**Statistics**

Noble, W. S. (2009, December). How does multiple testing correction work? Retrieved from <https://www.nature.com/articles/nbt1209-1135.pdf>

Annotation: This article discusses how multiple test correction works, and applications that it can be used for. Correction is 0.05 / size of dataset.

Zach. (2020, March 24). How to perform a mann-whitney u test in excel. Retrieved March 08, 2021, from <https://www.statology.org/how-to-perform-a-mann-whitney-u-test-in-excel/>

Annotation: A **Mann-Whitney U test** is used to compare the differences between two samples when the sample distributions are not normally distributed, and the sample sizes are small. It is considered to be the nonparametric equivalent to the two-sample t-test. Examples are included in the document.

Meeta Mistry, R. (2017, October 13). Gene-level differential expression analysis with deseq2. Retrieved March 08, 2021, from <https://hbctraining.github.io/DGE_workshop_salmon/lessons/05_DGE_DESeq2_analysis2.html>

Abstract: Differential expression analysis with DESeq2: model fitting and hypothesis testing. It also covers hypothesis testing

**Data models**

The Toulmin model of argument. (2018, February 12). Retrieved March 08, 2021, from <https://englishcomposition.org/advanced-writing/the-toulmin-model-of-argument/>

Abstract: The Toulmin Model is a tool for analyzing and constructing arguments.  It was created by British philosopher Stephen Toulmin and consists of the following six parts: claim, grounds, warrant, backing, rebuttal, and qualifier.

Training, validation, and test sets. (2021, March 03). Retrieved March 08, 2021, from <https://en.wikipedia.org/wiki/Training,_validation,_and_test_sets>

Abstract: In machine learning, a common task is the study and construction of algorithms that can learn from and make predictions on data. Such algorithms function by making data-driven predictions or decisions, through building a mathematical model from input data.

Li, Y., Liu, Y., Juedes, D., Drews, F., Bunescu, R., & Welch, L. (2019, September 17). Set cover-based methods for motif selection. Retrieved March 08, 2021, from <https://academic.oup.com/bioinformatics/article/36/4/1044/5570984>

Abstract: De novo motif discovery algorithms find statistically over-represented sequence motifs that may function as transcription factor binding sites. Current methods often report large numbers of motifs, making it difficult to perform further analyses and experimental validation. This can be used to express the overall coverage of a dataset by the data.

Zheng, P. (2018, August 07). An overview of methods to address the multiple comparison problem. Retrieved March 14, 2021, from https://towardsdatascience.com/an-overview-of-methods-to-address-the-multiple-comparison-problem-310427b3ba92?gi=553c637ad8c4

Abstract: This article lists various ways of multiple test correction.

**CIMP relations in cancer**

Miller, B., Sánchez-Vega, F., & Elnitski, L. (2016, November 22). The emergence of pan-cancer cimp and its elusive interpretation. Retrieved March 10, 2021, from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5197955/>

Abstract: Epigenetic dysregulation is recognized as a hallmark of cancer. a CpG island methylator phenotype (CIMP) has been documented in tumors originating from different tissues. Includes a synopsis of the history of CIMP and describe the pattern of DNA methylation that defines the CIMP phenotype in different cancer types.

Levine, D. (2013, May 01). Integrated genomic characterization of endometrial carcinoma. Retrieved March 08, 2021, from <https://www.nature.com/articles/nature12113>

Abstract: We performed an integrated genomic, transcriptomic and proteomic characterization of 373 endometrial carcinomas using array- and sequencing-based technologies. Tumour samples and corresponding germline DNA were collected from 373 patients, including 307 endometrioid and 66 serous (53) or mixed histology (13) cases. Local Institutional Review Boards approved all tissue acquisition.

KD. Robertson, A., R. Jaenisch, A., A. Portela, M., Baylin, S., SB. Baylin, P., M. Rodríguez-Paredes, M., . . . MS. Cline, B. (1970, January 01). Pan-cancer stratification of solid HUMAN Epithelial tumors and cancer cell lines REVEALS commonalities and tissue-specific features of the CPG ISLAND METHYLATOR PHENOTYPE. Retrieved March 08, 2021, from <https://epigeneticsandchromatin.biomedcentral.com/articles/10.1186/s13072-015-0007-7>

Abstract: The term CpG island methylator phenotype (CIMP) has been used to describe widespread DNA hypermethylation at CpG-rich genomic regions affecting clinically distinct subsets of cancer patients. We analyze genome-wide patterns of CpG island hypermethylation in 5,253 solid epithelial tumors from 15 cancer types from TCGA and 23 cancer cell lines from ENCODE. We identify differentially methylated loci that define CIMP+ and CIMP− samples, and we use unsupervised clustering to provide a robust molecular stratification of tumor methylomes for 12 cancer types and all cancer cell lines. With a minimal set of 89 discriminative loci, we demonstrate accurate pan-cancer separation of the 12 CIMP+/− subpopulations, based on their average levels of methylation.

**Biological Implications**

Reimand, J., Isserlin, R., Voisin, V., Kucera, M., Tannus-Lopes, C., Rostamianfar, A., . . . Bader, G. (2019, February). Pathway enrichment analysis and visualization OF omics data Using G:profiler, GSEA, CYTOSCAPE and enrichmentmap. Retrieved March 08, 2021, from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6607905/>

Abstract: Pathway enrichment analysis helps researchers gain mechanistic insight into gene lists generated from genome-scale experiments. This method identifies biological pathways that are enriched in a gene list more than would be expected by chance. The protocol comprises three major steps: definition of a gene list from omics data, determination of statistically enriched pathways, and visualization and interpretation of the results.

Dees, N., Zhang, Q., Kandoth, C., Wendl, M., Schierding, W., Koboldt, D., . . . Ding, L. (2012, August). Music: Identifying mutational significance in cancer genomes. Retrieved March 14, 2021, from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3409272/>

Abstract: we present a packaged suite of comprehensive, user-friendly tools designed to determine mutational significance in cancer (MuSiC). The primary goal of MuSiC is to separate the significant events which are likely drivers for disease from the passenger mutations present in mutational discovery sets using a variety of statistical methods. We assessed multiple methods of calculating summarized P-values, including a convolution test (CT), a Fisher's combined P-value test (FCPT), and the likelihood ratio test (LRT), using a partially simulated data set (this data set and the associated test simulations are described in the Supplemental Material). By this approach, we determined that the P-value distribution obtained using the CT method most closely resembled the uniform distribution expected under the null (in this case, the null is such that no gene is truly significantly mutated), while the FCPT and LRT methods produced slightly inflated or deflated P-values, respectively (Supplemental Fig. S1). During the SMG test, a false discovery rate (FDR) also is calculated. We evaluate our SMG test results by establishing a P-value or FDR threshold (threshold typically 0.2 or less for FDR), and then appropriately filtering the test output.