A couple notes about assignments

- It doesn't matter how you get the answer as long as you get the answer
- If you get a few questions wrong, that's fine. What's important is that you understand the material. If you don't understand something, please ask questions!
- Answer keys with example solutions (there are ALWAYS multiple ways to answer these questions) will be posted after each assignment is due. There's no plan to go over assignments with the group.
  - With the exception of the first assignment. Either you were able to commit or you didn't so there wasn't really anything to post solutions for

How to Troubleshoot Code
and
How to Read a Scientific Paper

How to Troubleshoot Code

What should you do when you get an error?

1. Check for typos
   - Did you close all (), “”, ‘’, []?
   - Did you spell everything correctly? Data table, function name, column name?
2. Run the code line by line so you can figure out what piece of code is causing the error
3. Read the messages
   - What does the error message say? Have you seen it before?
   - Is RStudio giving you any messages or warnings? It has some built-in error-checking
4. Check the help documentation
   - ?function()
   - Google the function name and check more extensive documentation online
5. Google the error message; you are not the first person to have the problem and someone will have a solution.
6. Ask for help! During this program from us, but in the future, you can ask friends, professors, colleagues, Stack Overflow

DEMONSTRATE RSTUDIO, GOOGLE

How to Read a Scientific Paper

Anatomy of a Paper

Title: Anatomy of a Paper

Keith, K, Calendo G, Madzo J

Abstract: This cartoon paper is to demonstrate how a scientific paper is laid out

Introduction: the paper will describe any necessary background.

Results: Here's where the paper will walk you through its data. It will explain what that data means and illustrate it with figures (graphs, photographs) and tables

Discussion: The paper will summarize its findings and state the conclusions.

Methods: Traditionally, the methods come after the introduction, but increasingly journals are choosing to bury them at the end of the paper. This section describes all the steps the authors took to describe the experiments.

References:

1. Fosmire, M and Edmonson A. How to Read a Scientific Paper. https://www.lib.purdue.edu/sites/default/files/libraries/engr/Tutorials/Newest%20Scientific%20Paper.pdf

What Section To Read First?

- Once you decide you want to read the paper, you don't have to read it in order. There are arguments to be made for the introduction, the discussion, the results, the methods.
- I suggest going through in order when you're starting out
  - When you're not familiar with the field, the introduction will have crucial information you'll need to understand the rest of the paper.
  - Why NOT the other sections?
    - For bioinformatics, the methods descriptions are usually useless. Even if there's code available, it takes too long to go through to make it worth it
    - If you don't know the background, the results and discussion are very hard to read

How to Read a Paper (Or How I Read a Paper)

- Look up any words you don't understand and define them for yourself
- Highlight important points as you go along
- Restate the main points of each section as you go along
- Take time looking at the figures
- At the end summarize for yourself the 2-3 main points of the paper

- IgG antibodies bind to short tandem repeats (STRs), particularly CA repeats
- competition for binding b/t mark of interest and STRs
- STR binding doesn't seem to happen in intact chromatin (possibly because STRs are inaccessible)

nature methods

ANALYSIS

https://doi.org/10.1038/s41592-018-0038-7

A reassessment of DNA-immunoprecipitation-based genomic profiling

Antonio Lentini[1], Cathrine Lagerwall[1], Svante Vikingsson[2,3], Heidi K. Mjoseng[4], Karolos Douvlataniotis[1], Hartmut Vogt[1], Henrik Green[2,3], Richard R. Meehan[4], Mikael Benson[1,5] and Colm E. Nestor[1,5*]

DNA immunoprecipitation followed by sequencing (DIP-seq) is a common enrichment method for profiling DNA modifications in mammalian genomes. However, the results of independent DIP-seq studies often show considerable variation between profiles of the same genome and between profiles obtained by alternative methods. Here we show that these differences are primarily due to the intrinsic affinity of IgG for short unmodified DNA repeats. This pervasive experimental error accounts for 50–99% of regions identified as 'enriched' for DNA modifications in DIP-seq data. Correction of this error profoundly altered DNA-modification profiles for numerous cell types, including mouse embryonic stem cells, and subsequently revealed novel associations among DNA modifications, chromatin modifications and biological processes. We conclude that both matched input and IgG controls are essential in order for the results of DIP-based assays to be interpreted correctly, and that complementary, non-antibody-based techniques should be used to validate DIP-based findings to avoid further misinterpretation of genome-wide profiling data.

The ability to establish and maintain DNA-methylation patterns is essential for normal development in mammals, and aberrant DNA methylation is observed in numerous diseases, including all forms of cancer[1]. Comprehensive mapping of DNA methylation (5-methylcytosine (5mC)) in multiple species has established the relevance of methylation dynamics to gene regulation and chromatin organization[2–4]. An effective method for the generation of genome-wide 5mC profiles couples antibody-based enrichment of methylated DNA fragments (methyl-DNA immunoprecipitation (MeDIP)) with hybridization to DNA microarrays or high-throughput sequencing[5,6]. Relevant MeDIP sequencing information is contained not in the read sequence itself, but in the enrichment or depletion of sequencing reads that map to specific regions of the genome[7,8]. Consequently, appropriate control samples are required, which typically correspond to the 'input' genomic DNA before enrichment. More recently, DIP-seq has been extended to chart the genomic location of additional DNA modifications,

between 50% and 99% of enriched regions in DIP-seq data are false positives, the removal of which markedly affects the perception of methylation dynamics in mammals. Our findings will substantially improve the accuracy of future DIP-seq experiments and allow new insights to be gained from the wealth of existing DIP-seq data.

Results

IgG antibodies bind short tandem repeats in mammalian DNA. To simplify the comparison of DIP-seq results from separate studies, we used a uniform computational pipeline (Methods) to analyze published DIP-seq profiles of 5mC, 5hmC, 5fC and 5caC (hereinafter referred to collectively as 5modC) in mouse embryonic stem cells (mESCs). Supplementary Table 1 presents an outline of all analyzed datasets and their relationship to the figures. This approach revealed striking enrichment at STRs in all 5modC DIP-seq datasets (Fig. 1a and Supplementary Fig. 1) that could not be explained by nonspecific binding of the antibodies to other modifications, as the

DEMO

Journal Club Assignment

- 20 minute presentation summarizing the main points of one of the six review papers posted on the course website
- 3 presentations next Friday 7/17 and 3 presentations the following Friday 7/24
- Sign up sheet for the paper and the date you want is in this GoogleDoc https://docs.google.com/spreadsheets/d/1TVXcyu1PKn4d_Wk2hEjOWLUxLkwE4U_xLLDlVM2k2CU/edit?usp=sharing (link also on course website). First come, first served!
- There are example slides from a journal club I gave at Coriell, as well as a link to a YouTube video where you can listen to an example Journal Club on the course website as well. Also, please keep in mind that there are many more slides than you'll need in the example PDF because it was for an hour-long presentation.