

**BGGN 213**  
**Foundations of Bioinformatics**  
 Barry Grant  
 UC San Diego  
<http://thegrantlab.org/bgg213>

**HELLO**  
my name is  
**BARRY**  
[bjgrant@ucsd.edu](mailto:bjgrant@ucsd.edu)

**HELLO**  
HIS name is  
**KEVIN**  
[kkchau@ucsd.edu](mailto: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

<b>Course Logistics</b>	Website, screencasts, survey, ethics, assessment and grading.
<b>Learning Objectives</b>	What you need to learn to succeed in this course.
<b>Course Structure</b>	Major lecture topics and specific learning goals.
<b>Introduction to Bioinformatics</b>	Introducing the <i>what, why</i> and <i>how</i> of bioinformatics?
<b>Bioinformatics Database</b>	<b>Hands-on</b> exploration of several major databases and their associated tools.

http://thegrantlab.org/bggn213/

UC San Diego

## 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**

### Bioinformatics (BGGN 213, Spring 2018)

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

**Instructional Assistant**  
Yuansheng Zhou (Email: [yuz461@ucsd.edu](mailto: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?

UC San Diego

## 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**

### 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!

**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

**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	<a href="#">Welcome to Bioinformatics</a> 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	<a href="#">Sequence alignment fundamentals, algorithms and applications</a> 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

**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	<a href="#">Welcome to Bioinformatics</a> 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	<a href="#">Sequence alignment fundamentals, algorithms and applications</a> 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

**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**

The screenshot shows a GitHub repository page for BGGN213. The left sidebar contains a navigation menu with items like Overview, Lectures, Computer Setup, Learning Goals, Assignments & Grading, Ethics Code, and Screen Cast Videos. The main content area is titled 'Homework:' and includes a 'Questions' section with a red box around it, a 'Readings' section with links to PDFs and a New York Times article, and a 'Screen Casts' section featuring a video thumbnail for 'Foundations of Bioinformatics' by Barry Grant.

# Homework

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

The screenshot shows a Google Docs form titled 'BGGN213 Lecture 1 Homework'. A red banner across the top reads 'Homework is due before the next weeks class!'. The form asks the user to provide their name/email address and answer a question: 'Which of the following operating systems is most frequently used for bioinformatics tool development?'. The options are Windows, iOS, Unix, and Perl.

# Projects

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

The screenshot shows a GitHub repository page for a supervised learning mini-project. The left sidebar contains a navigation menu with items like Overview, Lectures, Computer Setup, Learning Goals, Assignments & Grading, Ethics Code, and Screen Cast Videos. The main content area is titled '9: Unsupervised learning mini-project' and includes 'Topics', 'Goals', and 'Material' sections with links to lecture slides, lab worksheets, and data files.

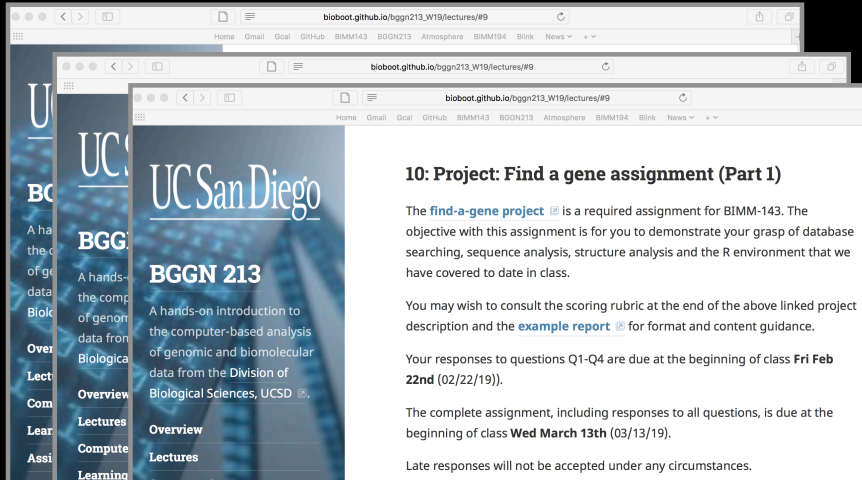
# Projects

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

The screenshot shows a GitHub repository page for a cancer genomics project. The left sidebar contains a navigation menu with items like Overview, Lectures, Computer Setup, Learning Goals, Assignments & Grading, Ethics Code, and Screen Cast Videos. The main content area is titled '18: Cancer genomics' and includes 'Topics', 'Material', and 'N.B.' sections with links to lecture slides, lab worksheets, and data files.

# Projects

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



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

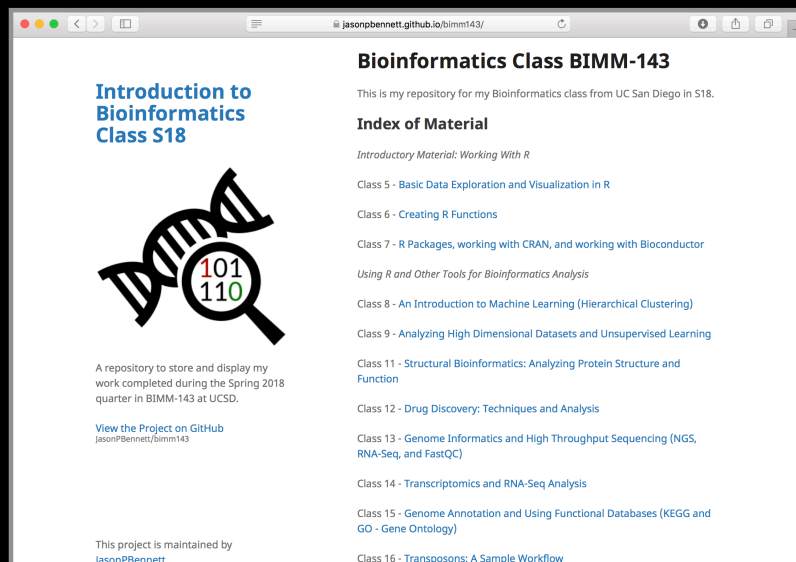
The complete assignment, including responses to all questions, is due at the beginning of class **Wed March 13th (03/13/19)**.

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!



**Bioinformatics Class BIMM-143**

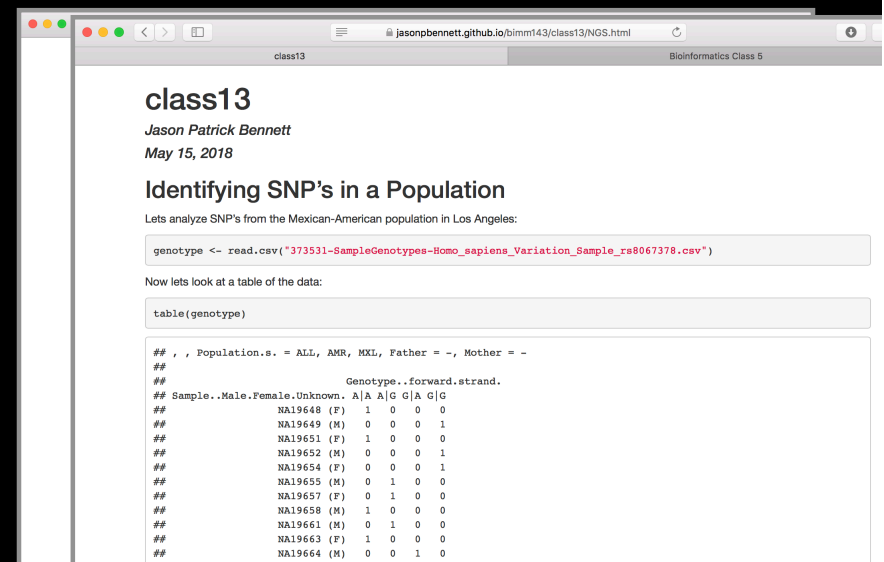
This is my repository for my Bioinformatics class from UC San Diego in S18.

**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

Online portfolio of **your** bioinformatics work!



**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:

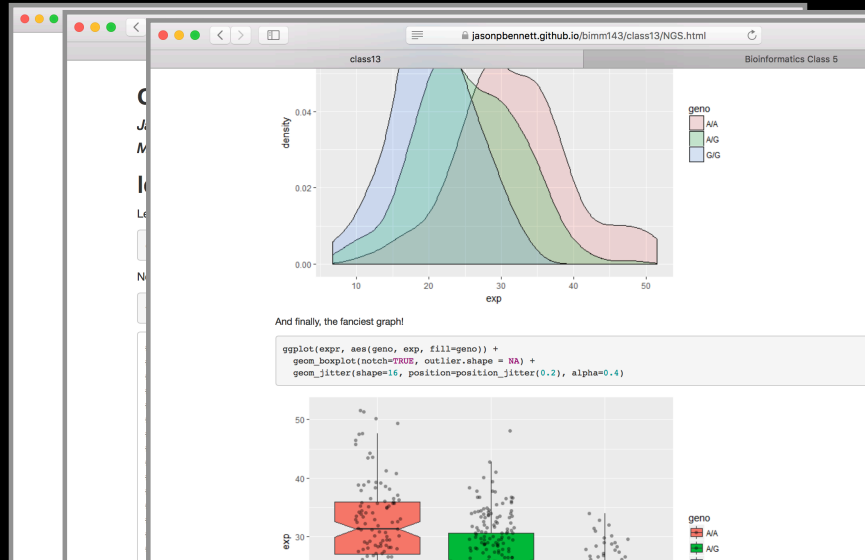
```
genotype <- read.csv("373531-SampleGenotypes-Homo_sapiens_Variation_Sample_rs8067378.csv")
```

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 (M) 1 1 1 1
```

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



## Bonus:

### Bioinformatics & Genomics in industry

**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 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!**

## BGGN-213 Learning Goals....

### Advanced UNIX and R based learning goals

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	13, 14, 15

# BGGN-213 Learning Goals....

Delve deeper into “real-world” bioinformatics

UC San Diego  
**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

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

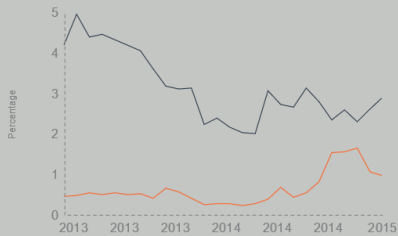
Language Rank	Types	Spectrum Ranking
1. C		100.0
2. Java		98.1
3. Python		98.0
4. C++		95.9
5. R		87.9
6. C#		86.7
7. PHP		82.8
8. JavaScript		82.2
9. Ruby		74.5
10. Go		71.9

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

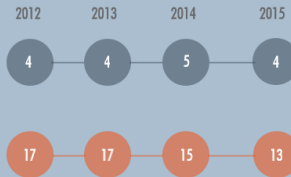
# R and Python: The Numbers

## Popularity Rankings

R and 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)



## Jobs And Salary?

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



\$115,531



\$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 website interface. A notification bell icon in the top right corner is circled in red, and a dropdown menu is open, listing several assignments with their completion dates. The main content area shows a course titled 'Introduction to Spark in R using...' and a 'DAILY PRACTICE' section at the bottom.

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

The screenshot shows the RStudio IDE interface. The 'Possible Answers' section for the question 'What is an IDE anyway?' is visible. The 'Integrated Development Environment' option is selected and circled in red. A 'Submit Answer' button is also circled in red. The console shows the R version and platform information.



< <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 dark sidebar on the left with a 'What is an IDE anyway?' title. A large 'Exercise Completed' message is displayed, with a 'Continue' button circled in red. Below the message, there are instructions to 'Submit Answer' using 'Ctrl + Shift + Enter'. The main content area shows a terminal window with R version 3.3.1 information and a file explorer.

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

Homework assignments will be via DataCamp

The screenshot shows a DataCamp exercise page for 'PCA analysis'. The URL is <https://www.datacamp.com/exercises/pca-analysis>. The page includes a 'script.R' editor with the following code:

```
1 # Transform the normalized counts
2 vsd_smc2 <- vst(dds_smc2, blind = TRUE)
3
4 # Plot the PCA of PC1 and PC2
5 ...(..., intgroup=...)
```

Below the code, there are 'Run Code' and 'Submit Answer' buttons. The R console shows an error: `Error: object 'vsd_smc2' not found`. The page also includes instructions and a 'Take Hint (-15 XP)' button.

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

The screenshot shows a DataCamp course page for 'Foundations of Bioinformatics (BGGN-213)'. The URL is <https://www.datacamp.com/groups/foundations-of-bioinformatics-bgg-213/details>. The page features a 'Groups' tab and a 'Leaderboard' section. The leaderboard table is as follows:

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 Schiknis	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

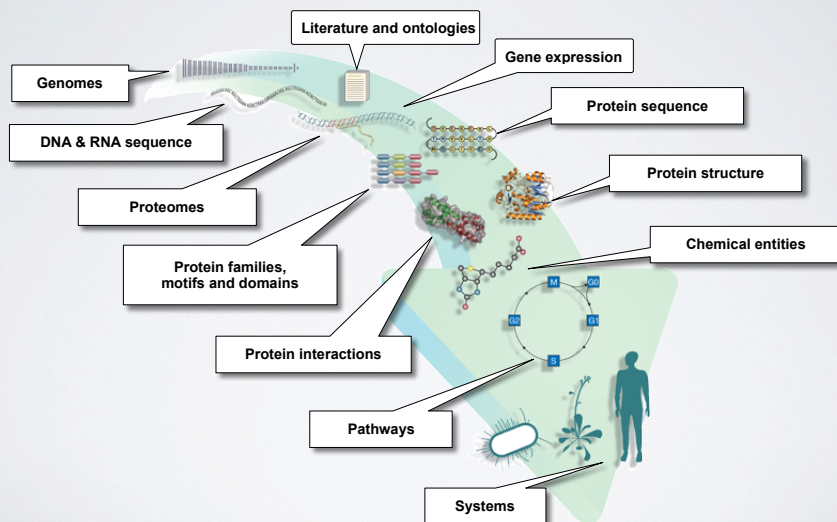
- ▶ “Bioinformatics is conceptualizing biology in terms of **macromolecules** and then applying “informatics” techniques (derived from disciplines such as applied maths, computer science, and statistics) to **understand** and **organize** the information associated with these molecules, on a **large-scale**.  
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

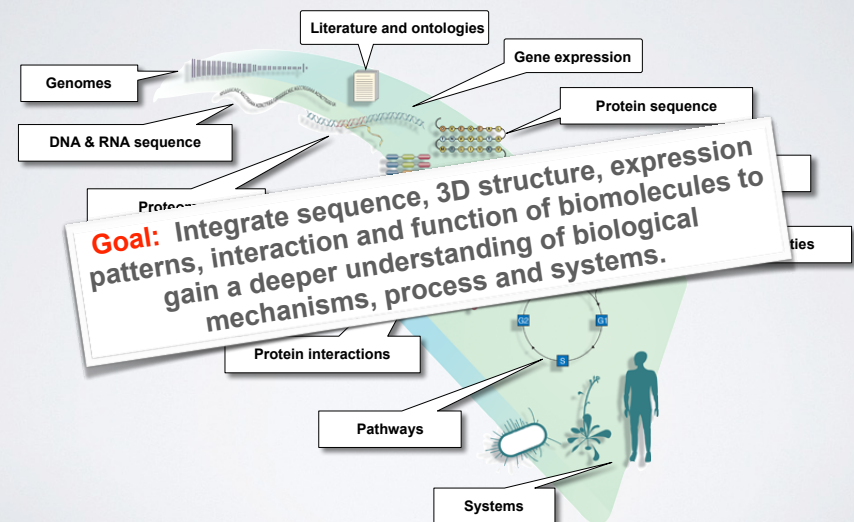
- ▶ “Bioinformatics is conceptualizing biology in terms of **macromolecules** and then applying “informatics” techniques (derived from disciplines such as applied maths, computer science, and statistics) to **understand** and **organize** the information associated with these molecules, on a **large-scale**.  
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> )

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

## Major types of Bioinformatics Data

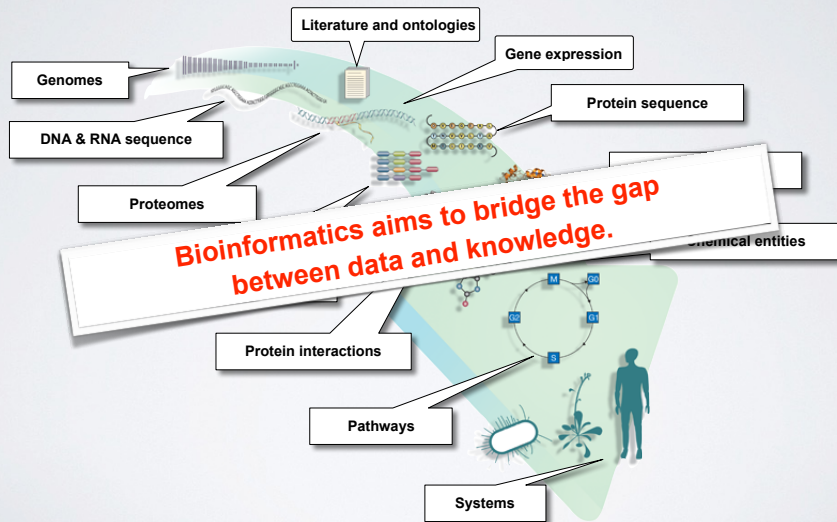


## Major types of Bioinformatics Data



**Goal: Integrate sequence, 3D structure, expression patterns, interaction and function of biomolecules to gain a deeper understanding of biological mechanisms, process and systems.**

## Major types of Bioinformatics Data



## BIOINFORMATICS RESEARCH AREAS

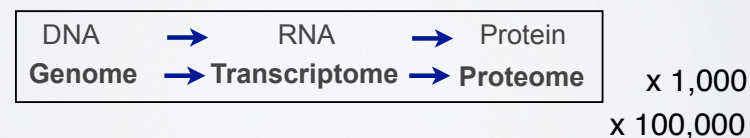
Include but are not limited to:

- 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?

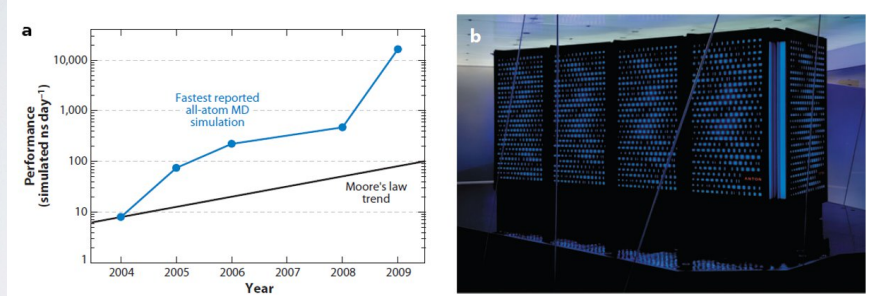
### Pre-packaged tools and databases

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

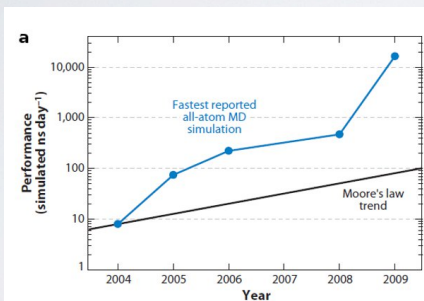
### Advanced tool application & development

- 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 AND GPUS



## SIDE-NOTE: SUPERCOMPUTERS AND GPUS



## HOW COMPUTERS HAVE CHANGED

DATE	COST	SPEED	MEMORY	SIZE
1967	\$409	0.1 MHz	1 MB	HALL
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 shoebox.

# NSF Extreme Science and Engineering Discovery Environment (XSEDE)

The screenshot shows the XSEDE website with a navigation bar and a main content area. The page title is 'Curriculum and Educator Programs'. The main text states: '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

**Key Points**

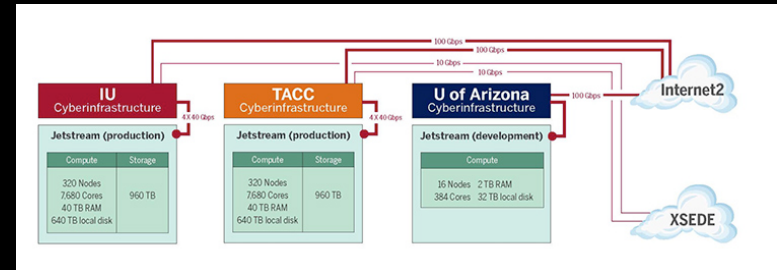
- XSEDE sponsors full-semester online courses
- Collaborations with faculty at participating institutions
- Campus visits offer guidance concerning course content

**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 the Jetstream tutorial page for UC San Diego. The page title is 'Starting a Jetstream Computer Instance!'. The main text states: '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 a 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

**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

## Jetstream tutorials

Developed user friendly labs for Jetstream basics

The screenshot shows the Jetstream tutorial page for UC San Diego. The page title is 'Request to log in to the Jetstream Portal'. The main text states: 'First, go to the Jetstream application at: https://use.jetstream-cloud.org/application. Now click the login link in the upper right.' The screenshot shows a screenshot of the Jetstream portal interface with a red circle around the 'login' link in the upper right corner.

**BGGN 213**

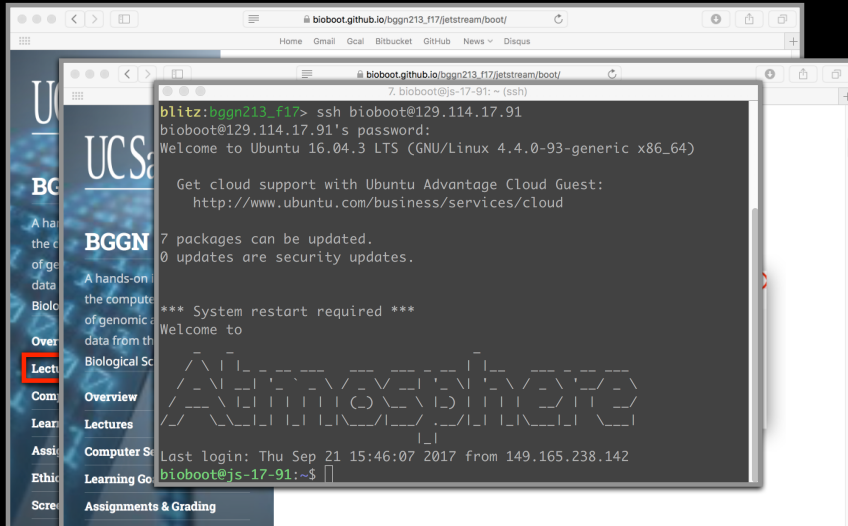
A hands-on introduction to the computer-based analysis of genomic and biomolecular data from the Division of Biological Sciences, UCSD.

**Overview**

- Lectures**
- Computer Setup
- Learning Goals
- Assignments & Grading

# 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)

Protein BLAST: search protein databases using a protein query

blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blast&BLAST\_PROGRAMS=blast&PAGE\_TYPE=BlastSearch&SHOW\_DEFAULTS=on&LINK\_LOC=blasthome

**General Parameters**

- Max target sequences: 500
- Short queries:  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 scoring

**Filters and Masking**

- Filter:  Low complexity regions
- Mask:  Mask for lookup table only,  Mask lower case letters

**PSI/PHI/DELTA BLAST**

- Upload PSSM: Choose File (no file selected)
- PSI-BLAST Threshold: 0.005
- Pseudocount: 0

**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)

**Related tools with different terminology**

DNA STRAND	HISTOGRAM	FILTER	STATISTICAL ESTIMATES
N/A	no	none	Regress

**SCORES**

ALIGNMENTS	SEQUENCE RANGE	DATABASE RANGE	MULTI HSPs
50	START-END	START-END	no

**SCORE FORMAT**

Default

Even Blast has many settable parameters

## Key Online Bioinformatics Resources: NCBI & EBI

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

National Center for Biotechnology Information

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

The European Bioinformatics Institute

Part of the European Molecular Biology Laboratory

<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

[www.ncbi.nlm.nih.gov](http://www.ncbi.nlm.nih.gov)

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 bioassays.

**Popular Resources**

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

**NCBI Announcements**

New version of Genome Workbench available

An integrated, downloadable application



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

The screenshot shows the NCBI homepage with a 'Popular Resources' dropdown menu. The menu items are: PubMed (with a red arrow), Bookshelf, PubMed Central, PubMed Health, BLAST (with a red arrow), Nucleotide (with a red arrow), Genome, SNP, Gene (with a red arrow), Protein (with a red arrow), and PubChem. The background shows the main navigation and search area of the website.

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

The screenshot shows the NCBI homepage with a white text box overlay. The text inside the box reads: 'Notable NCBI databases include: **GenBank**, **RefSeq**, **PubMed**, **dbSNP** and the search tools **ENTREZ** and **BLAST**'. Below the text box, the website's navigation and search area are visible.

## 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 the 'Welcome to NCBI' message and various resource links. The URL <http://www.ncbi.nlm.nih.gov> is displayed at the bottom.

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

The screenshot shows the EBI homepage with the 'The European Bioinformatics Institute' header and various service links. The URL <https://www.ebi.ac.uk> is displayed at the bottom.

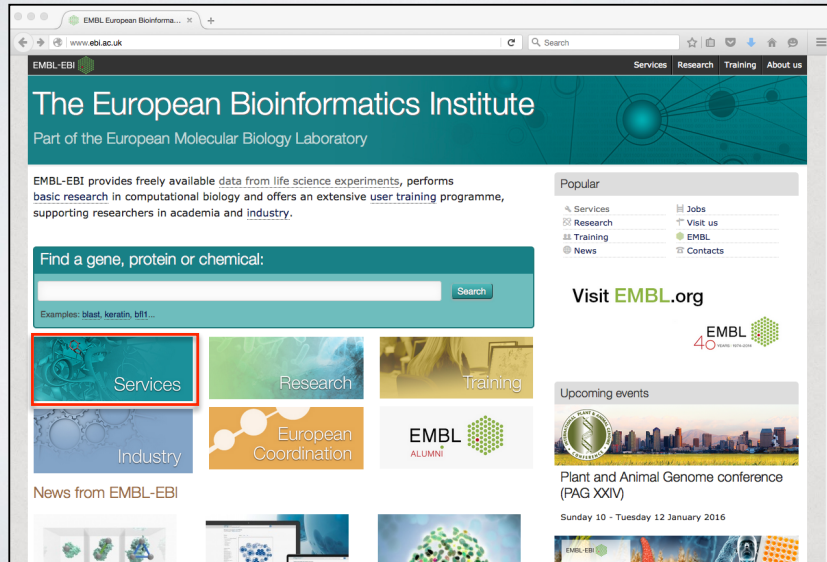
<https://www.ebi.ac.uk>

## European Bioinformatics Institute (EBI)

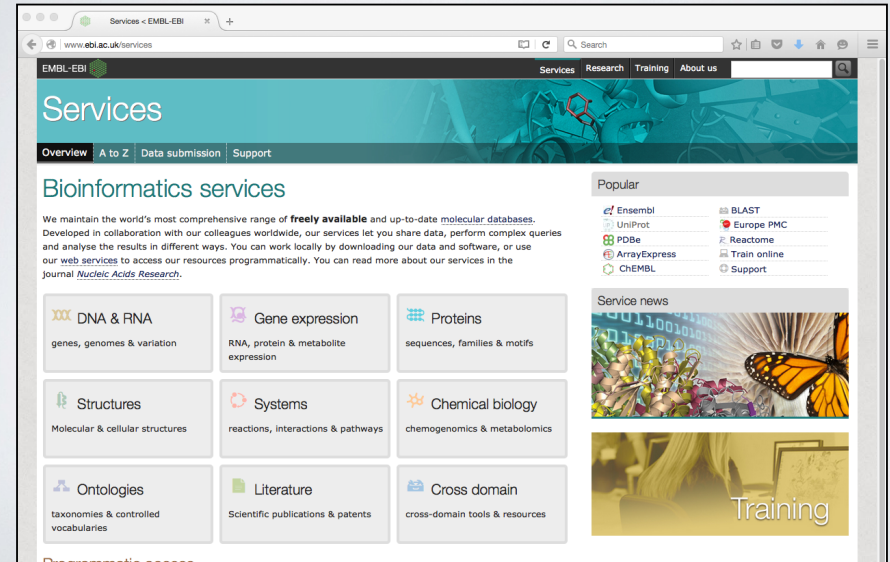
- Created in 1997 as a part of the European Molecular Biology Laboratory (EMBL)
- EBI's mission includes:
  - ▶ providing freely available **data and bioinformatics services**
  - ▶ and providing advanced **bioinformatics training**
- We will briefly cover several EBI databases and tools that have advantages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



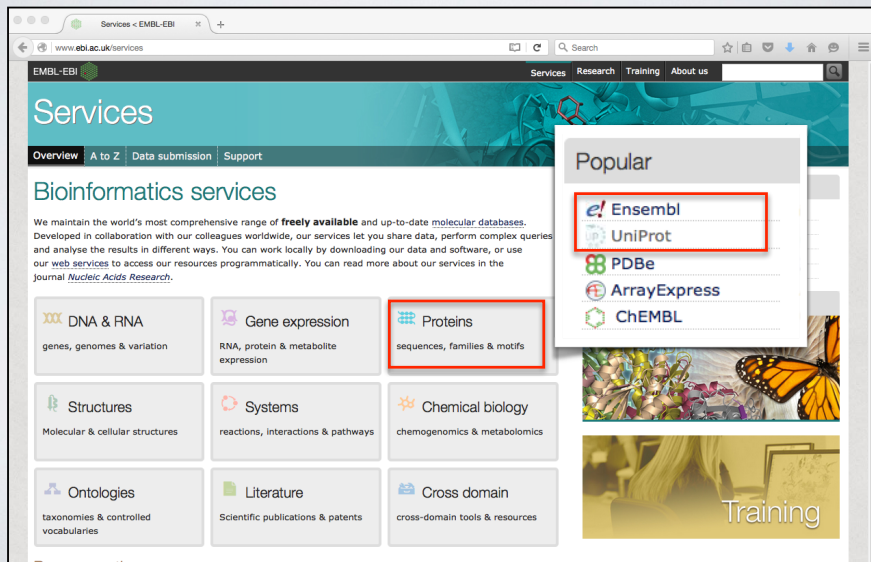
The EBI maintains a number of high quality curated **secondary databases** and associated tools



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

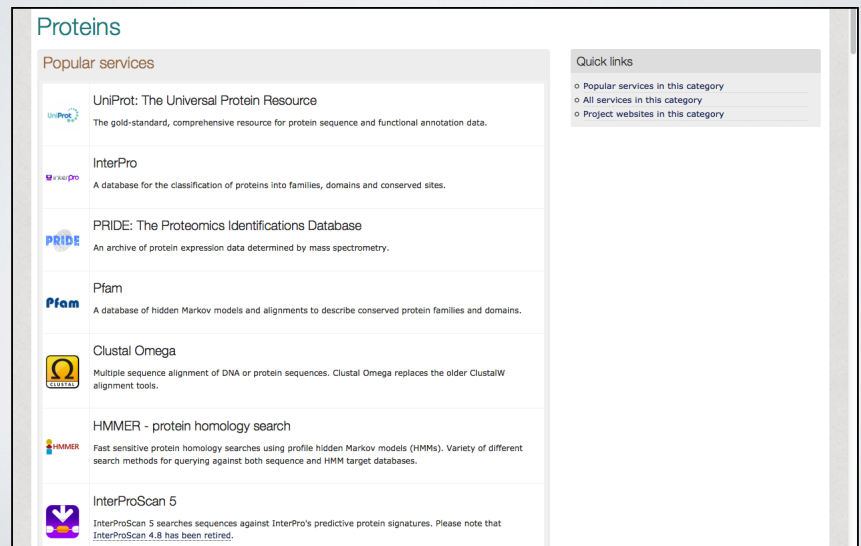


The EBI maintains a number of high quality curated **secondary databases** and associated tools

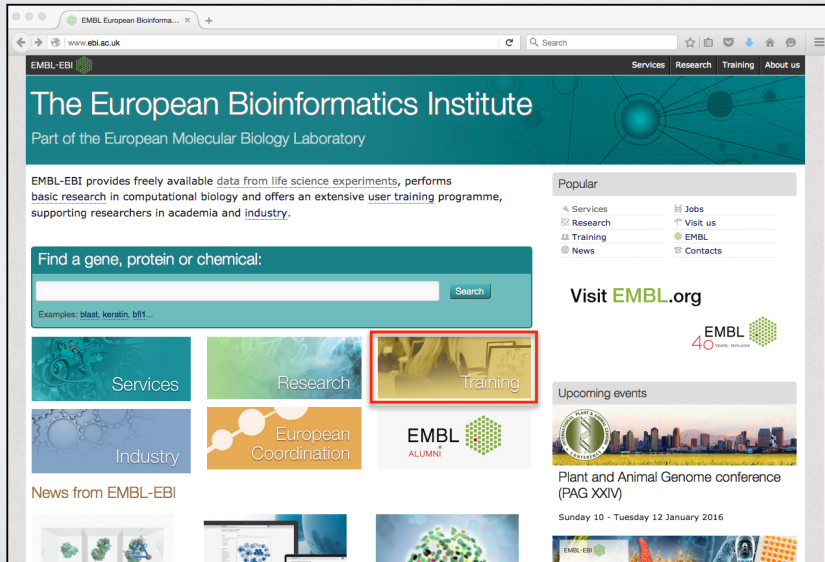


<https://www.ebi.ac.uk>

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



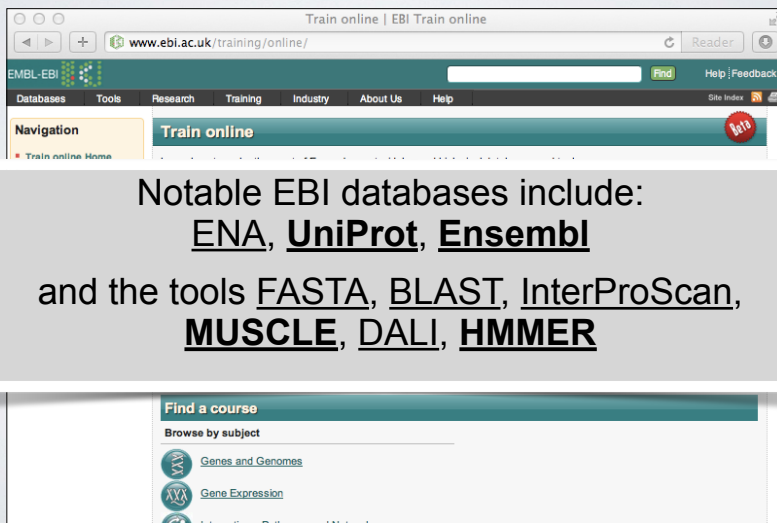
The EBI also provides a growing selection of **online tutorials** on EBI databases and tools



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



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**

**Next Class...**

**MAJOR BIOINFORMATICS  
DATABASES AND ASSOCIATED  
ONLINE TOOLS**

## Bioinformatics Databases

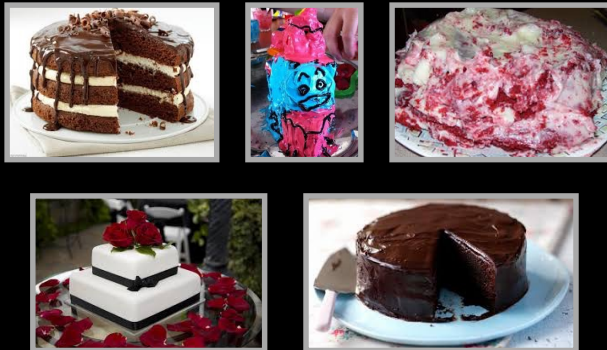
AATDB, AceDb, ACUTS, ADB, AFDB, AGIS, AMSdb, ARR, AsDb, BBDB, BCGD, Bearref, Biolmage, BioMagResBank, BIOMDB, BLOCKS, BovGBASE, BOVMAP, BSORF, BTKbase, CANSITE, CarbBank, CARBHYD, CATH, CAZY, CCDC, CD4OLbase, CGAP, ChickGBASE, Colibri, COPE, CottonDB, CSNDB, CUTG, CyanoBase, dbCFC, dbEST, dbSTS, DDBJ, DGP, DictyDb, Picty\_cDB, DIP, DOGS, DOMO, DPD, 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, HIBD, 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, 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, BCGD, Bearref, 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, 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, HIBD, 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, 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 ..... !!!!

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

# Your Turn!

[https://bioboot.github.io/bgg213\\_S18/lectures/#1](https://bioboot.github.io/bgg213_S18/lectures/#1)

The screenshot shows a web browser displaying the course page for BGGN 213. The page includes a navigation menu on the left with links for Overview, Lectures, Computer Setup, Learning Goals, Assignments & Grading, and Ethics Code. The main content area lists 'Goals' and 'Material'. Two items in the 'Material' list are highlighted with red boxes: 'Lab: Hands-on section worksheet' and 'Feedback: Muddy Point Assessment'.

## BGGN-213: FOUNDATIONS OF BIOINFORMATICS (Lecture 1)

**Bioinformatics Databases and Key Online Resources**  
[https://bioboot.github.io/bgg213\\_S18/lectures/#1](https://bioboot.github.io/bgg213_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
ATGGTGCATCTGACTCTCTGGGAGAAGCTCGCCCTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAG
TTGGTGTGGAGCCCTGGGAGGCTGCTGGTGGTCTACCTTGGACCCAGAGGTTCTTTGAGTCTTTGG
GGACTCTGCTACTCTGCTGCACTTATGGGCAACCTTAAGTGGAGCTCAATGGCAGAAAGTCTCGGT
GCCTTTAGTGATGGCTGCTCACCTGGACAACCTCAAGGGCACTTTGCACACAGTGAAGTGGCTGCACT
GTGACAGCTGACAGTGGATCTGAGAACTTCAGGCTCTGGGCAACGTGCTGCTGTGTGGTGGCCCA
TCACCTTGGCAAGAATTCAACCCACCAGTGCAGGCTGCTATCAGAAAGTGGTGGCTGGTGGCTTAAT
GCCCTGGCCACAAAGTATCACTAAGCTGGCTTTCTGCTGTCAAATTT
```

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:<br>[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/bgg213/>

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

THANK YOU