Class Projects

Single Molecule Sequencing: Methods, Training, and Applications

I1: Building on the ideas from Homework 3, the student would perform a descriptive study of current signals from either -002 or -004 flow cells, attempting to parse out differences between the chemistry on the basis of upstream and downstream quantitative metrics, from both the raw signal current and potential downstream alignments. The student would summarize these findings in a descriptive study, determining potentially what features cause -004 to exhibit a higher accuracy.

I2: The Elephant Genome project has sequenced the genome to a depth of 7X (https://www.broadinstitute.org/elephant/elephant-genome-project). Updated reference genomes and transcriptomes can be found at this address (https://ftp.ncbi.nlm.nih.gov/genomes/refseq/vertebrate_mammalian/). Another recent group sequenced Elephant samples with both traditional short-read RNA sequencing and PacBio long-read RNA sequencing (https://onlinelibrary.wiley.com/doi/10.1111/acel.13917). The task for this student will be to re-construct reference transcriptomes based on direct-RNA sequencing data for the Elephant, utilizing StringTie, and compare these to existing reference transcriptomes with Gffcompare. The student should then present preliminary analysis regarding the functional annotation of the transcriptome.

I3: This course is producing the first Nanopore direct-RNA sequencing samples of elephant samples. Choose a conserved gene with large regions of homologous sequence across multiple species, and generate aligned signal files for these genes. Ideally, there is reliable epitranscriptomic information available for at least one of those species. What do the differences in current signal between the species tell you about the state / status of the RNA in each of those species? Perform downstream quantitative analysis via basecalling, error analysis, and alignment; what do the results suggest?

I4: Consider the principles of Hardigan's Dip Test: "The dip test measures multimodality in a sample by the maximum difference, over all sample points, between the empirical distribution function, and the unimodal distribution function that minimizes that maximum difference." Develop a strategy on this basis to recognize epitranscriptomic modifications within the elephant direct-RNA sequencing samples. Choose one high-coverage gene to visualize and statistically ascertain the presence of a potential modification. Propose a strategy for performing this analysis in high-throughput fashion.

I5: Generate simulated current signal and slow5 files, base-calling them in order to examine the result. Propose a hypothesis for how changing the input current leads to a resulting output during base-calling. Use both the regular base-calling model and the m6a-DRACH base-calling model.

I6: Examine annotated full length reads, along with truncated reads of similar kmers, to recognize current changes associated with 7-methyl guanosine. Can 7-methyl guanosine be detected in the Nanopore? Propose a strategy for developing a 7-methyl guanosine basecaller and possible significance.

I7: Analyze mitochondrial RNA reads from multiple flow cells to ascertain whether there are separate populations of currents which could be classified / grouped according to their modification status. Evaluate the current state of the field with regard to epitranscriptomic modifications on mitochondrial RNA; propose your own theory on the basis of direct-RNA seq analysis.

I8: The calibration strand is a synthetic RNA added to all direct-RNA sequencing samples (https://help.nanoporetech.com/en/articles/6632031-what-is-rna-cs-rcs). Suggest some possible controls for how this control could be used in cross-sample comparison, and demonstrate the utility of this control in an analysis between two biological conditions.

I9: Run the Dorado m6a basecalling model and benchmark m6anet according to the results. Determine accuracy and precision for m6anet.

I10: Take signals aligned to a high-coverage gene [>2,000 reads if possible]. Use pytorch to generate a machine learning model that classifies whether a signal belongs to a gene or not. Run all signals from the sample through the machine learning model. Calculate accuracy and precision statistics. Base-call the incorrect calls. Are there any trends for the kinds of signals the model got wrong?

I11: Implement Hardigan's dip test for the Inspiration4 *m6anet* outputs. Re-interpret the biological conclusions, and integrate with the expression-based counts to develop relevant longitudinal measurements.

I12: Implement Hardigan's dip test on an assortment of *m6anet* outputs. Perform a GWAS-style analysis between m6a probabilities and single-nucleotide polymorphisms, interpreting the biological significance of the results.

I13: Implement Hardigan's dip test on an assortment of *m6anet* outputs. Perform a GWAS-style analysis between m6a probabilities and expression profiles, interpreting the biological significance of the results.

I14: Your own idea! Feel free to email tmn2126@columbia.edu and chm2042@med.cornell.edu.

Logistics: These are individual projects; posters are due according to the syllabus deadline. Please feel free to collaborate with others.

Milestones:
May 7th, 2024: Check-In Meeting with tmn2126@columbia.edu during Lab Session
May 14th, 2024: Data Exploratory Analysis Due [before Class]
May 21st, 2024: Signal-Level Results Summary Due [before Class]
May 28th, 2024: Final Poster First Draft Due [before Class]
May 31st, 2024: Final Poster Due