BIMM 143

Introduction to Bioinformatics

Barry Grant

UC San Diego

http://thegrantlab.org/bimm143

HELLO
my name is

BARRY

bjgrant@ucsd.edu

HELLO
HIS my name is

ALEX

ajweitze@ucsd.edu

HELLO
HER my name is

YUSI

cyusi@ucsd.edu

05:00

Introduce Yourself!

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

Today's Menu

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

http://thegrantlab.org/bimm143/

bioboot.github.io/bimm143_W20/

Home Gmail Gcal GitHub BIMM143 BGGN213 Atmosphere BIMM194 Blink News + +

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/

bioboot.github.io/bimm143_W20/

Home Gmail Gcal GitHub BIMM143 BGGN213 Atmosphere BIMM194 Blink News + +

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?

bioboot.github.io/bimm143_W18/goals/

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

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!

bioboot.github.io/bimm143_W18/goals/

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

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

Specific Learning Goals

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

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

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

Course Structure

Derived from specific learning goals

bioboot.github.io/bimm143_S18/lectures/

Home Gmail Gcal Bitbucket GitHub News BIMM143_S18 BGGN213_S18 Disqus BIMM-194 Atmosphere Blink GDocs Galaxy +

UC San Diego

BIMM 143

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

- Overview
- Lectures
- Computer Setup
- Learning Goals
- Assignments & Grading
- Ethics Code

Lectures

All Lectures are Tu/Th 9:00-12:00 pm in Warren Lecture Hall 2015 (WLH 2015) (Map). Clicking on the class topics below will take you to corresponding lecture notes, homework assignments, pre-class video screen-casts and required reading material.

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

Course Structure

Derived from specific learning goals

bioboot.github.io/bimm143_S18/lectures/

Home Gmail Gcal Bitbucket GitHub News BIMM143_S18 BGGN213_S18 Disqus BIMM-194 Atmosphere Blink GDocs Galaxy +

UC San Diego

BIMM 143

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

- Overview
- Lectures
- Computer Setup
- Learning Goals
- Assignments & Grading
- Ethics Code

Lectures

All Lectures are Tu/Th 9:00-12:00 pm in Warren Lecture Hall 2015 (WLH 2015) (Map). Clicking on the class topics below will take you to corresponding lecture notes, homework assignments, pre-class video screen-casts and required reading material.

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

Class Details

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

bioboot.github.io/bimm143_W18/lectures/#1

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

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, Leaning 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**

bioboot.github.io/bimm143_W18/lectures/#1

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

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:

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

Homework

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

bioboot.github.io/bimm143_W18/lectures/#1

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

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:

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 email address and UCSD PID number so you can receive credit for your responses.

* Required

Email address *

Your email

UCSD PID number (exam number)

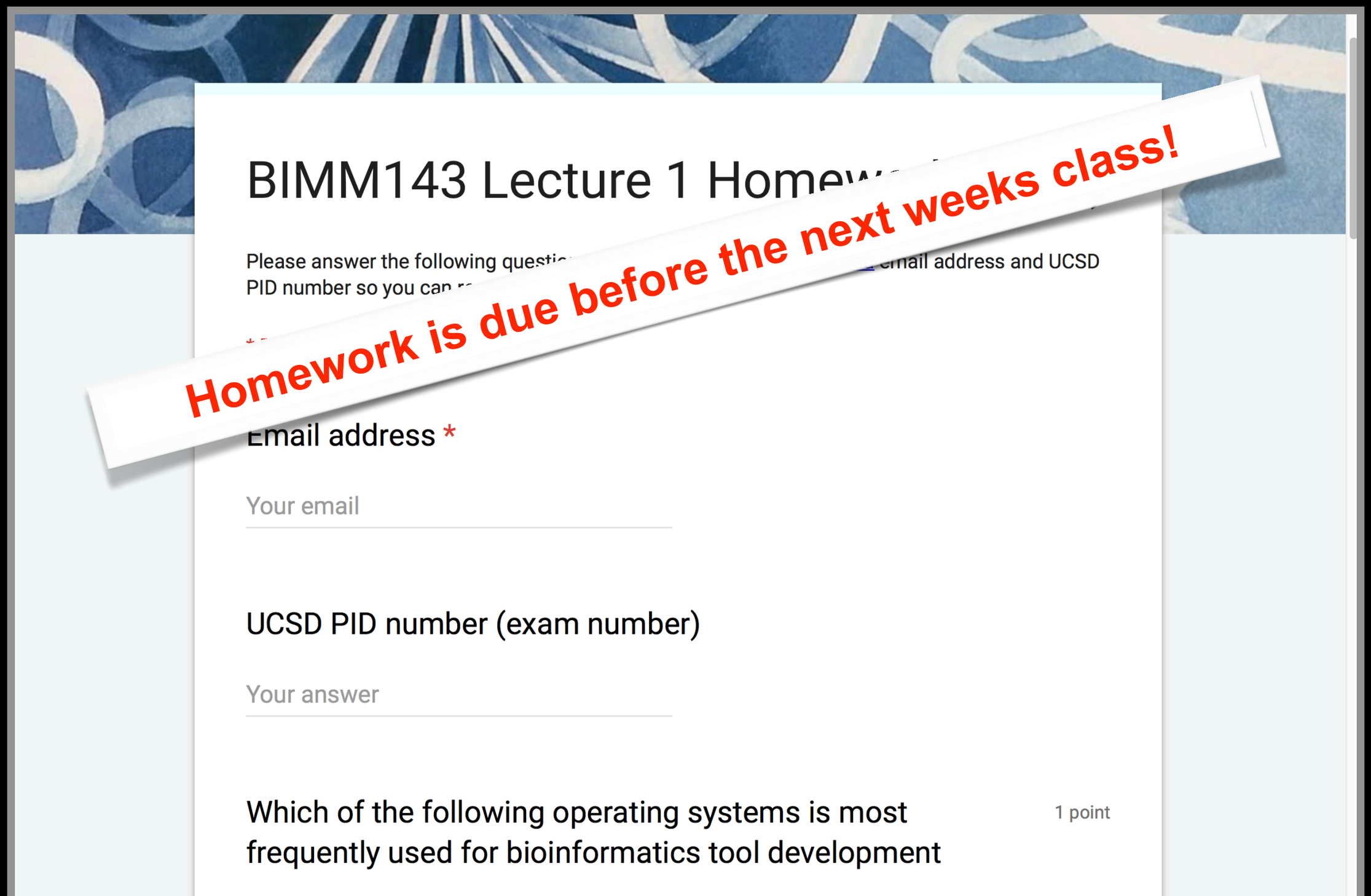
Your answer

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**



Projects

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

bioboot.github.io/bimm143_W19/lectures/#9

Home Gmail Gcal GitHub BIMM143 BGGN213 Atmosphere BIMM194 Blink News

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

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:

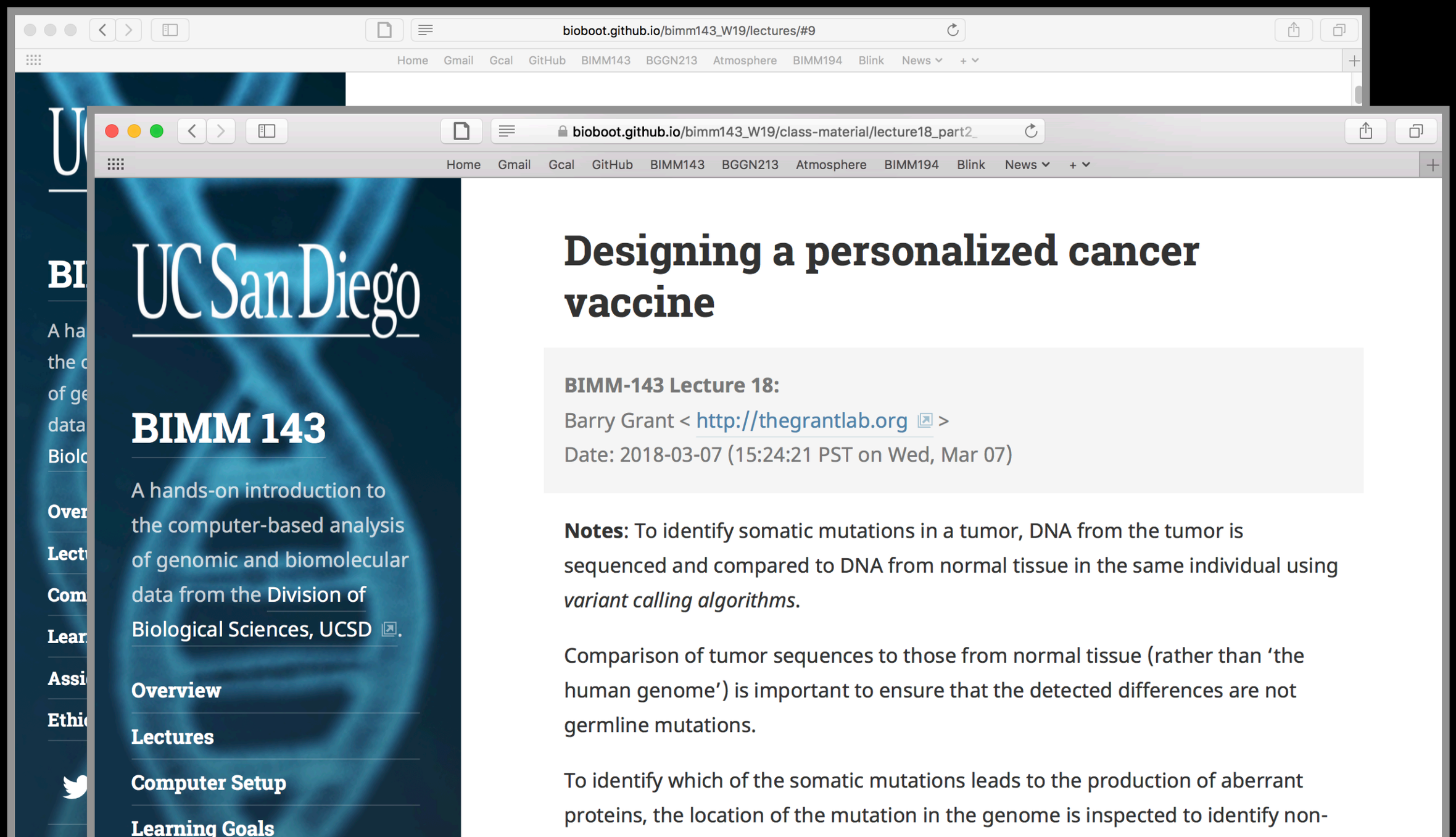
- Be able to import data and prepare for unsupervised learning analysis.
- Be able to apply and test combinations of PCA, k-means and hierarchical clustering to high dimensional datasets and critically review results.

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.

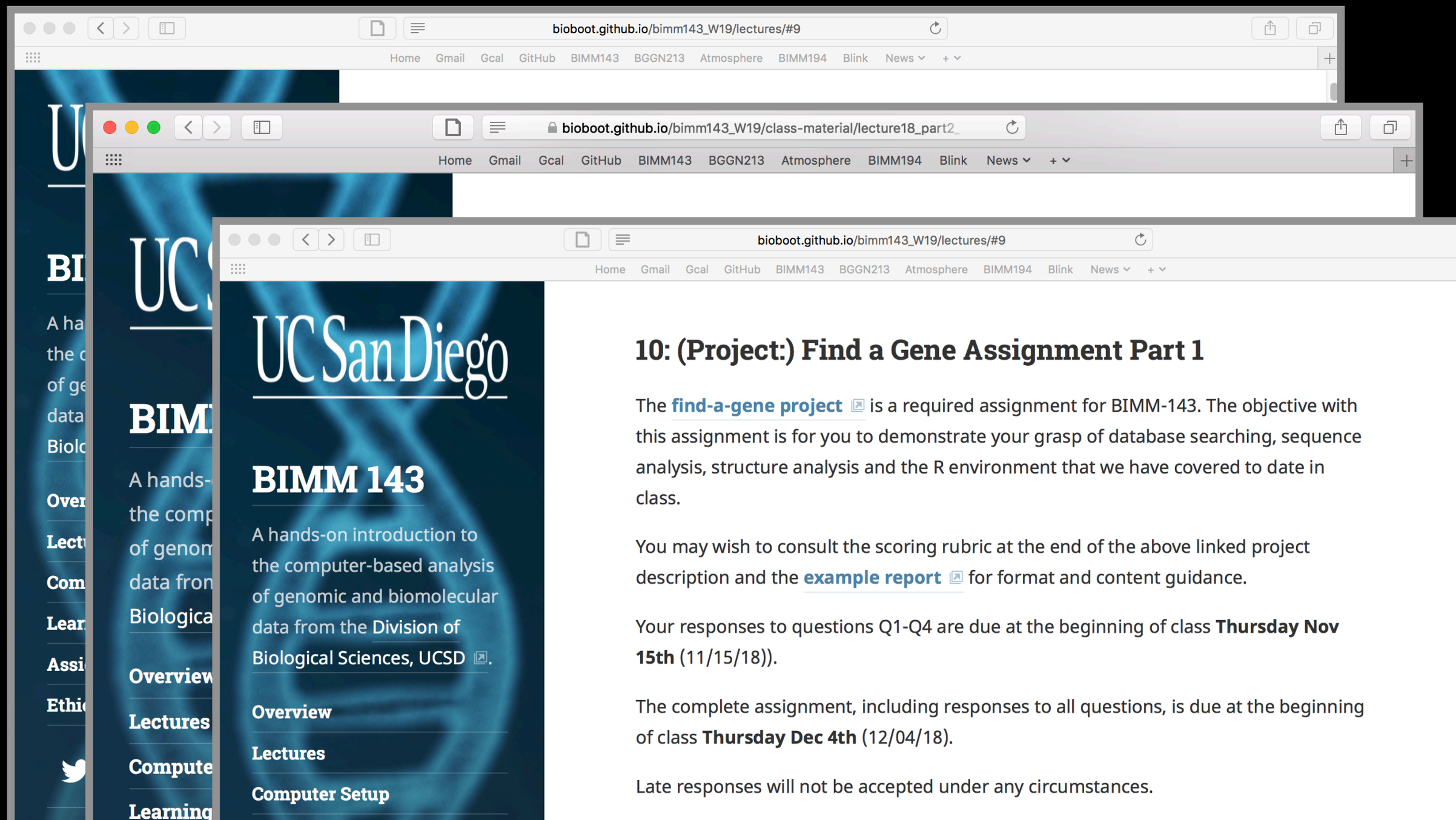
Projects

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



Projects (20% of course grade)

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



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)

bioboot.github.io/bimm143_W19/lectures/#18

Home Gmail Gcal GitHub BIMM143 BGGN213 Atmosphere BIMM194 Blink News + +

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

20: Final Exam

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

Please remember to:

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

Bonus:

Online portfolio of **your** bioinformatics work!

jasonpbennett.github.io/bimm143/

Introduction to Bioinformatics Class S18

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

View the Project on GitHub
JasonPBennett/bimm143

This project is maintained by
JasonPBennett

Bioinformatics Class BIMM-143

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

Index of Material

Introductory Material: Working With R

Class 5 - Basic Data Exploration and Visualization in R

Class 6 - Creating R Functions

Class 7 - R Packages, working with CRAN, and working with Bioconductor

Using R and Other Tools for Bioinformatics Analysis

Class 8 - An Introduction to Machine Learning (Hierarchical Clustering)

Class 9 - Analyzing High Dimensional Datasets and Unsupervised Learning

Class 11 - Structural Bioinformatics: Analyzing Protein Structure and Function

Class 12 - Drug Discovery: Techniques and Analysis

Class 13 - Genome Informatics and High Throughput Sequencing (NGS, RNA-Seq, and FastQC)

Class 14 - Transcriptomics and RNA-Seq Analysis

Class 15 - Genome Annotation and Using Functional Databases (KEGG and GO - Gene Ontology)

Class 16 - Transposons: A Sample Workflow

Bonus:

Online portfolio of **your** bioinformatics work!

jasonpbennett.github.io/bimm143/class13/NGS.html

class13 | Bioinformatics Class 5

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)
```

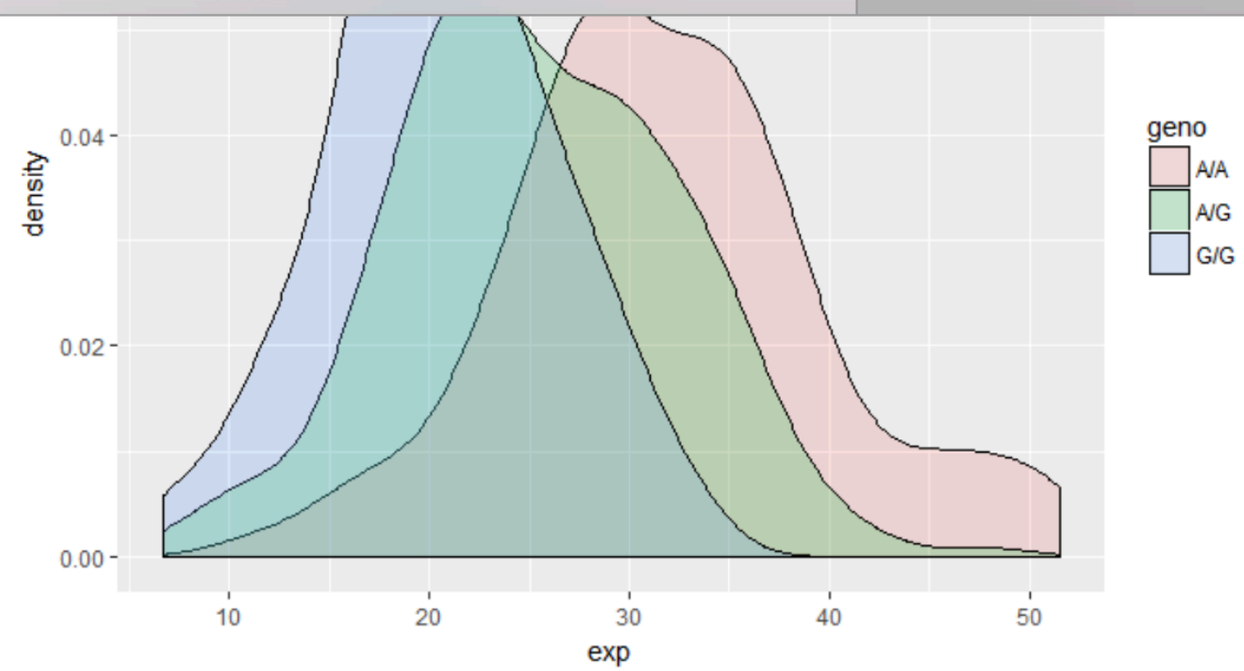
```
## , , Population.s. = ALL, AMR, MXL, Father = -, Mother = -
##
##                                Genotype..forward.strand.
## Sample..Male.Female.Unknown. A|A A|G G|A G|G
##                   NA19648 (F)   1   0   0   0
##                   NA19649 (M)   0   0   0   1
##                   NA19651 (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
```

Bonus:

Online portfolio of **your** bioinformatics work!

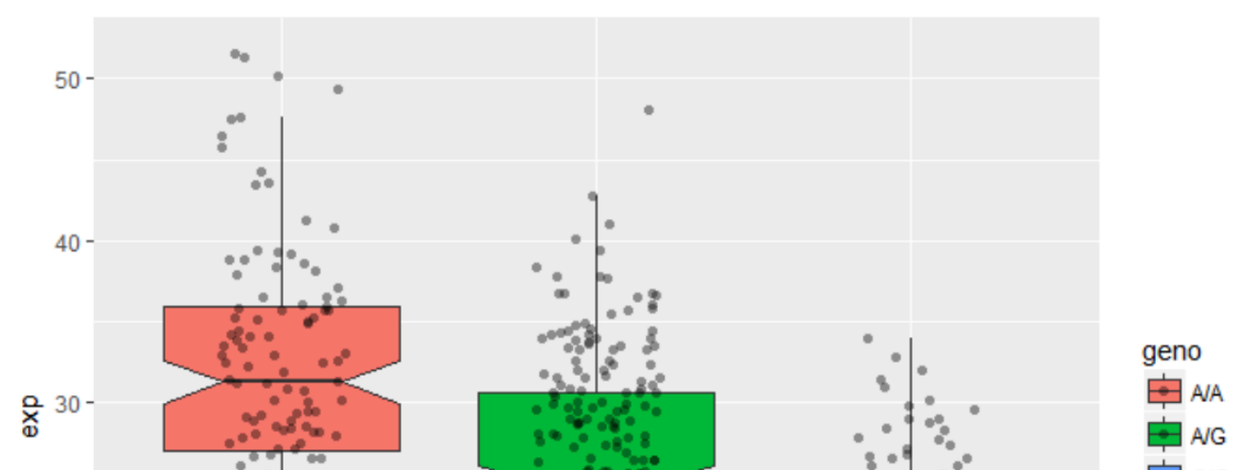
jasonpbennett.github.io/bimm143/class13/NGS.html

class13 | Bioinformatics Class 5



And finally, the fanciest graph!

```
ggplot(expr, aes(geno, exp, fill=geno)) +
  geom_boxplot(notch=TRUE, outlier.shape = NA) +
  geom_jitter(shape=16, position=position_jitter(0.2), alpha=0.4)
```



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 programing experience or high level math skills.

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

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 programing experience or high level math skills.

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

BIMM-143 Learning Goals….

Data science R based learning goals

bioboot.github.io/bimm143_W18/goals/

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

UC San Diego

BIMM 143

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

- Overview
- Lectures
- Computer Setup
- Learning Goals
- Assignments & Grading
- Ethics Code

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

BIMM-143 Learning Goals….

Delve deeper into “real-world” bioinformatics

bioboot.github.io/bimm143_W18/goals/

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

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

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

These support a major learning objective

At the end of this course students will:

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

Why use R?

Productivity
Flexibility
Genomic data analysis

IEEE 2016 Top Programming Languages

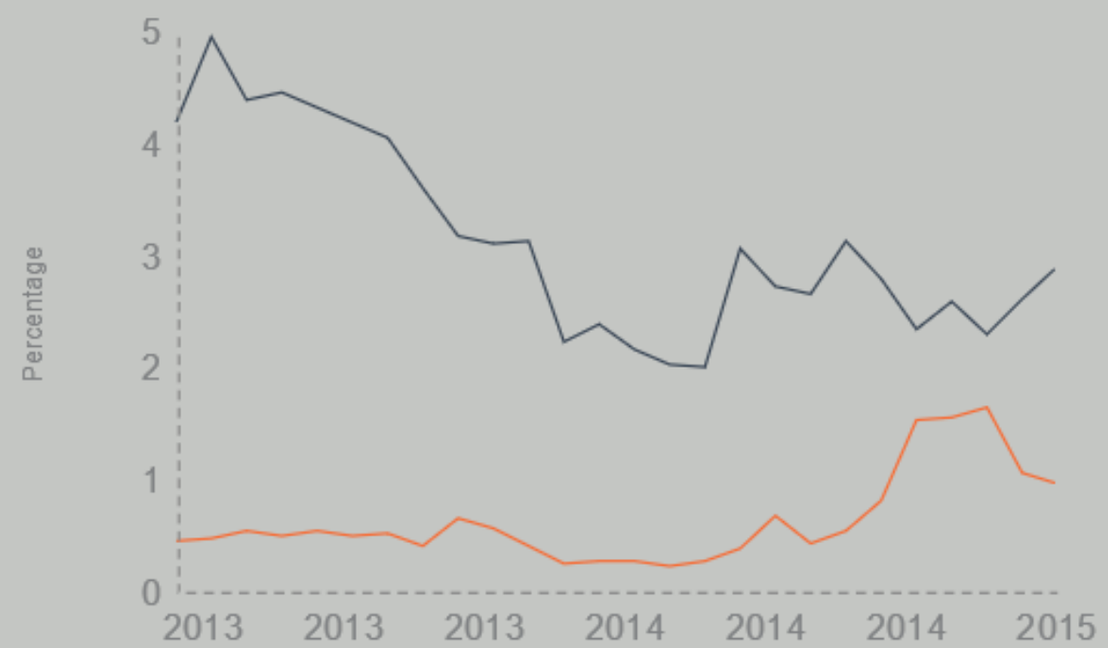
Language Rank	Types	Spectrum Ranking
1. C		100.0
2. Java		98.1
3. Python		98.0
4. C++		95.9
5. R		87.9
6. C#		86.7
7. PHP		82.8
8. JavaScript		82.2
9. Ruby		74.5
10. Go		71.9

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

R and Python: The Numbers

Popularity Rankings

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



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

	2012	2013	2014	2015
Python	4	4	5	4
R	17	17	15	13

Jobs And Salary?

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

$ 115,531

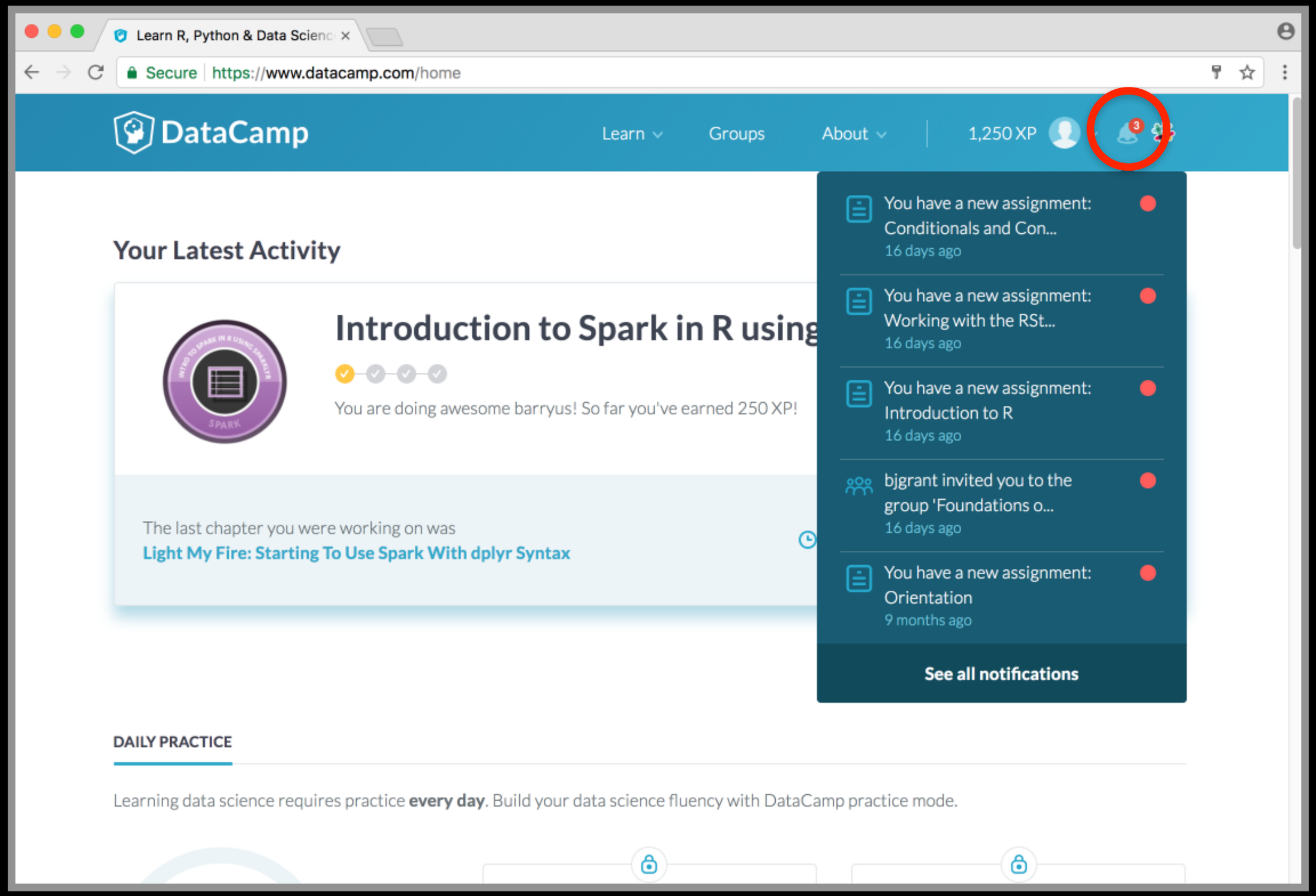
$94,139

http://www.kdnuggets.com/2015/05/r-vs-python-data-science.html?utm_medium=email&utm_source=flipboard

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/ >



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

What is an IDE anyway? | R

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

DataCamp

Course Outline

What is an IDE anyway?

50xp

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

What does IDE stand for?

Possible Answers

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

Take Hint (-15xp)

Submit Answer

File Edit Code View Plots Session Build Debug Profile Tools Help barryus Sessions

Go to file/function Addins Project: (None) R 3.

Console ~/

```
R version 3.3.1 (2016-06-21) -- "Bug in Your Hair"
Copyright (C) 2016 The R Foundation for Statistical Comp
uting
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANT
Y.
You are welcome to redistribute it under certain conditi
ons.
Type 'license()' or 'licence()' for distribution detail
s.

  Natural language support but running in an English loc
ale

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

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

>
```

Environment History

Import Dataset List

Global Environment

Environment is empty

Files Plots Packages Help Viewer

New Folder Upload Delete Rename

Home

Name Size

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

What is an IDE anyway? | R

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

DataCamp

Course Outline

What is an IDE anyway?

Exercise Completed

50xp

Nice job! Move onto the next video to start learning more about the RStudio IDE!

PRESS ENTER TO Continue

Become a power user!

Submit Answer Ctrl + Shift + Enter

See all keyboard shortcuts

File Edit Code View Plots Session Build Debug Profile Tools Help barryus Sessions

Go to file/function Addins Project: (None) R 3.

Console ~/

```
R version 3.3.1 (2016-06-21) -- "Bug in Your Hair"
Copyright (C) 2016 The R Foundation for Statistical Comp
uting
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANT
Y.
You are welcome to redistribute it under certain conditi
ons.
Type 'license()' or 'licence()' for distribution detail
s.

  Natural language support but running in an English loc
ale

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

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

>
```

Environment History

Import Dataset List

Global Environment

Environment is empty

Files Plots Packages Help Viewer

New Folder Upload Delete Rename

Home

Name Size

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

Homework assignments will be via DataCamp

DataCamp

Course Outline

Exercise

PCA analysis

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

To assess the similarity of the `smoc2` samples using PCA, we need to transform the normalized counts then perform the PCA analysis. Assume all libraries have been loaded, the DESeq2 object created, and the size factors have been stored in the DESeq2 object, `dds_smoc2`.

Instructions 1/2 50 XP

- Run the code to transform the normalized counts.
- Perform PCA by plotting PC1 vs PC2 using the DESeq2 `plotPCA()` function on the DESeq2 transformed counts object, `vsd_smoc2` and specify the `intgroup` argument as the factor to color the plot.

Take Hint (-15 XP)

script.R RDocumentation

```
# Transform the normalized counts
vsd_smoc2 <- vst(dds_smoc2, blind = TRUE)

# Plot the PCA of PC1 and PC2
___(___, intgroup=___)
```

Run Code Submit Answer

R Console Slides

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

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

DataCamp

Secure | https://www.datacamp.com/groups/foundations-of-bioinformatics-bggn-213/details

DataCamp Learn Groups About 1,550 XP

Back to My Dashboard

Foundations of Bioinformatics (BGGN-213)

Leaderboard My Assignments

30 Days | 90 Days | Last Year

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

Today’s Menu

Course Logistics	Website, screencasts, survey, ethics, assessment and grading.
Learning Objectives	What you need to learn to succeed in this course.
Course Structure	Major lecture topics and specific leaning goals.
Introduction to Bioinformatis	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...

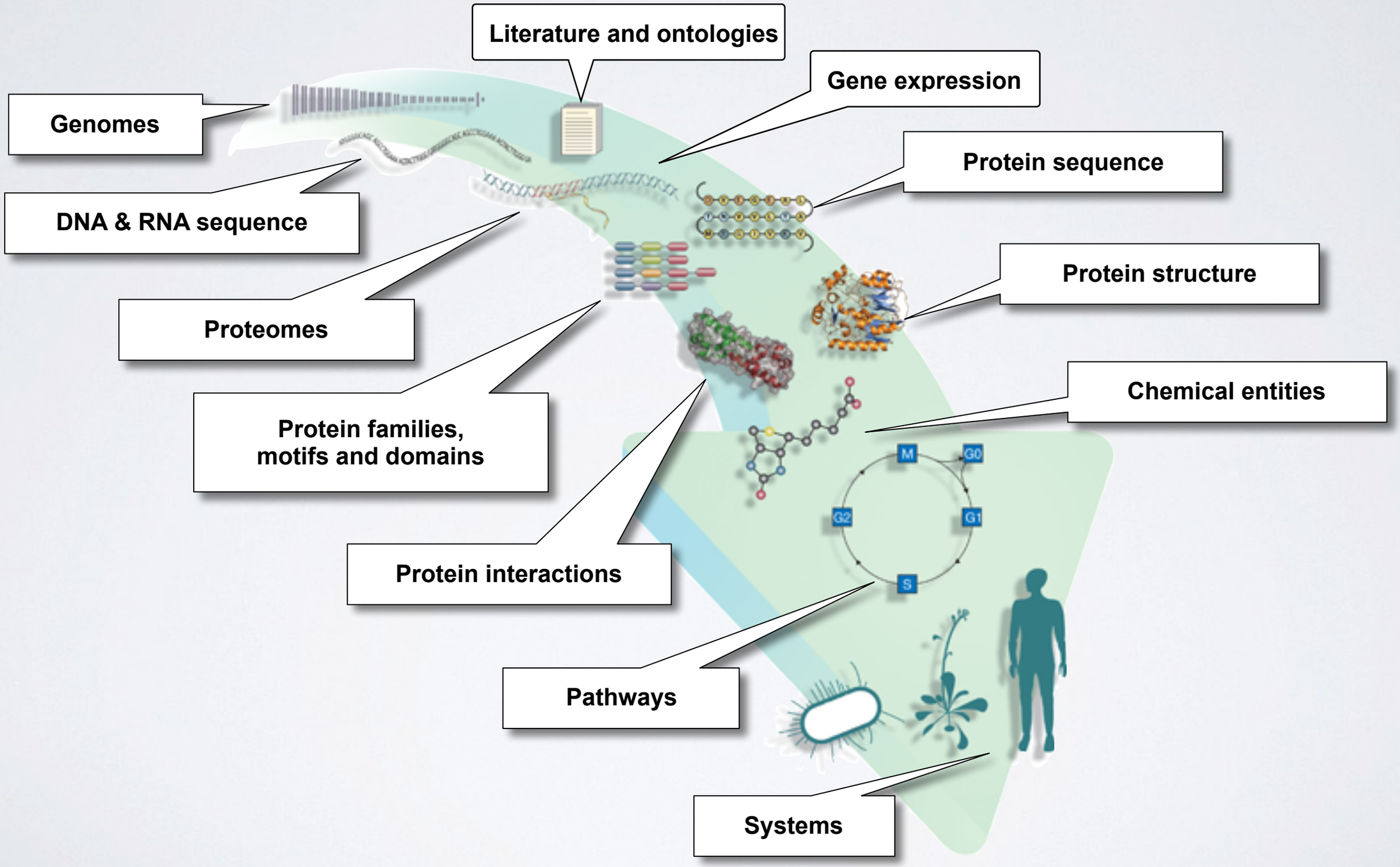
- "Computer based management and analysis of biological and biomedical data with useful applications in many disciplines, particularly genomics, proteomics, metabolomics, and re[illegible]d fields."
  (BIMM-143)

- "Bioinformatics is conceptualizing biolo[illegible] of macromolecules and then applyin[illegible]tics" techniques (derived from disciplines such as appli[illegible], computer science, and statistics) to understand [illegible]anize the information associated with these molecules [illegible]-scale."
  (Luscomb[illegible]01)

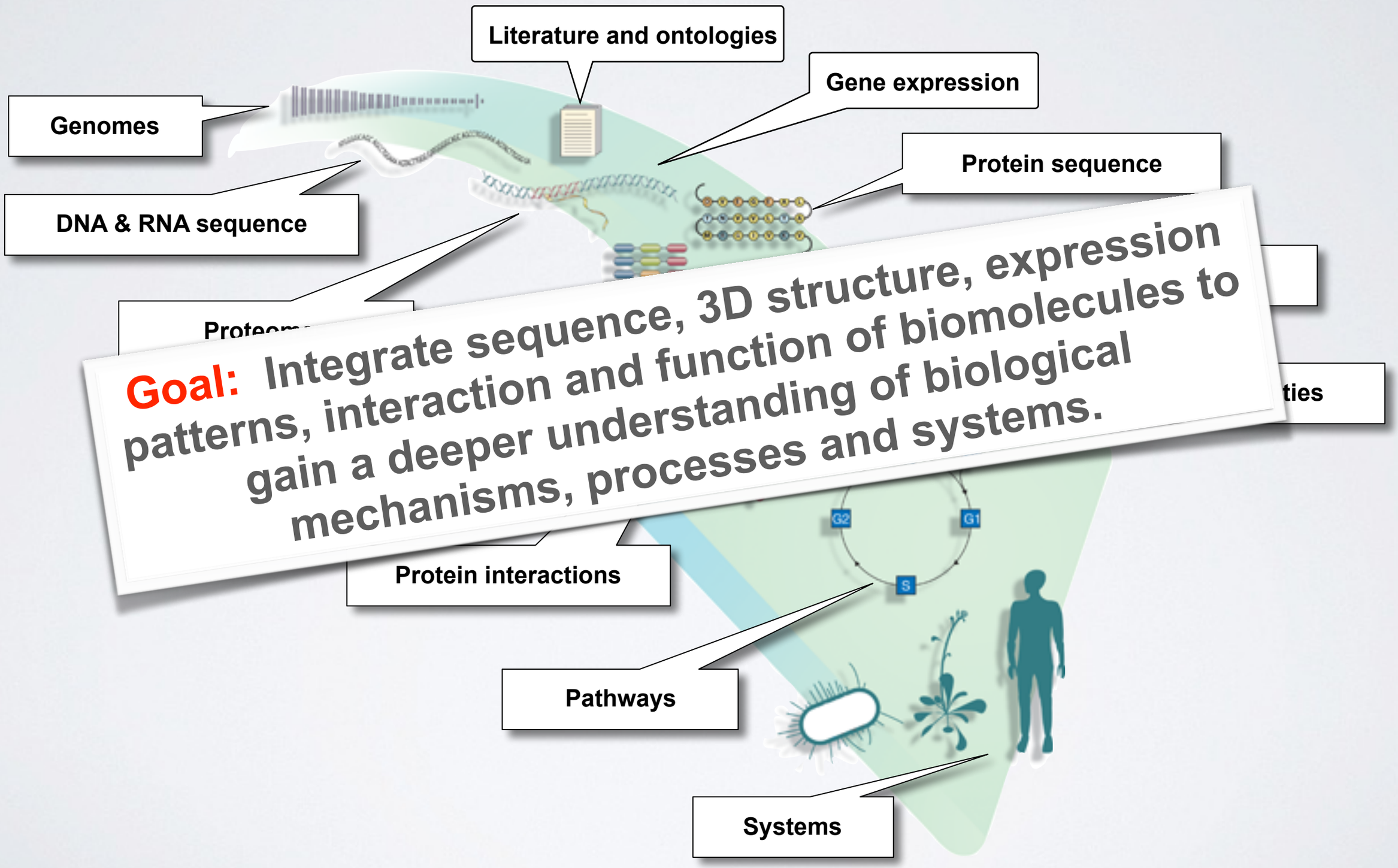
- "B[illegible]atics is research, development, or application of co[illegible]putational approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize and analyze such data ...<cut>..."
  (National Institutes of Health: http://tinyurl.com/l3gxr6b)

Key Point: Bioinformatics is Computer Aided Biology

Major types of Bioinformatics Data

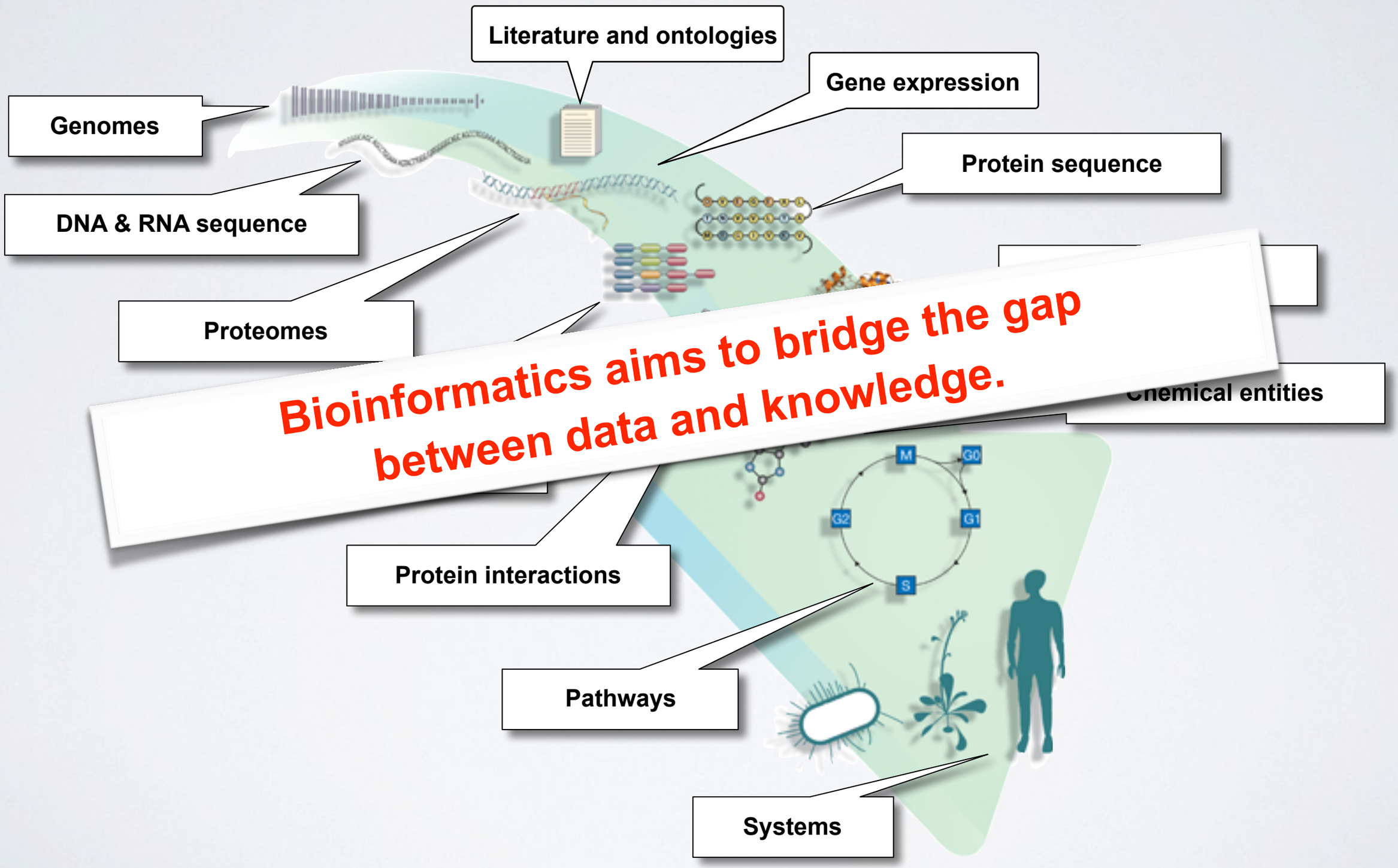


Major types of Bioinformatics Data



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



Bioinformatics aims to bridge the gap between data and knowledge.

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 <u>large collections</u> of these entities.

DNA	→	RNA	→	Protein
Genome	→	**Transcriptome**	→	**Proteome**

x 1,000

x 100,000

How do we *actually* do Bioinformatics?

Pre-packaged tools and databases

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

Advanced tool application & development

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

How do we *actually* do Bioinformatics?

Pre-packaged tools and databases

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

Advanced tool application & development

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

NSF Extreme Science and Engineering Discovery Environment (XSEDE)

www.xsede.org/community-engagement/educator-pr

Home Gmail Gcal Bitbucket GitHub News Disqus

XSEDE

About For Users Ecosystem Community Engagement News XUP

Curriculum and Educator Programs

XSEDE pursues innovation and collaboration in computational science education.

Campus Visits

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

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

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

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

Key Points

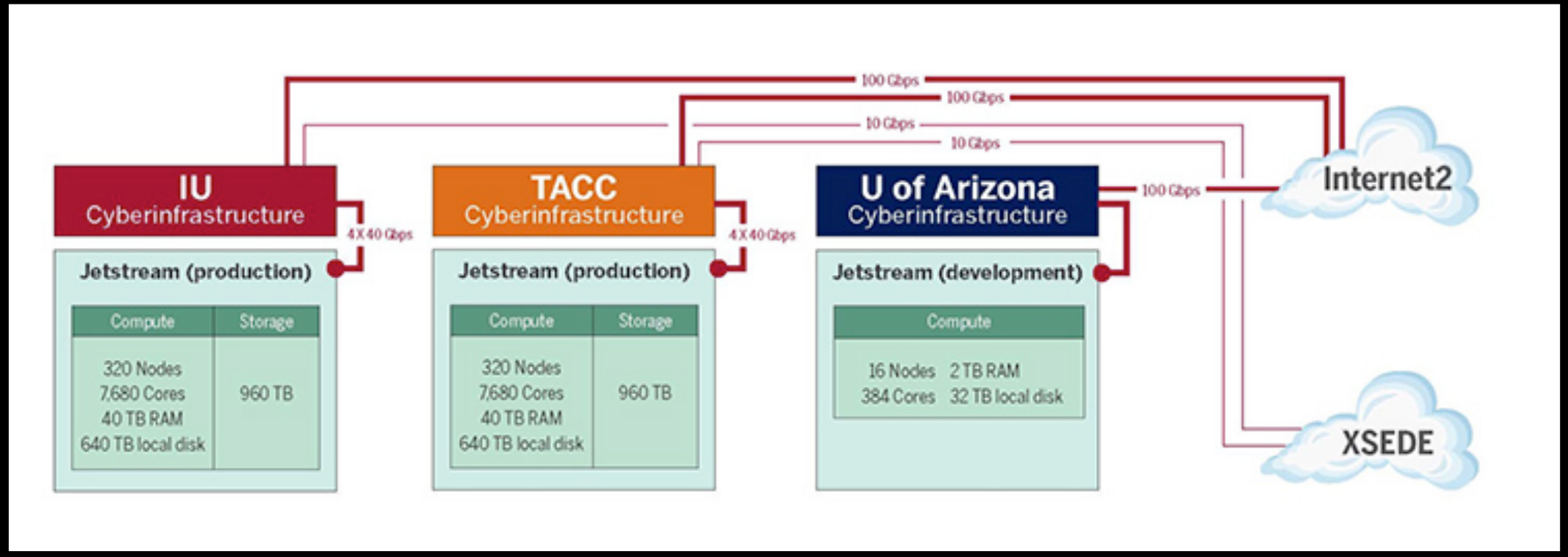
- XSEDE sponsors full-semester online courses
- Collaborations with faculty at participating institutions
- Campus visits offer guidance concerning course content

Related Links

Diversity and Inclusion
Student Engagement
Campus Champions
XSEDE Scholars Program

What is *Jetstream*?

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



Skepticism & Bioinformatics

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

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

  What does this model actually contribute?
- Avoid the miss-use of ‘black boxes’

Skepticism & Bioinformatics

Gunnar von Heijne in *"Sequence Analysis in Molecular Biology"* states:

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

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

Common problems with Bioinformatics

Confusing multitude of tools available

- Each with many options and settable parameters

Most tools and databases are written by and for nerds

- Same is true of documentation - if any exists!

Most are developed independently

Notable exceptions are found at the:

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

Protein BLAST: search protein databases using a protein query

blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&BLAST_PROGRAMS=blastp&PAGE_TYPE=BlastSearch&SHOW_DEFAULTS=on&LINK_LOC=blasthome

General Parameters

Max target sequences	500 — Select the maximum number of aligned sequences to display
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 sco

Filters and Masking

Filter	☐ Low complexity regions
Mask	☐ Mask for lookup table only ☐ Mask lower case letters

PSI/PHI/DELTA BLAST

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

Even Blast has many settable parameters

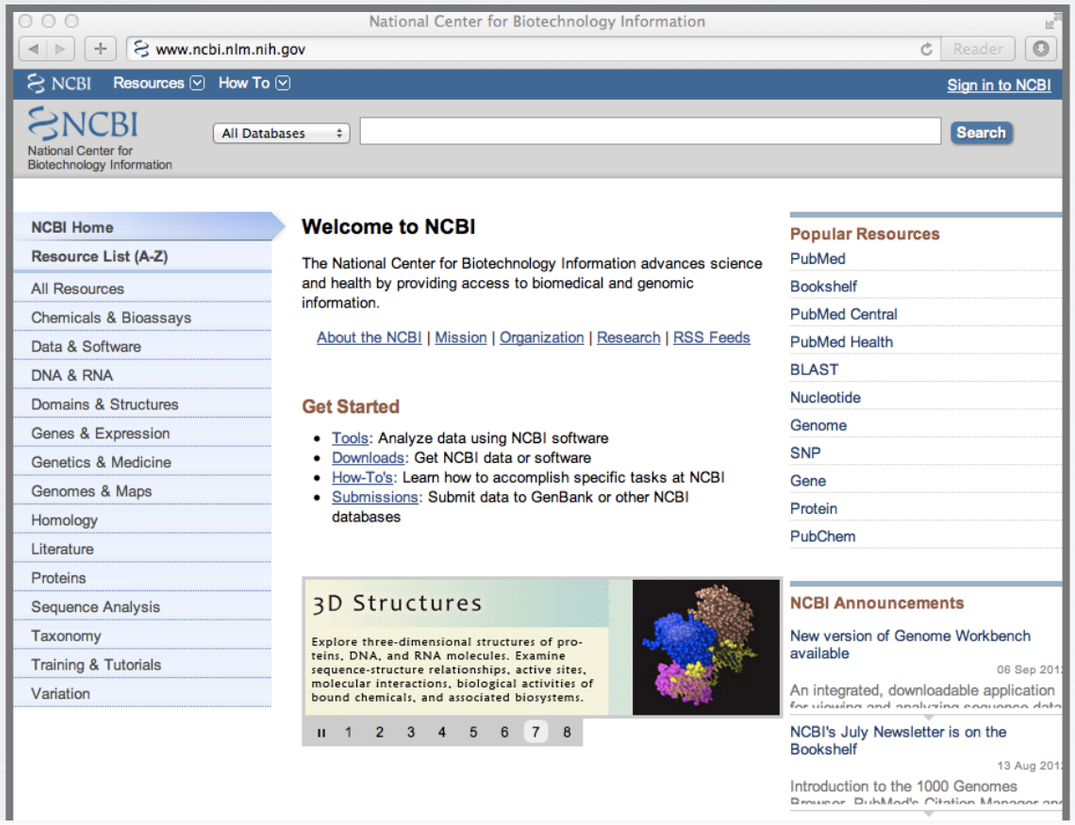
STEP 3 - Set you

Related tools with different terminology

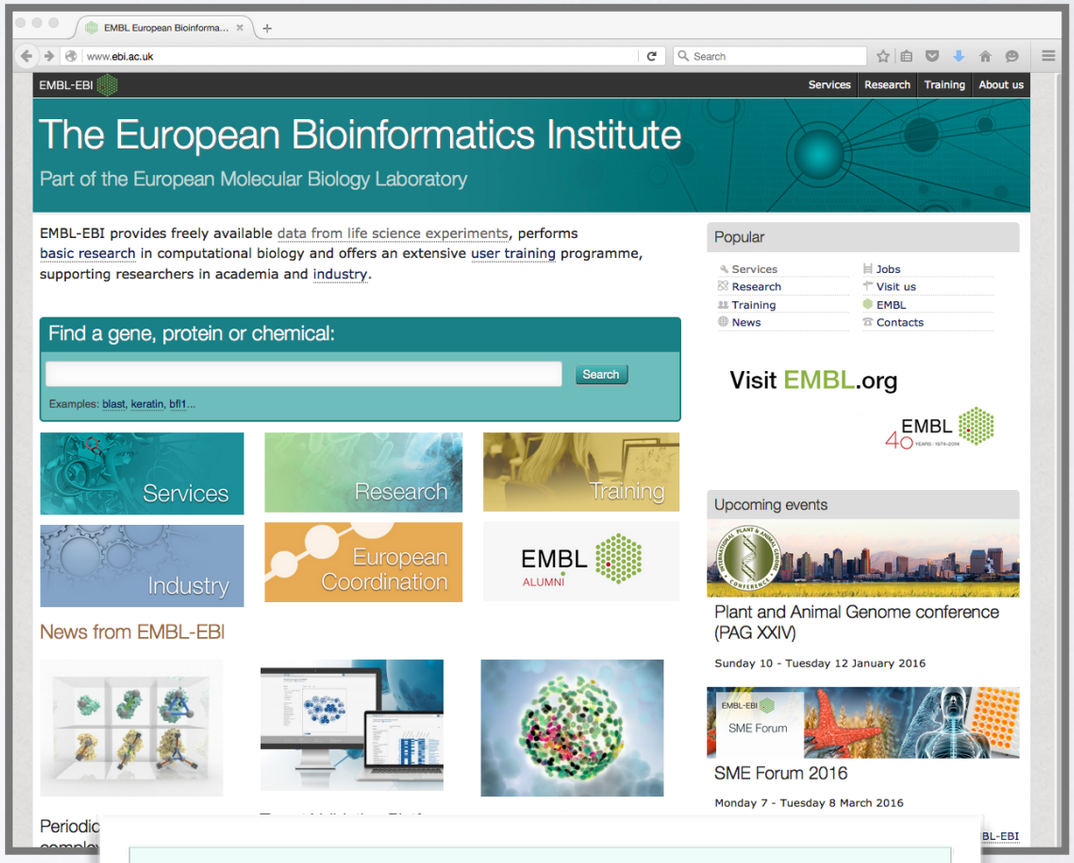
PROGRAM					
FASTA					
MATRIX	**GAP OPEN**	**GAP EXTEND**	**KTUP**	**EXPECTATION UPPER VALUE**	**EXPECTATION LOWER VALUE**
BLOSUM50	-10	-2	2	10	0 (default)
DNA STRAND	**HISTOGRAM**	**FILTER**		**STATISTICAL ESTIMATES**	
N/A	no	none		Regress	
SCORES	**ALIGNMENTS**	**SEQUENCE RANGE**	**DATABASE RANGE**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

Key Online Bioinformatics Resources: NCBI & EBI

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



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

National Center for Biotechnology Information

www.ncbi.nlm.nih.gov

NCBI Resources How To Sign in to NCBI

NCBI
National Center for Biotechnology Information

All Databases Search

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

Welcome to NCBI

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

About the NCBI | Mission | Organization | Research | RSS Feeds

Get Started

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

Popular Resources

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

3D Structures

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

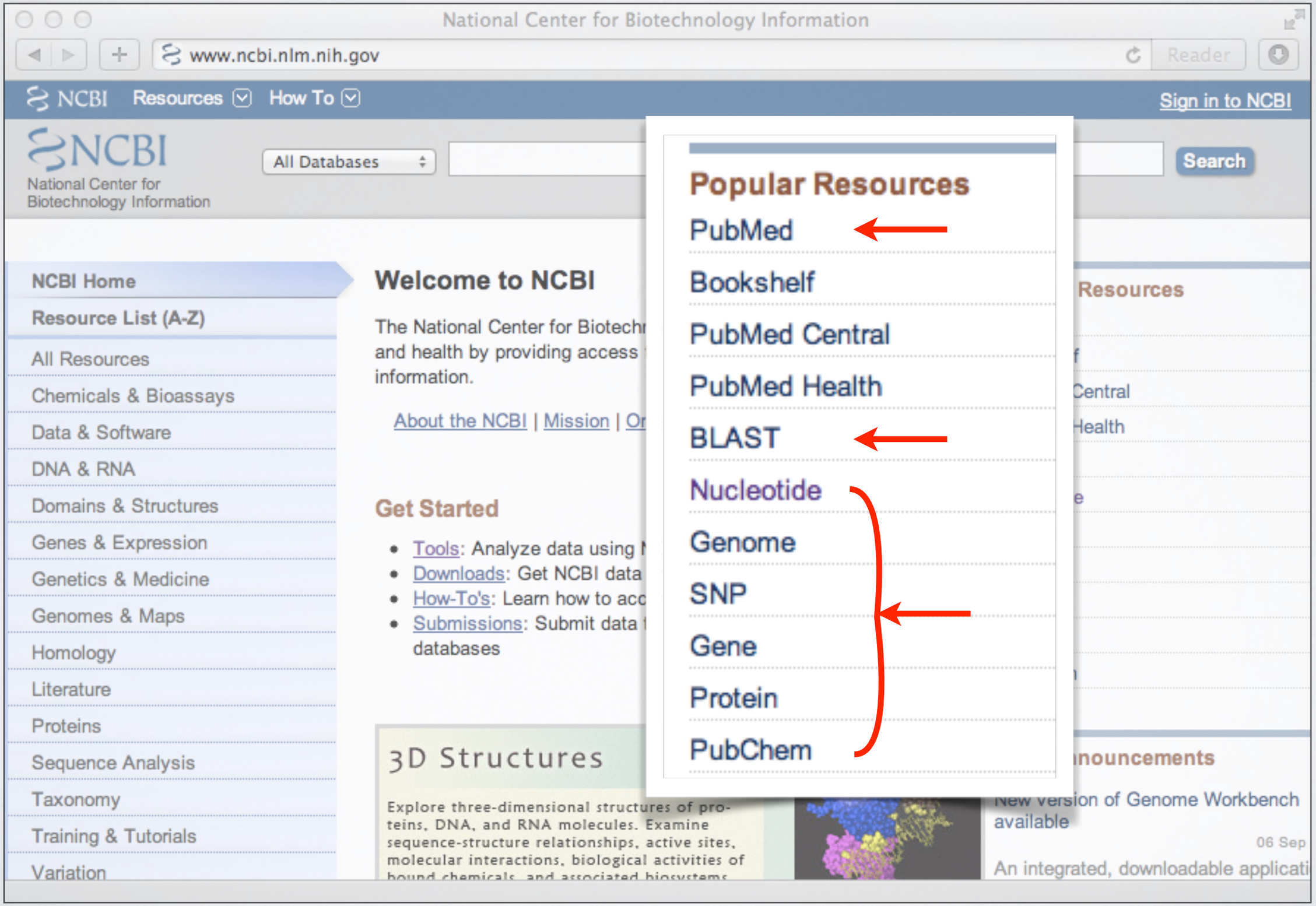
NCBI Announcements

New version of Genome Workbench available

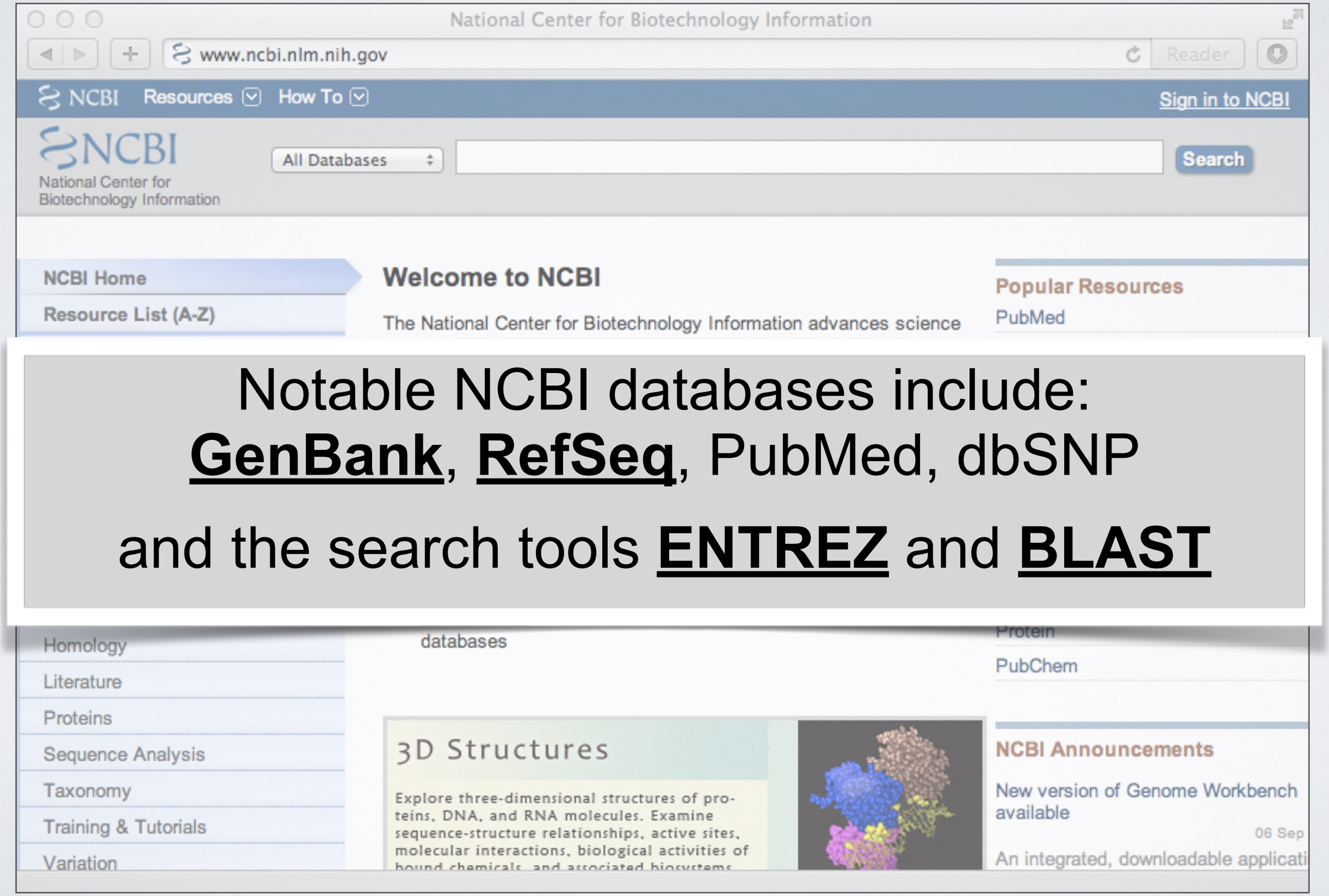
06 Sep

An integrated, downloadable applicati

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



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

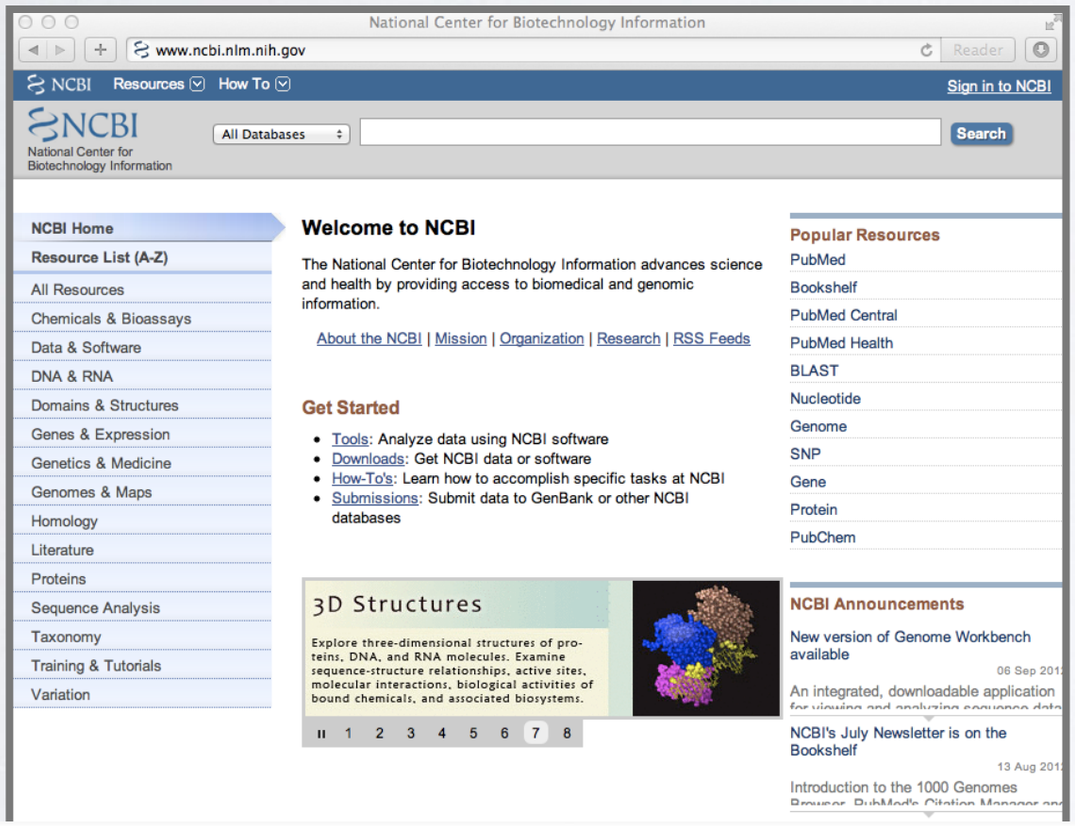


Notable NCBI databases include:
GenBank, **RefSeq**, PubMed, dbSNP

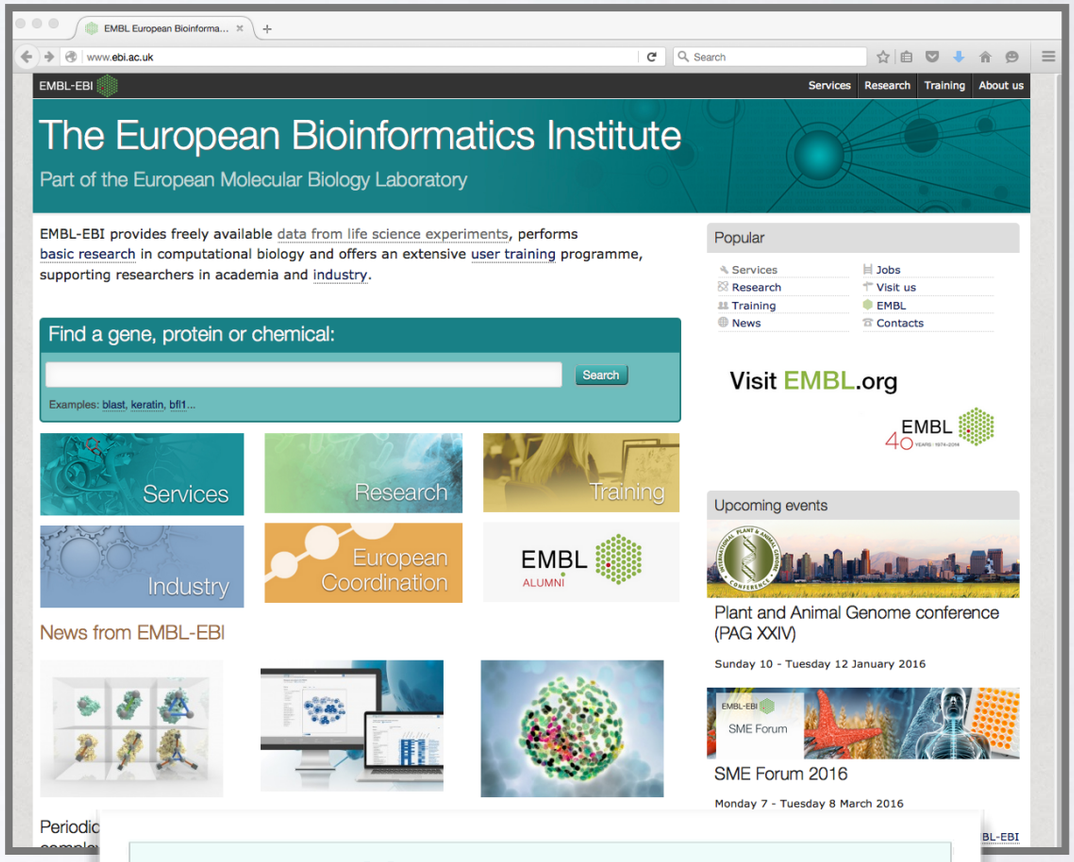
and the search tools **ENTREZ** and **BLAST**

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

EMBL European Bioinforma...

www.ebi.ac.uk

EMBL-EBI

Services | Research | Training | About us

The European Bioinformatics Institute

Part of the European Molecular Biology Laboratory

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

Find a gene, protein or chemical:

Search

Examples: blast, keratin, bfl1...

Services | Research | Training | Industry | European Coordination | EMBL ALUMNI

News from EMBL-EBI

Popular

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

Visit EMBL.org

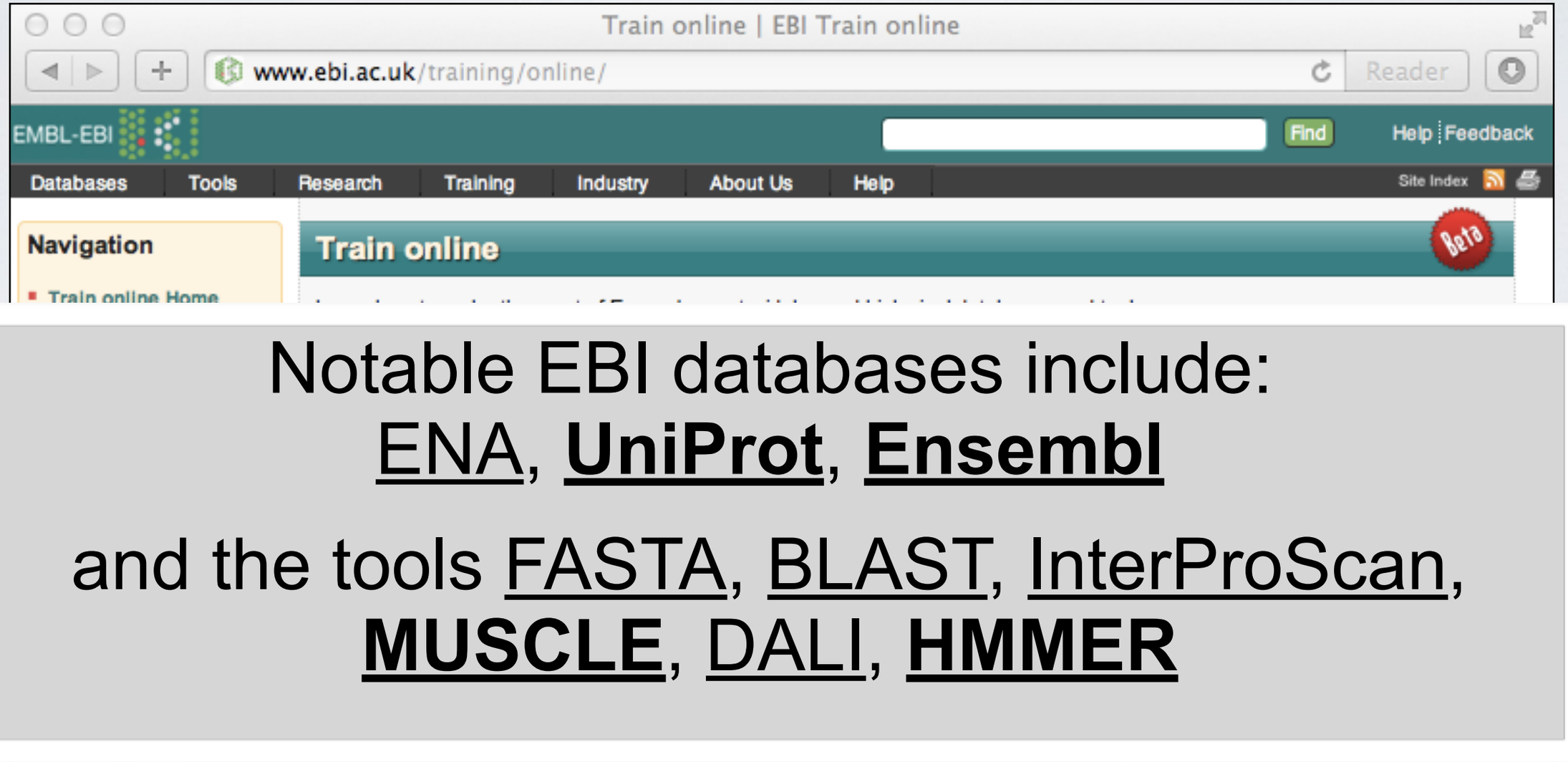
EMBL 40 YEARS | 1974-2014

Upcoming events

Plant and Animal Genome conference (PAG XXIV)

Sunday 10 - Tuesday 12 January 2016

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



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, BioImage, BioMagResBank, BIOMDB, BLOCKS, BovGBASE, BOVMAP, BSORF, BTKbase, CANSITE, CarbBank, CARBHYD, CATH, CAZY, CCDC, CD4OLbase, CGAP, ChickGBASE, Colibri, COPE, CottonDB, CSNDB, CUTG, CyanoBase, dbCFC, dbEST, dbSTS, DDBJ, DGP, DictyDb, Picty_cDB, DIP, DOGS, DOMO, DPD, DPInteract, ECDC, ECGC, EC02DBASE, EcoCyc, EcoGene, EMBL, EMD db, ENZYME, EPD, EpoDB, ESTHER, FlyBase, FlyView, GCRDB, GDB, GENATLAS, Genbank, GeneCards, Genlilesne, GenLink, GENOTK, GenProtEC, GIFTS, GPCRDB, GRAP, GRBase, gRNAsdb, GRR, GSDB, HAEMB, HAMSTERS, HEART-2DPAGE, HEXAdb, HGMD, HIDB, HIDC, HIVdb, HotMolecBase, HOVERGEN, HPDB, HSC-2DPAGE, ICN, ICTVDB, IL2RGbase, IMGT, Kabat, KDNA, KEGG, Klotho, LGIC, MAD, MaizeDb, MDB, Medline, Mendel, MEROPS, MGDB, MGI, MHCPEP5 Micado, MitoDat, MITOMAP, MJDB, MmtDB, Mol-R-Us, MPDB, MRR, MutBase, MycDB, NDB, NRSub, 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 ..... !!!!

Bioinformatics Databases

AATDB, AceDb, ACUTS, ADB, AFDB, AGIS, AMSdb, ARR, AsDb,BBDB, BCGD, Beanref, BioImage, BioMagResBank, BIOMDB, BLOCKS, BovGBASE, BOVMAP, CATKbase, CANSITE, CarbBank, CARBHYD, CATH, CAZY, CCAP, ChickGBASE, Colibri, COPE, CottonDB, CGAP, dbEST, dbSTS, DDBJ, DGP, DictyDb, Bi[illegible] CDC, ECGC, EC02DBASE, E[illegible] THER, FlyBase, Fl[illegible] Link, G[illegible] DB, HAEMB, H[illegible] HIVdb, HotMolecBase, H[illegible] IEZRGbase, IMGT, Kabat, KDNA, K[illegible] DB, Medline, Mendel, MEROPS, MGDB, MGI, MHC[illegible] PROTOMAP, MJDB, MmtDB, Mol-R-Us, MPDB, MRR, MutBase, Myc[illegible] db, 0-lycBase, OMIA, OMIM, OPD, ORDB, OWL, PAHdb, PatBase, PD[illegible] DD, Pfam, PhosphoBase, PigBASE, PIR, PKR, PMD, PPDB, PRESAGE, PRINTS, ProDom, Prolysis, PROSITE, PROTOMAP, RatMAP, RDP, REBASE, RGP, SBASE, SCOP, SeqAnaiRef, SGD, SGP, SheepMap, Soybase, SPAD, SRNA db, SRPDB, STACK, StyGene,Sub2D, SubtiList, SWISS-2DPAGE, SWISS-3DIMAGE, SWISS- MODEL Repository, SWISS-PROT, TelDB, TGN, tmRDB, TOPS, TRANSFAC, TRR, UniGene, URNADB, V BASE, VDRR, VectorDB, WDCM, WIT, WormPep, etc ..... !!!!

There are lots of Bioinformatics Databases

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

< Major_Databases.pdf >

Side-note: Databases come in all shapes and sizes





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

Today’s Menu

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

Hands-on section

Your Turn!

http://thegrantlab.org/bimm143/

bioboot.github.io/bimm143_W18/lectures/#1

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

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

Goals:

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

Material:

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

BIMM-143: INTRODUCTION TO BIOINFORMATICS (Lecture 1)

Bioinformatics Databases and Key Online Resources
https://bioboot.github.io/bimm143_W18/lectures/#1
Dr. Barry Grant
Jan 2018

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

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

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

Section 1
The following transcript was found to be abundant in a human patient's blood sample.

```
>example1
ATGGTGCATCTGACTCCTGTGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAG
TTGGTGGTGAGGCCCTGGGCAGGCTGCTGGTGGTCTACCCTTGGACCCAGAGGTTCTTTGAGTCCTTTGG
GGATCTGTCCACTCCTGATGCAGTTATGGGCAACCCTAAGGTGAAGGCTCATGGCAAGAAAGTGCTCGGT
GCCTTTAGTGATGGCCTGGCTCACCTGGACAACCTCAAGGGCACCTTTGCCACACTGAGTGAGCTGCACT
GTGACAAGCTGCACGTGGATCCTGAGAACTTCAGGCTCCTGGGCAACGTGCTGGTCTGTGTGCTGGCCCA
TCACTTTGGCAAAGAATTCACCCCACCAGTGCAGGCTGCCTATCAGAAAGTGGTGGCTGGTGTGGCTAAT
GCCCTGGCCCACAAGTATCACTAAGCTCGCTTTCTTGCTGTCCAATTT
```

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**: