Biomedical Data Science - Final Project (Spring 2026)

Analysis of Carl Zimmer's Personal Genome: Sex Chromosome Study (X or Y)

Project Overview

In this project, you will analyze real genomic data from science journalist Carl Zimmer, who has made his personal genome publicly available. Your team will focus on one assigned sex chromosome (X or Y) and investigate its germline variants — naturally occurring DNA differences in his genome. The project is divided into two parts: first, identifying genes with the highest mutational burden on your assigned chromosome and comparing variant patterns to the rest of the genome; second, performing a deeper biological analysis of those genes through one of four analytical lenses (expression, network, structure, or literature). This is an opportunity to apply the computational and analytical skills from the course to real human genomic data.

Key Due Dates:

- 5-minute recording and slides: Noon of April 21st , 2026 (Canvas submission)
- 5-minute recorded presentations: April 22nd , 2026 (in class)
- 20-minute discussion presentations: April 23rd and 24th, 2026 (discussion section)
- Written report and deliverables: May 3rd , 2026 (Canvas submission)

Group Assignment

- Students will work in teams on topics of interest. Team composition will be balanced by students enrolled in non-programming and programming modules. The teams have been assigned already. Email us ASAP if there are issues with your team.
- We encourage team members to work together in a collaborative environment on both the analysis and written parts of the project. If any student feels their voice was not heard while working on the project, please reach out to the TAs as soon as possible. At the end of the submitted write-up, please include each team member's contribution.

Documents and Deliverables

Each team is required to submit SIX documents as well as any supplementary files, all together in one zipped folder:

1. Written Report (PDF): Introduction, Methods, Results, Discussion, and Team Members' Contributions. Minimum 1000 words (excluding references and contributions).
2. 20-minute presentation slides (PDF): Under 15 slides. The final slide must be a one-slide summary of your methods and key findings.
3. Text File (5 lines):
   Line 1: Project Description
   Line 2: Section # and Group # (i.e., section 1, group 1)

Line 3: Assigned Chromosome (chrX or chrY)
Line 4: Top 10 Genes (comma-separated) (i.e. GENE1, GENE2, …)
Line 5: Key Variants if Applicable (separated by commas) (i.e. rs283843, rs2383828 OR chr17:3832322:C:A, chr17:28383839:T:G; etc.)

4. VCF File: *gene_variants_chrX.vcf* or *gene_variants_chrY.vcf* containing variants within selected genes.
5. Code: all the codes for the project. This could be a single bash/python/R file, or a zip folder including all the codes.
6. One-Slide Summary: Finally, the sixth document is a 1 slide summary of your project. Please use the exact layout and template found here: [One-slide Summaries]

This zip folder should be submitted on Canvas by May 3rd , 2026.

Presentations

You will make 2 presentations. **One 5-minute recorded presentation** to be played in the class, and **one 20-minute presentation in discussion session**. All team members must present and participate in both.

Note on Presentation Slides: The 5-minute and 20-minute presentations use separate slide decks, though you are welcome to reuse slides across both. Both decks must end with the one-slide summary (see *Documents and Deliverables* for the required template). Your 5-minute slides (due April 21st on Canvas) should be concise, while your 20-minute slides are part of the final deliverables zip folder (due May 3rd) and should provide a more comprehensive walkthrough of your analysis.

5-Minute Recorded Presentation

- Presentation day: **April 22nd, 2026 (Wednesday), 1:00 PM, BASS 305**. All group members come to the stage; the recording will be played, followed by a **3-minute Q&A** with the audience and instructors. Both the recording and Q&A will be graded.
- Submit the video (mp4) **and slides** on Canvas **before noon on April 21st**. If the file is too large, upload to YouTube or Google Drive and submit the link.
- Given the short format, not everyone is required to appear on camera — your group can decide how to organize the recording. However, all group members must be physically present on presentation day to participate in the Q&A.

20-Minute In-Person Presentation

- 20-presentation followed by a 5 minute Q&A.
- Delivered in discussion section on April 23rd or 24th.
- All group members are required to speak during this presentation.

Note:

- Carl Zimmer will join on presentation day to discuss your findings on his personal genome — this should be a unique and engaging experience for everyone.

- Since presentations are due before the final writing, prioritize getting your results ready first and focus on writing afterward.

Grading

Final grades will be based on the content and clarity of written summary, both presentations, analysis, and any submitted code.

Generally, group members will share the same grades. Unless otherwise noted, all team members will receive the same grade.

Grading Scheme

Component	Weight
5-minute recorded presentation	20%
20-minutes in-person presentation	25%
Written report	30%
Analysis & Code	20%
Extra Credit	+5%

Analysis Topics

Each team will analyze one assigned chromosome (X or Y). Carl's germline SNPs are found here under Germline SNP call set for subjectZ. Coordinates are based on the GRCh37 version of the human genome. The file is in VCF format. For more information about VCF, please see here.

Part 1: Gene Prioritization and Chromosome-Level Analysis

A. Gene Prioritization (10 Genes)

- Given the germline variant call (VCF), find 10 genes on the chromosome you are assigned with the highest mutational burden (i.e., number of mutations).
- List the genes and submit records of the variants you identified in the prioritized genes in a file called gene_variants_chr{i}.vcf, where i is the number of the chromosome your team is assigned.
- In your report, describe the steps you take to identify the variants in the genes of interest. Make sure to mention any database or software tool you use. If you write your own code, please make sure to include it in the final submission.

B. Chromosome-Level Statistical Comparison to Autosomes

Because sex chromosomes differ structurally and evolutionarily from autosomes, comparing them helps us understand whether their variant patterns are typical or biologically distinct.

- Using the VCF file, perform one quantitative comparison between your assigned chromosome (X or Y) and the autosomes (chromosomes 1–22).
- Define the metric you use to compare variant burden and explain how you computed it.
- Describe any gene-wise/ chromosome-wise normalization or adjustment applied (if applicable).
- Present your results clearly (table or figure) with sufficient explanation and discussion.

[Extra Credit] Chromosome-Specific Requirements

If assigned chromosome X:

Identify whether your analysis includes genes in PAR and/or non-PAR regions. Briefly comment on whether variant patterns appear different between these regions (if applicable). Discuss how the fact that chromosome X is hemizygous in males may influence interpretation of variant burden.

If assigned chromosome Y:

Determine whether any of your selected Y-linked genes have homologs on chromosome X. If applicable, briefly compare variant patterns between homolog pairs. Comment on any unexpected genotype patterns observed on chromosome Y (e.g., heterozygous calls) and suggest possible explanations.

Part 2: In-Depth Analysis of Selected Genes

Now that you selected 10 genes from Part 1, each team will choose one of the following areas and perform in-depth analysis on the prioritized genes.

1. Gene Expression Analysis.
   Using GTEx (https://gtexportal.org/home/) or similar resources:
   - Examine expression profiles of your prioritized genes across tissues.
   - Identify general patterns (e.g., tissue-specific, broadly expressed, low vs high expression).
   - Discuss what these patterns may suggest about biological function.
   - Support your interpretation with at least two references.

2. Network analysis.
   Choose ONE of the following approaches:
   a. Protein-protein interaction network (e.g., STRING), OR
   b. Pathway enrichment analysis (e.g., KEGG, Reactome, MSigDB)

   You must:
   - Present one clear network or pathway figure.
   - Explain the key biological processes represented.
   - Discuss how variants in your prioritized genes might influence these interactions or pathways.
3. Protein structure analysis.
   Using PDB or AlphaFold:
   - Identify available protein structures for your prioritized genes (if available).
   - Visualize at least one structure.
   - Highlight any amino acids affected by exonic variants (if applicable).
   - Discuss potential structural or functional implications.

4. Text mining analysis.
   Using PubMed:
   - Analyze at least 20 publications related to your prioritized genes.
   - Identify recurring biological themes or disease associations.
   - Compare your findings with annotations from UniProt or GeneCards.
   - Briefly discuss consistencies or discrepancies.

If you have questions regarding the final project, please contact the TFs at cbb752@gersteinlab.org