

**BIMM 143**  
**Introduction to Bioinformatics**  
 Barry Grant  
 UC San Diego  
<http://thegrantlab.org/bimm143>

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

**HELLO**  
*HIS* name is  
**ALEX**  
[ajweitze@ucsd.edu](mailto:ajweitze@ucsd.edu)

**HELLO**  
*HER* name is  
**YUSI**  
[cyusi@ucsd.edu](mailto: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

<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>	<b>Introducing the <i>what</i>, <i>why</i> and <i>how</i> of bioinformatics?</b>
<b>Bioinformatics Database</b>	<b>Hands-on</b> 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**

**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**

**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 the course page for BIMM 143 at UC San Diego. The 'Learning Goals' section is highlighted with a red box. The goals are listed in a table with corresponding lecture numbers.

		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, BLAST, BLAST, BLAST and protein structure based databases	5, 10

## Course Structure

Derived from specific learning goals

The screenshot shows the course page for BIMM 143 at UC San Diego. The 'Lectures' section is highlighted with a red box. The lecture topics for Spring 2018 are listed in a table.

#	Date	Topics for Spring 2018
1	Tu, 04/03	<b>Welcome to Bioinformatics</b> 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	<b>Sequence alignment fundamentals, algorithms and applications</b> 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

The screenshot shows the course page for BIMM 143 at UC San Diego. The 'Lectures' section is highlighted with a red box. The lecture topics for Spring 2018 are listed in a table.

#	Date	Topics for Spring 2018
1	Tu, 04/03	<b>Welcome to Bioinformatics</b> 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	<b>Sequence alignment fundamentals, algorithms and applications</b> 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

**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:**

- 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](#) [↗](#)

# Homework

## Goals, Class material, Screencasts & Homework

**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**

### Homework:

- [Questions](#) [↗](#),
- **Readings:**
  - [PDF1: What is bioinformatics? An introduction and overview](#) [↗](#),
  - [PDF2: Advancements and Challenges in Computational Biology](#) [↗](#),
  - [Other: For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights](#) [↗](#) [New York Times](#), 2014.

**Screen Casts:**

Welcome to "Foundations of Bioinformatics" (BGGN-2...)

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

# Homework

## Goals, Class material, Screencasts & Homework

**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**

### Homework:

- [Questions](#) [↗](#),
- **Readings:**
  - [PDF1: What is bioinformatics? An introduction and overview](#) [↗](#),
  - [PDF2: Advancements and Challenges in Computational Biology](#) [↗](#),
  - [Other: For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights](#) [↗](#) [New York Times](#), 2014.

**Screen Casts:**

Welcome to "Foundations of Bioinformatics" (BGGN-2...)

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 question. Your answer should be submitted to the email address and UCSD PID number so you can receive your grade.

**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.

**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

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.

### 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

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.

### 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).

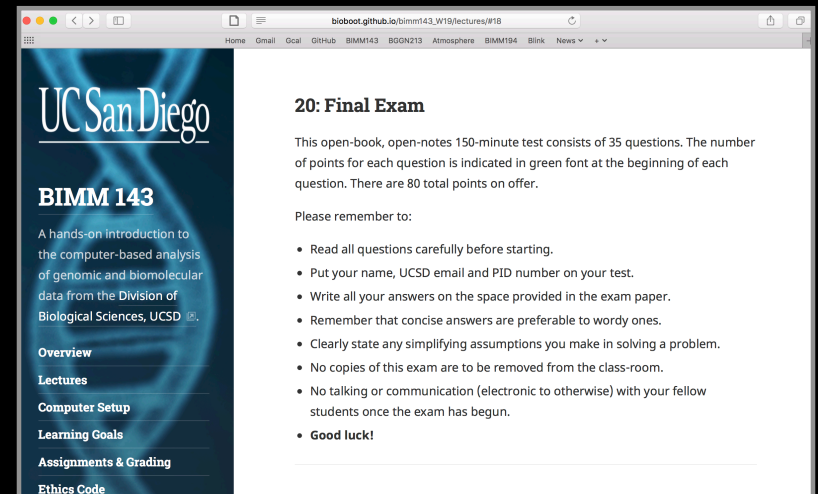
Late responses will not be accepted under any circumstances.

# 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)



**20: Final Exam**

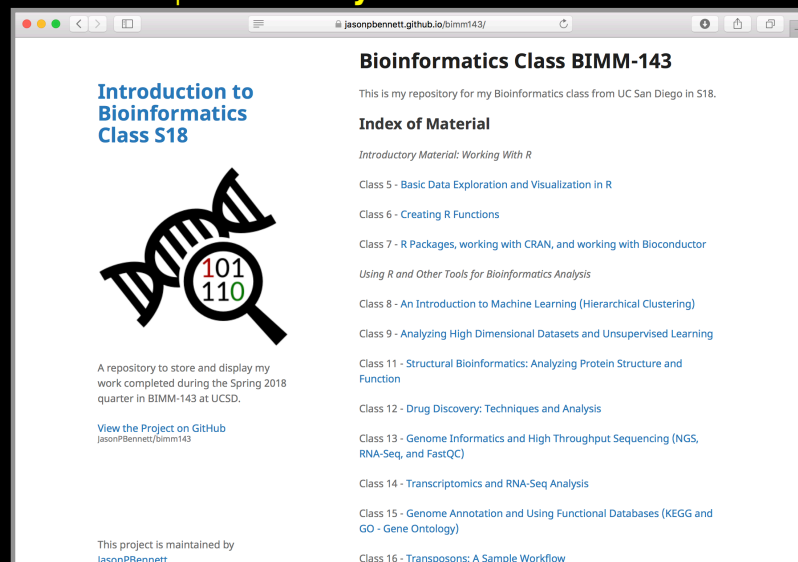
This open-book, open-notes 150-minute test consists of 35 questions. The number of points for each question is indicated in green font at the beginning of each question. There are 80 total points on offer.

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 or otherwise) with your fellow students once the exam has begun.
- **Good luck!**

# Bonus:

Online portfolio of **your** bioinformatics work!



**Bioinformatics Class BIMM-143**

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

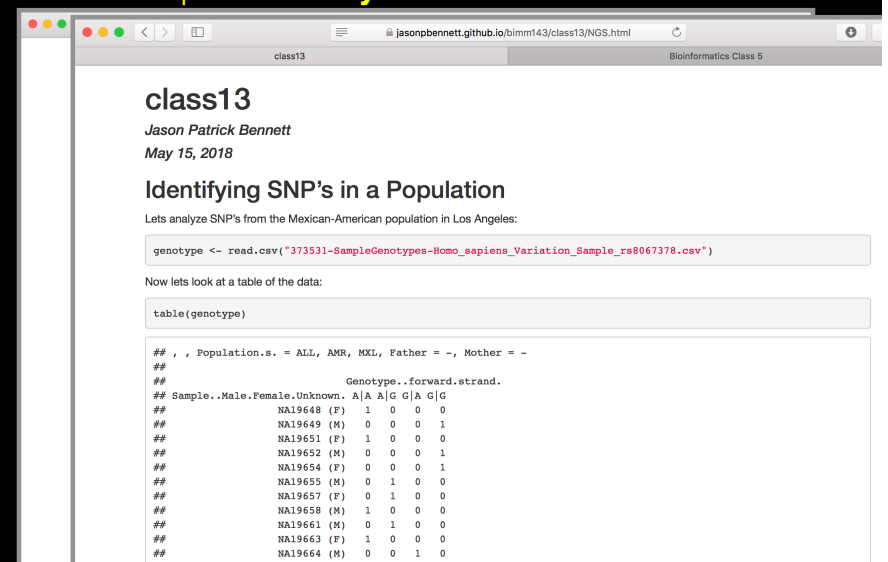
**Index of Material**

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

# Bonus:

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")
```

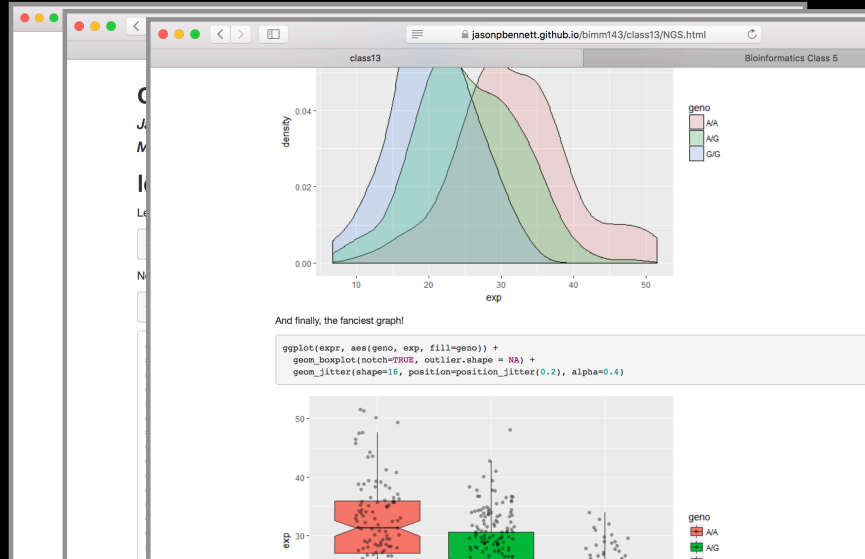
Now lets look at a table of the data:

```
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
```

## 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

Goal Number	Goal Description	Weeks
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

Goal	Description	Page
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

## 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

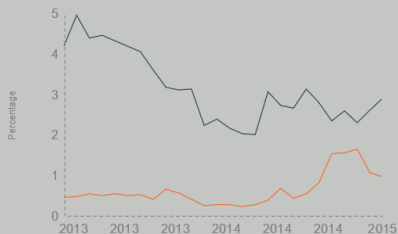
<http://spectrum.ieee.org/computing/software/the-2016-top-programming-languages>



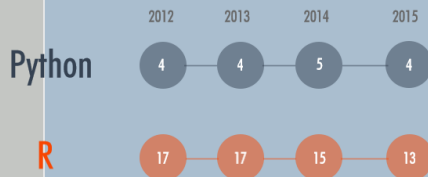
## R and Python: The Numbers

### 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)



### 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 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. The top navigation bar includes 'Learn', 'Groups', 'About', and '1,250 XP'. A notification bell icon in the top right corner is circled in red. The main content area features 'Your Latest Activity' with a card for 'Introduction to Spark in R using...' and a list of recent assignments. A 'DAILY PRACTICE' section is visible at the bottom.

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

The screenshot shows the RStudio IDE interface. The 'Possible Answers' section for a quiz question is circled in red. The 'Submit Answer' button is circled in yellow. The interface includes a console window on the right showing R version information and a file explorer at the bottom.

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

The screenshot shows a web browser displaying a DataCamp course page. On the left, a dark sidebar contains a message: "Exercise Completed" with a green checkmark and a "Stop" button. Below it, text says "Nice job! Move onto the next video to start learning more about the RStudio IDE!" and "PRESS ENTER TO Continue" with a "Continue" button. At the bottom of the sidebar, it says "Become a power user!" with "Submit Answer" and "Ctrl + Shift + Enter" buttons. The main content area shows the RStudio IDE interface with a console window displaying R version information and a file explorer.

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

Homework assignments will be via DataCamp

The screenshot shows a DataCamp exercise page titled "PCA analysis". The page includes instructions and a code editor. The instructions state: "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\_smc2." The code editor shows R code for transforming normalized counts and plotting PCA. The R console shows an error: "Error: object 'vds\_smc2' not found".

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

The screenshot shows a DataCamp course page for "Foundations of Bioinformatics (BGGN-213)". The page includes a "Groups" tab and a "Leaderboard" section. The leaderboard table shows the following data:

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?

## 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**

Side-Note:

## 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>)

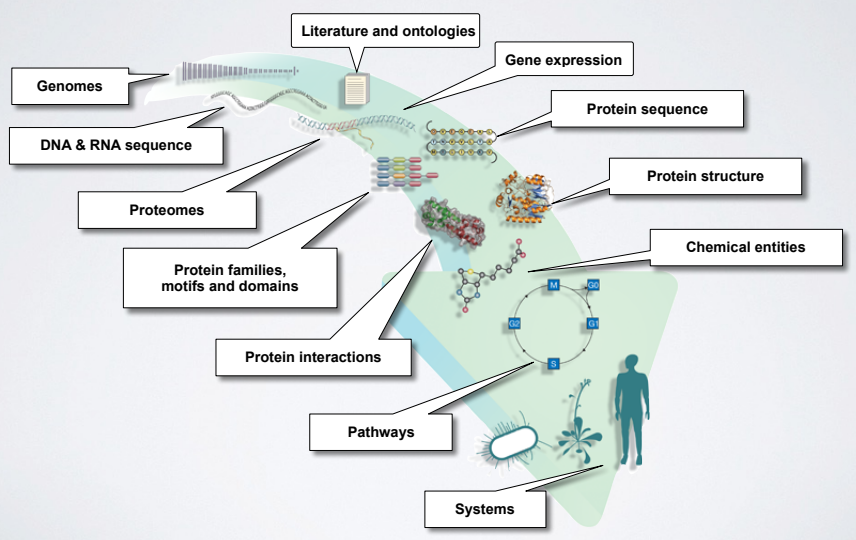
Side-Note:

## There are many useful definitions...

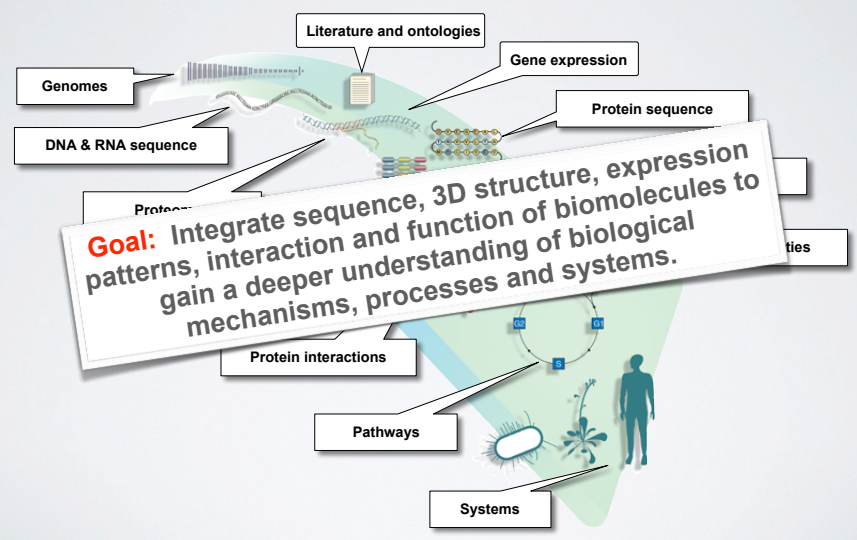
- "Computer based **management** and **analysis** of biological and biomedical data with useful applications in many disciplines, particularly **genomics**, **proteomics**, **metabolomics**, and related fields."  
(BIMM-143)
- "Bioinformatics is conceptualizing biology in terms of **macromolecules**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

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

## Major types of Bioinformatics Data



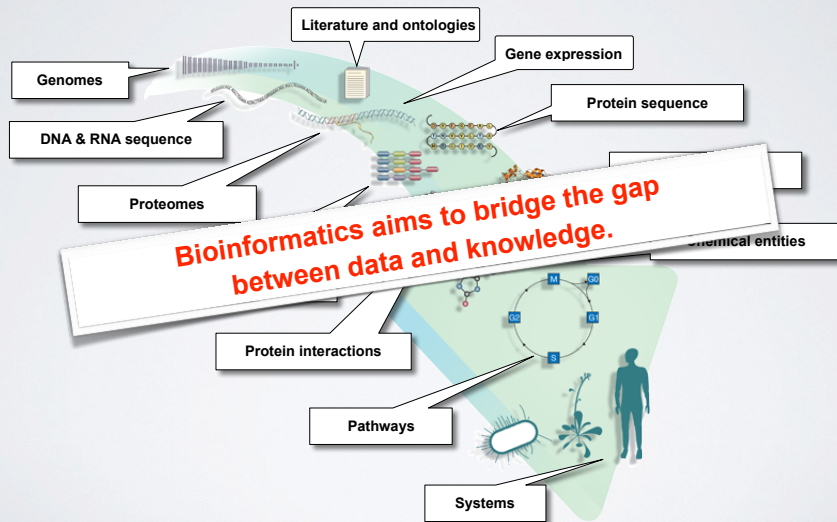
## Major types of Bioinformatics Data



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



## 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 the XSEDE website with a navigation bar and a main content area. The main content area is titled 'Curriculum and Educator Programs' and features a section for 'Campus Visits'. The text describes the purpose of campus visits and provides key points and related links.

**Key Points**

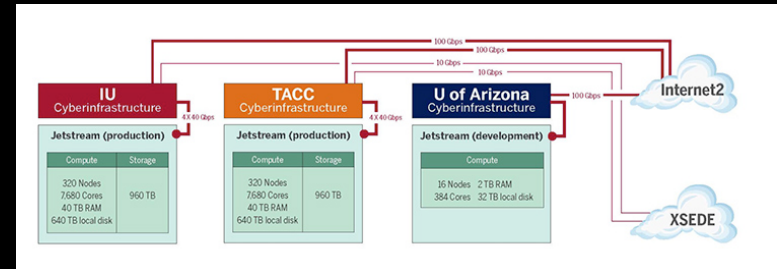
- XSEDE sponsors full-semester online courses
- Collaborations with faculty at participating institutions
- Campus visits offer guidance concerning course content

**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

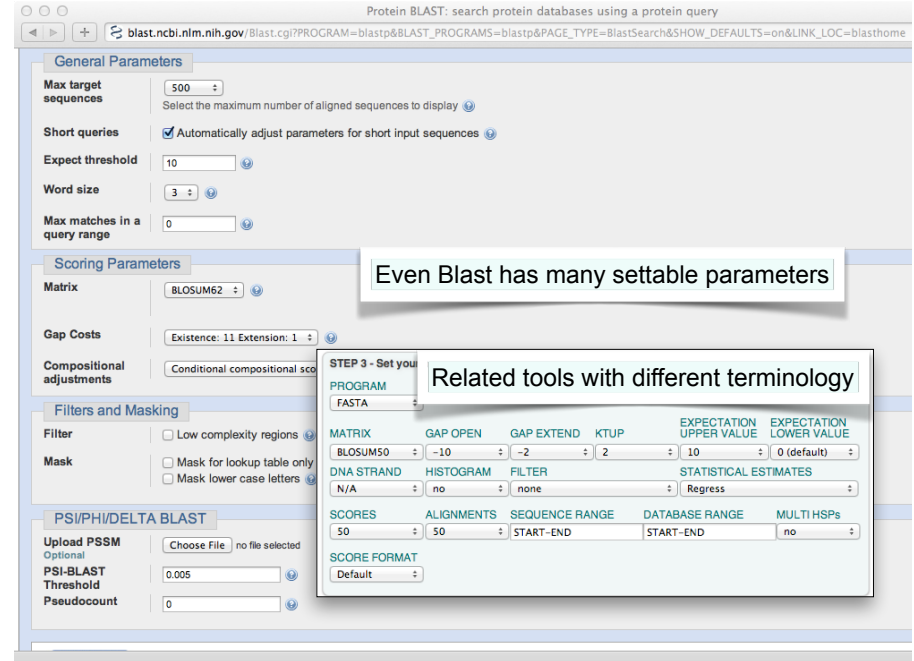
- ▶ Each with many options and settable parameters

Most tools and databases are written by and for nerds

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

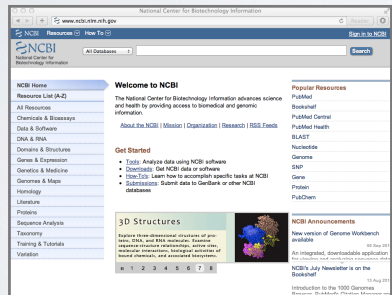
Most are developed independently

- Notable exceptions are found at the:
- **EBI** (European Bioinformatics Institute) and
  - **NCBI** (National Center for Biotechnology Information)

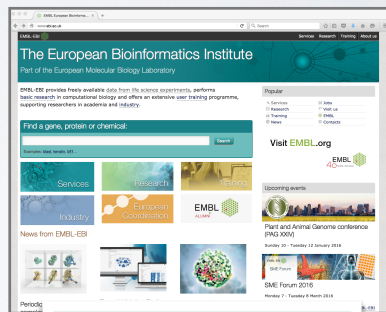


# Key Online Bioinformatics Resources: NCBI & EBI

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



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



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

# National Center for Biotechnology Information (NCBI)

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





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

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

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

## Key Online Bioinformatics Resources: NCBI & EBI

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

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

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

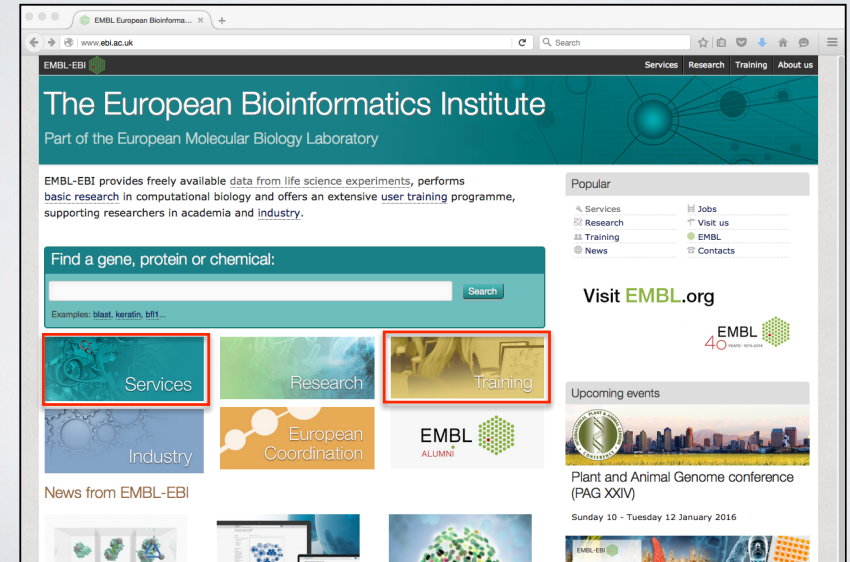


## European Bioinformatics Institute (EBI)

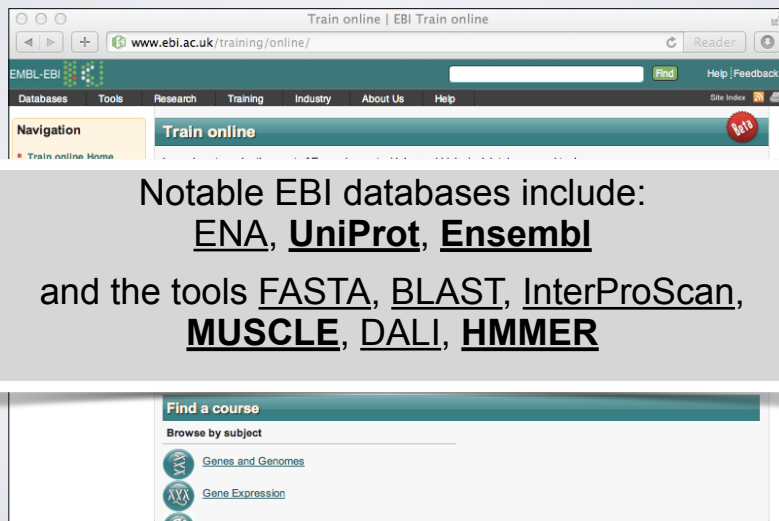
- Created in 1997 as a part of the European Molecular Biology Laboratory (EMBL)
- EBI's mission includes:
  - providing freely available **data** and **bioinformatics services**
  - and providing advanced **bioinformatics training**
- We will cover a number of EBI databases and tools that have advantages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



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



The 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**

## Bioinformatics Databases

AATDB, AceDb, ACUTS, ADB, AFDB, AGIS, AMSdb, ARR, AsDb, BBDB, BCGD, Beanref, Biomag, 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, HICD, HIVdb, HotMolecBase, HOVERGEN, HPDB, HSC-2DPAGE, ICN, ICTVDB, IL2RGbase, IMGT, Kabat, KDNA, KEGG, Klotho, LGIC, MAD, MaizeDb, MDB, Medline, Mendel, MEROPS, MGDB, MGI, MHCPEP5, Micado, MitoDat, MITOMAP, MJDB, MmtDB, Mol-R-Us, MPDB, MRR, MutBase, MycDB, NDB, NRSdb, 0-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 ..... !!!!



BIMM-143: INTRODUCTION TO BIOINFORMATICS (Lecture 1)

Bioinformatics Databases and Key Online Resources  
[https://bioboot.github.io/bimm143\\_W18/lectures/#1](https://bioboot.github.io/bimm143_W18/lectures/#1)  
Dr. Barry Grant  
Jan 2018

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

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

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

**Section 1**

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

>example1

```
ATGGTCATCTGACTCCTGTGGAGAAGTCGCGCTTACTGCCCTGTGGGCAAGGTGACCTGGATGAAG
TTGGTGTGAGGCGCTGGAGGCTGCTGAGGCTACCTGACCCAGAGGTTCTTGAAGTCTTTGG
GGACTCTGCACTCTGATGCACTTATGGGCAAGCTTAAGGTGAGGCTGATGGCAGAAGTGGCTGGT
GCCTTTAGTGTATGGCTGCTACCTGGACAACCTCAAGGGCACCTTGGCACACTGAGTGAGCTGCAC
GTGACAAGCTGCACGTGGATCCTGAGAAGTTCAGGCTCCTGGGCAAGCTGCTGTGTGTGGCCCA
TCACCTTGGCAAGAATTACCCCAAGTGCAGGCTGCTATCAGAAAGTGGTGGTGTGGCTAAT
GCCCTGGCCACAGTACTAAGCTGGCTTTCTTGTGCTCAATTT
```

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:

THANK YOU