional	19

A green rectangular box highlights the learning goals numbered 12 through 17. A red arrow with a dashed line above it points downwards from the right edge of the page.

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.



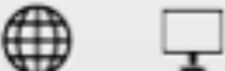





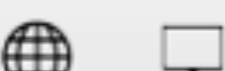

Why use R?

Productivity

Flexibility

Genomic data analysis

IEEE 2016 Top Programming Languages

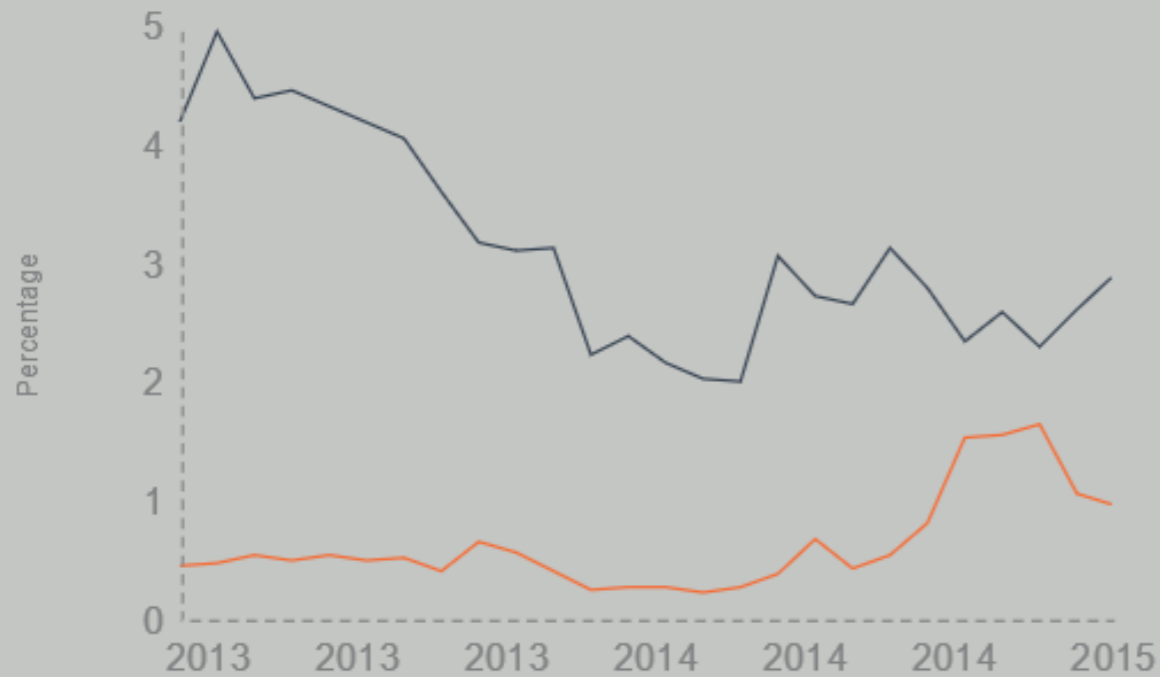
Language Rank	Types	Spectrum Ranking
1. C		100.0
2. Java		98.1
3. Python		98.0
4. C++		95.9
5. R		87.9
6. C#		86.7
7. PHP		82.8
8. JavaScript		82.2
9. Ruby		74.5
10. Go		71.9

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

R and Python: The Numbers

Popularity Rankings

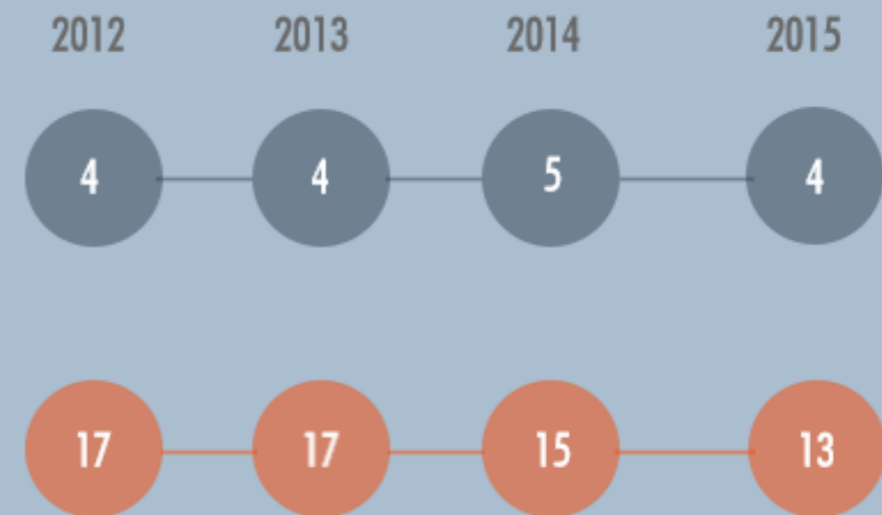
R and Python's popularity between 2013 and February 2015 (Tiobe Index)



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

Python

R



Jobs And Salary?

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



\$ 115,531



\$94,139

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 website interface. At the top, there is a navigation bar with the DataCamp logo, links for 'Learn', 'Groups', and 'About', and a user profile section showing '1,250 XP' and a notification icon with a red circle around it and a '3' badge. A dropdown menu is open from the notification icon, listing several 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), and 'You have a new assignment: Orientation' (9 months ago). At the bottom of the dropdown is a 'See all notifications' button. The main content area features a section titled 'Your Latest Activity' with a card for 'Introduction to Spark in R using...' showing progress and a message: 'You are doing awesome barryus! So far you've earned 250 XP!'. Below this, it says 'The last chapter you were working on was Light My Fire: Starting To Use Spark With dplyr Syntax'. At the bottom, there is a 'DAILY PRACTICE' section with the text: 'Learning data science requires practice every day. Build your data science fluency with DataCamp practice mode.'

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

The image shows a browser window displaying a DataCamp course page on the left and an RStudio IDE interface on the right.

Course Page (Left):

- Page title: "What is an IDE anyway? | R"
- URL: <https://campus.datacamp.com/courses/working-with-the-rstudio-ide-part-1/orientation?ex=2>
- Course title: "What is an IDE anyway?" (50xp)
- Text: "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?"
- Section: "Possible Answers"
- Options:
 - Intensive Design Environment
 - Integrated Document Environment
 - Independent Developer Ecosystem
 - Integrated Development Environment
- Buttons: "Take Hint (-15xp)" and "Submit Answer"

RStudio IDE (Right):

- Menu: File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help
- Version: R version 3.3.1 (2016-06-21) -- "Bug in Your Hair"
- Copyright: Copyright (C) 2016 The R Foundation for Statistical Computing
- Platform: x86_64-pc-linux-gnu (64-bit)
- Text: "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: Global Environment (Environment is empty)
- Files: Home directory view

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

The screenshot shows a web browser window displaying a DataCamp course page. The URL is <https://campus.datacamp.com/courses/working-with-the-rstudio-ide-part-1/orientation?ex=2>. The page features a blue header with the DataCamp logo and a 'Course Outline' button. A dark grey notification box on the left contains the text 'Exercise Completed' with a blue checkmark and '50xp' circled in red. Below this, it says 'Nice job! Move onto the next video to start learning more about the RStudio IDE!' and a yellow 'Continue' button is also circled in red. A smaller box below the notification says 'Become a power user!' and lists the keyboard shortcut 'Submit Answer Ctrl + Shift + Enter' with 'See all keyboard shortcuts' as a link. On the right, an RStudio terminal window is open, showing the R version 3.3.1 (2016-06-21) and the R Foundation copyright notice. The terminal text includes: '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. > |' The RStudio interface also shows an empty Environment pane and a file explorer pane.

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

Homework assignments will be via DataCamp

The screenshot shows a DataCamp exercise page for 'PCA analysis'. The left sidebar contains the exercise title and instructions. The main content area shows the R code editor with a script.R file containing code for transforming normalized counts and plotting PCA. The R Console at the bottom shows the execution of the code, resulting in an error: 'Error: object 'vsd_smoc2' not found'. The interface includes a 'Run Code' button and a 'Submit Answer' button.

DataCamp

Exercise

PCA analysis

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.

Take Hint (-15 XP)

```
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 plotPCA(vsd_smoc2, intgroup=___)
```

Run Code Submit Answer

R Console Slides

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


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

Back to My Dashboard

Foundations of Bioinformatics (BGGN-213)

Leaderboard | My Assignments

30 Days | [90 Days](#) | [Last Year](#)

	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 <i>what, why</i> and <i>how</i> 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...

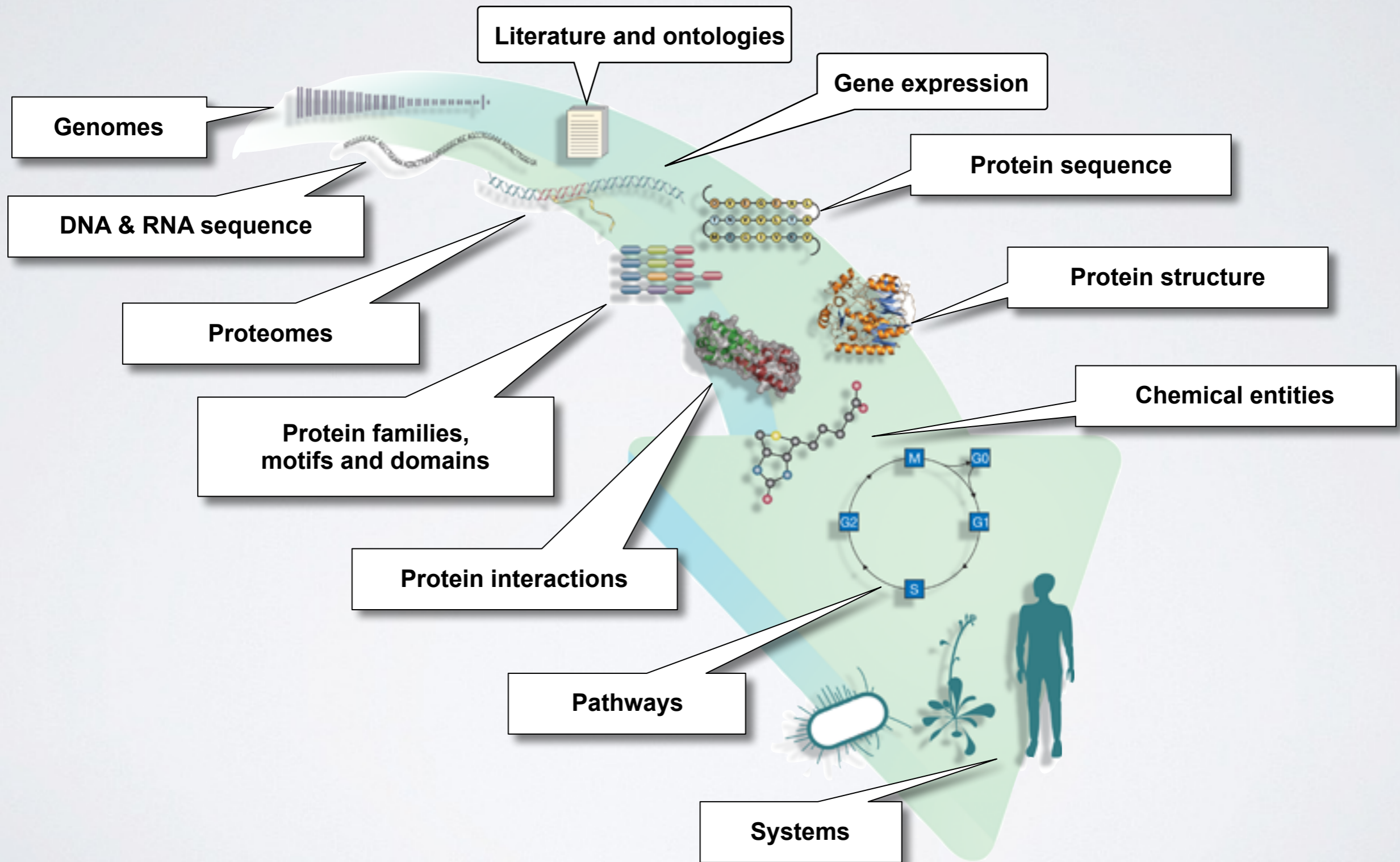
- "Computer based **management** and **analysis** of biological and biomedical data with useful applications in many disciplines, particularly **genomics**, **proteomics**, **metabolomics**, and related fields."
(BIMM-143)
- "Bioinformatics is conceptualizing biology in terms of **macromolecules** and 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...

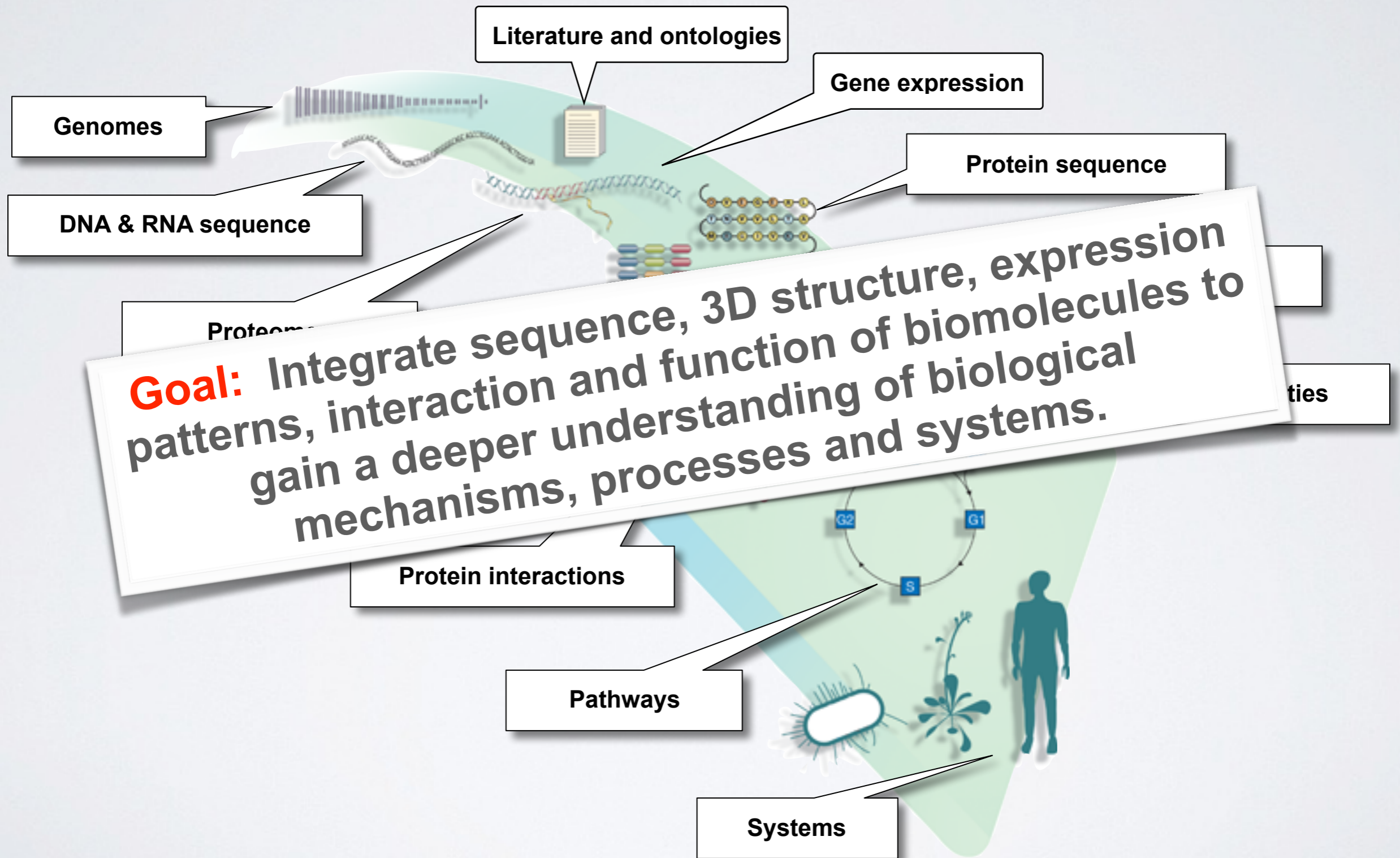
- "Computer based **management** and **analysis** of biological and biomedical data with useful applications in many disciplines, particularly **genomics**, **proteomics**, **metabolomics**, and related fields."
(BIMM-143)
- "Bioinformatics is conceptualizing biological data of **macromolecules** and 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 genome-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

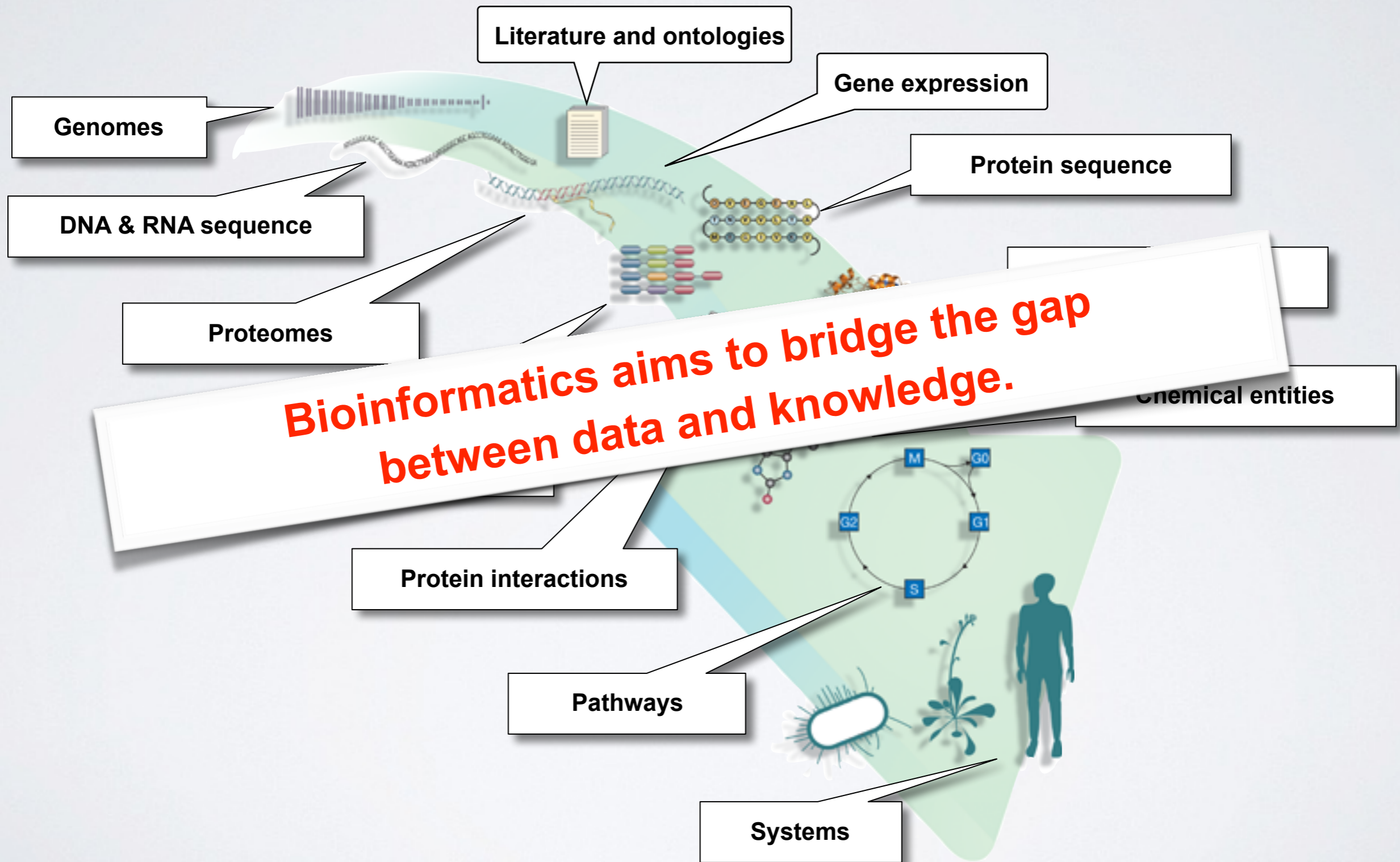
Major types of Bioinformatics Data



Major types of Bioinformatics Data

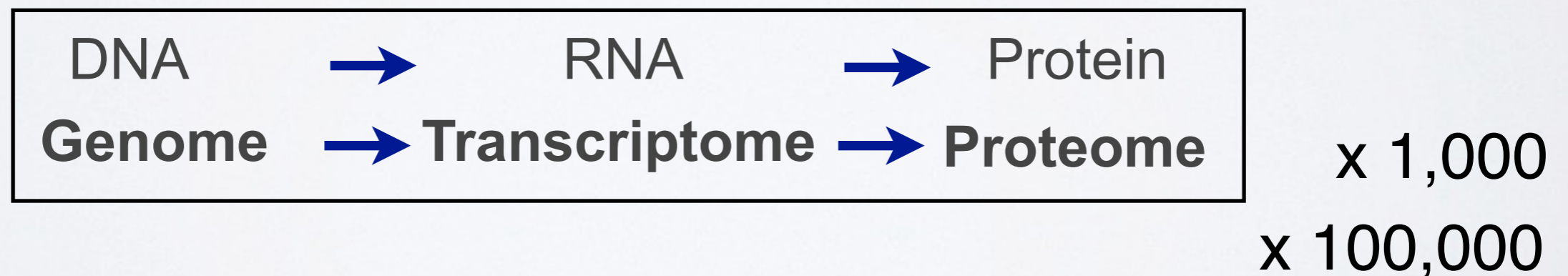


Major types of Bioinformatics Data



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

How do we *actually* do Bioinformatics?

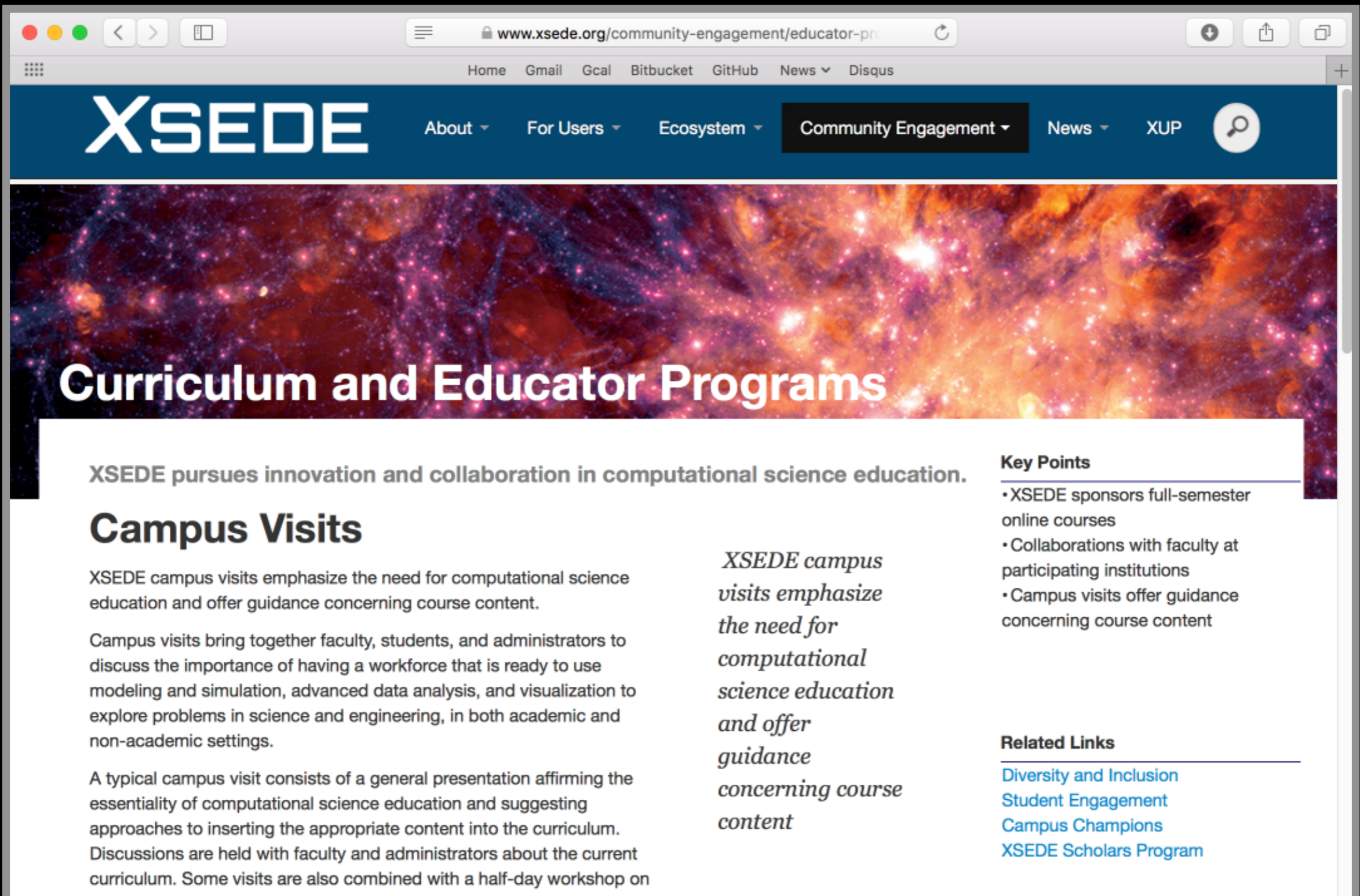
Pre-packaged tools and databases

- 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...

NSF Extreme Science and Engineering Discovery Environment (XSEDE)



The screenshot shows a web browser window with the URL www.xse.de.org/community-engagement/educator-pro. The browser's address bar and navigation buttons are visible. The website's header is dark blue with the XSEDE logo on the left and navigation links: Home, Gmail, Gcal, Bitbucket, GitHub, News, and Disqus. A secondary navigation bar contains: About, For Users, Ecosystem, Community Engagement (highlighted), News, and XUP, along with a search icon. The main content area features a large, vibrant image of a galaxy or nebula. Below this image is the section title 'Curriculum and Educator Programs'. The text below the title states: 'XSEDE pursues innovation and collaboration in computational science education.' The 'Campus Visits' section includes a paragraph about the need for computational science education and offers guidance on course content. A quote states: 'XSEDE campus visits emphasize the need for computational science education and offer guidance concerning course content.' The 'Key Points' section lists: 'XSEDE sponsors full-semester online courses', 'Collaborations with faculty at participating institutions', and 'Campus visits offer guidance concerning course content'. The 'Related Links' section includes: 'Diversity and Inclusion', 'Student Engagement', 'Campus Champions', and 'XSEDE Scholars Program'.

www.xse.de.org/community-engagement/educator-pro

Home Gmail Gcal Bitbucket GitHub News Disqus

XSEDE About For Users Ecosystem Community Engagement News XUP

Curriculum and Educator Programs

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

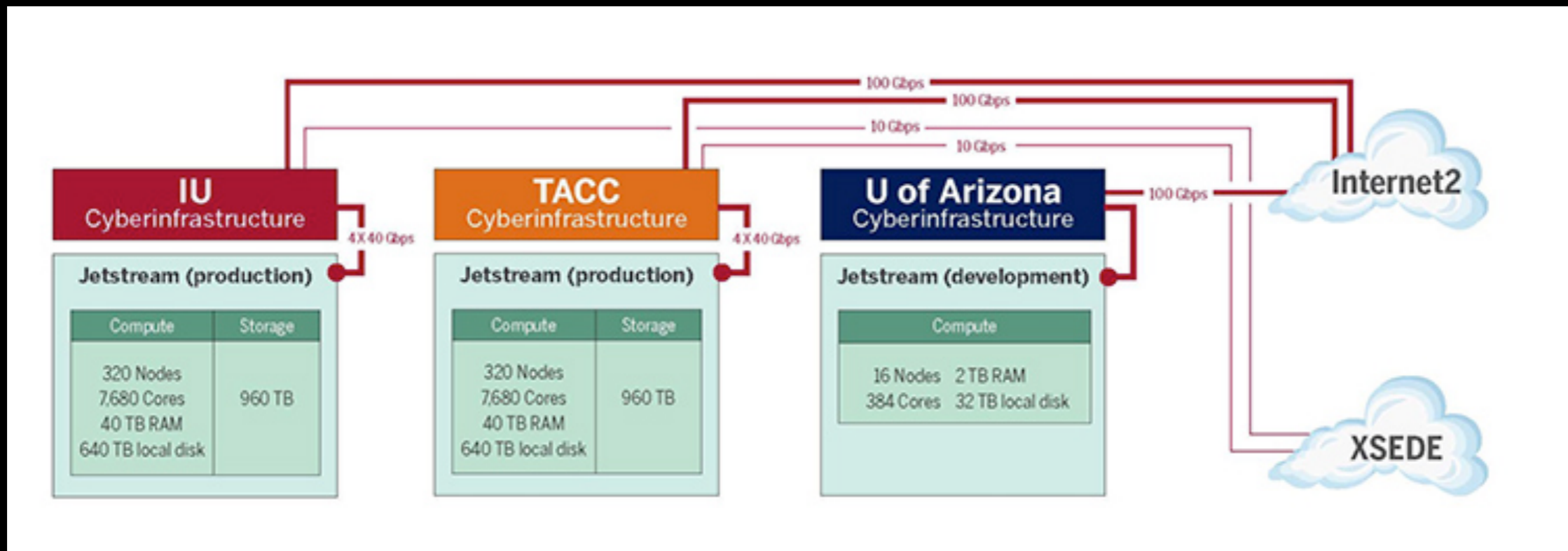
- XSEDE sponsors full-semester online courses
- Collaborations with faculty at participating institutions
- Campus visits offer guidance concerning course content

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 Select the maximum number of aligned sequences to display

Short queries Automatically adjust parameters for short input sequences

Expect threshold

Word size

Max matches in a query range

Scoring Parameters

Matrix

Gap Costs Existence: 11 Extension: 1

Compositional adjustments

Filters and Masking

Filter Low complexity regions

Mask Mask for lookup table only
 Mask lower case letters

PSI/PHI/DELTA BLAST

Upload PSSM no file selected

PSI-BLAST Threshold

Pseudocount

Even Blast has many settable parameters

Related tools with different terminology

STEP 3 - Set your parameters

PROGRAM

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

Key Online Bioinformatics Resources: NCBI & EBI

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

The screenshot shows the NCBI website homepage. At the top, there is a navigation bar with 'NCBI Resources' and 'How To' menus, and a search bar. The main content area is divided into several sections: 'Welcome to NCBI' with a brief description of the center's mission; 'Get Started' with links to 'Tools', 'Downloads', 'How-To's', and 'Submissions'; 'Popular Resources' listing services like PubMed, Bookshelf, and BLAST; and '3D Structures' with a visual representation of a protein structure. A sidebar on the left provides a 'Resource List (A-Z)' with categories like 'Chemicals & Bioassays', 'Data & Software', and 'Genomes & Maps'.

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

The screenshot shows the EMBL-EBI website homepage. The header features the EMBL-EBI logo and navigation links for 'Services', 'Research', 'Training', and 'About us'. The main heading reads 'The European Bioinformatics Institute' and 'Part of the European Molecular Biology Laboratory'. Below this, there is a search bar with the text 'Find a gene, protein or chemical:' and a search button. The page is organized into a grid of content blocks, including 'Services', 'Research', 'Training', 'Industry', 'European Coordination', and 'News from EMBL-EBI'. On the right side, there are sections for 'Popular' links, 'Visit EMBL.org' with a 40th anniversary logo, and 'Upcoming events' such as the 'Plant and Animal Genome conference (PAG XXIV)' 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

www.ncbi.nlm.nih.gov

NCBI Resources How To Sign in to NCBI

NCBI National Center for Biotechnology Information

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>

The image shows a screenshot of the National Center for Biotechnology Information (NCBI) website. The browser address bar displays 'www.ncbi.nlm.nih.gov'. The page features a navigation menu on the left with categories like 'NCBI Home', 'Resource List (A-Z)', and 'All Resources'. The main content area includes a 'Welcome to NCBI' message and a 'Get Started' section with links to 'Tools', 'Downloads', 'How-To's', and 'Submissions'. A 'Popular Resources' box is overlaid on the right side of the page, listing various services: PubMed, Bookshelf, PubMed Central, PubMed Health, BLAST, Nucleotide, Genome, SNP, Gene, Protein, and PubChem. Red arrows point to PubMed, BLAST, and SNP, while a red bracket groups Nucleotide, Genome, SNP, Gene, and Protein.

National Center for Biotechnology Information

www.ncbi.nlm.nih.gov

NCBI Resources How To Sign in to NCBI

NCBI National Center for Biotechnology Information

All Databases Search

Popular Resources

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

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 provides access to a wide range of biological and health-related information.

[About the NCBI](#) | [Mission](#) | [Our Services](#)

Get Started

- [Tools](#): Analyze data using NCBI tools
- [Downloads](#): Get NCBI data
- [How-To's](#): Learn how to access NCBI resources
- [Submissions](#): Submit data to 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.

Resources

Central Health

Announcements

New version of Genome Workbench available

06 Sep

An integrated, downloadable application

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

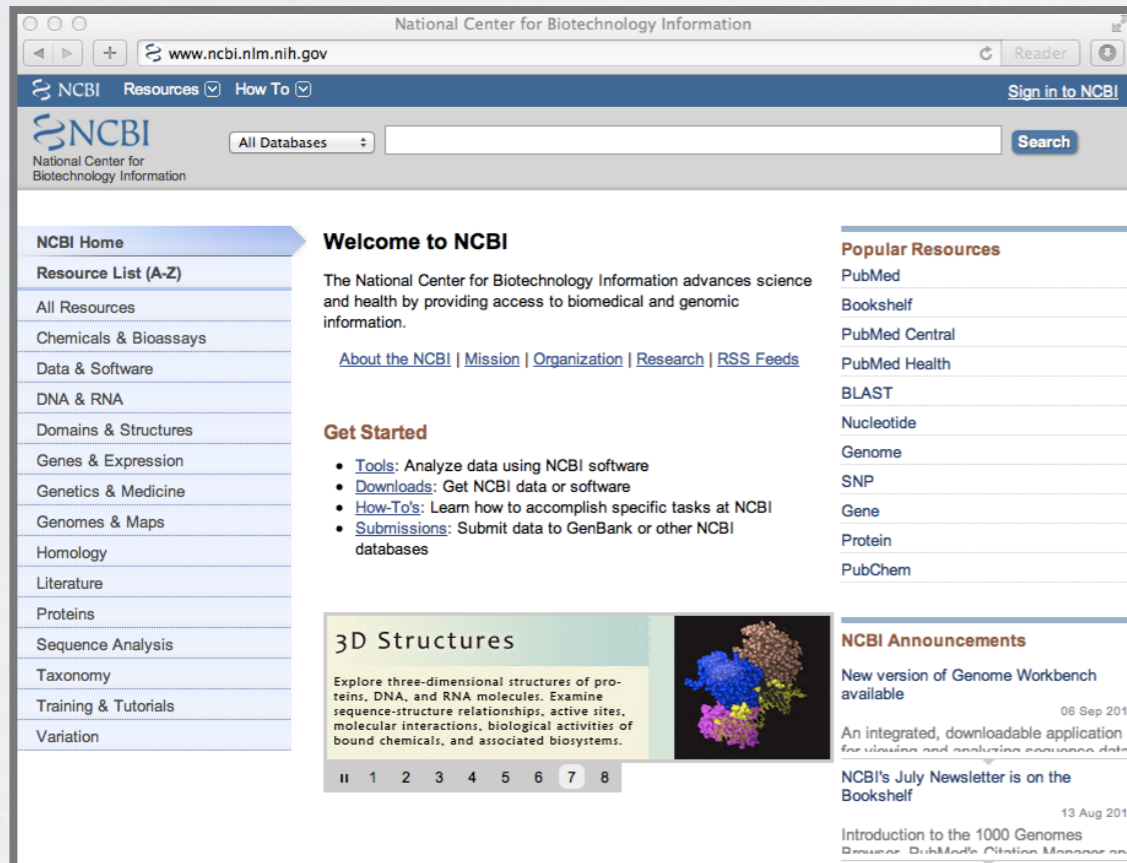
The screenshot shows the NCBI website homepage. At the top, there is a navigation bar with "NCBI", "Resources", and "How To" menus, and a "Sign in to NCBI" link. Below this is a search bar with a dropdown menu set to "All Databases" and a "Search" button. The main content area features a "Welcome to NCBI" message, a "Resource List (A-Z)" link, and a "Popular Resources" section with a link to "PubMed".

Notable NCBI databases include:
GenBank, **RefSeq**, PubMed, dbSNP
and the search tools **ENTREZ** and **BLAST**

This screenshot shows a section of the NCBI website with a sidebar on the left containing links to "Homology", "Literature", "Proteins", "Sequence Analysis", "Taxonomy", "Training & Tutorials", and "Variation". The main content area is titled "databases" and features a "3D Structures" section with a description: "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." To the right of this text is a 3D molecular model. Further right, there are links for "Protein" and "PubChem". At the bottom right, there is an "NCBI Announcements" section with a headline: "New version of Genome Workbench available" dated "06 Sep" and a sub-headline: "An integrated, downloadable applicati".

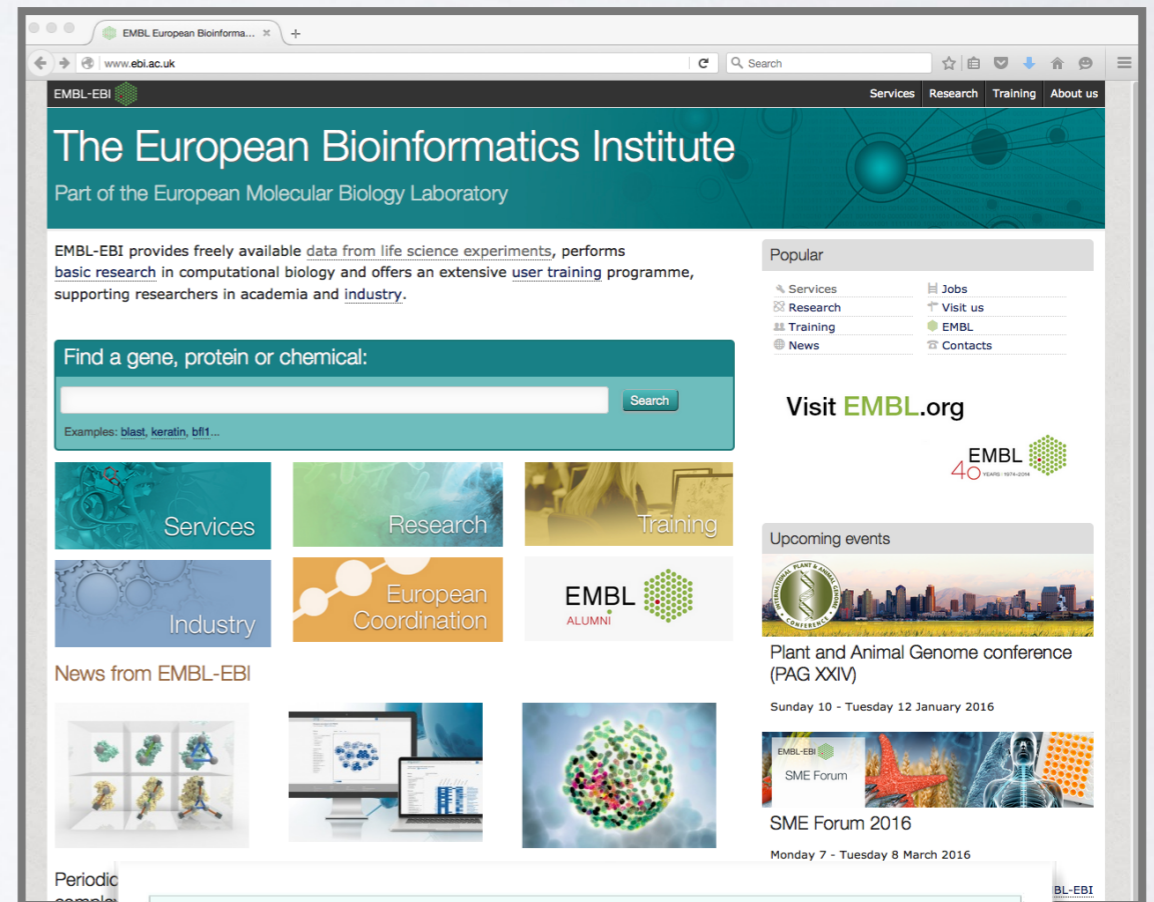
Key Online Bioinformatics Resources: NCBI & EBI

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



The screenshot shows the NCBI website homepage. The browser address bar displays 'www.ncbi.nlm.nih.gov'. The page features a navigation menu on the left with categories like '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'. The main content area includes a 'Welcome to NCBI' message, a 'Get Started' section with links to 'Tools', 'Downloads', 'How-To's', and 'Submissions', and a 'Popular Resources' list containing PubMed, Bookshelf, PubMed Central, PubMed Health, BLAST, Nucleotide, Genome, SNP, Gene, Protein, and PubChem. There is also a '3D Structures' section and an 'NCBI Announcements' section.

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



The screenshot shows the EBI website homepage. The browser address bar displays 'www.ebi.ac.uk'. The page features a navigation menu at the top with categories like 'Services', 'Research', 'Training', and 'About us'. The main content area includes a 'Find a gene, protein or chemical:' search bar, a 'Popular' section with links to 'Services', 'Research', 'Training', 'News', 'Jobs', 'Visit us', 'EMBL', and 'Contacts', and a 'Visit EMBL.org' section. There are also sections for 'Upcoming events', 'News from EMBL-EBI', and 'EMBL ALUMNI'.

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

European Bioinformatics Institute (EBI)

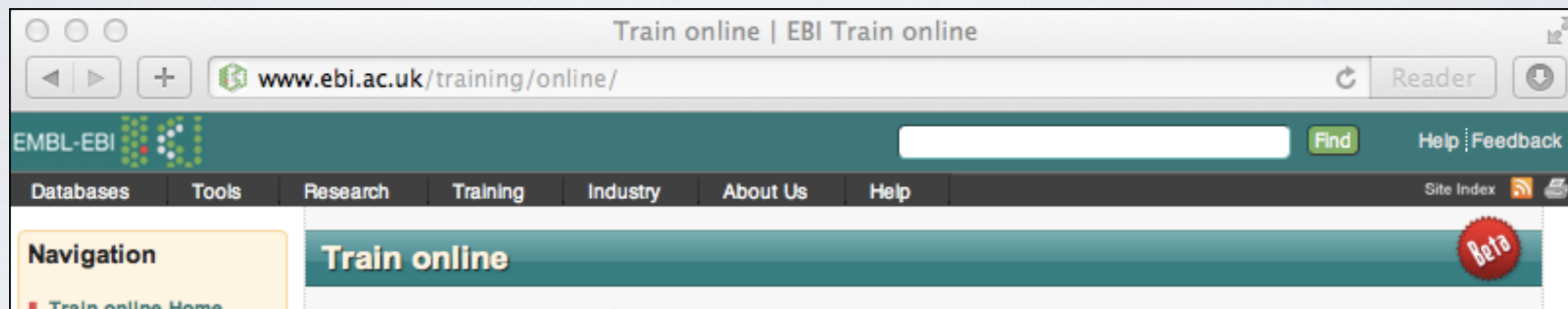
- Created in 1997 as a part of the European Molecular Biology Laboratory (EMBL)
- EBI's mission includes:
 - ▶ providing freely available **data and bioinformatics services**
 - ▶ and providing advanced **bioinformatics training**
- We will cover a number of EBI databases and tools that have advantages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



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

The screenshot shows the EMBL-EBI website homepage. At the top, the browser address bar displays 'www.ebi.ac.uk'. The main header features the EMBL-EBI logo and navigation links for 'Services', 'Research', 'Training', and 'About us'. A large teal banner contains the text 'The European Bioinformatics Institute' and 'Part of the European Molecular Biology Laboratory'. Below this, a paragraph states: '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.' A search bar is provided with the prompt 'Find a gene, protein or chemical:' and a 'Search' button. Below the search bar, there are six colored tiles: 'Services' (teal), 'Research' (green), 'Training' (yellow), 'Industry' (blue), 'European Coordination' (orange), and 'EMBL ALUMNI' (white with green logo). The 'Services' and 'Training' tiles are highlighted with red boxes. On the right side, a 'Popular' section lists links for 'Services', 'Research', 'Training', 'News', 'Jobs', 'Visit us', 'EMBL', and 'Contacts'. Below this is a 'Visit EMBL.org' section with the EMBL 40th anniversary logo (1974-2014). The 'Upcoming events' section features a banner for the 'Plant and Animal Genome conference (PAG XXIV)' held from Sunday 10 to Tuesday 12 January 2016. The bottom of the page shows a 'News from EMBL-EBI' section with several small image thumbnails.

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



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, Biolmage, 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, DPlInteract, 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-U's, 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, TelDB, 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, BCCP, Beanref, Biolmage, BioMagResBank, BIOMDB, BLOCKS, BovGBASE, BOVM, TKbase, CANSITE, CarbBank, CARBHYD, CATH, CAZY, AP, ChickGBASE, Colibri, COPE, CottonDB, bEST, dbSTS, DDBJ, DGP, DictyDb, CDC, ECGC, EC02DBASE, OTHER, FlyBase, Link, G, HAEMB, H, HZRGbase, IMG, Kabat, KDNA, K, DB, Medline, Mendel, MEROPS, MGDB, MGI, MHC, MAP, MJDB, MmtDB, Mol-R-U, MPDB, MRR, MutBase, Myc, O-lycBase, OMIA, OMIM, OPD, ORDB, OWL, PAHdb, PatBase, PD, 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 !!!!

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 <i>what</i> , <i>why</i> and <i>how</i> of bioinformatics?
Bioinformatics Database	Hands-on exploration of several major databases and their associated tools.

Hands-on section

Your Turn!

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

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

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:

- 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

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

```
ATGGTGCATCTGACTCCTGTGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAG
TTGGTGGTGAGGCCCTGGGCAGGCTGCTGGTGGTCTACCCTTGGACCCAGAGGTTCTTTGAGTCCTTTGG
GGATCTGTCCACTCCTGATGCAGTTATGGGCAACCCTAAGGTGAAGGCTCATGGCAAGAAAGTGCTCGGT
GCCTTTAGTGATGGCCTGGCTCACCTGGACAACCTCAAGGGCACCTTTGCCACACTGAGTGAGCTGCACT
GTGACAAGCTGCACGTGGATCCTGAGAACTTCAGGCTCCTGGGCAACGTGCTGGTCTGTGTGCTGGCCCA
TCACTTTGGCAAAGAATTCACCCCACAGTGCAGGCTGCCTATCAGAAAGTGGTGGCTGGTGTGGCTAAT
GCCCTGGCCCAAGTATCACTAAGCTCGCTTTCTTGCTGTCCAATTT
```

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:

