# Broad Summer Research Program Remote Computational Genomics Course

## The digital classroom

* Github Classroom
  + Group repository: <https://github.com/BSRPCompGenomics>
  + Classroom: <https://classroom.github.com/classrooms/73200083-bsrpcompgenomics-2021-classroom>
* Google Classroom
  + 2020: <https://classroom.google.com/u/2/c/MTExMjEzNTY1NjM4>

## References

Resources Chris used to develop the curriculum.

### Content

#### Primary

CS61A Berkeley <http://composingprograms.com/>

Data8 Berkeley <https://www.inferentialthinking.com/>

Analysis of Biological Data by Michael Whitlock. Datasets used in textbook: <https://whitlockschluter.zoology.ubc.ca/>

List of papers used can be shared via Mendeley groups.

##### Secondary

MIT CompBio <https://www.youtube.com/c/ManolisKellis1/videos>

HSPH Rafalab Biomedical Data Science <http://genomicsclass.github.io/book/>

HSPH Rafalab Data Science <https://rafalab.github.io/dsbook/>

### Pedagogy

MIT Digital Learning Toolkit in preparing a course <http://dltoolkit.mit.edu/best-practices/>

MIT PBL resources <http://web.mit.edu/jrankin/www/pbl/pbl.html>

MIT Blossoms PBL Team norms worksheet

<https://blossoms.mit.edu/sites/default/files/project/page_files/Team-Agreements-v2.pdf>

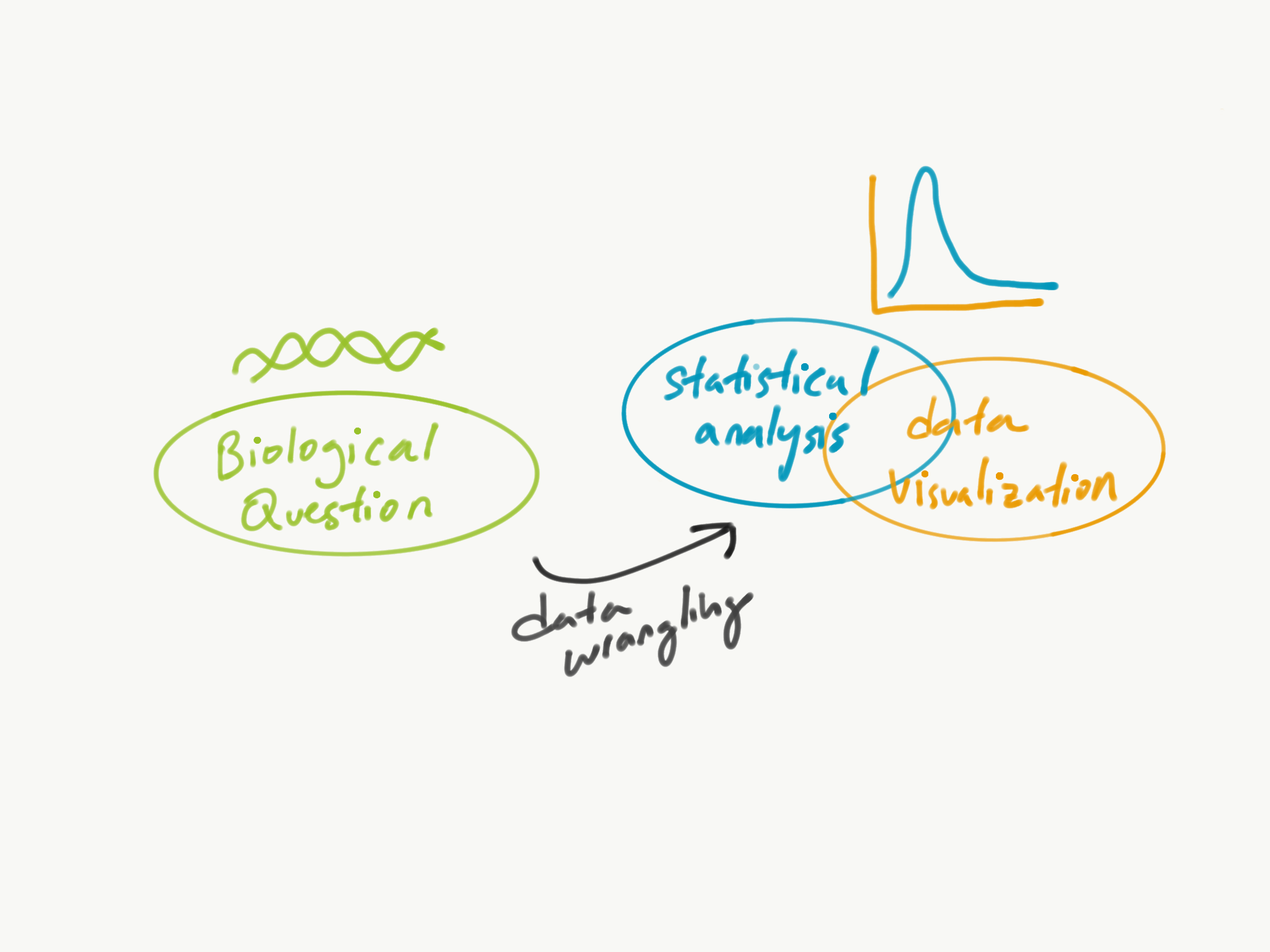
## Course themes

The course can be cut through two different lenses: a computational perspective, and a genomics perspective.

Key:

We were able to cover this in BSRP 2020.

### L1: You are a life science researcher. How can you answer biological questions in a different, new way via computation?



| Biological Question | Computational Analysis, Data Visualization |
| --- | --- |
| What is the summary of gene X in all of our samples? | Summary statistics, histograms |
| Is mutation of gene X associated with gene Y’s expression level? | T-test, boxplot |
| Which genes are most differentially expressed under gene X’s mutation? | Differential expression analysis, volcano plot |
| What biological pathways are in common to a set of genes? | String DB, connected graphs |
| Does gene X’s expression associate with gene Y’s expression? (or dependency) | Correlation, scatterplots, univariate regression |
| Is gene X’s mutation associated with gene Y’s dependency? | T-test, boxplot |

#### Computational takeaways for students (concepts and important examples)

* Computing as modular functions that reduces redundancy (Functions)
* Writing basic code (Loops, logical statements, scoping)
* Methods of encoding biological information (Mutation -> boolean, expression -> numeric)
* Methods of organizing biological information (Gene expression matrix, mutation MAFs)
* Wrangling complex datasets into a format ready for analysis (Tidyverse)
* Plotting multidimensional data visualization (ggplot)
* Exploring cancer biology via genomics (Cancer Genome Landscapes by Vogelstein et al., KRAS pathway)
* Analyze ways to summarize a sample (Summary statistics)
* Recognize randomness that comes with sampling, and how to infer it back to the population (Probability calculations, statistical inference)
* Interpret a statistical test, and the errors that it comes with (Hypothesis testing, false positive, false negatives)
* Recognizing the problems of confounding variables (Subset analysis)
* Association vs. causation vs. prediction (Association between genes vs. functional genomics vs. predicting a functional outcome based on genomics)

### L2: How can cancer genomics play a role in cancer precision medicine?

*Precision Medicine:* Prioritizing treatment using a patient’s genomic profile. From identifying candidate variants to validating its function to therapeutics.

samples → variants → genes → cell types → processes → disease

1. Characterization of cell lines (Week 1, Week 2)
   1. Revealing new characterizations of cancer genomics
   2. Understand omics technologies and sequencing
   3. Cell lines as a model system
2. Characterization of genes that are likely be casual for cancer, based on patterns of abundance (W2)
   1. Distinguishing driver and passenger mutations
   2. Characterizing oncogenes and tumor suppressor genes
3. Validating cancer drivers by using functional knock-out data (W3)
   1. Can our candidates for drivers be validated by functional knock-outs?
   2. Is there a drug that targets a candidate dependency?
   3. Why is genomics useful to associate dependency and drug sensitivity?
4. Genome-wide search space: what feature is most closely associated with a gene? (W4)
   1. How does looking across the genome help understand associations better? Ie. differential gene expression, genomewide correlation

## Timeline and Curriculum

Subject to change based on student progress, of course!

Instructions for students if they have interest to start Datacamp modules early.

[Setup for students once they arrive.](https://docs.google.com/document/d/1Z3ms7Ur_CIMTxQGj_OPk7rOPoEsqM6JBdgZEjpdzlSA/edit)

### Week 1

Monday is off for Memorial Day.

Goals: Introduction to genomics. FIrst lesson on coding in R. Get familiar with vectors and dataframes. Get setup in R and github classroom. First launch into looking at data with short practices of mean, median, standard deviation, and proportions.

* [Genomics: what does that look like, computationally?](https://docs.google.com/presentation/d/1me3ilXuLGsCHA77CRuWsB_WefOY1ANbu/edit?usp=drive_web&ouid=111930923017680744005&rtpof=true)
  + Watch [Eric Lander’s talk on the past, present and future of genomics](https://www.youtube.com/watch?v=VAvMSzBJaqQ&ab_channel=BroadInstitute). (Cooper/Chris)
  + Present scope of genomics and bioinformatics, and what we are going to focus on in this course.
  + Introduce course themes: how does one convert a biological problem to a computational problem? How does one understand a disease from variant identification to mechanism?
* [Introduction to computing:](https://docs.google.com/presentation/d/1KC8uWnNmw92JMXx5Nw5XuehH8xUGRpez/edit#slide=id.p1) (Chris)
  + Computation as a calculator
  + The power of abstraction in computing using variables and functions
  + Execution rules for evaluating a function, variable assignment, what is the global environment
  + Rstudio Demo
  + R: data types, data objects (vector, dataframe, matrix), properties of data objects, subsetting, and boolean statements
  + [Datacamp: Introduction to R.](https://learn.datacamp.com/courses/free-introduction-to-r) Skip sections 3, 4, and 6: matrix, factors, and lists.
* Github Classroom and Rstudio Demo
* [PSet 1:](https://github.com/BSRPCompGenomics/pset1-template) 
  + Short exercises
  + Longer practice: Look at CCLE metadata to pull out descriptive statistics.
* Friday Discussion: (Cooper; Chris on CCLE)
  + Main figures of CCLE paper (2019 or 2012) to convey the -omics profiling of CCLE. Save drug sensitivity and functional associations for later.
  + -Omics - what are the molecular techniques to generate the data?
  + Descriptive statistics: mean, median, standard deviation, and proportions.

### Week 2

Goals: Deeper into R with loops, functions, conditionals. Start doing data wrangling. Reduce the types of problems of data wrangling into: creating new columns, intersection, normalizing, and collapsing/grouping (summarizing by summary statistic). Intro to mutation/expression data and how people have looked at cancer driving genes computationally.

* Pset1 Discussion
  + How does one debug a program?
* Schedule 1:1s with students to get to know them better and their scientific interests to help form groups later.
* [Functions, loops, conditionals](https://docs.google.com/presentation/d/1_Ell2cQIH-H4lp3RV4HZuj8UxMrU3V9L/edit#slide=id.p1)
  + Writing functions
  + Variable scoping, global vs. local environments
  + Iteration and loops
  + Conditional statements
  + Datacamp: [Intermediate R](https://learn.datacamp.com/courses/intermediate-r), sections 1-3: Conditionals, Functions, Loops
* [Wrangling biological datasets](https://docs.google.com/presentation/u/1/d/1LNfAPH717Y6COU-xpcxHk8P7-UU8R4z7/edit?usp=drive_web&ouid=111930923017680744005&rtpof=true)
  + Dataframes as observations, variables, values
  + Tidyverse: mutate, group\_by, summarize, join, pipes
  + (Might need to cover later: long vs. wide formats
  + Datacamp: [Introduction to Tidyverse](https://learn.datacamp.com/courses/introduction-to-the-tidyverse), sections 1 and 3: data wrangling, grouping and summarizing
* Review sequencing technology: [Illumina’s review module.](https://support.illumina.com/content/dam/illumina-support/courses/sequencing-fundamentals/story_html5.html?iframe)
* [Demo of bam files, and IGV.](https://docs.google.com/document/d/1OV7lCS59YZws__eCH7QJEeUf49bxnrf6jBC7YYQvYxo/edit)
* [PSet2*:*](https://github.com/BSRPCompGenomics/pset2-template) 
  + Short exercises
  + Read KRAS mutation/expression paper by Stephens et al.
  + Read cancer genomics landscapes paper by Vogelstein et al.
  + Longer practice: Implement KRAS paper Figure 1 on CCLE data.
* [Friday Discussion: Cancer Genomics](https://docs.google.com/presentation/u/1/d/1xJllNfX1MMaWOpTmf_0R5J-mZ-aci3YV/edit?usp=drive_web&ouid=111930923017680744005&rtpof=true)
  + Cancer biology 101, discuss Vogelstein et al. cancer genomics review paper
  + DIscuss the history and importance of cancer precision medicine, using Bill Seller’s talk from 2019 with BSRP.
  + PSet2/Discuss Stephen et al. KRAS mutation/expression paper

### Week 3

Friday off for Juneteenth.

* [Data Visualization](https://docs.google.com/presentation/u/1/d/1L58mgM5iczzMv30a-3ZokG5id1e7uF5i/edit?usp=drive_web&ouid=111930923017680744005&rtpof=true)
  + ggplot
  + What makes a good scientific figure?
* [PSet3](https://github.com/BSRPCompGenomics/pset3-template)
  + Read the CCLE [press release.](https://www.broadinstitute.org/news/writing-book-cancer-knowledge)
  + Read Depmap [press release/history](https://www.broadinstitute.org/blog/effort-map-cancer%E2%80%99s-weak-spots-starting-reveal-new-drug-targets).
  + Short exercises
  + Longer practice: Interpret Figure 1 statistical tests, validate using dependency data

[Friday Discussion:](https://docs.google.com/presentation/d/105pLkGvhdWlOuGpdi20M1pD37YRjJmqM/edit#slide=id.p1)

* CRISPR and functional genomics
* Dependency data
* How is functional genomics useful for precision medicine? Associating genomics with CRISPR and drug sensitivity data
  + PSet3

### Week 4

* [Statistical inference and hypothesis testing](https://docs.google.com/presentation/u/1/d/1bilJClRIFgQmANDJDq0ZrPhJjHEYauu6/edit?usp=drive_web&ouid=111930923017680744005&rtpof=true)
  + Population, sample, randomness
  + Computing probabilities
  + Parameter, statistics
  + Sampling distribution
  + Law of large numbers
  + Confidence intervals
  + Hypothesis testing, p-value, rejecting/not rejecting the null hypothesis
  + T-test for sample mean, 2 sample T-test for difference in mean
  + Small population, biased sampling in medical research: Small sample sizes of rare diseases and underrepresented populations in medical research: [Count Me In](https://www.youtube.com/watch?v=-L4BnpEJd30), Rare Genomes Project, Eurocentric population in genetic studies by [Alisha Martin](https://aliciarmartin.com/),
  + Possible student reading: Chapter 4 (Estimating with Uncertainty) and Chapter 6 (Hypothesis Testing) from Analysis of Biological Data
* [Differential gene expression](https://docs.google.com/presentation/u/1/d/1sObs8B_VI4Yp0YPuzzTSPCv3P7p_B_Z9/edit?usp=drive_web&ouid=111930923017680744005&rtpof=true)
  + Errors in testing: False positives, False negatives
  + T-test Power calculator - <https://www.statskingdom.com/32test_power_t_z.html>
  + Multiple testing problem
  + Multiple testing correction methods: bonferroni correction, false discovery rate
  + Reproducibility crisis in science, p-hacking
* Pick and choose (can supplement these lectures in project weeks):
  + [Confounding](https://docs.google.com/presentation/u/1/d/1RUULzu1L47wQKUYfekZk5zWC_GAyhNA1/edit?usp=drive_web&ouid=111930923017680744005&rtpof=true)
* [Correlation and Regression](https://docs.google.com/presentation/d/1qNH2ffkjb3AETqjHETlbPe1V3mRa2aAr/edit?rtpof=true)
  + Pearson correlation
  + Anscombe's quartet
  + Univariate regression
  + Assumptions for regression, residuals
  + Connect back to dependency

Other topics:

* GSEA
  + String DB
  + Clustering
* [PSet4](https://github.com/BSRPCompGenomics/pset4-template)
  + Run power calculations for PSet3.
  + Run differential gene expression analysis (limma)
* Friday Discussion:
  + PSet4
  + Assignment of student groups

### Week 5

Friday is off for 4th of July.

* Student start their projects
* BSRP lab meeting, project meetings

### Week 6

Monday is off for 4th of July.

* Student projects
* BSRP lab meeting, project meetings

### Week 7

* Student projects
* BSRP lab meeting, project meetings

### Week 8

* Finalize analysis, put together to presentation
* Practice talks with Broad faculty

### Week 9

* Presentations!
* Wrap-up, reflections, what to do to keep learning