THE 13TH INTERNATIONAL SYMPOSIUM ON HEALTH INFORMATICS AND BIOINFORMATICS

# HIBIT 2020

22 - 23 OCTOBER 2020

# The 13th International Symposium on Health Informatics and Bioinformatics

22-23 October 2020
Virtual

# Contents

# Committees

## Conference Co-chairs

**Ogün Adebali** Sabancı University
**Öznur Taştan** Sabancı University

## Organization Committee

**Ercüment Çiçek** Bilkent University
**Hilal Kazan** Antalya Bilim University
**Ogün Adebali** Sabancı University
**Öznur Taştan** Sabancı University
**Stuart James Lucas** Sabancı University

## Program Committee

**Abdullah Kahraman** University Hospital Zurich
**Ali Çakmak** Istanbul Şehir University
**Alper Kucukural** University of Massachusetts Medical School
**Arzucan Özgur** Bogazici University
**Aslı Suner** Ege University
**Athanasia Pavlopoulou** Izmir Biomedicine and Genome Center
**Atilla Gürsoy** Koç University
**Aybar Acar** Middle East Technical University
**Barış Süzek** Muğla Sıtkı Koçman University
**Burcu Bakır-Güngör** Abdullah Gül University
**Burçak Otlu** UC San Diego
**Can Alkan** Bilkent University
**Cesim Erten** Antalya University
**Cigdem Gündüz-Demir** Bilkent University
**Emre Güney** Pompeu Fabra University
**Evren Koban** Ege University
**Ezgi Karaca** Dokuz Eylul University
**Ferhat Ay** La Jolla Institute
**Gökhan Karakülah** Dokuz Eylul University
**Hilal Kazan** Antalya University

## Poster Award Committee

**Nurcan Tunçbağ** Middle East Technical University
**Stuart J. Lucas** Sabancı University Nanotechnology Research & Application Center
**Tunca Doğan** Hacettepe University
**Volkan Atalay** Middle East Technical University
**Yavuz Oktay** Izmir Biomedicine Genome Center
**Özlen Konu** Bilkent University

## Student Organization Committee

**Afshan Nabi** Sabancı University
**Asuman Inan** Sabancı University
**Begüm Özemek Güner** Max Planck Institute for Molecular Genetics
**Berk Turhan** Sabancı University
**Berkay Selçuk** Sabancı University
**Burak Islek** Sabancı University
**Defne Çirci** Sabancı University
**Dilek Koptekin** Middle East Technical University
**Elif Öz** Acıbadem University
**Halil Tuvan Gezer** Sabancı University
**Halil İbrahim Kuru** Bilkent University
**İbrahim Berber** Antalya Bilim University
**İlayda Beyreli** Bilkent University
**Muhammet Edip Akay** Antalya Bilim University
**Oya Işılay Canik** Sabancı University
**Rana Kalkan** Sabancı University
**Sevilay Güleşen** Kadir Has University
**Yasin Kaya** Hacettepe University
**Zeynep Kılınç** Sabancı University

# Speakers

## Keynote Speakers

**Sunduz Keles** University of Wisconsin-Madison
**Janet Thornton** EMBL-EBI
**Igor Jouline** Ohio State University

## EMBO YIP Lecturers

**Ana Cvejic** University of Cambridge
**Fran Supek** IRB Barcelona

## Invited Speakers

**Elif Nur Firat Karalar** Koc University
**Ezgi Kaya** IBG & Dokuz Eylul University
**Ugur Sezerman** Acibadem University

# Welcome Address

The International Symposium on Health Informatics and Bioinformatics (HIBIT) is now in its thirteenth year. The symposium aims to bring together academics, researchers and practitioners who work in these popular and fulfilling areas and to create the much-needed synergy among medical, biological and information technology sectors. HIBIT is one of the few conferences emphasizing such synergy. HIBIT provides a forum for discussion, exploration and development of both theoretical and practical aspects of health informatics and bioinformatics and a chance to follow current research in this area by networking with other bioinformaticians.

This year, we expect 720 attendees from all over the world. The 4 days of solid science will include 3 keynote speakers, 5 invited speakers, 14 selected talks and 91 posters.

Welcome to HIBIT 2020!

**The Organizing Committee**
Ercüment Çiçek
Hilal Kazan
Ogün Adebali
Öznur Taştan
Stuart James Lucas

# Abstracts

## Characterizing the influence of genetic variation on the ESC proteome

Selcan Aydin[1], Duy Pham[1], Tian Zhang[2], Daniel A. Skelly[1], Matthew Pankratz[3], Devin K. Porter[3], Ted Choi[3], Steven Gygi[2], Laura G. Reinholdt[1], Christopher L. Baker[1], Gary A. Churchill[1] and Steven C. Munger[1]

[1]The Jackson Laboratory, Bar Harbor, ME 04609
[2]Harvard Medical School, Boston MA 02115
[3]Predictive Biology, Carlsbad, CA 92009

Genetic background is known to affect pluripotency in embryonic stem cells (ESCs) but most studies to date have been conducted on a limited number of cell lines. We recently performed a genetic analysis of gene expression and chromatin accessibility in a large panel of ESCs derived from genetically heterogeneous Diversity Outbred (DO) mice, and identified thousands of loci that were associated with differences in chromatin state and transcript abundance among DO ESC lines. Here, we are further integrating the genetic analysis of the ESC proteome to extend our investigation into the role of genetic variation on pluripotency. We used multiplexed proteomics to measure global protein abundance in each DO ESC line, and compared protein and transcript abundance across genes and lines. Overall, protein abundance was highly variable across cell lines, similar to our observation in the transcriptome. We identified genetic background and sex as major drivers of this variation. We integrated genotyping data to our proteomic dataset, and mapped thousands of quantitative trait loci that affect protein levels (pQTL). Integrated analysis of eQTL and pQTL analysis revealed two sets of genomic loci, where the loci in the first one influence both transcript and protein abundance, whereas the second set only impacts one. For the majority of genomic loci in the first set, the variation affects transcript and protein levels identically, meaning if a variant is associated with higher/lower expression of a gene, it also leads to more/less protein. We further validated Lifr expression as the causal regulator of trans-variation on chromosome 15, where identical genetic effects in protein and transcript abundance are observed for many pluripotency related genes. The second set of genomic loci mainly consists of variants that affect post-transcriptional regulatory mechanisms. Of note, we discovered a largely unknown and underappreciated class of variants by the stem cell field previous to this work which could not be identified by analyzing either the transcript or protein data alone. Moreover, mediation analysis identified Nedd4 protein abundance as the top candidate regulator for a pQTL-specific hotspot on Chr 9 that controlled the abundance of many protein with roles in trans-

lation. Future efforts will experimentally validate other candidate regulators and better define their specific roles in pluripotency maintenance. This analysis combining measurements across molecular levels from a large number of genetically diverse ESCs will allow us to improve our understanding of the robust regulatory circuitry governing pluripotency and differentiation, and characterize the effects of genetic variation on these critical cellular processes.

---

# A novel Python bioinformatics pipeline for sequence error correction for metabarcoding with Nanopore sequencing

Bilgenur Baloglu[1], Zhewei Chen[2], Vasco Elbrecht[3], Thomas Braukmann[1], Shanna MacDonald[1] and Dirk Steinke[1]

[1]Center for Biodiversity Genomics, University of Guelph, ON, N1G2W1, Canada
[2]Caltech, Pasadena, CA 91125
[3]Zoological Research Museum Alexander Koenig, Bonn, Germany 53113

Metabarcoding (identification of the plant, animal, and fungal taxa present in an environmental sample) rapidly gains importance in ecology, food safety, pest identification, and disease surveillance. It has a compelling advantage over traditional approaches for obtaining data on species distributions, however, it is often difficult to detect all the species present in a bulk sample using High-throughput Sequencing (HTS). This can – in parts – be attributed to the shorter read lengths most HTS instruments generate. Moreover, most HTS platforms are not portable, making in situ field-based sequencing not feasible. Oxford Nanopore sequencing platforms such as the MinION represent an exception to that and they are also known to provide longer reads albeit limited by rather high error rates ( 12-15%).

We used a freshwater mock community of 50 Operational Taxonomic Units (OTU) to test the capacity of the Oxford Nanopore MinION coupled with a rolling circle amplification protocol to provide long read metabarcoding results. We also propose a new Python pipeline that explores error profiles of nanopore consensus sequences, mapping accuracy, and overall community representation within a complex bulk sample. Using our molecular and bioinformatics workflow, we were able to estimate the diversity of the tested freshwater mock community with an average sequence accuracy of ¿99% for 1D2 sequencing on the nanopore platform. We also showed that the high error rates associated with long-read single molecule sequencing can be mitigated by using a rolling circle amplification protocol. Future bioassessment programs will tremendously benefit from such portable, highly accurate, species-level metabarcoding and it appears that we reached a point were cost-effective field-based DNA metabarcoding is possible.

---

# Automated post-processing to improve public health related tweet tetection

Emine Ela Küçük[1], Doğan Küçük[2], Nursal Arıcı[2] and Erkut Küçük[3]

[1]Department of Nursing, Giresun University, Giresun, Turkey 28340

[2]Department of Computer Engineering, Gazi University, Ankara, Turkey 06570

[3]Faculty of Medicine, Niğde Ömer Halisdemir University, Niğde, Turkey 51000

Public health surveillance on the Web is a significant research topic in public health [1]. Especially after the recent global Covid-19 outbreak, considerable research effort is being allocated to mining Web resources in order to obtain verified and useful information and news regarding this pandemic. Related literature on public health surveillance on the Web includes studies that utilize news portals, Web search engine query logs and social media posts to track incidents related to public health, among others. For instance in [2], the authors make use of a public health ontology to automatically detect public health related tweets within a tweet stream. Experiments are carried out on two distinct datasets of 1 million tweets each, and the detection algorithm marks 1,500 tweets from the first set and 1,455 tweets from the second set as related to public health. The proposed detection algorithm based on the existence of ontology concepts/terms in the tweets achieves an accuracy of 73.5% on the first dataset and an accuracy of 69.4% on the second dataset [2]. In the current study, we have automatically post-processed the retrieved (i.e., automatically marked) tweets from these two datasets to further improve the performance of public health related tweet detection. On both of the retrieved tweet sets, we have filtered out tweets including: (1) hashtags (those tokens beginning with #), (2) mentions (those tokens beginning with ), and (3) external links (those tokens beginning with http). After these filtering procedures, the accuracies on the retrieved tweet set from the first dataset are (1) 74.8% after filtering tweets with hashtags, (2) 75.1% after filtering tweets with mentions, and (3) 71% after filtering tweets with external links. The accuracies on the retrieved tweets from the second dataset are (1) 69.7% after filtering tweets with hashtags, (2) 71.4% after filtering tweets with mentions, and (3) 69.3% after filtering tweets with external links.

The accuracies obtained after automated post-processing to filter hashtags, mentions, and external links in tweets show that:

- Filtering out those tweets including mentions improves the performance of public health related tweet detection around 1-2% consistently for both datasets. Hence, this result suggests that a post-processing scheme based on mentions is a viable candidate filtering scheme for improved public health surveillance on social media.

- Filtering out the tweets including hashtags leads to very slight increases in accuracies. Therefore, further experiments are required to draw conclusions regarding filtering tweets with hashtags on the performance of public health related tweet detection.

- Filtering out the tweets with external links leads to decreases in accuracies. Hence, our study suggests that this filtering procedure may be avoided in automated public health surveillance systems on social media.

## References


1. Edo-Osagie, O., De La Iglesia, B., Lake, I., & Edeghere, O. (2020). A scoping review of the use of Twitter for public health research. Computers in Biology and Medicine, 103770.

2. Küçük, E. E., Yapar, K., Küçük, D., & Küçük, D. (2017). Ontology-based automatic identification of public health-related Turkish tweets. Computers in Biology and Medicine. 83, syf. 1-9.


---

# Automatic rumour detection and fact checking for enhanced text-based epidemic intelligence

Erkut Küçük[1], Emine Ela Küçük[2] and Dilek Küçük[3]

[1]Faculty of Medicine, Niğde Ömer Halisdemir University, Niğde, Turkey 51000
[2]Department of Nursing, Giresun University, Giresun, Turkey 28340
[3]TUBITAK Energy Institute, Kocaeli, Turkey 41400

Epidemic intelligence systems are usually defined as early information and warning systems for disease outbreaks in order to decrease prospective public health problems in a society [1]. Text-based epidemic intelligence systems analyze formal and informal textual data to determine disease outbreaks and other common public health issues. Yet, when particularly conducting related research on social media texts, differentiating verified information and misinformation is an important issue [2]. Automatic rumour classification [3] and automatic fact checking [4] are two recent and significant research problems in social media analysis and text mining. In this study, we point out open research topics on the use of automatic rumour classification and fact checking systems for enhanced text-based epidemic intelligence.

Particularly with the widespread use of social media, unverified information (regarding different domains including health) is reported to spread instantly [2]. There are several related terms utilized in this context including rumours, fake news, and misinformation. Misinformation is defined as false information that does not aim to be harmful [2] while rumour is defined as a piece of circulating information that has not yet been verified [3]. Fake news is similarly defined as fabricated information that seem like news content but lacks the required news characteristics [4]. Fake news detection has recently become a very significant research problem as there are related competitions such as Fake News Challenge that aim to solve this problem using artificial intelligence techniques [5]. Considered as a step towards alleviating the effects of these types of unverified information, automatic fact checking is defined as the task of assessing the truth of given statements [6].

It is reported in the literature that misinformation is common in social media particularly regarding vaccines and infectious diseases [2]. Detection and resolution of rumours and fake news are critical for many domains including public health and different high-performance machine learning and deep learning methods are being used to solve these problems. Therefore, such instances of misinformation as well as rumours in social media should be detected before analyzing the content of the social posts and using these analysis results during health-related decision-making.

Text-based epidemic intelligence systems are designed and developed to facilitate public health monitoring and hence to ensure timely employment of measures to protect public health. In order to achieve this objective, these systems should include modules to detect rumours and then apply rumour resolution and fact checking methods to filter out those undesirable information sources. Different artificial intelligence algorithms, and particularly deep learning algorithms can be employed and tested for these problems on health-related textual data. Thereby, the resulting enhanced epidemic intelligence systems are expected to produce more reliable and useful results for health professionals.

**References:**


1. Joshi, A., Karimi, S., Sparks, R., Paris, C., & MacIntyre, C. R. (2019). Survey of text-based epidemic intelligence: A computational linguistics perspective. ACM Computing Surveys, 52(6), 1-19.

2. Wang, Y., McKee, M., Torbica, A., & Stuckler, D. (2019). Systematic literature review on the spread of health-related misinformation on social media. Social Science & Medicine, 240, 112552.

3. Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M., & Procter, R. (2018). Detection and resolution of rumours in social media: A survey. ACM Computing Surveys, 51(2), 1-36.

4. Vlachos, A., & Riedel, S. (2014). Fact checking: Task definition and dataset construction. ACL Workshop on Language Technologies and Computational Social Science (pp. 18-22).

5. Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., et al. (2018). The science of fake news. Science, 359(6380), 1094-1096.

6. Fake news challenge (2017) http://www.fakenewschallenge.org/


---

# Network-based discovery of Molecular targeted agent treatments in hepatocellular carcinoma

Rumeysa Fayetörbay, Nurcan Tunçbağ and Rengül Atalay
Middle East Technical University, Ankara, Turkey 06800

Hepatocellular carcinoma (HCC) is one of the most-deadly cancers and the most common type of primary liver cancer. Multikinase inhibitor Sorafenib is one of FDA approved targeted agents in HCC treatment. PI3K/AKT/mTOR pathway is altered in about 51% of HCC; hence, understanding how Sorafenib and PI3K/AKT/mTOR pathway inhibitors act at signaling level is crucial for targeted therapies and to reveal the off-target effects. In this work, we use gene expression profiles (GEPs) of HCC cells (Huh7 and Mahlavu) which were treated with seven different agents and their combination. Our aim is to reveal the important targets and modulators in agent treatments by inferring the dysregulation of Interactome. In other words, we search for the mechanism of action of the agents in a network context beyond the list of genes. For this purpose, we use the DeMAND (Detecting Mechanism of Action based on Network Dysregulation) algorithm developed by Califano Lab. DeMAND compares GEPs and assesses the change in the individual interactions from weighted interactome obtained from STRING database. As a result, we reconstructed 18 agent-specific networks from each GEPs. Each gene and interaction within these networks have a value signifies how strongly these genes are affected from the chemical network perturbation. Then, we found enriched pathways in each network. We initially compared the networks of single agents and their combination; i.e. PI3Ki-$\alpha$, Sorafenib and their combined treatment. Then, we compared all networks simultaneously. The simultaneous comparison of the reconstructed networks at gene and pathway levels shows that several pathways and proteins are commonly affected across agent treatments (e.g., Wnt, HIF-1, Notch pathways and MCM proteins, mTOR). On the other hand, some pathways are only affected in a specific agent treatment (e.g., SNARE interactions).

---

## Potpourri: An epistasis test prioritization agorithm via diverse SNP selection

Gizem Caylak[1], Oznur Tastan[2] and A. Ercument Cicek[1]

[1]Department of Computer Engineering, Bilkent University, Ankara, Turkey 06800
[2]Department of Computer Engineering, Sabancı University, İstanbul, Turkey 34956

Genome-wide association studies explain only a fraction of the underlying heritability of genetic diseases. Investigating epistatic interactions between two or more loci help to close the missing heritability gap. Unfortunately, the sheer number of loci combinations to be tested pose computational and statistical challenges. The epistasis test prioritization algorithms rank SNP pairs and focus on the highest rank pairs to limit the number of tests. Yet, the currently available algorithms still suffer from very low precision due to multiple hypothesis test correction. That is, due to still high number of tests performed, the selected pairs do not pass the adjusted significance threshold. It was shown in the literature that selecting SNPs that are individually correlated with the phenotype and diverse with respect to genomic location, leads to better phenotype prediction due to genetic complementation. Here, we provide evidence that an algorithm that pairs SNPs from such diverse regions and ranks them can improve the prediction power. We propose an epistasis test prioritization algorithm that optimizes a submodular set function to select a diverse and complementary set of genomic regions that span the underlying genome. SNP pairs from these regions are then further ranked with respect to their co-coverage of the case cohort. As shown in Figure 1, we compare our algorithm with the state-of-the-art on three GWAS and show that (i) we substantially improve precision (from 0.003 to 0.652) while maintaining the significance of selected pairs, (ii) decrease the number of tests by 25 folds, and (iii) decrease the runtime by 4 folds. We also show that promoting SNPs from regulatory/coding regions improves the performance (up to 0.8). Potpourri is available at http:/ciceklab.cs.bilkent.edu.tr/potpourri. This paper has been published in the proceedings of RECOMB 2020. The full paper can be found at https://www.biorxiv.org/content/10.1101/830216v3. This project was supported by TUBITAK 3501 Career Grant 116E148 to AEC.

---

# Machine learning assisted intraoperative assessment of brain tumor margins using HRMAS NMR spectroscopy

Doruk Cakmakci[1], E. Onur Karakaslar[1], Elisa Ruhland[2], Marie-Pierre Chenard[2], Francois Proust[2], Martial Piotto[3], Izzie Jacques Namer[2] and A. Ercument Cicek[1]

[1]Department of Computer Engineering, Bilkent University, Ankara, Turkey 06800
[2]University Hospitals of Strasbourg, Strasbourg, France
[3]Bruker Biospin, Wissembourg, France

Complete resection of the tumor is important for survival in glioma patients [1]. Even if the gross total resection was achieved, left-over micro-scale tissue in the excision cavity risks recurrence. High Resolution Magic Angle Spinning Nuclear Magnetic Resonance (HRMAS NMR) technique can distinguish healthy and malign tissue efficiently using peak intensities of biomarker metabolites. The method is fast, sensitive and can work with small and unprocessed samples, which makes it a good fit for real-time analysis during surgery. However, only a targeted analysis for the existence of known tumor biomarkers can be made and this requires a technician with chemistry background, and a pathologist with knowledge on tumor metabolism to be present during surgery. Here, we show that we can accurately perform this analysis in real-time and can analyze the full spectrum in an untargeted fashion using machine learning. We propose a pipeline for tumor margin assessment during brain tumor surgery based on machine learning methods (see Figure 1). We work on a new and large HRMAS NMR dataset of glioma and control samples (n = 565), which are also labeled with a quantitative pathology analysis. Our results (see Figure 2) show that a random forest based approach can distinguish samples with tumor cells and controls accurately and effectively with a median AUC of 85.6% and AUPR of 93.4%. We also show that we can further distinguish benign and malignant samples with a median AUC of 87.1% and AUPR of 96.1%. We analyze the feature (peak) importance for classification to interpret the results of the classifier. We validate that known malignancy biomarkers such as creatine and 2-hydroxyglutarate play an important role in distinguishing tumor and normal cells and suggest new biomarker regions. The code is released at http://github.com/ciceklab/HRMAS_NC. This work is presented at RECOMB-CCB 2020 and is accepted for publication in PLoS Computational Biology. Full text of the paper is available at medRxiv:

https://www.medrxiv.org/content/10.1101/2020.02.24.20026955v1

---

# CSI NGS Portal: An online platform for automated NGS data analysis and sharing

Ömer An, Kar-Tong Tan, Ying Li, Jia Li, Chan-Shuo Wu, Bin Zhang, Polly Leilei Chen and Henry Yang

Cancer Science Institute of Singapore, National University of Singapore, Singapore 117599

Background: Next-generation sequencing (NGS) has been a widely-used technology in biomedical research for understanding the role of molecular genetics of cells in health and disease. A variety of computational tools have been developed to analyse the vastly growing NGS data, which often require bioinformatics skills, tedious work and significant amount of time. Methods: To facilitate data processing steps minding the gap between biologists and bioinformaticians, we developed CSI NGS Portal, an online platform which gathers established bioinformatics pipelines to provide fully automated NGS data analysis and sharing in a user-friendly website. Results: The portal currently provides 18 standard pipelines for analysing data from DNA, RNA, smallRNA, ChIP, RIP, 4C, SHAPE, circRNA, eCLIP, Bisulfite and scRNA sequencing, and is flexible to expand with new pipelines. The users can upload raw data in fastq format and submit jobs in a few clicks, and the results will be self-accessible via the portal to view/download/share in real-time. The output can be readily used as the final report or as input for other tools depending on the pipeline. Conclusions: Overall, CSI NGS Portal helps researchers rapidly analyse their NGS data and share results with colleagues without the aid of a bioinformatician. The portal is freely available at: https://csibioinfo.nus.edu.sg/csingsportal

---

# The effect of kinship in re-identification attacks against genomic data sharing beacons

Kerem Ayöz[1], Miray Aysen[1], Erman Ayday[1] and A. Ercument Cicek[1]

[1]Department of Computer Engineering, Bilkent University, Ankara, Turkey 06800
[2]Case Western Reserve University, Cleveland, OH 44106

Big data era in genomics promises a breakthrough in medicine, but sharing data in a private manner limits the pace of field. Widely accepted genomic data-sharing beacon protocol provides a standardized and secure interface for querying the genomic datasets. The data is only shared if the desired information (e.g., a certain variant) exists in the dataset. Various studies showed that beacons are vulnerable to re-identification (or membership inference) attacks. As beacons are generally associated with sensitive phenotype information, re-identification creates a significant risk for the participants. Unfortunately, proposed countermeasures against such attacks have failed to be effective, as they do not consider the utility of beacon protocol. In this study, for the first time, we analyze the mitigation effect of the kinship relationships among beacon participants against re-identification attacks. We argue that having multiple family members in a beacon can garble the information for attacks since a substantial number of variants are shared among kin-related people. Using family genomes from HapMap and synthetically-generated datasets, we show that having one of the parents of a victim in the beacon (Figure 1) causes (i) significant decrease in the power of attacks and (ii) substantial increase in the number of queries needed to confirm an individual's beacon membership. We also show how the protection effect attenuates when more distant relatives, such as grandparents are included alongside the victim. Furthermore, we quantify the utility loss due adding relatives and show that it is smaller compared to flipping based techniques. We define the utility as the proportion of the flipped beacon responses (due to the proposed mitigation technique) and we show that the proposed mitigation technique does not cause a significant decrease in utility (especially for SNPs with low MAF values). This work will appear in the proceedings of the 19th European Conference on Computational Biology (ECCB 2020) and the full paper is available at https://doi.org/10.1101/2020.01.30.926907. Research reported in this abstract was supported by NLM of the NIH under award number R01LM013429.

# Elucidating the roles of naturally occurring silent mutations in Polycystic Ovary Syndrome (PCOS)

Aslı Kutlu, Şuara Şahin, Dilara Gümüşgül, Banu Taktak Karaca and Hatice Kübra Turan

Biruni University, İstanbul, Turkey 34010

Polycystic ovary syndrome (PCOS) is a complex genetic disorder, which often causes infertility in women. Most of the PCOS-related genes are also linked with metabolic diseases, inflammatory responses, or certain types of cancer. GWAS studies connect single nucleotide polymorphisms (SNPs) with these diseases, emphasizing their critical role in understanding the impacts of SNPs that exchange nucleotides. However, their response to changes in protein conformation and corresponding functional characteristics remains ambiguous. The present study provides evidence that elucidates for the role of specific SNPs known as silent mutations by studying impacts on INSR (insulin receptor), FST (follistatin), and AR (androgen receptor) genes, all strongly associated with PCOS pathogenesis. Although silent mutations do not change the amino acid type, their effects on protein expression may cause disease progression. To assess the potential impacts of silent mutations, we first calculated the anticodon availabilities through RFMapp. To avoid the controversial issues regarding anti-codon stability for certain SNPs, MFE (minimum free energy) was calculated as a measure of the thermodynamic stability of newly formed mRNA structures to provide more information. For rs2059806 and rs1799817 located in the INSR gene, we reported depletion in anti-codon availabilities of newly existed codons. For rs6152 and rs2059806 cases, we calculated both increase and decrease in mRNA stabilities, which were noted as instabilities of the tRNA-mRNA complex. With the results obtained, we referred to the case of codon optimality for providing a plausible explanation between the existence of these SNPs and PCOS. In this study, the effects of silent mutations on complex PCOS diseases have been demonstrated. In future studies, MD simulation will be conducted with missense mutations existed in INSR, FT and AR proteins to constitute a theoretical basis for functional aspects of missense mutations.

---

# Revealing the structural impacts of point mutations on MeCP2 protein associated with Rett Syndrome via MD Simulations

Ahmet Melih Öten and Aslı Kutlu
Biruni University, İstanbul, Turkey 34010

Rett Syndrome (RTT) is a rare disease, which is seen in 1 in 30,000 people. RTT is a clinically postnatal neurodevelopmental disease, and it is characterized by the slowing and loss of neurodevelopmental abilities in early childhood. Most cases of RTT have been associated with the MBD domain of the MECP2 gene on the X chromosome. MBD domain belongs to the family of proteins that bind to the methylated cytosine nucleotides found on the CpG islands. Proteins of the methyl-CpG-binding domain (MBD) family are primary candidates for the readout of DNA methylation as they recruit chromatin remodelers, histone deacetylases and methylases to methylated DNA associated with gene repression.

The Arg106Trp and Thr158Met are two of the point mutations that cause RTT. In this project, we analyzed the structural effects of these two mutations via MD simulations. We built three simulative systems for the native and the mutated proteins and run them for 50 ns. The results were analyzed under seven different topics, and structural changes caused by mutations were observed. As a result of our analysis, we theoretically conclude that two mutations lead different structural changes to cause RTT. The Arg106Trp mutation causes the change in the structure of beta-sheets in the secondary structure of the protein, making this part of the protein looser and more flexible. On the other hand, the Thr158Met mutation causes the eight of the fourteen DNA-binding residues to move away from the DNA. Hence, this mutation leads a change in the DNA-binding affinity All of these analyzes were simulated and obtained on protein models with only the MBD domain. As a future perspective, firstly we will build the MeCP2 full-length protein model; then we will improve our simulations and analyzes by adding a few point mutations and truncated mutations to previous two mutations.

---

# PAMOGK: A pathway graph kernel based multi-omics approach for patient clustering

Yasin Tepeli[1], Ali Burak Ünal[2], Mustafa Furkan Akdemir[1] and Oznur Tastan[1]

[1]Department of Computer Engineering, Sabancı University, İstanbul, Turkey 34956

[2]Department of Computer Engineering, Bilkent University, Ankara, Turkey 06800

Accurate classification of patients into molecular subgroups is critical for the development of effective therapeutics and for deciphering what drives these subgroups to cancer. The availability of multi-omics data catalogs for large cohorts of cancer patients provides multiple views into the molecular biology of the tumors with unprecedented resolution. We develop PAMOGK (Pathway based Multi Omic Graph Kernel clustering) that integrates multi-omics patient data with existing biological knowledge on pathways. We develop a novel graph kernel that evaluates patient similarities based on a single molecular alteration type in the context of a pathway. To corroborate multiple views of patients evaluated by hundreds of pathways and molecular alteration combinations, we use multi-view kernel clustering. Applying PAMOGK to kidney renal clear cell carcinoma (KIRC) patients results in four clusters with significantly different survival times p-value = 1.24e-11).

When we compare PAMOGK to eight other state-of-the-art multi-omics clustering methods, PAMOGK consistently outperforms these in terms of its ability to partition KIRC patients into groups with different survival distributions. The discovered patient subgroups also differ with respect to other clinical parameters such as tumor stage and grade, and primary tumor and metastasis tumor spreads. The pathways identified as important are highly relevant to KIRC. PAMOGK is available at https://github.com/tastanlab/pamogk. This paper has been presented in the proceedings of RECOMB CCB 2020 and published in Oxford Bioinformatics journal. The full paper can be found at https://doi.org/10.1093/bioinformatics/btaa655.
This project was supported by TUBITAK under Grant #117E140.

---

# CEN-tools: An integrative platform to identify the 'contexts' of essential genes

Cansu Dincer[1], Sumana Sharma[2], Paula Weidemüller[1], Gavin J. Wright[3] and Evangelia Petsalaki[1]

[1]EMBL-EBI, Cambridgeshire, UK

[2]The MRC Weatherall Institute of Molecular Medicine, University of Oxford, Oxford, UK

[3]Wellcome Sanger Institute, University of Cambridge, Saffron Walden, UK

The advances in genome editing technologies and in particular large-scale CRISPR (clustered regularly interspaced short palindromic repeats) screens provide the possibility to interrogate differences in genetic networks in a variety of genetic backgrounds or perturbations. Identification of genes that are only essential in specific situations, called context-dependent essential genes, is a useful strategy for defining gene function and also developing novel therapeutic interventions. Here, we present 'CEN-tools' (Context-specific Essentiality Network-tools), a website and python package, in which users can identify context-dependent essential genes from large-scale CRISPR screens, and additionally pinpoints these contexts including tissue of origin, mutation profiles, and expression levels. The statistical associations between genes and given contexts are visualised with dependency networks called CENs. We also integrate the CENs with human interactome to infer the context-dependent essential cellular pathways which were rewired in cancer cells. Moreover, we have a python package called pyCEN which offers users to navigate interested contexts rather than pre-calculated ones with different data. By this work, we illustrate the utility of CEN-tools to define the dependency map.

---

# Pathogenic impact of transcript isoform switching in 1209 cancer samples covering 27 cancer types using an isoform-specific interaction network

Tülay Karakulak[1,2,3], Abdullah Kahraman[1,2,3], Damian Szklarczyk[1,2,3] and Christian von Mering[1,2,3]

[1]Swiss Institute of Bioinformatics, Lausanne, Switzerland
[2]University Hospital Zurich, Zurich, Switzerland
[3] University of Zurich, Zurich, Switzerland

Alternative splicing regulation is often disturbed in various cancers leading to cancer-specific switches in the Most Dominant Transcripts (cMDT). To understand how these switches drive oncogenesis, we have analyzed isoform-specific protein interaction disruptions in the Pan-Cancer Analysis of Whole Genomes (PCAWG) project. Our study identified large variations in the number of cMDT with the highest frequency in cancers of female reproductive organs. Surprisingly, in contrast to the mutational load, cancers arising from the same primary tissue showed similar numbers of cMDT. Some cMDT were found in almost all samples of a cancer type rendering them as ideal diagnostic biomarkers. Other cMDT tended to be located at densely populated protein network regions disrupting interactions next to pathogenic cancer gene products in enzyme signalling, protein translation, and RNA splicing pathways. The highlighted common and distinct patterns of alternative splicing deregulations constitute new avenues for novel therapeutic targets in the fight against cancer.

---

# Integrative analysis of DNA methylation and RNA-sequencing data for identifying diagnostic cancer markers

Ezgi Demir Karaman and Zerrin Işık

Department of Computer Engineering, Dokuz Eylül University, İzmir, Turkey 35220

Integration of transcriptome and methylome data can work as a powerful strategy for understanding cancer mechanisms and has attracted clinical interest to identify biomarkers for diagnosis, prognosis, and predictive purposes in various cancer types.This study aims to identify DNA methylation changes in distal regulatory regions, correlate these signatures with mRNA expression of nearby genes and eventually to find similar genomic regions in different cancer types.

We used data of four cancer types (BRCA, LUSC, COAD, KIRC) available in the GDC portal.The integrative methylation-expression analysis was performed using the ELMER package in R.We focused on distal enhancers, so we selected specific methylation probes that have more than +/- 2Kbp distance to known transcription start sites.Furthermore, the differential analysis was performed to identify the hypermethylated ($\geq 0.7$) and hypomethylated ($\leq 0.3$) CpGs among the tumor and normal samples of four cancer types. In order to identify target genes regulated by these distal regulatory elements, we analyzed mRNA expression data for 10 genes both upstream and downstream of each distal regulatory element; these 20 nearby genes constituted the candidate target genes.For each probe, we statistically determined whether any of the nearby genes are affected by probe's DNA methylation change; then we extracted significant probe-gene pairs.Finally, we analyzed these significant probe-gene pairs across four cancer types.

We identified common 47 and 17 DNA methylation-driven genes which are in vicinity of the hypomethylated and hypermethylated regions for all cancers.Some of these 47 genes are annotated with the mitotic nuclear division GO-BP; identified in breast, cervix, and liver carcinoma. On the other hand, some of 17 genes play roles in various phases of cell cycle; they are mostly related with breast carcinoma.We will continue integrative analysis of methylation-driven genes by introducing a network-based approach.

# MTPpilot: an interactive software for NGS results analysis for molecular tumor boards

Fabian Arnold [1] and Abdullah Kahraman [1,2]

[1]University Hospital Zurich, Zurich, Switzerland
[2] University of Zurich, Zurich, Switzerland

Information from Next Generation Sequencing (NGS) cancer panels have become a key corner-stone for today's clinical decision making in oncology. In this context, NGS sequencing results are often discussed at molecular tumor boards, where amongst other oncologists, pathologists, bioinformaticians and geneticists discuss these results to decide on the best treatment options for a patient. However, with the increasing size and complexity of NGS cancer panels, NGS results have become challenging to interpret, especially if presented merely in the form of a written report. Manual analysis of several mutations from a comprehensive NGS cancer panel are time consuming and incomplete.

To address these short-comings, we developed the software Molecular Tumor Profiling Pilot (MTPpilot), which provides automated annotation, linking and interactive visualization to support the interpretation of NGS results at molecular tumor boards.

## Transcriptome analysis of regenerating zebrafish brain unravels a key role for Canonical Wnt signaling at early wound healing stage

Gökhan Cucun, Yeliz Demirci, Yusuf Kaan Poyraz and Güneş Özhan
Izmir Biomedicine and Genome Center, İzmir, Turkey 35340

Regeneration ability is highly limited in the mammalian central nervous system. On the contrary, one of the teleost models zebrafish has enormous brain regeneration capacity among vertebrates. Taking the brain regeneration ability of adult zebrafish as an advantage, elucidation of molecular mechanisms of brain regeneration has a pivotal role in regenerative medicine research. In our previous study, we have shown that Canonical Wnt signaling is activated at a very early stage of regenerating brain. Currently in our study, we performed stab wound assay in the brain of adult zebrafish, separate injured and uninjured hemispheres of transgenic Wnt reporter zebrafish line to identify Canonical Wnt target genes by conducting bulk RNA-Seq. Our finding reveals that 119 novel genes positively regulated by the Canonical Wnt pathway at the early wound healing stage. Additionally, we measure 6 positively regulated genes to validate our RNA-Seq results.

---

# Integrated analysis of transcriptomic and proteomic data to understand the effect of aneuploidy on cancer genomes

Gökçe Senger and Martin Schaefer

European Institute of Oncology, Milan , Italy P.I. 08691440153

Aneuploidy, whole chromosomal or chromosome arm level changes, is a hallmark of human cancer cells, but its role in cancer still remains to be fully elucidated. In this work, we focus on developing an understanding of how cancer cells deal with the excess amount of expression at both transcriptome and proteome level induced by chromosome gains, and how the excess expression affects protein complex stoichiometry. For 298 tumor samples, for which we have aneuploidy, transcriptomic and proteomic data made available by TCGA and CPTAC consortia, we first identified cancer-type specific chromosomes that are altered at higher frequencies than would be expected by chance. Then we profiled transcriptomic changes in response to chromosome number changes. To our surprise, we found that a relatively small number of genes on the aneuploid chromosomes changed expression while many expression changes happened on other chromosomes. Those differentially expressed genes on other chromosomes often form complexes and, even more, are often in the same complexes as differentially expressed genes on aneuploid chromosomes. These observations are even more pronounced on proteome level. To further investigate the differential co-regulation between co-complex members, we calculated protein level correlations between proteins of aneuploid chromosomes and their partner proteins of other chromosomes. We found that proteins involved in a smaller number of complexes have stronger correlations with their partners, highlighting the importance of compensation for stoichiometric imbalance in protein complexes. Aggregation-prone complex members also show stronger expression correlations suggesting that proteotoxicity of unpaired complex members make this compensation necessary. Our ongoing efforts focus on deciphering the regulatory control of gene expression of complex members (both on transcriptome and proteome level) to understand the molecular mechanisms of cancer cell adaptation to aneuploidy.

---

# Analysis of a potential microRNA network that co-regulate autophagy and epithelial to mesenchymal transition under nutrient restriction

Aliye Ezgi Güleç[1,2], Hepşen Hazal Hüsnügil[1], Ilir Sheraj[1], Ayşe Elif Erson Bensan[1] and Sreeparna Banerjee[1]

[1]Middle East Technical University, Ankara, Turkey 06800
[2]Baskent University, Ankara, Turkey 06810

Autophagy is a catabolic process that involves the degradation of cytoplasmic materials and organelles primarily in response to nutrient stress. Increasing evidence suggests that in the context of chemotherapy-induced stress response, autophagy mainly functions as a cell survival mechanism. Epithelial to mesenchymal transition (EMT) is a highly complex sequence of events that includes loss of contact between adherent epithelial cells and enhanced motility, accompanied by resistance to chemotherapy drugs. EMT-related resistance is one of the major causes of treatment failure. A cross-talk between autophagy and EMT has been suggested; however, whether one activates or inhibits the other is highly context and tumor type dependent. In the current study, we hypothesized that under nutrient restriction, the cross talk between autophagy and EMT may be regulated by a network of microRNAs (miRNAs). Caco-2 and T84 colorectal cancer (CRC) cells were subjected to nutrient restriction for 48h in a pre-optimized medium containing low serum, low glucose and low L-glutamine. The changes in autophagy and EMT markers were analyzed with western blot and quantitative real-time PCR (qRT-PCR), respectively. Small RNA sequencing was carried out to determine the differential expression of miRNAs in nutrient-restricted cells compared to controls. With nutrient restriction, both cell lines underwent autophagy induction as indicated by an increase in LC3-II and Beclin-1 levels; however, autophagic flux was inhibited as indicated by enhanced levels of the p62 cargo protein. The levels of both epithelial markers (E-cadherin and Occludin) and mesenchymal markers (Vimentin and Snail-1) were increased in both cell lines under nutrient restriction compared to controls. This phenotype in the starved cells is suggestive of the induction of partial EMT, which is associated with enhanced stemness and drug resistance compared to complete EMT. Small RNA sequencing results showed that there were 38 significantly differentially expressed annotated miRNAs in the starved cells. Of those, 11 miRNAs were associated with autophagy, 14 miRNAs were associated with EMT and 10 miRNAs were common to both terms supporting the notion that many common miRNAs can regulate both autophagy and EMT. Our results shows a potential network of miRNAs that are differentially regulated in response to nutrient stress and involved in the regulation of autophagy and EMT. This miRNA-regulated interplay between autophagy and partial EMT may allow to unveil new targets and avenues for improved treatment of therapy-resistant tumors.

## Evolution of genetic diseases in Turkey

Mehmet Çetin, Şevval Aktürk and Mehmet Somel
Middle East Technical University, Ankara, Turkey 06800

Many relatively common genetic diseases are observed in Anatolia today, including but not limited to celiac diseases, Behçet's disease, and cystic fibrosis. Considering their impacts on health, their relatively high frequency is surprising. However, for the most part, it is not known why that is the case. We sought to answer this question by analyzing genomes from Anatolia from 10,000 years ago, and comparing them with a published sample of 16 individuals' genomes from the modern-day Turkish population. We compared the frequency of disease-associated SNPs between the two time points and found a statistically significant change in the frequency of 6 disease associated alleles. In order to understand the underlying causes for these changes, we compared them with neutral alleles and performed neutral evolution simulations using Slim 3. As a result, we conclude that these changes occurred through neutral processes and migration. We also note that our small sample size severely limits our power to detect possible selection events, and call for new genome sequencing initiatives in Turkey.

---

# Robust prediction of genetic mutation effects by homology analysis

Alperen Taciroglu[1], Yesim Aydin Son[1] and Ogun Adebali[2]

[1]Department of Health Informatics, Graduate School of Informatics, Middle East Technical University, Ankara, Turkey 06800

[2]Molecular Biology, Genetics and Bioengineering Program, Faculty of Engineering and Natural Sciences, Sabancı University, İstanbul, Turkey 34956

Paralogous genes occur during gene duplication events which a copy of a gene is copied to another location within DNA. Origination of the second copy of a gene lifts most of the evolutionary pressure on one of the duplicated pairs. Thus, one of the pairs gains an opportunity to develop new functionalities by accumulating mutations. Variations in these duplicated pairs are assessed by their property of being benign, deleterious by automated tools like SIFT or PolyPhen. However, these tools disregard paralogy information and yield incorrect results. Project is designed to find out about homology including those that are distant homologs to selected genes with the aim to identify and characterize genetic variation in a robust manner. With the project is still ongoing, until this point we worked on Human Obscurin gene, which is encoded into a large protein mostly expressed in striated muscle cells. Although it requires further functional evaluation, mutations within Obscurin sequence are documented to be related with disorders like cardiomyopathies and tumorigenesis. In order to further document these mutations, we performed phylogenetic analysis, scanned for functional domains on Obscurin protein sequence then sequence based clustered these domains. Cluster-wise multiple alignment of the domains of closely related proteins allowed us to reconstruct Obscurin protein to serve as a backbone for conservation analysis.

---

# Integration of machine learning and entropy analysis as a Post-GWAS analysis approach

Burcu Yaldız, Onur Erdogan, Cem İyigünand Yeşim Aydın Son

Department of Health Informatics, Graduate School of Informatics, Middle East Technical University, Ankara, Turkey 06800

Non-linear relationships between genotypes play an essential role in understanding the genetic interactions of complex disease traits. Various statistical approaches such as entropy-based methods have been suggested for revealing these non-additive interaction effects between variants in recent years. We proposed a new ensemble workflow that integrates the machine learning algorithms and entropy-based 3-way interaction information method for capturing the hidden patterns resulting from non-linear relationships between genotypes. As the case study Late-Onset Alzheimer's Disease (LOAD) after PLINK-RF-RF analysis of LOAD GWAS datasets were used for the discovery of early and differential diagnosis markers. For curating interacting SNP sets through 3WI purely in these datasets, an entropy-based test statistic is used to identify 3-way and 2-way interactions that show a significant difference between case and control groups. Then triplets that involve significant SNP pairs according to 2-way mutual information gain test statistics are eliminated. Acquired p-values are adjusted for multiple comparisons using Benjamini and Hochberg procedure. Selected triplets of SNPs that show a significant difference between case and control groups in terms of 3WII are proposed as candidate biomarkers for a genotyping based diagnosis of LOAD.

---

# A meta-analysis of gene expression data for pathway enrichment in food allergy research

Asuman İnan[1], Öznur Taştan[1] and Stuart J. Lucas[2]

[1]Department of Computer Engineering, Sabancı University, İstanbul, Turkey 34956
[2]Sabanci University Nanotechnology Research and Application Center, İstanbul, Turkey 34956

Food allergy is developed as a reaction of immune system to certain allergens, leading to mild to life-threatening responses which affect 5% of adults and 8% of children[1]. As there is no cure for food allergies, allergic individuals can only be encouraged to avoid any possible exposure to allergens including cross-contamination. So far, oral immunotherapy - controlled introduction of the allergen to increase the threshold for sensitization - has been the only approved therapy. Besides medical research, there are only a handful of studies including transcriptome analysis of patients with food allergies. To our knowledge, there are 7 different datasets in the literature and out of which only 4 are currently available for re-analysis. All these studies focus on egg and peanut allergies, as they are two of the most common food allergies, and comprise bulk RNAseq, single-cell RNAseq, and microarray datasets. In our project, we first identify differentially expressed genes (DEGs) using the Hisat2-StringTie-Ballgown pipeline on the available RNA-seq datasets. Due to variations in study designs and techniques used, we perform DEG analysis starting from the raw data, taking different parameters from each study into account and aim to obtain a standard output format to be used as an input for meta-analysis. Our results are expected to indicate genes that are consistently differentially regulated during food allergy, despite differences in allergens and study design. Eventually, by using and comparing available meta-analysis approaches and pathway analysis software, we will be able to identify common food allergy-related pathways as targets for future intervention. REFERENCES [1] S. H. Sicherer and H. A. Sampson, "Food allergy: A review and update on epidemiology, pathogenesis, diagnosis, prevention, and management," J. Allergy Clin. Immunol., vol. 141, no. 1, pp. 41–58, Jan. 2018, doi: 10.1016/j.jaci.2017.11.003.

# Inference Attacks Against Differentially-Private Query Results from Genomic Datasets Including Dependent Tuples

Nour Alserr[1], Erman Ayday[1,2] and Ozgur Ulusoy[1]

[1]Department of Computer Engineering, Bilkent University, Ankara, Turkey 06800
[2]Case Western Reserve University, Cleveland, OH 44106

Thanks to the fast-paced throughput sequencing technologies which result in a large-scale datasets and biobanks. The number of sequenced human genomes has been increasing at an exponential rate, and now we are at about 2.5 million sequenced genomes around the world. This is projected to reach 105 million and this number can reach a lot more in 2025, especially after the COVID 19 pandemic, where many countries decide to study the genomic data in a population scale. These rich troves of data can empower the scientific advances. However, according to the sensitive nature of the genetic information, sharing the genomic datasets which include sensitive genetic or medical information for individuals can be misused if it lands in the wrong hands. Hence, for the hope of sharing the genomic dataset to gain better understanding of the human genetics, differential privacy (DP) is one of the privacy concepts proposed for sharing the summary statistics of genomic datasets in a private manner. DP mechanism provides a rigorous mathematical foundation for preserving privacy, but it does not consider the dependency of the data tuples in the dataset, which is a common situation for genomic datasets due to the inherent correlations between genomes of family members. We show how kin relationships between individuals in a genomic dataset cause a significant reduction in the privacy guarantees of traditional DP-based mechanisms. We formulate this as an attribute inference attack and show the privacy loss using differentially-private results of minor allele frequency (MAF) and chi-square queries over two real-life genomic datasets. Our results show that using the results of differentially-private MAF queries and exploiting the dependency between tuples, an adversary can reveal up to 50% more sensitive information about the genome of a target (compared to original privacy guarantees of standard DP-based mechanisms), while differentially-privacy chi-square queries can reveal up to 40% more sensitive information. Furthermore, we show that these inferred genomic records (as a result of the attribute inference attack) can be utilized to perform successful membership inference attacks to other statistical genomic datasets (e.g., associated with a sensitive trait). Using a log-likelihood-ratio (LLR) test, our results also show that the inference power of the adversary can be significantly high in such an attack even by using inferred (and hence partially incorrect) genomes. This work presented at the 28th conference of Intelligent Systems for Molecular Biology (ISMB2020). The full paper is available at: https://doi.org/10.1093/bioinformatics/btaa475

---

## Optimizing pipeline combinations for cancer sequencing

Batuhan Kısakol[1], Sahin Sarihan[1], Mehmet Arif Ergun[1] and Mehmet Baysan[2]
[1]Marmara University, İstanbul, Turkey 34722
[2]Istanbul Şehir University, İstanbul, Turkey 34865

Mapping and variant calling are the two major phases in DNA-Sequencing analysis pipelines. The presence of different approaches and many algorithms in these steps makes benchmarking of different sequencing pipelines an active field of research. In this study, we tested the performance of 12 different pipelines (3 aligners: BWA, Bowtie2, and Novoalign with 4 variant callers: Mutect2, Strelka, Varscan, and Somaticsniper) individually and more importantly as combinations to determine the ideal analyses workflows for cancer sequencing projects in different experimental conditions. For test data, we used a dataset that we recently constructed which included 50 samples from a single patient [1]. This dataset had primary tumor samples, in vitro polyclone samples cultured from the primary tumor, in vivo polyclone samples obtained from mouse xenografts, and in vitro monoclone samples which were obtained from in vitro polyclone samples through isolation of single cells. The availability of these related samples from different experimental conditions provided us a unique opportunity to test sequencing pipelines at different heterogeneity levels. Mutations were declared v̈alidatedẅhen they were detected in two independent samples since the probability of false detection of a specific variant in two independent samples is extremely low. Our initial findings revealed that mapping and variant calling algorithms perform differently for different heterogeneity levels and most of the pipelines record good precision scores but suffer bad recall scores. Additional to the individual evaluation of pipelines, we also tested the performance of pipeline combinations. In these analyses, we observed that certain pipelines complement each other much better than others and display superior performance than individual pipelines. This suggests that adhering to a single pipeline is not optimal for cancer sequencing analysis and sample heterogeneity should be considered in sequencing workflow optimization.

---

# Scalable classification of organisms into a taxonomy using hierarchical supervised learners

Gihad Sohsah[1], Ali Reza Ibrahimzada[2], Huzeyfe Ayaz[2] and Ali Cakmak[3]
[1]BlackStone eIT, Seattle, WA
[2]Marmara University, İstanbul, Turkey 34722
[3]İstanbul Technical University, İstanbul, Turkey 34467

Accurately identifying organisms based on their partially available genetic material is an important task to explore the phylogenetic diversity in an environment. Specific fragments in the DNA sequence of a living organism have been defined as DNA barcodes and can be used as markers to identify species efficiently and effectively. The existing DNA barcode-based classification approaches suffer from three major issues: (i) most of them assume that the classification is done within a given taxonomic class and/or input sequences are pre-aligned, (ii) highly performing classifiers, such as SVM, cannot scale to large taxonomies due to high memory requirements, (iii) mutations and noise in input DNA sequences greatly reduce the taxonomic classification score. In order to address these issues, we propose a multi-level hierarchical classifier framework to automatically assign taxonomy labels to DNA sequences. We utilize an alignment-free approach called spectrum kernel method for feature extraction. We build a proof-of-concept hierarchical classifier with two levels, and evaluated it on real DNA sequence data from Barcode of Life Data Systems. We demonstrate that the proposed framework provides higher f1-score than regular classifiers. Besides, hierarchical framework scales better to large datasets enabling researchers to employ classifiers with high classification performance and high memory requirement on large datasets. Furthermore, we show that the proposed framework is more robust to mutations and noise in sequence data than the non-hierarchical classifiers.

---

## Systematic analysis of phosphorylation structure

Altuğ Kamacıoğlu[1], Nurhan Ozlu[1] and Nurcan Tuncbag[2]
[1]Koç University, İstanbul, Turkey 34450
[2]Department of Health Informatics, Graduate School of Informatics, Middle East Technical University, Ankara, Turkey 06800

Phosphorylation is an essential post-translational modification for the regulation of almost all cellular processes. Several phosphorylation-sites for diverse cellular mechanisms and their corresponding kinases and quantitative change in phosphorylation is revealed with widespread quantitative phosphoproteomics analyses and even though the structure of a single protein and its phosphorylation-sites are studied, no systematic analysis concerning the structure of whole phosphoproteomics has been performed. In this study, we focused on the structural mechanism of phosphorylation to detect the respective location of phospho-sites through relative solvent accessibility of the phospho-sites and their characteristic features based on their location. We build on the data from all phosphorylation regions in current databases and a selected paper which filter false positive phosphorylation via quality-control. We find that a certain part of phosphorylation-sites locates in core part of protein with extremely low solvent accessibility and we observed that core phosphorylation-sites are highly found in false-positive phosphorylation-sites in databases. Core phosphorylation-sites are significantly less functional and more rigid than other type of phosphorylation. We found out that some of core phosphorylation-sites are very dynamic and highly functional. Lastly, we performed same analysis in Karayel et al. paper which include phosphorylation regulation in cell division, and almost all core phosphorylation regulated throughout cell division are detected as dynamic.

---

# Evolutionary history of complex human traits in the light of time-serial genetic data

Dilek Koptekin[1], H. Melike Donertas[2], Can Kosukcu[3], Idil Yet[3], Erdem Karabulut[3], Ismail Kudret Saglam[4], Anders Gotherstrom[5], Mehmet Somel[1], Fusun Ozer[3], Yilmaz Selim Erdal[3] and Gulsah Merve Kilinc[3]

[1]Middle East Technical University, Ankara, Turkey 06800
[2]EMBL-EBI, Cambridgeshire, UK
[3]Hacettepe University, Ankara, Turkey
[4]Koç University, İstanbul, Turkey 34450
[5]Stockholm University, Stockholm, Sweden

Last two decades have seen a steady increase in research that have touched on the stories hidden in human genomes either living today or who lived in the past. Genome wide association studies from modern genomes have provided a comprehensive catalogue of present-day human genetic variation and their relation to complex human traits while archaeogenetic research have revealed a fascinating snapshot of the prehistoric world dating back to last ten thousand years. However, how genetic variation associated with complex human traits have changed over evolutionary time and the extent to which this variation has been shaped by demography is poorly understood. Evaluating the temporal dynamics of loci known to be associated with complex human traits and understanding the influence of geography, demographic history and cultural backgrounds on this variation holds great promise for understanding the evolution of complex diseases. To this end, we investigated the changes in frequency of genetic variants with small and intermediate effects under different demographic models. We ran forward-time genetic simulations using SLiM for two different demographic scenarios mimicking the genetic and demographic changes in the last 10,000 years in Anatolia and Scandinavia as captured by ancient DNA data, and we further evaluated allele frequency trajectories over time. Our approach considering the spatiotemporal distribution of disease risk alleles is not only important to understand the evolution of the complex diseases but has potential to guide translational research through the exploration of the local differences in genetic disease risk factors.

---

## DriveWays: A method for identifying possibly overlapping driver pathways in cancer

Ilyes Baali, Cesim Erten and Hilal Kazan
Antalya Bilim University, Antalya, Turkey 07190

The majority of the previous methods for identifying cancer driver modules output non-overlapping modules. This assumption is biologically inaccurate as genes can participate in multiple molecular pathways. This is particularly true for cancer associated genes as many of them are network hubs connecting functionally distinct set of genes. It is important to provide combinatorial optimization problem definitions modeling this biological phenomenon and to suggest efficient algorithms for its solution. We provide a formal definition of the Overlapping Driver Module Identification in Cancer (ODMIC) problem. We show that the problem is NP-hard. We propose a seed-and-extend based heuristic named DriveWays that identifies overlapping cancer driver modules from the graph built from the IntAct PPI network. DriveWays incorporates mutual exclusivity, coverage, and the network connectivity information of the genes. We show that DriveWays outperforms the state-of-the-art methods in recovering well-known cancer driver genes performed on TCGA pan-cancer data. Additionally, DriveWay's output modules show a stronger enrichment for the reference pathways in almost all cases. Overall, we show that enabling modules to overlap improves the recovery of functional pathways filtered with known cancer drivers, which essentially constitute the reference set of cancer-related pathways. The data, the source code, and useful scripts are available at: https://github.com/abucompbio/DriveWays.

---

# Validation of LOAD-RF-RF selected risk SNVs for the early and differential diagnosis of Alzheimer's disease

Sevda Rafatov, Hüseyin Cahit Burduroğu, Yavuzhan Çakır, Onur Erdoğan, Cem İyigün and Yeşim Aydın Son

Department of Health Informatics, Graduate School of Informatics, Middle East Technical University, Ankara, Turkey 06800

Late-Onset Alzheimer's Disease (LOAD) is the most common type of dementia in the aging populations, characterized by deterioration of memory and other cognitive domains. The complex genetic etiology of the LOAD is still unclear, which restrains the early and differential diagnosis of LOAD. Genome-Wide Association Studies (GWAS) allows exploration of the statistical interactions of individuals variants, but the univariate analysis oversees interactions between variants. The machine learning algorithms can capture hidden, novel, and significant patterns considering nonlinear interactions between variants for the understanding of the genetic predisposition for the complex genetic disorders, where multiple variants determine the risk. We developed in-silico LOAD models based on genotyping data from three different datasets from ADNI and dbGAP initiatives, through controlled access. GWAS datasets provided by ADNI (210 controls and 344 cases), and GenADA (777 controls and 798 cases), and NCRAD by dbGaP (1310 controls and 1289 cases) are analyzed. In the first step, GenADA, NCRAD, and ADNI datasets analyzed independently, and after preprocessing, PLINK is used for GWAS and followed by p-value filtering for the initial dimension reduction. For each dataset, two-step Random Forest (RF) is implemented with 5-fold cross-validation (CV) using the RANGER R package after GWAS with PLINK. Test performances of LOAD-RF models of ADNI, NCRAD, and GenADA datasets were 72,9%, 68,8%, and 92,4%, respectively. 390 SNVs from ADNI, 1740 from NCRAD, and 434 from GenADA datasets selected by the individual LOAD-RF models considering permutation importance of variants at 95% confidence. There were no consensus variants, but 62 genes common in at least two datasets are identified. Additionally, six genes were common in all 3 LOAD-RF models is identified. The test performances of LOAD-RF-RF models of ADNI, NCRAD and GenADA datasets were 74,0%, 72,1%, and 85,1% respectively. 32 SNVs from ADNI, 581 from NCRAD, and 107 from GenADA datasets selected by the individual LOAD-RF-RF models considering permutation importance of variants at 95% confidence. The LOAD-RF-RF analysis identified the SNVs that are highly significant and six SNVs are selected for experimental validation with pyrosequencing. Initially, we have genotyped 41 LOAD patients for the SPOCK1 variant and observed the minor allele frequency as 0.317 , which is significantly higher than the expected global minor allele frequency of 0.154. The experimental validation of the rest of the LOAD-RF-RF selected risk variants is still ongoing. SNVs identified and validated in this study will be utilized for the development of a genotyping kit for the early and differential diagnosis of LOAD. The kit will support the clinician's decision in the early and differential diagnosis of LOAD and benefit the patients and their families for the planning of the treatment and support strategies.

## Integrating omics to unravel hepatocellular carcinoma using single-cell sequencing

Muntadher Jihad and Idil Yet

Deparment of Bioinformatics, Institute of Health Sciences, Hacettepe University, Ankara, Turkey 06100

Studying cancer cells at the level of single-cell resolution dissect the heterogeneity of cancer cell population. Analyzing the different omics simultaneously could help our understanding of the mechanism of how these different omics regulate each other in individual cancer cells. The main aim of this study is to understand this biological mechanism by linking copy number variation (CNV), DNA methylation and gene expression in Hepatocellular carcinoma (HCC) using Bayesian networks (BN). In this study, dataset of 25 single-cell sequencing data were used to detect causality between omics. First, we analyzed the three omics separately by using Tophat and Cufflinks for gene expression, Bismark tool for DNA methylation and HMMcopy R package for CNV calculation and we calculated how they correlate with each other. Then, we built three BN models representing the alternative hypotheses of the causal relations between CNV, gene expression and DNA methylation. The parameters of these networks were estimated by using MLE. Afterwards, we examined the compatibility of these structures with the data by using the AIC score. In summary, we applied a new approach on how the different omics regulate each other and detected several cases of three-way associations where either genetically driven DNA methylation levels impact gene expression profiles, or genetically driven gene expression traits impact DNA methylation levels. Moreover, we inferred the active and passive effect of DNA methylation on gene expression levels.

---

## Age-related diseases share common genetic associations

Handan Melike Donertas[1], Daniel K Fabian[1], Matias Fuentealba Valenzuela[2], Linda Partridge[2,3] and Janet M. Thornton[1]

[1]EMBL-EBI, Cambridgeshire, UK
[2]Institute of Healthy Aging, University of College London, UK
[3]Max Planck Institute for Biology of Ageing, Cologne, Germany

Ageing is the major risk factor for many diseases. With the rise in life expectancy, the overall burden of ageing-related diseases increases. The molecular link between ageing and age-related diseases, however, remains elusive. In this study, we test whether diseases with similar age-of-onset share a genetic component that is also implicated in ageing. We perform GWAS on UK Biobank data, which includes genomic, medical and lifestyle measures for almost half a million participants. Our analysis comparing 116 diseases suggested four disease clusters defined by their age-of-onset. We found that diseases with the same onset profile are genetically more similar, suggesting a common aetiology. Moreover, this similarity cannot be explained by disease categories (e.g. cardiovascular, endocrine), co-occurrences, or disease cause-effect relationships. Two of the clusters showed an age-dependent profile, starting to increase in prevalence after the age of 20 and 40 years. These clusters had genetic risk factors associated with senescence regulators and targets of the pro-longevity drugs. However, they had distinct functional enrichment and risk allele frequency distributions. We also tested predictions of mutation accumulation and antagonistic pleiotropy theories of ageing and found support for both. We are now working on a drug repurposing approach to find drugs targeting the common genetics between age-related diseases. This approach has the potential to identify drugs targeting multiple diseases simultaneously and alleviate the effects of multimorbidity and polypharmacy in late ages.

---

# SLPred: A multi-view subcellular localization prediction tool for multi-location proteins

Gökhan Özsari[1], Ahmet Süreyya Rifaioğlu[1], Tunca Doğan[2], Rengül Çetin-Atalay[1] and Volkan Atalay[1]

[1]Middle East Technical University, Ankara, Turkey 06800

[2]Department of Computer Engineering, Hacettepe University, Ankara, Turkey 06800

Identifying subcellular localization of a protein presents initial information about its functions and interactions. Protein subcellular localizations are traditionally identified by in vivo and in vitro methods, which are expensive and time-consuming. As an alternative, subcellular localization of proteins can be predicted by computational methods. Although various computational methods are proposed to predict the subcellular localization of proteins in the last decade, there is still room for significant improvement in this area, especially considering the proteins that localize to multiple subcellular compartments. In this study, we propose SLPred, an amino acid sequence-based multi-label, and a multi-view subcellular localization prediction method. The proposed method consists of nine independently constructed classification models, each of which produces binary predictions for the query protein sequence in terms of nine subcellular locations: Cytoplasm, Nucleus, Cell membrane, Mitochondrion, Endoplasmic Reticulum, Secreted, Golgi apparatus, Lysosome, and Peroxisome, as described by UniProt's subcellular locations (SL) vocabulary. We employ multiple types of protein descriptors to extract the features of the input sequences, and we select the best-performing one for each subcellular location out of 40 descriptors that span various types of protein attributes, including physicochemical and evolutionary properties. We employed multiple support vector machine (SVM) classifiers, where each individual classifier is trained with a selected descriptor to present an independent result. The predictions are finalized by weighted mean voting and thresholding operations (Figure 1.a). With the aim of improving UniProt's SL vocabulary, we re-organized the SL hierarchy using Gene Ontology cellular component ontology. Another contribution of our study is the benchmark dataset of human proteins that we constructed, which can be used for training and testing in future SL prediction studies as a standard dataset. We trained and tested SLPred using multiple benchmark datasets and compared our performance against five state-of-the-art machine learning-based SL prediction methods: CELLO2.5, MultiLoc, LocTree2, SubCons, and DeepLoc, and found that SLPred performs consistently better (Figure 1.b). SLPred is freely available as a ready-to-use stand-alone command-line tool at for researchers to analyze their unknown protein sequences in terms of their compartmentalizations.

---

# A comparison of deep CNN and BLSTM based RNA splice site prediction models

Amin Zabardast[1], Elif Güney Tamer[1], Yeşim Aydın Son[1] and Arif Yılmaz[2]

[1]Department of Health Informatics, Graduate School of Informatics, Middle East Technical University, Ankara, Turkey 06800

[2]TUBITAK Space Technologies Research Institute, Ankara, Turkey 06800

RNA splicing is a cellular process that pre-mRNAs are processed into mature mRNAs. This provides the production of multiple mRNA transcripts from a single gene. This process is important for proper protein synthesis. Thanks to advanced sequencing technologies, many splice site variants have been reported to be associated with the diseases. RNA splice site prediction is essential for gene finding and genome annotation, as well as identifying potential variants that are biomarkers and disease-causing variants. In the past few years, Deep Learning has become a prominent pattern recognition tool. Many successful algorithms have been designed to extract complex patterns from data with different characteristics. Most promising approaches for genomic data structures are based on Convolutional Neural Networks (CNNs) or Bidirectional Long Short-Term Memory (BLSTM) cells. CNNs are sensible choices because, in genomic data, each nucleotide has a correlation with other bases in its vicinity. LSTM is also a sensible class of methods since genomic data are encoded as a one-dimensional sequence. Our CNN and BLSTM models have resulted in 98 % and 96 % accuracy, respectively.

---

## Identification of the subfamily-specific functional residues of aminergic GPCRs through evolutionary analyses

Berkay Selçuk, Ogun Adebali
Sabancı University, İstanbul, Turkey 34956

The aminergic subfamily of GPCRs includes receptors for dopamine, serotonin, epinephrine, histamine, trace amine, and acetylcholine which take part in many various physiological functions and targets of many different drugs. Throughout evolution, functionally important residues are highly conserved among orthologous receptors. Some of the residues are specifically conserved in receptor subfamilies. Although the deeply conserved positions are well established in a wide range of receptors, specifically conserved residues are mostly underexplored. Here, we established the evolutionary history of aminergic receptors and identified distinct subfamilies. We explored receptor specific conservation profiles among aminergic GPCRs. We hypothesized that subfamily specific conserved residues play a role in homodimerization, G-protein coupling selectivity, and ligand binding selectivity. We aimed to identify novel deleterious variants that may disrupt homodimer formation or G-protein coupling. Our methodology is applicable to all kinds of protein families and it can be used to identify many other evolutionary conservations related to a biological function.

---

# Development of text-mining tool (SEDA) and its specific application on cardiomyopathy disease

Dilara Karaoğlu, Seda Serttürk, Evren Ata and Aslı Kutlu
Biruni University, İstanbul, Turkey 34010

Personalized medicine is one of the popular approaches in medical and biological sciences, and its good integration with bioinformatics tools has expanded the application fields. Offering the most effective treatment for each patient is one of the main motivations of the personalized medicine approach, and following the recent literature becomes a key step in the decision-making process. Due to the great attention in personalized medicine, there exists a need for health decision algorithms developed through high-level programming languages that already compromised the statistical analyses and numerical computations. In the scope of this project, we present a tool, called as SEDA, that enables us to facilitate making research in the PubMed database and provide a classification of scientific literature utilizing the text mining approach with the published abstracts. The tool is created in different programming languages: R and Python enabling users to filter specific terms and dates, and further tested on cardiomyopathy disease since the great expansion in the number of publications in recent decade create a need for catching up recent literature with a precise manner. First, SEDA tool has been employed to retrieve the cardiomyopathy related genes, counted as 33 genes, and these genes are further analyzed in the String tool to reveal the relationships between them. It is revealed that there are found experimentally verified 12 interacted gene pairs. These interacted gene partners are also verified with KEGG and Reactome pathway databases and listed as cardiac muscle contraction, hypertrophic cardiomyopathy, dilated cardiomyopathy, arrhythmogenic right ventricular cardiomyopathy, adrenergic signaling in cardiomyocytes and calcium signaling pathways, as being strongly associated with cardiomyopathy. Indeed, we present how combining text-mining approach with bioinformatics tools is an effective approach for both catching up the updated literature before making decisions about treatment and revealing newly identified genes or pathways for cardiomyopathy disease.

---

## Comparative assessment of 11 predictors of free energy change upon mutation on a dataset of 1390 mutations from 72 different proteins

Narod Kebabci and Emel Timuçin

Department of Biostatistics and Medical Informatics Faculty of Medicine, Acibadem University, Istanbul, Turkey 34752

Predicting the effect of a mutation on protein stability which is usually given by the change in Gibbs free energy upon folding (DDG) is a crucial bioinformatics task, refining our understanding of folding. Many computational methods have been developed for this purpose. These methods use various approaches, such as biophysical modeling and machine learning or combinations thereof. Despite the vast availability and easy-to-use nature of several DDG predictors, their results may conflict, leading to difficulty interpreting modelers' conclusions. This study evaluates the performances of eleven different predictors, including the novel algorithms against a large mutation dataset used for PON-Tstab, which consists of 1390 number of mutations from 72 different proteins. Scoring performances were assessed by correlation, which suggests DeepDDG as the best predictor. Because linear correlation may include bias, we also analyzed whether the signs of mutations (stabilizing or destabilizing) were accurately predicted. Mathew's correlation and ROCAUC metric were calculated for each tool. We also recruited a p53 mutation subset from of 10 mutations to a workflow that performs 100 ns of molecular dynamics simulations and subsequent MM-PBSA based stability prediction. Although this workflow may arguably be considered robust as it is computationally more demanding than other tested predictors, we found that its scoring accuracy is merely as good as the medium level predictor, reflecting the necessity for thorough optimization for this workflow. Overall computationally demanding MD steps may not contribute to a more accurate prediction of relative DG upon mutation. Thus other fast predictors, particularly those use deep learning, may be tried for a similar if not higher accuracy.

---

# ChemBoost: A chemical language based approach for protein - ligand binding affinity prediction

Rıza Özçelik[1], Hakime Öztürk[1], Arzucan Özgür[1] and Elif Ozkirimli[2]

[1]Department of Computer Engineering and [2]Department of Chemical Engineering, Bogazici University, Istanbul, Turkey 34342

Identification of high affinity drug-target interactions is a major research question in drug discovery. Proteins are generally represented by their structures or sequences. However, structures are available only for a small subset of biomolecules and sequence similarity is not always correlated with functional similarity. We propose ChemBoost, a chemical language based approach for affinity prediction using SMILES syntax. We hypothesize that SMILES is a codified language and ligands are documents composed of chemical words. These documents can be used to learn chemical word vectors that represent words in similar contexts with similar vectors. In ChemBoost, the ligands are represented via chemical word embeddings, while the proteins are represented through sequence-based features and/or chemical words of their ligands. Our aim is to process the patterns in SMILES as a language to predict protein-ligand affinity, even when we cannot infer the function from the sequence. We used eXtreme Gradient Boosting to predict protein-ligand affinities in KIBA and BindingDB data sets. ChemBoost was able to predict drug-target binding affinity as well as or better than state-of-the-art machine learning systems. When powered with ligand-centric representations, ChemBoost was more robust to the changes in protein sequence similarity and successfully captured the interactions between a protein and a ligand, even if the protein has low sequence similarity to the known targets of the ligand.

---

## Characterization of primary cilium as a mediator of gastrointestinal stem cell – niche communication at a cellular level

Deniz Esen, Müge Bozlar, Nagihan Gizay Gönüllü and Bahar Degirmenci Uzun
Department of Computer Engineering, Bilkent University, Ankara, Turkey 06800

Wnt/$\beta$-catenin signaling is one of the key regulators in tissue homeostasis and renewal. Wnt molecules are able to establish gradients in intestinal and liver tissues, leading to highly heterogeneous cell populations with specific gene expression profiles. The mesenchymal portion of the gut emerges as the only Wnt source for the epithelial layer of the colon, and in the small intestine, it is a secondary source of Wnt next to Paneth cells. In the liver, endothelial cells in the central vein regions have been proposed as the stem cells of the organ due to active Wnt signaling in these cells, as well as compliance with other stem cell characteristics. However, the hydrophobic nature of Wnt molecules necessitates a relay structure in order to establish coordination between cells. Primary cilia are small organelles protruding from cell membranes in a wide range of tissues and could potentially act as the antennae for Wnt and Hedgehog signaling molecules, enabling cross-talk between the stem cell niche and terminally differentiated cells. Here we investigate publicly available single-cell RNA sequence (sc-RNAseq) data from various human and mouse intestine and liver studies in order to characterize changes in cell expression profiles in homeostasis and disease conditions. Arl13b and Ift88 transcripts, which mark the presence of primary cilia, were found to be expressed in all cell types throughout the investigated tissues. Interestingly however, only a limited number of cells within each population appear to possess primary cilia. We aim to identify the common and cell type specific gene expression patterns of the primary cilia expressing cells. Subsequently, we aim to uncover their role in homeostasis and response to injury, and how they may be contributing to coordination of tissue renewal.

---

# Prediction of the effects of single amino acid variations on protein functionality with structure and residue-level annotation centric modeling

Fatma Cankara[1] and Tunca Dogan[2]

[1]Department of Bioinformatics, Middle East Technical University, Ankara, Turkey 06800
[2]Department of Computer Engineering, Hacettepe University, Ankara, Turkey 06800

Genomic variations may cause deleterious effects on protein functionality and perturb biological processes via various mechanisms. Elucidating the effects of variations is important for developing novel treatments for persistent diseases of genetic origin. Computational approaches have been aiding the experimental work in this field by providing variant effect predictions. However, today, predictive performance of the state-of-the-art methods is still moderate and new approaches are required. In this study, we propose a new machine-learning-based method to predict the function changing capabilities of single amino acid variations (SAV) with unknown consequences. Considering the novelty in this work, we evaluated the variations' function altering capabilities within a protein residue-level-annotation and structure centric approach, since functional sites/regions of proteins are more sensitive to changes. For this, we extracted the correspondence between the mutated residue and 30 different sequence feature annotations obtained from the UniProtKB (active/binding/lipidation/glycosylation sites; calcium, DNA binding, inter/transmembrane regions, etc.), together with 3-D structural features such as; protein domains, categorization according to the location of variation (core/interface/surface), and the change in physicochemical properties (polarity, composition, molecular volume), due to the variation. We also mapped both the mutated residue and the annotated residues on the 3-D structure of the corresponding protein and calculated the spatial distance between these residues, since a proximal variation (e.g., in the same pocket) may also affect the functionality. We constructed a 67-D feature vector for each variation datapoint using above-mentioned properties and trained our model via random forest classifier on 100,000 SAVs with experimentally known consequences. According to our benchmarks, our method displays competing and complementary performance against widely-used state-of-the-art predictors. This led to combining our approach with these methods under a hybrid/ensemble tool and increasing the prediction performance of the state-of-the-art, which can be utilized by researchers working on topics related to the molecular mechanisms of diseases.

---

## A novel probabilistic approach for detecting acceptable amino acid substitutions

Nurdan Kuru, Aylin Bircan and Ogün Adebali

Molecular Biology, Genetics and Bioengineering Program, Faculty of Engineering and Natural Sciences, Sabancı University, İstanbul, Turkey 34956

Selective pressure at amino acid sites in a protein varies based on the importance of a site in protein structure and function. Amino acids which are crucial for structure and function such as those at ligand binding sites or dimerization sites are conserved over evolutionary time. Detecting these amino acids, and acceptable substitutions of them to any other amino acid is necessary to reveal the functional properties of a protein as well as disease-causing mutations. In this study, we built a novel taxonomy dependent probabilistic function based on the maximum likelihood phylogenetic tree and ancestral sequence reconstruction of homologs of a protein to identify acceptable substitutions of an amino acid in humans. We calculated the probability of every amino acid at each site in a protein by considering whether the substitution events are independent during evolution and the distance of species to human in which the substitution is observed. We try our method on different protein families and compare it with programs like PolyPhen-2 and SIFT.

---

# Unraveling genome-wide interactions between genome structure and nucleotide excision repair

Sezgi Kaya and Ogün Adebali

Molecular Biology, Genetics and Bioengineering Program, Faculty of Engineering and Natural Sciences, Sabancı University, İstanbul, Turkey 34956

R-loop is a type of non-canonical structures on DNA which form during transcription when the nascent RNA anneals to a template DNA strand, leaving non-template strand as single strand DNA. Recent evidence show that R-loops affect gene expression by protecting the promoters from methylation and ensuring efficient transcription termination at transcription end sites. On the other hand, they are processed into double-strand breaks (DSBs) and cause genome instability. Our project aims to discover how R-loops influence nucleotide excision repair (NER), which is a crucial pathway for the repair of bulky DNA lesions induced by ultraviolet (UV) light as well as chemotherapeutic agents and chemical carcinogens. To assess the potential impact of R-loops on NER, we have processed DRIP-seq and ssDRIP-seq data and obtained the R-loop-forming regions on the genome of HeLa cells, together with Damage-seq and XR-seq data to see the distributions of UV-induced damage and its repair on the same genome. We have intersected the locations of R-loops, UV-induced damage and repair sites. By normalizing the repair with the damage abundance, we have found the repair rate of UV-induced damages around R-loops. The results suggested a lower repair rate at the center of the R-loops. Interestingly, higher repair rate was observed on R-loops around highly expressed genes when compared to that of on R-loops around genes with lower expression. In addition, we have intersected R-loop-forming locations with DSB locations on the genome. The DSB regions that contained R-loops had higher repair rate than DSB regions with no R-loops. Moreover, an R-loop abundance was observed in gene bodies, upstream of transcription start sites (TSS), downstream of transcription end sites (TES) and around DSBs. These results provide an insight about the distribution of R-loops on the genome and their effects on genome stability in terms of DNA repair. Although the exact mechanisms how R-loops implement their functions is still unclear, our research highlights the importance of R-loops and paved the way for further research.

---

## Identification of unique features of drugs via integrating fluxome and transcriptome

Hilal Taymaz-Nikerel
Istanbul Bilgi University, İstanbul, Turkey 34060

Unveiling the mode of action of drugs is essential to understand the response of the drug, develop new drugs, adjust the dose of the drug, propose combinatorial therapies and for repurposing studies. Doxorubicin and imatinib are chemotherapy agents used as an anthracycline antibiotic, and targeted therapy, respectively. Although imatinib is known to inhibit activity of the tyrosine kinase enzyme, while doxorubicin to interact with DNA through intercalation, exact mechanisms of action for both drugs have not been resolved. In order to reveal the functional changes resulting from the exposure to a certain drug, cellular responses to this drug might well be quantified. In this work, long-term effects of doxorubicin and imatinib were investigated in Saccharomyces cerevisiae, a eukaryotic model organism, through exploring responses on the genome-wide transcriptome and genome-wide fluxes. Differentially co-expressed genes under the presence of two different drugs were identified; induced and repressed pathways, in the metabolic network, associated with the response of cells to long-term exposure of these two different chemotherapeutic drugs were further categorized. Then, fluxes through the reactions catalyzed by these differentially co-expressed genes were compared to elucidate the unique features of imatinib and doxorubicin. Adverse response of a certain gene at the transcriptional versus flux level indicated possible regulatory mechanisms, which offer new prospects towards detection of distinctive drug targets in cancer therapy.

---

# Genome-wide effects of DNA replication on nucleotide excision repair of UV-induced DNA lesions

Cem Azgari[1], Jinchuan Hu[2], Yi-Ying Chiou[3], Aziz Sancar[4] and Ogun Adebali[1]

[1]Molecular Biology, Genetics and Bioengineering Program, Faculty of Engineering and Natural Sciences, Sabancı University, İstanbul, Turkey 34956

[2]Fifth People's Hospital of Shanghai and Institute of Biomedical Sciences, Fudan University, Shanghai 200032, China

[3]National Chung Hsing University, Taichung, Taiwan

[4]Department of Biochemistry and Biophysics, University of North Carolina, Chapel Hill, North Carolina 27599

Replication can cause unrepaired DNA damages to turn into mutations that might lead to cancer. Nucleotide excision repair is the leading repair mechanism that prevents melanoma cancers by removing UV-induced bulky adducts. However, the role of replication on nucleotide excision repair, in general, is yet to be clarified. Recently developed methods Damage-seq and XR-seq map damage formation and nucleotide excision repair events respectively, in various conditions. Here, we applied Damage-seq and XR-seq methods to UV-irradiated HeLa cells synchronized at two stages of the cell cycle: early S phase, and late S phase. We analyzed the damage and repair events along with replication origins and replication domains of HeLa cells. We found out that in both early and late S phase cells, early replication domains are more efficiently repaired relative to late replication domains. The results also revealed that repair efficiency favors the leading strands around replication origins. Moreover, we observed that the repair efficiency of the strands around replication origins is inversely correlated with the number of melanoma mutations.

---

## Mitochondrial DNA haplogroups in Central Anatolia: a small-scale research

M. Arda Temena, Oğuz Çilingir, Ebru Erzurumluoğlu Gökalp, Duygu Çınar, Ezgi Susam, Beyhan Durak Aras and Sevilhan Artan

Medical Genetics Department, Faculty of Medicine, Eskisehir Osmangazi University, Eskisehir, Turkey 26040

Mitochondrial DNA (mtDNA) has been one of the important subjects for recent researches, considering that treatments regarding genetic variations has started to offer promising solutions. The variants detected in mtDNA have been associated with both structural and functional complications at organs with high energy demand such as heart and brain. In addition, these variants could be useful to determine maternal haplogroups as these could also be beneficial to predict regional predispositions to various diseases. In this study, we aimed to investigate mtDNA haplogroups of both mitochondrial suspected twenty-seven patients mainly involving a heart condition and thirty-one healthy demographically-matched control subjects from Central Anatolia region. Not only variations in control region of mtDNA but also other variants in the remain part were analyzed by whole mtDNA sequencing. No significant differences were established; yet, there were found a novel variant in MT-RNR1, in a subject. More than half of the subjects are grouped into three main haplogroups H, U and J, which are 32%, 16% and 11% of the sum, respectively, which is similar to the findings of one previous attempt before in Central Anatolia. While there has been considerable recent progress in studying mtDNA variation, little data of Anatolian peninsula are available. With ongoing investigations like ours, it is certain that such efforts could be used to determine fundamental characteristics of local populations and contribute to the vast knowledge that is valuable for decision-making processes in clinical aspects as well.

---

# GROMACS performance optimization on the Turkish National Grid Resources TRUBA

Büşra Savaş[1] and Ezgi Karaca[2]
[1]Kadir Has University, Istanbul, Turkey 34083
[2] Dokuz Eylül University, Izmir, Turkey

The molecular dynamics (MD) simulations is a valuable tool to explore time-dependent motions of atoms according to Newton's equation of motion [1, 2]. This setup allows us to simulate biological systems with hundreds to millions of atoms for the sake of investigating biomolecular motion. While MD simulations of small systems, such as drug molecules, could be carried out on a standard laptop computer in a manageable runtime, simulating larger systems, such as heterogeneous protein-DNA complexes require big computing resources. For studying such systems efficiently, national and international consortia offer researchers an access point to high performance computing (HPC) resources. To that end, TUBITAK ULAKBIM offers the indispensable TRUBA HPC center to the use of Turkish computational biology community. In this work, to aid TRUBA's community-wide efforts in addressing Turkey's HPC needs, we have benchmarked the performance of the famous MD package GROMACS on a large protein-DNA complex (composed of 8693 number of atoms) [3, 4]. While benchmarking, we have run two different GROMACS versions (GROMACS 5.1.4 and GROMACS 2020) on three CPU/GPU clusters (levrek-cuda, barbun-cuda and akya-cuda), offering different CPU/GPU ratios. On a single node, levrek-cuda offers 24/2, barbun-cuda 40/2, akya-cuda 40/4 CPU/GPU ratios of different generations (for more, please check TRUBA WIKI). As a result, we have observed the best performance with GROMACS 2020 version on the akya-cuda machine, and the worst one was obtained with GROMACS 5.1.4 version on the levrek-cuda. The optimum CPU/GPU ratio was determined as 40/1, since using more GPU cores has led to a decrease in the simulation performance. When we have selected the same CPU/GPU ratio on barbun-cuda and akya-cuda; akya-cuda resulted in 1.8 times higher performance with GROMACS 2020. This performance difference between barbun/akya-cuda depends on the GPU card model. Since, the maximum performance is achieved with akya-cuda, we encourage researchers to use this computing cluster for more efficient and faster simulations. To aid the community in benchmarking their own systems on TRUBA or on their local resources, we are sharing our TRUBA parameter and GROMACS run and log files over Github at .

# Understanding the mechanism of PAD2 using multiple microsecond long molecular dynamics simulations

Erdem Çiçek and Fethiye Aylin Sungur
Istanbul Technical University, Istanbul, Turkey

Highly conserved histone proteins are primary protein components of chromatin fiber complexesserving as the scaffold for DNA in eukaryotic cells. Structural modifications in histone moleculescauses loss of interaction with DNA and other nuclear proteins which affects major chromatin func-tions like transcriptional activation/inactivation, chromosome packaging, and DNA damage/repair.Such structural modifications belong to a set of post-translational modifications, including phosphorylation, methylation, acetylation, ubiquitination, and citrullination. Thus, the control overgenes within such a highly compact environment still is a challenging question in cell biology. Peptidylarginine Arginine Deiminase (PAD) enzymes, which are commonly found in mammalian cells, catalyze the post-translational hydrolysis of peptidylarginine to form peptidyl-citrulline in a process termed deimination or citrullination.PADs are calcium-dependent enzymes that use a nucleophilic cysteine to hydrolyze guanidinium groups on arginine residues to form citrulline. This reaction results in the loss of positive charge, thereby affecting proteinfunction and altering protein–protein and protein–nucleic acid interactions. PAD enzymes have garnered significant attention over the past several years because PAD activity is dysregulated in cancer and a host of inflammatory diseases (e.g., rheumatoid arthritis,lupus, ulcerative colitis, Alzheimer's disease, and multiple sclerosis). In this study, we report multiple MD simulations of the PAD2-BAEE homodimer complex to obtain statistically significant information about protein dynamics. In addition to its biological compatibility, we have chosen this protein as model systems for research because the availability of experimental results makes it possible to compare it with our simulations. To assess the effect of protonation states of the selected active site residues on ligand binding, five protein-ligand systems were designed. In this study, we analyzed the conformations sampled and the resulting conformational transitions in a series of 50 independent simulations of 1 microsecond length in total.

# Heterogeneous COVID-19 knowledge graphs in comprehensive resource of biomedical relations (CROssBAR) system

Tunca Dogan[1], Heval Ataş[2], Vishal Joshi[3], Ahmet Atakan[2], Ahmet Süreyya Rifaioğlu[2], Esra Nalbat[2], Andrew Nightingale[3], Rabie Saidi[4], Vladimir Volynkin[3], Hermann Zellner[3], Rengul Atalay[2], Maria Martin[3] and Volkan Atalay[2]

[1]Department of Computer Engineering, Hacettepe University, Ankara, Turkey 06800
[2]Middle East Technical University, Ankara, Turkey 06800
[3]EMBL-EBI, Cambridgeshire, UK
[4]UniProt, European Bioinformatics Institute, Cambridge, UK

Systemic analysis of available biological/biomedical data is critical for developing novel and effective treatment approaches against both complex diseases and rapidly emerging outbreaks (e.g., COVID-19). Owing to the fact that different sections of the biomedical data are produced by different organizations/institutions using various technologies, the data is scattered across individual resources without any explicit relations/connections, hindering comprehensive multi-omics-based analysis. We aimed to address this issue by constructing a comprehensive biological/biomedical resource, CROssBAR, with large-scale data integration from various data sources, enriching this data with deep learning-based prediction of relations, and its presentation via cutting-edge knowledge graph (KG) representations in our open-access web-service at https://crossbar.kansil.org. Starting from late 2019, the new coronavirus pandemic has wreaked havoc and brought along nearly 850K deaths. Systemic evaluation of the current biomedical knowledge about SARS-CoV-2 infection is expected aid researchers in developing effective drugs and vaccines. With the aim of contributing to this endeavor, we have constructed two COVID-19 KGs (https://crossbar.kansil.org/covid_main.php) using the CROssBAR system; (i) large-scale version including the entirety of related information on various CROssBAR-integrated resources, and (ii) simplified version distilled to include only the most relevant terms, ideal for fast interpretation. CROssBAR COVID-19 KGs incorporate relevant virus and host genes/proteins, interactions, pathways, phenotypes and other diseases, as well as drugs/compounds, some of which are new. These new drugs have been incorporated to the KGs either due to our network analysis-based pipeline or predicted by our deep-learning-based tools. We conducted a literature-based validation study and found that many of these drugs are now being experimented at preclinical/clinical stages against COVID-19. It is interesting to observe direct/indirect relations between the phenotypes/diseases in the KGs and COVID-19 over the incorporated host genes/proteins and enriched pathways, and between COVID-19 and our computationally predicted drugs/compounds, as they may reveal further evidence to be utilized against this disease.

# Head and neck cancer: Performing functional gene enrichment study to discover the new potentials as biomarker

Evren Atak[1], Ahmet Melih Oten[1], Seda Sertturk[1], Oyku Irigul Sonmez[2] and Aslı Kutlu[1]

[1]Biruni University, İstanbul, Turkey 34010
[2]AYA RD Biotechnology Inc, İstanbul

Head and neck cancers (HNCs) are known as a heterogeneous disease that accounts for 9% of all body cancers. HNCs consist of a diverged group of tumors located in the larynx, hypopharynx, oropharynx, oral cavity and nasopharynx. Even CT and PET are used to detect and to stage HNC; the novel diagnostic approaches are required for the early diagnosis of HNC with highest accuracy and specificity, especially clarifying the subtypes of HNC. Among many, the discovery and further development of biomarker is seen as the most popular approach with highest specificity in the field of genomics and proteomics. In our study, we basically aim to perform functional gene enrichment and clustering to determine the significantly expressed and/or suppressed genes as a potential of biomarker that play a crucial role for the development and staging of HNC in terms of function, mechanism and metabolic processes. It is stated in the literature that FRMD5, PCMT1, PDGFA, TMC8, YIPF4 and ZNF324B genes are playing a role in HNC. First, we extended our gene pool by identifying miRNAs regulated by the listed genes above, and then find the corresponding genes to these identified miRNAs to obtain the extended gene pool. 3 different geo data sets were analyzed through R-program to identify the up-regulated and down-regulated genes according to logFC values. Through hierarchical clustering, these data sets were clustered, accordingly, and we tried to identify the common pathways based on clustering results. As a result of our bioinformatics studies, we identified 6 important genes in 3 geodata sets. These are TMPRSS2, NKX3-1, PAX8, ITGA2, HNRNPK and SLC44A3 genes. As a further study, we will prove our study by analyzing these genes experimentally.

# Completing the partially resolved N-Myc in the crystal complex structure of Aurora Kinase A / N-Myc by molecular modeling: insights into the molecular targets of N-Myc overexpressing tumors

Pinar Altiner[1], Süleyman Selim Çınaroğlu[2] and Emel Timuçin[1]

[1]Acibadem University, Istanbul, Turkey 34752

[2]Department of Biochemistry, University of Oxford, South Parks Road, Oxford OX1 3QU, UK

A recent crystallization study resolved the complex of Aurora kinase A (AurA) (PDB ID: 5G1X) which prevents ubiquitination of the oncogene N-Myc, leading to elevated N-Myc levels in cells. Given the fact that over-expression of N-Myc is frequently encountered in human tumors especially in neuroblastoma, this complex represents an indirect target for fighting enhanced N-Myc levels in such tumors. In this complex, N-Myc is formed by two MYC box (MB) motifs namely MB0 and MBI which span the residues from 19-47 and 61-89 respectively. These MB motifs form disordered structures and as a result they are usually difficult to spot by crystallization. Although the construct in the crystallization experiment covers both motifs, only MBI is validated while MB0 is missing from the structure. In this study, we directly analyzed the 5G1X complex by molecular dynamics simulations to assess the stability of the complex with the missing MB0 motif. Repeated simulations (n=4) and subsequent free energy calculations suggested that AurA-N-Myc complex with a single MB motif was unstable. N-Myc either partially or completely dissociated from AurA. Next, we have constructed the missing MB0 of N-Myc in the complex and simulated it similarly. Furthermore, we generated another complex in which the missing MB0 is replaced by a short protein TPX2 which also interacts with AurA. As revealed by trajectory analysis and binding free energy predictions, both complexes with the reconstructed MB0 or fused TPX2 were more stable than N-Myc in the crystal complex. Particularly, N-terminus of MBI and the interconnecting loop between two motifs showed the lowest flexibility and highest contribution to binding free energy. Consequently, these models are useful for future studies targeting N-Myc complexes, reflecting the promise of molecular modeling as a complementary method to crystallization.

---

## Identification of Novel Endochitinase class I Based Allergens

Yesim Yilmaz Abeska and Levent Cavas
Dokuz Eylül University, Izmir, Turkey

Bioinformatics analysis have been becoming important research tools in the field of life science since 2000. New tools have also been developed for detection of allergens in food sources. These tools provide important contribution to the wet-lab based experiments. In this study, potential allergen proteins similar to the endochitinase class 1(Pers a 1) protein found in Persea americana were investigated using bioinformatics analysis tools. Pers a 1 is known to cause allergic reactions. The potential allergens were examined with four different in silico tools. Clustal W for multiple sequence comparison, Swiss-MODEL for homology modelling, ProtParam for comparing protein parameters and AlgPred for allergen prediction were used. 9 of the 10 potential allergen species investigated are plants consumed in daily life and these species have been identified as allergens. According to our study results, people who are allergic to avocado (Persea americana) may be advised to be careful while consuming some foods such as Phoenix dactylifera, Allium sativum, Citrus unshiu, Diospyros kaki, Spinacia oleracea.

---

# Gene regulatory network construction and transcriptomic profiling of neuronal differentiation under the regulation of ETS transcription factor family

Yigit Koray Babal and Isıl Kurnaz
Gebze Technical University, Kocaeli, Turkey 41400

Major Transcription Factor super family called ETS family, characterized that highly evolutionary conserved DNA-binding domain about 80 amino acid residues, unique for metazoan lineage. ETS family have around 30 members and 12 subfamilies which mainly play role as transcriptional activator and inhibitor. When the literature is examined, various biological processes such as cell differentiation, proliferation, cancer progression, are regulated by ETS transcription factor family. For instance, ERF and ETV3L are required for balance between proliferation and differentiation of primary neurogenesis by interacting with retinoic acid. Another study suggests that Pea3 subfamily expression can control late progression of neuronal differentiation by regulating branching and target invasion. Transcriptomic profiling of the neuronal differentiation is emerging area to understand regulation during nervous system development and finding novel therapeutic approaches for neurodegenerative diseases. According to literature, ETS transcription factor family members are associated with neuronal differentiation with widely diverse aspects. To investigate the potential role of ETS family members on the neuronal differentiation, RNAseq study related with neuronal differentiation (GSE56785) was analyzed by using Galaxy Platform. After that downstream analysis (Enrichment analysis, network construction, protein-protein interaction) were performed to investigate neuronal differentiation under the regulation of ETS family by using R programming with related packages. Additionally, DNA binding motif of Pea3 subfamily members (ETV1, ETV4 and ETV5) were searched by using Eukaryotic Promoter Database motif search tool (FindM) on the 1 kilobase-pair upstream region of transcription start site of human genome, then transcription factor-gene interactions were constructed. With integration of transcriptomic data, transcription factor-gene interaction and protein-protein interaction, edge weighted gene regulatory network associated with Pea3 subfamily were constructed by using Cytoscape Software to investigate neuronal differentiation under the ETS family regulation. As a result, ETV4 and ETV5 may crucial role in neuronal differentiation on the both early and late stage. According to edge weighted gene regulatory network, ETV5 is the most effective gene regulation on the differentiation related gene cluster. All these information and results suggest that gene regulatory network of Pea3 subfamily play critical role divergence between maintenance of stem cells and differentiation.

# PhyloMAF: Next generation phylogenetic microbiome analysis framework

Farid Musa and Efe Sezgin
Izmir Institute of Technology, Izmir, Turkey 35430

A growing appreciation of the microbial life around and within our bodies is caused by discoveries made in microbiome research at an unprecedented rate. Technological breakthroughs in DNA sequencing have pushed microbiome research to the new limits where previously unknown microbes can not only be identified but also quantified. Moreover, studies on microbial interactions between the host genome and its internal gut microbiome, have unveiled yet unknown effects of our microbiota on our health, boosting the call for more studies. Subsequently, just as the research interest for microbiome studies grows, so does the demand for novel data analysis methods and software tools. Although the research community responds to the emerging demands with new methods and tools for microbiome data analysis, the problem of meta-analysis is yet to be addressed. Typically, samples sequenced by different sequencing platforms and processed via different OTU-picking software such as QIIME or mothur can produce significantly different abundances of operational taxonomic units (OTUs). Also, taxonomic classification databases like Greengenes, SILVA, or RDP can have considerable levels of disparity between each other; hence, resulting in incomparable taxonomy assignments for OTUs. Here we introduce our novel microbiome meta-analysis framework developed to address these shortcomings. PhyloMAF is based on Python programming language and is designed to provide a flexible and extendable framework that can satisfy a wide range of microbiome data processing and analysis requirements. To demonstrate usage applications of PhyloMAF, here we used real-life microbiome datasets to produce multidimensional scalings on beta-diversity analysis results via conventional raw dataset processing approach and using PhyloMAF to merge OTU-tables produced from different reference databases. Moreover, we demonstrate how PhyloMAF can be used to analyze and visualize microbial interactions based on OTU-tables adopting a phylogenetic perspective.

---

# A pan-cancer evaluation of NADPH generating enzymes using the TCGA cohort

Ilir Sheraj, Sreeparna Banerjee and Tulin Guray
Middle East Technical University, Ankara, Turkey 06800

Adenosine Triphosphate (ATP) is the main energy source involved in most biochemical reactions and maintenance of cellular function and physiology. Another equally important energy source that has started to gain a lot of attention recently is Nicotinamide Adenine Dinucleotide Phosphate (NADPH), which is used in many cellular processes such as biosynthesis, especially lipid and cholesterol, oxidative stress mitigation, and detoxification reactions by CYP450 enzymes. All these processes have major implications in cancer progression and response to chemotherapeutic drugs. Due to its importance, NADPH is produced by many pathways in the cell, the major one being pentose phosphate pathway (PPP). Besides that, Malic Enzyme (ME1) and Isocitrate Dehydrogenase 1 (IDH1) also produce NADPH through anaplerotic reactions. Yet another major contributor to cellular NADPH pool was shown recently to the one-carbon cycle (OCC), which takes place simultaneously in cytosol and mitochondria amd can contribute to the mitochondrial NADPH pool. In this study, we have used a bioinformatics approach to study the expression of the enzymes involved in NADPH synthesis across 20 different tumor types by using the publicly-available TCGA RNA Sequencing data. We have compared the expression of these enzymes in tumors and their adjacent normal samples, and have determined the clinical importance by checking their impact on overall patients survival. Our data indicate that the expression of PPP and OCC enzymes are significantly increased in almost all tumors, while ME1, ME3, IDH1, IDH2 and and NADK are upregulated only in some. In addition, clinical data show that higher expression of some of these enzymes are associated with poor patient prognosis for certain tumors, indicating at their importance in cancer therapeutic approaches.

---

# MAPK pathway and ETS family potential interactions in Parkinson's disease

Ekin Sönmez, Yiğit Koray Babal and Işıl Aksan Kurnaz
Institute of Biotechnology, Gebze Technical University, Kocaeli, Turkey 41400

Parkinson's Disease (PD) is one of the progressive neurodegenerative disorders. The degeneration of dopaminergic (DA) neurons is the motor symptoms of PD. The studies showed that the long- term exposure to paraquat (1,1'-dimethyl-4-4'-bipyridinium, PQ) is related with the risk of PD. Drosophila is one of the model system to identify the genetic factors, which are involved in the several pathways, and provided therapeutic targets of PD. A Drosophila model has been already developed, based on PQ exposure. ERK- ETS signaling axis is one of the potential targets for PD. The inhibitor of ERK-ETS signaling were tested in Drosophila to determine the new therapeutic targets for PD. In our study, we aim to illustrate upstream and downstream of the MAPK signaling network in Drosophila PD model by combination of bioinformatic analysis of several transcriptomic data from in vitro and in vivo Drosophila studies. The RNAseq raw data were retrieved from GEOMNIBUS and processed with the tools from GALAXY platform and downstream analysis (GO and KEGG analysis) were performed by using R programming. Network analysis were created by Cytoscape Software. As results, the differentially expressed genes from PQ treatment have role in metabolic process, drug metabolism and various pathways specially MAPK pathway. MAPK related genes expression were effected, which are upstream and downstream of Rolled (ERK), Bsk (JNK), D-p38 (p38). ETS transcription factors, Pnt (ETS1 and ETS2) and Aop (ETV6 and ETV7) are two antagonizing factors of MAPK signaling, whose RNAi transcriptomic data were investigated and DEGs were identified to analyze the ERK- ETS signaling axis. This antagonistic relation may play important role in PD. Dissection of MAPK pathway by bioinformatic approach, may shed light into molecular mechanisms of PD. Transcriptome studied from toxin-based PD Drosophila model may compare with human toxin-based studies to disclose multiple therapeutic strategies to prevent PD.

# Evaluation of Aldo-Keto Reductases as prognostic biomarkers in colon cancer

Esin Gülce Seza[1], Seçil Demirkol Canlı[2], Ilir Sheraj[1], Ali Osmay Güre[3] and Sreeparna Banerjee[1]

[1]Department of Biological Sciences, Middle East Technical University, Ankara, Turkey 06800

[2]Molecular Pathology Application and Research Center, Hacettepe University, Ankara, Turkey 06100

[3]Department of Molecular Biology and Genetics, Bilkent University, Ankara, Turkey 06800

ldo-Keto Reductases (AKRs) are a superfamily of oxidoreductases that use nicotinamide adenine dinucleotide (phosphate) (NAD(P)H) and are involved in many critical metabolic reactions in health and disease. AKR1B1 and AKR1B10, the best-studied AKRs, are very similar in structure but prefer different substrates for reduction. AKR1B1 catalyzes excess glucose to sorbitol, which is important in diabetic pathophysiology. AKR1B10 is a poor reductant of glucose but can reduce retinals and cytotoxic aldehydes. We have previously shown that AKR1B1 and AKR1B10 play highly divergent roles in the progression of colorectal cancer (CRC). High expression of AKR1B1 (AKR1B1HIGH) and low expression of AKR1B10 (AKR1B10LOW) was associated with enhanced motility and an inflammatory signature in CRC cells. Additionally, CRC tumors with AKR1B1LOW and AKR1B10HIGH showed the best prognosis indicating that these genes could serve as potential biomarkers. In this study, using publicly available microarray and RNA-seq datasets and ex vivo gene expression analysis (n = 51, Ankara cohort), we have validated our previous in silico finding that AKR1B1HIGH was associated with worse overall survival. A combined signature of AKR1B1HIGH and AKR1B10LOW was significantly associated with worse recurrence-free survival in microsatellite stable patients and a stronger mesenchymal signature when compared with AKR1B1LOW/AKR1B10HIGH tumors. According to consensus molecular subtypes (CMS) analysis, AKR1B1HIGH/AKR1B10LOW samples were primarily classified as CMS4 (mesenchymal characteristics) while AKR1B1LOW/ AKR1B10HIGH samples were classified as CMS3 (metabolic deregulation). Analysis of Reverse Phase Protein Array data from the TCGA cohort indicated that AKR1B1HIGH/AKR1B10LOW tumors showed metabolic deregulation. In vitro confirmation with colon cancer cell lines also suggested that AKR1B1HIGH/AKR1B10LOW was associated with aberrant activation of nutrient-sensing pathways. To conclude, our data suggest that the AKR1B1HIGH/ AKR1B10LOW signature may be predictive of poor prognosis, aberrant activation of metabolic pathways, and can be considered as a novel biomarker for colon cancer prognostication.

# Differential response of SW620 and SW480 colon cancer cells to a combination of L-glutamine starvation and Metformin treatment

Çağdaş Ermiş, İsmail Güderer and Sreeparna Banerjee
Department of Biological Sciences, Middle East Technical University, Ankara, Turkey 06800

Aberrant metabolism is considered to be hallmark of cancer. SW620 colorectal cancer (CRC) cells were shown to rely primarily on mitochondrial ATP generation compared to SW480 cells although these lines were derived respectively from primary and lymph node metastasis specimens from the same patient. We therefore hypothesized that compared to SW480, SW620 cells may be more vulnerable to the mitochondrial respiration inhibitor Metformin. By growing these cell lines in glucose and L-glutamine deprived medium and treating with Metformin we aimed to mechanistically unravel these differences in vulnerability. CEL files for the microarray dataset GSE89523 showing differential gene expression of 3D-cultured SW480 and SW620 cells was downloaded, normalized with RMA and analyzed in R 3.6.1 with oligo and limma packages. SW480 and SW620 cells were cultured in presence or absence of L-glutamine and treated with pre-optimized doses of Metformin. Changes in the expression of target genes and proteins were analyzed with western blot and quantitative real-time PCR. Compared to SW480 cells, L-glutamine deprived SW620 cells showed robust phosphorylation of GSK-3$\beta$, which was exacerbated with Metformin. GSK-3$\beta$ can cross-talk with the Wnt pathway. We observed that the expression of the Wnt target c-Myc was upregulated in the L-glutamine-starved cell lines, but metformin treatment increased c-Myc expression drastically only in SW620 cells. Analysis of GSE89523 indicated that SW620 cells showed stronger expression of the stem cell markers CD133, CD24, ALDH1A1 and Lgr5 compared to SW480 cells. Moreover, cell dormancy-related markers CDH1 and SOX2 were upregulated and the dormancy activation marker cyclin D1 was downregulated in SW620 cells. Since c-Myc is also a dormancy activation marker, our data suggest that metformin treated SW620 cells grown with L-glutamine deprivation may induce dormancy activation. Future studies will show whether dormancy activated cells can be killed more effectively with a combination of metformin and chemotherapy drugs.

# Sheep or goat? A comparative tool for taxon identification of low coverage ancient genomes

Gözde Atağ[1], Kıvılcım Başak Vural[1], Damla Kaptan[1], Mustafa Özkan[1], Dilek Koptekin[2], Ekin Sağlıcan[1], Mehmet Somel[1] and Füsun Özer[3]

[1]Department of Biological Sciences and [2]Department of Health Informatics, Middle East Technical University, Ankara, Turkey 06800

[3]Department of Anthropology, Hacettepe University, Ankara, Turkey 06800

Being amongst the first species subject to domestication, the origins of domestic sheep and goat have long been of interest. One major challenge in this field, is to morphologically distinguish these two closely related species' remains, especially using small bone fragments. Ancient DNA (aDNA) can theoretically help here. A straightforward strategy would be shotgun sequencing of aDNA extracted from archeological remains and comparing the mapping frequencies to the sheep and goat reference genomes. In practice, however, the high genomic similarity between sheep and goat, the quality differences between the two reference genomes, the degree of poor DNA preservation in ancient samples and consequent limited amount of data (e.g. 0.01x genome coverage) may limit this approach. Here, we propose an alternative method which focuses on mitochondrial DNA (mtDNA), to distinguish between sheep and goat using ancient DNA sequences. For this, we utilise n=216 mtDNA sites that differ between the two species, i.e. mismatch positions obtained from the alignment between sheep (Oar_v3.1) and goat (ARS1) mtDNA reference genomes. Following the alignment of sample sequences to both mitochondrial references, we assess the genotypes at the mismatch positions as reference and alternative, for each alignment. Based on the genotype composition, the shared reads are assigned as sheep or goat. Finally, the proportion of sheep and goat reads are used to determine the sample taxon using a binomial test. The analysis is restricted to transversions, aiming to reduce the impact of postmortem transitions. We applied our method on n=6 sheep and n=4 goat samples of known species identity, which yielded 100% accuracy. We further estimated accuracy using simulated ancient genome data at different coverages from the two genomes. Our method provides an easy but elaborate way to classify closely related species which are compelling to distinguish through anthropological methods.

---

# Transcriptomic analysis of Pea3 and potential miRNA interactions in neurons.

İrem Sinem Acınan and Başak Kandemir

Department of Molecular Biology and Genetics, Faculty of Science and Letters, Baskent University, Ankara Turkey 06810

ETS domain transcription factors play a role in the regulation of tissue patterning during development. Pea3 transcription factors belonging to the subfamily of the ETS superfamily are particularly crucial for oncogenesis and branching morphogenesis as well as for nervous system development. In a recent transcriptomic study, axon guidance and neurogenesis related novel target genes of Pea3 were revealed. Interestingly, over-expression of Pea3, known as an activator, suppressed many genes. This suggests that suppression may be related to the potential miRNA mechanism. This study focused on the investigation of potential relationship of miRNA with Pea3 in neurons. Initially, mRNA expression study and genome wide miRNA expression study in normal neural stem cells and adult glial stem cells were analyzed. The studies performed in glia cells were used as a control and only changes in the expressions of mRNA and miRNAs in neurons were determined using R-programming. Possible target mRNAs and miRNAs related to axon guidance and neural differentiation were identified by enrichment analysis. The target promoters of mRNAs and miRNAs were analyzed for presence of a consensus Pea3-binding ets motif and the selected mRNA and miRNA promoters with binding motifs were accepted as possible targets. The interaction between Pea3 transcription factor and the miRNA-target gene was visualized, providing the first evidence of potential down-regulation mechanism in neurons. In the future, the results of this study will be confirmed with experimental studies, and provide a better understanding of Pea3 related axon guidance mechanism.

---

# Novel methods and tools for predictive modeling of RNA-sequencing data

Gökmen Zararsiz, Vahap Eldem, Dincer Göksülük, Bernd Klaus, Selcuk Korkmaz, Gözde Ertürk Zararsiz, Ahu Durmuscelebi and Ahmet Öztürk

[1]Department of Biostatistics, School of Medicine, Erciyes University, Kayseri, Turkey
[2]Department of Biology, Faculty of Science, Istanbul University, Istanbul, Turkey
[3]European Molecular Biology Laboratory, Heidelberg, Germany

RNA-sequencing (RNA-Seq) is the state-of-the-art technique in quantifying gene expression. Using gene expression data, one major task is to identify the subset of genes and build predictive models for molecular diagnosis of diseases. Using microarray technology, numerous machine-learning methods have been developed or applied for gene-expression based classification. Since the nature of RNA-Seq data is different than microarrays, most of these methods are not directly applicable. To classify RNA-sequencing data, either microarray-based methods should be used after proper data transformation or new models should be developed directly based on the count RNA-Seq data. We provided discriminant analysis solutions considering both situations. Transformation based methods include voom based nearest shrunken centroids, voom based diagonal discriminant classifiers and voom based diagonal quadratic classifiers, while raw count based methods include sparse negative binomial linear discriminant analysis classifier. Comprehensive simulation studies resulted that the proposed methods provides fast, sparse and accurate results than the other methods including Poisson linear discriminant analysis, support vector machines and random forests classifiers. MLSeq R package and voomDDA web tool are provided for the implementation of the proposed methods. Currently, MLSeq is the commonly used software for building machine-learning models using RNA-Seq data. Using this package, researchers are able to preprocess raw RNA-Seq count data and implement hundreds of machine-learning algorithms including both the proposed methods as well as other popular approches such as deep neural networks.

---

# Regression analysis with Bootstrap confidence intervals in method comparison studies

Gözde Ertürk Zararsiz[1], Gökmen Zararsiz[1], Dincer Göksülük[1], Cengiz Bal[1], Ahmet Öztürk[1] and Gabi Kastenmuller[2]

[1]Department of Biostatistics, School of Medicine, Erciyes University, Kayseri, Turkey
[2]Institute of Bioinformatics and Systems Biology, Helmholtz Zentrum München, Neuherberg, Germany

Measurements of candidate analytical methods produced in vitro diagnostic companies are often compared with those of accepted reference methods. Linear regression models are commonly used in these comparisons. The presence of a systematic error among the methods can be determined by examining the confidence intervals of the coefficient estimates obtained from these regression models. Although bootstrap confidence intervals are widely used in statistical analysis literature, the use of these confidence intervals in method comparison studies is limited. In this study, we performed a comprehensive simulation and evaluated the performance of bootstrap confidence intervals in method comparison studies. The performance of the regression methods and confidence interval approaches are investigated and compared with each other in the simulation scenario including different measurement range, sample size, measurement distribution, analytical standard deviation ratio, whether or not the measurement error rate is known, whether there is influential observation, and the existence of constant and proportional error. Ordinary Least Squares (OLS), Deming (DR) and Passing-Bablok (PB) regression methods and analytical, jackknife, bootstrap percentile, bootstrap student, bootstrap Bca and bootstrap t confidence interval approaches were included. In these comparisons, the type-I error rates and power of the methods were determined as the evaluation criteria. For this purpose, each simulation was repeated 5,000 times. The number of bootstrap is set as 999. The performances of DR and PB methods were found to be much better than the OLS method. The results revealed that the methods had the best performance when jackknife, bootstrap percentile, bootstrap Bca and bootstrap t confidence intervals are used with DR method; bootstrap percentile and bootstrap Bca confidence intervals were used with the PB method. In conclusion, It is recommended to use bootstrap confidence intervals, especially in cases where the sample size is low.

---

## Implementation of KronaTools into QIIME 2

Kaan Büyükaltay
Department of Bioinformatics, Middle East Technical University, Ankara, Turkey 06800

Microbiome studies mainly use two different approaches: metagenomic or amplicon analysis. The metagenomic analysis is based on a whole-genome shotgun approach. It is possible to acquire taxonomic resolution to species and strain level, but requires high computational power and is expensive. In the amplicon analysis approach, marker genes like 16S, 18S, and ITS are used to identify the sequences. These genes are highly conserved but contain hypervariable regions, so allow researchers to acquire taxonomic resolution at the genus level. The downside of this approach is due to the biases introduced with PCR, its high error rate. QIIME 2 is a tool consisting of many plugins for amplicon analysis. QIIME 2 team has a decentralized development approach so that third parties can develop new plugins and share them with the community. Although there are different options for visualizations in QIIME 2, there is no tool for visualizing the contents of the features table for a single sample. On the other hand, KronaTools is a tool that visualizes hierarchical data as multi-layered pie charts. While this visualization is desired, it is not a straightforward task, as the QIIME 2 and KronaTools do not share the same data formats. Our study aims to fill this gap by developing a plugin, which will transfer QIIME2 outputs into KronaTools to ease the visualization of QIIME2 analysis results with KronaTools without additional script or manipulation during metagenomics analysis.

---

## Genotyping macro-satellites in the human population

Marzieh Eslami Rasekh and Gary Benson
Boston University, Boston, MA 02215

Macrosatellite repeats (MSRs) are DNA patterns of 100 bp or longer that repeat tandemly throughout the genome. MSRs that change copy number are called variable number tandem repeats (VNTRs), which have been predicted to have biological effects and have been linked to diseases. However, MSRs have not been studied in a high-throughput fashion. Therefore, we have developed a computational tool named Macro-Satellites Using Depth (MaSUD) to genotype MSR loci in the human genome. To predict copy number changes, MaSUD compares the number of reads mapping inside each MSR locus to a background distribution of similarly simulated reads of the reference allele. The performance of MaSUD was demonstrated on simulated datasets (precision >90% and recall>50%) and validated using long PacBio reads (linear regression p-value$< 2e-16$ and r2=0.55, correlation=74.76%). We ran MaSUD on 2,504 genomes from five super-populations of the 1000 Genomes Project using 3,875 reference MSR loci. MaSUD predicted that >95% of these MSRs have a copy number variant in at least one individual and that, on average, a locus was variant in 1,457 individuals. A total of 2,512 VNTRs overlapped with 1,190 genes that were enriched in pathways related to cancer, diabetes, neuron differentiation, and neurogenesis. To identify VNTRs affecting gene expression, we compared the mean B-cell mRNA expression levels from 448 individuals using probes overlapping VNTRs (t-test, FDR<5%). Expression of 84 genes was significantly correlated with the corresponding VNTR allele. Top genes correlated with VNTRs include FANCA, AMFR, SPG7, INPP5E, DPYSL4, GPR35, PIGN, PEX5, PRPF6, EXOC2, MXRA7, and LRCH3. Alternative Splicing was among the UniProt keywords enriched for these genes (FDR=5e-3). In addition, unsupervised clustering shows that VNTRs separate human super-populations, and using a Random Forest model we could predict ancestry with 78% accuracy. This represents the first high-throughput analysis of macrosatellites in humans.

---

# A framework with randomized encoding for a fast privacy preserving calculation of non-linear kernels for machine learning applications in precision medicine

Ali Burak Ünal, Mete Akgün and Nico Pfeifer
University of Tuebingen Tübingen, Germany 72074

For many diseases it is necessary to gather large cohorts of patients with the disease in order to have enough power to discover the important factors. In this setting, it is very important to preserve the privacy of each patient and ideally remove the necessity to gather all data in one place. Examples include genomic research of cancer, infectious diseases or Alzheimer's. This problem leads us to develop privacy preserving machine learning algorithms. So far in the literature there are studies addressing the calculation of a specific function privately with lack of generality or utilizing computationally expensive encryption to preserve the privacy, which slows down the computation significantly. In this study, we propose a framework utilizing randomized encoding in which four basic arithmetic operations (addition, subtraction, multiplication and division) can be performed, in order to allow the calculation of machine learning algorithms involving one type of these operations privately. Among the suitable machine learning algorithms, we apply the oligo kernel and the radial basis function kernel to the coreceptor usage prediction problem of HIV by employing the framework to calculate the kernel functions. The results show that we do not sacrifice the performance of the algorithms for privacy in terms of F1-score and AUROC. Furthermore, the execution time of the framework in the experiments of the oligo kernel is comparable with the non-private version of the computation. Our framework in the experiments of radial basis function kernel is also way faster than the existing approaches utilizing integer vector homomorphic encryption and consequently homomorphic encryption based solutions, which indicates that our approach has a potential for application to many other diseases and data types.

---

# New methods for clustering RNA-sequencing data

Ahu Durmuşçelebi[1] and Gökmen Zararsız[2]

[1]Department of Biology, Faculty of Science, Istanbul University, Istanbul, Turkey
[2]Department of Biostatistics, School of Medicine, Erciyes University, Kayseri, Turkey

Clustering gene-expression data is a common major task for discovering the disease subtypes. A large number of studies have been carried out using microarray technology. Next-generation RNA-Sequencing (RNA-Seq) technique has replaced microarrays as the technology of choice and microarray based methods cannot be directly used due to the discrete nature of RNA-Seq data. Moreover, RNA-Seq data is overdispersed. Thus it is necessary to understand the mean-variance relationship and to deal with over-dispersion problem before performing analysis to RNA-Seq data. In earlier studies, researchers applied various transformation techniques to use microarray-based clustering methods. In the latter studies, researchers turned to identify appropriate discrete models for clustering RNA-Seq data. In this study, we aimed to develop new clustering algorithms by combining the existing clustering algorithms with the voom transformation method. These algorithms are referred as voomPW (voom+precision weight) and voomQW (voom+quality weight) and basically work with the weighted distance matrices that are obtained using voom transformation. Both algorithms have been applied to both raw and normalized & transformated RNA-Seq data. Then, hierarchial clustering, k-means and k-medoid clustering methods were applied with the obtained weighted distance matrices. We compared the model performances with rlog and vst transformation based clustering, Poisson, model-based and edgeR negative Binomial clustering methods. Rand and adjusted rand statistics were used as evaluation criteria. Analyzes were applied to ten simulation and six real RNA-seq data. All applications are implemented in various packages of R (https://www.r-project.org) and Python (https://www.python.org/) languages. The proposed approaches performed compatibly or much better than other available methods. In this study, new algorithms have been introduced for gene-expression based clustering analysis. Researchers can access the web based implementation of the presented methods at http://opensoft.turcosa.com.tr/voomCluster/ and access all source code and data at https://github.com/gokmenzararsiz/voomCluster.

---

# Investigation of type 4 pili protein's inhibition mechanism and discovery of corresponding natural inhibitory drugs

Aslıhan Özcan Yöner[1], Halil İbrahim Özdemir[1], Özlem Keskin Özkaya[2], Berna Sarıyar Akbulu[1]t and Pemra Özbek Sarıca[1]

[1]Marmara University, İstanbul, Turkey 34722, [2]Koç University, İstanbul, Turkey 34450

Nowadays, the rapid spread of bacterial diseases and the inability to prevent it, is a big problem. The rapidly increasing resistance to antibiotics that are used in the treatment of these diseases and the slowing down of the discovery of antibiotic classes leads to different therapeutic approaches [1], [2]. An inviting strategy is to identify molecules that target bacterial virulence as an alternative to traditional antibiotics with declining effects [3]. Since existing protein structures or their functioning mechanisms are not fully known and new target proteins are constantly available, the mechanism of inhibiting the virulence factor has not been fully unraveled [4]. Current work involves identification of type 4 pili's (T4P) inhibition mechanism. At the same time, this work focuses on natural product solutions of the already mentioned inhibitory mechanism. The main purpose is to make the virulence factor dysfunctional by preventing from elongation and/or the adhesion of the elongated pili. For this intent, the PilB protein of P. aeruginosa and PilF protein of N. meningitidis were studied. The binding regions in the target structures have been determined by metaPocket 2.0, primarily to understand inhibition region(s) of relevant microorganisms [5]. It has been observed that the regions where the ATP molecule binds play an important role in the inhibition mechanism of T4P. Then, in order to specify natural products, a virtual screening method was performed. In near future, inhibitors determined as a result of these methods will have the potential to be used in therapies in combination with existing drugs and/or other virulence factor inhibitors. The methods and analyses developed within the framework of this work will surely to catch attention of others that are interested in solving protein inhibition related issues.

# Robust inference of kinase activity using functional networks

Serhan Yılmaz[1], Marzieh Ayati[3], Daniela Schlatzer[2], A. Ercument Cicek[4], Mark Chance[1] and Mehmet Koyuturk[1]

[1]Case School of Engineering and [2]Center for Proteomics and Bioinformatics, Case Western Reserve University, Cleveland, Ohio 44106

[3]University of Texas Rio Grande Valley, Edinburg, TX 78539

[4]Department of Computer Engineering, Bilkent University, Ankara, Turkey 06800

Mass spectrometry enables high-throughput screening of phospho-proteins across a broad range of biological contexts. When complemented by computational algorithms, phospho-proteomic data allows the inference of kinase activity, facilitating the identification of dysregulated kinases in various diseases including cancer, Alzheimer's disease and Parkinson's disease. To enhance the reliability of kinase activity inference, we present a network-based framework, RoKAI, that integrates various sources of functional information to capture coordinated changes in signaling. Through computational experiments, we show that phosphorylation of sites in the functional neighborhood of a kinase are significantly predictive of its activity. The incorporation of this knowledge in RoKAI consistently enhances the accuracy of kinase activity inference methods while making them more robust to missing annotations and quantifications. This enables the identification of understudied kinases and will likely lead to the development of novel kinase inhibitors for targeted therapy of many diseases. RoKAI is available as web-based tool at http://rokai.io.

---

# A dynamical model based-on side chain relaxations provide the mechanism of action of resistance conferring mutants

Ebru Cetin, Ali Rana Atilgan and Canan Atilgan

Faculty of Engineering and Natural Sciences, Sabancı University, İstanbul, Turkey 34956

Proteins are complex nanomachines operating in a narrow energy range and on different time scales [1,2]. While available studies show that properties deduced from protein dynamics such as hydrogen bond networks, sulphur bridges, root mean squared deviations of backbone dynamics can shed some light on this complex picture, the functioning mechanism is still an enigma for most proteins. Studies conducted in the last decade show that protein functioning is deeply affected by entropic factors. NMR and fluorescence spectroscopy are experimental techniques having unique capabilities for understanding conformational dynamics of a protein at atomic resolution, thus showing traces of these entropic effects. In this work, we have adopted the "Model-Free Approach" by Lipari and Szabo developed for the quantification of NMR data of C-H vectors [1] to the study of the relaxation behavior of side chain C-C bond vectors. We analyze molecular dynamics trajectories and separate out the motions driven by side-chain/solvation dynamics and backbone dynamics time scales. The distribution of relaxation times of the internal time is modeled into the stretch exponent ($0< <1$), whereby the smaller the value, the broader is the distribution [2]. We have used this approach to analyze the effect of resistance conferring mutants of the E. coli DHFR towards trimethoprim on the dynamics of the enzyme. We find that our model well-describes the relaxations ($R^2 = 0.98$ for 90% of the residues). Findings of the analysis led us through to the mechanism of action of L28R mutation and the underlying effects of the addition of L28R mutation to A26T mutation [3,4].

**References:**

[1] Lipari, G., & Szabo, A. (1982). Model-Free Approach to the Interpretation of Nuclear Magnetic Resonance Relaxation in Macromolecules. 2. Analysis of Experimental Results. Journal of the American Chemical Society, 104(17), 4559–4570. https://doi.org/10.1021/ja00381a010

[2] Okan, O. B., Atilgan, A. R., & Atilgan, C. (2009). Nanosecond motions in proteins impose bounds on the timescale distributions of local dynamics. Biophysical Journal, 97(7), 2080–2088. https://doi.org/10.1016/j.bpj.2009.07.036

[3] Tamer, Y. T., Gaszek, I. K., Abdizadeh, H., Batur, T. A., Reynolds, K. A., Atilgan, A. R., Atilgan, C., & Toprak, E. (2019). High-order epistasis in catalytic power of dihydrofolate reductase gives rise to a rugged fitness landscape in the presence of trimethoprim selection. Molecular Biology and Evolution, 36(7), 1533–1550. https://doi.org/10.1093/molbev/msz086

[4] Abdizadeh, H., Tamer, Y. T., Acar, O., Toprak, E., Atilgan, A. R., & Atilgan, C. (2017). Increased substrate affinity in the Escherichia coli L28R dihydrofolate reductase mutant causes trimethoprim resistance. Physical Chemistry Chemical Physics, 19(18), 11416–11428. https://doi.org/10.1039/c7cp01458a

---

## Using the male death ratio to estimate COVID-19 burden among excess Istanbul deaths

Mehmet Somel[1], Meriç Erdolu[1], Yetkin Alıcı[1], Pavlos Pavlidis[3], Yiannis Kamarianakis[4] and Erol Taymaz[2]

[1]Department of Biological Sciences and [2]Department of Economics, Middle East Technical University, Ankara, Turkey 06800

[3]Institute of Computer Science and [4]Institute of Applied and Computational Mathematics, Foundation for Research and Technology-Hellas, Heraklion, Crete GR

Death records during March and April 2020 from Istanbul, Turkey, suggested a surge in death numbers, reaching twice as much as the seasonal death rate, and which could not be accounted for by the official COVID-19 death toll of the same period. This was interpreted by local and international media as COVID-19 deaths having gone unrecorded. The Health Ministry instead claimed the surge was an artifact caused by how deaths were recorded during the quarantine period. A third possibility is that an abnormal rise in deaths did occur but not directly caused by COVID-19 infection, rather by secondary effects, such as limited access to health services during the quarantine period. To resolve the issue, here we use a maximum likelihood model to estimate the proportion of direct COVID-19-caused deaths within Istanbul's March-May 2020 death toll, leveraging the fact that male deaths occur at higher rate due to COVID-19 than background deaths.

---

# MDeePred: Novel multi-channel protein featurization for deep learning based binding affinity prediction in drug discovery

Ahmet Süreyya Rifaioğlu[1], Rengul Cetin-Atalay[3], Deniz Cansen Kahraman[2], Tunca Dogan[4], Maria Martin[5] and Volkan Atalay[1]

[1]Department of Computer Engineering and [2]Graduate School of Informatics, Middle East Technical University, Ankara, Turkey 06800

[3]University of Chicago, Chicago, IL 60637

[4]Department of Computer Engineering, Hacettepe University, Ankara, Turkey 06800

[5]EMBL-EBI, Cambridgeshire, UK

Identification of interactions between compounds and target proteins is important for the discovery of novel drugs, drug repurposing and revealing off-target effects. However, it is not feasible to conduct experiments to cover all interactions among proteins and compounds due to the huge chemical and protein space. Therefore, computational methods are needed to aid drug discovery. Several machine learning and deep learning-based methods have been proposed in the last years for predicting bioactive compounds against known targets, however, there are still room for significant improvements in this field. Here, we present a novel proteochemometric approach, called MDeePred, for binding affinity prediction. In this approach, we proposed a multi-channel input representation of protein sequences on 2D space where each channel represents a different property of a protein such their sequence, structural, evolutionary and physicochemical features (Figure 1-1). The constructed multi-channel features of proteins were fed to a hybrid deep neural network along with the binary fingerprints of compounds where the output is the binding affinities of the input compound-protein pairs (Figure 1-2). The performance of MDeePred were evaluated on well-known datasets (i.e, Davis, Filtered Davis, PDBBind Refined) and the results were compared with state-of-the-art methods. We showed that the performance of MDeePred was significantly better than the other methods in majority of the cases. Furthermore, family-specific models were trained for kinase and GPCR protein families which were also tested on unseen compound-target pairs that were obtained from our new in-house temporal split dataset. The results showed that our models can also be used for predicting drug-target interactions even the drugs/targets are not available in the training datasets. The proposed multi-channel protein representation may also be used in other fields of bioinformatics and cheminformatics studies.

---

# SmartBioGraph: A graph database of integrated biomedical knowledge along with Web application for querying and visualisation

Buğra Aker Yılmaz[1], Ekin Tire[3], Yunus Emre Uzun[1], Ahmet Dara Vefa[3], Volkan Atalay[1], Ahmet S. Rifaioğlu, Tunca Doğan[4] and Rengül Çetin-Atalay[2]

[1]Department of Computer Engineering and [2]Graduate School of Informatics, [3]Middle East Technical University, Ankara, Turkey 06800

[4]Department of Computer Engineering, Hacettepe University, Ankara, Turkey 06800

Biological data is scattered through multiple resources, each with a focus on a limited number of biological entities. However, finding out relationships among entities from multiple sources is laborious, since each resource is able to identify entities in their own accounts and each of them has different user interfaces, advanced programming interfaces (API), database designs, etc. "Comprehensive Resource of Biomedical Relations (CROssBAR)" project (https://cansyl.metu.edu.tr/crossbar) aims to address this problem by fetching and integrating large-scale heterogeneous biological data from various resources in terms of genes/proteins, diseases/phenotypes, pathways/mechanisms/functions and drugs/compounds, and displaying the resulting information to the user via state-of-the-art knowledge graph representations to be used for systems biology research. However, database design of CROssBAR-DB, the project's document-oriented database, is complex for researchers for the task of finding the relationships among biological entities. Also, CROssBAR-DB is not suitable to apply graph-based machine learning that is an emerging field. In this study, we propose a novel graph-based system entitled SmartBioGraph, which consists of two parts: SmartBioGraphDB, a graph database mainly based on data from CROssBAR-DB, and SmartBioSearch Interface, a web application that enables users to query and visualise SmartBioGraphDB. Together, they provide an end-to-end framework for researchers without a programming background to inspect the data gathered from several popular public resources for drug discovery and systems biology. SmartBioSearch Interface also provides a novel functionality named graph search for users to construct customized knowledge graphs. In this functionality, users draw a customized abstract tree around a node of their interest as an input to query the database. To construct a knowledge graph from this input, multiple optimized queries are generated and enrichment analysis-based filtering operations are performed. SmartBioGraphDB can be a useful data source in graph-based computational systems biology research on a large scale. SmartBioGraph is available at https://smartbiograph.kansil.org.

---

## Coupled dynamics around the ribosomal tunnel focusing on macrolide discrimination

Merve Yuce, Pelin Guzel and Ozge Kurkcuoglu
İstanbul Technical University, İstanbul, Turkey 34467

Across all kingdoms of life, ribosomal complexes synthesize proteins, which are major antibiotic targets in bacteria. The nascent polypeptide grows to the solvent side through the ribosomal tunnel, the wall of which is mainly composed of the conserved portions of the 23S rRNA, and the loops of uL4 and uL22 form a constriction region in the tunnel. uL4 has an additional eukaryote-specific loop making the constriction narrower than prokaryotes, which may be responsible for the decrease of macrolide affinity to human ribosomal complexes. The ribosomal tunnel is also known to be a gateway actively taking role in translation, mostly using allostery. At this point, investigating the allosteric regulation at the ribosomal tunnel focusing on the macrolide binding regions in bacteria is highly critical to understand the discrimination mechanism of macrolides in human. Here, 500 ns long Coarse-Grained Molecular Dynamics simulations are conducted for the ribosomal structures of T. thermophilus and H. sapiens. We used ProDy to perform perturbation response scanning analysis on the covariance matrixes to identify the effectors and sensors having high possibility to take a role in allosteric regulation in the ribosomal tunnel. Sensitivity profiles of macrolide binding regions in T. thermophilus (A2058-A2059, A2062, U2585-C2586) and additional loop residues of uL4 in H. sapiens (Arg80-Gln89) are analyzed. The method captures residues/nucleotides, which are known to communicate by allostery. Among the analyzed macrolide binding regions, there is no significant difference in coupled-dynamics profiles; however, analysis of the additional loop of uL4 indicates that the sensitivity profiles of the tip residues (Gly81-His85) of H. sapiens differs, suggesting their possible role in macrolide discrimination mechanism between bacterial and human ribosomal complexes. Our findings suggest that the macrolide discrimination is not only due to the steric hindrance of additional loop of uL4, but also the difference in perturbation transmission of residues/nucleotides. This approach is highly useful to predict new antibiotics binding sites for rational drug design.

---

# Impact of Pan-Cancer mutation profiles in signaling pathways through phosphorylation events

Esra Basaran[1] and Nurcan Tuncbag[2]

[1]Department of Molecular Biology and Genetics and [2]Department of Health Informatics, Graduate School of Informatics, Middle East Technical University, Ankara, Turkey 06800

Signaling pathways are mediated by phosphorylation events; however, some of these events are inhibited or over-activated in cancer tissues. In this study, we focused on patient-specific mutations in TCGA that overlaps with residues having potential to be phosphorylated such as Serine, Tyrosine, Threonine and leverage these mutations to understand the pathway-level signaling alterations in cancer. This change may be from S, Y or T to any amino acid which we call 'inactivating mutation' or from any amino acid to S, Y or T which we called 'activating mutation'. Our main aim is to reveal the perturbed pathways and upstream kinases that are inhibited or activated because of these mutations. For this purpose, we integrated patient-specific mutation profiles of more than 10,000 tumor samples with PhosphoSitePlus. For activating mutations, we applied a rigorous motif search method with the help of GPS algorithm and found the potential kinases. As a result we found 20541 unique positions that are changed from S, Y, T to any aminoacid that can not be phosphorylated. In this dataset, 725 out of 20541 substrates are cataloged in PhosphositePlus. GSK3B, and SRC were found as the most affected kinases. Many phosphorylation sites are not annotated with their kinase regulators yet. Therefore, we will continue to search for more annotations of these inactivated positions. We next analyzed this data according to the cancer subtypes and found that the prevalence of inactivating and activating mutations is the highest in a high median value in Skin Cutaneous Melanoma (SKCM) cancer type. In our ongoing work, we will construct the patient-specific perturbed signaling networks and through a detailed comparison of these networks we will reveal the subtype-specific affected pathways.

# Integrative predictive modeling of miRNA markers in melanoma metastasis

Aysegul Kutlay and Yeşim Aydın Son

Department of Health Informatics, Graduate School of Informatics, Middle East Technical University, Ankara, Turkey 06800

Abstract- Transcriptional regulation is one of the critical mechanisms underlying cancer development. Even though mRNA, microRNA, and DNA methylation mechanisms have a critical impact on the metastatic outcome, there are no comprehensive data mining models that combine all aspects of transcriptional regulation for metastasis prediction. So, in this study, we focused on identifying the regulatory impact of genetic biomarkers for monitoring metastatic molecular signatures of melanoma by investigating the consolidated impact of miRNA, mRNA, and DNA methylation. We used the TCGA melanoma dataset to predict metastatic melanoma samples by assessing a set of predictive models. Throughout the study, a combination of differentially expressed miRNA, mRNA, and methylation signatures are used as input features. The selected biomarkers of the highest performing models are further analyzed for the biological interpretation of functional enrichment and to determine regulatory networks. According to the combination of our results of three biomarkers predicted metastatic outcome with the F-Score of %92 (sensitivity= 92 %, Specify = 93,5) and produced a stable model with low variance across multiple trials. Also, selected biomarkers showed enrichment on the metastasis-associated pathways of melanoma, such as 'Osteoclast', 'Rap1 Signaling' and 'Chemokine Signaling' Pathways. Here we presented that miRNA plays an essential role in the metastatic progression of primary melanoma and predicts metastasis outcomes with high accuracy with an F Score of 81%. Moreover, our results indicate that the integrated evaluation of miRNA with mRNA and methylation biomarkers increases the model's predictive power.

---

# Rational design of small molecules targeting PD-1/PD-L1 interaction

Baris Kalem and Ozlem Ulucan
Istanbul Bilgi University, Istanbul, Turkey 34060

Immune checkpoints are immune system regulators that maintain immune homeostasis and prevent autoimmunity. Cancer cells often manipulate immune checkpoint mechanisms to escape anti-tumor immune response by overexpressing the immune checkpoint ligands. Binding of those ligands to their receptors on T cells will cause constitutive activation of immune checkpoints, which will result in T cell exhaustion. Blocking the immune checkpoint receptor-ligand binding will release the brakes on the immune system and will reinvigorate exhausted T cells by restoring anti-tumor immune response. For this reason, the interactions between the immune checkpoint receptors and their ligands caught attention and were proven to be effective targets in treating cancer. In this study we have combined several computational approaches to target the interaction between negative immune check point receptor PD-1 and its ligand PD-L1 by designing small molecules that bind effectively to the ligand PD-L1. Using triple and quadruple combinations of the interface residues on PD-1, 35 different pharmacofore models were constructed that used later for scanning ZINC15 database. In total 14000 small molecules were retrieved from ZINC15 database using the pharmacofore models constructed from the PD-1 interfacial residues. The retrieved small molecules were further assessed using molecular docking calculations. We performed molecular dynamics simulations of top 20 protein-ligand complexes that we obtained from molecular docking calculations. Binding free energies for those molecules were computed using the well-known MMGBSA approach where the entropy was estimated based on normal mode analysis. Free energy calculations put forward 6 molecules whose free energies of binding varying from -7.1 to -18.1 kcal/mol. We further examined the promising 6 molecules using structural analysis. Combining several computational approaches, we have discovered 6 promising molecules that bind to PD-L1 with varying binding free energy. However experimental investigations are required for validation.

---

# Comparison of the impact of protein-protein interaction networks and local variant features for pathogenicity of non-synonymous single-nucleotide variants

Kazım Kıvanç Eren, Hamza Umut Karakurt and Yağmur Ceren Dardağan
Idea Technology Solutions, Istanbul, Turkey 34398

Non-synonymous Single Nucleotide Variants (nsSNVs) occurring in the coding region can lead to amino acid change that plays a huge role of occurrence of various diseases. This amino acid substitutions may change the structure of related protein and this causes the biological processes involved in the protein to be interrupted. For discovering the effect of nsSNPs, many machine learning-based models designed that uses only features related with variant itself. However, except the features associated with a variant, the interactions of this protein with other proteins may be related to the pathogenicity of the variant. In this work, we are discovering the effect features extracted from protein-protein interaction networks (PPINs) and compare with sequential features of variants. We collected 22,115 non-conflict nsSNVs from ClinVar database and merge them with the sequential features (conservation scores, allele frequencies etc.) from CADD database and PPIN features (Closeness, Betweenness centrailty etc.) extracted from HumanNet v2. We built 3 different models based on local region features, ppin features and combination of local region and ppin features. We trained different models with different feature combinations and measure which features are important for prediction of impact of nsSNVs in human diseases.

# The impact of protein-DNA force fields in the prediction of nucleosomal DNA dynamics

Ayşe Berçin Barlas, Burcu Özden and Ezgi Karaca
İzmir Biomedicine and Genome Center, Izmir, Turkey 35340

The nucleosomes compact genomic material, a single unit of which consists of a histone octamer and 147 bp long DNA. While compacting the genomic material in the nucleus of eukaryotic cells, nucleosomes serve as the main template for diverse vital processes, such as transcription and replication [1]. Therefore, understanding the nucleosomal dynamics holds the potential to disclose the hidden rules of gene regulation. Within this scope, molecular dynamics approaches have been commonly used to simulate nucleosomal dynamics. Though, these efforts have always focused on dissecting the contribution of histone dynamics on the nucleosome function. So, there hasn't been any comprehensive effort focusing how much the available molecular dynamics (MD) force fields (FFs) could reproduce the experimentally observed nucleosomal DNA geometries. To close this gap, in this work, we compare two state-of-the-art protein-DNA FFs, i.e., AMBER parmbsc1 [2] and CHARMM36 [3] in exploring the nucleosomal DNA dynamics. For this, with each FF, we have generated 2 $\mu$s long MD simulations using the most stable nucleosomal DNA sequence, i.e. the 601 nucleosome (pdb id: 3lz0) [4]. Afterwards, we have calculated the DNA groove geometries (minor and major groove) and its base pair step parameters (shift, slide, rise, tilt, roll and twist) by using 3DNA suite [5]. In the case of the groove width comparison, the average minor groove geometries predicted by AMBER oscillates around the experimentally determined minor groove values, while CHARMM36 reflect larger minor groove values than expected. CHARMM36 also leads to a flat major groove profile, while AMBER could successfully reproduce the 601 major groove profile. For the base pair step parameter comparison, both FFs reflect similar slide and shift trends, while CHARMM36 turns out to produce resolving DNA ends (end fraying), emphasized in rise, roll, tilt and twist base pair step parameters. The complete base pair step parameter comparisons can be found at https://github.com/CSB-KaracaLab/NucDNADynamics . As a conclusion, our results show that CHARMM36 results in wider conformational variations than AMBER. It also reveals that CHARMM36 leads to end fraying while AMBER is capable of keeping the nucleosomal DNA intact (Figure 1). Our comparative study provides a basis for the researchers who aim to understand nucleosomal DNA's role in gene regulation by using MD.

# Modeling the impact of designer mutations on SARS-CoV-2 and ACE2 interactions

Eda Şamiloğlu[1], Ayşe Berçin Barlas[1], Mehmet Erguven[1], Mehdi Koşaca[1], Büşra Savaş[2], Burcu Özden[1] and Ezgi Karaca[1]

[1]Izmir Biomedicine and Genome Center, Izmir, Turkey 35340

[2]Department of Computational Biology and Bioinformatics, Kadir Has University, Istanbul, Turkey 34083

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is a global pandemic, reaching 19,500,000 cases and 700,000 deaths worldwide [1]. SARS-CoV-2infection is initiated upon the recognition of SARS-CoV-2 Spike protein on the host Angiotensin-converting 2 (ACE2) enzyme. In early 2020, the high-resolution structure of ACE2:Spike interaction was resolved. Since then, several studies have been published on the optimization of ACE2:Spike interface upon modifying the available human ACE2 sequence. These approaches are crucial since a modified ACE2 holds therapeutic potential for neutralizing SARS-CoV-2 before it reaches ACE2 of the host. Expanding on this, in this work, we aim to benchmark the prediction capacity of the available computation tools in estimating the impact of recently published ACE2 designer mutations [2]. To expand our benchmark set, we have included recently published Spike designer mutations in our data set too [3]. On this complete set, we have been testing the performance of two widely used binding affinity predictors, i.e., FoldX [4] and HADDOCK [5]. Our preliminary analysis with haddock has shown that HADDOCK score could predict the direction of change (whether a mutation improves or worsens the ACE2: Spike binding) on a small subset of cases with an 86%success rate [6]. We are currently at the stage of expanding this preliminary data on our whole designer mutation set. Our benchmarking efforts will showcase the state of the available mutation modeling/binding affinity prediction tools in the development of COVID-19-combating therapeutics.

## Traces of human adaptive evolution in Mediterranean region

F. Rabia Fidan[1], Evrim Fer[2], N. Ezgi Altınışık[3], Ömer Gökçümen[4] and Mehmet Somel[1]

[1]Department of Biological Sciences, Middle East Technical University, Ankara, Turkey, 06800

[2]Department of Genetics, University of Arizona, Tucson, Arizona 85721

[3]Department of Anthropology, Hacettepe University, Ankara, Turkey 06800

[4]Department of Biological Sciences, University at Buffalo, Buffalo, NY 14260

Positive selection has always been a focus of interest owing to its direct link to adaptive evolution. Humans inhabit a wide range of habitats with varying environmental conditions with respect to temperature, elevation, humidity, oxygen level, light-dark cycle lengths, or main nutrition sources. Accordingly, signatures of local adaptations are expected to be found in the genomes of different human populations. Indeed, lactase persistence and skin pigmentation adaptations in Northern European populations and high altitude adaptations in Tibetans exemplify such local adaptations in humans. Similarly, adaptations involving fatty acid metabolism, malaria, diabetes and climate were reported in various populations around the world. In this study, we focus on the Central and Eastern Mediterranean region, which offers a distinct environment in terms of diet, humidity and climate. Specifically, we seek parallel signals of positive selection in the genomes of individuals from Turkey and Italy. We use three datasets: the 1000 Genomes data, the Turkish Genome Project dataset, and published Neolithic genomes from Anatolia. We employ two computational tools to scan the genomes for positive selection: the population branch statistic (PBS), which detects allele frequency differentiation in a specific population relative to a sister population and an outgroup; and the cross-population extended haplotype homozygosity (XP-EHH), which excels in finding selective sweeps close to fixation. In PBS analyses we compare Turkey or Italy with present-day Central Europe and Neolithic Anatolia, using Africa as an outgroup. In XP-EHH, we compare Turkey or Italy with other present-day Eurasians. Our preliminary results suggest that present-day populations of Italy and Turkey share specific genomic regions that differentiate both of them from other populations, past or present.

---

# Performance evaluation of automated machine-learning algorithms in omics data

Meltem Ünlüsavuran[1], Cem Sönmez[2], Ahu Durmuşçelebi[3], Vahap Eldem[4], Gözde Ertürk Zararsız[2], Funda İpekten[2] and Gökmen Zararsız[2]

[1]Faculity of Pharmacy, Ankara University, Ankara, Turkey 06560
[2]Department of Biostatistics, School of Medicine, Erciyes University, Kayseri, Turkey
[3]Department of Biostatistics, School of Medicine, Hacettepe University, Ankara, Turkey
[4]Department of Biology, Faculty of Science, Istanbul University, Istanbul, Turkey

The importance of omics studies has increased considerably in recent years. Omics technologies have become routinely used tools to reveal many health problems. Statistical methods are used to make sense of the data obtained using these technologies. Machine-learning methods are frequently used in omics studies, mostly for classification related problems. A wide range of machine-learning algorithms are available for classification of omics data. However, training these models to omics data may be a daunting task for clinical researchers. Automated machine learning methods (AutoML) have been presented for this purpose. These methods automate the following model building stages: (i) preprocessing, (ii) feature selection, (iii) model selection and (iv) parameter optimization. In this study, we investigated the performance of automatic machine learning methods in omic datasets. A total of 29 datasets were used, including 16 microarrays, 6 RNA-sequencing and 7 metabolomic datasets. All of these datasets are obtained from real studies for classification purposes. We included H2O and TPOT automatic machine learning methods and random forests (RF), support vector machines (SVM) and nearest shrunken centroids (NSC) machine learning methods to our experiments. These methods were compared with each other in terms of accuracy, cost and usability. Using H2O and TPOT AutoML methods, the highest classification performance was achieved with a total of 22 data, 11 microarrays, 5 RNA-seq, 6 metabolomics. Using RF, SVM and NSC methods, the highest classification performance was obtained in 11 omics datasets. AutoML methods provided compatible or better performance as compared to the classical approaches. However, as a result, no best performing approach could be found for each data. More dataset and simulation studies need to be carried out to make more precise evaluations. Finally, specialized efforts are required to automate the preprocessing steps of omics data.

---

## Benchmarking kinship estimation tools for ancient genomes using pedigree simulations

Mehmet Çetin[1], Şevval Aktürk[2], Igor Mapelli[1], Reyhan Yaka[1], Seda Çokoğlu[1], Douaa Zakaria[1], Francisco Ceballos[1], Ayshin Ghalichi[1], N. Ezgi Altınışık[6], Dilek Koptekin[3], Kıvılcım Başak Vural[1], Yılmaz Selim Erdal[6], Çiğdem Atakuman[4], Anders Götherström[7], Füsun Özer[6], Elif Sürer[5] and Mehmet Somel[1]

[1]Department of Biology, [2]Department of Molecular Biology and Genetics, [3]Department of Health Informatics, [4]Institute of Social Sciences and[5] Department of Modeling and Simulation, Graduate School of Informatics, Middle East Technical University, Ankara, Turkey 06800

[6]Department of Anthropology, Hacettepe University, Ankara, Turkey 06800

[7]Department of Archaeology and Classical Studies, Stockholm University, Stockholm, Sweden 114 19

There is growing interest in identifying genetic kinship levels among ancient individuals buried in physical proximity, from graveyards to archaeological crime scenes. Ancient genome data produced using massively parallel sequencing allow such estimation, albeit with significant limitations owing to the extremely low coverage of ancient genomes, frequently ranging between 0.2x - 0.01x. Estimating the kinship coefficient reliably using such sparse data is a challenge. Both likelihood-based and non-parametric tools have been recently developed to address this question, but their efficiency has not yet been systematically and comparatively studied. Here we present work where we compared the accuracy of three most commonly employed tools, using ancient genome data produced from simulated pedigrees. We studied accuracy with respect to both kinship coefficient estimation, and also Cotterman coefficients. Our results show that genotype data that include >5,000 SNPs allow close kinship coefficients to be relatively reliably estimated, although accuracy falls dramatically beyond the 3rd degree. In addition, we confirm that pedigree relationship estimation (e.g. distinguishing between parent-child pairs vs. siblings) using Cotterman coefficients is a noteworthy problem, and will require alternative, and holistic approaches to address.

---

# A novel network-centric framework for evaluating epistasis in cancer

Rafsan Ahmed, Cesim Erten, Hilal Kazan and Cansu Yalcin
Antalya Bilim University, Antalya, Turkey 07190

Recent cancer genome projects have revealed a widespread phenomenon that genes from the same functional pathway show mutually exclusive mutation profiles, that is simultaneous alterations of those genes in the same samples do not tend to occur. Given the importance of this observation, a number of computational approaches have been developed to detect mutual exclusivity. However, it is difficult to assess the accuracy of these methods due to the lack of ground truth and the use of different input data by each method. Here, we propose a novel network-centric framework to compare the performances of five different mutual exclusivity detection algorithms on TCGA data: DISCOVER [1], Fisher's Exact Test, MEGSA [2], MEMO [3], WeXT [4]. Assuming that cancer driver genes in the same pathway are more likely to exhibit mutual exclusivity, we utilize the interactome to systematically compare mutual exclusivity detection algorithms via unbiased statistics such as F1 score. We find that WEXT and DISCOVER perform better than the other approaches. When we insert the mutual exclusivity predictions of these methods as input to an existing cancer driver module finding method named MEXCOwalk, we observe significant improvement in the recovery of known cancer driver genes. Lastly, we show that incorporating network knowledge helps reduce the mutation load confounding problem introduced by van de Haar et. al. [5]. Such an effect makes our framework a more viable alternative to employing subtype-stratification in dissecting true epistasis relationships from among a set of pairwise significance values as computed by some appropriate statistical mutual exclusivity test. This is especially significant in settings where the subtype information for the cohort under study is not abundant.

## References


1. S. Canisius, J. W. M. Martens, and L. F. A. Wessels, "A novel independence test for somatic alterations in cancer shows that biology drives mutual exclusivity but chance explains most co-occurrence," Genome Biology, vol. 17, Dec. 2016.

2. X. Hua, P. L. Hyland, J. Huang, L. Song, B. Zhu, N. E. Caporaso, M. T. Landi, N. Chatterjee, and J. Shi, "MEGSA: A powerful and flexible framework for analyzing mutual exclusivity of tumor mutations," The American Journal of Human Genetics, vol. 98, pp. 442–455, Mar. 2016.

3. G. Ciriello, E. Cerami, C. Sander, and N. Schultz, "Mutual exclusivity analysis identifies oncogenic network modules," Genome Research, vol. 22, pp. 398–406, Sept. 2012.

4. M. D. Leiserson, M. A. Reyna, and B. J. Raphael, "A weighted exact test for mutually exclusive mutations in cancer," Bioinformatics, vol. 32, pp. i736–i745, Sept. 2016.

5. J. van de Haar, S. Canisius, M. K. Yu, E. E. Voest, L. F. Wessels, and T. Ideker, "Identifying epistasis in cancer genomes: A delicate affair," Cell, vol. 177, pp. 1375–1383, May 2019.

# Cardiac atrial transcriptomic landscaping reveals defects in various pathways in patients with ischemic heart disease or heart failure

Arda Eskin[1], Severi Mulari[2], Nurcan Tunçbağ[1] and Esko Kankuri[2]

[1]Department of Health Informatics, Graduate School of Informatics, Middle East Technical University, Ankara, Turkey 06800

[2]Department of Pharmacology, University of Helsinki, Helsinki, Finland 00014

Ischemic heart disease (IHD), causing high morbidity and mortality, continues to be the leading cause of death worldwide. In this study, samples of right atrial appendage were collected for transcriptomic profiling from 40 patients with IHD undergoing elective coronary artery bypass grafting (CABG) surgery. Additionally, 8 samples from patients with solitary valvular disease undergoing corrective valvular surgery were harvested to serve as controls. Clinical and follow-up data including medication, laboratory measurements are also collected for each patient. We obtained the transcriptomic data of healthy right atrial appendage tissue from GTEx (n = 429). Our aim in this study is to find novel associations and genes related to IHD and to predict the risk of having IHD by integrating transcriptomic, clinical and interactome data. We found 357 upregulated and 310 downregulated DEGs in IHD samples compared to healthy tissues ($FDR < 0.05$ and $|logFC| > 2$). Among these, genes from protocadherin gamma subfamily were found to be significantly different between patient group who has an ejection fraction lower than 55% which represents the percentage of blood leaving the heart each time it contracts. ($p - value < 0.05$). We inferred the most critical pathways from the list of DEGs and found that agrin interactions at neuromuscular junction, epithelial adherens junction signaling, sirtuin signaling and oxidative phosphorylation are significantly enriched. DEGs associated with oxidative phosphorylation are downregulated. Additionally, functional analysis of miRNAs and their targets that have significantly different expression values between patient groups, resulted with the enrichment of lipid metabolism. Overall, our results provide a transcriptome level understanding into processes reactive to IHD and the association of gene level data to phenotypic information.

## Dynamics of the homotrimeric TolC transmembrane protein

Isik Kantarcioglu, Ali Rana Atilgan and Canan Atilgan
Faculty of Engineering and Natural Sciences, Sabancı University, İstanbul, Turkey 34956

Bacteria produce bacteriocins to inhibit activity of other bacterial strains. The TolC protein on the outer membrane of gram-negative bacteria prevents bacteriocin and viral phage entrance to bacteria and enables efflux of bile salts and antibiotics. TolC protein is a homotrimer, each monomer containing 428 amino acid residues [1]. The protein complex is inserted in the outer membrane and extends into the periplasmic space. Open and closed conformations of the TolC conduct survival and death mechanism. In this study, we aim to probe the dynamics of TolC in the membrane environment via 100 ns long molecular dynamics simulations. We observe that the extracellular loops and periplasmic helices of the protein are highly fluctuating which may have important roles in the uptake and efflux mechanisms. To determine additional allosteric residues that reside in other parts of the protein complex, we utilize the perturbation response scanning (PRS) method [2] developed to relate forces acting on select locations to experimentally identified conformational changes. We discuss the role of residues residing on extracellular loops and periplasmic helices in manipulating the functioning of TolC. Findings are related to function altering mutants determined via deep mutagenesis scanning of this protein.

**References:**

1. Koronakis, V.; Sharff, A.; Koronakis, E.; Luisi, B.; Hughes, C., Crystal structure of the bacterial membrane protein TolC central to multidrug efflux and protein export. Nature 2000, 405 (6789), 914-919

2. Atilgan, C.; Atilgan, A. R., Perturbation-Response Scanning Reveals Ligand Entry-Exit Mechanisms of Ferric Binding Protein. Plos Comput Biol 2009, 5 (10).

---

# Identifying potential drug treatments for COVID-19 with a deep learning model

Alperen Bağ[1], Berk Atıl[1], Rıza Özçelik[1], Elif Özkırımlı[2] and Arzucan Özgür[1]

[1]Department of Computer Engineering and [2]Department of Chemical Engineering, Bogazici University, Istanbul, Turkey 34342

SARS-CoV-2 infected millions of people around the world and had severe implications. However, an effective treatment is yet to be found. To accelerate drug discovery and to identify lead molecules, drug repurposing is an attractive approach. In this study, we use DeepDTA (Öztürk et. al.), a state-of-the-art deep drug-target affinity pre- diction model that represents ligands by their SMILES strings and proteins by their amino-acid sequences, to virtually scan the existing drugs to suggest drug candidates for the treatment of the coronavirus disease (COVID-19).

We first trained DeepDTA with $\approx$ 510K interactions from BindingDB (Liu, T. et. al.), for which Ki or Kd values are reported. Then, we manually identified 13 human proteins that have been suggested as targets for COVID-19 treatment in the literature and down- loaded their sequences from UniProt (Apweiler et. al.). We screened the approved drugs using their SMILES representations and used DeepDTA to identify the drugs that can interact with any of the selected targets.

Our results showed that drugs that have already been reported to inhibit SARS-CoV-2 proteins in the literature, are predicted by DeepDTA to interact with the targets that we examined. These results suggest that our DeepDTA based approach is promising for identifying COVID-19 drugs. We will present some of the drugs identified by DeepDTA but not presented in previous literature.

# Predicting capsid and tail proteins in bacteriophage genomes by using deep learning

Rıdvan Cakci[1], Yeşim Aydın Son[1] and Arif Yilmaz[2]

[1]Department of Health Informatics, Graduate School of Informatics, Middle East Technical University, Ankara, Turkey 06800

[2]TUBITAK Space Technologies Research Institute, Ankara, Turkey 06800

Bacteriophages (BPs) are a group of viruses that can infect and kill bacteria without damaging human or animal cells. They are the most widespread biological substances that can be found in soil, seawater, and extreme environments. Phages are getting increasing attention primarily due to having important implications in dealing with antibiotic resistance in bacteria where phages can be introduced to infect and kill pathogenic bacteria. The capsid and tail proteins are the main structural proteins of phages. These proteins are also considered as their signature since they exist only in phage genomes. These protein sequences generally lack sequence conservation due to the high mutation rate in viral genomes; therefore, they are highly diverse and thus let a significant difficulty in identifying and annotating their sequences. Additionally, more than 70% of phage sequences in the viral Reference Sequence database encode proteins have unknown functions based on FASTA annotations. In our study, we explore deep neural network methods for the prediction of the unannotated viral proteins. We construct a model composed of convolutional and hidden layers composed of training and validation datasets. Here, we summarize the challenges for the prediction of the viral proteins by using deep neural network methods and present our preliminary results.

# Optimization of the HADDOCK sampling for the rapid modeling of interfacial mutations

Mehdi Koşaca, Eda Şamiloğlu, Mehmet Erguven and Ezgi Karaca
Izmir Biomedicine and Genome Center, Izmir, Turkey 35340

HADDOCK is a docking platform poised to predict diverse range of biomolecular complexes, in the form of protein-protein, protein-ligand, or protein-DNA. HADDOCK can also be used to structurally model mutations observed across complex interfaces if the structure of the wild type complex is at hand. For the interfacial mutation modeling, HADDOCK generates different conformers of the imposed mutation, the number of which is determined by the user. At the end of this sampling, the mutant models are ranked according to a linear sum of their electrostatic, van der Waals, and desolvation energies. In this work, we aim to determine the optimal number of mutation sampling parameters in HADDOCK for the rapid and accurate modeling of interfacial mutations. To this end, we have chosen the training dataset of iSEE 1 (Interface structure, evolution, and energy), a binding affinity prediction tool, which includes 1102 single point mutations, acquired from 57 protein-protein complexes. The iSEE set also contains experimental binding affinities and physicochemical characteristics of the presented mutations. Within this dataset, we have selected 20 cases as a preliminary working set, while making sure that we have covered all the mutation types. To ensure the maximum amount of sampling per case, we have generated 500 models for each complex through the HADDOCK2.2 guru interface2. After analyzing the HADDOCK energies calculated for the complexes, we have checked the sampling point where the model with the minimum HADDOCK score is generated. This approach has shown that sampling of 250 models per case will be optimal to produce the lowest energy model for most of the cases. Currently, we are exploring the relationship between the binding affinity change ($\Delta\Delta G$) and the minimum HADDOCK score obtained for our use cases. We hope to expand our study to set a baseline for the accurate structural modeling of disease-causing mutations.

# Modeling the tumor specific network rewiring through aternative isoforms of proteins

Habibe Cansu Demirel and Nurcan Tuncbag

Department of Health Informatics, Graduate School of Informatics, Middle East Technical University, Ankara, Turkey 06800

Alternative splicing is a post-transcriptional regulation which is important for the diversity of the proteome and interactome. It enables the production of multiple proteins from a single gene with different structures. The resulting protein isoforms are important for understanding tumor driven changes in both protein and network level. In this work, we collected tumor and normal RNA-seq sequencing reads from the Cancer Genome Atlas (TCGA) and found transcript expressions with StringTie, a transcriptome assembler. Using log fold changes between each tumor sample with respect to a pool of normal samples and supporting it with another comparison with available canonical isoforms, we labelled some isoforms as dominant. As a frequently studied transcription factor, a comparison of gene expressions between two groups of samples based on the presence of TP53 dominant isoforms revealed some up and downregulation patterns. After comparing dominant isoforms with patient mutations coming from TCGA and known protein domains to reveal common patterns, we focused on the interaction losses they brought with. By mapping the missing regions on these dominant isoforms onto protein-protein, protein-DNA and protein-drug interfaces, we found potentially lost interactions. Results included interesting interactors such as cancer drug – targets and transcription factor – DNA couples. By removing lost protein-protein interactions, we constructed tumor-specific interactomes where only the interactions of expressed proteins were kept per sample. Using them, we created two sets of interaction networks by prioritizing proteins with dominant isoforms using Omics Integrator which solves the prize-collecting Steiner forest problem. Finally, we compared these two sets of networks to highlight striking changes between them while also examining all tumor specific networks simultaneously to reveal pathway, interaction and protein patterns that can cluster the tumors. The results of our analysis will contribute to the elucidation of tumor mechanisms and will help for target selection and developing therapeutic strategies.

# Exploring allosteric mechanisms of CXCR4 and implications in drug design

Tugce Inan and Ozge Kurkcuoglu Levitas

Department of Chemical Engineering, Chemical and Metallurgical Engineering Faculty, Istanbul Technical University, Istanbul, Turkey 34469

Chemokine receptors like CXCR4 and CXCR7 are crucial for autoimmune, central nervous system diseases, and cancer. Particularly, CXCR4 and its ligand SDF-1 are over-expressed in cancer cells like breast and lung cancer, ovarian, prostate, melanoma, and hematopoietic malignancies. Homodimerization and heterodimerization of CXCR4 with CXCR7 are accepted as a mechanism for modulating CXCR4 function [1]. However, dimerization of GPCRs is unclear. All of these indicate CXCR4 as a potential drug target. Here, we aim to understand the functional mechanisms of CXCR4 and to elucidate new drug binding sites for the inhibition of intracellular signaling. Thus, the Gaussian Network Model is performed to reveal its functional collective motions, and hinge residues that coordinate rigid domains. Also, allosteric communication pathways of CXCR4 are investigated using the Residue Network Model and the K Shortest Path Method, where the protein structure is described as a unidirectional weighted graph [2]. These methods are systematically applied to CXCR4 monomer, homodimer, and CXCR4/CXCR7. For each case, potential allosteric drug binding regions are revealed and suboptimal pathways between potential allosteric sites and active sites are calculated. Transmembrane domain 2 of CXCR4, suggested as a dimerization site with CXCR7, contains hub residues with high betweenness, indicating that it may be a potential drug-binding site.

**References:**

[1] Décaillot, F. M., Kazmi, M. A., Lin, Y., Ray-Saha, S., Sakmar, T. P., & Sachdev, P. (2011). The Journal of biological chemistry, 286(37), 32188–32197.

[2] Guzel, P., & Kurkcuoglu, O. (2017). Biochimica et Biophysica Acta, 1861(12), 3131–3141.

[3] Evans, A. E., Tripathi, A., LaPorte, H. M., Brueggemann, L. I., Singh, A. K., Albee, L. J., Byron, K. L., et al. (2016). International Journal of Molecular Sciences, 17(6), 1–19.

# Drug Respositioning in Colorectal Cancer by using Co-expression Networks of P-glycoprotein

Hande Beklen[1], Gizem Gulfidan[1], Kazım Yalcın Arga[1], Adil Mardınoglu[2] and Beste Turanli[3]

[1]Department of Bio-Engineering, Marmara University, Istanbul, Turkey 34722
[2]Science for Life Laboratory, KTH—Royal Institute of Technology, Stockholm, Sweden 114 28
[3]Department of Bio-Engineering, Istanbul Medeniyet University, Istanbul, Turkey 34720

Colorectal cancer (CRC) is the third most fatal type of cancer that is seen in both men and women and found in Turkey and worldwide. The high heterogeneity of the cancer leads diffuculty in explaining the biology and behavior of this cancer. Although several drugs are developed for the treatment of CRC, some patient may be resistant to these drugs. P-glycoprotein (P-gp) (MDR1) which is expressed by ABCB1 gene may cause failure of chemotherapy in several cancer types due to its high expression. MDR1 throws the drug molecules out of the cancer cells, thus drug concentration decreases and causes resistance against drugs. Therefore drug-resistant is the major problem in treatment of cancer and new strategies are needed for effective therapy. Identification of specific biomarkers and potential drug targets of repurposing drugs can be efficient strategies by using network-based approaches in drug-resistant cancers. So drug repositioning is the promising method to find new therapeutic target by using existing drugs due to its high efficiency and low cost. In this study, co-expression network of ABCB1 gene with different network sizes (50, 100, 150, 200 edges) were used to understand the drug resistance of MDR-1 in CRC by applying drug repositioning for effective treatment of CRC. Molecular docking were performed to determine potential physical interaction of the candidate drugs and MDR1protein. Besides inhibitors of ABCB1 gene were taken as a positive control in molecular docking. Also, prognostic and diagnostic features of four networks were evaluated and biological functions of the genes involved in the networks determined by functional enrichment analysis in CRC. Finally, differentially expressed genes of drug resistant (i.e oxaliplatin, methotrexate, SN38) HT29 cell lines were found and used for repurposing drugs with reversal gene expressions. As a result, all networks showed high diagnostic and prognostic performance and 7 candidate drugs were determined with high binding affinity. Candidate drugs were highly interacted with MDR1 protein when compared to its positive controls. All these results can shed light on the development of effective diagnosis, prognosis, and treatment strategies for drug resistance in CRC.

# New solutions to old problems: Mitigating data loss and bias in ancient genome data processing

Dilek Koptekin[1], Etka Yapar[2], Ekin Sağlıcan[2], Can Alkan[3] and Mehmet Somel[2]

[1]Department of Health Informatics and [2]Department of Biology, Middle East Technical University, Ankara, Turkey 06800

[3]Department of Computer Engineering, Bilkent University, Ankara, Turkey 06800

DNA in ancient samples is highly fragmented due to decay after death, has exogenous contamination and contains a low amount of endogenous DNA. Consequently, ancient DNA processing usually involves studying genome data with ¡1x coverage, composed of short reads with frequent C-to-T transitions at their ends. These create two types of challenges. One is the inability to call full diploid genotypes. Solutions include pseudo-haploidization, and genotype likelihood methods. However, it has been observed that such ancient genome data is "reference biased", i.e. contain more reference alleles than alternatives at heterozygous positions. This appears to be caused by loss of alternative allele-bearing reads due to their slightly lower mapping quality. Second challenge is to avoid confusing postmortem C-to-T transitions with authentic variation. The solution is to use variants identified in worldwide populations instead of de novo calls. Further, one may use only transversions, or both transversions and transitions but after trimming 2-10 nucleotides of read ends where postmortem damage accumulates. Unfortunately, the former approach means not using c.67% of SNP data, while the latter means losing up to 30% of data due to short read lengths. Here we propose solutions to mitigate these effects in ancient genome data preprocessing. The first addresses reference bias. We show that aligning read data to a graph genome, or aligning to a linear reference genome but after masking common polymorphic sites in the reference, effectively removes reference bias in ancient genotype data. The second involves avoiding postmortem damage effects and minimizing data loss. Here, instead of trimming read ends, we mask potential sites where the read's genotype can be affected by postmortem cytosine deamination. Our primary analysis increases genotyping by 15% especially in the lowest coverage samples without compromising accuracy, thereby significantly boosting statistical power in downstream population genetics analyses.

---

# Empowering SVM-RCE with user specified ranking function to classify gene expression data

Amhar Jabeer[1], Burcu Bakir-Gungor[1] and Malik Yousef[2]

[1]Department of Computer Engineering, Faculty of Engineering, Abdullah Gul University, Kayseri, Turkey 38080

[2]Galilee Digital Health Research Center (GDH), Zefat Academic College, Safed, İsrail 1320611

The algorithms that are developed for classifying gene expression data is mainly based on statistical models and they mostly disregard the focus of gene expression. One of the algorithms that has piqued extensive interest in the bioinformatics community is SVM – RCE (Support Vector Machine - Recursive Cluster Elimination) since it performed significant improvement over its competitors. SVM-RCE combines a clustering method that is used to identify correlated gene clusters, with an SVM classifier that is used to identify and score (rank) those gene clusters for the purpose of classification. Here, we have extended this algorithm by integrating a user specified ranking function. While the original SVM-RCE algorithm uses accuracy measure to rank clusters for elimination, this new version makes use of different performance metrics such as accuracy, sensitivity, specificity, f-measurement, area under curve and precision as weights. These weights correspond to which metric should be considered as higher priority during the ranking stage of the algorithm. We have tested our algorithm on 10 different Gene Expression Omnibus datasets with 6 different user specified ranking functions. We have found that by using a ranking function, we can improve the results of SVM-RCE. We have tested the influence of the ranking function with different values. The results show that a significant impact accrued, reaching improvement of about 16% in some cases over the standard rank used in the original version. Finally, we have implemented our algorithm in KNIME, which allows anyone to use it without going through a steep learning curve and hence, makes our tool a user-friendly tool.

---

# Computational prediction of the activity of metabolic reactions in Alzheimer's Disease using personalized metabolic network models

Hatice Büşra Lüleci[1], Vijay R Varma[2], Anup M. Oommen[3], Sudhir Varma[4], Jackson A. Robert[2], Madhav Thambisetty[2] and Tunahan Çakır[1]

[1]Department of Bioengineering, Gebze Technical University, Kocaeli, Turkey 41400
[2]National Institute on Aging (NIA),National Institutes of Health (NIH), Baltimore, Maryland
[3]Department of Mechanical and Biomedical Engineering, National University of Ireland Galway, Galway, Ireland
[4]HiThru Analytics, Princeton, NJ

Alzheimer's disease (AD) is a type of dementia that causes impairment in memory, reasoning and thinking. Progression of AD is related to the cholesterol metabolism in the brain. Our aim is to predict changes in the activity of the reactions associated with cholesterol and bile acid metabolisms for four different brain regions, which are Hippocampus (HIP), Entorhinal cortex (EC), Posterior cingulate (PC) and Primary visual cortex (VCX) by using an optimization algorithm that predicts activation of the metabolic reactions based on the gene expression level of healthy people and AD patients. The most recent version of human genome-scale metabolic network, Human-GEM 1.3.2 was used to create personalized metabolic networks for all samples from HIP, EC, PC and VCX regions of GSE5281 and GSE48350 datasets retrieved from Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/). The metabolic network includes 13417 reactions associated with 3628 genes. The RMA-normalized and sex- and age- corrected transcriptomic data were mapped on Human-GEM network separately for each control and disease sample using Integrative Metabolic Analysis Tool (iMAT), leading to personalized genome-scale metabolic networks. These personalized metabolic networks spanned the range between 6801-7896 reactions associated with 2654-2867 genes. Different number of reactions and genes in the personalized metabolic networks indicate the inherent heterogeneity of control and AD samples, justifying our personalized approach. The personalized networks were compared between AD and control samples to identify cholesterol-associated reactions significantly associated with AD or control. iMAT predicts a set of relevant reactions to be significantly affected in AD despite the lack of any AD-specific constraints in our analysis. The results show significantly inactive de novo cholesterol biosynthesis in AD for both EC and HIP. A more pronounced inactivation of the reactions was specifically predicted in pre-squalene pathway of HIP whereas the post-squalene pathway was more significantly affected in EC. Mapping transcriptome data on genome-scale metabolic networks to predict condition-specific activity of reactions give us crucial perspective about changes in cholesterol metabolism in AD.

## Ligand switching mutations in PDZ domain explained by centrality of amino acids

Tandac Guclu, Canan Atilgan and Ali Rana Atilgan
Faculty of Engineering and Natural Sciences, Sabancı University, İstanbul, Turkey 34956

Mutations occasionally affect protein structure and/or function, and these changes are important alterations in ligand specificity that may have significant consequences, such as emergence of antibiotic resistance or disruption in cell signaling. Here we study PDZ3 domain which has an important role in mammal neural cell signaling. PDZ domains construct the PSD-95 complex by binding CRIPT (ligand I) and T-2F (ligand II) ligands. Previously, its specific mutations have been demonstrated to display preferred ligand specificity: Wild-type(WT) protein has higher binding affinity to ligand I and G330T mutation binds to both ligands I/II while the H372A mutation and the G330T-H372A double-mutation(DM) tend to bind only to ligand II. To scrutinize the emergent structural features due to the mutations, we conducted network analyses on the snapshots from the 400-ns long molecular dynamics simulations. Then, we utilized betweenness centrality (BC) to find the nodes which act as hubs for information communication in biological function. $\Delta$BC results show that the N-terminus has an impact on the formation of H372AL2. Furthermore, we employed Girvan-Newman algorithm to investigate the modularity of PDZ3 protein. The results indicate that N and C termini of the structure are in the same community, while N-terminus and the ligand tend to be located in the same community only in favorable WT and the single mutation cases. We explain how the changes of the residue centralities by perturbations introduced in the form of mutations lead to the ligand switching behavior in the PDZ domain, and discuss why this behavior is governed by N-terminus region.

---

# Identification of signature pathways and reactions using human metabolic model and transcriptome data for subtyping of lung cancer

Ezgi Tanıl and Emrah Nikerel

Genetics and Bioengineering Department, Yeditepe University, Istanbul, Turkey 34755

Genome scale metabolic models became increasingly popular computational tool to predict intracellular fluxes, allowing simulation of different metabolic scenarios, strain design within industrial biotechnology context as well as predicting the disease states allowing a gateway to the personalized medicine. These models are typically underdetermined sets of algebraic equations, each representing a biochemical reaction, encompassing thereby available information on the metabolism. Within the fluxome analysis framework, typically infinite solution space for the above equation set needs to be constrained by either integrating various –omic data (thereby generating tissue/patient/disease specific model) and/or optimizing for a given objective, e.g. maximum growth. Lung cancer (LC) is the cause of more than 1 million cancer-related deaths in the world. Patient specific treatment strategies heavily rely on correct subtyping and staging while due to the complexity of oncogenic mechanisms, available biomarkers at genome (e.g. PIK3CA, TP53, PTEN, KRAS, ALK, EGFR genes etc.) or proteome level (e.g. haptoglobin $\beta$chain, serum amyloid A), are far from being LC specific. Cancer cells are known to have altered metabolism. For instance, reactive oxygen species (ROS) were reported to have both tumor-promoting and tumor-inhibiting properties. For example, during tumor progression (and also initiation) antioxidant pathways, involving several metabolic and non-metabolic enzymes, were reported to be upregulated [1]. These enzymes, reactions and pathways are key in subtyping such that the fluxome analysis presents a key step in subtyping LC. Aim of this study is to identify signature pathways and reactions for subtyping LC in (i) adenoma and adenocarcinoma and (ii) squamous cell neoplasms. This will be achieved via the differential analysis of flux distribution for each subtype, calculated using human genome scale metabolic model (Recon3) with available, a priori subtype-labelled transcriptome data integrated.

**References:**

[1] Gorrini et al., Modulation of oxidative stress as an anticancer strategy. Nat Rev Drug Discov. 2013 Dec;12(12):931-47.

---

# Exploring orthogonal gene expression for the identification of signature genes for subtyping Glioblastoma Multiforme

Nehir Kızılilsoley and Emrah Nikerel

Genetics and Bioengineering Department, Yeditepe University, Istanbul, Turkey 34755

Glioblastoma multiforme (GBM) is the most malignant form of malignant glioma with extreme rates of mortality and recurrence [1, 2]. Due to its genetic (in)stability, morphological and genetic heterogeneity, and unpredictable clinical course and heterogeneity in their transcriptional profile, characterization of GBM and its subtypes (classical, mesenchymal, neural and proneural) is highly challenging [2-4]. This in turn, hampers the development of advanced, personalized treatment strategies and precise predictions on disease evolution and prognosis. Since transcriptome profiles in GBM cases differ vastly and non-linearly among subtypes, searching signature gene sets that would allow differentiating these subtypes is of great interest. Such signature genes would preferably be orthogonal in their expression, i.e. differentially expressed only in one single subtype and should be easy to detect using either high or low throughput methods. The non-linear nature of the resulting transcriptome data calls for machine learning techniques for (supervised) classification methods. By employing large-scale datasets from literature and databases and a set of ML methods, we aim to generate a computational pipeline that would result in such a signature gene set for subtyping GBM. The offered pipeline, starts by analysis of differential expression, normalization of the gene expressions across samples, checking the (near) orthogonality for gene expression vectors for each subtype with respect to each other and classification using ANN or SVM. By using a greedy approach, we start with a small set and iteratively enlarge signature gene sets using the above described pipeline. The resulting signature gene set is also compared with literature information. We conclude that a priori search for orthogonal gene sets improves the prediction accuracy of the constructed ANN or SVM. The offered pipeline can readily be adapted and various related datasets.

---

# miRcorrNet: Machine learning based integration of miRNA and mRNA expression profiles for classification and detecting targets

Gokhan Goy[1], Burcu Bakir-Gungor[1] and Malik Yousef[2]

[1]Department of Computer Engineering, Faculty of Engineering, Abdullah Gul University, Kayseri, Turkey 38080

[2]Galilee Digital Health Research Center (GDH), Zefat Academic College, Safed, İsrail 1320611

In the field of gene expression due to recent technology one is able to obtain mRNA gene expressions and miRNA gene expressions. Most studies consider only mRNA or microRNAs expression data to investigate these mechanisms. However, understanding the complex structures of complex diseases using one type of omics data poses challenges. Most of the approaches that integrate miRNA and mRNA are based on statistical methods, such as Pearson correlation, combined with enrichment analysis. To the best of our knowledge, there are two existing tools that serve the researchers for analyzing microRNA and gene expression profiles simultaneously. The other studies just use different packages to perform this task. In this study, we developed a novel tool called miRcorrNet, which performs machine learning-based integration for the analysis of miRNA and mRNA gene expression profiles. miRcorrNet groups mRNA genes based on their correlation to the miRNA expressions. Then these groups are subject to a rank function, which estimates its contribution to the classification task. We have tested our tool on the miRNA-seq and mRNA-seq expression profiles that we have downloaded from TCGA data portal for 11 solid tumor types. In our experiments, we have reported the average performance measures of 100-fold Monte Carlo Cross-Validation. Additionally, we have comparatively evaluated miRcorrNet with maTE and SVM-RFE. The performance results show that the tool is working as good as other tools in terms of accuracy measurements, reaching over 95% AUC values. Moreover, a deep biological analysis shows that the results are very meaningful in term of their association to cancer. We hope that our tool will serve to the bioinformatics community in order to more precisely identify the target genes for each microRNA using microRNA and gene expression profiles simultaneously.

---

# ProNetView-ccRCC: An interactive visual exploration portal for clear cell renal cell carcinoma proteogenomics networks

Selim Kalayci, Francesca Petralia, Pei Wang and Zeynep H. Gümüş
Icahn School of Medicine at Mount Sinai, New York, NY 10029

Recent advances in analytic technologies are enabling the large-scale integrative proteomic and genomic (proteogenomic) study of cancers. The National Cancer Institute's Clinical Proteomics Tumor Analysis Consortium (CPTAC) has so far performed comprehensive proteogenomic characterizations of several cancers, including those of the breast, colon, rectum, ovaries, and kidneys, while the analysis of several other cancer types is in progress. To identify drivers of cancer progression from these massive multi-omics CPTAC datasets, our team is characterizing co-expression patterns between protein pairs using machine-learning algorithms. We then represent these co-expressed pairs as networks to visually communicate the results. However, in order to better infer the network, our algorithms increasingly borrow information from multiple types of massive -omics datasets generated by CPTAC on the same tumor samples, leading to large and complex networks that are challenging to visually communicate. Thus, to better understand and interpret the multi-omics data represented by these networks, effective visual exploration tools are needed. Here, we introduce ProNetView-ccRCC, an interactive web-based portal to explore a phosphopeptide co-expression network we have inferred using random forest based network construction. Specifically, the network characterizes the co-expression patterns from 20,976 tumor phosphopeptides in 103 participants. The inferred ccRCC tumor phosphoprotein co-expression network is comprised of 3,614 genes, which participate in 30 functional pathway-enriched network modules. ProNetView-ccRCC en-ables quick, user-intuitive visual interactions for users to explore the network and conveniently query for association between abundance of each phosphopeptide in the network and clinical variables (e.g. tumor grade). We anticipate that ProNetView-ccRCC will facilitate researchers to conduct their own analyses on the rich CPTAC ccRCC network data, and share their results. ProNetView-ccRCC and associated net-works are available at http://ccrcc.cptac-network-view.org/.

# Detection of sequencing error patterns of Illumina sequencing by deep learning based natural language processing algorithms

Emre Taylan Duman and Pınar Pir

Department of Bioengineering, Gebze Technical University, Kocaeli, Turkey 41400

Genome sequencing is the one of the key technologies that can reveal the power of the personalized medicine. Sequencing by synthesis is one of the mostly used approach to extract information of the genetic material. Working principle of the approach relies on the detection of the radiations by the sensor that is triggered by the laser inside of the sequencing devices of Illumina. The output of sequencing technologies is not error free and few factors are known to affect the error rates. The samples are prepared by using TruSeq library preparation kits provided by the Illumina, TruSeq kits are considered as one of the sources of sequencing errors by favoring G nucleotides. Also, it is shown that, there are certain patterns that cause incorrect basecall in sequencing devices. Further, reduced quality scores in reads, cause poor mapping coverage in the important gene sections and reduced mapping coverage may lead to incorrect SNP discoveries. In this project, we aim to better understand the quality variations on some k-mers in Illumina-sequenced data. Deep learning-based algorithms will be used to find important patterns that cause quality score fluctuations in the individual 3'mer or more k-mers. Natural language processing (NLP) techniques, mostly used for language pattern identification and developing models from the sequential data as text or speech. Popular NLP models as BERT and GPT are also prepared and trained for fastq files. Representation of the expression data by implementation of the quality scores, would potentially reveal the reasons behind quality score drop in the central locations of the sequence reads.

---

# Evolutionary Analysis reveals Unique Features of Frizzled 4 receptor

Burak Islek and Ogun Adebali

Molecular Biology, Genetics and Bioengineering Program, Faculty of Engineering and Natural Sciences, Sabancı University, İstanbul, Turkey 34956

G protein-coupled receptors (GPCRs) are classified into five subgroups by GRAFS classification (A to F). The subfamily-specific features of more than 800 GPCRs are yet to be comprehensively established. In this study, we aim to understand the evolution, function, interaction, and significance of F class GPCRs (also termed Frizzled) through comparative genomics. We performed BLAST using the sequences of all frizzled family members in humans (n=10) against complete proteomes of the entire eukaryotic lineage. Phylogenetic analysis and sequence homology clustering demonstrate that Frizzled proteins consist of the following 5 subgroups: FZD1, FZD2, FZD7; FZD3, FZD6; FZD4, FZD9, FZD10 and FZD5, FZD8. BLAST results were grouped in this order and complete sequences of their intersected targets were retrieved from UniProt database. Multiple sequence alignment was performed using MAFFT. The clade consisting of FZD4, FZD9, and FZD10 was investigated through their domains, secondary structures, and subfamily specific residues using Spial web-server. In the light of initial results, we identified a region with 16 specific residues on the FZD4 extracellular cysteine-rich domain (CRD; aa 40-161). Specific residues on the FZD4-CRD region are clustered prominently different from the other members of this clade. Two residues were reported as binding sites of Norrin and FZD4 before (FZD4 K109 and M157), and some mutations (M105V and M157V) have been associated with Familial Exudative Vitreoretinopathy (FEVR). Because FZD4 is the only Frizzled family member that can interacts with Norrins and plays a role in the regulation of the retinal vascular development, these 16 residues may have a potential role in the FZD4-Dishevelled protein interaction and WNT signaling. Further functional studies of FZD4 are required to assess whether a change in these residues can plays any functional role in FEVR.

---

## Investigating sex specific molecular differences in female and male patient groups of bladder cancer

Emine Ezel Cilek

Biotechnology Institute, Ankara University, Ankara, Turkey 06135

Bladder cancer is one of the most common cancers that occur in men more often than it does in women. It begins with fundamental abnormalities in urothelial cells, which may result loss of cell differentiation and eventually last with tumor formation. Exploring sex based differences in molecular mechanisms may provide knowledge of distinct molecular sub-types and lead to personalized research of bladder cancer. In this study, we compared mRNA expression, mutated genes and copy-number alterations in female of male groups by using TCGA bladder cancer data (Cell,2017) in cBioPortal platform. Our brief analyses showed the differential expression and involvement of sexually dimorphic genes such as DDX3, YKDM5D, EIF1AY in male groups whereas TSIX, XIST in female groups. We suggested that this study might help to understand sex disparities in bladder cancer, which is observed more aggressive in female groups and help to gain insight into distinct molecular pathogenesis in two groups.

---

# Integrative network modelling of drug responses for revealing mechanism of action

Seyma Unsal Beyge and Nurcan Tuncbag

Department of Health Informatics, Graduate School of Informatics, Middle East Technical University, Ankara, Turkey 06800

Cancer is the second leading cause of death globally (WHO, 2015), and expected to rise 70% in two decades. Molecular heterogeneity across tumors and within the tumors makes the treatment and diagnosis difficult. Since cancer is a complex disease, a multi-stage transformation of a normal cell into malignant cells, network of interacting proteins and genes are involved in disease progression. Revealing potential disease networks becomes the important and challenging step for designing treatment strategies for cancer. In this study, our aim is to stratify the effectiveness of the drugs at network level beyond the list of altered molecules and to reveal signaling network models by using reverse engineering principles. Thus, it will be possible to examine different responses to the same drug molecule in different cancer types at the level of pathways. For this purpose, we used omic dataset in the P100 collection of CMap (ConnectivityMap). The P100 collection consists of both P100 (phosphoproteomic) and L1000(transcriptomic) data for 6 different core cell types, treated with the same 90 drugs (Abelin et al., 2016)(Lamb et al., 2006). We collected drug-target interactions from DrugBank & PubChem. We applied a network reconstruction approach that combines link-prediction algorithm with the solution of prize-collecting Steiner tree problem so that at both node level and edge level we find an optimal network solution for each cell line and drug pairs. At the reconstruction step, each network is oriented from the targets of each drugs. All pair distance-based network comparison of each cell line and drug pair revealed that some drugs has a significant similarity at network level despite having totally different targets. One example is lenalidomide and IKK-inhibitor drug pairs. Both drugs have impact PIK3-AKT and Ras signaling pathways, but they differentiate in the former additionally impacts Rap1 signaling and the latter has impact on proteolysis.

# Immunoinformatic prediction of regionally-specific candidate epitopes in SARS-CoV-2 proteins based on updated South American HLA frequencies

David Requena[1], Aldhair Medico[2], Ruy D. Chacón[3], Manuel Ramírez[4] and Obert Marín-Sánchez[4]

[1]Laboratory of Cellular Biophysics (Simon Lab) - The Rockefeller University, New York, NY 10065

[2]Laboratorio de Bioinformática, Biología Molecular y Desarrollos Tecnológicos, Laboratorios de Investigación y Desarrollo, Facultad de Ciencias y Filosofía, Universidad Peruana Cayetano Heredia, Lima, Peru

[3]Departamento de Patologia, Faculdade de Medicina Veterinária e Zootecnia, Programa Interunidades em Biotecnologia, Universidade de São Paulo, São Paulo, Brazil

[4]Departamento Académico de Microbiología Médica, Facultad de Medicina, Universidad Nacional Mayor de San Marcos, Lima, Peru

The COVID-19 pandemic is caused by SARS-CoV-2. South American countries have been severely affected, reporting more than 6.4 million cases and 205,000 deaths as of September 2020. Efficient epitope recognition mediated by the HLA system is a key element of the cellular and humoral immunity against the disease. Therefore, global control efforts include epitope-based vaccines and immunodiagnostic tests. Thus, information about the HLA frequencies and its regional distribution is crucial, being the Allele Frequency Net Database (AFNDB) the most important repository. Nevertheless, this database does not actively collects data from scientific literature, underrepresenting South America. We have complemented this information with an extensive collection of datasets from scientific articles in PubMed, resulting in more than 12 million new datapoints. We collected studies with HLA frequencies at 4-digit resolution, matching nomenclature and technology. We calculated weighted average frequencies by country, generating the first South American integrated map of HLA allelic frequencies. Then, linear T epitopes were predicted in SARS-CoV-2 proteins for the most abundant HLA-I and II alleles in South America (frequency$\geq$5%). We used machine-learning-based prediction software: NetMHCpan-v4.0 and MHCflurry-v1.6.0 for HLA-I, and NetMHCIIpan-v4.0 for HLA-II. Predicted epitopes were filtered by its binding score to South American alleles, conservation, and structural accessibility (for Class-II only). Candidate epitopes (27 for HLA-I and 34 for HLA-II) epitopes were selected. Fourteen HLA-I and four HLA-II candidates had experimental evidence reported in other coronaviruses, while the other candidates represent novel candidates with even better scores. Remarkably, two Class-II candidates covered all South American alleles. Recent similar studies have presented candidate epitopes in SARS-CoV-2 extrapolating experimentally-determined epitopes from SARS by sequence similarity. These approaches selected potential epitopes for worldwide coverage. Nevertheless, we show that these candidates have deficient coverage for South America, whereas our candidates provide 100% coverage.