Title

Workflow development for GDC cancer data of AML and ALL samples

Authors

Olamide Adefioye[1] and Jonathon E. Mohl[1,2]

[1]Bioinformatics Program and [2]Department of Mathematical Sciences, The University of Texas at El Paso, El Paso, TX

Abstract

Acute Myeloid Leukemia (AML) and Acute Lymphoid Leukemia (ALL) are two distinct hematological malignancies characterized by uncontrolled proliferation of undifferentiated myeloid and lymphoid cells, respectively. While phenotypic and cytogenetic methods are traditionally employed for classification and prognosis, molecular profiling offers deeper insights into the disease's biology, potentially unveiling novel therapeutic targets. This study sought to elucidate the differential gene expression patterns between different groups of AML and ALL patients, utilizing RNA sequencing (RNA-seq) data. The Genomic Data Commons (GDC) is an initiative of the National Cancer Institute (NCI) that provides a unified platform for sharing genomic, transcriptomic, and clinical data among cancer researchers. Our workflow approach was centered on downloaded sequencing data processed by the program "Spliced Transcripts Alignment to a Reference (STAR)" aligner obtained from the GDC repository. Python was used for the preliminary data cleaning, followed by differential expression analysis in R using DESeq2. This workflow is important because it can also be used to compare subgroups between AML or ALL or other types of cancer datasets from the GDC.

Title

# Cumulative functional effects of pathogenic single nucleotide variants on protein-coding genes

Authors

Amanda Bataycan,[1] Jonathon E. Mohl,[1,2] and Ming-Ying Leung[1,2]

[1]Computational Science Program and [2]Department of Mathematical Sciences, The University of Texas at El Paso, El Paso, TX

Abstract

The purpose of this work is to devise a quantitative assessment of the cumulative effects of single nucleotide variants (SNVs) on the protein-coding genes in patients with acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL), aiming to find novel candidate leukemia-related genes for future studies. With published project data as primary resources, SNV information from whole-exome data of 149 patients with AML and 603 patients with ALL was obtained. The overall numbers of distinct SNVs found in the AML and ALL datasets were 136,084 and 181,998 respectively, with the vast majority occurring only in tumor samples. Both AML and ALL favored mutations that reduce GC content in genes.

A functional effect analyzer, FATHMM-XF, was applied to obtain a numerical score that anticipates the deleterious effect for each SNV. Predictions are made using models formulated from interactions between data sources, where pathogenic SNVs are identified with predicted scores 0.5 or higher. The FATHMM-XF predictions were implemented within a gene scoring formula, called Q(gene), to assess the cumulative effect on a gene caused by pathogenic variants found in its protein-coding transcripts. In total there were 9,761 AML and 15,544 ALL pathogenic variants distributed respectively among 6,325 and 7,054 protein-coding genes. Using the top 1% Q(gene) scores, approximately 5% of AML and 18% of ALL genes matched with known AML and ALL-related genes in public literature. The performance of Q(gene) in selecting leukemia-related genes is expected to improve by incorporating additional functional effect evaluations of the variants.

Title

Identifying UCEs to design baits for phylogenetic analysis of monogonont rotifers

Authors

Ratna Jyothi Bolem,[1] Elizabeth J. Walsh,[1,2] and Jonathon E Mohl[1,3]

[1] Bioinformatics Program, [2]Department of Biological Sciences, and [3]Department of Mathematical Sciences, The University of Texas at El Paso, El Paso, TX

Abstract

Rotifers, the microscopic invertebrates, play essential roles in freshwater ecosystems. With approximately 2,200 known species and many more awaiting discovery, comprehending the genetic diversity and evolutionary history of rotifers poses a daunting challenge with past attempts using traditional genes (e.g., 18S RNA, COI) resulting in an unresolved phylogeny of the phylum. In this study, we harness the power of Ultra-Conserved Elements (UCEs) as indispensable molecular markers. These highly conserved DNA sequences transcend species boundaries, offering a promising avenue for unraveling the intricate genetic landscape of rotifers. Our goal is to revolutionize rotifer species identification and clarify evolutionary relationships. To identify the UCEs, we employed the versatile Phyluce software package, known for its comprehensive workflow that encompasses preprocessing, aligning the base genome, conserved locus identification, conserved locus validation, and final bait design. This workflow seamlessly identifies and processes UCEs, unveiling their pivotal role in our research. Beyond identification, Phyluce's bioinformatics tools empowered us to pinpoint target loci of UCEs with precision. In our study encompassing 15 rotifer species, we successfully identified a total of 258 target loci. These target loci will serve as the foundation upon which we will design baits—customized sequences for capturing UCE loci. This crucial step represents a strategic leap forward in our exploration of rotifer species complexity.

Title

# Predictive modeling of cancer stage and location in colorectal cancer patients using transcriptomic data

Authors

Robert Diaz,[1] Brian Grajeda,[2] and Sourav Roy[1,2]

[1]Bioinformatics Program and [2]Department of Biological Sciences, The University of Texas at El Paso, El Paso, TX

Abstract

Colorectal cancer (CRC) was ranked as the third most prevalent cancer globally in 2020, with nearly 1.93 million new cases. In the United States, ethnicity significantly influences the incidence and mortality rates of CRC; notably, Hispanic populations exhibit lower early detection rates compared to Non-Hispanic Whites (NHW). In 2018, CRC accounted for 11% and 9% of all cancer-related fatalities in Hispanic men and women, respectively, alongside 12% and 8% of all new cancer diagnoses. Early detection of CRC elevates survival rates by 90%; however, merely 39% of CRC patients are diagnosed early, largely due to inadequate screening, especially among Hispanics. Limited healthcare access further exacerbates this disparity. Consequently, Hispanics often face advanced-stage CRC diagnoses, correlating with diminished survival rates. In this study, transcriptomic data from 13 Hispanic and 15 NHW CRC patients were generated using high-throughput sequencing platforms like Illumina. Raw data were processed to attain sequenced reads via CASAVA base recognition, stored in FASTQ format. The reads were then aligned and annotated using the Rsubread library against a human genomic reference from NCBI. The generated transcriptomic counts, alongside variable descriptors for each sample, formed the basis of our analytical dataset. Utilizing the tidymodels library in R, the clean data were partitioned into testing (25%) and training sets (75%), followed by cross-validation. Through specifying the target variable and predictors, a Random Forest model facilitated a 95% prediction accuracy in determining cancer stage and location, showcasing a promising avenue for enhancing early CRC detection and intervention, particularly among disadvantaged ethnic groups.

Title

Comparing prediction algorithms for RNA secondary structures with pseudoknots

Authors

Dan Du,[1] Khodeza Begum,[2] and Ming-Ying Leung[1,2,3]

[1]Computational Science Program, [2]Border Biomedical Research Center, and [3]Department of Mathematical Sciences, The University of Texas at El Paso, El Paso, TX

Abstract

The secondary structures of Ribonucleic Acid (RNA) encompass two fundamental categories: stem-loops and pseudoknots. Both patterns play significant roles in vital biological processes, including gene expression and regulation. Pseudoknots are now widely acknowledged as prevalent motifs with diverse and crucial functions. Despite the substantial contributions of computational RNA secondary structure predictions to our understanding of RNA's molecular mechanisms, accurately predicting pseudoknots remains a computationally intensive challenge. Over the past two decades, tools such as PKnots and pKiss, utilizing thermodynamic free energy minimization, have been pivotal in predicting RNA structures, particularly those involving pseudoknots. Recent advancements, like the application of deep learning technologies in tools such as SPOT-RNA and UFold, showcase the ongoing progress in this field. Our current work involves developing an assessment scheme for evaluating four distinct RNA sequence secondary structure prediction methods that possess the capability to predict pseudoknots. The predictions for the 398 known pseudoknot structures in the PseudoBase++ (pseudobaseplusplus.utep.edu) database demonstrate that PKnots and pKiss achieve higher prediction accuracies, whereas SPOT-RNA and UFold exhibit superior runtime efficiency and accommodate longer sequence lengths. These results are consistent with predictions made for an experimentally confirmed pseudoknot from SARS-CoV-2. This pseudoknot is 78 nucleotides in length, situated within the ORF1ab gene, which is expressed as a result of frameshifting.

Title

# The effect of interaction of heptadecafluorooctanesulfonic acid solution with β-lactoglobulin (LG), hemoglobin (Hb) and myoglobin (Mb): Influence on protein structure and function stability

Authors

Lucky Kofi Gbeda[1] and Mahesh Narayan[1,2]

[1]Bioinformatics Program and [2]Department of Chemistry and Biochemistry, The University of Texas at El Paso, El Paso, TX

Abstract

Heptadecafluorooctanesulfonic acid ($C_8F_{17}KO_3S$) is also known as Perfluorooctanesulfonic acid (PFOS). PFOS is a group of perfluoroalkyl sulfonic acids, which is among the most prominent xenobiotics contaminants in human tissues. The study aims at understanding the effects of $C_8F_{17}KO_3S$ on the structural conformation, stability, and physiological functions of β-Lactoglobulin (LG), hemoglobin (Hb) and Myoglobin (Mb) and the implications in relation to disease emergence. The study is focused on using UV–vis, and fluorescence spectroscopic methods, circular dichroism (CD), and molecular modeling to understand how $C_8F_{17}KO_3S$ affects the stability and the conformation of changes in hemoglobin. Currently, preliminary work has been done on the binding of 8-anilinonaphthalene-1-sulfonic acid (ANS) to LG indicated that the secondary structure of LG has been affected. This gave a greenlight to further investigate the binding of $C_8F_{17}KO_3S$ to LG, Hb and Mb, which may reveal significant mechanist explanation for the longer biological half-life of $C_8F_{17}KO_3S$ in human tissues. Through this study, useful information associated with the toxicity of $C_8F_{17}KO_3S$ will be revealed. Studying the effects of $C_8F_{17}KO_3S$ binding to LG, Hb and Mb could provide insight into emergence of many diseases and putative approaches to be employed to solve these challenges concerning human health. The binding of $C_8F_{17}KO_3S$ to these proteins could alter the structure, physicochemical, and functional properties of these proteins, resulting in complex medical conditions. Hence, it is important to understand the effects of $C_8F_{17}KO_3S$, so that proper preventive and corrective measures can be developed.

Title

Detection of early-introduced RNA post-transcriptional modifications of an early-stage large subunit ribosomal intermediate

Authors

Luis A. Gracia Mazuca,[1] Gyan Narayan,[2] Samuel S. Cho,[3,4] Jonathon E. Mohl,[1,5] and Eda Koculi[2]

[1]Bioinformatics Program, The University of Texas at El Paso, El Paso, TX 79968, USA;

[2]Department of Chemistry and Biochemistry, The University of Texas at El Paso, El Paso, TX 79968, USA;

[3]Department of Physics, Wake Forest University, Winston-Salem, NC 27109, USA;

[4]Department of Computer Science, Wake Forest University, Winston-Salem, NC 27109, USA;

[5]Department of Mathematical Sciences, The University of Texas at El Paso, El Paso, TX 79968, USA.

Abstract

Protein production by ribosomes is fundamental to life. The ribosome requires numerous modifications to be assembled correctly. The ribosomal RNA (rRNA), which is post-transcriptionally modified, provides one part of ribosome assembly complex. Thus, a complete understanding of ribosome assembly requires the determination of the RNA post-transcriptional modifications in all the ribosome assembly intermediates. There are 26 RNA post-transcriptional modifications in 23S rRNA of the mature Escherichia coli (*E. coli*) large ribosomal subunit. The levels of these modifications have been investigated extensively only for a small number of large subunit intermediates and under a limited number of cellular and environmental conditions. The 27S intermediate is one of three large subunit intermediates accumulated in *E. coli* cells lacking the DEAD-box RNA helicase DbpA and expressing the helicase inactive R331A DbpA construct. In this study we determined the level of incorporations of 2-methyl adenosine, 3-methyl pseudouridine, 5-hydroxycytosine and seven pseudouridines in an early-stage *E. coli* large subunit assembly intermediate with a sedimentation coefficient of 27S. Through extensive bioinformatics analysis employing Illumina next-generation sequencing and tools such as ShapeMapper, modifications on the 27S intermediate were identified. Utilizing discoveries from earlier intermediates the current list of modifications, the known modified sites, at positions 1915, 1919, and 1921, were concluded to not be present in the 27S intermediate and to be incorporated at the later stages of ribosome assembly.

Title

Single nucleotide variant data in patients with thyroid cancer

Authors

Purvi Kanfade,[1] Amanda Bataycan,[2] and Ming-Ying Leung[1,2,3]

[1]Bioinformatics Program, [2]Computational Science Program, and [3]Department of Mathematical Sciences, The University of Texas at El Paso, El Paso, TX

Abstract

According to the American Cancer Society, over 40,000 new cases of thyroid cancer (TC) are diagnosed each year, causing over 2,000 deaths. Although certain specific DNA mutations have been associated with TC, the pathogenesis pathways are not fully understood. The goal of this project is to organize a set of genomic single nucleotide variant (SNV) data obtained from patients with (TC) for downstream bioinformatics analysis to help provide insights into the molecular mechanisms of the disease.

Individual variant call format (VCF) data files with SNV information from both tumor and normal tissue samples of 497 patients with TC were downloaded from The Cancer Genome Atlas (TCGA) resources. These files were converted to more readable CSV files and the SNV information within were subsequently merged into one single data file listing all the distinct SNVs. By integrating essential information from RefFlat gene annotations into our computational workflow, we further organized the SNVs by their genomic locations. While all 497 tumor samples provided useful SNV information, only 477 normal samples yielded reliable data after quality filtering. The resulting dataset contained 98,952 distinct SNVs, of which 95,627 occurred only in tumor samples, 7 only in normal samples, and 3325 in both. The SNVs in tumor samples involved 20,308 unique genes with 21,703 possible transcripts, while those in normal samples involved 2,672 unique genes with 2,718 possible transcripts. Our next step will be to evaluate of the functional effects of the SNVs with the aim of finding potential TC-associated genes.

Title

Processing rotifer sequences for phylogenetic analysis and unraveling evolutionary relationships

Authors

Sri Varsha Kodiparthi,[1] Ratna Bolem,[1] Elizabeth J Walsh,[1,2] and Jonathon E Mohl[1,3]

[1]Bioinformatics Program, [2]Department of Biological Sciences, and [3]Department of Mathematical Sciences, The University of Texas at El Paso, El Paso, TX

Abstract

Rotifers make up a phylum of microscopic and near-microscopic invertebrate animals. These aquatic organisms have long captivated researchers with their unique biology and ecological significance. In recent times, advancements in the field of molecular biology have opened new avenues for gaining a better understanding of the evolutionary past of these animals, but rotifers still lack resolved phylogenies at most taxonomic levels. To reconstruct the evolutionary relationships among rotifer species, genetic biomarkers are used. To isolate these biomarkers DNA extraction and sequencing are used. High-throughput sequencing has revolutionized the field by providing extensive genetic data from various rotifer species, primarily targeting genetic markers such as nuclear 18S rRNA and internal transcribed spacer regions and mitochondrial COI. Alignment algorithms are then employed to align sequences, facilitating the comparison of genetic variation and shared ancestry among species. Phylogenetic trees are constructed using various inference methods, such as maximum likelihood and Bayesian analysis to estimate the evolutionary relationships among rotifers. The resulting phylogenetic trees offer insights into the evolutionary history, biogeography, and diversification of rotifers, shedding light on their adaptation to various environments and ecological roles. Here we developed a simple Python script to facilitate a rigorous quality control process, including trimming, filtering, and error correction. This script was used in processing the 146 sequencing results to ensure reliable downstream results.

Title

# Examining genomic data to understand functional effects of genetic variants in ovarian cancer

Authors

Omodolapo Nurudeen,[1] Amanda Bataycan,[2] and Ming-Ying Leung[1,2,3]

[1]Bioinformatics Program, [2] Computational Science Program, and [3]Department of Mathematical Sciences, The University of Texas at El Paso, El Paso, TX

Abstract

Cancer is a global health priority, with ovarian tumors being the fifth in cancer deaths among women. Ovarian cancer poses serious dangers to patients' well-being, with the year relative survival rate at 50.8.%. The goal of this research is to identify yet unreported possible ovarian cancer-related single nucleotide variants (SNVs) for downstream bioinformatics and experimental investigations to gain better understanding of the molecular underpinnings of this disease. We developed custom Python codes to convert variant call format (VCF) files obtained from the Genomic Data Commons Portal into more readable CSV files that provided critical information of the SNVs, including their genomic locations and associated transcript IDs as well as the reference and mutated bases. After processing and cleaning the CSV files from 426 patient records, we compiled 213,894 distinct variants that occur in tumor samples only, 78 in normal samples only, and 8,858 common to both. Next, we used the FATHMM-XF web server, which is based on a supervised machine learning approach, to score the SNVs on a scale of 0 to 1, with 1 being most pathogenic. Out of all the SNVs, 52,802 were found within coding regions, and 17,909 of them were classified as pathogenic with a score of above 0.5. In the near future, we will incorporate other functional effect assessment tools such as SNPnexus and PROVEAN with FATHMM-XF to establish a scoring scheme for evaluating the cumulative impacts of SNVs on genes, with the ultimate aim of revealing novel gene targets for cancer therapy.

Title

# Using a SEIR model to develop bivalent booster allocation strategies against emerging SARS-CoV-2 variants in US cities with large Hispanic communities

Authors

Francis Owusu Dampare[1] and Anass Bouchnita[1,2]

[1]Bioinformatics Program and [2]Department of Mathematical Sciences, The University of Texas at El Paso, El Paso, TX

Abstract

COVID-19 is a disease that disproportionately impacts the Hispanic population, due to the prevalence of certain risk factors and the high number of essential workers in this community. In this work, we analyze the vaccination strategies that would minimize the COVID-19 health disparities in El Paso County, TX, in the context of the emergence of a new highly transmissible and immune-escaping SARS-CoV-2 variant. We stratify an age-structure stochastic SEIR model that tracks the evolution of immunity derived from infections and vaccination according to Hispanic vs non-Hispanic ethnicity and parameterize it to the demographic, health and immunization data of El Paso County, TX. We did this using curated data from CDC; modifying the relevant variables, and making stochastic projections in Python. After fitting the model, the results show that increasing vaccination with bivalent boosters by five-fold in anticipation of highly transmissible and immune escaping variants would decrease the cumulative hospital admissions and mortality from Mar 1, 2023, to Dec 31, 2023, by 62.72% and 61.41%, respectively. Additionally, allocating 50% of the doses administered to non-Hispanic individuals to the Hispanic community would eliminate the disparities in hospitalizations. Our findings can guide public health officials in US cities with large Hispanic communities and help them design vaccination strategies that minimize COVID-19 health disparities caused by emerging variants using specific vaccination strategies.

Title

# A synergistic approach to high-throughput detection of RNA modifications using direct RNA sequencing and algorithmic analysis

Authors

Salvador Rodarte,[1] Luis Gracia Mazuca,[1] Isaac Weislow,[2] Davor Beltran,[2] Jonathon E. Mohl,[1,3] and Eda Koculi[2]

[1]Bioinformatics Program, [2]Department of Chemistry and Biochemistry, and [3]Department of Mathematical Sciences, The University of Texas at El Paso, El Paso, TX, 79968

Abstract

RNA post-transcriptional modifications play critical roles in regulating various cellular processes, and when aberrant, can lead to severe health conditions such as cancer and neurodegenerative diseases. Of roughly 300 identified classes of RNA modifications, there remains a significant demand for methods that combine high throughput and single nucleotide resolution detection. In this research, the potential of the RNA direct-sequencing platform developed by Oxford Nanopore Technologies was harnessed, focusing on bacterial ribosomal RNA as a model molecule. While efforts led to the pinpointing of several classes of RNA post-transcriptional modifications, including pseudouridine (Ψ) and numerous nucleotide methylations, certain modifications like 5-hydroxycytosine (5HC) and others displaying consistent mutation rates between control and sample sets remain elusive. To enhance the method's robustness and accuracy, the capabilities of two pre-existing tools, Nanopore-PSU and EpiNano, were incorporated. Nanopore-PSU, developed for Ψ site prediction, exploits native content training and machine learning modeling to map Ψ modifications across varied transcriptomes. Simultaneously, EpiNano in its EpiNano-Error mode was utilized to derive crucial features from RNA sequencing reads—such as current intensity and mismatch frequency—that boost the reliability and accuracy of RNA modification detection. This approach, which integrates custom algorithms and external tools, not only augments comprehension of RNA modifications but also underscores the efficacy of employing diverse analytical methods to progress the domain of post-transcriptional RNA modification detection.

Title

Prediction of thrombin generation thresholds for coagulation initiation under flow using machine learning surrogate models

Authors

Anass Bouchnita,[1,2] Kanishk Yadav,*[2,] Jean-Pierre Llored,[3] Alvaro Gurovich,[4] and Vitaly Volpert[5,6]

[1]Department of Mathematical Sciences, The University of Texas at El Paso, El Paso, TX, USA;
[2]Bioinformatics Program, The University of Texas at El Paso, El Paso, TX, USA;
[3]Ecole Centrale Casblanca, Ville Verte Bouskoura, Casablanca, Morocco;
[4]Department of Physical Therapy and Movement Sciences, The University of Texas at El Paso, El Paso, TX, USA;
[5]Institut Camille Jordan, University Lyon, France;
[6]S.M. Nikolsky Mathematical Institute, Peoples Friendship University of Russia (RUDN University), Moscow, Russia.

Abstract

In veins, clotting initiation displays a threshold response to flow intensity and injury size. Mathematical models can provide insights into the patient-specific conditions leading to clot growth initiation under flow. However, it is hard to determine the thrombin generation curves that favor coagulation initiation in a fast manner, especially when considering a wide range of conditions related to flow and injury size. In this work, we propose to address this challenge by using a neural network model trained with the numerical simulations of a validated 2D model for clot formation. Our surrogate model approximates the results of the 2D simulations, reaching an accuracy of 94% on the test dataset. Our study tackles the challenge of identifying the conditions initiating clot growth in veins, influenced by flow intensity and injury size. We employ artificial intelligence, particularly deep neural networks, to streamline the assessment of thrombin generation curves favoring coagulation initiation, even under diverse flow and injury conditions. Various machine learning algorithms were explored, with deep neural networks, support vector machines, and decision trees consistently achieving accuracies exceeding 90%, while higher-level boosting algorithms like XGBoost and CatBoost reaching above 95%. Deep neural networks were preferred due to their proficiency with high-dimensional data, anticipating future research expansions. In summary, our innovative approach utilizing artificial neural networks serves as a proof-of-concept tool for estimating bleeding risk in patients based on their Thrombin Generation Assay results. This promises to enhance our understanding of clot formation under various conditions, revolutionizing clinical assessments and treatments.

*****Presenting Author