



THE 19TH INTERNATIONAL CONFERENCE ON

bioinformatics

NOV 25 - 29, 2020, VIRTUALLY, WHEREVER YOU ARE

“Bioinformatics and the Translation of Data-Driven Discoveries”

Conference Proceedings



**Proceedings of the 19th International Conference On Bioinformatics
(InCoB)
Virtual Conference, 25 – 29 November 2020**

Edited by:

Mohammad Asif Khan
Hilyatuz Zahroh

Chong Li Chuin
Hilal Hekimoğlu

Scientific Committee:

Brian Chen
Chia-Lang Hsu
Chih (Steve) Lee
Chinh SU Tran To
Christian Schoenbach
Daniele Santoni
Durai Sundar
Francesco Pappalardo
Guanglan Zhang
H.A.Nagarajaram
Hiroshi Mamitsuka
Jiajie Peng
Jie Li
Jinyan Li
Jonathan Hoyin Chan
Juanying Xie
Khang Tsung Fei

Kurochkin Igor
Kwoh Chee Keong
Lin Gao
Michael Gromiha
Ming Chen
Mohd Shahir Shamsir Omar
Paul Kennedy
Peng Chen
Puneet kumar Singh
Ruiting Lan
Sakshi Piplani
Shandar Ahmad
Shanfeng Zhu
Shinji Kondo
Shuigeng Zhou
Sorayya Malek
Susumu Goto

Tao Liu
Tatsuya Akutsu
Tetsuo Shibuya
Vikash Fang-Rong Hsu
Vladimir Brusic
Wentian Li
Wing Kai Hon
Wing-Kin Sung
Y-h. Taguchi
Yinglei Lai
Yingqiu Xie
Yongqun “Oliver” He
Yun Zheng
Zhao Liang
Zhongming Zhao

Editorial Office:

Asia Pacific Bioinformatics Network Limited
101 Cecil Street,
#25-04, Tong Eng Building,
Singapore 069533

Published, 2021

<https://incob.apbionet.org/incob20/>

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned. Nothing from this publication may be translated, reproduced, stored in a computerized system or published in any form or in any manner, including, but not limited to electronic, mechanical, reprographic or photographic, without prior written permission from the publisher.

The individual contributions in this publication and any liabilities arising from them remain the responsibility of the authors.

The publisher is not responsible for possible damages, which could be a result of content derived from this publication.

©Copyright 2021 by the International Conference on Bioinformatics (InCoB).

The individual abstracts remain the intellectual properties of the contributors.

Table of Contents

Preface	IX
About APBioNET	X
About InCoB	X
Keynote Talks	1
SARS-CoV-2: What we have learned so far	1
Genomes - from personal to populations and back	2
Global activities in Bioinformatics training and education	3
Computational methods for trans-omics and single cells	4
Mutational signatures: What caused the mutations in these cancers? Why do we care?	5
Omics resources and tools in under-represented animal models: A bird's eye view	6
Single-cell data analytics: Asian immune diversity and cancer cell states	7
Highlight	8
Comparative genome analysis provides shreds of molecular evidence for reclassification of <i>Leuconostoc mesenteroides</i> MTCC 10508 as a strain of <i>Leu. suionicum</i>	8
Oral Presentations	9
Inference of phosphopeptide binding affinity from 14-3-3s by QSAR-based prediction	9
Integrated regulatory network based on lncRNA-miRNA-mRNA-TF reveals key genes and sub-networks associated with dilated cardiomyopathy	10
The regulation of microRNA in each of cancer stage from two different ethnicities as potential biomarker for breast cancer	11
Convolutional neural networks with image representation of amino acid sequences for protein function prediction	12
Immunogenicity and structural efficacy of P41 of <i>Plasmodium sp.</i> as potential cross-species blood-stage malaria vaccine	13
Structural and immunogenicity analysis of reconstructed ancestral and consensus P48/45 for cross-species anti malaria transmission-blocking vaccine	14
Discovery of new inhibitor for the protein arginine deiminase type 4 (PAD4) by rational design of α -Enolase-derived peptides	15
Bipartite molecular approach for species delimitation and resolving cryptic speciation of <i>Exobasidium vexans</i> within the <i>Exobasidium</i> genus	16
Sub-structure-based screening and molecular docking studies of potential enteroviruses inhibitors	17
OriC-ENS: A sequence-based ensemble classifier for predicting origin of replication in <i>S. cerevisiae</i>	18
	IV

<i>In silico</i> design of potent inhibitor to hamper the interaction between HIV-1 integrase and LEDGF/p75 interaction using e-pharmacophore modelling, virtual screening, molecular docking and dynamics simulations	19
Discovery of network motifs based on induced subgraphs using a dynamic expansion tree	20
Prediction of protein-protein interaction between human and <i>Streptococcus pneumoniae</i> using logistic regression	21
Validation of predicted novel Myc motifs mediating important PPIs using computational approaches	22
Exploration the underlying mechanism of a traditional Chinese medicine formula Youdujing ointment for cervical cancer treatment	23
DNA methylation profiling reveals new potential subtype-specific gene markers for early-stage renal cell carcinoma in Caucasian population	24
Study of COVID-19 epidemic in India with SEIRD model	25
<i>In silico</i> study of race- and cancer stage-specific DNA methylation pattern in breast cancer patients based on TCGA dataset	26
A semi-supervised deep learning approach for predicting the functional effects of genomic non-coding variations	27
Comparative analysis and prediction of nucleosome positioning using integrative feature representation and machine learning algorithms	28
A computational model of mutual antagonism in the mechano-signaling network of RhoA and nitric oxide	29
Identifying genomic islands with deep neural networks	30
Mendelian randomization studies of brain MRI yield insights into the pathogenesis of neuropsychiatric disorders	31
Dissection of genetic association of anorexia nervosa and obsessive-compulsive disorder at network and cellular levels	32
Identification of copy number polymorphisms associated with early trauma in obsessive-compulsive disorder	33
MicroRNA profiles in sorghum exposed to individual or combined abiotic stresses	34
Characterizing promoter and enhancer sequences by a deep learning method	36
The evolutionary landscape of long non-coding RNA in green plants	37
Single-cell sequencing data analysis with dimension reduction based on robust and sensitive genes	38
Rapid screening and identification of viral pathogens in metagenomic data	39
Transcriptional dynamics of transposable elements when converting fibroblast cells of <i>Macaca mulatta</i> to neuroepithelial stem cells	40
Epigenetic interplay between methylation and miRNA in bladder cancer: Focus on isoform expression	41
Instance-based error correction for short reads of disease-associated genes	42
Data integration and evolutionary analysis of long non-coding RNAs in 25 flowering plants	43
SPECTRA – A tool for enhanced brain wave signal recognition	44
Boosting scRNA-seq data clustering by cluster-aware feature weighting	45

Forecasting the spread of COVID-19 using LSTM network	46
PIKE-R2P: Protein-protein interaction network-based knowledge embedding with graph neural network for single-cell RNA to protein prediction	47
BREC: An R package/Shiny app for automatically identifying heterochromatin boundaries and estimating local recombination rates along chromosomes	48
Identification of cell states using super-enhancer RNA	49
Temporal expression study of miRNAs in crown tissues of winter wheat grown under natural growth condition	50
Ontology-based annotation, modeling, and analysis of phenotypes and comorbidities in COVID-19 patients	52
Development of the International Classification of Diseases Ontology (ICDO) and its application for COVID-19 diagnostic data analysis	53
A multi-task CNN learning model for taxonomic assignment of human viruses	54
Analysis of StAR-related lipid transfer (START) domains across the rice pangenome reveals how ontogeny recapitulated selection pressures during rice domestication	55
Feature selection for topological proximity prediction of single-cell transcriptomic profiles in <i>Drosophila</i> embryo using Genetic Algorithm	56
PupStruct: Prediction of pupylated lysine residues using structural properties of amino acids	57
RAM-PGK: Prediction of lysine phosphoglycerylation based on residue adjacency matrix	58
Shared ancestry of core-histone subunits and non-histone plant proteins containing the histone fold motif (hfm)	60
Unsupervised tensor decomposition-based method to extract candidate transcription factors as histone modification bookmarks in post-mitotic transcriptional reactivation	61
Demos	62
NetVA: An R package for network vulnerability analysis	62
Understanding polygenic disease with BitEpi and EpiExplorer	63
BacEffluxPred: A two-tier system to predict and categorize bacterial efflux mediated antibiotic resistance proteins	64
Automated identification of SNP-genotypes in genomic datasets: SNiPSoL - An application to <i>Mycobacterium leprae</i>	65
Lightning Talks	66
Variants profiling of BRCA 1/2 genes through next-generation sequencing in young women with breast cancer	66
Galaxy Australia – A key partner in the global rapid response to the COVID-19 pandemic	67
APBioNetTalks: A platform to share bioinformatics talks, tutorials and trainings	68
Genetic ancestry in the hunt for disease genes and the fight against COVID-19	69

Workshops	70
Machine learning approaches for ascertaining transcriptomics data using T-bioinfo & code playground	70
Galaxy – A platform for life science analyses (more than just genomics)	71
Visualisation and analysis of complex networks in biology	72
Poster Presentations	73
Misannotation of coproporphyrinogen III oxidases HemN in the mycobacterial genome	73
Genomic surveillance reveals SARS-CoV-2 lineage B.6 is the major contributor to transmission in Malaysia	74
scMontage: Fast and robust gene expression similarity search for massive single-cell data	75
MPLANABASE™, a repository of <i>Metisa plana</i> transcriptome data for gene discovery	76
Deciphering the molecular interactions of kaempferol with three carrier proteins	77
Discovery of potential new inhibitors of <i>Mycobacterium tuberculosis</i> CYP121 from drug repositioning database	78
Cis-regulatory regions of pathogenic bacteria are associated with functionally conserved G-quadruplex motifs	79
A study of Fascioliasis from semi wild ruminants from two biological hotspots of India: A molecular approach using ribosomal ITS2 and mitochondrial CO1 genes	80
Phylogeography and population genetics of the rat tapeworm, <i>Hymenolepis diminuta</i> : An inference based on mtCO1 gene	81
Molecular docking study on the anti-staphylococcal activity of <i>Rumex nepalensis</i> (Spreng.)	82
Structural, functional and molecular dynamics analysis of CASR gene SNVs associated with tropical calcific pancreatitis	83
Interaction mechanism of Withanone and Withaferin-A from <i>Withania somnifera</i> with lysozyme, serum albumin and haemoglobin – A molecular dynamics approach	84
Identification of oil palm simple sequence repeat markers associated with basal stem rot disease	85
<i>In silico</i> docking actives compounds of betel leave (<i>Piper betle L.</i>) as antimalarial against plasmepsin 1 and plasmepsin 2	86
Multiplex primer design and optimization for effective detection of pathogenic bacteria <i>Aeromonas hydrophila</i>	87
Community structure and bacterial diversity in digestive tract of cultivated elver eel based on metagenomic analysis	88
Design and analysis of peptide inhibitors against the G12D KRAS protein in colon cancer	89
Exploring the patterns of codon usage and host adaptation in Sin Nombre virus	90
Comparing linear and non-linear machine learning models for predicting the pathogenicity of rare missense variants in hereditary cancer	91
Molecular docking of cobra venom cytotoxin with death receptors	92
Elastic SCAD SVM cluster for the selection of informative functional connectivity in autism spectrum disorder classification	93

Whole genome sequencing and analysis of Indian isolate of <i>Leptospira</i>	94
Whole genome sequencing and de novo assembly of three virulent Indian isolates of <i>Leptospira</i>	95
Preliminary structure-based drug discovery on Cathepsin S as an anti-inflammatory target	96
Identifying <i>Drosophila</i> cis-regulatory modules by a deep convolutional neural network on multiple transcriptional regulatory features	97
Automatic transcriptional factor-gene interaction literature evidence extraction via temporal convolutional neural networks	98
Melanoma detection via deep transfer learning	99
Novel biological metrics for evaluating the functional significance of RNA secondary structure predictions	100
Systems pharmacology approach to identify the activity of bioactive molecules against cervical cancer	101
Identification and study of specific genomic variants of the Kazakh population using comparative population genomics analysis	102
Validation of predicted novel Myc motifs mediating important PPIs using computational approaches	103
Role of surface-exposed charged basic amino acids (Lys, Arg) and guanidination in insulin on the interaction and stability of insulin–insulin receptor complex	104
Early investigation into the origin and evolution of SARS-CoV-2	105
ARIMA modelling for predicting Covid-19 in Indonesia: Integrated moving average, IMA(1,1), for modelling confirmed and cured cases of Covid-19 in Indonesia	106
Understanding the carbon concentrating mechanism in <i>Chlamydomonas reinhardtii</i> : A systems biology approach	107
Computational approaches to understand role of Argonautes in host-encoded resistance against vector borne plant RNA viruses	108
Application of whole exome-trio analysis in the elucidation of genetic basis of congenital pouch colon	109
Identification and analysis of class specific residues across rice Argonaute family	110
Predicting lncRNA and protein interactions in prostate cancer using a combination of computational and biophysical methods	111
Improving diagnostic approach to amyotrophic lateral sclerosis in India via two-stage NGS panel design	112
Extended mining of the oil biosynthesis pathway in biofuel plant <i>Jatropha curcas</i> by combined analysis of transcriptome and gene interactome data	113
UNIQmin: An alignment-independent tool for the study of pathogen sequence diversity at any given rank of taxonomy lineage	114

Preface

This book includes the abstracts of the keynote, highlight, oral, demo, lightning talk, workshop and poster presentations delivered at the 19th International Conference on Bioinformatics (InCoB), with the theme of “Bioinformatics and the Translation of Data-Driven Discoveries”. The conference was held from the 25th to 29th November 2020 and organized virtually, by the Asia-Pacific Bioinformatics Network (APBioNET), for the first time. The conference aimed to bring together researchers, decision makers, educators and scholars from the Asia-Pacific region and beyond for a scientific discourse and exchange on the state-of-the-art in the field, while providing for networking opportunities.

Despite the challenges presented by the COVID-19 pandemic, InCoB 2020 successfully attracted 181 registrants, with as many as 134 presentation submissions. A total of 71 were research manuscripts submitted for publication consideration in as many as 10 partner journals (the highest ever):

- BMC Genomics
- BMC Medical Genomics
- BMC Bioinformatics
- BMC Molecular and Cell Biology
- Computational Biology and Chemistry (CBAC)
- PeerJ
- MDPI Genes
- Journal of Bioinformatics and Computational Biology (JBCB)
- Quantitative Biology
- Frontiers Journals Research Topic Collection

We thank all the participants, the members of the organizing and scientific committees, the volunteers, and not forgetting our partners [Global Organisation for Bioinformatics Learning, Education & Training \(GOBLET\)](#), [International Society for Computational Biology \(ISCB\)](#), [Galaxy](#), [Pine.Bio](#), [MDPI Genes](#), and [Birunisoft PLT](#) for contributing to the success of the conference.

We look forward to your continued support in the future.

Mohammad Asif Khan, Ph.D.

Prashant Suravajhala, Ph.D.

InCoB 2020 Co-chairs

About APBioNET

The Asia-Pacific Bioinformatics Network (APBioNET; www.apbionet.org) is a nonprofit, non-governmental, international organization founded in 1998 that focuses on the promotion of bioinformatics in the Asia-Pacific region. APBioNET's mission, since its inception, has been to pioneer the growth and development of bioinformatics awareness, training, education, infrastructure, resources, and research among member countries and economies. Its work includes technical coordination, liaison, and/or affiliation with other international scientific bodies, such as the European Molecular Biology network (EMBnet), the International Society for Computational Biology (ISCB), and Global Organisation for Bioinformatics Learning, Education and Training (GOBLET), among others. APBioNET has members from over 12 countries in the region, from industry, academia, research, government, investors, and international organizations. APBioNET is spearheading a number of key bioinformatics initiatives in collaboration with international organizations, such as the Asia-Pacific Advanced Network (APAN), the Association of South-East Asian Nations (ASEAN), the Asia-Pacific Economic Cooperation (APEC), and the Asia-Pacific International Molecular Biology Network (A-IMBN), and industry partners. Many of the initiatives and activities have been initiated through its flagship conference, the International Conference on Bioinformatics (InCoB). In 2012, APBioNET was incorporated in Singapore as a public limited liability company to ensure quality, sustainability, and continuity of its mission to advance bioinformatics across the region and beyond. The five key thrust areas of APBioNET are i) Connecting the Bioinformatics Community in the Region, ii) Advancing Standards for Bioinformatics Activities, iii) Bioinformatics Education and Training, iv) Database/Computational Services and Resources, and v) Policy and Awareness. We invite members to actively contribute to each of these thrust areas.

About InCoB

The International Conference on Bioinformatics (InCoB) is a conference series first started in Bangkok, Thailand in 2002. Since then, the Asia Pacific Bioinformatics Network (APBioNet) has adopted this conference and grown the conference to become one of Asia's oldest and largest conferences on bioinformatics, with special focus on bioinformatics amongst life scientists. After the first InCoB, the conference has gone places, held in Penang (Malaysia), Auckland (New Zealand), Busan (Korea), Delhi (India), Hong Kong, Taipei (Taiwan), Singapore, Tokyo (Japan), Kuala Lumpur (Malaysia), Bangkok (Thailand), New South Wales (Australia), Odaiba (Japan), Shenzhen (China), and Jakarta (Indonesia). This year, due to the COVID-19 pandemic, InCoB 2020 was organized virtually for the first time. InCoB is one of the main drivers of APBioNET's key thrust area of "Connecting the Bioinformatics Community in the Region".

Keynote Talks

KN-1

SARS-CoV-2: What we have learned so far

Yoshihiro Kawaoka¹

¹University of Wisconsin-Madison, Madison, Wisconsin, United States

²University of Tokyo, Japan

Abstract

Late in 2019, a new coronavirus emerged in Wuhan, China and spread worldwide. The causative virus, SARS-CoV-2, continues to have a devastating impact on human lives. In an effort to develop therapeutics and preventive measures, we are performing numerous research projects with this virus. In this presentation, I will discuss our findings regarding animal models and vaccine development.

KN-2

Genomes - from personal to populations and back

Vinod Scaria¹

¹CSIR Institute of Genomics and Integrative Biology, Delhi, 110025, India

Abstract

The last decade has seen tremendous developments in the capability to sequence genomes. This has seen the unprecedented growth of personal genomics spilling over to population-scale genome initiatives across the world which now has provided insights which can significantly add value to interpreting personal genomes. One of the areas in modern medicine that has immensely been impacted by these developments have been clinical genetics - today impacting the diagnosis and potential precise treatment of thousands of patients and families suffering from rare genetic diseases. We have over the last decade from the initial personal genomes, build GUARDIAN, a clinical network for undiagnosed and rare diseases in India - today impacting thousands of families through genomic diagnosis. The followup initiatives as part of the IndiGen initiative for population genomics have provided insights and the much needed basal data to start implementing genomic medicine in India. This would only be possible with close collaboration and partnership towards enabling Predictive, Preventive, Precise, Personalised and Participatory Medicine.

KN-3

Global activities in Bioinformatics training and education

Nicola Mulder¹

¹Computational Biology Division, Department of Integrative Biomedical Sciences, Institute for Infectious Disease and Molecular Medicine, University of Cape Town, Cape Town, South Africa

Abstract

The increasing need for skills in bioinformatics, computational biology and data science, has resulted in a growing community of trainers involved in training and education in these topic areas. This community has started working closely together to build resources for trainers, develop curricula, and pioneer novel training approaches. Global initiatives such as GOBLET and the ISCB Education Committee and COSI, and more regional projects, such as the ELIXIR training activities in Europe and the H3ABioNet training program in Africa have convened during two Education summits to build a community of trainers and develop competency frameworks, best practices documents, guidelines and endorsement processes. The talk will highlight some of these activities and outputs and then provide specific use cases with the H3ABioNet consortium. H3ABioNet, the Pan African bioinformatics network for H3Africa, is mandated to build human capacity in bioinformatics for a broad list of stakeholders, from bioinformatics researchers to systems administrators, software developers, and life science researchers. The network uses a wide range of training modalities and implements competency-based curriculum development, standards for training material curation, and follows best practices developed by the bioinformatics global training community. Like many others, our training program has also had to adapt during the COVID-19 pandemic, switching to a more online focus. Some of these activities and adaptations will be discussed.

KN-4

Computational methods for trans-omics and single cells

Pengyi Yang¹

¹Charles Perkins Centre, School of Mathematics and Statistics, University of Sydney, Sydney, New South Wales, Australia

Abstract

In this talk, I will introduce various computational and machine learning methods used for integrating trans-omics data for trans-regulatory network reconstruction and characterisation. I will first focus on methods for cell signalling analysis using phosphoproteomics data. Then, I will showcase the methods that we have developed for classifying single cells and characterising gene properties on the single-cell level. Together, these works exemplify our work on computational method development and their application in making systems biological discovery.

KN-5

Mutational signatures: What caused the mutations in these cancers? Why do we care?

Steven G. Rozen^{1,2,3}

¹Programme in Cancer & Stem Cell Biology, Duke-NUS Medical School, Singapore

²Centre for Computational Biology, Duke-NUS Medical School, Singapore

³Duke-NUS Institute of Precision Medicine, National Heart Centre Singapore, Singapore

Abstract

Mutational signature analysis looks for patterns of somatic mutations with the ultimate aim of deducing what caused the mutations. Mutational signatures are important for understanding DNA damage and repair, for understanding the cellular origins of cancer, for detecting mutagenic exposures that cause cancer, and for preventing and treating cancer. Aristolochic acid, a mutagen found naturally in herbs that are widely used as medicine, is a straightforward example. In the last 7 years, mutational signatures have implicated aristolochic acid as a common cause of liver, urinary tract, and oesophageal cancers. More involved analyses require machine learning, which can infer mutational signatures from the mutations in large numbers of cancers. Multiple approaches for this have been developed, but the process still requires substantial human guidance and interpretation. Furthermore, existing approaches are poorly-suited to answering common questions, such as “Does this set of tumours contain any novel mutational signatures?” and “What is the evidence that a particular signature is present in a given tumor?” I will propose some new computational approaches to answering these questions.

KN-6

Omics resources and tools in under-represented animal models: A bird's eye view

Guojun Sheng¹

¹International Research Center for Medical Sciences (IRCMS), Kumamoto University, Kumamoto, Japan

Abstract

From Darwin's theory of evolution to seasonal flu vaccines, birds occupy an important niche in modern biological research. The first bird genome, of the chicken *Gallus gallus*, was sequenced in 2004. This was followed by that of the zebra finch in 2010 and today there are over 50 avian genomes that have been sequenced and are publicly accessible. However, avian models are in general not well suited for genetics-based research and as a consequence, tools for functional annotations of the avian omics resources, based on both wet and dry analyses, are limited. For example, for *G. gallus*, the most updated version of genome assembly (galGal6) still lacks information on six microchromosomes, and data on gene models, full-length RNA-seq, non-coding RNAs, transcription start sites, intron-exon boundaries and epigenetics-level information are very limited and poorly integrated with the genome assembly data. For avian labs more focused on investigating biological, rather than bioinformatic problems, it is challenging to make the best use of available, fragmented resources in their respective study. I will use our lab's work on avian early development as an example to highlight how our research projects can be helped by better collaboration and integration with bioinformatics-minded experts.

KN-7

Single-cell data analytics: Asian immune diversity and cancer cell states

Shyam Prabhakar¹

¹Genome Institute of Singapore (GIS), Agency for Science, Technology and Research (A*STAR), Singapore

Abstract

In recognition of the fact that cellular properties can vary systematically across individuals, the Human Cell Atlas-Asia consortium has initiated a flagship single cell project to generate an Asian Immune Diversity Atlas (AIDA). AIDA will profile transcriptome and epigenome variation in peripheral blood within and across population groups, and thereby characterize the influence of age, sex, genetics and environment on immune cell types in Asia. To execute such a geographically dispersed study without succumbing to lab-specific biases, we have standardized protocols across the consortium, centralized primary data processing and leveraged a common set of control samples across study sites. To reduce cost and minimize technical variation, we pool samples before single cell encapsulation (mux-seq). Early results suggest that these measures yield data that can seamlessly be integrated across the three initial sites.

Cohort-scale single-cell analysis requires industrial-strength data analysis, i.e. scalable algorithms that work uniformly well on all datasets at constant parameter settings. We have therefore developed next-gen algorithms for clustering cells by major cell type (RCA2) and then performing feature selection (DUBStepR) to sensitively identify subtypes within each major cell type. RCA2 combines the robustness of reference-based (supervised) clustering with the scalability of graph-based methods to detect common and rare cell types in large scRNA-seq datasets, despite profound batch effects and technical variation. DUBStepR exploits gene-gene correlations and a novel measure of cell clumping in expression space to identify the optimal gene set for sensitive and accurate cell clustering. Importantly, DUBStepR also generalizes to data types such as scATAC-seq that resist conventional feature selection approaches.

We used these methods to characterize molecular states in colorectal cancer, based on scRNA-seq analysis of >200,000 cells from tumor and matched normal. Notably, each patient contributed at least one unique epithelial cell cluster, indicating a remarkable diversity of cancer cell states across patients. In contrast, stromal and immune clusters were consistently shared across patients. Intriguingly, the clear distinctions between stromal cell types in normal colorectal tissue were smeared into an unbroken continuum of cell states in tumors. In a similar single cell study of bone marrow samples from chronic myeloid leukemia, we found that prevalence of certain cell states at diagnosis was strongly predictive of response to tyrosine kinase inhibitors.

Highlight

HL-1

Comparative genome analysis provides shreds of molecular evidence for reclassification of *Leuconostoc mesenteroides* MTCC 10508 as a strain of *Leu. suionicum*

Girija Kaushal^{1,2}, Sudhir P. Singha¹

¹Center of Innovative and Applied Bioprocessing, S.A.S. Nagar, Sector-81 Knowledge City, Mohali 140 306, India

²Department of Microbial Biotechnology, Panjab University, Chandigarh, India

Corresponding Author: Sudhir Singh

Abstract

This study presents the whole-genome comparative analysis of a *Leuconostoc sp.* strain, previously documented as *Leu. mesenteroides* MTCC 10508. The ANI, dDDH, dot plot, and MAUVE analyses suggested its reclassification as a strain of *Leu. suionicum*. Functional annotation identified a total of 1971 genes, out of which, 265 genes were mapped to CAZymes, evincing its carbohydrate transforming capability. The genome comparison with 59 *Leu. mesenteroides* and *Leu. suionicum* strains generated the core and pan-genome profiles, divulging the unique genes in *Leuconostoc sp.* MTCC 10508. For the first time, this study reports the genes encoding alpha-xylosidase and copper oxidase in a strain of *Leu. suionicum*. The genetic information for any possible allergenic molecule could not be detected in the genome, advocating the safety of the strain. The present investigation provides the genomic evidence for reclassification of the *Leuconostoc sp.* strain and also promulgates the molecular insights into its metabolic potential.

Oral Presentations

CBAC-1

Inference of phosphopeptide binding affinity from 14-3-3s by QSAR-based prediction

Ying Fan¹, Xiaojun Wang¹, Chao Wang^{2,3}

¹Shenzhen Institute of Information Technology, Shenzhen 518172, China

²HKBU Institute for Research and Continuing Education, Shenzhen 518057, China

³School of Chinese Medicine, Hong Kong Baptist University, Hong Kong SAR, China

Corresponding author: Chao Wang

Abstract

14-3-3s present in a multiple-isoforms fashion in mammalian cells and are involved in mediating many signal transduction activities by binding to the phosphorylated ligands. It is also demonstrated that 14-3-3s act as a key factor in inducing chemoresistance of tumorigenesis from the current findings. Thus, 14-3-3s are regarded as a novel candidate for biomarker development and cancer therapy. In this work, we developed the predictive models that can determine the binding affinity of the phosphopeptide fragments against 14-3-3s by the computational methods. We found that the hydrophobic property of the residues in the specific positions has a direct connection with the binding affinity of the phosphopeptides against 14-3-3s. The conserved patterns of 14-3-3 binding motifs were verified by our prediction results. A group of peptide sequences was predicted with high binding affinity and high sequence conservation, which had an agreement with 14-3-3s ligands. Overall, our results demonstrate that how the residues are likely to function in 14-3-3s interaction and the computational methods we introduced may contribute to further research.

Keywords: QSAR modeling, computational peptidology, peptide microarray, protein-ligand binding

CBAC-2

Integrated regulatory network based on lncRNA-miRNA-mRNA-TF reveals key genes and sub-networks associated with dilated cardiomyopathy

Sona Charles¹, Jeyakumar Natarajan¹

¹Data Mining and Text Mining Laboratory, Department of Bioinformatics, Bharathiar University, Coimbatore, Tamilnadu, India

Corresponding author: Jeyakumar Natarajan

Abstract

Dilated Cardiomyopathy (DCM) is a multifactorial condition often leading to heart failure in many clinical cases. Due to the high number of DCM incidence reported as familial, a gene level network based study was conducted utilizing high throughput next generation sequencing data. We exploited the exome and transcriptome sequencing data in NCBI-SRA database to construct a high confidence scale-free regulatory network consisting of lncRNA, miRNA, mRNA and Transcription Factors (TFs). Analysis of RNA-Seq data revealed 477 differentially expressed coding transcripts and 77 lncRNAs. 268 miRNAs regulated either lncRNAs or mRNAs. Out of the 477 coding transcripts that are deregulated 82 were TFs. We identified three major hub nodes lncRNA (XIST), miRNA (hsa-miR-195-5p) and mRNA (NOVA1) from the network. We also found putative disease associations of DCM with diabetes and DCM with hypoventilation syndrome. Five highly connected modules were also identified from the network. The hubs showed significant connectivity with the modules. Through this study we were able to gain insights into the underlying lncRNA-miRNA-mRNA-TF network. From a high throughput dataset we have isolated a handful of probable targets that may be utilized for studying the mechanisms of DCM development and progression to heart failure.

Keywords: Dilated cardiomyopathy, heart failure, transcriptomics, miRNA-lncRNA-gene networks

CBAC-3

The regulation of microRNA in each of cancer stage from two different ethnicities as potential biomarker for breast cancer

Kevin Nathanael Ramanto¹, Kresnodityo Jatiputro Widiyanto¹, Stefanus Satrio Hadi Wibowo¹,
David Agustriawan¹

¹Department of Bioinformatics, School of Life Sciences, Indonesia International Institute for Life Sciences, Jakarta, Indonesia

Corresponding author: David Agustriawan

Abstract

Previous studies have identified ethnic-specific miRNA, which could be a potential biomarker for breast cancer in a specific population. However, miRNA regulation is not only affected by a single factor. It is said that miRNA regulation is different in each cancer stage and could be used to predict specific cancer stages in different ethnicities. The present study identified specific miRNA biomarkers from two distinct ethnicities (non-Hispanic white and non-Hispanic black) using the TCGA dataset. Patient classification and miRNA selection were made by using R studio. Aberrant miRNAs identified by using the edgeR package. Aberrant miRNA specification was visualized using a Venn diagram, while the differential expression between normal and tumor patients was visualized using boxplot. miRNA-gene interaction was identified by using Spearman correlation analysis. The percentage of mutation cases and the protein-protein network of target genes were observed. The involvement of potential biomarkers in cancer was analyzed by KEGG functional enrichment analysis. Lastly, the diagnostic performance and prognosis were determined by using ROC and Kaplan-Meier survival analysis, respectively. Eleven unique aberrant miRNAs were selected as potential biomarkers based on its log fold change. The involvement of selected miRNAs in cancer was validated. Lastly, the result showed four of the miRNAs (hsa-mir-495, hsa-mir-592, hsa-mir-6501, and hsa-mir-937) are significantly detrimental to breast cancer diagnosis and prognosis. The result revealed miRNA regulation could differentiate the cancer stage within a specific population and provide valuable information to explore the role of miRNA in each cancer stage between non-Hispanic white and non-Hispanic black.

Keywords: breast cancer, miRNA regulation, epigenetics, miRNA profiling, TCGA, aberrant miRNA

CBAC-4

Convolutional neural networks with image representation of amino acid sequences for protein function prediction

Samia Tasnim Sara¹, Md Mehedi Hasan^{2,3}, Ahsan Ahmad¹, Swakkhar Shatabda¹

¹Department of Computer Science and Engineering, United International University Plot-2, United City, Madani Avenue, Badda, Dhaka-1212, Bangladesh

²Department of Bioscience and Bioinformatics, Kyushu Institute of Technology, 680-4 Kawazu, Iizuka, Fukuoka 820-8502, Japan

³Japan Society for the Promotion of Science, 5-3-1 Kojimachi, Chiyoda-ku, Tokyo 102-0083, Japan

Corresponding author: Swakkhar Shatabda

Abstract

Proteins are one of the most important molecules that govern the cellular processes in organisms. Various functions of the proteins are of paramount importance to understand the basics of life. Several supervised learning approaches are applied to this field to predict the functionality of proteins. In this paper, we propose a convolutional neural network based approach ProtConv to predict the functionality of proteins by converting the amino-acid sequences to a two dimensional image. We have used a protein embedding technique using transfer learning to generate the feature vector. Feature vector is then converted into a square sized single channel image to be fed into a convolutional network. The neural network architecture used here is a combination of convolutional layers and average pooling layers followed by dense fully connected layers to predict a binary function. We have performed experiments on standard benchmark datasets taken from two very important protein function prediction tasks: proinflammatory cytokines and anticancer peptides. Our experiments show that the proposed method, ProtConv achieves state-of-the-art performances on both of the datasets. All necessary details about implementation with source code and datasets are made available at: <https://github.com/swakkhar/ProtConv>.

Keywords: feature representation, convolutional neural network, protein function prediction, transfer learning 2010 MSC: 00-01, 99-00

CBAC-5

Immunogenicity and structural efficacy of P41 of *Plasmodium sp.* as potential cross-species blood-stage malaria vaccine

Kevin Nathanael Ramanto¹, Rizky Nurdiansyah¹

¹Department of Bioinformatics, School of Life Sciences, Indonesia International Institute for Life Sciences, Jakarta, Indonesia

Corresponding author: Rizky Nurdiansyah

Abstract

Vaccine based strategies offer a promising future in malaria control by generating protective immunity against natural infection. However, vaccine development is hindered by the *Plasmodium sp.* genetic diversity. Previously, we have shown P41 protein from 6-Cysteine shared by *Plasmodium sp.* and could be used for cross-species anti-malaria vaccines. Two different approaches, ancestral, and consensus sequence, could produce a single target for all human-infecting *Plasmodium*. In this study, we investigated the efficacy of ancestral and consensus of P41 protein. Phylogenetic and time tree reconstruction was conducted by RAXML and BEAST2 package to determine the relationship of known P41 sequences. Ancestral and consensus sequences were reconstructed by the GRASP server and Unipro Ugene software, respectively. The structural prediction was made using the Psipred and Rosetta program. The protein characteristic was analyzed by assessing hydrophobicity and post-translational modification sites. Meanwhile, the immunogenicity score for B-cell, T-cell, and MHC was determined using an immunoinformatic approach. The result suggests that ancestral and consensus have a distinct protein characteristic with high immunogenicity scores for all immune cells. We found one shared conserved epitope with phosphorylation modification from the ancestral sequence that has the potential as a target for the cross-species vaccine. Thus, this study provides detailed insight into P41 efficacy for the cross-species antimalaria blood-stage vaccine.

Keywords: malaria, vaccine, 6-cysteine, protein structure, immunogenicity

CBAC-6

Structural and immunogenicity analysis of reconstructed ancestral and consensus P48/45 for cross-species anti malaria transmission-blocking vaccine

Kevin Nathanael Ramanto¹, Rizky Nurdiansyah¹

¹Department of Bioinformatics, School of Life Sciences, Indonesia International Institute for Life Sciences Jakarta, Indonesia

Corresponding author: Rizky Nurdiansyah

Abstract

The development of the anti-malaria vaccine holds a promising future in malaria control. One of the anti-malaria vaccine strategies known as the transmission-blocking vaccine (TBV) is designed to inhibit the parasite transmission between human and mosquito by targeting the parasite gametocyte. Previously, we found that P48/45 that include in the 6-cysteine protein family shared by Plasmodium sp. We also detected vaccine properties that are possessed by all human-infecting Plasmodium and could be used as a cross-species anti-malaria vaccine. In this study, we investigated the efficacy of P48/45 through the ancestral and consensus reconstruction approach. P48/45 phylogenetic and time tree analysis was done by RAXML and BEAST2. GRASP server and Ugene software were used to reconstruct ancestral and consensus sequences, respectively. The protein structural prediction was done by using a psipred and Rosetta program. Each protein characteristic of P48/45 was analyzed by assessing hydrophobicity and post-translational modification sites. Meanwhile, the Epitope sequence for B-cell, T-cell, and MHC was determined using an immunoinformatics approach. Lastly, molecular docking simulation was done to determine native binding interactions of P48/45-P230. The result showed a distinct protein characteristic of ancestral and consensus sequences. The immunogenicity analysis revealed the number of epitopes in the ancestral sequence is greater than consensus sequence. The study also found a conserved epitope located in the binding site and consists of specific post-translational modification sites. Hence, our research provides detailed insight into ancestral and consensus P48/45 efficacy.

Keywords: malaria, vaccine development, 6-cysteine protein family, ancestral and consensus reconstruction

CBAC-7

Discovery of new inhibitor for the protein arginine deiminase type 4 (PAD4) by rational design of α -Enolase-derived peptides

Izzuddin Ahmad Nadzirin^{1,4}, Adam Leow Thean Chor², Abu Bakar Salleh³, Mohd Basyaruddin Abdul Rahman¹, Bimo A. Tejo¹

¹Department of Chemistry, Faculty of Science, Universiti Putra Malaysia, 43400 UPM Serdang, Malaysia

²Departments of Cell and Molecular Biology, Faculty of Biotechnology and Biomolecular Sciences, Universiti Putra Malaysia, 43400 UPM Serdang, Malaysia

³Department of Biochemistry, Faculty of Biotechnology and Biomolecular Sciences, Universiti Putra Malaysia, 43400 UPM Serdang, Malaysia

⁴Department of Biomedical Science, Faculty of Allied Health Science, International Islamic University Malaysia, 25200 Kuantan, Pahang, Malaysia

Corresponding author: Bimo A. Tejo

Abstract

Rheumatoid arthritis (RA) is an inflammatory autoimmune disease affecting about 0.5–1.0% of the world population. Protein arginine deiminase type 4 (PAD4) is believed to be responsible for the occurrence of RA by catalyzing citrullination of proteins. The citrullinated proteins act as autoantigens by stimulating an immune response. Citrullinated α -enolase has been identified as one of the autoantigens for RA. Hence, α -enolase serves as a good template for design of potential peptide inhibitors against PAD4. The binding affinity of α -enolase-derived peptides and PAD4 was virtually determined using PatchDock and HADDOCK docking programs. Synthesis of the designed peptides was performed using a solid phase peptide synthesis method. The inhibitory potential of each peptide was determined experimentally by PAD4 inhibition assay and IC₅₀ measurement. PAD4 assay data show that the N-P2 peptide is the most favourable substrate among all peptides. Further modification of N-P2 by changing the Arg residue to canavanine (P2 (Cav)) rendered it an inhibitor against PAD4 by reducing the PAD4 activity to 35% with IC₅₀ 1.39 mM. We conclude that P2 (Cav) is a potential inhibitor against PAD4 and can serve as a starting point for the development of even more potent inhibitors.

Keywords: drug design, PAD4, peptide inhibitor, rheumatoid arthritis

CBAC-8

Bipartite molecular approach for species delimitation and resolving cryptic speciation of *Exobasidium vexans* within the *Exobasidium* genus

Chayanika Chaliha¹, V. Chandra Kaladhar², Robin Doley¹, Praveen Kumar Verma², Aditya Kumar¹, Eeshan Kalita¹

¹Department of Molecular Biology and Biotechnology, Tezpur University, Tezpur, Assam, India

²Plant Immunity Laboratory, National Institute of Plant Genome Research, Aruna Asaf Ali Marg, New Delhi, India

Corresponding author: Eeshan Kalita

Abstract

Exobasidium vexans, a basidiomycete pathogen, is the causal organism of blister blight disease in tea. The molecular identification of the pathogen remains a challenge due to the limited availability of genomic data in sequence repositories and cryptic speciation within its genus *Exobasidium*. In this study, the nuclear internal transcribed spacer rDNA region (ITS) based DNA barcode was developed for *E. vexans*, to address the problem of molecular identification within the background of cryptic speciation. The isolation of *E. vexans* strain was confirmed through morphological studies followed by molecular identification utilizing the developed ITS barcode. Phylogenetic analysis based on Maximum Parsimony (MP), Maximum Likelihood (ML) and Bayesian Inference (BI) confirmed the molecular identification of the pathogen as *E. vexans* strain. Further, BI analysis using BEAST mediated the estimation of the divergence time and evolutionary relationship of *E. vexans* within genus *Exobasidium*. The speciation process followed the Yule diversification model wherein the genus *Exobasidium* is approximated to have diverged in the Paleozoic era. The study thus sheds light on the molecular barcode-based species delimitation and evolutionary relationship of *E. vexans* within its genus *Exobasidium*.

Keywords: *E. vexans*, ITS, cryptic speciation, maximum parsimony, maximum likelihood, bayesian inference

CBAC-9

Sub-structure-based screening and molecular docking studies of potential enteroviruses inhibitors

Stephen Among James^{1,2,†}, Wai Keat Yam¹

¹Centre for Bioinformatics, School of Data Sciences, Perdana University, Selangor Darul Ehsan, Malaysia

²Department of Biochemistry, Faculty of Science, Kaduna State University, 800211, Kaduna, Nigeria

[†]This author contributed at this institution while on study

Corresponding author: Wai Keat Yam

Abstract

Rhinoviruses (RV), especially Human rhinovirus (HRVs) have been accepted as the most common cause for upper respiratory tract infections (URTIs). Pleconaril, a broad spectrum anti-rhinoviral compound, has been used as a drug of choice for URTIs for over a decade. Unfortunately, for various complications associated with this drug, it was rejected, and a replacement is highly desirable. In silico screening and prediction methods such as sub-structure search and molecular docking have been widely used to identify alternative compounds. In our study, we have utilised sub-structure search to narrow down our quest in finding relevant chemical compounds. Molecular docking studies were then used to study their binding interaction at the molecular level. Interestingly, we have identified 3 residues that is worth further investigation in upcoming molecular dynamics simulation systems of their contribution in stable interaction.

Keywords: pleconaril, molecular docking, sub-structure, Rhinoviruses, antiviral agents

CBAC-10

OriC-ENS: A sequence-based ensemble classifier for predicting origin of replication in *S. cerevisiae*

Sayed Mehedi Azim¹, Md. Rakibul Haque¹, Swakkhar Shatabda¹

¹Department of Computer Science and Engineering, United International University, Plot-2, United City, Madani Avenue, Badda, Dhaka-1212, Bangladesh

Corresponding author: Swakkhar Shatabda

Abstract

DNA Replication plays the most crucial part in biological inheritance, ensuring an even flow of genetic information from parent to offspring. The beginning site of DNA Replication which is called the Origin of Replication (ORI), plays a significant role in understanding the molecular mechanisms and genomic analysis of DNA. Hence, it is paramount to accurately identify the origin of replication in order to gain a more accurate understanding of the biochemical and genomic properties of DNA. In this paper, we have proposed a new approach named OriC-ENS that uses sequence-based feature extraction techniques, K-mer, K-gapped Mono-Di and Di Mono, and an ensemble classification technique that uses majority voting for the identification of Origin of Replication. We have used three SVM classifiers, one for the K-mer features and two more for K-Gapped Mono-Di and K-Gapped Di-mono features. Finally, we used majority voting to combine the prediction by each predictor. Experimental results on the *S. Cerevisie* dataset has shown that our method achieves an accuracy of 91.62% which outperforms other state-of-the-art methods by a significant margin. We have also tested our method using other metrics such as Matthews Correlation Coefficient (MCC), Area Under Curve (AUC), Sensitivity, and Specificity, where it has achieved a score of 0.83, 0.98, and 0.90 respectively.

Keywords: -

CBAC-11

***In silico* design of potent inhibitor to hamper the interaction between HIV-1 integrase and LEDGF/p75 interaction using e-pharmacophore modelling, virtual screening, molecular docking and dynamics simulations**

Umesh Panwar¹, Sanjeev Kumar Singh¹

¹Computer Aided Drug Design and Molecular Modelling Lab, Department of Bioinformatics, Alagappa University, Karaikudi-630 004, Tamil Nadu, India.

Corresponding Author: Sanjeev Kumar Singh

Abstract

The rapid increase of HIV-1 infection throughout the globe has a high demand for a superior drug with lesser side effects. LEDGF/p75, the human Lens Epithelium-Derived Growth Factor is identified as promising cellular cofactor with integrase in facilitating viral replication in early stage by acting as a tethering factor for the pre-integration to the chromatin. Therefore, the present study was designed to identify a potent inhibitor by applying an E-pharmacophore based virtual screening, molecular docking, and dynamics simulation studies. Our investigation identified two potent molecules ZINC22077550 and ZINC32124441 with the efficient binding affinity, strong hydrogen bonding and acceptable pharmacological properties to hamper the interaction between integrase and LEDGF/p75. Finally, the DFT and MDS were also analyzed, and shown a favorable energetic state and dynamics stability in comparison to the reference compound. In conclusion, we suggest that these findings could be a novel therapeutics in future and may increase the lifespan of an individual suffering with viral infection.

Keywords: HIV-1 integrase, LEDGF/p75, e-pharmacophore, virtual screening, docking, DFT, MD simulation

CBAC-12

Discovery of network motifs based on induced subgraphs using a dynamic expansion tree

Sabyasachi Patra¹

¹Bioinformatics Lab, Department of Computer Science, IIT, Bhubaneswar, India

Abstract

Biological networks are powerful representations of topological features in biological systems. Finding network motifs in biological networks is a computationally hard problem due to their huge size and abrupt increase of search space with the increase of motif size. Motivated by the computational challenges of network motif discovery and considering the importance of this topic, an efficient and scalable network motif discovery algorithm based on induced subgraphs in a dynamic expansion tree is proposed. This algorithm uses a pruning strategy to overcome the space limitation of the static expansion tree. The proposed algorithm is able to identify large network motifs up to size 15 by significantly reducing the computationally expensive subgraph isomorphism checks. Further, the present work avoids the unnecessary growth of patterns that do not have any statistical significance. The runtime performance of the proposed algorithm outperforms that of most of the existing algorithms for large network motifs.

Keywords: biological network, network motif, induced subgraph, subgraph isomorphism, static expansion tree, dynamic expansion tree

CBAC-13

Prediction of protein-protein interaction between human and *Streptococcus pneumoniae* using logistic regression

Vivitri Dewi Prasasty¹, Reinhart Gunadi², Dewi Yustika Sofia², Ernawati Sinaga³

¹Faculty of Biotechnology, Atma Jaya Catholic University of Indonesia, Jakarta 12930, Indonesia

²Department of Biology, Faculty of Life Sciences, Surya University, Tangerang, Banten 15143, Indonesia

³Faculty of Biology, Universitas Nasional, Jakarta 12520, Indonesia

Corresponding author: Vivitri Dewi Prasasty

Abstract

Streptococcus pneumoniae remains a leading cause of mortality in children under five years old. In recent years, the emergence of antibiotic-resistant strains of *S. pneumoniae* increases the threat level of this pathogen. For that reason, the exploration of *S. pneumoniae* protein virulence factors should be considered in developing new drugs or vaccines, for instance, by the analysis of host-pathogen protein-protein interaction (HP-PPI). In this research, the prediction of protein-protein interactions was performed with a logistic regression model with the number of protein domain occurrences as features. By utilizing HP-PPIs of three different pathogens as training data, the model achieved 57-77% precision, 64-75% recall, and 96-98% specificity. Prediction of human-*S. pneumoniae* protein-protein interactions using the model yielded 5823 interactions involving thirty *S. pneumoniae* proteins and 324 human proteins. Pathway enrichment analysis showed that most of the pathways involved in the predicted interactions are immune system pathways. Network topology analysis revealed β -galactosidase (BgaA) as the most central among the *S. pneumoniae* proteins in the predicted HP-PPI network, with a degree centrality of 1.0 and a betweenness centrality of 0.451853. Further experimental studies are required to validate the predicted 76 interactions and examine their roles in *S. pneumoniae* infection.

Keywords: host-pathogen interaction, network centrality, pathway enrichment, pneumococcal infection

CBAC-14

Validation of predicted novel Myc motifs mediating important PPIs using computational approaches

Debangana Chakravorty¹, Abhirupa Ghosh¹, Sudipto Saha¹

¹Division of Bioinformatics, Bose Institute, Kolkata, India

Corresponding author: Sudipto Saha

Abstract

Myc has a large stretch of disordered region whose structure is unknown and it contains many short conserved regions that mediate Protein-Protein Interactions (PPIs). Previously, many such experimentally validated and predicted linear motifs have been identified in Myc. Two novel predicted linear motifs which mediate important interactions were selected to be validated. The interaction of Myc motifs and their PPI partners were explored and visualized using protein-protein docking servers followed by the identification of binding interfaces. Computational Alanine scanning was performed to identify key residues in the motif and the effects of mutation in these residues were explored by prediction servers. One key residue mutation in each motif was studied by Molecular Dynamics simulation. These key residues may prove to be of importance for mediating the PPI if its mutation leads to destabilization of the interaction.

Keywords: motif, PPI partner, web servers, interaction interface, MD simulation

QB-1

Exploration the underlying mechanism of a traditional Chinese medicine formula Youdujing ointment for cervical cancer treatment

Lei Zhang¹, Jian Cheng Lv¹, Ming Xiao¹, Le Zhang¹

¹College of Computer Science, Sichuan University, Chengdu 610065, China

Corresponding author: Le Zhang

Abstract

A traditional Chinese medicine formula Youdujing (YDJ) ointment was widely used for HPV-related diseases (e.g. cervical cancer) treatment. However, the underlying mechanisms of YDJ for cervical cancer treatment are still unclear. For this reason, we develop a comprehensive network pharmacology approach to explore its key mechanisms by integrating potential target identification, network analysis, and enrichment analysis into classical molecular docking procedure. Firstly, we use the network and enrichment analysis to screen out four potential therapeutic targets: ESR1, NFKB1, TNF and AKT1. Secondly, we employ molecular docking to have four potential effective compounds: E4, Y2, Y20 and Y21. Thirdly, we employ network-based study to turn out that Y2-E4 and Y21-E4 are potential drug combinations. Y2 or Y21 can work alone or together with E4 to trigger apoptotic cascades via mitochondrial apoptotic pathway and estrogen receptors. In summary, our study can not only demonstrate why YDJ are effective for cervical cancer treatment, but also offer a general procedure to investigate the underlying mechanism for traditional Chinese medicine formula.

Keywords: -

QB-2

DNA methylation profiling reveals new potential subtype-specific gene markers for early-stage renal cell carcinoma in Caucasian population

Alvaro Filbert Liko¹, Edward Ciputra¹, Nathaniel Alvin Sanjaya¹, Priskila Cherisca Thenaka¹, David Agustriawan²

¹Department of Biomedicine, School of Life Sciences, Indonesia International Institute for Life Sciences, Jakarta, Indonesia.

²Department of Bioinformatics, School of Life Sciences, Indonesia International Institute for Life Sciences, Jakarta, Indonesia.

Corresponding author: David Agustriawan

Abstract

Introduction: Renal cell carcinoma (RCC) is among the top adult cancers worldwide, with a challenging management due to lack of early diagnosis, therapy resistance, and diverse molecular background. Aberrant DNA methylation has been associated with RCC development due to transcription deregulation. We discovered potential DNA methylation-based biomarkers for stage I RCC in Caucasian population from The Cancer Genome Atlas (TCGA) database.

Methods: Patients' clinical, methylation beta-value, and mRNA expression data were retrieved. Differential methylation and expression analysis were conducted to obtain differentially methylated CpG-gene pairs. Inversely correlated CpG-gene pairs between their expression and methylation levels were selected using Pearson's correlation test and then screened for somatic mutations using the Catalogue of Somatic Mutations in Cancer (COSMIC) database. Their biomarker capacities were analyzed using the Kaplan-Meier and receiver operating characteristic analysis, followed by protein interactomics and functional enrichment analysis.

Results: We obtained differentially methylated CpGs in clear cell (KIRC) and papillary RCC (KIRP) but not chromophobe RCC. Six inversely correlated CpG-gene pairs with no reported cancer-associated mutations were selected. Prognostic values were found in ATXN1 and RFTN1 for KIRC, along with GRAMD1B and TM4SF19 for KIRP, while diagnostic values were found in VIM and RFTN1 for KIRC, along with TNFAIP6 and TM4SF19 for KIRP. Both subtypes showed enrichment of immune and metabolism-related pathways.

Conclusions: We discovered potential DNA methylation-based prognostic and diagnostic markers for early-stage RCC in Caucasian population. Validation by wet laboratory analysis and adjustments for confounding variables might be needed, considering our study limitation to specific race.

Keywords: -

QB-3

Study of COVID-19 epidemic in India with SEIRD model

Rudra Banerjee¹, Srijit Bhattacharjee¹, Pritish Kumar Varadwajz¹

¹Indian Institute of Information Technology Allahabad, Jhalwa, Uttar Pradesh

Corresponding author: Pritish Kumar

Abstract

Coronavirus pandemic (COVID-19) is causing a havoc to the entire world due to the newly discovered virus SARS-CoV-2. In this study, the dynamics of COVID-19 for India and a few selected states with different demographic structures has been analyzed using a SEIRD epidemiological model. A systematic estimation of the basic reproductive ratio R_0 is made for India and for each of the selected states. The study has analysed and predicted the dynamics of the temporal progression of the disease in India and the selected eight states: Andhra Pradesh, Chhattisgarh, Delhi, Gujarat, Madhya Pradesh, Maharashtra, Tamil Nadu, and Uttar Pradesh. For India, the model shows, peak of infection is expected to appear near the end of October. Further, we compare the model scenario with a Gaussian fit of daily infected cases and show the peak of infections will also appear around middle of October this year. A comparison of the infection dynamics with two countries Italy and Russia has been displayed. This shows an early imposition of lockdown has reduced the number of infected cases but delayed the appearance of peak significantly.

Keywords: -

QB-4

***In silico* study of race- and cancer stage-specific DNA methylation pattern in breast cancer patients based on TCGA dataset**

Jeremias Ivan¹, Gabriella Patricia¹, David Agustriawan¹

¹Department of Bioinformatics, School of Life Sciences, Indonesia International Institute for Life Sciences

Correspondence author: David Agustriawan

Abstract

Breast cancer is one of the most common types of cancer, particularly among women. As current breast cancer treatments are still ineffective, we assess the methylation pattern of breast cancer patients across race and cancer stage based on The Cancer Genome Atlas (TCGA) dataset. Significant hypermethylation and hypomethylation can regulate the gene expression, thus becoming potential biomarkers in breast cancer tumorigenesis.

BMC-1

A semi-supervised deep learning approach for predicting the functional effects of genomic non-coding variations

Hao Jia¹, Sung-Joon Park^{1,2}, Kenta Nakai^{1,2}¹Department of Computer Science, the University of Tokyo, Japan²Human Genome Center, the Institute of Medical Science, the University of Tokyo, Japan**Corresponding author:** Kenta Nakai

Abstract

Background: Understanding the functional effects of non-coding variants is important as they are often associated with gene-expression alteration and disease development. Over the past few years, many computational tools have been developed to predict their functional impact. However, the intrinsic difficulty in dealing with the scarcity of data leads to the necessity to further improve the algorithms. In this work, we propose a novel method, employing a semi-supervised deep-learning model with pseudo labels, which takes advantage of learning from both experimentally annotated and unannotated data.

Results: We prepared known functional non-coding variants with histone marks, DNA accessibility, and sequence context in GM12878, HepG2, and K562 cell lines. Applying our method to the dataset demonstrated its outstanding performance, compared with that of existing tools. Our results also indicated that the semi-supervised model with pseudo labels achieves higher predictive performance than the supervised model without pseudo labels. Interestingly, a model trained with the data in a certain cell line is unlikely to succeed in other cell lines, which implies the cell-type-specific nature of the non-coding variants. Remarkably, we found that DNA accessibility significantly contributes to the functional consequence of variants, which suggests the importance of open chromatin conformation prior to establishing the interaction of non-coding variants with gene regulation.

Conclusions: The semi-supervised deep learning model coupled with pseudo labeling has advantages in studying with limited datasets, which is not unusual in biology. Our study provides an effective approach in finding non-coding mutations potentially associated with various biological phenomena, including human diseases.

Keywords: -

BMC-2

Comparative analysis and prediction of nucleosome positioning using integrative feature representation and machine learning algorithms

Qi Li¹, Guosheng Han², Ying Li²

¹Department of Environmental Science and Engineering, Xiangtan University, Xiangtan 411105, China

²School of Mathematics and Computational Science, Xiangtan University, Hunan, 411105, China

Corresponding authors: Qi Li and Guosheng Han

Abstract

Background: Nucleosome plays an important role in the process of genome expression, DNA replication, DNA repair and transcription. Therefore, the research of nucleosome positioning has invariably received extensive attention. Considering the diversity of DNA sequence representation methods, we tried to integrate multiple features to analyze its effect in the process of nucleosome positioning analysis. This process can also deepen our understanding of the theoretical analysis of nucleosome positioning.

Results: Here, we not only used frequency chaos game representation (FCGR) to construct DNA sequence features, but also integrated it with other features and adopted the principal component analysis (PCA) algorithm. Simultaneously, support vector machine (SVM), extreme learning machine (ELM), extreme gradient boosting (XGBoost), multilayer perceptron (MLP) and convolutional neural networks (CNN) are used as predictors for nucleosome positioning prediction analysis, respectively. The integrated feature vector prediction quality is significantly superior to a single feature. After using principal component analysis (PCA) to reduce the feature dimension, the prediction performance of *H. sapiens* dataset has been significantly improved.

Conclusions: Comparative analysis and prediction results on *H. sapiens*, *C. elegans*, *D. melanogaster* and *S. cerevisiae* datasets demonstrate that the application of FCGR to nucleosome positioning is feasible, and we also found that integrative feature representation would be better.

Keywords: -

BMC-3

A computational model of mutual antagonism in the mechano-signaling network of RhoA and nitric oxide

Akila Surendran¹, C. Forbes Dewey², Jr., Boon Chuan Low^{3,4}, Lisa Tucker-Kellogg⁴

¹National Institute of Speech and Hearing (NISH), Centre for Assistive Technology and Innovation (CATI), India

²Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA

³Mechanobiology Institute, National University of Singapore, 5A Engineering Drive 1, Singapore 117411, Singapore

⁴Department of Biological Sciences, National University of Singapore, 16 Science Drive 4, Singapore 117558, Singapore

⁵Cancer & Stem Cell Biology, and Centre for Computational Biology Duke-NUS Medical School, Singapore

Corresponding author: Lisa Tucker-Kellogg

Abstract

Background: RhoA is a master regulator of cytoskeletal contractility, while nitric oxide (NO) is a master regulator of relaxation, e.g. vasodilation. There are multiple forms of cross-talk between the RhoA/ROCK pathway and the eNOS/NO/cGMP pathway, but previous work has not studied their interplay at a systems level. Literature review suggests that the majority of their cross-talk interactions are antagonistic, which motivates us to ask whether the RhoA and NO pathways exhibit mutual antagonism in vitro, and if so, to seek the theoretical implications of their mutual antagonism.

Results: Experiments found mutual antagonism between RhoA and NO in epithelial cells. Since mutual antagonism is a common motif for bistability, we sought to explore through theoretical simulations whether the RhoA-NO network is capable of bistability. Qualitative modeling showed that there are parameters that can cause bistable switching in the RhoA-NO network, and that the robustness of the bistability would be increased by positive feedback between RhoA and mechanical tension.

Conclusions: We conclude that the RhoA-NO bistability is robust enough in silico to warrant the investment of further experimental testing. Tension-dependent bistability has the potential to create sharp concentration gradients, which could contribute to the localization and self-organization of signaling domains during cytoskeletal remodeling and cell migration.

Keywords: -

BMC-4

Identifying genomic islands with deep neural networks

Fangfang Xia¹, Rick Stevens^{2,3,4}, Rida Assaf⁵

¹Department of Chemistry, College of Environmental Sciences, Nanjing Xiaozhuang University, Jiangsu Province, Nanjing 211171, PR China

²Computation Institute, University of Chicago, Chicago, IL, USA

³Computing, Environment and Life Sciences, Argonne National Laboratory, Argonne, IL, USA

⁴Department of Computer Science, University of Chicago, Chicago, IL, 60637, USA

⁵Department of Computer Science, University of Chicago, Chicago, IL 60637, USA

Corresponding author: Rida Assaf

Abstract

Horizontal gene transfer is the main source of adaptability for bacteria, through which genes are obtained from different sources including bacteria, archaea, viruses, and eukaryotes. This process promotes the rapid spread of genetic information across lineages, typically in the form of clusters of genes referred to as genomic islands (GIs). Different types of GIs exist, often classified by the content of their cargo genes or their means of integration and mobility. Various computational methods have been devised to detect different types of GIs, but no single method currently is capable of detecting all GIs. We propose a method, which we call Shutter Island, that uses a deep learning model (Inception V3, widely used in computer vision) to detect genomic islands. The intrinsic value of deep learning methods lies in their ability to generalize. Via a technique called transfer learning, the model is pre-trained on a large generic dataset and then re-trained on images that we generate to represent genomic fragments. We demonstrate that this image-based approach generalizes better than the existing tools. We used a deep neural network and an image-based approach to detect the most out of the correct GI predictions made by other tools, in addition to making novel GI predictions. The fact that the deep neural network was retrained on only a limited number of GI datasets and then successfully generalized indicates that this approach could be applied to other problems in the field where data is still lacking or hard to curate.

Keywords: -

BMC-5

Mendelian randomization studies of brain MRI yield insights into the pathogenesis of neuropsychiatric disorders

Weichen Song¹, Wei Qian¹, Weidi Wang^{1,2}, Shunying Yu^{1,2}, Guan Ning Lin^{1,2}

¹Shanghai Mental Health Center, Shanghai Jiao Tong University School of Medicine, School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai 200030, China

²Shanghai Key Laboratory of Psychotic Disorders, Shanghai 200030, China

Corresponding authors: Shunying Yu and Guan Ning Lin

Abstract

Background: Observational studies have identified various associations between neuroimaging alterations and neuropsychiatric disorders. However, whether such associations could truly reflect causal relations remains still unknown.

Results: Here, we leveraged genome-wide association studies (GWAS) summary statistics for 1) 11 psychiatric disorders (sample sizes varied from $n=9,725$ to $91,331,010$); 2) 110 diffusion tensor imaging (DTI) measurement (sample size $n=17,706$); and 3) 101 region-of-interest (ROI) volumes, and investigate the causal relationship between brain structures and neuropsychiatric disorders by two-sample Mendelian randomization. Among all DTI-Disorder combinations, we observed a significant causal association between the superior longitudinal fasciculus (SLF) and Anorexia nervosa (AN) ($\beta=-0.48$, 95% confidence interval: $-0.69 \sim -0.27$, $15P=6.4 \times 10^{-6}$). Similar significant associations were also observed between the body of 16th corpus callosum (fractional anisotropy) and Alzheimer's disease ($\beta=0.07$, 95% CI: $170.03 \sim 0.10$, $P=4.1 \times 10^{-5}$). By combining all observations, we found that the overall p-value for DTI-Disorder associations was significantly elevated compared to the null distribution (Kolmogorov-Smirnov $P=0.009$, inflation factor $\lambda=1.37$), especially for 20 DTI-Bipolar disorder (BP) ($\lambda=2.64$) and DTI-AN ($\lambda=1.82$). In contrast, for ROI-Disorder combinations, we only found a significant association between the brain region of pars triangularis and Schizophrenia ($\beta=-0.73$, 95% CI: $-1.08 \sim -0.37$, $1P=5.9 \times 10^{-5}$) and no overall p-value elevation for ROI-Disorder analysis compared to the null expectation.

Conclusion: As a whole, we show that SLF degeneration may be a risk factor for AN, while DTI variations could be causally related to some neuropsychiatric disorders, such as BP and AN. In addition, the white matter structure might have a larger impact on neuropsychiatric disorders than subregion volumes.

Keywords: neuroimaging, neuropsychiatric disorders, dysconnectivity, anorexia nervosa, superior longitudinal fasciculus

BMC-6

Dissection of genetic association of anorexia nervosa and obsessive-compulsory disorder at network and cellular levels

Weichen Song¹, Weidi Wang^{1,2}, Shunying Yu^{1,2}, Guan Ning Lin^{1,2}

¹Shanghai Mental Health Center, School of Medicine, School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai 200030, China

²Shanghai Key Laboratory of Psychotic Disorders, Shanghai 200030, China

Corresponding authors: Shunying Yu and Guan Ning Lin

Abstract

Background: Anorexia Nervosa (AN) and Obsessive-Compulsory Disorder (OCD) exhibited a high comorbidity rate, similar symptoms, and a shared genetic basis. However, the understanding of specific underlying mechanisms of these commonalities is currently limited.

Method: Here, we collected Genome-Wide Association Analysis results for AN and OCD, and obtained genes hit by top SNPs as risk genes. We then carried out an integrative co-expression network analysis to explore the convergence and divergence of AN and OCD risk genes.

Result: At first, we observed that risk genes of AN enriched in co-expression modules that involved extracellular matrix functions and highly expressed in postnatal brain, limbic system, and non-neuronal cell types, while OCD risk genes were enriched in modules of synapse function, prenatal brain, cortex layers and neuron. Next, by comparing expressions from eating disorder and OCD postmortem patient brain tissues, we observed both disorders have similar prefrontal cortex expression alterations influencing the synapse transmission, which suggests that two diseases could have similar functional pathways.

Conclusion: We found that AN and OCD risk genes had distinct functional and spatiotemporal enrichment patterns, but carrying similar expression alterations as a disease mechanism, which may be one of the key reasons why they had similar but not identical clinical phenotypes.

Keywords: anorexia nervosa, comorbidity, network analysis, obsessive-compulsory disorder, risk genes

BMC-7

Identification of copy number polymorphisms associated with early trauma in obsessive-compulsive disorder

Guan Ning Lin^{1,2}, Xuemei Wang^{1,3}, Weichen Song¹, Margit Burmeister⁴, Haipeng Li⁵, Chen Ming⁵, Qing Fan¹, Weidong Tian⁶, Xinran Dong⁶, Jue Chen¹, Kaida Jiang¹, Zeping Xiao¹, Donghong Cui^{1,2}

¹Shanghai Mental Health Center, Shanghai, Jiao Tong University School of Medicine, School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, PR China

²Shanghai Key Laboratory of Psychotic Disorders, Shanghai, PR China

³The First Affiliated Hospital of Soochow University, Suchow, PR China

⁴Molecular & Behavioral Neuroscience Institute, University of Michigan, Ann Arbor, Michigan, USA

⁵CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, PR China

⁶Institute of Biostatistics, School of Life Sciences, Fudan University, Shanghai, PR China

Corresponding authors: Guan Ning Lin, Zeping Xiao, Donghong Cui

Abstract

Background: Both hereditary and environmental factors have a role in Obsessive-compulsive disorder (OCD) etiology. However, the exact impact of copy number variations (CNVs) and early trauma experience on the development of OCD remains to be quantified.

Results: To address these challenges, We first investigated the role of genome-wide CNVs in 30 male early-onset OCD subjects and 30 well-matched male healthy controls. Candidate polymorphic CNVs potentially associated with OCD were then validated in an independent cohort of 577 OCD subjects and 608 healthy controls. A copy number gain of a fragment at 1p31.1 (near the neuronal growth regulator 1 gene, NEGR1) and a copy number loss of a fragment at 20p13 (within the signal-regulatory protein beta 1 gene, SIRPB1) were found associated with OCD. In addition to the genetic vulnerability, OCD subjects, especially those with the early-onset, showed significantly higher scores on the Early Trauma Inventory-Short Form (ETI-SF). We also found a significant three-way interaction between the CNVs in 1p31.1 and 20p13 and emotional abuse in the development of OCD.

Conclusion: These results demonstrate that CNVs in 1p31.1 and 20p13 and their interaction with early trauma experience confer the risk of OCD.

Keywords: obsessive-compulsive disorder, copy number variation, childhood abuse, trauma, 1p31.1

BMC-8

MicroRNA profiles in sorghum exposed to individual or combined abiotic stresses

Ramanjulu Sunkar¹, Chandra Obul Reddy Puli¹, Yun Zheng², Yong-Fang Li¹, Guru Jagadeeswaran¹, Angbaji Suo², Bingbing Jiang², Pradeep Sharma¹, Robert Mann¹, Govindan Ganesan¹, Nirmali Gogoi¹, Asha Srinivasan¹, Aparna Kakani³, Vijaya Gopal Kakani³, Abdelali Barakat⁴

¹Department of Biochemistry and Molecular Biology, Oklahoma State University, Stillwater, OK 74078 USA

²Faculty of Life Science and Technology, Kunming University of Science and Technology 727, South Jingming Road, Kunming, Yunnan, China

³Department of Plant and Soil Sciences, Oklahoma State University, Stillwater, OK 74078 USA

⁴Department of Biology, University of South Dakota, Vermillion, SD 57069 USA

Corresponding author: Ramanjulu Sunkar

Abstract

Background: Sorghum is largely grown for food, fodder and for biofuel production. It is mostly grown in semi-arid tropical and sub-tropical regions where the drought or high temperature or their combination co-occur in the field. In plants, microRNAs (miRNAs) are integral to the gene regulatory networks that control almost all biological processes including adaptation to stress conditions. Thus far, plant miRNA profiles under separate drought or heat stresses have been reported but not under combined drought and heat. In this study, we report miRNA profiles in leaves of sorghum exposed to individual drought or heat, and more importantly to the combined drought and heat.

Results: The altered miRNA profiles in response to individual stresses (drought or heat) or combined drought and heat in sorghum were identified by constructing and sequencing small RNA libraries. The bioinformatic analysis of small RNAs has revealed the expression of approximately 30 conserved miRNA families represented by 81 individual miRNAs, 47 new homologs of less conserved or sorghum-specific miRNA families as well as 11 novel miRNA families. Of these, 26 miRNAs (21 conserved miRNAs belonging to 15 families and five novel miRNAs) were found to be differentially regulated in response to stress treatments. Interestingly, the miRNA profiles were almost similar between heat and the combined drought and heat stresses. By contrast, under drought, the extent of regulation was small compared to either heat alone or combined drought and heat treatments. We also have sequenced and analyzed degradome profiles to identify targets for the miRNAs in sorghum. On the basis of observed characteristic cleavages on the target transcripts, 48 genes (39 for the conserved miRNA families, and additional 9 for six other non-conserved miRNAs) were identified as targets for the miRNAs in sorghum.

Conclusions: The comparative profiling revealed that the miRNA regulation was stronger under heat or combination of heat and drought compared to the drought alone. Using degradome sequencing, 48 genes were confirmed as targets for the miRNAs in sorghum. Overall, this study provides a frame work for understanding of the miRNA-guided gene regulations under individual drought or heat or their combination.

Keywords: sorghum, drought, miRNAs, heat, degradome

BMC-9

Characterizing promoter and enhancer sequences by a deep learning method

Xin Zeng¹, Sung-Joon Park², Kenta Nakai^{1,2}

¹Department of Computational Biology and Medical Science, the University of Tokyo, Japan

²Human Genome Center, the Institute of Medical Science, the University of Tokyo, Japan

Corresponding author: Kenta Nakai

Abstract

Background: Promoters and enhancers are well known regulatory elements modulating gene expression. As confirmed by high-throughput sequencing technologies, these regulatory elements are bidirectionally transcribed. That is, promoters produce stable mRNA in the sense direction and unstable RNA in the antisense direction, while enhancers transcribe unstable RNA in both directions. Although it is thought that enhancers and promoters share a similar architecture of transcription start sites (TSSs), how the transcriptional machinery distinctly uses these genomic regions as promoters or enhancers remains unclear. To address this issue, we developed a deep learning method by utilizing a Convolutional Neural Network (CNN) and the saliency algorithm.

Results: In comparison with other classifiers, our CNN presented higher predictive performance, suggesting the overarching importance of the high-order sequence features, captured by the CNN. Moreover, our method revealed that there are substantial sequence differences between the enhancers and promoters. Remarkably, the 20–120bp downstream regions from the center of bidirectional TSSs seemed to contribute to the RNA stability. These regions in promoters tend to have a larger number of guanines and cytosines, compared with those in enhancers, and this feature contributed to the classification of the regulatory elements.

Conclusions: Our CNN-based method can capture the complex TSS architectures. We found that the genomic regions around TSSs for promoters and enhancers contribute to RNA stability and show some GC-biased characteristic as a critical determinant for promoter TSSs.

Keywords: -

BMC-10

The evolutionary landscape of long non-coding RNA in green plants

Xiangna Hong^{1,3}, Yan Zhu^{1,2}, Xuan Li^{1,2}

¹Key Laboratory of Synthetic Biology, CAS Center for Excellence in Molecular Plant Sciences/Institute of Plant Physiology and Ecology, Chinese Academy of Sciences, Shanghai 200032, China

²University of Chinese Academy of Sciences, Beijing 100049, China

³Henan University, Kaifeng 475001, China

Corresponding author: Xuan Li

Abstract

Long non-coding RNAs (lncRNAs) have been widely identified in plants. Though the number of lncRNAs identified in plants was increasing, only a few of them were found to play roles in biological processes in model plants, whereas most of them remained function unknown. An integrated analysis using whole transcriptome sequencing on eight plant species was conducted to reveal the distribution, molecular features, and evolutionary patterns of lncRNAs in plant. LncRNAs have more simple structures than mRNA. Although lncRNA probably used the same splicing mechanism with mRNA, inefficient splicing of lncRNA was common across plant species. LncRNAs were more tolerant for TE inserting than protein gene coding sequences, and the transcription of some lncRNAs may be driven by TE promoters. Among divergent plant species, lncRNAs were consistent on expression patterns and molecular features, such as transcript length, exon number, GC content, expressional levels, etc. However, the sequence conservation was limited even in closely related species. Some lncRNAs contained conserved patches, which may play important biological roles in plants. Most highly and specially expressed lncRNAs formed co-expression pattern with protein genes, in which their functions were closely related to their expression tissues. Our study offered insight into the functions, origin and evolution of lncRNAs across the evolution landscape of divergent plant species.

Keywords: -

BMC-11

Single-cell sequencing data analysis with dimension reduction based on robust and sensitive genes

Zechuan Chen^{1,2}, Zeruo Yang³, Yingying Cao⁴, Xiaojun Yuan¹, Xiaoming Zhang², Pei Hao²

¹College of Life Sciences, Shanghai University, Shanghai, China

²Key Laboratory of Molecular Virology & Immunology, Institut Pasteur of Shanghai, Chinese Academy of Sciences, Shanghai, China

³Natural Medicine Institute of Zhejiang YangShengTang Co., Ltd. No. 181, Geyazhuang, Xihu District, Hangzhou, Zhejiang, China

⁴Bioinformatics and Computational Biophysics, Faculty of Biology and Center for Medical Biotechnology, University of Duisburg-Essen, Essen, Germany

Corresponding author: Pei Hao

Abstract

Background: Single-cell RNA sequencing (scRNA-seq) is a powerful technology in studying cell development and differentiation. Cell types and differentiation states are often distinguished via differential gene expression, for which many scRNA-seq bioinformatics tools are available. These tools utilize the concept of highly variable genes (HVGs) to differentiate cell types and states. However, we have discovered that a group of genes, sensitive to environmental stimuli, have high coefficients of variation (CV), and often generate overwhelming expression change signals that are noises which negatively impact cell type grouping.

Result: In this study, we developed a method to identify such noises, and we termed this group of genes as the “sensitive genes” by incorporating CV-ranking within unsupervised clustering and the Shannon index. We demonstrated that after the removal of these genes before cell type clustering, the result of unsupervised clustering was closer to the true cell-type labels when compared to the first-time clustering result. We also validated the reliability of our method in 11 different types of human tissues’ scRNA-seq data sets, and the results showed that the sensitive genes were enriched in pathways related to cellular stress response in most of the data sets.

Conclusion: Our study revealed the prevalence of stochastic gene expression patterns in most types of cells, compared the differences among cell marker genes, housekeeping genes (HK genes), and sensitive genes, demonstrated the similarities of functions of sensitive genes in various scRNA-seq data sets, and improved the results of unsupervised clustering towards the ground-truth labels. We hope our method will provide new insights into the reduction of data noise in scRNA-seq data analysis and contribute to the development of better scRNA-seq unsupervised clustering algorithms in the future.

Keywords: -

BMC-12

Rapid screening and identification of viral pathogens in metagenomic data

Shiyang Song^{1,†}, Liangxiao Ma^{2,†}, Xintian Xu¹, Han Shi³, Xuan Li³, Yuanhua Liu¹, Pei Hao¹

¹Key Laboratory of Molecular Virology and Immunology, Institut Pasteur of Shanghai, Center for Biosafety Mega-Science, Chinese Academy of Sciences, Shanghai 200031, China

²Bio-Med Big Data Center, Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institute of Nutrition and Health, Shanghai 200031, China

³Key Laboratory of Synthetic Biology, CAS Center for Excellence in Molecular Plant Sciences, Chinese Academy of Sciences, Shanghai 200032, China

[†]These authors contributed equally to this work

Corresponding authors: Yuanhua Liu and Pei Hao

Abstract

Background: Virus screening and viral genome reconstruction are urgent and crucial for rapid identification of a viral pathogen, tracing its source, and understanding its pathogenesis when a viral outbreak occurs. The Next-generation sequencing (NGS) nowadays has become the standard technique for metagenomic study, and is an essential approach for detection and identification of viral pathogen from patient and environment samples. Despite the availability of many software, data analysis requires human operations step by step. A mature pipeline is necessary when thousands of samples are waiting for quick identification of viral sequences reconstruction.

Results: In this paper, 1) we present a rapid and accurate workflow to screen metagenomic sequencing data for viral pathogens and other compositions, and enable a reference-based assembler to reconstruct viral genomes. 2) We tested our workflow using several metagenomic datasets, i.e. SARS-CoV-2 patient sample NGS data, pangolins tissues NGS data, MERS-infected cells NGS data, etc. The workflow was demonstrated to detect and identify the target viruses from large NGS metagenomic data with accuracy and efficiency. Our workflow is flexible to work with a broad range of NGS datasets from small (kb) to large (100 Gb), which took from a few minutes to hours to complete its task. At the same time, our workflow automatically generates a report that incorporates visual feedback on the statistics of metagenomic data quality, the identities and compositions of host and viral sequences, identified viral pathogens and compositions, and assembled viral pathogen genomes based on their closest references.

Conclusions: Overall our system enables the rapid screen and identification of viral pathogens from metagenomic data, providing a critical mean in support of viral pathogen research in time of pandemic. The visualized report contains information from raw sequences' quality to a reconstructed viral sequences, which enables non-professional people screening for viruses in their samples by themselves. With the speed advantage, this pipeline can also be used in large amounts of daily viruses screening tasks.

Keywords: -

BMC-13

Transcriptional dynamics of transposable elements when converting fibroblast cells of *Macaca mulatta* to neuroepithelial stem cells

Dahai Liu^{1,†}, Li Liu^{2,†}, Kui Duan^{2,†}, Junqiang Guo³, Shipeng Li², Zhigang Zhao², Xiaotuo Zhang⁴, Nan Zhou³, Yun Zheng^{2,3}

¹Department of Basic Medicine and Biomedical Engineering, School of Stomatology and Medicine, Foshan University, Foshan, Guangdong 528000, China

²Yunnan Key Lab of Primate Biomedicine Research, Institute of Primate Translational Medicine, Kunming University of Science and Technology, Kunming, Yunnan 650500, China

³Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, Yunnan 650500, China

⁴State Key Laboratory of Genetic Engineering, Collaborative Innovation Center for Genetics and Development, Institute of Plant Biology, School of Life Sciences, Fudan University, Shanghai 200438, China

[†]These authors contributed equally to this work

Corresponding author: Yun Zheng

Abstract

Background: Transposable elements (TE) account for more than 50% of human genome. It has been reported that some types of TEs are dynamically regulated in the reprogramming of human cell lines. However, it is largely unknown whether some TEs in *Macaca mulatta* are also regulated during the reprogramming of cell lines of monkey.

Results: Here, we systematically examined the transcriptional activities of TEs during the conversion of *Macaca mulatta* fibroblast cells to neuroepithelial stem cells (NESCs). Hundreds of TEs were dynamically regulated during the reprogramming of *Macaca mulatta* fibroblast cells. Furthermore, 48 Long Terminal Repeats (LTRs), as well as some integrase elements, of *Macaca* endogenous retrovirus 3 (MacERV3) were transiently activated during the early stages of the conversion process, some of which were further confirmed with PCR experiments. These LTRs were potentially bound by critical transcription factors for reprogramming, such as KLF4 and ETV5.

Conclusion: These results suggest that the transcription of TEs are delicately regulated during the reprogramming of *Macaca mulatta* fibroblast cells. Although the family of ERVs activated during the reprogramming of fibroblast cells in *Macaca mulatta* is different from those in the reprogramming of human fibroblast cells, our results suggest that the activation of some ERVs is a conserved mechanism in primates for converting fibroblast cells to stem cells.

Keywords: -

BMC-14

Epigenetic interplay between methylation and miRNA in bladder cancer: Focus on isoform expression

Manu Shivakumar^{1,2,†}, Seonggyun Han^{3,†}, Younghee Lee^{3,4}, Dokyoon Kim^{1,2,4}

¹Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States

²Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA, United States

³Department of Biomedical Informatics, University of Utah School of Medicine, Salt Lake City, Utah

⁴Huntsman Cancer Institute, Salt Lake City, Utah

[†]These authors contributed equally to this work

Corresponding authors: Dokyoon Kim and Younghee Lee

Abstract

Background: Various epigenetic factors are responsible for the non-genetic regulation on gene expression. The epigenetically dysregulated oncogenes or tumor suppressors by miRNA and/or DNA methylation are often observed in cancer cells. Each of these epigenetic regulators has been studied well in cancer progressions; however, their mutual regulatory relationship in cancer still remains unclear. In this study, we propose an integrative framework to systematically investigate epigenetic interactions between miRNA and methylation at the alternatively spliced mRNA level in bladder cancer. Each of these epigenetic regulators has been studied well in cancer progressions; however, their mutual regulatory relationship in cancer still remains unclear.

Methods: DNA methylation, miRNA, mRNA isoform, and clinical datasets for bladder cancer were retrieved from Xena browser. The methylation and miRNA expression data were used to develop an integrative framework to identify epigenetic interactions that are associated with isoform expression. Further, the identified interactions were split based on the high and low expression of miRNA and methylation to examine a clinical implication among the group. Additionally, the methylation and miRNA were categorized based on correlation to the isoform expression and location of the methylation probe site to investigate different patterns observed in the interactions.

Results: The integrative analyses yielded 136 significant combinations (methylation, miRNA and isoform). Further, overall survival analysis on the 136 combinations based on methylation and miRNA, high and low expression groups resulted in 13 combinations associated with survival. Additionally, different interaction patterns were examined.

Conclusions: Our study provides a higher resolution of molecular insight into the crosstalk between two epigenetic factors, DNA methylation and miRNA. Given the importance of epigenetic interactions and alternative splicing in cancer, it is timely to identify and understand the underlying mechanisms based on epigenetic markers and their interactions in cancer, leading to alternative splicing with primary functional impact.

Keywords: -

BMC-15

Instance-based error correction for short reads of disease-associated genesXuan Zhang¹, Yuansheng Liu¹, Zuguo Yu², Michael Blumenstein³, Gyorgy Hutvagner³,
Jinyan Li¹¹Advanced Analytics Institute, Faculty of Engineering and IT, University of Technology Sydney, Ultimo, 2007 NSW, Australia²Key Laboratory of Intelligent Computing and Information Processing of the Ministry of Education and Hunan Key Laboratory for Computation and Simulation in Science and Engineering, Xiangtan University, 411105 Xiangtan, China³Faculty of Engineering and IT, University of Technology Sydney, 2007 NSW, Australia**Corresponding author:** Jinyan Li**Abstract**

Background: Genomic reads from sequencing platforms contain random errors. Global correction algorithms have been developed, aiming to rectify all possible errors in the reads using generic genome-wide patterns. However, the non-uniform sequencing depths hinder the global approach to conduct effective error removal. As some genes may get under-corrected or over-corrected by the global approach, we conduct instance-based error correction for short reads of disease-associated genes or pathways. The paramount requirement is to ensure the relevant reads, instead of the whole genome, are error-free to provide significant benefits for SNP (single-nucleotide polymorphism) or variant calling studies on the specific genes.

Results: To rectify possible errors in the short reads of disease-associated genes, our novel idea is to exploit local sequence features and statistics directly related to these genes. Extensive experiments are conducted in comparison with state-of-the-art methods on both simulated and real datasets of lung cancer associated genes (including single-end and paired-end reads). The results demonstrated the superiority of our method with the best performance on precision, recall and gain rate, as well as on sequence assembly results (*e.g.* N50, the length of contig and contig quality).

Conclusion: Instance-based strategy makes it possible to explore fine-grained patterns focusing on specific genes, providing high precision error correction and convincing gene sequence assembly. SNP case studies show that errors occurring at some traditional SNP areas can be accurately corrected, providing high precision and sensitivity for investigations on disease-causing point mutations.

Keywords: -

BMC-16

Data integration and evolutionary analysis of long non-coding RNAs in 25 flowering plants

Shiye Sang^{1,2}, Wen Chen¹, Di Zhang^{1,2}, Xuan Zhang^{1,2}, Wenjing Yang^{1,2}, Changning Liu^{1,3,4}

¹CAS Key Laboratory of Tropical Plant Resources and Sustainable Use, Xishuangbanna Tropical Botanical Garden, Chinese Academy of Sciences, Menglun, Mengla, Yunnan 666303, China

²College of Life Sciences, University of Chinese Academy of Sciences, Beijing 100049, China

³Center of Economic Botany, Core Botanical Gardens, Chinese Academy of Sciences, Menglun, Mengla, Yunnan 666303, China

⁴The Innovative Academy of Seed Design, Chinese Academy of Sciences, Menglun, Mengla, Yunnan 666303, China

Corresponding author: Changning Liu

Abstract

Background: Long non-coding RNAs (lncRNAs) play vital roles in many important biological processes in plants. Currently, a large fraction of plant lncRNA studies center at lncRNA identification and functional analysis. Only a few plant lncRNA studies focus on understanding their evolutionary history, which is crucial for an in-depth understanding of lncRNAs. Therefore, the integration of large volumes of plant lncRNA data is required to deeply investigate the evolution of lncRNAs.

Results: We present a large-scale evolutionary analysis of lncRNAs in 25 flowering plants. In total, we identified 199,796 high-confidence lncRNAs through data integration analysis, and grouped them into 5,497 lncRNA orthologous families. Then, we divided the lncRNAs into groups based on the degree of sequence conservation, and quantified the various characteristics of 756 conserved *Arabidopsis thaliana* lncRNAs. We found that compared with non-conserved lncRNAs, conserved lncRNAs might have more exons, longer sequence length, higher expression levels, and lower tissue specificities. Functional annotation based on the *A. thaliana* coding-lncRNA gene co-expression network suggested potential functions of conserved lncRNAs including autophagy, locomotion, and cell cycle. Enrichment analysis revealed that the functions of conserved lncRNAs were closely related to the growth and development of the tissues in which they were specifically expressed.

Conclusions: Comprehensive integration of large-scale lncRNA data and construction of a phylogenetic tree with orthologous lncRNA families from 25 flowering plants was used to provide an oversight of the evolutionary history of plant lncRNAs including origin, conservation, and orthologous relationships. Further analysis revealed a differential characteristic profile for conserved lncRNAs in *A. thaliana* when compared with non-conserved lncRNAs. We also examined tissue specific expression and the potential functional roles of conserved lncRNAs. The results presented here will further our understanding of plant lncRNA evolution, and provide the basis for further in-depth studies of their functions.

Keywords: -

BMC-17

SPECTRA – A tool for enhanced brain wave signal recognition

Shiu Kumar¹, Tatsuhiko Tsunoda^{2,3,4}, Alok Sharma^{2,3,5,6}

¹School of Electrical and Electronics Engineering, Fiji National University, Suva, Fiji

²Laboratory of Medical Science Mathematics, RIKEN Center for Integrative Medical Sciences, Yokohama 230-0045, Japan

³Department of Medical Science Mathematics, Medical Research Institute, Tokyo Medical and Dental University (TMDU), Tokyo 113-8510, Japan

⁴Laboratory of Medical Science Mathematics, Department of Biological Sciences, Graduate School of Science, University of Tokyo, Tokyo 113-0033, Japan

⁵School of Engineering and Physics, The University of the South Pacific, Suva, Fiji

⁶Institute for Integrated and Intelligent Systems, Griffith University, Nathan, Brisbane, QLD, Australia

Corresponding author: Shiu Kumar

Abstract

Introduction: Brain wave signal recognition has gained increased attention in neuro-rehabilitation applications. This has driven the development of brain-computer interface (BCI) systems. Brain wave signals are acquired using electroencephalography (EEG) sensors, processed and decoded to identify the category to which the signal belongs. Once the signal category is determined, it can be used to control external devices. However, the success of such a system essentially relies on significant feature extraction and classification algorithms. One of the commonly used feature extraction techniques for BCI systems is the common spatial pattern (CSP).

Methods: In this study, we propose an effective spatial-frequency-temporal feature extraction (SPECTRA) predictor that utilizes the CSP-TSM (tangent space mapping) approach for obtaining features that are more separable. CSP-TSM and common spatio-spectral pattern (CSSP)-TSM features are extracted from multiple temporal delayed windows and significant features are selected based on the F-score ranking. The significant features are then used to recognize the MI electroencephalography (EEG) signal using a support vector machine (SVM) classifier.

Results: The performance of the proposed method is analysed using three public benchmark datasets. Our proposed predictor outperformed other competing methods achieving lowest average error rates of 8.55%, 17.90% and 20.26%, and highest average kappa coefficient values of 0.829, 0.643 and 0.595 for BCI Competition III dataset IVa, BCI Competition IV dataset I and BCI Competition IV dataset IIb, respectively.

Conclusions: Our proposed SPECTRA predictor effectively finds features that are more separable and shows improvement in brain wave signal recognition that can be instrumental in developing improved real-time BCI systems that are computationally efficient.

Keywords: -

BMC-18

Boosting scRNA-seq data clustering by cluster-aware feature weightingRui-Yi Li¹, Jihong Guan¹, Shuigeng Zhou²¹Department of Computer Science and Technology, Tongji University, 4800 Caoan Road, 201804 Shanghai, China²Shanghai Key Lab of Intelligent Information Processing, and School of Computer Science, Fudan University, 220 Handan Road, 200433 Shanghai, China**Corresponding author:** Shuigeng Zhou**Abstract**

Background: The rapid development of single-cell RNA sequencing (scRNA-seq) enables the exploration of cell heterogeneity, which is usually done by scRNA-seq data clustering. The essence of scRNA-seq data clustering is to group cells by measuring the similarities among genes/transcripts of cells. And the selection of features for cell similarity evaluation is of great importance, which will significantly impact clustering effectiveness and efficiency.

Results: In this paper, we propose a novel method called CaFew to select genes based on cluster-aware feature weighting. By optimizing the clustering objective function, CaFew obtains a feature weight matrix, which is further used for feature selection. The genes have large weights in at least one cluster or the genes whose weights vary greatly in different clusters are selected. Experiments on 8 real scRNA-seq datasets show that CaFew can obviously improve the clustering performance of existing scRNA-seq data clustering methods. Particularly, the combination of CaFew with SC3 achieves the state-of-art performance. Furthermore, CaFew also benefits the visualization of scRNA-seq data.

Keywords: -

BMC-19

Forecasting the spread of COVID-19 using LSTM network

Shiu Kumar¹, Ronesh Sharma¹, Tatsuhiko Tsunoda^{2,3,4,†}, Thirumananseri Kumarevel^{5,†}, Alok Sharma^{3,4,6,†}

¹School of Electrical and Electronics Engineering, Fiji National University, Suva, Fiji

²Laboratory for Medical Science Mathematics, Department of Biological Sciences, Graduate School of Science, University of Tokyo, Tokyo 113-0033, Japan

³Laboratory for Medical Science Mathematics, RIKEN Center for Integrative Medical Sciences, Yokohama 230-0045, Japan

⁴Department of Medical Science Mathematics, Medical Research Institute, Tokyo Medical and Dental University (TMDU), Tokyo 113-8510, Japan

⁵Laboratory for Transcription Structural Biology, RIKEN Center for Biosystems Dynamics Research, 1-7-22 Suehiro, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan

⁶Institute for Integrated and Intelligent Systems, Griffith University, Nathan, Brisbane, QLD, Australia

†Last Authors

Corresponding author: Shiu Kumar

Abstract

Background: The novel coronavirus (COVID-19) is caused by severe acute respiratory syndrome coronavirus 2, and within a few months, it has become a global pandemic. This forced many affected countries to take stringent measures such as complete lockdown, shutting down businesses and trade, as well as travel restrictions, which has had a tremendous economic impact. Therefore, having knowledge and foresight about how a country might be able to contain the spread of COVID-19 will be of paramount importance to the government, policy makers, business partners and entrepreneurs.

Method: To help social and administrative decision making, we have developed a computational model using a long short-term memory network that uses currently available data on daily new cases for forecasting when a country might be able to contain the spread of COVID-19. The prediction is based on the current trend and does not account for future restrictions that may be placed on said countries.

Results: The results obtained are promising as we validate our prediction model using New Zealand's data since they have been able to contain the spread of COVID-19 and bring the daily new cases tally to zero. Our proposed forecasting model was able to correctly predict the dates within which New Zealand was able to contain the spread of COVID-19. Similarly, the proposed model has been used to forecast the dates when other countries would be able to contain the spread of COVID-19.

Keywords: -

BMC-20

PIKE-R2P: Protein-protein interaction network-based knowledge embedding with graph neural network for single-cell RNA to protein prediction

Xinnan Dai¹, Fan Xu¹, Shike Wang¹, Piyushkumar A. Mundra², Jie Zheng¹

¹School of Information Science and Technology, Shanghai Tech University, Shanghai, China, 393 Middle Huaxia Road, Pudong District, 201210 Shanghai, China.

²Molecular Oncology Group, Cancer Research UK Manchester Institute, The University of Manchester, Alderley Park, Manchester, United Kingdom.

Corresponding author: Jie Zheng

Abstract

Computational prediction of protein abundance using bulk RNA expression data is challenging as both entities are generally not measured on the collected sample, introducing intrinsic measurement errors. Recent advances in simultaneous measurements of RNA expressions and abundance of multiple proteins at single-cell level provide a unique opportunity to build machine learning models predicting protein abundance from the RNA expression data. Here, we present this task in a multi-label RNA to protein prediction framework where multiple proteins are linked together at a single-cell state level. The proposed algorithm, PIKE-R2P, incorporates protein-protein interactions and prior knowledge embedding into a graph neural network, and has shown significant improvement in the simultaneous prediction of the abundances of multiple proteins.

Keywords: -

BMC-21

BREC: An R package/Shiny app for automatically identifying heterochromatin boundaries and estimating local recombination rates along chromosomes

Yasmine Mansour^{1,2}, Annie Chateau², Anna-Sophie Fiston-Lavier¹

¹Genomics Dept, Institute of Evolution Science of Montpellier (ISEM), Montpellier, France

²Informatics Dept, Laboratory of Computer Science, Robotics and Microelectronics of Montpellier (LIRMM), Montpellier, France

Corresponding author: Yasmine Mansour

Abstract

Background: Meiotic recombination is a vital biological process playing an essential role in the genome's structural and functional dynamics. Genomes exhibit highly various recombination profiles along chromosomes associated with several chromatin states. However, heterochromatin boundaries are not available nor easily provided for non-model organisms, especially for newly sequenced ones. Hence, we miss accurate local recombination rates necessary to address evolutionary questions.

Results: Here, we propose an automated computational tool, based on the Marey maps method, allowing to identify heterochromatin boundaries along chromosomes and estimating local recombination rates. Our method, called BREC (heterochromatin B oundaries and REC ombination rate estimates) is non-genome-specific, running even on non-model genomes as long as genetic and physical maps are available. BREC is based on pure statistics and is data-driven, implying that good input data quality remains a strong requirement. Therefore, a data pre-processing module (data quality control and cleaning) is provided. Experiments show that BREC handles different markers' density and distribution issues.

Conclusions: BREC's heterochromatin boundaries have been validated with cytological equivalents experimentally generated on the fruit fly *Drosophila melanogaster* genome, for which BREC returns congruent corresponding values. Also, BREC's recombination rates have been compared with previously reported estimates. Based on the promising results, we believe our tool has the potential to help bring data science into the service of genome biology and evolution. We introduce BREC within an R-package and a Shiny web-based user-friendly application yielding a fast, easy-to-use, and broadly accessible resource. BREC R-package is available at the GitHub repository <https://github.com/ymansour21/BREC>.

Keywords: -

BMC-22

Identification of cell states using super-enhancer RNA

Yueh-Hua Tu^{1,2,3}, Hsueh-Fen Juan^{1,4}, Hsuan-Cheng Huang³

¹Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, Taipei 106, Taiwan

²Bioinformatics Program, Taiwan International Graduate Program, Institute of Information Science, Academia Sinica, Taipei, Taiwan

³Institute of Biomedical Informatics, National Yang-Ming University, Taipei 112, Taiwan

⁴Department of Life Science, National Taiwan University, Taipei 106, Taiwan

Corresponding authors: Hsueh-Fen Juan and Hsuan-Cheng Huang

Abstract

Background: A new class of regulatory elements called super-enhancers, comprised of multiple neighboring enhancers, have recently been reported to be the key transcriptional drivers of cellular, developmental, and disease states.

Results: Here, we defined super-enhancer RNA as highly expressed enhancer RNAs that are transcribed from a cluster of localized genomic regions. Using the cap analysis of gene expression sequencing data from FANTOM5, we systematically explored the enhancer and messenger RNA landscapes in hundreds of different cell types in response to various environments. Applying non-negative matrix factorization (NMF) to superenhancer RNA profiles, we found that different cell types were well classified. In addition, through the NMF of individual time-course profiles from a single cell-type, super-enhancer RNAs were clustered into several states with progressive patterns. We further investigated the enriched biological functions of the proximal genes involved in each pattern, and found that they were associated with the corresponding developmental process.

Conclusions: The proposed super-enhancer RNAs can act as a good alternative, without the complicated measurement of histone modifications, for identifying important regulatory elements of cell type specification and identifying dynamic cell states.

Keywords: -

BMC-23

Temporal expression study of miRNAs in crown tissues of winter wheat grown under natural growth condition

Menglei Wang^{1,†}, Chenhui Yang^{1,†}, Kangning Wei^{1,†}, Miao Zhao¹, Liqiang Shen, Jie Ji¹, Li Wang^{1,2}, Daijing Zhang¹, Junqiang Guo⁴, Yun Zheng⁵, Juanjuan Yu^{1,2}, Mo Zhu^{1,2}, Haiying Liu¹, Yong-Fang Li^{1,2}

¹College of Life Sciences, Henan Normal University, Xinxiang, Henan, 453007, China

²Henan International Joint Laboratory of Agricultural Microbial Ecology and Technology, Henan Normal University, Xinxiang 453007, China

³Jindal School of Management, University of Texas at Dallas, 800 W Campbell RD, Richardson, 14 TX, 75080, USA

⁴Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, Yunnan, 650500, China

⁵Yunnan Key Laboratory of Primate Biomedical Research, Institute of Primate Translational Medicine, Kunming University of Science and Technology, Kunming, 650500, China

[†]These authors contributed equally to this work

Corresponding author: Yong-Fang Li

Abstract

Background: Winter wheat requires prolonged exposure to low temperature to initiate flowering (vernalization). Shoot apical meristem of crown is the site of cold perception, which produces leaf primordia during vegetative growth and then develops into floral primordia at the initiation of reproductive phase. Although many crucial genes essential for winter wheat cold acclimation and floral initiation have been revealed, the importance of microRNA (miRNA) mediated post-transcriptional regulation in the crown is not well understood. To understand the potential roles of miRNAs in crown tissue, we performed a temporal expression study of miRNAs in crown tissues at three-leaf stage, winter dormancy stage, spring greenup stage and jointing stages of winter wheat grown in field under natural growth condition.

Results: Totally, 348 miRNAs belonging to 298 miRNA families were identified in wheat crown tissues. 92 differentially expressed miRNAs (DEMs) were found significantly regulated from three-leaf stage to jointing stage. Most of these DEMs were highly expressed at three-leaf stage and winter dormancy stage, then declined in later stages. Six DEMs, including miR156a-5p, were significantly induced at winter dormancy stage. Eleven DEMs, including miR159a.1, miR390a-5p, miR393-5p, miR160a-5p and miR1436, were highly expressed at the green-up stage. Twelve DEMs were significantly induced at jointing stage, such as miR394a, miR172a-5p, miR319b-3p and miR9676-5p. Moreover, 14 novel target genes of wheat or Poideae specific miRNAs were verified using RLM-5'RACE assay, notably, 6 mTERF s and 2 Rf1 genes, which are associated with mitochondrial gene expression, were confirmed as targets of 3 wheat specific miRNAs.

Conclusions: The present study not only confirmed the known miRNAs associated with cold acclimation and floral initiation, but also identified a number of wheat or Pooideae specific miRNAs critical for winter wheat cold acclimation and floral development. Most importantly, this study provides experimental evidence that miRNA could regulate mitochondria gene expression via targeting mTERF and Rf1 genes. Our study provides valuable information for further exploration of the mechanism of miRNA mediated post-transcriptional regulation during winter wheat vernalization and inflorescent initiation.

Keywords: -

BMC-24

Ontology-based annotation, modeling, and analysis of phenotypes and comorbidities in COVID-19 patients

Yang Wang^{1,2}, Fengwei Zhang¹, Hong Yu^{1,2}, Yongqun He³, Xianwei Ye^{1,2}

¹Guizhou University School of Medicine, Guiyang, Guizhou 550025, China

²Department of Pulmonary and Critical Care Medicine, Guizhou Provincial People's Hospital and NHC Key Laboratory of Immunological Diseases, People's Hospital of Guizhou University, Guiyang, Guizhou 550002, China

³Center for Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, United States

Corresponding authors: Yongqun He and Xianwei Ye

Abstract

Background: COVID-19 pandemic has caused an unpredictable and devastating disaster to the public health worldwide. Common phenotypes initially identified include fever, cough, shortness of breath, and fatigue. With more cases identified, other clinical phenotypes (e.g., loss of smell, loss of tastes) have been gradually recognized. Compared with discharged or cured patients, severe or died patients often have one or more comorbidities, such as hypertension, diabetes, and cardiovascular disease. Ontology provides a standardized integrative way for phenotype and disease modeling. In this study, we systematically surveyed the literature, collected clinical phenotypes and comorbidities in COVID-19 patients, and ontologically modeled and analyzed their associations.

Results: Commonly occurring 18 phenotypes in COVID-19 patients were first classified into different groups based on the Human Phenotype Ontology (HPO). Fever, cough, and the loss of smell and taste were ranked as the highest phenotype in China, the United States, and Italy, respectively. The patients from Europe and USA appeared to have higher nervous phenotypes (loss of smell, loss of taste, and headache) and abdominal phenotypes (nausea, vomiting, abdominal pain, and diarrhea) than patients from Asia. A total of 22 comorbidities were found to commonly exist in COVID-19 patients. Patients with the comorbidities such as diabetes and kidney failure had worse outcomes compared with those without these comorbidities. The knowledge of the phenotypes and comorbidities shown in COVID-19 patients was further modeled and represented in the Coronavirus Infectious Disease Ontology (CIDO), and semantic relations were generated to logically link COVID-19 with the phenotypes and comorbidities. A SPARQL query was generated to demonstrate the usage of CIDO for querying typical phenotype groups.

Conclusions: COVID-19 phenotypes and comorbidities were surveyed and ontologically classified using HPO. Many differential phenotypes and comorbidities were found in COVID-19 patients in different countries. The semantic relations among COVID-19, phenotypes, and comorbidities were modeled and represented in CIDO, supporting advanced reasoning and knowledge query.

Keywords: -

BMC-25

Development of the International Classification of Diseases Ontology (ICDO) and its application for COVID-19 diagnostic data analysis

Ling Wan^{1,2}, Justin Song³, Virginia He⁴, Yongqun He¹

¹University of Michigan Medical School, Ann Arbor, MI 48109, USA

²OntoWise, Nanjing, Jiangsu, China

³Cranbrook Kingswood Upper School, Bloomfield Hills, MI 48304, USA

⁴Huron High School, Ann Arbor, MI 48105, USA

Corresponding author: Yongqun He

Abstract

Background: The 10th and 9th revisions of the International Statistical Classification of Diseases and Related Health Problems (ICD10 and ICD9) have been adopted worldwide as a well-recognized norm to share codes for diseases, signs and symptoms, abnormal findings, etc. The international Consortium for Clinical Characterization of COVID-19 by EHR (4CE) website stores diagnosis COVID-19 disease data using ICD10 and ICD9 codes. However, the ICD systems are difficult to decode due to their many shortcomings, which can be addressed using ontology.

Results: We have developed an ICD ontology (ICDO) to logically and scientifically represent ICD terms and their relations among different ICD terms. Different from existing disease ontologies, all ICD diseases in ICDO are defined as disease processes to describe its occurrence with different properties. The ICDO decomposes each disease term into different components, including anatomic entities, process profiles, etiological causes, and output phenotype, etc. Over 500 ICD terms have been represented in ICDO. Many ICDO terms are presented in both English and Chinese. As a use case, the diagnosis data of over 27,000 COVID-19 patients from 5 countries were extracted from the 4CE organization website, and the disease codes associated with the data were coded using ICD10/ICD9. Approximately 400 COVID-19-related disease codes, each of which were associated with 10 or more cases in the 4CE dataset, were mapped to ICDO and further analyzed using the ICDO logical annotations. Our study showed that COVID-19 targeted multiple systems and organs such as lung, heart, and kidney. Different acute and chronic kidney phenotypes were identified. Some kidney diseases appeared to result from other diseases such as diabetes. Some of the findings could only be easily found using ICDO instead of ICD9/10.

Conclusions: ICDO was developed to ontologize ICD10/10 codes and applied to study COVID-19 patient diagnosis data. Our findings showed that ICDO provides a semantic platform for more accurate detection of disease profiles.

Keywords: -

BMC-26

A multi-task CNN learning model for taxonomic assignment of human viruses

Haoran Ma¹, Tin Wee Tan^{1,2}, Kenneth Hon Kim Ban^{1,2}

¹Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, 117592, Singapore

²National Supercomputing Centre (NSCC), 138632, Singapore

Corresponding author: Kenneth Hon Kim Ban

Abstract

Taxonomic assignment is a key step in the identification of human viral pathogens. Current tools for taxonomic assignment from sequencing reads based on alignment or alignment-free k-mer approaches may not perform optimally in cases where the sequences diverge significantly from the reference sequences. Furthermore, many tools may not incorporate the genomic coverage of assigned reads as part of overall likelihood of a correct taxonomic assignment for a sample. In this paper, we describe the development of a pipeline that incorporates a multi-task learning model based on convolutional neural network (MT-CNN) and a Bayesian ranking approach to identify and rank the most likely human virus from sequence reads. For taxonomic assignment of reads, the MT-CNN model outperformed Kraken 2, Centrifuge, and Bowtie 2 on reads generated from simulated divergent HIV-1 genomes and was more sensitive in identifying SARS as the closest relation in four RNA sequencing datasets for SARSCoV-2 virus. For genomic region assignment of assigned reads, the MT-CNN model performed competitively compared with Bowtie 2 and the region assignments were used for estimation of genomic coverage that was incorporated into a naïve Bayesian network together with the proportion of taxonomic assignments to rank the likelihood of candidate human viruses from sequence data of candidate human viruses from sequence data.

Keywords: -

BMC-27

Analysis of StAR-related lipid transfer (START) domains across the rice pangenome reveals how ontogeny recapitulated selection pressures during rice domestication

Sanjeet Kumar Mahtha^{1,†}, Ravi Kiran Purama^{1,†}, Gitanjali Yadav^{1,2}

¹Computational Biology Laboratory, National Institute of Plant Genome Research, New Delhi, India

²Department of Plant Sciences, University of Cambridge, Downing Street, Cambridge CB23EA, United Kingdom

[†]These authors contributed equally to this work

Corresponding author: Gitanjali Yadav

Abstract

Background: START proteins, encoded by a plant amplified family of evolutionary conserved genes, play important roles in lipid binding, transport, signaling, and modulation of transcriptional activity in the plant kingdom, but there is limited information on their evolution, duplication and associated sub- or neo-functionalization.

Results: Here we perform a comprehensive investigation of this family across the rice pangenome, using ten wild and cultivated varieties. Conservation of START domains across all ten rice genomes suggests low dispensability and critical functional roles for this family, further supported by chromosomal mapping, duplication and domain structure patterns. Analysis of synteny highlights a preponderance of segmental and dispersed duplication among STARTs, while transcriptomic investigation of the main cultivated variety *Oryza sativa* var. japonica reveals sub-functionalization amongst genes family members in terms of preferential expression across various developmental stages and anatomical parts, such as flowering. Ka/Ks ratios confirmed strong negative/purifying selection on START family evolution, implying that ontogeny recapitulated selection pressures during rice domestication.

Conclusions: Our findings provide evidence for high conservation of START genes across rice varieties in numbers, as well as in their stringent regulation of Ka/Ks ratio, and showed strong functional dependency of plants on START proteins for their growth and reproductive development. We believe that our findings advance the limited knowledge about plant START domain diversity and evolution, and pave the way for more detailed assessment of individual structural classes of START proteins among plants and their domain specific substrate preferences, to complement existing studies in animals and yeast.

Keywords: -

MDPI-1

Feature selection for topological proximity prediction of single-cell transcriptomic profiles in *Drosophila* embryo using Genetic Algorithm

Shruti Gupta^{1,2}, Ajay Kumar Verma³, Shandar Ahmad⁴

¹Harvard Medical School, Boston, MA

²Division of Renal Medicine, Brigham and Women's Hospital, Boston, MA

³Department of Pulmonary and Critical Care Medicine, King George's Medical University, Lucknow, Uttar Pradesh, India

⁴School of Computational and Integrative Sciences, Jawaharlal Nehru University, New Delhi 110 067, India

Corresponding author: Shandar Ahmad

Abstract

Single-cell transcriptomics data when combined with in-situ hybridisation patterns of specific genes can help recover the lost spatial information during cell isolation sequencing. Dialogue for Reverse Engineering Assessments and Methods (DREAM) consortium conducted a crowd-sourced competition known as DREAM Single Cell Transcriptomics Challenge (SCTC) to identify top 60,40 and 20 genes which contain the most spatial information out of 84 in-situ gene patterns known in *Drosophila* embryo. We applied a Genetic Algorithm (GA) to predict the most important genes that carry positional and proximity information of the single cell origins. Resulting gene selection was found to perform well and was ranked among top 10 in two of the three subchallenges. However, the details of the method did not make it to the main challenge publication due to an intricate aggregation ranking. In this work, we discuss the detailed implementation of GA and potential areas where GA-based approaches of gene set selection for topological association prediction may be improved to be more effective. We believe this work provides additional insights into the feature selection strategies and their relevance to single cell similarity prediction and will form a strong addendum to the recently published work from the consortium.

Keywords: -

MDPI-2

PupStruct: Prediction of pupylated lysine residues using structural properties of amino acids

Vineet Singh¹, Alok Sharma², Abdollah Dehzangi³, Tatushiko Tsunoda⁴

¹Faculty of Biotechnology, School of Life Sciences, SARI, Jeju National University, Jeju, 63243, Republic of Korea

²MCPHS University, Worcester/Manchester, NH, USA

³Department of Computer Science, Morgan State University, Baltimore, MD, 21251, USA

⁴Cancer Genomics Project, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan

Corresponding authors: Vineet Singh and Alok Sharma

Abstract

Post-translational modification (PTM) is a critical biological reaction which adds to the diversification of proteome. With numerous known modifications being studied, pupylation has gain focus in the scientific community due to its significant role in regulating biological processes. The traditional experimental practice to detect pupylation sites proved to be expensive and requires a lot of time and resources. Thus, there have been many computational predictors developed to challenge this issue, however, performance is still limited. In this study, we propose another computational method named, PupStruct, which uses the structural information of amino acids with radial basis kernel function SVM to predict pupylated lysine residues. We compared PupStruct with three state-of-the-art predictors from the literature where PupStruct has validated significant improvement in performance over them with statistical metrics such as sensitivity (0.9234), specificity (0.9359), accuracy (0.9296), Precision (0.9349) and Mathew's correlation coefficient (0.8616) on a benchmark dataset.

Keywords: -

MDPI-3

RAM-PGK: Prediction of lysine phosphoglycerylation based on residue adjacency matrix

Abel Avitesh Chandra^{1,†}, Alok Sharma^{1,2,3,†}, Abdollah Dehzangi^{4,5}, Tatushiko Tsunoda^{2,6,7}

¹School of Engineering & Physics, University of the South Pacific, Fiji

²Laboratory for Medical Science Mathematics, RIKEN Center for Integrative Medical Sciences, Yokohama 230-0045, Japan

³Institute for Integrated and Intelligent Systems, Griffith University, Brisbane, QLD 4111, Australia

⁴Department of Computer Science, Rutgers University, Camden, NJ 08102, USA

⁵Center for Computational and Integrative Biology, Rutgers University, Camden, NJ 08102, USA

⁶Laboratory for Medical Science Mathematics, Department of Biological Sciences, Graduate School of Science, The University of Tokyo, Tokyo 113-0033, Japan

⁷Department of Medical Science Mathematics, Medical Research Institute, Tokyo Medical and Dental University, Tokyo 113-8510, Japan

†These authors contributed equally to this work

Corresponding authors: Abel Avitesh Chandra and Alok Sharma

Abstract

Background: Post-translational modification (PTM) is a biological process that is associated with the modification of proteome, which results in alteration of normal cell biology and pathogenesis. There have been numerous PTM reports in recent years, out of which, lysine phosphoglycerylation has come out as one of the recent developments. The traditional method of identifying phosphoglycerylated residues, which are experimental procedures such as mass spectrometry, have shown to be time-consuming and cost-inefficient, despite the abundance of proteins being sequenced in this post-genomic era. Due to these drawbacks, computational techniques are being sought to establish an effective identification system of phosphoglycerylated lysine residues. Development of a predictor for phosphoglycerylation prediction is not a first, but it is necessary as the latest predictor falls short in adequately detecting phosphoglycerylated and non-phosphoglycerylated lysine residues.

Results: In this work, we are introducing a new predictor named RAM-PGK, which uses sequence based information relating to amino acid residues to predict phosphoglycerylated and non-phosphoglycerylated sites. A Benchmark dataset is employed for this purpose that contains experimentally identified phosphoglycerylated and non-phosphoglycerylated lysine residues. From the dataset, we have extracted the residue adjacency matrix pertaining to each lysine residues in the protein sequences and converted them into feature vectors, which is used to build the phosphoglycerylation predictor.

Conclusion: RAM-PGK, which is based on sequential feature and support vector machine classifier, has shown a noteworthy improvement in terms of performance in comparison to some of the recent prediction methods. The performance metrics of RAM-PGK predictor are: 0.5741 sensitivity, 0.6436 specificity, 0.0531 precision, 0.6414 accuracy, and 0.0824 Mathews correlation coefficient. The data and software package of this work can be found at <https://github.com/abelavit/RAM-PGK> or www.alok-ai-lab.com.

Keywords: -

JCBC-1

Shared ancestry of core-histone subunits and non-histone plant proteins containing the histone fold motif (hfm)

Amish Kumar¹, Gitanjali Yadav²

¹National Institute of Technology Durgapur, Durgapur, India

²Computational Biology Laboratory, National Institute of Plant Genome Research, New Delhi, India

Corresponding author: Gitanjali Yadav

Abstract

The three helical Histone Fold Motif (HFM) of core histone proteins provides an evolutionarily favoured site for the protein-DNA interface. Despite significant variation in sequence, the HFM retains a distinctive structural fold that has diversified into several non-histone protein families. In this work, we explore the ancestry of non-histone HFM containing families in the plant kingdom. A sequence search algorithm was developed using iterative profile Hidden Markov Models to identify remote homologs of core-histone proteins. The resulting hits were functionally annotated, classified into families, and subjected to comprehensive phylogenetic analyses via Maximum likelihood and Bayesian methods. We have identified 4390 HFM containing proteins in the plant kingdom that are not histones, mostly existing as diverse transcription factor families, distributed widely within and across taxonomic groups. Patterns of homology suggest that core histone subunit H2A has evolved into newer families like NF-YC and DRAP1, whereas the H2B subunit of core histones shares a common ancestry with NF-YB and DR1 class of TFs. Core histone subunits H3 and H4 were found to have evolved into DPE and TAF proteins, respectively. Taken together these results provide insights into diversification events during the evolution of the histone fold motif, including sub-functionalization and neo-functionalization of the HF.

Keywords: -

Frontiers-1

Unsupervised tensor decomposition-based method to extract candidate transcription factors as histone modification bookmarks in post-mitotic transcriptional reactivation

Y-H. Taguchi¹, Turki Turki²

¹Department of Physics, Chuo University, Tokyo, Japan

²Department of Computer Science, King Abdulaziz University, Jeddah, Saudi Arabia

Corresponding Author: Y-H. Taguchi

Abstract

The histone group added to a gene sequence must be released during mitosis to halt transcription during the DNA replication stage of the cell cycle. However, the detailed mechanism of this transcription regulation remains unclear. In particular, it is not realistic to reconstruct all appropriate histone modifications throughout the genome from scratch after mitosis. Thus, it is reasonable to assume that there might be a type of “bookmark” that retains the positions of histone modifications, which can be readily restored after mitosis. We developed a novel computational approach comprising tensor decomposition (TD)-based unsupervised feature extraction (FE) to identify transcription factors (TFs) that bind to genes associated with reactivated histone modifications as candidate histone bookmarks. To the best of our knowledge, this is the first application of TD-based unsupervised FE to the cell division context and phases pertaining to the cell cycle in general. The candidate TFs identified with this approach were functionally related to cell division, suggesting the suitability of this method and the potential of the identified TFs as bookmarks for histone modification during mitosis.

Keywords: advanced unsupervised learning, tensor decomposition, histone modification, bookmark, mitosis, transcription

Demos

Demo-1

NetVA: An R package for network vulnerability analysis

Swapnil Kumar¹, Vaibhav Vindal¹

¹Department of Biotechnology & Bioinformatics, University of Hyderabad, Gachibowli, Hyderabad, 500046, India.

Corresponding author: Vaibhav Vindal

Abstract

In biological network analysis, the identification of key molecules plays a decisive role in the development of potential therapeutic targets. Among various approaches of network analysis, network vulnerability analysis is quite important. As it comprises an assessment of significant associations between functional essentiality and topological properties of the network. Keeping this in mind, we aimed to develop an R package named NetVA, which performs the vulnerability analysis of networks. To demonstrate the application and relevance of our package in network analysis, previously published protein-protein interaction networks of humans were analyzed. This resulted in the identification of some most vulnerable proteins, which were essential or hub proteins as well as neither essential nor hubs but was consistent with previously reported experimental evidence. Thus, the package assists in the prediction of putative drug targets by exploring network topological features through vulnerability analysis at the systems level.

Demo-2

Understanding polygenic disease with BitEpi and EpiExplorer

Arash Bayat¹, Brendan Hosking¹, Yatish Jain¹, Cameron Hosking¹, Milindi Kodikara¹, Daniel Reti¹, Natalie Twine¹, Denis Bauer¹

¹CSIRO Australia

Corresponding author: Natalie Twine

Abstract

Polygenic diseases are driven by a large number of Single Nucleotide Variations (SNVs) and many of these interact in complex ways. Identifying these interactions is difficult due to computational complexity, especially in the case of higher-order interactions where more than two SNVs are involved. Here we introduce BitEpi, a fast and accurate method to test all SNVs and combinations of up to four SNVs. BitEpi introduces a novel bitwise algorithm that is 2.1 and 56 times faster than a 3-SNV search with MPI3SNP and 4-SNV search with MDR respectively. Prior to the development of BitEpi, MPI3SNP was the fastest exhaustive 3-SNV search tool and MDR was the only software to perform an exhaustive 4-SNV search. BitEpi uses a novel test to identify statistically relevant SNVs and interactions. Our method is 44% more accurate than BOOST and MPI3SNP when identifying interactive SNVs. BitEpi is compatible with standard genomic format and offers p-value-based significance testing. To aid in the visualisation of statistically significant SNVs from BitEpi, our novel tool, EpiExplorer, utilizes an interactive Cytoscape graph. EpiExplorer uses various visual elements to facilitate the discovery of the underlying biology in a complex polygenic environment. For example, it is possible to layout the graph to separate genomic regions with different functionality or highlight part of the graph based on a query.

Demo-3

BacEffluxPred: A two-tier system to predict and categorize bacterial efflux mediated antibiotic resistance proteins

Deeksha Pandey¹, Bandana Kumari¹, Neelja Singhal¹, Manish Kumar¹

¹Department of Biophysics, University of Delhi South Campus, New Delhi, India

Corresponding author: Manish Kumar

Abstract

Efflux proteins are transport proteins, which normally transport different substrates from the bacterial cell to the external environment. In any bacterial species, efflux pump proteins constitute between 6–18% of the total transporters. The efflux mechanism and efflux pumps are a major reason underlying emerging rampant antibiotic resistance (AR) in microbes. To reduce the resources required and time of identification, characterization and classification of bacterial efflux proteins, we have developed a fast and accurate machine learning based two-tier prediction system, BacEffluxPred (Bacterial Efflux Prediction). Our method is based on support vector machine (SVM) algorithm and it can predict bacterial efflux proteins responsible for AR (at tier-I) and identify their corresponding families (at tier-II). We used leave-one-out cross-validation procedure to train and evaluate the performance of SVM models. The accuracy to discriminate bacterial AR efflux from non-AR efflux was 85.81% (at tier-I) while accuracies for prediction of efflux pump families like ABC, MFS, RND and MATE family were found 92.13%, 85.39%, 91.01% and 99.44%, respectively (tier-II). Benchmarking of BacEffluxPred on an independent dataset composed of efflux proteins, non-efflux and antibiotic resistance proteins also showed comparable accuracy for prediction of bacterial AR efflux pumps and their families. This is the first *in-silico* tool for predicting bacterial AR efflux proteins and their families and is freely available as both web-server and standalone versions. We believe that the described tool would help the scientific community in their quest to combat antibiotic resistance in clinical and environmental settings.

Demo-4

Automated identification of SNP-genotypes in genomic datasets: SNIpSoL - An application to *Mycobacterium leprae*

Fenil Ganatra¹, Deven Matkari¹, Purna Dwivedi^{1,2}, Mukul Sharma², Pushendra Singh^{1,2}

¹The Maharaja Sayajirao University of Baroda, Vadodara, Gujarat, India

²ICMR-National Institute of Research in Tribal Health, Jabalpur, MP, India

Corresponding author: Pushendra Singh

Abstract

With the advent of sequencing technologies, pathogens have been found to have many strains based on their phylogeography and drug resistance. Genomic analyses have allowed identification of SNPs/Insertions/Deletions for genotyping and enabled the study of transmission dynamics and emergence of drug resistance in these pathogens. Existing typing schemes involve identification from large datasets of thousands of SNPs, making this a labour-intensive and error-prone exercise. A program able to compare a given set of SNPs with inbuilt genomic reference datasets can make this process faster and more accurate. Therefore, we have developed an automated program (using custom script and BlueJ) called SNIpSoL (SNP-based Strain-typing of Leprosy) to identify the genotype of *M. leprae* strains. The emergence of drug resistant strains of *Mycobacterium leprae* (the causative agent of leprosy) has been reported in recent years. Next generation sequencing based whole genome analysis has also revealed phylogeographic markers. Based on user inputs, the program identifies SNPs at key genomic loci and assigns a genotype from a comprehensive dataset of SNPs of all strains. The program can also identify the predominant genotypes by input of just 1-2 SNP loci and hence can be customised based on the geographic location and strain prevalence. The utility of this program lies in its ability to accurately, rapidly and reproducibly identify strains and their molecular pattern of drug resistance, without the user having to go through research articles and databases for reliable strain typing schemes and could potentially make the process of choosing appropriate drugs for treatment easier.

Lightning Talks

LT-1

Variants profiling of BRCA 1/2 genes through next-generation sequencing in young women with breast cancer

Sonar Soni Panigoro¹, Rafika Indah Paramita^{2,3}, Kristina Maria Siswiandari¹, Fadilah Fadilah^{2,3}, Linda Erlina^{2,3}

¹Surgical Oncology Division, Department of Surgery, Faculty of Medicine, Universitas Indonesia

²Department of Medical Chemistry, Faculty of Medicine, Universitas Indonesia

³Bioinformatics Core Facilities - IMERI, Faculty of Medicine, Universitas Indonesia

Corresponding author: Rafika Indah Paramita

Abstract

Mutations in either BRCA1 or BRCA2 genes play an important roles in early-onset breast cancer. Testing for BRCA1/2 mutations in breast cancer patients is important because knowledge of these mutations helps clinicians to make a decision about options of risk-reduction strategies. The purpose of this study was to describe the profile of BRCA1/2 variants in young women with breast cancer. Young women patients (< 45 years old) with breast cancer (BC) familial history (n=18) and sporadic breast cancer (n=57) were included as subjects. Patients' blood samples were taken and their DNA sequenced using Illumina NextSeq 500 platform. Various bioinformatic tools for variant calling were implemented with the hg38 sequence as human genome reference. The VarSome software was used for in silico analysis of novel variants. In the BRCA1/2 gene variants, silent mutation were dominant among other variants, followed by missense variants, intron variants, and frameshift variants. Patients with BC familial history had more benign mutation in BRCA2 gene (48%), whereas patients with sporadic BC had more benign mutation in BRCA1 gene (80%). Pathogenic variants of BRCA 1/2 genes were found in BC Familial history patients (22%) which one of them was novel variant. Pathogenic variants were also found in patients with sporadic BC (8.8%) which three of them were novel variants. BRCA1/2 variants profiling can be obtain using of the new generation technologies which of these variants can help to make a decision about various risk-reduction strategy options.

LT-2

Galaxy Australia – A key partner in the global rapid response to the COVID-19 pandemic

Simon Gladman¹, Gareth Price², Andrew Lonie³

¹Melbourne Bioinformatics, Faculty of Medicine, University of Melbourne, Australia

²QCIF Facility for Advanced Bioinformatics, Institute of Molecular Biology, University of Queensland, Australia

³Australian BioCommons, University of Melbourne, Australia

Corresponding author: Gareth Price

Abstract

Galaxy Australia, as one member of the usegalaxy.* ecosystem was pivotal in a global call to enable clear and reproducible data sharing and analytics of the rapidly growing collection of sequencing data pertaining to severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Working jointly with the Australian national supercomputer centres, Pawsey and NCI, Galaxy Australia deployed a series of dedicated computer nodes for the analysis of SARS-CoV-2 DNA sequence. This complemented parallel activities at Galaxy Main (usegalaxy.org) and Galaxy Europe (usegalaxy.eu) with all three services enabling analysis of SARS-CoV-2. Given the rapid responses to the COVID-19 pandemic this presentation will highlight the work on the publication but also showcase the next evolution of the global analysis infrastructure; covid19.galaxyproject.org. Here Galaxy Main, Europe, Belgium, France and Australia support analysis of SARS-CoV-2 through genomic, evolution, cheminformatics, proteomics, direct RNA-Seq and Artic (amplicon based) analyses.

LT-3

APBioNetTalks: A platform to share bioinformatics talks, tutorials and trainings

Hilyatuz Zahroh¹, Mohammad Asif Khan^{2,3}

¹Genetics Research Centre, Universitas YARSI, Jakarta, Indonesia

²School of Data Science, Perdana University, Kuala Lumpur, Malaysia

³Beykoz Institute of Life Sciences and Biotechnology, Bezmîâlem Vakıf Üniversitesi, Turkey

Corresponding author: Mohammad Asif Khan

Abstract

APBioNetTalks is an education programme to be launched by Asia Pacific Bioinformatics Network (APBioNet) by the mid of October 2020. It will provide open access, video-based bioinformatics learning resources for the public. The platform will host and live stream bioinformatics related talks, which could comprise research presentations, tutorials, hands-on practicals, demos and trainings, among others. The programme aims to facilitate the discovery and sharing of bioinformatics. Experts, early career researchers/scientists, and students of the field, from different countries and institutions will be invited to share their knowledge and skills with the audience. The programme also encourages contribution from communities and institutional partners. Pre-recorded or live videos will be made available on the APBioNet YouTube channel. More information on the programme is available at <https://apbtalks.apbionet.org>. The programme is also listed on the Open Life Science Project (<https://openlifesci.org/posts/2020/09/01/ols2-announcement/>).

LT-4

Genetic ancestry in the hunt for disease genes and the fight against COVID-19

Natalie Twine¹

¹Transformational Bioinformatics Group, CSIRO, Australia

Abstract

Knowing your family history is important for diagnosing and treating genetic diseases. However, very rarely do people know who their distant 4th cousin is. We have recently developed 'TRIBES', a population-scale cloud-based software that enables the discovery of distant relatives based on their genome. Dr Twine will showcase how TRIBES has been able to identify distant relatives and novel disease genes in Australian patients suffering from the devastating neurodegenerative disease, Amyotrophic Lateral Sclerosis (ALS). Leveraging the scalability of TRIBES, Dr Twine will also present novel findings from the largest international ALS cohort, ProjectMinE. Finally, Dr Twine will explain how we have used genetic relationships with our cloud-native technology to advance the COVID-19 response.

Workshops

WS-1

Machine learning approaches for ascertaining transcriptomics data using T-bioinfo & code playground

Mohit Mazumder¹, Harpreet Kaur¹

¹Pine Biotech, Inc., 1441 Canal St. Suite 411, New Orleans, LA 70112, USA

Corresponding author: Mohit Mazumder

Abstract

With growing data availability, its heterogeneity, and complexity, it is not sufficient to have access to data. To go from patterns to insights, modern-day researchers in academia and industry need to be empowered to work with data independently. This innovative workshop will address the major gaps in the processing and analysis of high-throughput biomedical data and finding meaningful information from it. This will be achieved by participants applying algorithms to quantify gene expression and detect statistically significant variation from RNA-seq projects in oncology.

WS-2

Galaxy – A platform for life science analyses (more than just genomics)

Gareth Price¹, Simon Gladman²

¹QCIF Facility for Advanced Bioinformatics, Institute of Molecular Biology, University of Queensland, Australia

²Melbourne Bioinformatics, Faculty of Medicine, University of Melbourne, Australia

Corresponding author: Gareth Price

Abstract

Australia hosts a nationally accessible Galaxy instance; the Galaxy Australia platform. The platform is in its third year of national operation and is strongly aligned to the usegalaxy.* ecosystem of global Galaxy platforms. This demonstration will showcase the generic features of the Galaxy platform that enables it to be a robust nationally hosted web-accessible platform that lets you conduct accessible, reproducible, and transparent computational biological research. Tri-yearly Galaxy code base is updated and we will highlight features of the platform that demonstrate the clear commitment of the global Galaxy developer community to delivering functionality relevant to the user base. These will include streamlined data ingest functionality (both files and metadata), interactive tools for rich data exploration, improved workflow management. In addition, we will demonstrate how the underlying architecture of Galaxy services can be deployed to enable increasing complex data analysis, such as plant derived genome assemblies, de novo genome reconstruction and large cohort metagenomics, through the use of high memory computer nodes. Finally, the demonstration will highlight the broadening fields of research support enabled through the Galaxy beyond genomics, including proteomics, metabolomics, ecology and climate modelling.

WS-3

Visualisation and analysis of complex networks in biology

Gitanjali Yadav^{1,2}

¹Complex Networks and Machine Learning Laboratory, National Institute of Plant Genome Research (NIPGR), New Delhi, India.

²Department of Plant Sciences, University of Cambridge, U.K

Corresponding author: Gitanjali Yadav

Abstract

This workshop tutorial is an Introduction to Biological Networks, their types, and applications. It will include two of the most commonly used open source Network Visualisation Platforms (R-igraph and Cytoscape) with step-wise protocols for creating and visualising your own data as a network. It will present some of the major layout algorithms, visual styles and tips for effective visualisation, with examples from Genome Biology revealing how these can improve analysis and provide insights. The topics covered will be Biological networks, R-igraph, Cytoscape, Data visualisation and clustering. After this workshop, you should be able to (a) conceptualise your own data as a network, (b) create a simple network using Cytoscape or R platform (c) Add node/edge attributes to your Network (d) Perform MCODE Clustering on your Network in Cytoscape. During this course you will also learn about how networks can improve data analysis and provide useful insights, some of the major network layout algorithms & visual styles, as well as tips for effective visualisation.

Poster Presentations

P1

Misannotation of coproporphyrinogen III oxidases HemN in the mycobacterial genome

Mukul Sharma¹, Yash Gupta², Purna Dwivedi¹, Prakasha Kempaiah², Pushpendra Singh¹

¹Indian Council of Medical Research-National Institute of Research in Tribal Health, India

²Loyola University Medical Center, Maywood, Illinois, USA

Corresponding author: Pushpendra Singh

Abstract

Mycobacterium lepromatosis, a newly identified causative agent of leprosy, was sequenced in 2015(1), wherein a gene MLPM_5000 was detected whose corresponding sequences are missing in *Mycobacterium leprae*, the well-known causal agent of leprosy. Thus MLPM_5000 is a specific genomic template for *M. lepromatosis* (annotated as hemN, coproporphyrinogen III oxidase based on available annotations in other mycobacterial species). The goal of this study is to explain the structural features of MLPM_5000 and its mycobacterial orthologues (currently annotated as HemN) by using sequence alignment, modelling, MD simulations and phylogenetic analysis. We observed that the amino acid sequences of mycobacterial HemN sequences in different species have much higher degree of similarity with *E. coli* HemW (>30 %) in comparison to the *E. coli* HemN (<25 %). Additionally, the fourth cysteine of the mycobacterial HemN CX3CX2CXC motif is replaced by a phenylalanine, which indeed is a hallmark of *E. coli* HemW. Phylogenetic analysis showed that mycobacterial HemN forms a divergent phylogenetic clade with the HemW proteins in other species such as *E. coli* (2) and *L. lactis* (3). Homology modelling and MD simulation studies showed that the residues of conserved HemW HNXXYW motif, present in *M. lepromatosis* MLPM_5000 may have a role in the binding of heme, suggesting it to be HemW instead of HemN. Together, we found that MLPM_5000 and its mycobacterial orthologs are unlikely to exhibit coproporphyrinogen III dehydrogenase (CPDH) activity. Therefore, the presently mis-annotated mycobacterial HemN sequences need to be corrected as heme chaperone HemW in various mycobacterial databases.

P2

Genomic surveillance reveals SARS-CoV-2 lineage B.6 is the major contributor to transmission in Malaysia

Yoong Min Chong¹, I-Ching Sam^{1,2}, Jennifer Chong², Maria Kahar Bador^{1,2}, Sasheela Ponnampalavanar³, Sharifah Faridah Syed Omar³, Adeeba Kamarulzaman³, Vijayan Munusamy³, Chee Kuan Wong³, Fadhil Hadi Jamaluddin⁴, Yoke Fun Chan¹

¹Department of Medical Microbiology, Faculty of Medicine, University of Malaya, Malaysia

²Department of Medical Microbiology, University of Malaya Medical Centre, Malaysia

³Department of Medicine, Faculty of Medicine, University of Malaya, Malaysia

⁴Department of Anesthesiology, Faculty of Medicine, University of Malaya, Malaysia

Corresponding author: Yoke Fun Chan

Abstract

Coronavirus disease (COVID-19) caused by SARS-CoV-2 has spread globally. In Malaysia, the main wave of infection was associated with a religious (tabligh) mass gathering held in Kuala Lumpur at the end of February. We generated whole genome sequences from 62 COVID-19 patients. We performed bioinformatics analysis (SNV detection and phylogenetic analysis) on these and a further 46 available Malaysian sequences in the GISAID database. Among the 108 Malaysian sequences, nine different SARS-CoV-2 lineages (A, B, B.1, B.1.1, B.1.1.1, B.1.36, B.2, B.3 and B.6) were observed. The lineage B.6, which is relatively rare (1.4%; n=970/67,557) among globally reported strains, was the most predominant (62.9%; n=68/108) in Malaysia. This lineage was first reported a week after the mass gathering and temporally linked to tabligh-associated cases. Phylogenetic analysis also revealed that B.6 lineage spread to Southeast Asian countries, India and Australia. Altogether, 95.3% (n=924/970) of global B.6 sequences originated in Asia or Australia. We also reported the presence of a nsp3-C6310A substitution found in 40.5% (n=393/970) of global B.6 sequences which is associated with reduced sensitivity in a commercial qPCR assay. About 70% (n=28/40) of non-B.6 lineages were mainly associated with travel and showed limited onward transmission in Malaysia. In conclusion, lineage B.6 became the predominant in Malaysia after likely introduction during a religious mass gathering, and subsequent spread of B.6 viruses regionally. The use of genomic data is important to rapidly identify possible transmission chains and provide a framework for the response to COVID-19.

P3

scMontage: Fast and robust gene expression similarity search for massive single-cell data

Naila Shinwari¹, Tomoya Mori², Wataru Fujibuchi¹

¹Center for iPS Cell Research and Application (CiRA), Kyoto University, Japan

²Bioinformatics Center, Institute for Chemical Research, Kyoto University, Japan

Corresponding author: Wataru Fujibuchi

Abstract

Single-cell RNA-seq (scRNA-seq) analysis is widely used to characterize cell types or to detect heterogeneity of cell states at much higher resolutions than ever before. Here, we introduce scMontage (<https://scmontage.stemcellinformatics.org/>), a renovated system for searching gene expression databases for cells similar to the query gene expression profile. The scMontage search is based on Spearman's rank correlation coefficient and its robustness is ensured by introducing Fisher's Z-transformation and Z-test. Furthermore, search results are linked to a human cell database SHOGoiN (<http://shogoin.stemcellinformatics.org>), which enable users to fast access to additional cell-type specific information. Currently, the scMontage server provides human and mouse scRNA-seq data and allows users to quickly access cell-type-specific biological information, such as cell taxonomy, lineage map, cell marker, and so on. Fast search, Easy to use, no need of normalization of data are the attributes of scMontage that makes it better than the existing gene similarity tools. The scMontage is available not only as a web server but also as a stand-alone application for user's own data, and thus it enhances the reliability and throughput of cell analysis and helps users gain new insights into massive scRNA-seq data.

P4

MPLANABASETM, a repository of *Metisa plana* transcriptome data for gene discovery

Nor-Muhammad, N.A.¹, Rahmat, N.L.¹, Zifruddin, A.N.¹, Zainal-Abidin, C.M.R.², Hassan, M.¹

¹Institute of Systems Biology, Universiti Kebangsaan Malaysia (UKM), 43600 UKM, Bangi, Selangor, Malaysia

²Felda Global Venture R&D Sdn. Bhd., 26400, PPP Tun Razak, Jengka, Pahang, Malaysia

Corresponding author: Hassan, M.

Abstract

Metisa plana (Lepidoptera: Psychidae) is one of the major leaf defoliators of oil palm trees. After two years of infestation, oil palm productivity could be reduced by 43% due to damaged leaves. However, little is known on the molecular and physiological mechanisms of the pest. Here we present MPLANABASETM, a repository of *M. plana* transcriptome data that can be used for gene discovery and pathway validation works. The *M. plana* transcriptome was from four developmental stages; eggs, third instar larvae, pupae and female adults. Illumina Hi-Seq sequencing, 219,702,646 raw reads were processed, and de novo assembled into 352,477 unigenes using Trinity 2.8.5 assembler. The unigenes were functionally annotated against UniProt, Pfam, eggNOG, and KEGG databases as well as SignalP and TMHMM tools. 11,046 unigenes were functionally annotated. The database is built using the Laravel PHP framework with MySQL as the backend. A Digital Ocean droplet is used to host the Apache server at <http://mplanabase.org>. The database allows for other researchers to work on *M. plana* unigenes and hypothetical proteins sequences to understand *M. plana* molecular workings and a valuable genomic resource for further functional genomics studies on *M. plana*.

P5

Deciphering the molecular interactions of kaempferol with three carrier proteins

Zaved Hazarika¹, Anupam Nath Jha¹

¹Department of Molecular Biology and Biotechnology, Tezpur University, India

Corresponding author: Anupam Nath Jha

Abstract

In recent time, efforts were laid to explore solutions for different diseases from plant based natural products namely phytochemicals. Phytochemicals have profound implications in a wide array of health conditions such as neurodegenerative disorders, cardiovascular diseases, inflammatory disorders to name a few. One such class of phytochemicals are flavonoids, which itself are part of polyphenols and historically been utilized in traditional medicines. One of the known flavonoid is Kaempferol (3,5,7-trihydroxy-2-(4-hydroxyphenyl)chromen-4-one) (KMP) which is easily obtained from edibles like tea, broccoli, cabbage, beans, strawberries and grapes. It has been reported to be a potential anti-inflammatory, anti-proliferative and anti-oxidant molecule. Therefore, owing to the diverse potential roles of KMP, it is noteworthy to investigate its molecular recognition process by different carrier proteins. In our work we have utilized the computational techniques of molecular docking and molecular dynamics simulations to decipher the binding modes of KMP with three carrier proteins viz. human serum albumin, lysozyme and haemoglobin. The putative binding modes and the level of interactions between KMP and the mentioned proteins were analysed and compared. KMP was able to bind into the plausible binding cleft of the three proteins with a good number of non-bonded interactions. Additionally druglikeness property of the molecule KMP has been assessed wherein it satisfies the Lipinski's rule of 5. The stability of the bound complexes have been evaluated with help of MD simulations. The atomic level insights from our work have further implications to the development of the molecule as a potential compound.

P6

Discovery of potential new inhibitors of *Mycobacterium tuberculosis* CYP121 from drug repositioning database

Tarek El Moudaka¹, Bimo A. Tejo²

¹Department of Biotechnology, Faculty of Applied Sciences, UCSI University, Malaysia

²Department of Chemistry, Faculty of Science, Universiti Putra Malaysia

Corresponding author: Bimo A. Tejo

Abstract

Tuberculosis (TB) has been responsible for over a billion deaths in the past 200 years. Despite high effectiveness of BCG vaccine, tuberculosis continues to be a serious global health threat with the emergence of drug-resistant tuberculosis. The increasing number of deaths associated with *M. tuberculosis* over the last 25 years has highlighted the limitations of currently available anti-TB drugs. Cytochrome P450 CYP121 is considered the most promising anti-TB drug target due to its prominent substrate specificity. Phylogenetic analysis reveals that CYP121 is exclusive to *M. tuberculosis*, and this is presumably to accomplish the unique C–C bond-forming catalysis that is not required by other bacteria. This study aims to find new *M. tuberculosis* CYP121 inhibitors by screening more than 8000 molecules from a drug repositioning database RepoDB. The selection of CYP121 potential inhibitors is based on two criteria in which the new molecule must bind to CYP121 with a stronger affinity than that of its respective active ligand, and it must interact with residues that are catalytically important for the function of CYP121. Antrafenine (DB01419), which is an approved drug is found to strongly bind to CYP121 with a binding affinity of -12.6 kcal/mol. Antrafenine interacts with CYP121 by establishing bonds with the amino acid residues especially Arg386 and HEM402 group in which they both play an important role in CYP121 catalytic activity. Antrafenine is commercially available and can be acquired for further biological testing.

P7

Cis-regulatory regions of pathogenic bacteria are associated with functionally conserved G-quadruplex motifs

Upalabdha Dey¹, Sharmilee Sarkar¹, Venkata Rajesh Yella², Aditya Kumar¹

¹Department of Molecular Biology and Biotechnology, Tezpur University, Tezpur, Assam, India

²Department of Biotechnology, Koneru Lakshmaiah Education Foundation, Guntur, Andhra Pradesh, India

Corresponding author: Aditya Kumar

Abstract

G quadruplex is one of the most well evident non-B-DNA structures of nucleic acid in different domains of life i.e. ranging from prokaryotes to eukaryotes. Experimental studies have identified myriad cellular functions of G-quadruplexes in replication, transcription, translation, recombination in the last three decades explaining the significance of these secondary structures. Few recent in vitro studies used G-quadruplex motifs as potential drug targets in several human diseases by modulating their stability inside cellular milieu using various synthetic ligands. In this manuscript, we report a detailed bioinformatic analysis of the functional conservation of G-quadruplex motifs to understand its possible role in the pathogenic bacterial system. Biologically important positional preferences in the genome, gene orthology, ontology, and relevance in gene regulatory networks of G-quadruplexes were studied extensively. The conspicuous findings of the analysis are (a) Putative G-quadruplex forming sequences (PG4) are abundant in regulatory regions of the gene nullifying the role of GC content bias (b) Significant conservation of G4 motifs in the regulatory region of 512 orthologous genes has been observed across seven different bacterial species. (c) Published ChIP-Seq data of 113 DNA-binding proteins of the *M. tuberculosis* genome revealed regulatory regions with PG4 motifs show some characteristic transcription factor binding activity (d) Genes associated with PG4 motifs are linked to pathogenicity in *Mycobacterium tuberculosis*. In toto our study applies positional-functional relationship computation to delve into the cis-regulation of G-quadruplex structures in the context of gene orthology in pathogenic bacteria.

P8

A study of Fascioliasis from semi wild ruminants from two biological hotspots of India: A molecular approach using ribosomal ITS2 and mitochondrial CO1 genes

Damanbha Lyngdoh¹, Sunil Sharma², Bishnupada Roy³, Veena Tandon⁴

¹Department of Zoology, St. Anthony's College, Shillong 793001, Meghalaya, India

²Biotech Hub, St. Edmund's College, Shillong 793003, Meghalaya, India

³Department of Zoology, North Eastern Hill University, Shillong 793022, Meghalaya, India

⁴Biotech Park, Kursi Road, Lucknow 226021, Uttar Pradesh, India

Corresponding author: Veena Tandon

Abstract

Fascioliasis (a trematodiasis) is one of the many neglected tropical diseases reported from ruminant hosts in Asia, Africa and Europe and is mostly attributed to two species of the liverfluke *Fasciola* - *F. gigantica* and *F. hepatica*. Fascioliasis is poorly studied in India and is predominantly caused by *F. gigantica*. With the discovery of the hybrid/ intermediate forms of the liverfluke from regions in and around Northeast India, it is quite difficult to morphologically ascertain the causative species of fascioliasis. Molecular analyses using both nuclear ITS2 and mtCO1 gene markers in combination with bio-informatic tools such as MEGA, Jmodeltest, Mesquite, MrBayes, ProfDist, *etc.*, helped us in rapid identification and delineating the phenotypically identical species. Additionally, using the 4SALE software, a detailed study of the ITS2 secondary structure including helical length, comparative motif analysis and base changes clearly enhanced the characterization of these flukes and established the existence of two types of liverflukes in the indigenous semi wild ruminants - mithun (*Bos frontalis*) and yak (*Bos grunniens*) that abound the hotspot regions of Northeast India.

P9

Phylogeography and population genetics of the rat tapeworm, *Hymenolepis diminuta*: An inference based on mtCO1 gene

Sunil Sharma¹, Samrat Adhikari¹, Veena Tandon²

¹Advanced Level Biotech Hub, St. Edmund's College, Meghalaya-793003, India

²Biotech Park, Lucknow, India

Corresponding author: Sunil Sharma

Abstract

Hymenolepiasis is a neglected zoonotic disease of widespread occurrence. One of the causative agents, *Hymenolepis diminuta* is known to be a cosmopolitan parasite highly prevalent among rodent populations. In the present study, phylogeography and population genetics of *H. diminuta* was examined using the mitochondrial CO1 gene. For the purpose, all the mtCO1 gene sequences of *H. diminuta* available in database till date were retrieved and analysed. Phylogenetic analyses demarcated the *H. diminuta* population into two divergent clades, based on which the presence of two distinct genotypes *i.e.*, Euro-American and an Asian genotype is speculated. For phylogeographic study, a median joining haplotype network was constructed which revealed a typical star-like pattern. Such pattern indicates the divergence of *H. diminuta* from a common ancestral pool, signature of a rapid demographic expansion. The existence of a distinct Euro-American and an Asian genotype was also supported by the haplotype network. Few Asian isolates were found to be divergently related to the rest of the isolates in both phylogenetic and network analyses, indicative of the fact that these isolates could possibly represent the lineage of cryptic species in *H. diminuta*. With regard to the diversity of *H. diminuta*, the Asian and African populations were found to be relatively diversified as seen from the higher haplotype and nucleotide diversities in these populations. The 'Within the population' differentiation factor was attributed for most of the diversity in the population. Pairwise F_{st} was apparently higher between most of the populations except those between Europe and America, indicative of significant gene flow occurring between the two populations.

P10

Molecular docking study on the anti-staphylococcal activity of *Rumex nepalensis* (Spreng.)

Ninni Sutradhar¹, Sunil Sharma¹, Yogesh Negi¹, Sylvanus Lamare², Samrat Adhikari¹

¹Advanced Level Biotech Hub Facility, Department of Biotechnology, St. Edmund's College, Shillong-793003, Meghalaya, India

²Principal, Edmund's College, Shillong-793003, Meghalaya, India

Corresponding author: Ninni Sutradhar

Abstract

Rumex nepalensis is a commonly occurring medicinal plant routinely used in traditional healing system. To detect the therapeutic property of the plant, firstly, the crude extract obtained from the leaves of *R. nepalensis* was tested for antimicrobial activity against *Staphylococcus aureus* using the well diffusion assay and further processed for phytochemical screening. The crude powder was then subjected to LC-HRMS analysis for detection and identification of the compounds present. Following this, the compounds identified were evaluated for antimicrobial activity using *in-silico* analyses. The results of the well diffusion assay elucidated the inhibiting effect of the leaf extract on the growth of *S. aureus*. A number of compounds were identified based on LC-HRMS data, out of which, seven potential bioactive compounds namely, which passed through the test for drug likeness and were thus selected for molecular docking studies. A search through literature revealed the occurrences of certain putative drug targets in *S. aureus*; however, four common drug targets were selected for the present study. The results from molecular docking showed that chrysophanol interacted with all the four drug targets in *S. aureus*. Hastatusides A, schisandrin and pinoresinol also showed considerable interaction with the drug targets. Each of these interactions was well supported with low binding energy and inhibition constant values. Thus, the result of the present study revealed chrysophanol as the putative compound conferring antimicrobial property to the leaf of the plant *R. nepalensis* can be of immense significance in drug design and pharmacology.

P11

Structural, functional and molecular dynamics analysis of CASR gene SNVs associated with tropical calcific pancreatitis

Ashish Shrivastava¹, Garima Singh¹, Rohit Verma¹, Madhusudhan Chinthakindi², Sri Krishna Jayadev Magani¹, Ashutosh Singh¹

¹Translational Bioinformatics and Computational Genomics Research Lab, Department of Life Sciences, Shiv Nadar University, G.B. Nagar, Uttar Pradesh 201314, India.

²Department of Surgical Gastroenterology, Osmania General Hospital, Hyderabad, India

Corresponding author: Ashutosh Singh

Abstract

Tropical Calcific Pancreatitis (TCP) is a juvenile form of chronic calcific non-alcoholic pancreatitis, seen almost exclusively in the emergent countries of the tropical world. This disease's further progression can lead to pancreatic diabetes, called fibro-calcious pancreatic diabetes (FCPD), followed by pancreatic cancer. CASR gene encodes for a protein called a Calcium sensing receptor (CaSR), expressed in the kidney's parathyroid gland and cells of the kidney. It detects small changes in circulating calcium concentration and couples these details to intracellular signaling, which leads to the regulation of PTH secretion and mineral ion homeostasis. It plays a vital role in protease activation in the pancreas. SNVs related to the CASR gene were reported to increase the risk of chronic pancreatitis (CP) since high intracellular calcium levels activate trypsinogen within the acinar cells. The Association of SPINK1 mutations can increase the risk of TCP. Therefore the CASR gene, identified as a putative candidate gene in TCP. There are four novel missense mutations (i.e., p.P163R, p.I427S, p.D433H, and p.V477A) in the extracellular domain (ECD) of CaSR protein associated with TCP reported in the mutTCPdb database. The functional importance of missense mutations was analyzed by in-silico prediction algorithms such as SIFT, PolyPhen, and CADD, along with Molecular Dynamics simulation analysis to understand the atomic details of the CaSR ECD structure. Results from this study may suggest the impact of these novel missense variants on the structure of CaSR ECD, and consequence on the function of this protein and its disease association.

P12

Interaction mechanism of Withanone and Withaferin-A from *Withania somnifera* with lysozyme, serum albumin and haemoglobin – A molecular dynamics approach

Sanchaita Rajkhowa¹

¹Centre for Biotechnology and Bioinformatics, Dibrugarh University, Dibrugarh-786004, Assam, India

Corresponding author: Sanchaita Rajkhowa

Abstract

In December 2019, novel coronavirus SARS-CoV-2 was reported from the Wuhan city of China. It is a kind of pneumonia having an unknown aetiology, has become a pandemic with high fatality rate and caused a halt in everyone's normal life. Although the characterization of the complete sequence was completed in January 2020, there is no definitive cure or vaccine available for this virus. While a number of new lines of drug and vaccine development has been initiated world-wide, but considering the present scenario of high infection rate, the disease severity and high morbidity and mortality, repurposing of the existing drugs is heavily explored. Indian Ayurvedic herb, Ashwagandha (*Withania somnifera*) has been used in traditional medicines. It is known to boost the immune function, possess a variety of prophylactic and therapeutic activities. SARS-CoV-2 cell entry depends on ACE2 and trans-membrane protease serine 2 (TMPRSS2) receptor. Recently, withaferin-A and withanone, two bioactive compounds derived from Ashwagandha, are reported to bind with the catalytic site of TMPRSS2 and ACE2 receptor and they significantly decrease the electrostatic component of binding free energies of ACE2-RBD complex. Here, I attempted to understand the molecular recognition process of withaferin-A and withanone with three carrier proteins namely, lysozyme, human serum albumin and haemoglobin using molecular docking and dynamics simulations. These studies revealed that withanone has better interaction as compared to withaferin-A with the carrier proteins. Thus, this study has provided a substantial insight into the binding of the bioactive compounds withanone and withaferin-A with the carrier proteins.

P13

Identification of oil palm simple sequence repeat markers associated with basal stem rot disease

Mohd Amin Ab Halim¹, Rozana Rosli¹, Leslie Low Eng Ti¹

¹Malaysian Palm Oil Board, Malaysia

Corresponding author: Leslie Low Eng Ti

Abstract

Basal stem rot (BSR) disease is the most devastating disease in oil palm plantations in South East Asia. This study aims to understand the genetic diversity of oil palm populations with different tolerance and susceptibility levels to BSR disease. The study was initiated with the identification of 131 putative oil palm resistance gene analogues (RGA) in the oil palm hypomethylated GeneThresher® libraries using the protein domain composition of known resistance genes in other plants. Resistance genes play important roles in plant defence mechanisms, especially in the initial interaction and detection of a pathogen invasion. Out of the 131 putative genes, 34 simple sequence repeats (SSR) were identified, of which primer pairs were successfully designed for ten SSR. The primer pairs were screened on four populations with different susceptibility/tolerance levels to BSR infection, and the segregation patterns were evaluated and scored. Phylogenetic analysis shows distinct patterns that differentiate tolerant samples. Based on this study, the designed primer sets have the potential to predict and select oil palm progenies with higher tolerance levels before planting. Nevertheless, the study needs to be expanded to study the effectiveness of the markers on diverse populations.

P14

***In silico* docking actives compounds of betel leave (*Piper betle L.*) as antimalarial against plasmepsin 1 and plasmepsin 2**

Fatima Wali¹, Billy Kepel², Widdhi Bodhi², Trina Ekawati Tallei³

¹Pharmacy Study Program, Faculty of Mathematics and Natural Sciences, Sam Ratulangi University, Manado, Indonesia

²Department of Chemistry, Faculty of Medicine, Sam Ratulangi University, Manado, Indonesia

³Department of Biology, Faculty of Mathematics and Natural Sciences, Sam Ratulangi University, Manado, Indonesia

Corresponding authors: Fatima Wali and Trina Ekawati Tallei

Abstract

This study aims to determine the potential inhibition of bioactive compounds contained in betel leaf (*Piper betle L*) through the *in silico* approach compared to artemisin and chloroquine as comparative antimalarial compounds against the inhibition of plasmepsin 1 and plasmepsin 2. The dry powder of betel leaf was extracted using n-hexane and ethyl acetate solvents then analyzed using Gas Chromatography-Mass Spectrometer (GC-MS) to obtain information about the compounds contained in the betel leaf. Molecular docking in silico was used over the ANT protein-ligand system using Autodock Vina and compared it with artemisin and chloroquine. The results showed that there were 46 compounds detected in n-hexane and methanol extract of betel leaf using GC-MS analysis. Molecular docking studies showed that the five active compounds that have the strongest binding affinity for the 1LEE receptor are Androstan-17-one, 3-ethyl-3-hydroxy -, (5.alpha.) - ; Torreyol ; t-muurolol ; delta- Cadinene ; and epiglobulol, with binding affinity of -8, -6.6, -6.5, -6.4, -6.4 kcal/mol and five active compounds that have the strongest binding affinity to the 3QS1 receptor, are Androstan-17-one, 3-ethyl-3-hydroxy-(5.alpha.) - ; alpha.-Longipinene ; t-muurolol ; Longipinocarveol trans- ; and alpha.-Longipinene with binding affinity of -9, -7,1, -7, -6,6, -6, 4 kcal/mol were compared with artemisinin and chloroquin with binding affinity of -6.7 and -5.4 to the 1LEE receptor and -7.2 and -6.1 to the 3QS1 receptor, respectively. The Androstan-17-one,3-ethyl-3-hydroxy-,(5.alpha.) steroid present in betel leaf show the stronger activity than artemisin and chloroquine to inhibit plasmepsin 1 and plasmepsin 2.

P15

Multiplex primer design and optimization for effective detection of pathogenic bacteria *Aeromonas hydrophila*

Siti Triani Rakhmirianti¹, Any Aryani¹, Trina Ekawati Tallei², Diah Kusumawaty¹

¹Departement of Biology Education, Faculty of Mathematics and Natural Science Education, Indonesia University of Education, Indonesia

²Department of Biology, Faculty of Mathematics and Natural Science, Sam Ratulangi University, Indonesia

Corresponding author: Diah Kusumawaty

Abstract

Aeromonas hydrophila is a microbial flora in fresh water fishes that can cause Motile Aeromonas Septicemia (MAS) disease if the fish has a low immune. The aim of this study is to design and optimize multiplex primers for effective detection of pathogenic *Aeromonas hydrophila*. Molecular assay were used to test the validity of *A. hydrophila* samples using Polymerase Chain Reaction (PCR) with 16s rRNA and nine *Aeromonas* virulent genes (*alt*, *act*, *ast*, *aerA*, *pla/lip*, *ahyB*, *asc-FG*, *lip*, and *hlyA*). The in silico multiplex PCR test was carried out then the in vitro test was performed. The result of validity test with 16s rRNA showed all of the samples amplified an amplicon of 1,500 bp. Seven of nine virulent genes were amplified, six of them were used as multiplex primers (mPCR-1: *act*, *alt*, *ast*; mPCR-2: *aerA*, *pla/lip*, *ahyB*). The results of multiplex PCR amplification showed that mPCR-1 primer sets (*act* and *ast*) with Ta 65°C and the same combination primers concentration also mPCR-2 primer sets (*aerA*, *pla/lip*, *ahyB*) with Ta 55°C and *ahyB* primer concentration increasing can be used. Based on the in silico multiplex PCR test, the mPCR-1 primer set designed in the study can be used as a solution to the amplification problem.

P16

Community structure and bacterial diversity in digestive tract of cultivated elver eel based on metagenomic analysis

Stella Melbournita Noor Augustine¹, Any Aryani¹, Trina Ekawati Tallei², Diah Kusumawaty¹

¹Biology, Department of Biology Education, Faculty of Mathematics and Natural Science Education, Indonesia University of Education, Indonesia

²Department of Biology, Faculty of Mathematics and Natural Science, Sam Ratulangi University, Indonesia

Corresponding author: Diah Kusumawaty

Abstract

Eel (*Anguilla bicolor bicolor*) is a very economical fish in both local and foreign markets. The content of vitamins and micronutrients in eel is very high. One aspect that contributes to eel's health is the balance of the microbiota that inhabit the digestive tract. This research aims to study the community structure, diversity, and abundance of bacteria that inhabit elver eels' digestive tract. The method used to investigate the bacterial community's information is by taking total bacterial genomic DNA from the digestive tract of elver eel. The concentration and purity of the DNA obtained were analyzed using the OD ratio of 260/280. The examination of the samples was carried out by targeting the V3-V4 regions of the 16S rRNA gene using Next Generation Sequencing method. Data analysis was performed using Mothur software and PAST v.3.26 for calculating alpha diversity. Proteobacteria (64%) was found as the dominant phylum followed by Firmicutes (29%), and other phyla. Digestive tract microbiota of elver eels was dominated by the genus *Plesiomonas*. Information obtained from the digestive tract's microbiota can be useful for determining health and improving maintenance conditions in eel farming.

P17

Design and analysis of peptide inhibitors against the G12D KRAS protein in colon cancer

Ayesha Fatima¹, Wong Leong Yung¹

¹Faculty of Pharmacy, Quest International University Perak, Malaysia

Corresponding author: Ayesha Fatima

Abstract

KRAS protein remains an elusive protein till date as there are not many approved drugs for treating mutated KRAS induced cancers. Several efforts over the past years have focused on finding upstream or downstream inhibitors in the KRAS signalling pathway. G12D is the most common mutation found in patients suffering from colon cancer. This study attempted to design inhibitory peptides using structure based drug designing method and determined the binding potential of the designed peptides against mutated G12D KRAS protein by protein-protein docking studies using the HawkDock server. A total of 108 potential inhibitory peptides were designed and shortlisted based on docked pose in the GDP binding site of mutated KRAS G12D protein. 9 potential inhibitory peptides were selected which bound to the site could be useful in preventing the conversion GDP to GTP. The free binding energy of these selected inhibitory peptides was then measured using MM/GBSA method and re-ranked based on the output. The peptide with the amino acid sequence DCWRHRLCID folded in a simple loop like structure showed the highest free binding energy of -57.59 kcal/mol. Our study concluded that short peptides 10-aminoacid peptides could be designed efficiently and had a high probability to be further developed as a therapeutic option for mutated KRAS G12D protein.

P18

Exploring the patterns of codon usage and host adaptation in Sin Nombre virus

Himani Malhotra¹, Nitesh Kharga¹, Ayan Roy¹

¹Department of Biotechnology, coSchool of Bioengineering and Biosciences, Lovely Professional University, Punjab, India

Corresponding author: Ayan Roy

Abstract

Sin Nombre orthohantavirus (SNV), first isolated in the Four Corners region of United States, causes a severe pulmonary disease among humans commonly referred to as Hantavirus cardiopulmonary syndrome (HCPS). Deer mouse (*Peromyscus maniculatus*) acts as the natural reservoir of the virus and develops persistent infection with SNV. Human beings serve as the “dead-end” host for SNV and develop acute viral pneumonia with an alarmingly high mortality rate of around 50%. The present study intends to comprehensively investigate the complex codon usage patterns of the viruses and assess the signatures of host adaptation. Multivariate statistical analysis revealed compositional constraint and host selection pressure to be instrumental in shaping the viral codon usage patterns. Interestingly, the viral adaptation, as estimated by codon adaptation index (CAI), was observed to be significantly higher in human host compared to the natural reservoir *P. maniculatus* which might point towards the rapid progression of infection in humans. Higher similarity index of the SNV with respect to *P. maniculatus*, in contrast to the human genome, might indicate towards the co-evolution of SNV with its natural reservoir *P. maniculatus*. Similar trend of dinucleotide abundance among the viral pathogen and human host reinforced the patterns of high adaptation of SNV in human cellular niche. Dinucleotides UpG and CpA were noted to be over-represented whereas, dinucleotide CpG was observed to be strictly avoided among the SNV and human genomes. Strong restraint from the usage of CpG dinucleotide among SNV might be attributed to the strategy of evading human immune responses.

P19

Comparing linear and non-linear machine learning models for predicting the pathogenicity of rare missense variants in hereditary cancer

Seonhwa Kim¹, Wonseok Yoo¹, Changbum Hong¹, Kyu-Baek Hwang²

¹Research Center, Software Division, NGeneBio, Seoul, Korea

²School of Computer Science and Engineering, Soongsil University, Seoul, Korea

Corresponding authors: Changbum Hong and Kyu-Baek Hwang

Abstract

Many germline variants are responsible for hereditary cancers. For example, pathogenic variants of BRCA1 and BRCA2 increase the risk of breast, ovarian, and prostate cancers. Next-generation sequencing (NGS)-based gene panel tests are a comprehensive means to find variants and identify individuals at risk for inherited cancer susceptibility. However, a substantial number of rare missense variants called by NGS are of uncertain significance, impeding clear clinical decision-making. Thus, prediction of the pathogenicity of variants of uncertain significance (VUS) would increase the clinical utility of gene panel tests. For this, we applied machine learning methods to predict the pathogenicity of VUS based on their similarities to known pathogenic and benign variants. More precisely, machine learning models which classify rare missense variants as pathogenic or benign were built using high-confidence ClinVar variants annotated as pathogenic or benign. Three linear and one non-linear classification methods, i.e., the ridge, the least absolute shrinkage and selection operator (lasso), the elastic net, and the random forest were employed. We used a number of features such as reported minor allele frequency, exon location in the transcript, and multiple computational-prediction scores of functional impact and conservation. In a 5-fold cross validation experiment, the non-linear model (random forest) showed the best classification performance (area under the precision-recall curve: 0.9879 ± 0.0092 , area under the receiver operating characteristics curve: 0.9745 ± 0.0063) among the four methods. Our results suggest that machine learning is effective in improving the utility of NGS-based gene panel tests and the random forest has higher priority than others in this regard.

P20

Molecular docking of cobra venom cytotoxin with death receptors

Nurhamimah Misuan¹, Saharuddin Mohamad², Michelle Khai Khun Yap¹

¹School of Science, Monash University Malaysia, Bandar Sunway, Malaysia

²Institute of Biological Sciences, Faculty of Science, University of Malaya, Kuala Lumpur, Malaysia.

Corresponding author: Michelle Khai Khun Yap

Abstract

Cytotoxin is a three-finger toxin present predominantly only in cobra venom. The functional site of the toxin is located at its three hydrophobic loop tips. The actual mechanistic action of cytotoxicity remains inconclusive as few conflicting hypotheses were proposed besides direct cytolytic effects. The present work aims to investigate the interaction between cytotoxin with death receptor families such as Fas-ligand and tumor necrosis factor (TNF) receptors via in-silico molecular docking. A conserved cytotoxin sequence was constructed by multiple sequences alignment. The three-dimensional structure of the universal cytotoxin was later determined using homology modelling and quality validation by Ramachandran plot. The proximity of toxin interaction with death receptors was visualised using HADDOCK with the calculation of intermolecular forces. Our results showed that most of the death receptors interacted with cytotoxin. However, TNF receptor 1 (TNFR1) was the best receptor interacting with cytotoxin due to binding of all three functional loops with the receptor, high HADDOCK score, low z-score and RMSD value for TNFR1. The possible intermolecular interactions between cytotoxin and TNFR1 were Van der Waals forces and hydrogen bonding. Our findings suggest a possibility of cytotoxin triggers caspase-mediated apoptosis through non-covalent interactions with TNFR1. Further experimental receptor binding assays will be conducted to confirm the interaction of cytotoxin with TNFR1.

P21

Elastic SCAD SVM cluster for the selection of informative functional connectivity in autism spectrum disorder classification

Sin Yee Yap¹, Weng Howe Chan^{1,2}

¹School of Computing, Faculty of Engineering, Universiti Teknologi Malaysia

²Artificial Intelligence and Bioinformatics Group, School of Computing, Faculty of Engineering, Universiti Teknologi Malaysia

Corresponding author: Sin Yee Yap

Abstract

One of the proposed biomarkers for psychiatric disorders such as autism spectrum disorder (ASD) is the functional connectivity which can be extracted from magnetic resonance images (MRI). As the functional connectivity for the ASD studies is extracted based on huge number of brain regions, this leads to the high dimensionality of the functional connectivity data. This poster demonstrates a cluster framework using multiple Elastic SCAD SVM (ES-SVM) to reduce the dimensionality of the data, the accumulation of each set of selected functional connectivity from different permutation of training data covers more inclusive informative functional connectivity for classification purpose. Experiments had been done using 144 MRI from Autism Brain Imaging Data Exchange (ABIDE) dataset. The 5-fold cross validation results comparing the existing SVM and the proposed cluster shows that the latter performs better in terms of accuracy (+25.06%), sensitivity (+28.29%) and specificity (+22.06%). Additionally, the selected informative functional connectivity set is associated with the pairs of brain regions which are closely related to ASD, for instances, precentral gyrus with middle temporal gyrus, superior frontal gyrus with inferior frontal gyrus, and paracentral lobule with postcentral gyrus. However, there are several important considerations in this approach to generate a more reliable input data, which includes the diversity of the subject pool, the pre-processing of MRI, and the extraction method of functional connectivity. Therefore, more inclusive experiments can be done in future based on these considerations, as well as applying this approach onto other psychiatric disorders which can be studied using functional connectivity.

P22

Whole genome sequencing and analysis of Indian isolate of *Leptospira*

Kumari Snehkant Lata^{1,2}, Vibhisha Vaghasia^{1,2}, Shivarudrappa B. Bhairappanavar¹, Swapnil Kumar¹, Garima Ayachit², Saumya Patel², Jayashankar Das^{1,3}

¹Gujarat Biotechnology Research Centre, Department of Science & Technology, Government of Gujarat, Gandhinagar – 382011, India

²Department of Botany, Bioinformatics and Climate Change, University School of Sciences, Gujarat University, Ahmedabad – 380009, India

³Centre for Genomics and Biomedical Informatics, IMS and SUM Hospital, Siksha “O” Anusandhan (Deemed to be University), Bhubaneswar, Odisha – 751003, India

Corresponding author: Jayashankar Das

Abstract

Leptospirosis, caused by pathogenic *Leptospira spp*, is a re-emerging zoonotic disease that affects both public health and the economy worldwide. The genomic study of new *Leptospira* isolates from different geographic locations may illuminate and expand our limited understanding of the pathogenesis and evolutionary mechanism of *Leptospira*. There is a dearth of whole-genome sequence data of *Leptospira* isolates from India and their comparison with already available strains. In this study, we performed whole-genome sequencing and analysis of the most abundant and virulent species “*Leptospira interrogans* serovar Pyrogenes strain Salinem” from India. For sequencing, the genomic DNA library was prepared with a fragment size of 400 bp using the IonXpress Plus Fragment library kit and loaded onto the Ion 530TM Chip and sequenced using the Ion S5TM system. Consequently, genome sequences were de novo assembled followed by scaffolding and gap filling. Functional annotation, structural comparison and variant analysis were also performed. The structural analysis of genome showed potential evidence of inversions and structural rearrangement when compared with closely related strain. The variant calling analysis revealed 1151 single nucleotide polymorphisms (SNPs) with 272 missense and 316 synonymous mutations. It was observed that among these mutations the most mutated locus was “LEP1GSC019_1675” which contains 63 synonymous and 24 missense mutations. This gene encodes “virulence plasmid 65kDa B” protein which is involved in bacterial pathogenesis. Thus, the genomic data analyzed and presented here may prove to be a valuable add-on resource for deeper understanding of global diversity, evolution, and pathogenesis of *Leptospira spp*.

P23

Whole genome sequencing and de novo assembly of three virulent Indian isolates of *Leptospira*

Kumari Snehkant Lata^{1,2}, Vibhisha Vaghasia^{1,2}, Shivarudrappa B. Bhairappanavar¹, Swapnil Kumar¹, Garima Ayachit², Saumya Patel², Jayashankar Das^{1,3}

¹Gujarat Biotechnology Research Centre, Department of Science & Technology, Government of Gujarat, Gandhinagar – 382011, India

²Department of Botany, Bioinformatics and Climate Change, University School of Sciences, Gujarat University, Ahmedabad – 380009, India

³Centre for Genomics and Biomedical Informatics, IMS and SUM Hospital, Siksha “O” Anusandhan (Deemed to be University), Bhubaneswar, Odisha – 751003, India

Corresponding author: Jayashankar Das

Abstract

Leptospirosis is a re-emerging bacterial zoonosis caused by pathogenic *Leptospira*, with a worldwide distribution and becoming a major public health concern. Prophylaxis of this disease is difficult due to several factors such as non-specific variable clinical manifestation, presence of a large number of serovar, species and asymptomatic reservoir hosts, lack of proper diagnostics and vaccines. Despite its global importance and severity of the disease, knowledge about the molecular mechanism of pathogenesis and evolution of pathogenic species of *Leptospira* remains limited. In this study, we sequenced and analyzed three highly pathogenic species of Indian isolates of *Leptospira* (*interrogans*, *santarosai*, and *kirschneri*). Additionally, we identified some virulence-related and CRISPR-Cas genes. The virulent analysis showed 232 potential virulence factors encoding proteins in *L. interrogans* strain Salinem and *L. santarosai* strain M-4 genome. While the genome of *L. kirschneri* strain Wumalasena was predicted to encode 198 virulence factor proteins. The variant calling analysis revealed 1151, 19,786, and 22,996 single nucleotide polymorphisms (SNPs) for *L. interrogans* strain Salinem, *L. kirschneri* strain Wumalasena and *L. santarosai* strain M-4, respectively, with a maximum of 5315 missense and 12,221 synonymous mutations for *L. santarosai* strain M-4. The structural analyses of genomes indicated potential evidence of inversions and structural rearrangement in all three genomes. The availability of these genome sequences and *in silico* analysis of *Leptospira* will provide a basis for a deeper understanding of their molecular diversity and pathogenesis mechanism, and further pave a way towards proper management of the disease.

P24

Preliminary structure-based drug discovery on Cathepsin S as an anti-inflammatory target

Christian Kai Cheng Yee¹, Manoj Valappil¹, Wai Keat Yam²

¹Perdana University - Royal College of Surgeons in Ireland School of Medicine, Perdana University, Serdang, Selangor, Malaysia

²Centre for Bioinformatics, School of Data Sciences, Perdana University, Serdang, Selangor, Malaysia

Corresponding author: Wai Keat Yam

Abstract

Cathepsins are protease enzymes and can be categorized into 3 subtypes - serine, cysteine and aspartic. There are currently 15 known classes of cathepsins (Cathepsins A, B, C, D, E, F, G, H, K, L, O, S, V, W and Z). The cysteine cathepsins are a group of proteases from the family of papain-like cysteine proteases, found in lysosomes. Among these major cysteine cathepsins is Cathepsin S, which can be found in the lysosomes of antigen presenting cells. Cathepsin S has been shown to be involved in the production of IL-1 β , a key cytokine responsible for mediating inflammatory response to sterile particles resulting in diseases like silicosis, gout and atherosclerosis. The main objective of this study is to identify potential anti-inflammatory inhibitors that could target Cathepsin S. The trifluoromethylphenyl P2 motif was previously reported in the design and synthesis of highly potent novel ketoamide-based cathepsin S inhibitors. Therefore, to establish a valid positive control for this docking study, a control dock was performed on Cathepsin S with trifluoromethylphenyl P2 motif. It was then followed by virtual screening with compounds from ZINC database. Virtual screening results showed a list of compounds with the lowest free energy binding values ranging from -9.52 to -1.18 kcal/mol. The control docking study showed lowest free energy binding of -5.45 kcal/mol and hence the filtering of compounds were done based on this range of free energy binding values. Analysis and characterizing of the molecular interactions between the target and potential inhibitors was done using Protein-Ligand Interaction Profiler. These results showed preliminary observations to further ascertain Cathepsin S as a potential anti-inflammatory drug target.

P25

Identifying *Drosophila* cis-regulatory modules by a deep convolutional neural network on multiple transcriptional regulatory features

Jing-Xi Xu¹, Zong-Xiao Yang¹, Tzu-Hsien Yang¹

¹Department of Information Management, National University of Kaohsiung, Taiwan

Corresponding author: Tzu-Hsien Yang

Abstract

The modular DNA regulatory sequences, or cis-regulatory modules (CRMs), help control metazoan transcription regulation in differential gene expression under different developmental stages. Thus, understanding the distribution of CRMs in the genomic scale is an important task. In this research, five different site-by-site experimental transcriptional regulatory data, including histone modification ChIP-seq signals, chromatin binding protein ChIP-seq signals, TFBS odds ratio data calculated from the position weight matrices, sequence conservation and nucleosome occupancy DNase-seq signals, were integrated to build a deep convolution neural network that identifies genome-wide *Drosophila* CRMs. We adopted 24,523 verified CRMs from the REDFly database as our ground truth dataset for model training and hyperparameter selection. Our method first processed the aforementioned five features by five independent feature encoders and then concatenated these encoded features to form a five-channelled “genomic image”. Secondly, we use the convolutional filtering to help extract regulatory patterns from the five-channelled genomic images. Then these regulatory patterns were used to check if the given sequence is a functional CRM. In the 5-fold cross-validation, our model obtained an AUC = 86% (area under curve) in the receiver operating characteristic (ROC) curve, showing good CRM discrimination. We further collected extra CRMs from CRM Activity Database (CAD) as an independent test set. And the results of our method in the independent test set showed better performance in AUC than other existing CRM prediction methods. Therefore, we believe that the proposed method can help biologists find valuable CRM candidates for subsequent transcription regulation experiments.

P26

Automatic transcriptional factor-gene interaction literature evidence extraction via temporal convolutional neural networks

Yu-Huai Yu¹, Jing-Xi Xu¹, Chi-Fen Liao¹, Tzu-Hsien Yang¹

¹Department of Information Management, National University of Kaohsiung, Taiwan

Corresponding author: Tzu-Hsien Yang

Abstract

Cells control gene expression via rebuilding and regulating their cellular gene transcription programs. And correct gene transcriptional regulation requires suitable transcription factor (TF)-gene interactions. Therefore, it is of great interest for biologists to find out the detailed molecular mechanisms elucidating the interactions between TFs and genes. Numerous TF-gene interaction results have been published in the literature. These results include direct TF-gene binding evidence and indirect TF-gene regulation evidence. Due to the blooming of next-generation sequencing and nanopore techniques, more and more literature evidence concerning TF-gene interactions are piled these days. However, the TF-gene interaction information is still fragmentary and cannot be easily referenced. To solve this problem, in this research, we designed a deep learning tool to automatically extract the evidence of TF-gene interactions from the literatures deposited in PubMed. A deep learning model extended from BERT and temporal convolutional neural networks to automatically dig out literatures related to direct TF-gene binding and indirect TF-gene regulation were constructed in our tool. We trained the literature mining model on the manually curated TF-gene interaction evidence from YEASTRACT. Using the techniques of random training/validation/test splits, the results of the proposed deep learning model showed good performance in literature labelling for yeast TF-gene interactions. Further, the proposed tool presents a nice and clear web interface for users to obtain the literature abstracts with marked evidence sentences. The tool will be freely available online for academic usages.

P27

Melanoma detection via deep transfer learning

Dong-Ying Guo¹, Tzu-Hsien Yang¹

¹Department of Information Management, Nation University of Kaohsiung, Taiwan

Corresponding author: Tzu-Hsien Yang

Abstract

Skin cancer recognition is now one of the demanding tasks in medical diagnosis support systems because of its high fatality rate. And patients' survival rate can be greatly increased if the malignant melanocytes are found in their earlier stages. Currently, it is hard to correctly distinguish melanoma from other skin lesions because of factors such as skin conditions, lesion sites, and the varied appearances of lesions. In this study, we focused on the task of melanoma detection using deep transfer learning. We designed a convolutional neural network extended from the InceptionV3 model architecture to identify melanoma in regular skin images. We trained our model using a dataset consisting of 4,522 melanoma images and 20,809 normal/benign lesion skin images. The model architecture consists of 8 convolutional layers followed by pooling layers, two fully-connected layers and the softmax layer. Using the technique of 5-fold cross-validation, our model achieved a high value of area under curve (AUC = 0.852) in the receiver operating characteristic (ROC) curve plot, showing the good melanoma detection performance of the proposed model. The proposed model was further tested on another independent test of 584 melanoma images and 32,108 normal skin pictures. The results on the independent set also indicated that our model is applicable to unseen skin images and shows good melanoma detection performance.

P28

Novel biological metrics for evaluating the functional significance of RNA secondary structure predictions

Yu-Cian Lin¹, Tzu-Hsien Yang¹

¹Department of Information Management, National University of Kaohsiung, Taiwan

Corresponding author: Tzu-Hsien Yang

Abstract

It is now known that non-coding RNA molecules play important roles in cellular functions. And lots of functions of ncRNAs depend on their secondary structures or tertiary structures. Therefore, understanding RNA folding is an important issue in the study of the functions of ncRNA molecules. Today, there are many tools developed to help researchers obtain RNA structures from the given RNA sequences. However, effective biological metrics to evaluate the biological significance for each of the predicted RNA secondary structures are still missing. Thus, in this study, we designed four novel biological metrics (functional fitness significance metric, RNA-protein interaction significance metric, translational regulation significance metric, post-transcriptional regulation significance metric) that indicate the 4 non-mutually exclusive functional aspects for different predicted RNA structures to be potentially involved in based on different high-throughput RNA experimental data. To verify the proposed biological metrics, a test set of 125 yeast ncRNA sequences with verified structures from BRALibaseII and the literature was gathered. 22 structure prediction tools were used to obtain the predicted structures for each of the 125 yeast sequences and the designed four biological metrics for each of predicted structure were calculated. For each of the 125 ncRNAs, the predicted structures with the best significance among the 4 biological metrics obtained higher average F1 values than any of the 22 tools. These results demonstrated that a more biologically favourable structure can be selected by considering the functional aspects of the predicted structures.

P29

Systems pharmacology approach to identify the activity of bioactive molecules against cervical cancer

Aarthy Murali¹, Sanjeev Kumar Singh¹

¹CADD and Molecular Modelling Lab, Alagappa University, Karaikudi, Tamil Nadu, India

Corresponding author: Sanjeev Kumar Singh

Abstract

The immunological behaviour of viral infections in cervical cancer at molecular levels is very much essential for the therapeutic development and needs to be elucidated clearly. In such cases, Systems pharmacology analyses and multi omics helps in unravelling the multi-targeted mechanisms of novel biologically active compounds to treat cervical cancer. The Immunotranscriptomic dataset of healthy and infected cervical cancer patients were retrieved from array express. Secondly, the phytochemicals from the plant species were collected from the Pubchem databases. A total of 132 immune genes which is differentially expressed responsible for the cervical cancer is derived through Network Analyst 3.0. Among 75 compounds reported in plants for treating cervical cancer, only minimal compounds were targeting the immune genes of cervical cancer. The significant genes responsible for the supremacy in the cervical cancer are identified in this study. Along with this, the pharmacological properties of these phytochemicals were compared with the reported compounds and some compounds are found to be efficient with better bioactive scores. The virogenomic signatures observed from the cervical cancer caused by the oncoproteins serve as a therapeutic targets and the identified compound can serve for the anti-HPV drug deliveries. In future, the experimental validation of the obtained results will be carried out on the optimized small molecules to be treated as commercial drug candidates.

P30

Identification and study of specific genomic variants of the Kazakh population using comparative population genomics analysis

Askhat Molkenov¹, Daniyar Karabayev¹, Ainur Seisenova¹, Asset Daniyarov¹, Aigul Sharip¹, Ulykbek Kairov¹

¹Laboratory of Bioinformatics and Systems Biology, Center for Life Sciences, National Laboratory Astana, Nazarbayev University, Nur-Sultan, Kazakhstan

Corresponding authors: Askhat Molkenov and Ulykbek Kairov

Abstract

The modern development of high-throughput genomic technologies opens up new opportunities for detailed studying the human genome. Large-scale genomic research data together with the active development of bioinformatics makes it possible to create detailed databases and comprehensively study genomic data. One of the contemporary tasks is to study and identify specific genomic variants of a population by detailed analysis of whole-genome and whole-exome data in comparison with open large genomic datasets of populations. The materials of the study are 14 whole-genomes and 125 whole-exomes of Kazakhstani individuals. Our dataset was replenished with data from large whole-genome population datasets (The Simons Genome Diversity Project, Jorde lab data, Human Genome Diversity Project, and 1000 Genomes Project) for comparative population genomics and to search and identify specific genomic variants. For replenished datasets formed a general map of all variants, which were then excluded from the total number variants found for Kazakh sampling to search for specific genomic variants. Then the filtered variants were annotated and interpreted. For Kazakh whole-exomes were found 9 heterozygous or mutant variants unique among formed genomic databases. 7 variants located on the intron region, 1 on the upstream, and the last variant frameshift deletion on the exonic region. For the Kazakh whole-genomes were found 144 mutant variants that were presented among all Kazakh samples. Only 8 SNVs are located at the exonic region: 4 synonymous SNV, 3 nonsynonymous SNV, and 1 frameshift deletion. We have discovered several unique genomic variants specific for now to the Kazakh individuals.

P31

Validation of predicted novel Myc motifs mediating important PPIs using computational approaches

Debangana Chakravorty¹, Abhirupa Ghosh¹, Sudipto Saha¹

¹Division of Bioinformatics, Bose Institute, P-1/12 CIT Road Scheme VIIM, Kolkata 700054, WB India.

Corresponding author: Sudipto Saha

Abstract

Myc has a large stretch of disordered region whose structure is unknown and it contains many short conserved regions that mediate Protein-Protein Interactions (PPIs). Previously, many such experimentally validated and predicted linear motifs have been identified in Myc. Two novel predicted linear motifs which mediate important interactions were selected to be validated. The interaction of Myc motifs and their PPI partners were explored and visualized using protein-protein docking servers followed by the identification of binding interfaces. Computational Alanine scanning was performed to identify key residues in the motif and the effects of mutation in these residues were explored by prediction servers. One key residue mutation in each motif was studied by Molecular Dynamics simulation. These key residues may prove to be of importance for mediating the PPI if its mutation leads to destabilization of the interaction.

P32

Role of surface-exposed charged basic amino acids (Lys, Arg) and guanidination in insulin on the interaction and stability of insulin–insulin receptor complex

Vannajan Sanghiran Lee¹, Sri Devi Sukumaran¹, Tan Pak Kheong², Umah Rani Kuppasamy²,
Bavani Arumugam²

¹Department of Chemistry, Drug Design Development Research Group, Center of Theoretical and Computational Physics (TCP), Faculty of Science, University of Malaya, 50603 Kuala Lumpur, Malaysia

²Department of Biomedical Science, Faculty of Medicine, University of Malaya, 50603 Kuala Lumpur, Malaysia

Corresponding author: Bavani Arumugam

Abstract

Naturally occurring proteins are emerging as novel therapeutics in protein-based biopharmaceutical industry for the treatment of diabetes and obesity. However, proteins are not suitable for oral delivery due to short half-life, reduced physical and chemical stability and low permeability across the membrane. Chemical modification has been identified as a formulation strategy to enhance the stability and bioavailability of protein drugs. The present study aims to study the effect of charge-specific modification of basic amino acids (Lys, Arg) and guanidination on the interaction of insulin with its receptor using molecular modelling. Our investigation revealed that the guanidination of insulin (Lys-NHC=NHNH₂) enhanced and exerted stronger binding of the protein to its receptor through electrostatic interaction than native insulin (Lys-NH₃⁺). Point mutations of Lys and Arg (R22, K29; R22K, K29; R22, K29R; R22K, K29R) were attempted and the effects on the interaction and stability between insulin/modified insulins and insulin receptor were also analyzed in this study. The findings from the study are expected to provide a better understanding of the possible mechanism of action of the modified protein at a molecular level before advancing to real experiments.

P33

Early investigation into the origin and evolution of SARS-CoV-2

Chulochana Sivanantham¹, Christian Yee Kai Cheng¹, Nur Atiqah Azhar², Mohammad Asif Khan^{2,3}, Manoj Valappil¹

¹Perdana University-Royal college of Surgeons in Ireland School of Medicine, Malaysia

²School of Data Sciences, Perdana University, Malaysia

³Bezmialem Vakıf Üniversitesi, Turkey

Corresponding author: Manoj Valappil

Abstract

Introduction: Determining the origin and evolution of emerging pathogens is important for its surveillance, control and developing diagnostic, therapeutic, and preventative strategies.

Objectives: To explore the molecular divergence of SARS-CoV-2 from other coronaviruses and track its molecular evolution between December 2019 and February 2020.

Methods: Whole genome sequence (WGS) of 13 coronavirus from different species (including SARS CoV, MERS CoV, 4 human coronaviruses [HCoV-OC43, HCoV-229E, HCoV-NL63, HCoVHKU1], the first released SARS-CoV-2 WGS from China, and SARS-CoV2 from 52 other countries available at the time were retrieved from NCBI GenBank and GISAID databases. Multiple sequence alignment was performed using MEGA X84 software, with MUSCLE algorithm. The alignment was quality checked using BioEdit and phylogenetic analysis was conducted using MEGA X84, with Maximum Likelihood, Tamura-Nei and Bootstrap methods as the statistical, substitution and test of phylogeny methods respectively. AVANA software was used to undertake variant and diversity analyses.

Results: Bat coronavirus (Bat RaTG13) had the closest similarity to SARS-CoV-2 followed by the GD-Pangolin-CoV. Comparison of coronaviruses known to cause human infections showed that SARS-CoV-2 was more closely related to SARS CoV and MERS-CoV. Analysis of genomes across 53 countries showed that SARS-CoV-2 had evolved into three distinct clades, 24: 26: 3 countries. SARS-CoV-2 from Malaysia was closely similar with SARS-CoV-2 sequence reported from Israel. Diversity analysis revealed that the untranslated protein region (UTR), spike (S) protein and open reading frame (ORF) 1ab contained a position each with the highest entropy values. Variant analysis of the 53 WGSs identified a number of mutations in several regions of SARS-CoV-2 genome. The mutations detected in the sequence from Malaysia were found in ORF1ab, S and membrane (M) proteins. A-to-G nucleotide mutation at position 23,403 resulting in the D164G amino acid change and its commonly associated mutations (positions 241, 3,037, 14,408) conferring high infectivity, were not detected in sequences from any of the 53 countries, at the time.

Conclusions: Although the results are suggestive of Bats being the likely natural hosts of SARS-CoV-2, close similarity of certain regions with CoVs from other species (e.g. Pangolin) raises the possibility of transmission through another intermediate host or independent recombination events. Additional work is needed to clarify the clinical significance of the spatio-temporal evolution of different SARS-CoV-2 clades and elucidate the diagnostic and therapeutic implications of the emerging mutations.

P34

ARIMA modelling for predicting Covid-19 in Indonesia: Integrated moving average, IMA(1,1), for modelling confirmed and cured cases of Covid-19 in Indonesia

Nanda Rizqia Pradana Ratnasari¹

¹Bioinformatics Department, Indonesia International Institute for Life Sciences, Jakarta, Indonesia

Abstract

Covid-19 modelling is needed to help people understanding the distribution or pattern of the data and doing the prediction. The data used for modelling in this study was ‘confirmed cases’ and ‘cured cases’ of Covid-19 in Indonesia recorded from March 2 to August 23, 2020. Model obtained from analysis was IMA(1,1) for both confirmed cases and cured cases. The estimated parameters are -1 and standard error 0.0146 for confirmed cases. Meanwhile, for cured cases, the coefficient and standard error parameters are -1 and 0.0166 respectively. Maximum Likelihood was the method conducted to estimate the parameters. The model was appropriate to predict the actual data of Covid-19 cases in Indonesia as the models can exhibit similar prediction patterns for actual data.

P35

Understanding the carbon concentrating mechanism in *Chlamydomonas reinhardtii*: A systems biology approach

Citu Citu¹, Gitanjali Yadav¹, Sanjeet Kumar Mahtha¹

¹National Institute of Plant Genome Research, India

Corresponding author: Gitanjali Yadav

Abstract

Maintaining food security is a global concern due to increasing population and a finite agricultural land area. One approach is to increase the productivity of crop plants, that can be correlated with increasing the efficiency of photosynthetic CO₂ uptake by Ribulose-1,5-bisphosphate Carboxylase/Oxygenase (RuBisCO). Oxygen competes with CO₂ for the active site of RuBisCO, resulting in a loss of fixed CO₂ via photorespiration. Experimental elevation of CO₂ in the vicinity of C₃ crops in the field has been shown to increase yields by suppressing the RuBisCO oxygenase activity. Biophysical Carbon Concentrating Mechanism (CCM) is a process that increases CO₂ fixation and the efficiency of inorganic carbon uptake in cyanobacteria and eukaryotic algae. It elevates CO₂ level in the vicinity of RuBisCO enclosed in a micro compartment called pyrenoid, thus enhancing photosynthetic efficiency. Therefore, one potential approach for improving yields of C₃ crops is the transfer of biophysical CCM into higher plants in order to increase CO₂ fixation rates. Present work focuses on investigating gene networks that regulate fluxes within and between core components of biophysical CCM in the model alga *Chlamydomonas reinhardtii*, to identify molecular interactions and pathways that regulate CCM efficiency. Towards this, pyrenoid-enriched components have been selected and their expression profiles from available diurnally synchronized transcriptome in carbon-deprived condition were analyzed. The categorization of these important CCM components into different time zones was done based on the cell cycle of *C. reinhardtii*. Co-expressing and interacting partners of pyrenoid-enriched components have been identified, visualization of their respective networks at every time point has been done and nodes of these networks have been annotated to decipher their corresponding biological processes. Promoter sequences for all component genes have been extracted and assessed for statistically significant cis-binding sites as well as transcription factors regulating these components. This work will further help to identify the novel components of CCM by reducing data dimensionality.

P36

Computational approaches to understand role of Argonautes in host-encoded resistance against vector borne plant RNA viruses

Vinita Lamba¹, Gitanjali Yadav^{1,2}, Ravi Kiran Purama¹

¹Computational Biology Laboratory, National Institute of Plant Genome Research, Aruna Asaf Ali Marg, New Delhi, 110067 India

²Department of Plant Sciences, University of Cambridge, Downing Site, Cambridge, CB2 3EA UK

Corresponding author: Gitanjali Yadav

Abstract

Background: The Argonaute (AGO) protein family plays a central role in RNA induced silencing complex (RISC)-mediated gene regulation in plants. Small RNAs are known to guide AGOs to their specific targets through sequence complementarity, which typically leads to silencing of the target mostly by post-transcriptional inhibition or mRNA degradation.

Methods: We have identified Argonautes & Argonaute-like proteins in viridiplantae, followed by functional annotation and classification into clades, by means of detailed sequence analyses. We also performed structure based Evolutionary trace studies to identify class specific residues of AGO proteins and these two methodological approaches were complemented by transcriptional regulatory inferences derived from gene expression datasets in model plants Rice and Arabidopsis using Geneinvestigator and gene co-expression analyses.

Results: AGO identification, annotation and classification enabled us to quantify the expansion, and diversification of this critical and less explored family in the plant kingdom. Ev-trace studies helped us to identify class specific residues of AGO proteins that may have hitherto unidentified roles in target acquisition or functional interaction. Transcriptional regulatory inferences helped to understand host response against plant viruses at key regulatory pathways, particularly for R-genes (AGO proteins), dicers, and RDRPs of the RISC repertoire proteins. This response was observed in *A. thaliana* and *O. sativa*, in a tissue-specific and developmental stage specific manner. Co-expression studies provided a comprehensive understanding of the organization, function and evolution of plant RISC-complex genes, and enabled us to disentangle the complex interactions of RISC-complex proteins which play a significant role in plant defense at various stages of growth/development.

Conclusions: Taken together, our results demonstrate how gene regulatory networks and crucial cellular interactions can be extrapolated with accuracy from large scale sequence, structure and gene expression data, using the case of RISC proteins from plants that are specifically involved in development of viral resistance or proteins that prevent establishment of viral infection in plants.

P37

Application of whole exome-trio analysis in the elucidation of genetic basis of congenital pouch colon

Sonal Gupta¹, Praveen Mathur², Ashwani Kumar Mishra³, Krishna Mohan¹, Obul Reddy Bandapalli⁴, Prashanth Suravajhala¹

¹Birla Institute of Scientific Research, Statue Circle, Jaipur 302021, RJ India

²Department of Pediatric surgery, SMS medical college and Hospital, JLN Marg, Jaipur 302004, RJ India

³DNA Xperts, Noida, UP, India

⁴German Cancer Research Center (DKFZ), Heidelberg, Germany

Corresponding author: Prashanth Suravajhala

Abstract

Anorectal malformations (ARM) are individually common but Congenital Pouch Colon (CPC), a rare anorectal anomaly causes a dilated pouch in the whole or part of colon with invariably fistula in the genitourinary tract. We have earlier attempted to understand the clinical genetic makeup of CPC and identified genes responsible for the disease using whole exome sequencing (WES). Here we report our studies of CPC, by identifying heterozygous missense mutations in 16 proband-parent trios and further discover variants of unknown significance which could provide insights into CPC manifestation and its aetiology. Our study confirms candidate mutations in genes, viz. AK9, SQSTM1 and FRG1 besides emphasizing the role of hypothetical genes or open reading frames causing this developmental disorder. Variant validation revealed disease causing mutations associated with CPC and genitourinary diseases which could close the gaps of surgery in bringing intervention in therapies.

P38

Identification and analysis of class specific residues across rice Argonaute family

Ravi Kiran¹, Vinita Lamba¹, Gitanjali Yadav¹

¹National Institute of Plant Genome Research, India

Corresponding author: Gitanjali Yadav

Abstract

Argonautes (AGOs) play a pivotal role in gene regulation through accommodating a repertoire of proteins that forms RNA induced silencing complex (RISC). AGO proteins are structurally divided into four domains: an N-terminal heterogeneous domain, a PAZ domain (3'OH-binding) which binds small RNAs, a mid domain (5' P-binding) and a catalytically active C-terminal PIWI domain. Noncoding RNAs such as miRNAs, siRNAs and Piwi-interacting RNAs (piRNAs) acts as their substrates. The siRNA bound AGO binary complex scans through the target mRNAs to find the complementary region and forms the ternary complex which silence the target post-transcriptionally or causes degradation. Biological roles of very few plant AGO proteins have been elucidated so far. We found four AGO clades in the literature. AGO1, AGO2, AGO4, and AGO5 have been shown to prefer small RNAs possessing specific residues at their 5'terminus which demonstrates their functional diversity. Here, we have identified Argonautes & Argonaute-like proteins from two model plants *Arabidopsis thaliana* and *Oryza sativa* var. *japonica* which possess 10 AGOs and 22 AGOs, respectively. We have seen these 32 AGOs forming different phylogenetic clusters based on their diversification which further classified into four major clades. Further, we used an evolutionary trace (ET) analysis, a phylogenomic method that mimics experimental mutational scanning based on Kitsch algorithm to identify and sort the important amino acids in protein sequences into class specific and non-class specific residues that correlates to the evolutionary divergence of AGOs and later we mapped them to the only available plant AGO-mid domain 3D protein structure.

P39

Predicting lncRNA and protein interactions in prostate cancer using a combination of computational and biophysical methods

Nidhi Shukla¹, Bhumandeep Kour², Ayam Gupta¹, Renuka Suravajhala³, Maneesh Vijay⁴,
Devendra Sharma⁴, Sugunakar Vure⁵, Prashanth Suravajhala¹

¹Birla Institute of Scientific Research, Statue Circle, Jaipur 302021, RJ India

²Lovely Professional University, Jaipur, India

³Manipal University Jaipur, Rajasthan, India

⁴Rukmani Birla Hospitals, Jaipur, India

⁵Lovely Professional University, Jalandhar, India

Corresponding author: Sugunakar Vure

Abstract

Prostate cancer (PCa) is one of the most prevalent cancers worldwide and third most causal cancer in India. Although studies have dealt on the genetics, genomics and the environmental influence in causal of PCa, no association of genotype and phenotype employing the next generation sequencing (NGS) approaches of PCa has been deliberated. Our lab is interested to identify the candidate driver mutations specific to PCa and to check this, we have earlier performed whole exome sequencing (WES) as a pilot on malignant and benign prostate hyperplasia (BPH) as controls in Indian phenotype. From BRCA1 and BRCA2 mutations among as many as 30 causal genes including DNA repair genes, we have ascertained a few long non coding RNAs (lncRNAs) in addition to exploring piwi-RNAs that might play a vital role in the pathogenesis of PCa. Keeping in view of the RNA-protein interactions playing a regulatory role, we have used a combination of computational and biophysical methods to validate the lncRNA-protein interaction pairs by means of virtual pulldown assays, docking complexes and microscale thermophoresis besides characterizing them using RT-PCR. Specifically, we were interested to study interaction of three lncRNAs, viz. NONHSAT053810.2, NONHSAT079881.2 and NONHSAT103724.2 associated with PCa which would provide us possible leads towards discovery of biomarkers and development of novel therapies.

P40

Improving diagnostic approach to amyotrophic lateral sclerosis in India via two-stage NGS panel design

Kanikah Mehndiratta¹

¹University of Glasgow, UK

Abstract

Amyotrophic Lateral Sclerosis, a type of motor neuron disease is a progressive fatal neurodegenerative disorder affecting both upper and lower motor neurons and having symptoms and genes overlapping with many other conditions like multiple sclerosis and dementia. Here, the complexity of the disease ranging from its phenotypic heterogeneity, to interlinked pathways, environmental triggers posing risk and undetermined causes in ALS patients are reviewed, emphasizing the need for better approaches to disease diagnosis and management. Majorly, methods have been intended towards designing a two-stage NGS gene panel specific to variants commonly reported in Indian population using various Bioinformatic tools that analyse 45 genes that were part of existing gene panels in private labs. The detailed description of gene ranging from associated proteins, to pathways, reported variants, disease associated phenotype and the extent of pathogenicity are covered using software like gnomAD, PanelApp, ClinVar, UCSC, GeneCards, String and KEGG. Furthermore, the genes are categorised mainly based on pathogenicity and reported contribution of variants to ALS cases into Definitive, Strong and Moderate categories and a suitable protocol is designed keeping in check with the updated El Escorial diagnostic criteria. Finally, limitations to such a test in India and common environmental triggers are covered to devise a better approach to wide scale application of such a testing service.

P41

Extended mining of the oil biosynthesis pathway in biofuel plant *Jatropha curcas* by combined analysis of transcriptome and gene interactome data

Xuan Zhang^{1,2,3}, Jing Li^{1,2,3}, Bang-Zhen Pan^{1,2,3}, Wen Chen¹, Maosheng Chen^{1,2,3,4}, Mingyong Tang^{1,2,3}, Zeng-Fu Xu^{1,2,3}, Changning Liu^{1,2,3}

¹CAS Key Laboratory of Tropical Plant Resources and Sustainable Use, Xishuangbanna Tropical Botanical Garden, Chinese Academy of Sciences, Kunming, Yunnan, 650223, China

²Center of Economic Botany, Core Botanical Gardens, Chinese Academy of Sciences, Menglun, Mengla, Yunnan 666303, China

³The Innovative Academy of Seed Design, Chinese Academy of Sciences, Kunming, Yunnan, 650223, China

⁴College of Life Sciences, University of Chinese Academy of Sciences, Beijing 100049, China

Corresponding authors: Zeng-Fu Xu and Changning Liu

Abstract

Background: *Jatropha curcas* L. is an important non-edible oilseed crop with a promising future in biodiesel production. However, little is known about the molecular biology of oil biosynthesis in this plant when compared with other established oilseed crops, resulting in the absence of agronomically improved varieties of *Jatropha*. To extensively discover the potentially novel genes and pathways associated with the oil biosynthesis in *J. curcas*, new strategy other than homology alignment is on the demand.

Results: In this study, we proposed a multi-step computational framework that integrates transcriptome and gene interactome data to predict functional pathways in non-model organisms in an extended process, and applied it to study oil biosynthesis pathway in *J. curcas*. Using homologous mapping against Arabidopsis and transcriptome profile analysis, we first constructed protein-protein interaction (PPI) and co-expression networks in *J. curcas*. Then, using the homologs of Arabidopsis oil-biosynthesis related genes as seeds, we respectively applied two algorithm models, random walk with restart (RWR) in PPI network and negative binomial distribution (NBD) in coexpression network, to further extend oil-biosynthesis-related pathways and genes in *J. curcas*. At last, using k-nearest neighbors (KNN) algorithm, the predicted genes were further classified into different sub-pathways according to their possible functional roles.

Conclusions: Our method exhibited a highly efficient way of mining the extended oil biosynthesis pathway of *J. curcas*. Overall, 27 novel oil-biosynthesis-related gene candidates were predicted and further assigned to 5 sub-pathways. These findings can help better understanding of the oil biosynthesis pathway of *J. curcas*, as well as paving the way for the following *J. curcas* breeding application.

P42

UNIQmin: An alignment-independent tool for the study of pathogen sequence diversity at any given rank of taxonomy lineage

Li Chuin Chong¹, Wei Lun Lim², Kenneth Hon Kim Ban³, Mohammad Asif Khan^{1,4}

¹Centre for Bioinformatics, School of Data Sciences, Perdana University, Jalan MAEPS Perdana, 43400 Serdang, Selangor Darul Ehsan, Malaysia

²Faculty of Computing and Informatics, Multimedia University, 63100 Cyberjaya, Malaysia

³Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, Singapore

⁴Beykoz Institute of Life Sciences and Biotechnology, Bezmialem Vakif University, Beykoz Istanbul, Turkey

Corresponding author: Mohammad Asif Khan

Abstract

Pathogenic microorganisms are a primary contributor to the global burden of death and disability due to infectious diseases. Sequence variation can expand the infective ability of a pathogen and result in immune escape, posing a key challenge to vaccine and drug design. We present UNIQmin, a tool that utilises an alignment-independent algorithm to generate the minimal set of pathogen sequences, as a method to study their diversity, across any rank of taxonomic lineage, which is impractical using alignment-dependent approaches. The minimal set is the smallest possible number of sequences required to capture the entire repertoire of pathogen peptidome diversity present in a non-redundant sequence dataset. The algorithm is scalable for big data and enables decoding of the minimal set of a pathogen peptidome at any rank of lineage, leading to a better understanding of structure-function and evolution, with possible applications to diagnostic, drug and vaccine target discovery. UNIQmin is implemented in Python 3 (3.7) and available at <https://github.com/ChongLC/MinimalSetofViralPeptidome-UNIQmin>

