USC Department Of Translational Genomics
Keck Medicine of USC

TRGN 527: Applied Data Science and Bioinformatics

UNIT I. Introduction and Basic Data Science

Week 1 – Lecture 1

Enrique I. Velazquez Villarreal, M.D., Ph.D., M.P.H., M.S. | **Assistant Professor**

Dept. of Translational Genomics

USC | Keck School of Medicine | Norris Comprehensive Cancer Center

Leader of the USC Bioinformatics Core – *USC CaRE2 Health Equity Center*

David W. Craig, Ph.D. | **Professor and Vice Chair**

Dept. of Translational Genomics

USC | Keck School of Medicine | Norris Comprehensive Cancer Center

Co-Director, Institute of Translational Genomics

Topics

- Introduction, course outline.

Units:	4
Term:	Spring
Date/Time:	1:00 - 2:50 PM Tue, Thu
Location:	NRT 2508
Instructors:	Enrique I. Velazquez-Villarreal, M.D., Ph.D. David W. Craig, Ph.D.
Office:	NRT 2517N
Office Hours:	Thursday 4-5 PM
Contact Info:	Enrique I. Velazquez-Villarreal, M.D., Ph.D. Assistant Professor, Translational Genomics. 1450 Biggy St. NRT 2517N Email: eivelazq@usc.edu Office: (323) 442-0411

TA Office Hours

TA:	Xianghui Li
Office:	NRT 2509L
Office hours:	Tuesdays 3 – 5 pm
	Thursdays 3 – 5 pm
	Fridays 10am - Noon
Contact info:	xianghui@usc.edu

University of Pittsburgh

HARVARD MEDICAL SCHOOL

BRIGHAM AND WOMEN'S HOSPITAL

Carnegie Mellon University

USC Department Of Translational Genomics
Keck Medicine of USC

Professional Background

GENOMICS

Postdoctoral Research Associate, Bioinformatics/Genomics	Aug 2015 -Sep 2017
Doctor of Philosophy (PhD), Human Genetics, Statistical and Computational Genetics	Aug 2011- Aug 2015
Master of Science (MS), Epidemiology, Population Genetics	May 2010 - Aug 2011
Master of Public Health (MPH), Multidisciplinary	Dec 2008 - Aug 2011
Certificate in Global Health (CGH), Bioinformatics	Apr 2010 – Aug 2011

MEDICINE

Medical Doctor (MD), Surgery, Obstetrics and General Practitioner (UANL), Mexico	Aug 1999 - Nov 2006
Medical Practitioner, National Institute of Neurology and Neurosurgery, Mexico	Aug 2006 - Aug 2007
Medical Fellowship, Brigham and Women's Hospital, Department of Surgery, HARVARD	Apr - Jul 2006
Medical Fellowship, Vall D'Hebron Hospital, Department of Neurosurgery, U. Barcelona	Feb- Mar 2004
Medical Practitioner, University Hospital, School of Medicine (UANL), Mexico	Sep 2003 - Aug 2006

EXPERIENCE - **Statistical & Computational Geneticist:**

Pittsburgh Supercomputer Center (PSC), Carnegie Mellon University (CMU)	Mar 2015 – Aug 2015
Institute for Personalized Medicine, University of Pittsburgh	Sep 2013 - Aug 2015
Center for Simulation and Modeling, University of Pittsburgh	Sep 2011 - Aug 2015
Center for Computational Genetics, University of Pittsburgh	Sep 2009 - Aug 2015

Assistant Professor, University of Southern California (USC)	Oct 2017
Leader Bioinformatics Core, Care2USC (USC)	Oct 2018
Adjunct Faculty, San Diego State University (SDSU)	Apr 2017
Adjunct Faculty, National University (NU) San Diego	June 2017 – Feb 2018

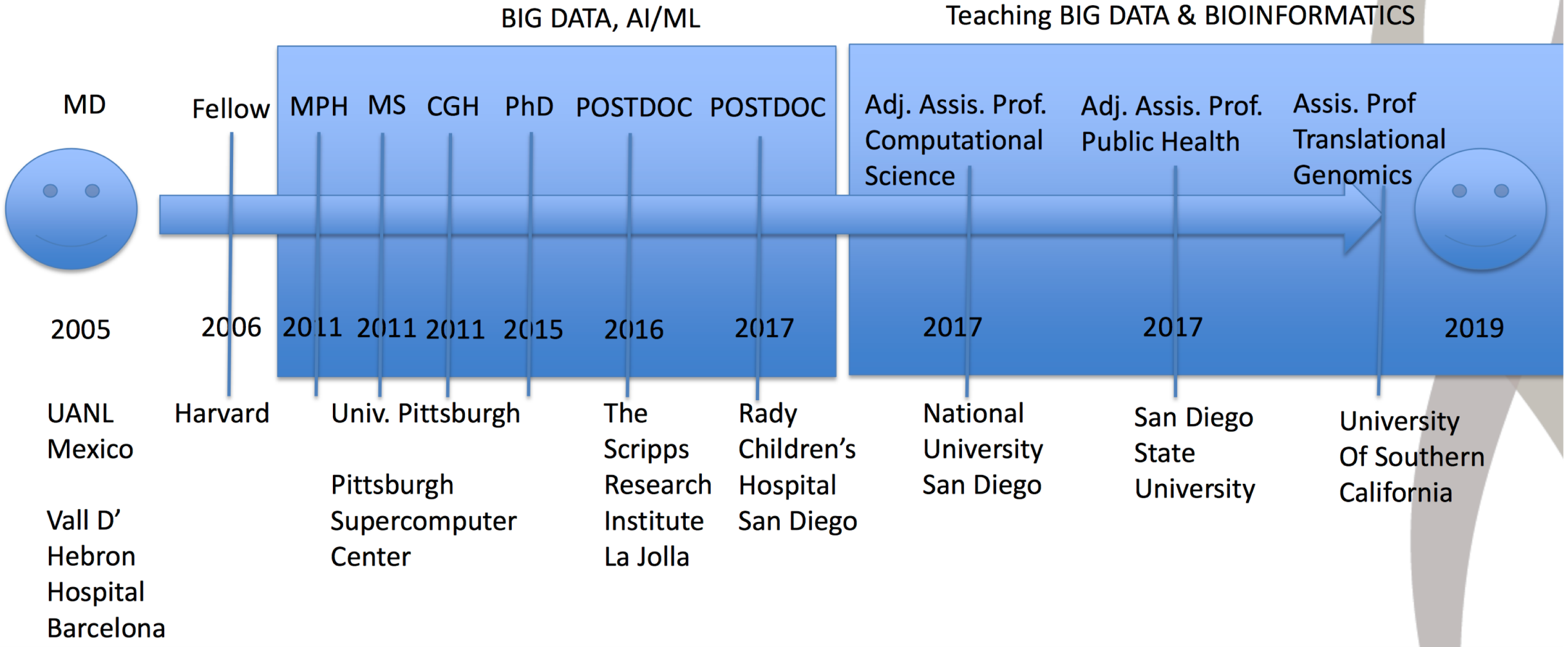
Vall d'Hebron Hospital

INSTITUTO NACIONAL DE NEUROLOGIA Y NEUROCIRUGIA

UANL
UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

Facultad de MEDICINA

My Timeline


BIG DATA, AI/ML
Teaching BIG DATA & BIOINFORMATICS
MD
Fellow
MPH
MS
CGH
PhD
POSTDOC
POSTDOC
Adj. Assis. Prof. Computational Science
Adj. Assis. Prof. Public Health
Assis. Prof Translational Genomics
2005
2006
2011
2011
2011
2015
2016
2017
2017
2017
2019
UANL Mexico
Vall D' Hebron Hospital Barcelona
Harvard
Univ. Pittsburgh
Pittsburgh Supercomputer Center
The Scripps Research Institute La Jolla
Rady Children's Hospital San Diego
National University San Diego
San Diego State University
University Of Southern California

Introduction

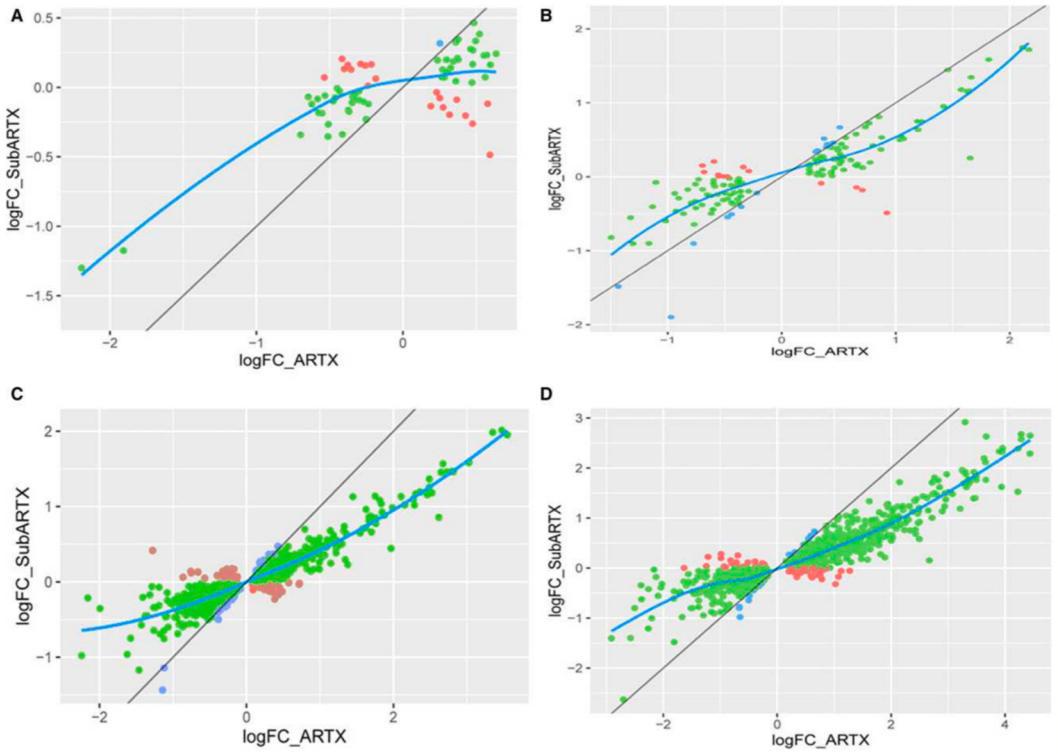
- The objective of this course will provide students from non-quantitative backgrounds with the skill sets for applying data science and bioinformatics tools in the study of human health and disease using R and Bioconductor.

- This course is intended for students who are not experts in either data Science and bioinformatics.

Introduction

- Students will practice data analysis and data visualization by examining challenges inherent in biomedical data, using common computational and statistical open source tools in data science.

- Teaching approaches will alternate between lecture and in-class analysis workshops that will focus on to the selection, application, and reproducible statistical analysis of large-scale multi-faceted 'omic' data from publicly available datasets, such as The Cancer Genome Atlas (TCGA) and ENCODE.



Comparison	cAR vs. TX	subAR vs. TX	Overlap (%)
Microarrays - Blood	2287	720	73 (3.2%*, 10.1%^)
NGS - Blood	2566	1647	143 (5.6%*, 8.7%^)
Microarrays - Biopsies	7376	2931	937 (12.7%*, 32.0%^)
NGS - Biopsies	8922	2565	1188 (13.3%*, 46.3%^)

* - Overlap compared to cAR vs. TX

^ - Overlap compared to subAR vs. TX

Figure 3: Scatterplots of the fold changes for cAR versus TX (x-axes) and subAR versus TX (y-axes) to demonstrate that cAR fold changes are of greater magnitude than subAR changes. (A) Microarrays, blood. (B) RNA-seq, blood. (C) Microarray, biopsies. (D) RNA-seq, biopsies. Green dots denote greater cAR versus TX fold changes, and blue dots denote greater subAR versus TX fold changes. The table shows the number of genes in each comparison and the overlapping genes that were plotted to create panels (A)–(D). cAR, clinical acute rejection; RNA-seq, RNA sequencing; subAR, subclinical acute rejection; TX, transplant with stable function.

Introduction

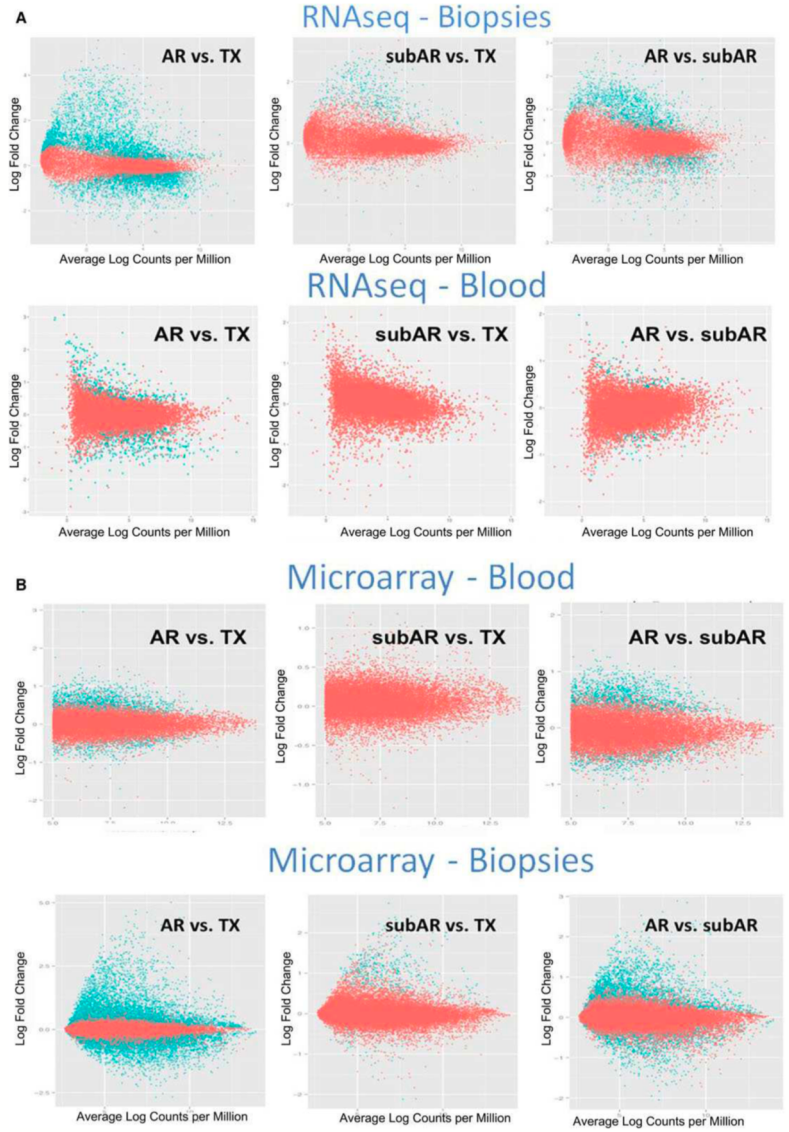


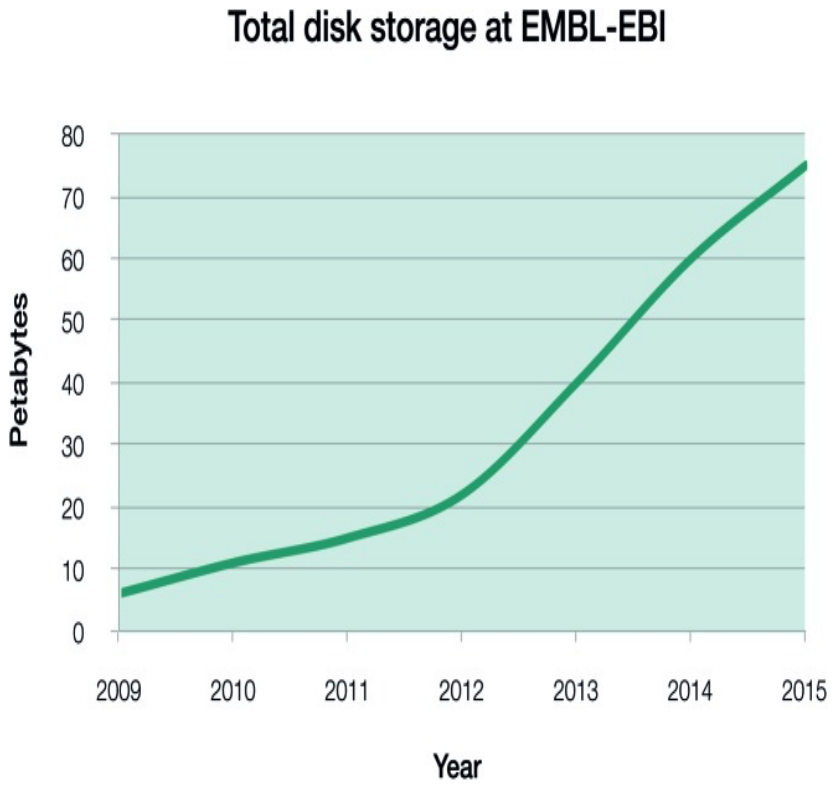
Figure 1: Correlation between the blood findings and the biopsy findings comparing the two analytical methodologies (microarrays and RNA-seq). The similarity between the technologies was also reflected in the consistently higher number of differentially expressed genes in the biopsy compared with the blood. The M (log ratios) and A (average) scale plots for all comparisons are shown for (A) RNA-seq (next-generation sequencing) and (B) microarrays. AR, acute rejection; RNA-seq, RNA sequencing; subAR, subclinical acute rejection; TX, transplant with stable function.

- Within this framework, topics will include basic statistics, hypothesis testing, both parametric and non-parametric analyses (e.g., such as hierarchal clustering and principal component analysis), linear regression analysis, data normalization, reproducibility/sensitivity analysis, multiple test correction, and power assessment.

- Finally, the course will provide an introductory exposure to command-line and Unix-based large-scale data processing, complementing the use of R and Bioconductor as tools for conducting and reproducing analysis frequently required in scientific journals.

Course Description

- The landscape of life-science research is increasingly voluminous, complex and integrative, and is increasingly computational. Bioinformatics skills have become an inherent component of life-science research, particularly 'omic'-based research (proteomics, genomics, metagenomics, etc.).



- The majority of life science researchers lack basic skills in data analysis and interpretation, and especially in data management, even though such skills are essential to many research projects today.

Course Description

- Many students thus progress to advanced life-science degrees without adequate foundations in computational science.

- Even if not all scientists involved in research need to become bioinformaticians, acquiring even a minimum level of bioinformatics understanding can help life- and computational scientists to communicate and interact with one another more effectively (whether in discussions about experimental design or particular analyses, or about specific technical requirements), as well as improve critical thinking about their research findings.

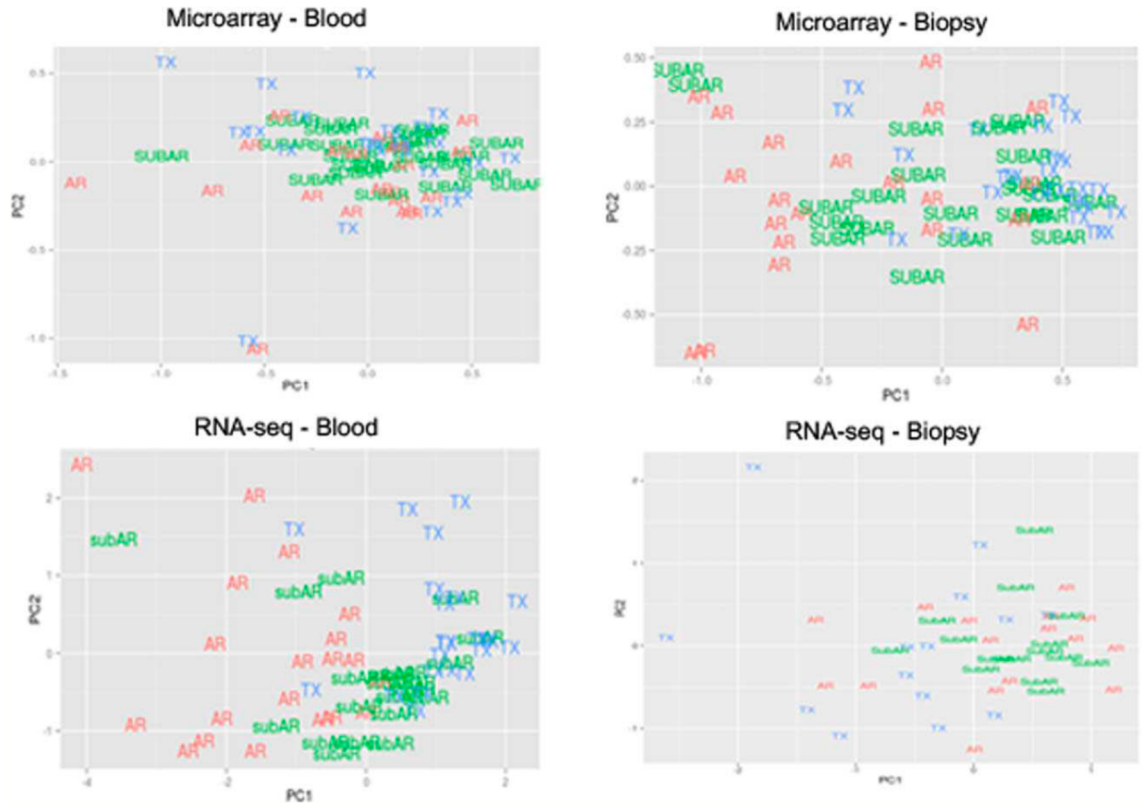
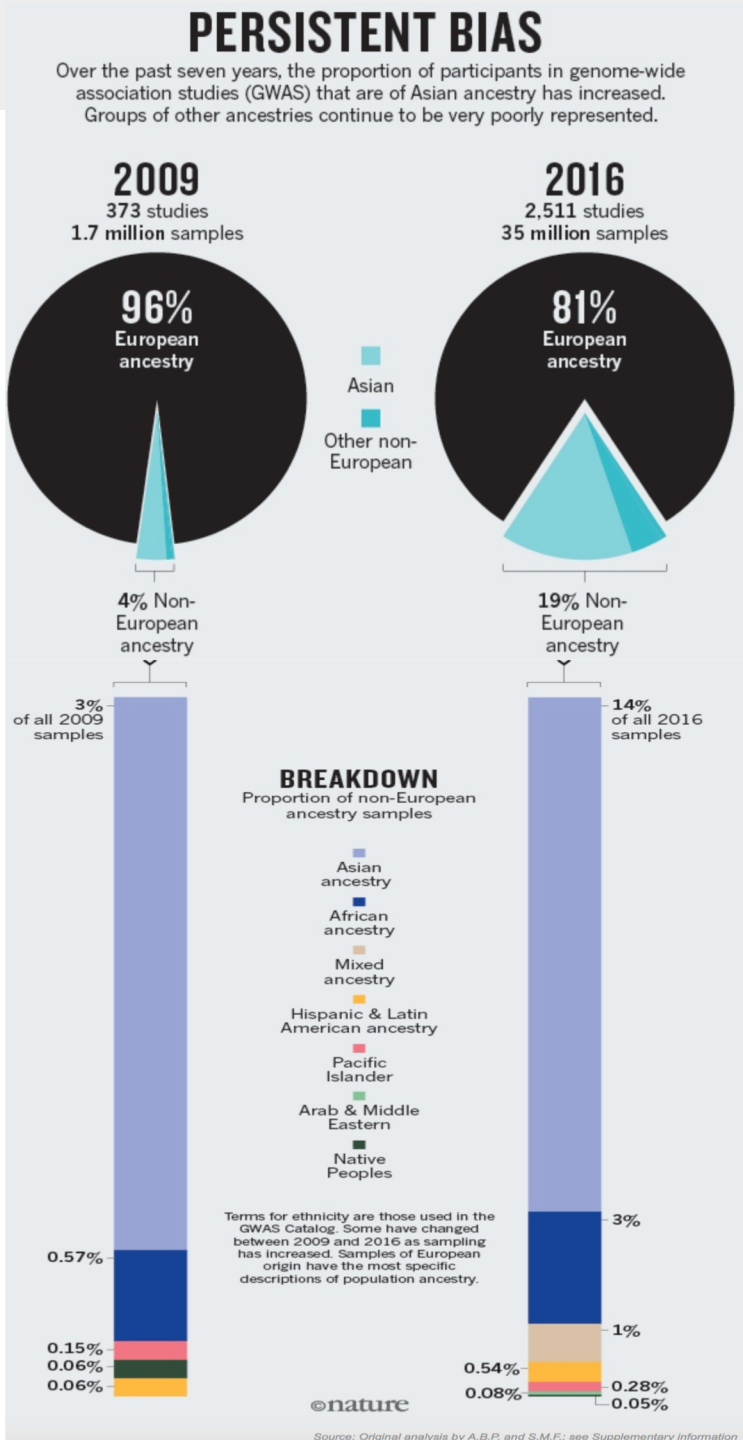


Figure 2: Supervised principal component analysis plots for the two tissues and technologies showing the separation of the phenotypes using the maximum set classifiers. AR, acute rejection; PC, principal component; RNA-seq, RNA sequencing; subAR, subclinical acute rejection; TX, transplant with stable function.

Course Description

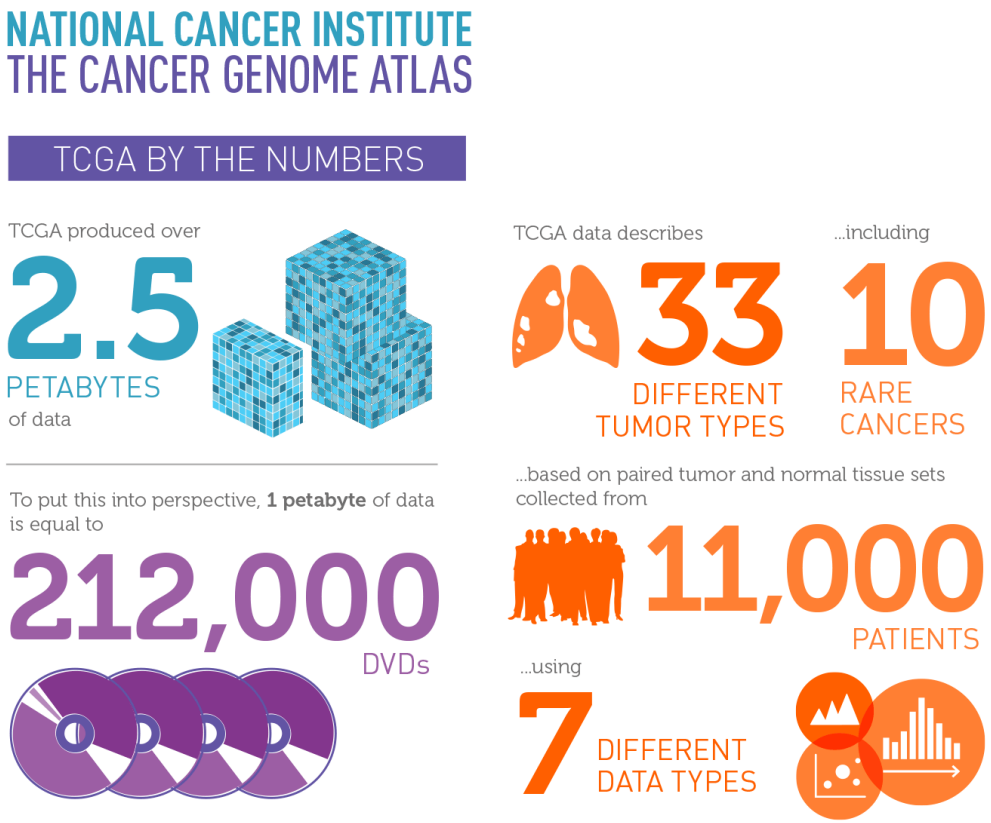
- The objective of this course is to provide advanced data science training in use of data wrangling, data collection, data integration, exploratory data analysis, predictive modeling, descriptive modeling and data visualization for analyzing big biomedical data.

- The course will predominantly utilize publicly available data derived from The Cancer Genome Atlas (TCGA) and Encode as a training set to assess biologics in clinical contexts with translational implications that is intended for students who are not experts in either data analysis and data visualization.



Course Description

- The course will consist of two primary areas 1) reading comprehension and genomic data management and 2) synthesis of concepts in data science and data analysis managing real public datasets with direct implications to multi-omic experimental datasets.

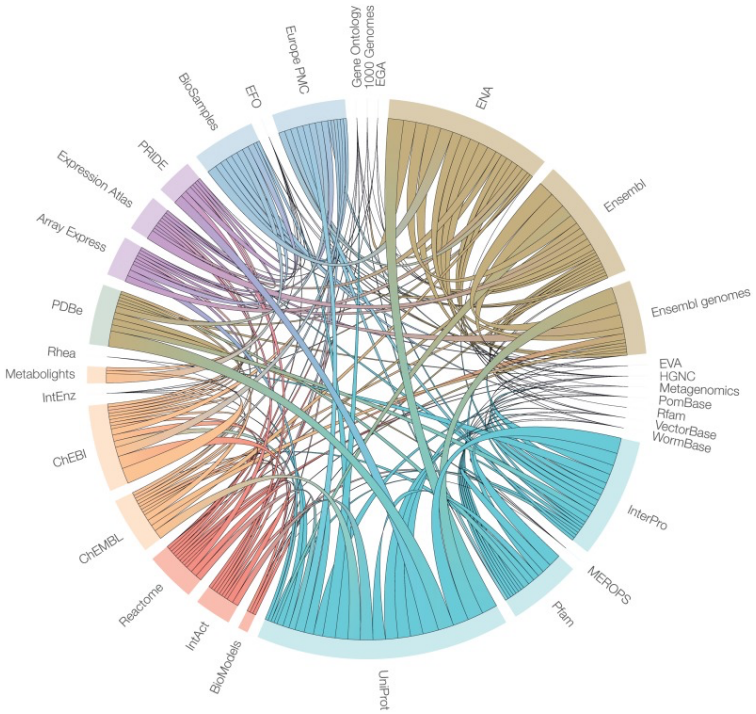


- A major part of this course will be on applied analytics where students will work in teams in a case- study based approach to identify appropriate data and analytic tools to analyze, visualize, evaluate data, and most importantly demonstrate biological and clinical interpretation of results in industry standard formats.

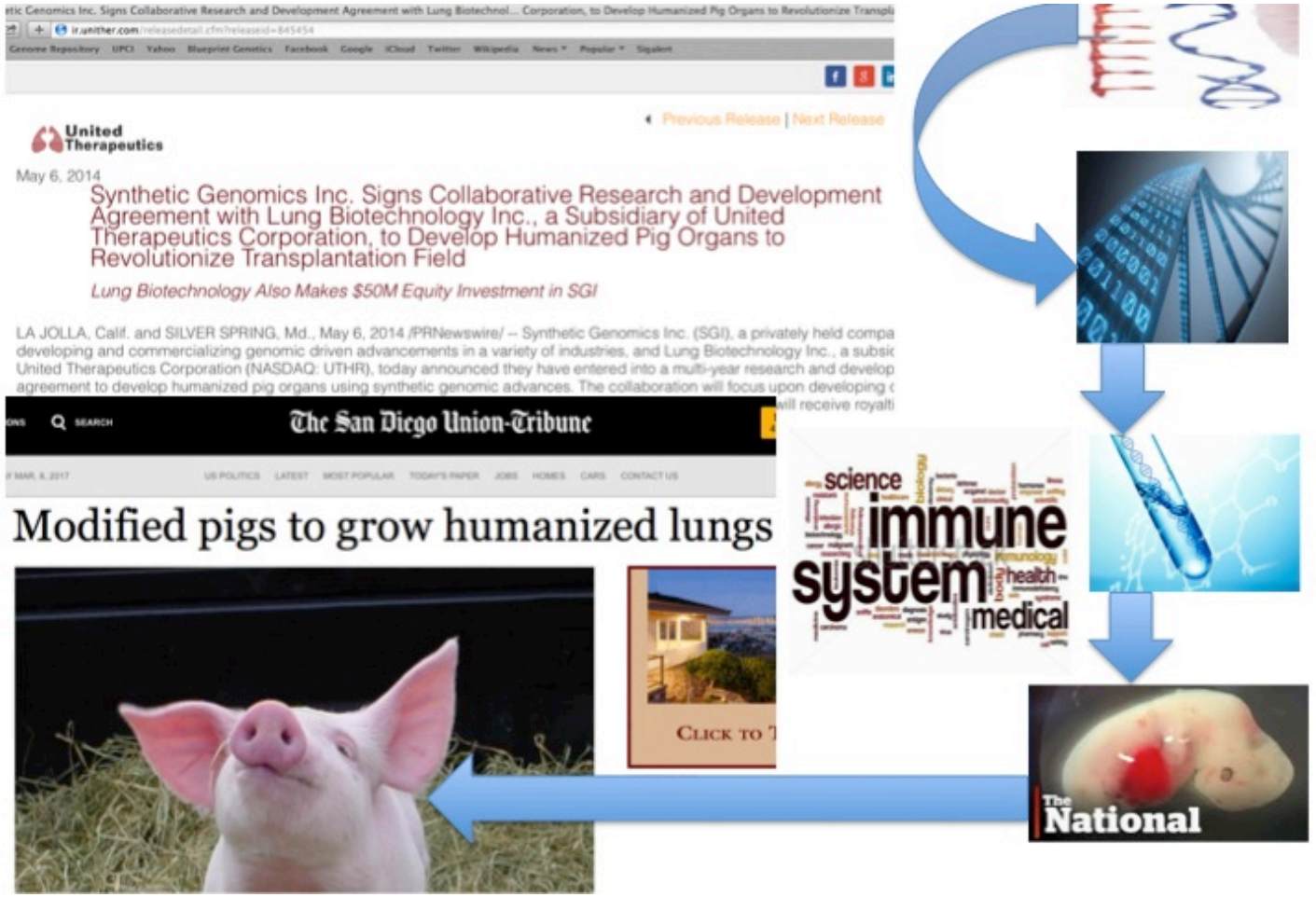
Course Description

- The course will enhance reading comprehension of big data studies and prepare students with essential R Bioconductor packages to analyze and visualize complex genomic data.

- Through this course, students will apply R programming packages to construct advanced genomic pipelines for managing and integrating biomedical data.



Course Description



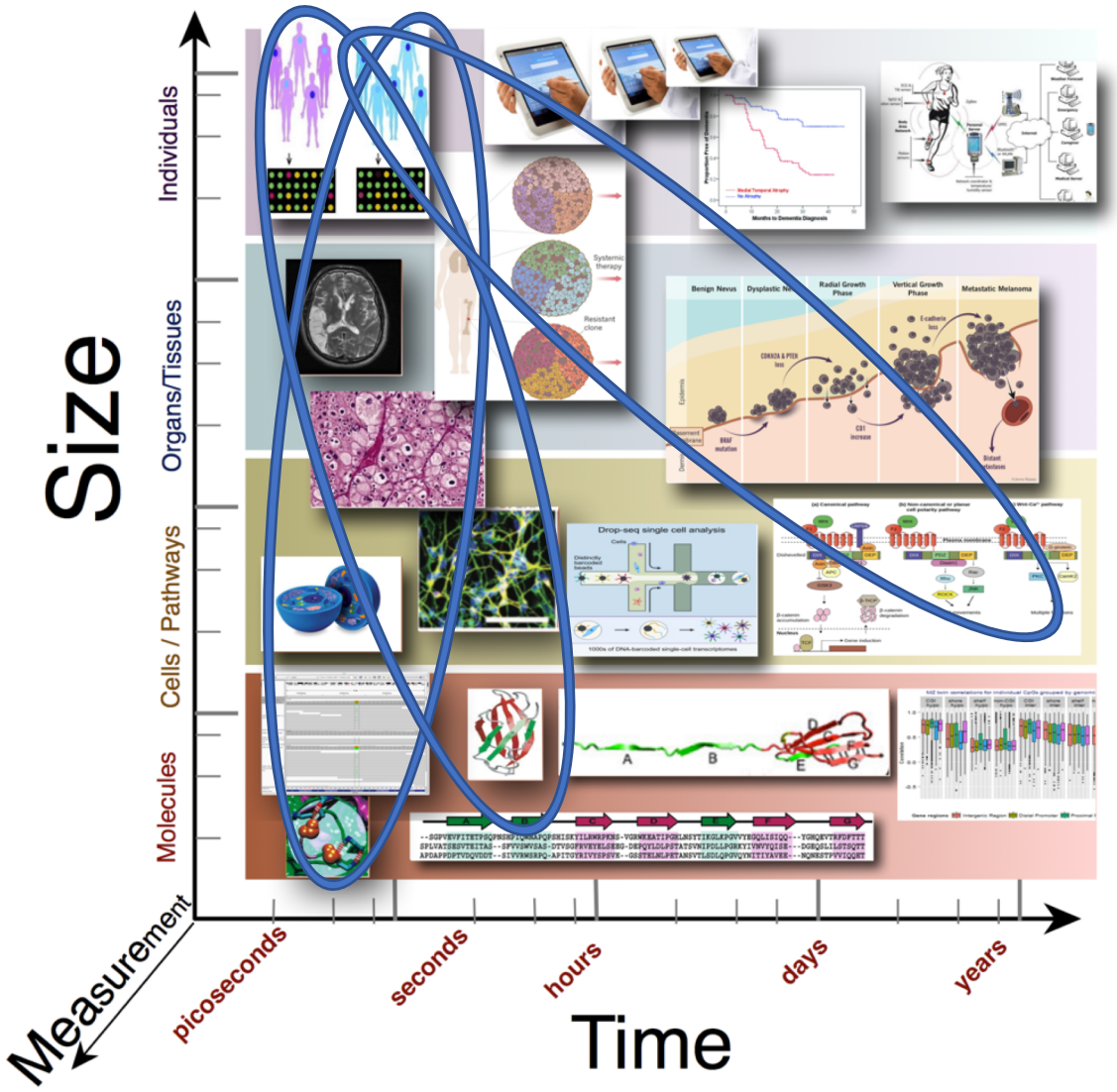
- Students will also be instructed in the proper use and application of data science concepts within R programming in the context of a disease model, research design, and data availability.
- The course will enable proficiency in reading comprehension of large studies with translational implications.
- The course will examine various aspects of data visualization challenges and the importance of certain genomic data analysis best practices and limitations.

Course Description

- Several important multi-omic data analysis standards and computational requirements will be examined to understand the infrastructure needed for certain analyses.

- At the end of the course, students are expected to be equipped with skills in the use of a plethora of open source codes to apply to their study of interest.

- A major goal of this course is for students to be able to apply their big data management skills to conceptually understand data structures and discern the appropriate statistical tools needed to generate high quality publication figures as applicable in the field of translational genomics.

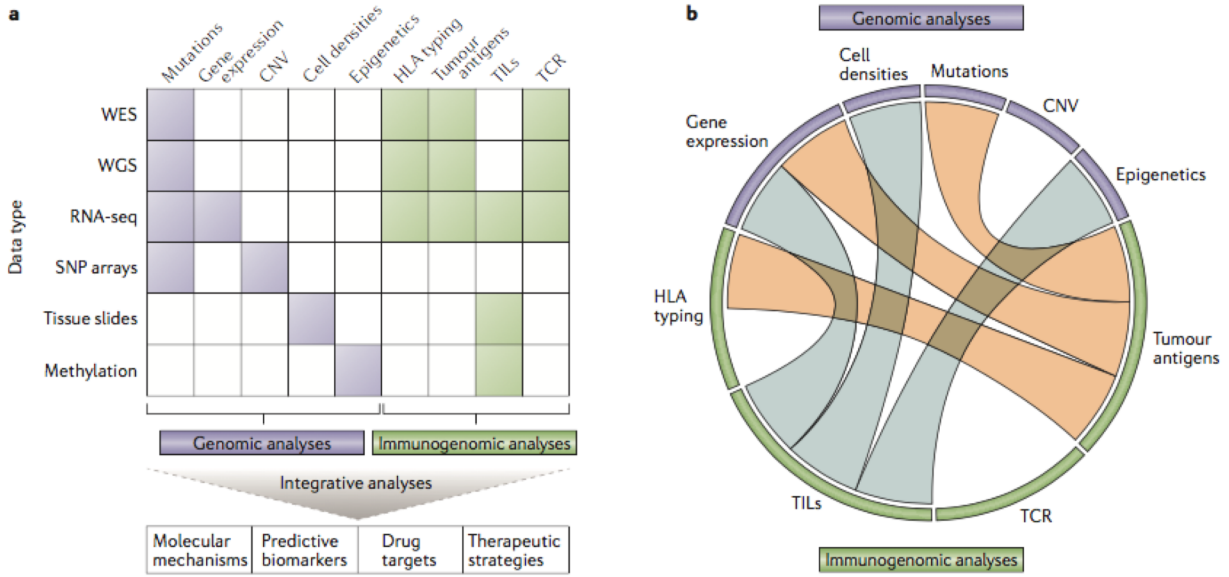


Learning Objectives

- At the end of this course, students will be able to:
  - Select, extract and integrate multiomic data, and other commonly used biomedical databases using open source software for data science
  - Select and apply the appropriate analysis using open source tools primarily from R Bioconductor packages, with an appreciation of different visualization tools frequently used in scientific journal publications.
  - Describe data analysis and data visualization challenges inherent to biomedical data
  - Effectively address scientific questions using common computational and statistical open source tools in data science.
- Prerequisite(s):
  - No pre-requisites needed.

Course Notes

- This course will follow an active learning pedagogy where students will have didactic instruction interspersed with activities performed in groups to allow practice in application of the course content.
- Students are expected to review materials and participate in group activities and homework on centralized servers.
- There will be in person testing and in person presentations at the Keck School of Medicine requiring attendance unless alternative arrangements are made.



Course Notes

- Students will learn by example within their own computers and using cloud-based servers designed to mirror data science work-environments.
- To ensure a focus on the data science challenges, students will be encouraged to have similar computational setups, allowing them the ability to better collaborate on problem solving and ensuring that one student's issue isn't an unusual result of their particular environment.
- Students will interact through Blackboard.
- Students will have specific requirements for software to use specific open source tools.

Blackboard

Course Notes

- A large majority of this course will be conducted using R and R Studio, which are available both on a Windows and a MacOs environment.

- Instructors will not have the ability to troubleshoot computing, and all course instructors and portions of the course will be conducted using demonstrations on a MacOs computer and thus, when possible a MacOs computer is recommended.

Technological Proficiency and Hardware/Software Required

- This course has specific hardware and open source software requirements for data analysis and data visualization. In order to optimize the ability for students to work together effectively, there are specific computing hardware requirements.
- Students will be required to have their own laptops.
- Students will be expected to be able to have access to computers capable of logging into shared servers through ssh, either standard to a MacOS terminal-space or using Putty/TeraTerm on a Windows PC.
- Operating systems should be capable of running R, Java, and should have administrative privileges if any of the versions need to be updated.
- If a loaner laptop is needed, it may be obtained from the USC Computing Center Laptop Loaner Program (details can be seen at https://itservices.usc.edu/spaces/laptoploaner/).

Required Readings and Supplementary Materials

- Weekly required readings will be provided and are described in the course syllabus. Material will be pulled from biomedical journals such as Nature, Science, Cell and other top tier journals available to students via the USC Library services.
- Most required material is generally available at no additional costs to the student, respecting the appropriate content license. For this course, we will utilize R studio for most of the sections.
  - R studio. https://www.rstudio.com
  - Bioconductor. https://bioconductor.org
  - R Cookbook, Proven Recipes for Data Analysis, Statistics, and Graphics. Paul Teetor, O'REILLY.
  - TCGA database. https://tcga-data.nci.nih.gov/docs/publications/tcga/?
  - DataCamp. Introduction to R. www.datacamp.com/courses/free-introduction-to-r
  - Ggplot 2 package. https://cran.r-project.org/package=ggplot2/ggplot2.pdf
  - R-Shiny tutorial. https://shiny.rstudio.com/tutorial GPL 3.0 License
  - Human genotype–phenotype databases: aims, challenges and opportunities. Nature Reviews Genetics 16, 702–715 (2015) doi:10.1038/nrg3932

Description and Assessment of Assignments

- Biomedical data analyses conducted in lecture halls are inherently difficult, especially for those who do not have prior experience.
- This course forms the framework with other concurrent courses, and early participation will be essential.
- The work load for this course will complement other concurrent courses, and the work-load expectations will be front- loaded to insure the foundations are provided within the first half of the course.
- While content will be available through on-line assignments, coursework requires timely iterative completion.
- It will be difficult to catch up, and the requirement for teamwork necessitates that deadlines cannot be individually altered.

Grading Breakdown

- 25% Assignments. Assignments typically consist of functional product, e.g. functional R markdown or documented functional code. 1/5th of the assignment grade is based on turning in results prior to the due date. Due dates are Sunday 11:59PM PT unless otherwise stated based on time-stamps.
- 15% Class participation. Class participation will be based on weekly participation that includes commentary on forums and code repositories of others.
- 25% Midterm Exam. The midterm exam will constitute 25% of the grade and cover key concepts.
- 35% Final Exam. A final exam constitutes 35% of the grade and cover key concepts.

Expectations on Student engagement

- Students are expected to act in a professional manner, meeting deadlines, solving problems, responding to questions from instructors voluntarily or when called upon, cooperating with classmates, and generally contributing in a positive way to the class.
- Working in the real world often means searching for solutions in a group context. Teamwork, listening, empathy, enthusiasm, emotional maturity, and consideration of other people's concerns are all essential to success.
- Please bring these qualities and values with you to class. It is as important to 'practice' these interpersonal skills as it is to learn new intellectual content.

Expectations on Student engagement

- Students are expected to provide feedback to instructors.
- This can be done informally during the semester through the course director or TA.
- It must be done formally by responding to surveys conducted where student anonymity is maintained to ensure that necessary changes may be made to the instructional material, presentation or assessments.

Additional Policies

- Using electronic devices such as smart phones, tablets or laptop computers for conducting personal business during class is prohibited.

Introduction

UNIT I. Introduction and Basic Data Science (w/ example multi-omic datasets)

Week 1

Topic. Course outline, Introduction to R, R studio, Bioconductor, open source software and Terminal. Basics of R, including install and configure software necessary for a statistical programming environment. Data aggregation and pivoting examples. Joining of tables. Loading in data, data frames, Data types (numerical, categorical, ordinal).

Reading Material.

- Class website reading material

Supplemental Reading Material

- R studio. https://www.rstudio.com
- Online material. Bioconductor. https://bioconductor.org
- R project. https://www.r-project.org
- Introduction to the command line.
- https://www.codecademy.com/courses/learn-the-command-line/lessons/navigation/exercises/your-first-command

Assignment #1. Data wrangling exercise using TCGA data

Week 2

Topic. Visualizing data, types of data, and data distributions. Data and data-types – categorical and continuous data. Kaplan Meier, Violin, heatmaps, etc. **D**istributions, normal, log distributions, sampling distributions, confidence intervals, correlations

Reading Material.

- Class website reading material

Supplemental Reading Material

- R Cookbook, Proven Recipes for Data Analysis, Statistics, and Graphics. Paul Teetor O'REILLY.
- TCGA database. https://tcga-data.nci.nih.gov/docs/publications/tcga/?
- DataCamp. Introduction to R. www.datacamp.com/courses/free-introduction-to-r
- Data types in R. https://www.statmethods.net/input/datatypes.html
- Dplyra Tutorial. http://genomicsclass.github.io/book/pages/dplyr_tutorial.html
- P values in R. https://www.cyclismo.org/tutorial/R/pValues.html
- Confidence interval in R. https://www.cyclismo.org/tutorial/R/confidence.html
- Line charts in R. https://www.statmethods.net/graphs/line.html
- Pie charts in R. https://www.statmethods.net/graphs/pie.html
- Scatter plots in R. https://www.statmethods.net/graphs/scatterplot.html

Datasets. TCGA Datasets

Assignment #2. Data visualization using TCGA data

Introduction

UNIT II. Descriptive Statistics (w/ example clinical genomics datasets)	
Week 3	**Topic.** Basic statistics, descriptive statistics, frequencies, data distribution and transformation, variance, and standard error. Prevalence, incidence. Sensitivity, specificity, analytical validity w/ ppv, false discovery rate. **Reading Material.** • Class website reading material • Cancer Genome Atlas Research Network. Comprehensive molecular characterization of human colon and rectal cancer; Nature 2012 **Datasets.** Cancer Genome Atlas Research Network. Comprehensive molecular characterization of human colon and rectal cancer; Nature 2012 **Supplemental Reading Material.** • Basic statistics in R. https://www.statmethods.net/stats/index.html • Descriptive statistics in R. https://www.statmethods.net/stats/descriptives.html • Frequencies in R. https://www.statmethods.net/stats/frequencies.html • Data distribution in R. https://www.statmethods.net/graphs/density.html **Assignment #3:** Descriptive statistics on clinical and genomic data. Analytical Validation in a clinical lab example.
Week 4	**Topic.** Distributions, normal, log distributions, sampling distributions, confidence intervals, correlations, Probability and simulations, Euclidean and other distance metrics. Data normalization and other transformations. **Reading Material.** • Class website reading material **Supplemental Reading Material** • R Cookbook, Proven Recipes for Data Analysis, Statistics, and Graphics. Paul Teetor, O'REILLY. • R for Datascience. https://r4ds.had.co.nz/ • TCGA database. https://tcga-data.nci.nih.gov/docs/publications/tcga/? **Assignment #4.** Data transformation, data plotting and normalization using TCGA data

Introduction

UNIT III. Supervised Statistical Tests (w/ example TCGA datasets)

Week 5

Topic. Tests in Data Science. Supervised vs. Unsupervised, Parametric and nonparametric statistical tests: T-test, Mann-Whitney test, ANOVA, Kruskal-Wallis

Reading Material: Class website reading material

Supplemental Reading.

- Statistical tests in R. http://r-statistics.co/Statistical-Tests-in-R.html
- T-test in R. https://www.statmethods.net/stats/ttest.html
- Non-parametric Statistics in R. https://www.statmethods.net/stats/nonparametric.html
- ANOVA in R. https://www.statmethods.net/stats/anova.html
- Assumptions in ANOVA. https://www.statmethods.net/stats/anovaAssumptions.html
- Cookbook, Paul Teetor, O'REILLY. Chapter 11. Linear Regression and ANOVA.
- Differential Expression: DESeq2: Differential gene expression analysis based on the negative binomial distribution. https://bioconductor.org/packages/release/bioc/html/DESeq2.html

Assignment #5. Statistical analysis on TCGA

Week 6

Topic. Correlation and regression analysis. Analyzing and visualizing breast cancer TCGA data. Build preprocessing heatmaps of gene expression data. Generate gene expression matrix.

Reading Material: Class website reading material

Supplemental Reading.

Berger AC, et al. A Comprehensive Pan-Cancer Molecular Study of Gynecologic and Breast Cancers. https://www.cell.com/cancer-cell/fulltext/S1535-6108(18)30119-3

Correlations in R. https://www.statmethods.net/stats/correlations.html

Regression Analysis in R. https://www.statmethods.net/stats/rdiagnostics.html

Cookbook, Paul Teetor, O'REILLY. Chapter 11. Linear Regression and ANOVA.

Assignment #6: Perform a differential expression analysis (DEA).

Introduction

UNIT IV. Un-supervised Statistical Tests, Multiple Testing (w/ example 1000 Genomes, Single cell datasets)

Week 7

Topic. Correlation, Unsupervised Clustering methods (hierarchical, PCA, tSNE).

Assignment #8: Conduct PCA analysis of 1000 genomes data to identify ancestry migration patterns; Identify batch effects;

Reading Material: Class website reading material

Supplemental Reading Materials.

- Brennan C.W. Verhaak R.G. McKenna A. Campos B. Noushmehr H. Salama S.R. Zheng S. ChakravartyD. Sanborn J.Z. Berman S.H. et al. The somatic genomic landscape of glioblastoma Cell 2013 155462 477
- Clustering Methods in R. https://www.datanovia.com/en/blog/types-of-clustering-methods-overview-and-quick-start-r-code/
- Resampling Statistics in R. https://cran.r-project.org/web/packages/resample/resample.pdf
- Hierarchical Clustering in R. https://www.r-bloggers.com/hierarchical-clustering-in-r-2/
- PCA in R. https://www.datacamp.com/community/tutorials/pca-analysis-r
- tSNE in R. https://www.analyticsvidhya.com/blog/2017/01/t-sne-implementation-r-python/

Week 8

Topic. Calculate and adjust p values, confidence intervals. Survival analysis. Data correlation of gene expression and survival analysis; multiple testing, Bonforonni, Benjami-Hochberg, and false discovery rate.

Assignment #9. Differentially Methylated regions analysis. Volcano plots. Heatmaps with cluster bars; Enrichment downstream analysis in breast cancer.

Reading Materials

- Berger AC, et al. A Comprehesive Pan-Cancer Molecular Study of Gynecologic and Breast Cancers. https://www.cell.com/cancer-cell/fulltext/S1535-6108(18)30119-3
- P values in R. https://www.cyclismo.org/tutorial/R/pValues.html
- Confidence interval in R. https://www.cyclismo.org/tutorial/R/confidence.html
- Line charts in R. https://www.statmethods.net/graphs/line.html
- Pie charts in R. https://www.statmethods.net/graphs/pie.html
- Scatter plots in R. https://www.statmethods.net/graphs/scatterplot.html
- Survival Analysis in R. https://www.datacamp.com/community/tutorials/survival-analysis-R

Heatmaps in R. https://www.r-graph-gallery.com/heatmap/

Introduction

UNIT V. Unsupervised Analysis, linear regression, enrichment analysis (w/ example 1000 Genomes, Single cell datasets)	
Week 10	**Topic.** Unsupervised Clustering methods (hierarchical, PCA, tSNE). **Assignment #8:** Conduct PCA analysis of 1000 genomes data to identify ancestry migration patterns; Identify batch effects; **Reading Material:** Class website reading material **Supplemental Reading Materials.** • Brennan C.W. Verhaak R.G. McKenna A. Campos B. Noushmehr H. Salama S.R. Zheng S. ChakravartyD. Sanborn J.Z. Berman S.H. et al. The somatic genomic landscape of glioblastoma Cell 2013 155462 477 • Clustering Methods in R. https://www.datanovia.com/en/blog/types-of-clustering-methods-overview-and-quick-start-r-code/ • Resampling Statistics in R. https://cran.r-project.org/web/packages/resample/resample.pdf • Hierarchical Clustering in R. https://www.r-bloggers.com/hierarchical-clustering-in-r-2/ • PCA in R. https://www.datacamp.com/community/tutorials/pca-analysis-r • tSNE in R. https://www.analyticsvidhya.com/blog/2017/01/t-sne-implementation-r-python/
Week 11	**Topic.** Correlation and linear regression analysis. Analyzing and visualizing breast cancer TCGA data. Build preprocessing heatmaps of gene expression data. Generate gene expression matrix. **Reading Material:** Class website reading material **Supplemental Reading.** • Berger AC, et al. A Comprehensive Pan-Cancer Molecular Study of Gynecologic and Breast Cancers. https://www.cell.com/cancer-cell/fulltext/S1535-6108(18)30119-3 • Correlations in R. https://www.statmethods.net/stats/correlations.html • Regression Analysis in R. https://www.statmethods.net/stats/rdiagnostics.html • Cookbook, Paul Teetor, O'REILLY. Chapter 11. Linear Regression and ANOVA. • Linear regression in R. http://r-statistics.co/Linear-Regression.html • Logistic Regression in R. http://r-statistics.co/Logistic-Regression-With-R.html **Assignment #6:** Perform a differential expression analysis (DEA).
Week 12	**Topic.** Enrichment analysis (EA): Hypergeometric, Binomial, Chi-squared, Fisher's exact test. Statistical normalization of genes. Quantile filter of genes. Generate boxplot of normalized and non-normalized data. Gene Ontology (GO) and Pathway enrichment bar plots **Assignment:** Differential expression downstream analysis in breast cancer. **Reading Materials** • Berger AC, et al. A Comprehensive Pan-Cancer Molecular Study of Gynecologic and Breast Cancers. Cancer Cell. Volume 33, Issue 4, P690-705.E9, April 09, 2018 https://www.cell.com/cancer-cell/fulltext/S1535-6108(18)30119-3 • Enrichment analysis in R. https://www.bioconductor.org/packages/devel/bioc/vignettes/topGO/inst/doc/topGO.pdf

Introduction

UNIT VI. Enrichment Analysis, Linear Regression (w/ example 1000 Genomes, Single cell datasets)	
Week 13	**Topic.** Case study of LLG. Array intensity correlation. Symmetric matrix of pearson correlation. Spearman and kendall correlation. Identify outliers. Within-lane normalization procedures. **Assignment:** Case Study: Intensity correlation and normalization in LLG. **Reading Material.** • Ceccarelli M. Et al. Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma. Cell. Volume 164, Issue 3, P550-563, January 28, 2016 • Matrix in R. https://www.rdocumentation.org/packages/base/versions/3.5.1/topics/matrix • Correlations in R. https://www.rdocumentation.org/packages/corrplot/versions/0.84 • Pearson, Spearman and Kendall correlation in R. https://www.statmethods.net/stats/correlations.html • Statistical test in R. http://r-statistics.co/Statistical-Tests-in-R.html • Outlier detection in R. http://r-statistics.co/Outlier-Treatment-With-R.html
Week 14	**Topic.** Loess robust local regression and global-scaling. GC-content effect. Preprocessing operations for clustering. Hierarchical clustering algorithm. Hierarchical cluster analysis **Assignment:** Case Study: Hierarchical downstream analysis in low grade glioma. **Reading Material.** • Ceccarelli M. Et al. Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma. Cell. Volume 164, Issue 3, P550-563, January 28, 2016 • Hierarchical clustering in R. https://www.datacamp.com/community/tutorials/hierarchical-clustering-R • Clusters in R. https://www.rdocumentation.org/packages/survival/versions/2.42-6/topics/cluster
Final Review	
Week 15	**Catchup Day**
Week 15	**Final Review EXAM**
Week 16	**FINAL EXAM**

Introduction

QUESTIONS?