The 17th Asia Pacific Bioinformatics Conference

- Conference Guide

10011001100110001100110011
10001100110011001100110
1001 0110001 1001
10011 0011001
10001110001111000110
011 1001 100101 11001

The 17th Asia Pacific Bioinformatics Conference

Wuhan, China

14-16 January 2019

Abstract Booklet

Welcome from the

Organizing Committees

The organizing committees welcome you to the 17th Annual Asia Pacific Bioinformatics Conference (APBC2019), organized by Huazhong Agriculture University College of Informatics.

The APBC is the largest bioinformatics conference in Asia. It promises to bring exciting and groundbreaking research. The scientific program of APBC2019 include 3 keynote, 8 highlight and 51 contributed talks. We are thankful to our esteemed invited and contributing speakers, who will shed light on the research and applications that shape our field today. There are also 33 contributed posters. We believe the coming together of diverse and world-class research expertise will bring bioinformatics studies a step closer towards new breakthrough and inventions.

We sincerely thank the local organizing institutes, China Computer Federation Task Force on Bioinformatics, and Genomics, Proteomics and Bioinformatics journal for their generous support that have made the conference achievable.

We hope you have an enjoying and highly successful meeting.

Conference Chair
David Sankoff, University of Ottawa, Canada

PC Committee Co-Chairs
Louxin Zhang, National University of Singapore, Singapore
Phoebe Chen, La Trobe University, Australia
Shaoliang Peng, Hunan University, China

Local Organizing Committee Co-Chairs
Guoliang Li, Huazhong Agricultural University, China
Hong-Yu Zhang, Huazhong Agricultural University, China

Tutorials and Poster Chair
Kang Ning, Huazhong University of Science and Technology, China

Program Committee

Daniel Huson, University of Tvebingen, Germany
Doheon Lee, KAIST, South Korea
Dongbo Bu, Institute of Computing Technology, CAS, China
Eric Ho, Lafayette College, USA
Fengfeng Zhou, Jilin University, China
Gabriel Valiente, Technical University of Catalonia, Spain
Giltae Song ,Pusan National University, South Korea
Glenn Tesler, University of California at San Diego, USA
Guoliang Li, Huazhong Agricultural University, China
Hiroshi Mamitsuka, Kyoto University and Aalto University
Hong-Yu Zhang, Huazhong Agricultural University, China
Hsien-Da Huang, National Chiao Tung University, ROC
Hsuan-Cheng Huang, National Yang-Ming University, ROC
Hui Lu, Shanghai Jiao Tong University, China
Jens Stoye, University of Bielefeld, Germany
Jialiang Yang, Icahn School of Med. at Mount Sinai, USA
Jian Ma, Carnegie Mellon University, USA
Jie Zheng, Nanyang Technological University, Singapore
Jijun Tang, University of South Carolina, USA
Jinbo Xu, Toyota Technological Institute at Chicago, USA
Juan Liu, Wuhan University, China
Kang Ning, Huazhong University of Science and Technology, China
Katharina Huber, University of East Anglia, UK
Kenta Nakai, University of Tokyo, Japan
Kui Lin, Beijign Normal University, China
Kyungsook Han, Inha University, South Korea
Laxmi Parida, IBM Research, USA
Lenore Cowen, Tufts University, USA
Liqing Zhang, Virginia Tech, USA
Louxin Zhang (Co-Chair), National University of Singapore, Singapore
Lu Tian, Stanford University, USA
Lusheng Wang, City University of Hong Kong, Hong Kong
Manuel Lafond, University of Ottawa, Canada
Min Li, Central South University, China
Min Xu, Carnegie Mellon University, USA
Mingfu Shao, Penn State University, USA
Minghua Deng, Peking University, China
Niko Beerenwinkel, ETH Zurich, Switzerland
Paola Bonizzoni, Università di Milano-Bicocca, Italy
Pawel Górecki, University of Warsaw, Poland
Phoebe Chen (Co-Chair), La Trobe University, Australia
Qian Zhu, Dana Farber Cancer Institute, USA
Quan Zou, Tianjin University, China
Rahul Siddharthan, Institute of Mathematical Sciences, India
S.M. Yiu, University of Hong Kong, Hong Kong

PC Committee

Sagi Snir, University of Haifa, Israel
Satoru Miyano, University of Tokyo, Japan
Shaojie Zhang, University of Central Florida, USA
Shaoliang Peng (Co-Chair), National University of Defense and Technology, China
Shihua Zhang, Academy of Mathematics and Systems Science, CAS, China
Shuai Cheng Li, The City University of Hong Kong, Hong Kong
Shuqin Zhang, Fudan University, China
Sun Kim, Seoul National University, South Korea
Tatsuya Akutsu, Kyoto University, Japan
Ting Chen, Tsinghua University, China
Weichuan Yu, Hong Kong University of Science and Technology, Hong Kong
Wing-Kin Sung, National University of Singapore, Singapore
Yann Ponty, CNRS, École Polytechnique and Inria Saclay, France
Yanni Sun, Michigan State University, USA
Yinglei Lai, George Washington University, USA
Yong Wang, Academy of Mathematics and Systems Science, CAS, China
Yoshihiro Yamanishi, Kyushu University, Japan
Yu Lin, Australian National University, Australia
Yu Xue, Huazhong University of Sci. and Tech., China
Yufeng Wu, University of Connecticut, USA
Zengyou He, Dalian University of Technology, China

Local Organizing Committee

Anyuan Guo, Huazhong University of Science and Technology
Binguang Ma, Huazhong Agricultural University
Dexin Kong, Huazhong Agricultural University
Guoliang Li (Co-chair), Huazhong Agricultural University
Hong-Yu Zhang (Co-chair), Huazhong Agricultural University
Juan Liu, Wuhan University
Weihua Chen, Huazhong University of Science and Technology
Xingpeng Jiang, Huazhong Normal University
Yaping Fang, Huazhong Agricultural University
Yu Xue, Huazhong University of Science and Technology

Tutorials and Poster Committee

Kang Ning (Chair), Huazhong University of Science and Technology
Qi Dai, Zhejiang Science and Technology University
Xuefeng Cui, Tsinghua University
Yusen Zhang, Shandong University at Weihai

APBC2019 Program

Day 0: Sunday, January 13, 2019		
08:20-08:40	Tutorial Registration	
08:40-10:10	Transcriptome Analysis (Part I) Cheng Li, Peking University	
10:10-10:30	Tea Break	
10:30-12:00	Transcriptome Analysis (Part II)	
12:00-14:00	Lunch	
14:00-15:30	Metagenomics (Part I) Rohan Williams National University of Singapore	Gene-Regulatory Network (Part I) Yong Wang Institute of Applied Math., CAS.
15:30-16:00	Tea Break	
16:00-17:30	Metagenomics (Part II)	Gene-Regulatory Network (Part II)
16:00-18:30	**Conference Registration** (IEC Lobby)	

† International Exchange Center Lobby

Day 1: Monday, January 14, 2019		
08:00-09:00	Registration (IEC Lobby)	
09:00-09:15	Welcome Speech (Main Hall)	
09:15-10:15 Plenary Session	Ming Li Discovering Neoantigens Chair TBA	
10:15-10:50	Group photo, tea break	
Venue	Main Hall (on the 1st floor)	Meeting Room (on the 6th floor)
Parallel Session	**Genome I** Chair Xuegong Zhang	**Data I** Chair Shaoliang Peng
10:50-11:10	(**P49**) Identification of trans-eQTLs using mediation analysis with multiple mediators	(**P54**) Improving the sensitivity of detecting long read overlaps using grouped short k-mer matches
11:10-11:30	(**Highlight 193**) Machine Learning Algorithms for Modeling 3D Chromosome Structures	(**P30**) Signal enrichment of metagenome sequencing reads using topological data analysis
11:30-11:50	(**P141**) A secure SNP panel scheme using homomorphically encrypted K-mers without SNP calling on the user side	(**P46**) Automatic localization and identification of Mitochondria in cellular electron Cryo-Tomography using Faster-RCNN
12:00-14:00	Lunch	
Parallel Session	**Proteins I** Chair Shuaicheng Li, Jianyang Zeng	**Genome II** Chair Laxmi Parida
14:00-14:20	(**P145**) De Novo glycan structural identification from mass spectra using tree merging strategy	(**Highlight 181**) Detection and analysis of ancient segmental duplications in mammalian genomes
14:20-14:40	(**P134**) Prediction of FMN Binding Sites in Electron Transport Chain based on 2-D CNN and PSSM Profiles	(**P117**) Sorting Signed Permutations by Inverse Tandem Duplication Random Losses
14:40-15:00	(**Highlight 211**) AuTom-dualx: a toolkit for fully automatic fiducial marker-based alignment of dual-axis tilt series with simultaneous reconstruction	(**P53**) A distance-type measure approach to the analysis of copy number variation in DNA sequencing data
15:00-15:20	(**P140**) Constructing optimal energy functions for protein structure prediction using reverse Monte Carlo sampling	(**P113**) Branching out to speciation with a Birth-and-Death model of fractionation: the *Malvaceae*
15:20-15:40	Tea break	
Parallel Session	**Systems I** Chair Kang Ning, Cheng Li	**Genes and RNAs I** Chair Yanni Sun, Yu Lin
15:40-16:00	(**P56**) Predicting drug-target interaction network using deep learning model	(**P149**) Detecting virus-specific effects on post infection temporal gene expression
16:00-16:20	(**P79**) Identifying mutated driver pathways in cancer by integrating multi-omics data	(**P9**) RCPred: RNA Complex Prediction as a constrained maximum weight clique problem
16:20-16:40	(**P18**) Modelling the role of dual specificity phosphatases in Herceptin resistant breast cancer cell lines	(**P25**) ReactIDR: Evaluation of the statistical reproducibility of high-throughput structural analyses for a robust RNA reactivity classification
16:40-17:00	(**P41**) Prediction of drug-disease associations based on ensemble meta paths and singular value decomposition	(**Highlight 191**) The functional study of non-coding elements by integrating of 3D and 1D genome information
17:00-18:00	**Poster Session** (posters with odd ID)	

† International Exchange Center Lobby

Day 2: Tuesday, January 15, 2019		
09:00-10:00 Plenary Session Main Hall	Fengzhu Sun Statistical and Computational Approaches for the Identification of Novel Viruses and Virus-host Interactions Chair TBA	
10:00-10:30	Tea break	
Venue	Main Hall (on the 1st floor)	Meeting Room (on the 6th floor)
Parallel Session	**Genome III** Chair Jie Zheng	**Data II** Chair Guoliang Li
10:30-10:50	(**P126**) Meta-Network: Optimized species-species network analysis for microbial communities	(**P108**) SplicedFamAlign: CDS-to-gene spliced alignment and identification of transcript orthology groups
10:50-11:10	(**P89**) Towards optimal decomposition of Boolean networks	(**P124**) Large-scale 3D chromatin re-construction from chromosomal contacts
11:10-11:30	(**P31**) Fusing gene expression and transitive protein-protein interaction for gene regulatory networks	(**P86**) Estimating the total genome length of a metagenomic sample using K-mers
11:30-11:50	(**P100**) Anti-TNF-α treatment-related path- ways and biomarkers revealed by trans-criptome analysis in Chinese Psoriasis patients	(**P67**) Ultrafast clustering of single-cell flow cytometry data using FlowGrid
12:00-14:00	Lunch	
Parallel Session	**Data III** Chair Yu Xue	**Genome IV** Chair Yufeng Wu
14:00-14:20	(**P5**) SCOUT: A new algorithm for the inference of pseudo-time trajectory using single-cell data	(**P87**) GPU accelerated sequence alignment with trace back for GATK Haplotype Caller
14:20-14:40	(**P80**) ENIGMA: An enterotype-like unigram mixture model for microbial association analysis	(**P20**) The unconstrained diameters of the duplication-loss cost and the loss cost
14:40-15:00	(**P109**) Extending liquid association to explore mediated co-varying dynamics in marine microbial community	(**P21**) Reconciliation reconsidered: In search of a most representative reconciliation in the Duplication-Transfer-Loss model
15:00-15:20	(**P150**) Microbiota in the apical root canal system of tooth with apical periodontitis	(**P104**) Multiple optimal reconciliations under the duplication-loss-coalescence model
15:20-15:40	Tea break	
Parallel Session	**Proteins II** Chair Dongbu Bu, Xuefeng Cui	**Genome V** Chair Kui Lin, Rohan Williams
15:40-16:00	(**P73**) Network-based characterization of drug-protein interaction signatures with a space-efficient approach	(**P129**) Genome-wide analysis of epigenetic dynamics across human developmental stages and tissues
16:00-16:20	(**Highlight 223**) iEKPD 2.0: an update with rich annotations for eukaryotic protein kinases, protein phosphatases and proteins containing phosphoprotein-binding domains	(**P148**) Identification of Hürthle cell cancers: solving a clinical challenge with genomic sequencing and a trio of machine learning algorithms
16:20-16:40	(**P19**) Protein complex detection based on flower pollination mechanism in multi-relation reconstructed dynamic protein networks	(**Highlight 199**) Resilience of human gut microbial communities for the long stay with multiple dietary shifts
16:40-17:00	(**P38**) A class imbalance-aware Relief algorithm for the classification of tumors using microarray gene expression data	(**P29**) DeepHistone: a deep learning approach to predicting histone modifications
17:00-18:00	**Poster Session** (posters with even ID)	
18:00-20:00	**Banquet and Business Meeting**	

Day 3: Tuesday, January 16, 2019		
09:00-10:00 Plenary Session Main Hall	Ron Shamir Integrated analysis of cancer data and precision medicine Chair TBA	
10:00-10:30	Tea break	
Venue	Main Hall (on the 1st floor)	Meeting Room (on the 6th floor)
Parallel Session	**Systems II** Chair Yong Wang	**Genes and RNAs II** Chair Hong-Yu Zhang
10:30-10:50	(**P143**) Boolean network modeling of β-cell apoptosis and insulin resistance in Type 2 diabetes mellitus	(**P33**) FCMDAP: Using miRNA family and cluster information to improve the prediction accuracy of disease related miRNAs
10:50-11:10	(**P139**) Discovery of perturbation gene targets via free text metadata mining in Gene Expression Omnibus	(**P26**) GNE: A deep learning framework for gene network inference by aggregating biological information
11:10-11:30	(**P127**) Automatic hierarchy classification in venation networks using directional morphological filtering for hierarchical structure traits extraction	(**Highlight 197**) De novo haplotype reconstruction in viral quasispecies using paired-end read guided path finding
11:30-11:50	(**Highlight 179**) Comprehensive identification of peptides in tandem mass spectra using an efficient open search engine	(**P23**) DeGSM: memory scalable construction of large scale de Bruijn graph
12:00-14:00	Lunch	
Parallel Session	**Genome VI** Chair Pawel Gorecki	**Systems III** Chair Juan Liu
14:00-14:20	(**P24**) MRCNN: A deep learning model for regression of genome-wide DNA methylation	(**P146**) A fast and efficient count-based matrix factorization method for detecting cell types from single-cell RNAseq data
14:20-14:40	(**P84**) TSEE: an elastic embedding method to visualize the dynamic gene expression patterns of time series single-cell RNA sequencing data	(**P34**) Predicting disease-related phenotypes using an integrated phenotype similarity measurement based on HPO
14:40-15:00	(**P62**) A new class of super-enhancers associated with fast recovery of 3D chromatin loops	
15:00-15:30	**Award Ceremony and Closing Remarks**	

Keynote I

Discovering neoantigens

Ming Li

University of Waterloo, Canada

Cancer immunotherapy is revolutionizing cancer treatment. A key step to enable personalized cancer immunotherapy is to discover neoantigens on the surface of the cancer cells of each patient, sensitively and efficiently. We show how to combine the most advanced mass spectrometry technology and deep learning to discover neoantigens.

Keynote II

Integrated analysis of cancer data and precision medicine

Ron Shamir

Tel Aviv University, Israel

Large biological datasets are currently available, and their analysis has applications to basic science and medicine. While inquiry of each dataset separately often provides insights, integrative analysis may reveal more holistic, systems-level findings. We demonstrate the power of integrated analysis in cancer on two levels: (1) in analysis of one omic in many cancer types together, and (2) in analysis of multiple omics for the same cancer. In both levels we develop novel methods and observe a clear advantage to integration. We also describe a novel method for identifying and ranking driver genes in an individual's tumor and demonstrate its advantage over prior art.

Keynote III

Statistical and computational approaches for the identification of novel viruses and virus-host interactions

Fengzhu Sun

Fudan University, China;
U of Southern California, USA

Viruses play important roles in controlling bacterial population size, altering host metabolism, and have broader impacts on the functions of microbial communities, such as human gut, soil, and ocean microbiomes. However, the investigations of viruses and their functions were vastly underdeveloped. Metagenomic studies provide enormous resources for the identifications of novel viruses and their hosts. We recently developed a k-mer based method, VirFinder, for the identification of novel virus contigs in metagenomic samples. Applications to a liver cirrhosis metagenomic data suggest that viruses play important roles in the development of the disease. We also developed an alignment-free statistic, VirHost-Matcher, for the identification of bacterial hosts of viruses. I will present other computational tools for metagenomics including local similarity analysis (LSA) for inferring microbial associations and COCACOLA for contig binning that were recently developed from my lab.

Highlight 179

Comprehensive identification of peptides in tandem mass spectra using an efficient open search engine

Hao Chi

Institute of Comput. and Tech., CAS, China

We present a sequence-tag-based search engine, Open-pFind, to identify peptides in an ultra-large search space that includes coeluting peptides, unexpected modifications and digestions. Our method detects peptides with higher precision and speed than seven other search engines. Open-pFind identified 70–85% of the tandem mass spectra in four large-scale datasets and 14,064 proteins, each supported by at least two protein-unique peptides, in a human proteome dataset.

Source: Nature Biotech., 36, 1059 (2018)

Highlight 181

Detection and analysis of ancient segmental duplications in mammalian genomes

Lianrong Pu

Fuzhou University, China

Although segmental duplications (SDs) represent hotbeds for genomic rearrangements and emergence of new genes, there are still no easy-to-use tools for identifying SDs. Moreover, while most previous studies focused on recently emerged SDs, detection of ancient SDs remains an open problem. We developed an SDquest algorithm for SD finding and applied it to analyzing SDs in human, gorilla, and mouse genomes. Our results demonstrate that previous studies missed many SDs in these genomes and show that SDs account for at lease 6.05% of the human genome (version hg19), a 17% increase as compared to the previous estimate. Moreover, SDquest classified 6.42% of the latest GRCh38 version of the human genome as SDs, a large increase as compared to previous studies. We thus propose to re-evaluate evolution of SDs based on their accurate representation across multiple genomes. Toward this goal, we analyzed the complex mosaic structure of SDs and decomposed mosaic SDs into elementary SDs, a prerequisite for follow-up evolutionary analysis. We also introduced the concept of the breakpoint graph of mosaic SDs that revealed SD hotspots and suggested that some SDs may have originated from circular extrachromosomal DNA (ecDNA), not unlike ecDNA that contributes to accelerated evolution in cancer

Source: Genome Research, 28, 901-909 (2018)

Highlight 191

The functional study of non-coding elements by integrating of 3D and 1D genome information

Yaping Fang

Huazhong Agricultural Univ., China

Non-coding elements are important to the study of biology, over 98% of the human genome are non-coding DNA including intros, noncoding RNA cis- and trans-regulatory elements, pseudogenes, repeats and transposons, et al. In history these non-coding DNA were termed as 'junk-DNA' and now they are known as genome's 'dark matter'. ENCODE project reported that 76% of the human genome's noncoding DNA elements were transcribed such as transcribed enhancer and most of the transcribed noncoding DNA elements were accessible to the regulatory proteins such as transcription factor and histone which can regulate the gene expressions and 3D genome architecture. However, the function study of non-coding elements is formidable as they cannot be translated into protein and are dynamic and interacted with each other. In our recent works, we have developed methods to study the non-coding elements (1) We developed a hidden Markov model model and software to study the cross-talks and dynamics of TFs (2) We developed a random forest based model to predict enhancers and a deep learning model to predict the activities of non-coding elements (3) We developed protocols to study the functions of lncRNAs and small RNAs.

Source: Bioinformatics, doi:10.1093 (2018)

Highlight 193

Machine learning algorithms for modeling 3D chromosome structures

Jianyang Michael Zeng

Tsinghua University, China

Decoding the spatial organizations of chromosomes has crucial implications for studying eukaryotic gene regulation. Recently, Chromosomal conformation capture based technologies, such as Hi-C, have been widely used to uncover the interaction frequencies of genomic loci in high-throughput and genome-wide manner and provide new insights into the folding of three-dimensional (3D) genome structure. In this project, we develop a novel manifold learning based framework, called GEM (Genomic organization reconstructor based on conformational Energy and Manifold learning), to reconstruct the three dimensional organizations of chromosomes by integrating Hi-C data with biophysical feasibility. Unlike previous methods, which explicitly assume specific relationships between Hi-C interaction frequencies and spatial distances, our model directly embeds the neighboring affinities from Hi-C space into 3D Euclidean space. Extensive validations demonstrated that GEM not only greatly outperformed other state-of-art modeling methods but also provided a physically and physiologically valid 3D representations of the organizations of chromosomes. Furthermore, we for the first time apply the modeled chromatin structures to recover long-range genomic interactions missing from original Hi-C data.

Highlight 197

De novo haplotype reconstruction in viral quasispecies using paired-end read guided path finding

Yanni Sun

City U of Hong Kong, China

A patient with chronic virus infection can generate billions of new viral particles each day. These viral particles may contain an unknown number of viral genomes, each of which can have different biological properties such as anti-viral drug resistance genes. Thus, the viral phenotype inference must be derived at strain level. The most powerful tool to study viral community is viral metagenomic sequencing. Despite the availability of a plethora of metagenomic analysis tools, there lack strain-level inference of viruses from viral metagenomic data. Assembling short reads into genome-scale strains is particularly challenging because of scarce reference viral genomes, the sequencing errors, high similarities between related strains, and heterogeneous sequencing coverage along each strain. In this recently published work, we developed a de novo strain assembly tool named PEHaplo. With a careful analysis of the longest common substrings (LCSs) between viral strains, our methods carefully utilize the properties of paired-end reads for viral haplotypes reconstruction. If was applied on both simulated and real quasispecies data, and the results were benchmarked against several recently published de novo haplotype reconstruction tools. The results showed that our tool is fast, and be able to produce significantly longer and high quality contigs than existing de novo assembly tools for virus quasispecies.

Source: Bioinformatics, 34, 2927-2935 (2018)

Highlight 199

Resilience of human gut microbial communities for the long stay with multiple dietary shifts

Ning Kang

Huazhong U of Sci. and Tech., China

Human gut microbial communities are important in diagnosis, treatment, and prevention of diseases, and the environmental changes of hosts, especially dietary shifts, usually leads to substantial dynamics of microbial communities. The effects of short-term environmental changes on human gut microbiota have received extensive studies. However, the dynamic patterns of microbial communities across a long time with multiple dietary shifts remain unclear. Here, we tracked a volunteer team who traveled from Beijing to Trinidad and Tobago (TAT), stayed there from December 2015 to June 2016 and then back. During the entire period of more than eight months, we recorded their dietary details and collected their faecal samples based on high density longitudinal sampling strategy, and obtained 287 faecal samples from 41 individuals and harvested 3.3 TB sequencing data. Analysis of the microbiome data reveals a highly plastic pattern (resilience) and the enterotypes could be recovered after the long stay, i.e., when the volunteer team entered TAT, the patterns (and enterotypes) of their gut microbial communities evolved quickly to that of the TAT natives; yet even after the long stay of six months, their gut microbial communities could quickly revert back to their original patterns (and enterotypes) when they returned to Beijing. And this bi-directional plastic pattern could be measured by our adaptation index quantitatively. We also found that the abundances of glycoside hydrolases were suppressed and those of glycosyltransferases and carbohydrate esterases were amplified during the long stay.

Source: Gut, doi: 10.1136 (2018)

Highlight 211

AuTom-dualx: a toolkit for fully automatic fiducial marker-based alignment of dual-axis tilt series with simultaneous reconstruction

Fa Zhang

Institute of Comput. Tech., CAS, China

Dual-axis electron tomography is an important 3D macromolecular structure reconstruction technology, which can reduce artifacts and suppress the effect of missing wedge. However, the fully automatic data process for dual-axis electron tomography still remains a challenge due to three difficulties: (i) how to track the mass of fiducial markers automatically; (ii) how to integrate the information from the two different tilt series; and (iii) how to cope with the inconsistency between the two different tilt series.

Here we develop a toolkit for fully automatic alignment of dual-axis electron tomography, with a simultaneous reconstruction procedure. The proposed toolkit and its workflow carries out the following solutions: (i) fully automatic detection and tracking of fiducial markers under large-field datasets; (ii) automatic combination of two different tilt series and global calibration of projection parameters; and (iii) inconsistency correction based on distortion correction parameters and the consequently simultaneous reconstruction. With all of these features, the presented toolkit can achieve accurate alignment and reconstruction simultaneously and conveniently under a single global coordinate system.

Source: Bioinformatics, doi: 10.1093 (2018)

Highlight 223

iEKPD 2.0: an update with rich annotations for eukaryotic protein kinases, protein phosphatases and proteins containing phosphoprotein-binding domains

Wankun Deng

Huazhong U of Sci. and Tech., China

We described the updated database iEKPD 2.0 (http://iekpd.biocuckoo.org) for eukaryotic protein kinases (PKs), protein phosphatases (PPs) and proteins containing phosphoprotein-binding domains (PPBDs), which are key molecules responsible for phosphorylation-dependent signalling networks and participate in the regulation of almost all biological processes and pathways. In total, iEKPD 2.0 contained 197,348 phosphorylation regulators, including 109,912 PKs, 23,294 PPs and 68,748 PPBD-containing proteins in 164 eukaryotic species. In particular, we provided rich annotations for the regulators of eight model organisms, especially humans, by compiling and integrating the knowledge from 100 widely used public databases that cover 13 aspects, including cancer mutations, genetic variations, disease-associated information, mRNA expression, DNA & RNA elements, DNA methylation, molecular interactions, drug-target relations, protein 3D structures, post-translational modifications, protein expressions/proteomics, subcellular localizations and protein functional annotations. Compared with our previously developed EKPD 1.0 (~0.5 GB), iEKPD 2.0 contains ~99.8 GB of data with an approximately 200-fold increase in data volume. We anticipate that iEKPD 2.0 represents a useful resource for further study of phosphorylation regulators.

Source: Nucleic Acids Research, gky1063 (2018)

Paper 5

SCOUT: a new algorithm for the inference of pseudo-time trajectory using single-cell data

Jiangyong Wei[1]
Tianshou Zhou[1]
Xinan Zhang[2]
Tianhai Tian[3]

[1]Sun Yat-sen University, China
[2]Central China Normal Univ., China
[3]Monash University, Australia

Single cell technology is a powerful tool to reveal intercellular heterogeneity and discover cellular developmental processes. When analyzing the complexity of cellular dynamics and variability, it is important to construct a pseudo-time trajectory using single-cell expression data to reflect the process of cellular development. Although a number of computational and statistical methods have been developed recently for single-cell analysis, more effective and efficient methods are still strongly needed. In this work we propose a new method named SCOUT for the inference of single-cell pseudo-time ordering with bifurcation trajectories. We first propose to use the fixed-radius near neighbors algorithms based on cell densities to find landmarks to represent the cell states, and employ the minimum spanning tree (MST) to determine the developmental branches. We then propose to use the projection of Apollonian circle or a weighted distance to determine the pseudo-time trajectories of single cells. The proposed algorithm is applied to one synthetic and two realistic single-cell datasets (including single-branching and multi-branching trajectories) and the cellular developmental dynamics is recovered successfully. Compared with other popular methods, numerical results show that our proposed method is able to generate more robust and accurate pseudo-time trajectories. The code of the method is implemented in Python and available at https://github.com/statway/SCOUT.

Paper 9

RCPred: RNA complex prediction as a constrained maximum weight clique problem

Audrey Legendre
Eric Angel
Fariza Tahi

Univ Evry, Universite Paris-Saclay, France

RNAs can interact and form complexes, which have various biological roles. The secondary structure prediction of those complexes is a first step towards the identification of their 3D structure. We propose an original approach that takes advantage of the high number of RNA secondary structure and RNA-RNA interaction prediction tools. We formulate the problem of RNA complex prediction as the determination of the best combination (according to the free energy) of predicted RNA secondary structures and RNA-RNA interactions.

We model those predicted structures and interactions as a graph in order to have a combinatorial optimization problem that is a constrained maximum weight clique problem. We propose an heuristic based on Breakout Local Search to solve this problem and a tool, called RCPred, that returns several solutions, including motifs like internal and external pseudoknots. On a large number of complexes, RCPred gives competitive results compared to the methods of the state of the art.

We propose in this paper a method called RCPred for the prediction of several secondary structures of RNA complexes, including internal and external pseudoknots. As further works we will propose an improved computation of the global energy and the insertion of 3D motifs in the RNA complexes.

Paper 18

Modelling the role of dual specificity phosphatases in Herceptin resistant breast cancer cell lines

Petronela Buiga
Ari Elson
Lydia Tabernero
Jean-Marc Schwartz

University of Manchester, UK

Breast cancer remains the most lethal type of cancer for women. A significant proportion of breast cancer cases are characterised by overexpression of the human epidermal growth factor receptor 2 protein (HER2). These cancers are commonly treated by Herceptin (Trastuzumab), but resistance to drug treatment frequently develops in tumour cells. Dual-specificity phosphatases (DUSPs) are thought to play a role in the mechanism of resistance, since some of them were reported to be overexpressed in tumours resistant to Herceptin.

We used a systems biology approach to investigate how DUSP overexpression could favour cell proliferation and to predict how this mechanism could be reversed by targeted inhibition of selected DUSPs. We measured the expression of 20 DUSP genes in two breast cancer cell lines following long-term (6 months) exposure to Herceptin, after confirming that these cells had become resistant to the drug. We constructed several Boolean models including specific substrates of each DUSP, and showed that our models correctly account for resistance when overexpressed DUSPs were kept activated. We then simulated inhibition of both individual and combinations of DUSPs, and determined conditions under which the resistance could be reversed.

Paper 19

Protein complex detection based on flower pollination mechanism in multi-relation reconstructed dynamic protein networks

Xiujuan Lei[1]
Ming Fang[1]
Ling Guo[1]
Fang-Xiang Wu[2]

[1]*Shaanxi Normal Univ., China*
[2]*Univ. of Saskatchewan, Canada*

Detecting protein complex in protein-protein interaction (PPI) networks plays a significant part in bioinformatics field. It enables us to obtain the better understanding for the structures and characteristics of biological systems.

In this study, we present a novel algorithm, named Improved Flower Pollination Algorithm (IFPA), to identify protein complexes in multi-relation reconstructed dynamic PPI networks. Specifically, we first introduce a concept called co-essentiality, which considers the protein essentiality to search essential interactions, Then, we devise the multi-relation reconstructed dynamic PPI networks (MRDPNs) and discover the potential cores of protein complexes in MRDPNs. Finally, an IFPA algorithm is put forward based on the flower pollination mechanism to generate protein complexes by simulating the process of pollen find the optimal pollination plants, namely, attach the peripheries to the corresponding cores.

The experimental results on three different datasets (DIP, MIPS and Krogan) show that our IFPA algorithm is more superior to some representative methods in the prediction of protein complexes.

Our proposed IFPA algorithm is powerful in protein complex detection by building multi-relation reconstructed dynamic protein networks and using improved flower pollination algorithm. The experimental results indicate that our IFPA algorithm can obtain better performance than other methods.

Paper 20

The unconstrained diameters of the duplication-loss cost and the loss cost

Pawel Gorecki
University of Warsaw, Poland
Oliver Eulenstein
Iowa State University, USA
Jerzy Tiuryn
University of Warsaw, Poland

Tree reconciliation costs are a popular choice to account for the discordance between the evolutionary history of a gene family (i.e., a gene tree), and the species tree through which this family has evolved. This discordance is accounted for by the minimum number of postulated evolutionary events necessary for reconciling the two trees. Such events include gene duplication, loss, and deep coalescence, and are used to define different types of tree reconciliation costs. For example, the duplication-loss cost for a gene tree and species tree accounts for the minimum number of gene duplications and losses necessary to reconcile these trees. Fundamental to the understanding of how gene trees and species trees relate to each other are the diameters of tree reconciliation costs. While such diameters have been well-researched, still absent from these studies are the unconstrained diameters for two of the classic tree reconciliation costs, namely the duplication-loss cost and the loss cost. Here, we show the essential mathematical properties of these diameters and provide efficient solutions for computing them. Finally, we analyze the distributions of these diameters using simulated datasets.

Paper 21

Reconciliation reconsidered: in search of a most representative reconciliation in the duplication-transfer-loss model

Melissa Grueter
Kalani Duran, Ramya Ramalingam
Ran Libeskind-Hadas
Harvey Mudd College, USA

Maximum parsimony reconciliation is a fundamental technique for studying the evolutionary histories of pairs of entities such as genes and species, parasites and hosts, and species and their biogeographical habitats. In these contexts, reconciliation is generally performed using the duplication-transfer-loss (DTL) model in a maximum parsimony framework. While efficient maximum parsimony reconciliation algorithms are known for the DTL model, the number of such reconciliations can grow exponentially with the sizes of the two phylogenetic trees. Choosing a maximum parsimony reconciliation arbitrarily may lead to conclusions that are not supported, and even contradicted, by other equally optimal reconciliations. This paper addresses the fundamental problem of how well a single reconciliation can represent the entire space of optimal reconciliations.

Paper 23

DeGSM: memory scalable construction of large scale de Bruijn graph

Hongzhe Guo
Yilei Fu
Yan Gao
Junyi Li
Yadong Wang
Bo Liu

Harbin Institute of Tech., China;

The de Bruijn graph, a fundamental data structure to represent and organize genome sequence, plays important roles in various kinds of sequence analysis tasks such as de novo assembly, high-throughput sequencing (HTS) read alignment, pan-genome analysis, metagenomics analysis, HTS read error correction, etc. We propose a lightweight parallel de Bruijn graph construction approach: de Bruijn Graph Constructor in Scalable Memory (deGSM). The main idea of deGSM is to efficiently construct the Burrows-Wheeler Transformation (BWT) of the unipaths of the de Bruijn graph in constant RAM space and transform the BWT into the original unitigs. It is mainly implemented by a fast parallel external sorting of k-mers, which uses a novel organization of the k-mers that requires only a part of k-mers kept in RAM. The experimental results demonstrate that, just with a commonly available machine, deGSM is able to handle very large genome sequence(s), e.g., the contigs (305 Gbp) and scaffolds (1.1 Tbp) recorded in GenBank database and Picea abies HTS dataset (9.7 Tbp). Moreover, deGSM also has faster or comparable construction speed compared with state-of-the-art approaches. With its high scalability and efficiency, deGSM has enormous potential in many large scale genomics studies. The deGSM is publicly available at: https://github.com/hitbc/deGSM.

Paper 24

MRCNN: A deep learning model for regression of genome-wide DNA methylation

Qi Tian
Jianxiao Zou
Shicai Fan
Jianxiong Tang
Yuan Fang
Zhongli Yu

University of Electronic Science and Tech. of China, China.

Determination of genome-wide DNA methylation is significant for both basic research and drug development. As a key epigenetic modification, this biochemical process can modulate gene expression to influence the cell differentiation which can possibly lead to cancer. Due to the involuted biochemical mechanism of DNA methylation, obtaining a precise prediction is a considerably tough challenge. Existing approaches have yielded good predictions, but the methods either need to combine plenty of features and prerequisites or deal with only hypermethylation and hypomethylation.

In this paper, we propose a deep learning method for prediction of the genome-wide DNA methylation, in which the Methylation Regression is implemented by Convolutional Neural Networks (MRCNN). Through minimizing the continuous loss function, experiments show that our model is convergent and more precise than the state-of-art method (DeepCpG) according to results of the evaluation. MRCNN also achieves the discovery of de novo motifs by analysis of features from the training process.

Genome-wide DNA methylation could be evaluated based on the corresponding local DNA sequences of target CpG loci. With the autonomous learning pattern of deep learning, MRCNN enables accurate predictions of genome-wide DNA methylation status without predefined features and discovers some de novo methylation-related motifs that match known motifs by extracting sequence patterns.

Paper 25

ReactIDR: evaluation of the statistical reproducibility of high-throughput structural analyses for a robust RNA reactivity classification

Risa Kawaguchi[1]
Hisanori Kiryu[2]
Junichi Iwakiri[2]
Jun Sese[1,3,4]

[1]National Inst. of Adv. Industrial Sci. and Tech.,
[2]University of Tokyo,
[3]AIST- Tokyo Tech Real World Big-Data Comput. Open Innovation Lab.,
[4]Humanome Lab Inc., Japan

Here, we introduced a statistical framework, reactIDR, which can be applied to the experimental data obtained using multiple HTS methodologies. Using this approach, nucleotides are classified into three structural categories, loop, stem/background, and unmapped. reactIDR uses the irreproducible discovery rate (IDR) with a hidden Markov model to discriminate between the true and spurious signals obtained in the replicated HTS experiments accurately, and it is able to incorporate an expectation-maximization algorithm and supervised learning for efficient parameter optimization. The results of our analyses of the real-life HTS data showed that reactIDR had the highest accuracy in the classification of ribosomal RNA stem/loop structures when using both individual and integrated HTS datasets, and its results corresponded the best to the three-dimensional structures.

We have developed a novel software, reactIDR, for the prediction of stem/loop regions from the HTS analysis datasets. For the rRNA structure analyses, reactIDR was shown to have robust accuracy across different datasets by using the reproducibility criterion, suggesting its potential for increasing the value of existing HTS datasets. reactIDR is publicly available at https://github.com/carushi/reactIDR.

Paper 26

GNE: a deep learning framework for gene network inference by aggregating biological information

Kishan K C
Rui Li
Feng Cui
Qi Yu
Anne Haake

Rochester Inst. of Tech., USA

The topological landscape of gene interaction networks provides a rich source of information for inferring functional patterns of genes or proteins. However, it is still a challenging task to aggregate heterogeneous biological information such as gene expression and gene interactions to achieve more accurate inference for prediction and discovery of new gene interactions. In particular, how to generate a unified vector representation to integrate diverse input data is a key challenge addressed here.

We propose a scalable and robust deep learning framework to learn embedded representations to unify known gene interactions and gene expression for gene interaction predictions. These low- dimensional embeddings derive deeper insights into the structure of rapidly accumulating and diverse gene interaction networks and greatly simplify downstream modeling. We compare the predictive power of our deep embeddings to the strong baselines. The results suggest that our deep embeddings achieve significantly more accurate predictions. Moreover, a set of novel gene interaction predictions are validated by up-to-date literature-based database entries.

The proposed model demonstrates the importance of integrating heterogeneous information about genes for gene network inference. GNE is freely available under the GNU General Public License and can be downloaded from GitHub (https://github.com/kckishan/GNE).

Paper 29

DeepHistone: a deep learning approach to predicting histone modifications

Qijin Yin
Mengmeng Wu
Hairong Lv
Rui Jiang

Tsinghua University, China

We proposed a deep learning framework to integrate sequence information and chromatin accessibility data for the accurate prediction of modification sites specific to different histone markers. Our method, named DeepHistone, outperformed several baseline methods in a series of comprehensive validation experiments, not only within an epigenome but also across epigenomes. Besides, sequence signatures automatically extracted by our method was consistent with known transcription factor binding sites, thereby giving insights into regulatory signatures of histone modifications. As an application, our method was shown to be able to distinguish functional single nucleotide polymorphisms from their nearby genetic variants, thereby having the potential to be used for exploring functional implications of putative disease-associated genetic variants.

DeepHistone demonstrated the possibility of using a deep learning framework to integrate DNA sequence and experimental data for predicting epigenomic signals. With the state-of-the-art performance, DeepHistone was expected to shed light on a variety of epigenomic studies. DeepHistone is freely available in https://github.com/QijinYin/ DeepHistone.

Paper 30

Signal enrichment of metagenome sequencing reads using topological data analysis

Aldo Guzmán-Sáenz[1]
Niina Haiminen[1]
Saugata Basu[2]
Laxmi Parida[1]

[1]Purdue University, USA;
[2]IBM T. J. Watson Res. Center, USA

A metagenome is a collection of genomes, usually in a micro-environment, and sequencing a metagenomic sample en masse is a powerful means for investigating the community of the constituent microorganisms. One of the challenges is in distinguishing between similar organisms due to rampant multiple possible assignments of sequencing reads, resulting in false positive identifications. We map the problem to a topological data analysis (TDA) framework that extracts information from the geometric structure of data. Here the structure is defined by multi-way relationships between the sequencing reads using a reference database.

Based primarily on the patterns of co-mapping of the reads to multiple organisms in the reference database, we use two models: one a subcomplex of a Barycentric subdivision complex and the other a Cech complex. The Barycentric subcomplex allows a natural mapping of the reads along with their coverage of organisms while the Cech complex takes simply the number of reads into account to map the problem to homology computation. Using simulated genome mixtures we show not just enrichment of signal but also microbe identification with strain-level resolution.

In particular, in the most refractory of cases where alternative algorithms that exploit unique reads (i.e., mapped to unique organisms) fail, we show that the TDA approach continues to show consistent performance. The Cech model that uses less information is equally effective, suggesting that even partial information when augmented with the appropriate structure is quite powerful.

Paper 31

Fusing gene expression and transitive protein-protein interaction for gene regulatory networks

Wenting Liu,

Univ. of California at Los Angles, USA

Jagath Rajapakse

Nanyang Tech. Univ., Singapore

We use a Gaussian Mixture Model (GMM) to soft-cluster GE data, allowing overlapping cluster memberships. Next, a heuristic method is proposed to extend sparse PPIN by incorporating transitive linkages. We then propose a novel way to score extended protein interactions by combining topological properties of PPIN and correlations of GE. Following this, GE data and extended PPIN are fused using a Gaussian Hidden Markov Model (GHMM) in order to identify gene regulatory pathways and refine interaction scores that are then used to constrain the GRN structure. We employ a Bayesian Gaussian Mixture (BGM) model to refine the GRN derived from GE data by using the structural priors derived from GHMM. Experiments on real yeast regulatory networks demonstrate both the feasibility of the extended PPIN in predicting transitive protein interactions and its effectiveness on improving the coverage and accuracy the proposed method of fusing PPIN and GE to build GRN.

The GE and PPIN fusion model outperforms both the state-of-the-art single data source models (CLR, GENIE3, TIGRESS) as well as existing fusion models under various constraints.

Paper 33

FCMDAP: using miRNA family and cluster information to improve the prediction accuracy of disease related miRNAs

Xiaoying Li[1]
Yaping Lin[1]
Changlong Gu[1]
Jialiang Yang[2]

[1]*Hunan University, China*
[2]*Icahn of Med. at Mount Sinai, USA*

We develop a novel method using multiple types of data to calculate miRNA and disease similarity based on mutual information, and add miRNA family and cluster information to predict human disease-related miRNAs (FCMDAP). This method not only depends on known miRNA-diseases associations but also accurately measures miRNA and disease similarity and resolves the problem of overestimation. FCMDAP uses the k most similar neighbor recommendation algorithm to predict the association score between miRNA and disease. Information about miRNA cluster is also used to improve prediction accuracy.

FCMDAP achieves an average AUC of 0.9165 based on leave-one-out cross validation. Results confirm the 100%, 98% and 96% of the top 50 predicted miRNAs reported in case studies on colorectal, lung, and pancreatic neoplasms. FCMDAP also exhibits satisfactory performance in predicting diseases without any related miRNAs and miRNAs without any related diseases.

In this study, we present a computational method FCMDAP to improve the prediction accuracy of disease related miRNAs. FCMDAP could be an effective tool for further biological experiments.

Paper 34

Predicting disease-related phenotypes using an integrated phenotype similarity measurement based on HPO

Hansheng Xue
Jiajie Peng
Xuequn Shang

Northwestern Polytech. U., China

Improving efficiency of disease diagnosis based on phenotype ontology is a critical yet challenging research area. Recently, Human Phenotype Ontology (HPO)-based semantic similarity has been affectively and widely used to identify causative genes and diseases. However, current phenotype similarity measurements just consider the annotations and hierarchy structure of HPO, neglecting the definition description of phenotype terms.

In this paper, we propose a novel phenotype similarity measurement, termed as DisPheno, which adequately incorporates the definition of phenotype terms in addition to HPO structure and annotations to measure the similarity between phenotype terms. DisPheno also integrates phenotype term associations into phenotype-set similarity measurement using gene and disease annotations of phenotype terms.

Compared with five existing state-of-the-art methods, DisPheno shows great performance in HPO-based phenotype semantic similarity measurement and improves the efficiency of disease identification, especially on noisy patients dataset.

Paper 38

A class imbalance-aware Relief algorithm for the classication of tumors using microarray gene expression data

Yuanyu He
Junhai Zhou
Yaping Lin
Tuanfei Zhu

Hunan University, China

DNA microarray data has been widely used in cancer research due to the significant advantage helped to successfully distinguish between tumor classes. However, typical gene expression data usually presents a high-dimensional imbalanced characteristic, which poses severe challenge for traditional machine learning methods to construct a robust classifier performing well on both the minority and majority classes. As one of the most successful feature weighting techniques, Relief is considered to particularly suit to handle high-dimensional problems. Unfortunately, almost all Relief-based methods have not taken the class imbalance distribution into account. This study identifies that existing Relief-based algorithms may underestimate the features with the discernibility ability of minority classes, and ignore the distribution characteristic of minority class samples. As a result, an additional bias towards being classified into the majority classes can be introduced. To this end, a new method, named imRelief, is proposed for efficiently handling high-dimensional imbalanced gene expression data. imRelief can correct the bias towards to the majority classes, and consider the scattered distributional characteristic of minority class samples in the process of estimating feature weights. This way, imRelief has the ability to reward the features which perform well at separating the minority classes from other classes. Experiments on four microarray gene expression data sets demonstrate the effectiveness of imRelief in both feature weighting and feature subset selection applications.

Paper 41

Prediction of drug-disease associations based on ensemble meta paths and singular value decomposition

Guangsheng Wu
Wuhan University, China
Juan Liu
Wuhan University, China
Xiang Yue
Ohio State University, USA

In the proposed EMP-SVD (Ensemble Meta Paths and Singular Value Decomposition), we introduce five meta paths corresponding to different kinds of interaction data, and for each meta path we generate a commuting matrix. Every matrix is factorized into two low rank matrices by SVD which are used for the latent features of drugs and diseases respectively. The features are combined to represent drug-disease pairs. We build a base classifier via Random Forest for each meta path and five base classifiers are combined as the final ensemble classifier. In order to train out a more reliable prediction model, we select more likely negative ones from unlabeled samples under the assumption that non-associated drug and disease pair have no common interacted proteins. The experiments have shown that the proposed EMP-SVD method outperforms several state-of-the-art approaches. Case studies by literature investigation have found that the proposed EMP-SVD can mine out many drug-disease associations, which implies the practicality of EMP-SVD.

The proposed EMP-SVD can integrate the interaction data among drugs, proteins and diseases, and predict the drug-disease associations without the need of similarity information. At the same time, the strategy of selecting more reliable negative samples will benefit the prediction.

Paper 46

Automatic localization and identification of mitochondria in cellular electron cryotomography using faster-RCNN

Ran Li[1]
Xiangrui Zeng[2],
Stephanie Siegmund[3]
Ruogu Lin[2]
Bo Zhou[2]
Chang Liu[2]
Kaiwen Wang[2]
Rui Jiang[1]
Zachary Freyberg[4]
Hairong Lv[1],
Min Xu[2]

[1]*Tsinghua University, China*
[2]*Carnegie Mellon University, USA*
[3]*Columbia University, USA*
[4]*University of Pittsburgh, USA*

Our experimental results were validated using in situ cyro-ET-imaged mitochondria data. Our experimental results show that our algorithm can accurately localize and identify important cellular structures on both the 2D tilt images and the reconstructed 2D slices of cryo-ET. When ran on the mitochondria cryo-ET dataset, our algorithm achieved Average Precision >0.95. Moreover, our study demonstrated that our customized pre-processing steps can further improve the robustness of our model performance.

In this paper, we proposed an automatic Cryo-ET image analysis algorithm for localization and identification of different structure of interest in cells, which is the first Faster-RCNN based method for localizing an cellular organelle in Cryo-ET images and demonstrated the high accuracy and robustness of detection and classification tasks of intracellular mitochondria. Furthermore, our approach can be easily applied to detection tasks of other cellular structures as well.

Paper 49

Identification of trans-eQTLs using mediation analysis with multiple mediators

Nayang Shan
Tsinghua University, *China*

Zuoheng Wang
Yale University, USA

Lin Hou
Tsinghua University, *China*

We proposed two hypothesis tests for testing the total mediation effect (TME) and the component-wise mediation effects (CME), respectively. We demonstrated in simulation studies that the type I error rates were controlled in both tests despite model misspecification. The TME test was more powerful than the CME test when the two mediation effects are in the same direction, while the CME test was more powerful than the TME test when the two mediation effects are in opposite direction. Multiple mediator analysis had increased power to detect mediated trans-eQTLs, especially in large samples. In the HapMap3 data, we identified 11 mediated trans-eQTLs that were not detected by the single mediator analysis in the combined samples of African populations. Moreover, the mediated trans-eQTLs in the HapMap3 samples are more likely to be trait-associated SNPs. In terms of computation, although there is no limit in the number of mediators in our model, analysis takes more time when adding additional mediators. In the analysis of the HapMap3 samples, we included at most 5 cis-gene mediators. Majority of the trios we considered have one or two mediators.

Trans-eQTLs are more likely to associate with multiple cis-genes than randomly selected SNPs. Mediation analysis with multiple mediators improves power of identification of mediated trans-eQTLs, especially in large samples.

Paper 53

A distance-type measure approach to the analysis of copy number variation in DNA sequencing data

Bipasa Biswas
FDA/CDRH/OSB-DBS, USA

Yinglei Lai
George Washington Univ., USA

In this study, we propose to analyze DNA-seq data based on the related distance-type measure. Distances are measured in base pairs (bps) between two adjacent alignments of short reads mapped to a reference genome. Our experimental data based simulation study confirms the advantages of distance-type measure approach in both detection power and detection accuracy. Furthermore, we propose artificial censoring for the distance data so that distances larger than a given value are considered potential outliers. Our purpose is to simplify the pre-processing of DNA-seq data. Statistically, we consider a mixture of right censored geometric distributions to model the distance data. Additionally, to reduce the GC-content bias, we extend the mixture model to a mixture of generalized linear models (GLMs). The estimation of model can be achieved by the Newton-Raphson algorithm as well as the Expectation-Maximization (E-M) algorithm. We have conducted simulations to evaluate the performance of our approach. Based on the rank based inverse normal transformation of distance data, we can obtain the related z-values for a follow-up analysis. For an illustration, an application to the DNA-seq data from a pair of normal and tumor cell lines is presented with a change-point analysis of z-values to detect DNA copy number alterations.

Our distance-type measure approach is novel. It does not require either a fixed or a sliding window procedure for generating count-type data. Its advantages have been demonstrated by our simulation studies and its practical usefulness has been illustrated by an experimental data application.

Paper 54

Improving the sensitivity of detecting long read overlaps using grouped short k-mer matches

Nan Du[1]
Jiao Chen[1]
Yanni Sun[2]

[1]Michigan State University, USA;
[2]City Univ. of Hong Kong, China

In this work, we designed and implemented an overlap detection program named GroupK, for third-generation sequencing reads based on grouped k-mer hits. While using k-mer hits for detecting reads' overlaps has been adopted by several existing programs, our method uses a group of short k-mer hits satisfying statistically derived distance constraints to increase the sensitivity of small overlap detection. Grouped k-mer hit was originally designed for homology search. We are the first to apply group hit for long read overlap detection. The experimental results of applying our pipeline to both simulated and real third-generation sequencing data showed that GroupK enables more sensitive overlap detection, especially for datasets of low sequencing coverage.

GroupK is best used for detecting small overlaps for third-generation sequencing data. It provides a useful supplementary tool to existing ones for more sensitive and accurate overlap detection. The source code is freely available at https://github.com/Strideradu/GroupK.

Paper 56

Predicting drug-target interaction network using deep learning model

Jiaying You
Robert D. McLeod
Pingzhao Hu

University of Manitoba, Canada

Traditional methods for drug discovery are time-consuming and expensive, so efforts are being made to repurpose existing drugs. To find new ways for drug repurposing, many computational approaches have been proposed to predict drug-target interactions (DTIs). However, due to the high-dimensional nature of the data sets extracted from drugs and targets, traditional machine learning approaches, such as logistic regression analysis, cannot analyze these data sets efficiently. To overcome this issue, we propose LASSO (Least absolute shrinkage and selection operator)-based regularized linear classification models and a LASSO-DNN (Deep Neural Network) model based on LASSO feature selection to predict DTIs. These methods are demonstrated for repurposing drugs for breast cancer treatment.

Experimental results showed that the LASSO-DNN over performed the SLG, LASSO, SVM and standard DNN models. In particular, the LASSO models with protein tripeptide composition (TC) features and domain features were superior to those that contained other protein information, which may imply that TC and domain information could be better representations of proteins. Furthermore, we showed that the top ranked DTIs predicted using the LASSO-DNN model can potentially be used for repurposing existing drugs for breast cancer based on risk gene information.

Paper 62

A new class of super-enhancers associated with fast recovery of 3D chromatin loops

Inkyung Jung
Hyunwoong Kim
Dongchan Yang
Jayoung Ryu

KAIST, South Korea

In this study, through a comprehensive analysis of super-enhancers in 30 human cell/tissue types, we identified a new class of super-enhancers which are constitutively active across most cell/tissue types. These 'common' super-enhancers are associated with universally highly expressed genes in contrast to the canonical definition of super-enhancers that assert cell-type specific gene regulation. In addition, the genome sequence of these super-enhancers is highly conserved by evolution and among humans, advocating their universal function in genome regulation. Integrative analysis of 3D chromatin loops demonstrates that, in comparison to the cell-type specific super-enhancers, the cell-type common super-enhancers present a striking association with rapidly recovering loops.

In this study, we propose that a new class of super-enhancers may play an important role in the early establishment of 3D chromatin structure.

Paper 67

Ultrafast clustering of single-cell flow cytometry data using FlowGrid

Xiaoxin Ye[1,2]
Joshua W K Ho[1,2,3]

[1]Victor Chang Cardiac Res. Inst.,
[2]U of New South Wales, Australia,
[3]University of Hong Kong, China

Flow cytometry is a popular technology for quantitative single-cell profiling of cell surface markers. It enables expression measurement of tens of cell surface protein markers in millions of single cells. It is a powerful tool for discovering cell sub-populations and quantifying cell population heterogeneity. Traditionally, scientists use manual gating to identify cell types, but the process is subjective and is not effective for large multidimensional data. Many clustering algorithms have been developed to analyse these data but most of them are not scalable to very large data sets with more than ten million cells.

Here, we present a new clustering algorithm that combines the advantages of density-based clustering algorithm DBSCAN with the scalability of grid-based clustering. This new clustering algorithm is implemented in python as an open source package, FlowGrid. FlowGrid is memory efficient and scales linearly with respect to the number of cells. We have evaluated the performance of FlowGrid against other state-of-the-art clustering programs and found that FlowGrid produces similar clustering results but with substantially less time. For example, FlowGrid is able to complete a clustering task on a data set of 23.6 million cells in less than 12 seconds, while other algorithms take more than 500 seconds or get into error.

FlowGrid is an ultrafast clustering algorithm for large single-cell flow cytometry data. The source code is available at https://github.com/VCCRI/FlowGrid.

Paper 73

Network-based characterization of drug-protein interaction signatures with a space-efficient approach

Yasuo Tabei[1]
Masaaki Kotera[2]
Ryusuke Sawada[3]
Yoshihiro Yamanishi[3]

[1]RIKEN,
[2]University of Tokyo,
[3]Kyushu Institute of Technology, Japan

Characterization of drug-protein interaction networks with biological features has recently become challenging in recent pharmaceutical science toward a better understanding of polypharmacology.

We present a novel method for systematic analyses of the underlying features characteristic of drug-protein interaction networks, which we call "drug-protein interaction signatures" from the integration of large-scale heterogeneous data of drugs and proteins. We develop a new efficient algorithm for extracting informative drug-protein interaction signatures from the integration of large-scale heterogeneous data of drugs and proteins, which is made possible by space-efficient representations for fingerprints of drug-protein pairs and sparsity-induced classifiers.

Our method infers a set of drug-protein interaction signatures consisting of the associations between drug chemical substructures, adverse drug reactions, protein domains, biological pathways, and pathway modules. We argue the these signatures are biologically meaningful and useful for predicting unknown drug-protein interactions and are expected to contribute to rational drug design.

Paper 79

Identifying mutated driver pathways in cancer by integrating multi-omics data

Jingli Wu
Qirong Cai
Jinyan Wang
Yuanxiu Liao

Guangxi Normal University, China

Since the driver pathway in cancer plays a crucial role in the formation and progression of cancer, it is very imperative to identify driver pathways, which will offer important information for precision medicine or personalized medicine. In this paper, an improved maximum weight submatrix problem model is proposed by integrating such three kinds of omics data as somatic mutations, copy number variations, and gene expressions. The model tries to adjust coverage and mutual exclusivity with the average weight of genes in a pathway, and simultaneously considers the correlation among genes, so that the pathway having high coverage but moderate mutual exclusivity can be identified. By introducing a kind of short chromosome code and a greedy based recombination operator, a parthenogenetic algorithm PGA-MWS is presented to solve the model. Experimental comparisons among algorithms GA, MOGA, iMCMC and PGA-MWS were performed on biological and simulated data sets. The experimental results show that, compared with the other three algorithms, the PGA-MWS one based on the improved model can identify the gene sets with high coverage but moderate mutual exclusivity and scales well. Many of the identified gene sets are involved in known signaling pathways, most of the implicated genes are oncogenes or tumor suppressors previously reported in literatures. The experimental results indicate that the proposed approach may become a useful complementary tool for detecting cancer pathways.

Paper 80

ENIGMA: an enterotype-like unigram mixture model for microbial association analysis

Ko Abe
Masaaki Hirayama
Kinji Ohno
Teppei Shimamura

Nagoya University, Japan

One of the major challenges in microbial studies is detecting associations between microbial communities and a specific disease. A specialized feature of microbiome count data is that intestinal bacterial communities form clusters called as “enterotype”, which are characterized by differences in specific bacterial taxa, making it difficult to analyze these data under health and disease conditions. Traditional probabilistic modeling cannot distinguish between the bacterial divergences derived from enterotype and those related to a specific disease.

We propose a new probabilistic model, named as ENIGMA (Enterotype-like uNIGram mixture model for Microbial Association analysis), which can be used to address these problems. ENIGMA enabled simultaneous estimation of enterotype-like clusters characterized by the abundances of signature bacterial genera and the parameters of environmental effects associated with the disease.

In the simulation study, we evaluated the accuracy of parameter estimation. Furthermore, by analyzing the real-world data, we detected the bacteria related to Parkinson's disease. ENIGMA is implemented in R and is available from GitHub (https://github.com/abikoushi/enigma).

Paper 84

TSEE: an elastic embedding method to visualize the dynamic gene expression patterns of time series single-cell RNA sequencing data

Shaokun An[1]
Lin Wan[1]
Liang Ma[2]

[1]Academy of Mathematics and Systems Science, CAS, China
[2]Beijing Institute of Genomics, CAS, China

In this study, we propose an algorithm, termed time series elastic embedding (TSEE), by incorporating experimental temporal information into the elastic embedding (EE) method, in order to visualize time series scRNA-seq data. TSEE extends the EE algorithm by penalizing the proximal placement of latent points that correspond to data points otherwise separated by experimental time intervals. TSEE is herein used to visualize time series scRNA-seq datasets of embryonic developmental processed in human and zebrafish. We demonstrate that TSEE outperforms existing methods (e.g. PCA, tSNE and EE) in preserving local and global structures as well as enhancing the temporal resolution of samples. Meanwhile, TSEE reveals the dynamic oscillation patterns of gene expression waves during zebrafish embryogenesis.

TSEE can efficiently visualize time series scRNA-seq data by diluting the distortions of assorted sources of data variation across time stages and achieve the temporal resolution enhancement by preserving temporal order and structure. TSEE uncovers the subtle dynamic structures of gene expression patterns, facilitating further downstream dynamic modeling and analysis of gene expression processes. The computational framework of TSEE is generalizable by allowing the incorporation of other sources of information.

Paper 86

Estimating the total genome length of a metagenomic sample using k-mers

Kui Hua
Xuegong Zhang

Tsinghua University, China

As an initial step toward understanding the complete composition of a metagenomic sample, we studied the problem of estimating the total length of all distinct component genomes in a metagenomic sample. We showed that this problem can be solved by estimating the total number of distinct k-mers in all the metagenomic sequencing data. We proposed a method for this estimation based on the sequencing coverage distribution of observed k-mers, and introduced a k-mer redundancy index (KRI) to fill in the gap between the count of distinct k-mers and the total genome length. We showed the effectiveness of the proposed method on a set of carefully designed simulation data corresponding to multiple situations of true metagenomic data. Results on real data indicate that the uncaptured genomic information can vary dramatically across metagenomic samples, with the potential to mislead downstream analyses.

We proposed the question of how long the total genome length of all different species in a microbial community is and introduced a method to answer it.

Paper 87

GPU accelerated sequence alignment with traceback for GATK HaplotypeCaller

Shanshan Ren
Nauman Ahmed
Koen Bertels
Zaid Al-Ars

Delft Univ. of Tech., Netherlands

We first analyze the characteristics of the semi-global alignment with traceback in GATK HC and then propose a new algorithm that allows for retrieving the optimal alignment efficiently on GPUs. For the first stage, we choose intra-task parallelization model to calculate the position of the optimal alignment score and the backtracking matrix. Moreover, in the first stage, our GPU implementation also records the length of consecutive matches/mismatches in addition to lengths of consecutive insertions and deletions as in the CPU-based implementation. This helps efficiently retrieve the backtracking matrix to obtain the optimal alignment in the second stage.

Experimental results show that our alignment kernel with traceback is up to 80x and 14.14x faster than its CPU counterpart with synthetic datasets and real datasets, respectively. When integrated into GATK HC (alongside a GPU accelerated pair-HMMs forward kernel), the overall acceleration is 2.3x faster than the baseline GATK HC implementation, and 1.34x faster than the GATK HC implementation with the integrated GPU-based pair-HMMs forward algorithm. Although the methods proposed in this paper is to improve the performance of GATK HC, they can also be used in other pairwise alignments and applications.

Paper 89

Towards optimal decomposition of Boolean networks

Cui Su
Jun Pang
Soumya Paul

University of Luxembourg, Luxembourg

In recent years, great efforts have been made to analyse biological systems to understand the long-run behaviours. As a well-established formalism for modelling real-life biological systems, Boolean networks (BNs) allow their representation and analysis using formal reasoning and tools. Most biological systems are robust - they can withstand the loss of links and cope with external environmental perturbations. Hence, the BNs used to model such systems are necessarily large and dense, and yet modular. However, existing analysis methods only work well on networks of moderate size. Thus, there is a great need for efficient methods that can handle large-scale BNs and for doing so it is inevitable to exploit both the structural and dynamic properties of the networks. In this paper, we propose a method towards the optimal decomposition of BNs to balance the relation between the structure and dynamics of a network. We show that our method can greatly improve the existing decomposition-based attractor detection by analysing a number of large real-life biological networks.

Paper 100

Anti-TNF-α treatment-related pathways and biomarkers revealed by transcriptome analysis in Chinese psoriasis patients

Lunfei Liu[1]
Wenting Liu[2]
Yuxin Zheng[3]
Jisu Chen[1]
Jiong Zhou[1]
Huatuo Dai[1]
Suiqing Cai[1]
Jianjun Liu[4]
Min Zheng[1]
Yunqing Ren[1]

[1]Zhejiang University, China;
[2]U of California at Los Angles, USA;
[3]Zhejiang Chinese Med. U, China;
[4]Genome Inst. of S'pore, Singapore

To better understand the molecular mechanisms of Anti-TNF-α therapy, we analysed the global gene expression profile (using mRNA microarray) in peripheral blood mononuclear cells (PBMCs) that were collected from 6 psoriasis patients before and 12 weeks after the treatment of etanercept. First, we identified 176 differentially expressed genes (DEGs) before and after treatment by using paired t-test. Then, we constructed the gene co-expression modules by weighted correlation network analysis (WGCNA), and 22 co-expression modules were found to be significantly correlated with treatment response. Of these 176 DEGs, 79 DEGs (M_DEGs) were the members of these 22 co-expression modules. Of the 287 GO functional processes and pathways that were enriched for these 79 M_DEGs, we identified 30 pathways whose overall gene expression activities were significantly correlated with treatment response. Of the original 176 DEGs, 19 (GO_DEGs) were found to be the members of these 30 pathways, whose expression profiles showed clear discrimination before and after treatment. As expected, of the biological processes and functionalities implicated by these 30 treatment response-related pathways, the inflammation and immune response was the top pathway in response to etanercept treatment, and some known TNF-α related pathways, such as molting cycle process, hair cycle process, skin epidermis development, regulation of hair follicle development, were implicated.

Paper 104

Multiple optimal reconciliations under the duplication-loss-coalescence model

Haoxing Du[1]
Yi Sheng Ong[1]
Marina Knittel[1]
Ross Mawhorter[1]
Nou Liu[1]
Gianluca Gross[2]
Reiko Tojo[1]
Ran Libeskind-Hadas[1]
Yi-Chieh Wu[1]

[1]Harvey Mudd College, USA;
[2]University of Pennsylvania, USA

Gene trees can differ from species trees due to a variety of biological phenomena, the most prevalent being gene duplication, horizontal gene transfer, gene loss, and coalescence. To explain topological incongruence between the two trees, researchers apply reconciliation methods, often relying on a maximum parsimony framework. However, while several studies have investigated the space of maximum parsimony reconciliations (MPRs) under the duplication-loss and duplication-transfer-loss models, the space of MPRs under the duplication-loss-coalescence (DLC) model remains poorly understood. To address this problem, we present new algorithms for computing the size of MPR space under the DLC model and sampling from this space uniformly at random. Our algorithms are efficient in practice, with runtime polynomial in the size of the species and gene tree when the number of genes that map to any given species is fixed, thus proving that the MPR problem is fixed-parameter tractable. We have applied our methods to a biological data set of 16 fungal species to provide the first key insights in the space of MPRs under the DLC model. Our results show that a plurality reconciliation, and underlying events, are likely to be representative of MPR space.

Paper 108

SplicedFamAlign: CDS-to-gene spliced alignment and identification of transcript orthology groups

Safa Jammali,
Jean-David Aguilar,
Esaie Kuitche Kamela,
Aïda Ouangraoua

Universite de Sherbrooke, Canada.

The experimental results show that SFA outperforms existing spliced alignment methods in terms of accuracy and execution time for CDS-to-gene alignment. We also show that the performance of SFA remains high for various levels of sequence similarity between input sequences, thanks to accounting for the splicing structure of the input sequences. It is important to notice that unlike all current spliced alignment methods that are meant for cDNA-to-genome alignments and can be used for CDS-to-gene alignments, SFA is the first method specifically designed for CDS-to-gene alignments.

We show the usefulness of SFA for the comparison of genes and transcripts within a gene family for the purpose of analyzing splicing orthologies. It can also be used for gene structure annotation and alternative splicing analyses. SplicedFamAlign was implemented in Python. Source code is freely available at https://github.com/UdeS-CoBIUS/SpliceFamAlign.

Paper 109

Explore mediated co-varying dynamics in microbial community using integrated local similarity and liquid association analysis

Dongmei Ai[1]
Xiaoxin Li[1]
Hongfei Pan[1]
Jiamin Chen[1]
Jacob A. Cram[2]
Charlie Li Xia[2]

[1]U of Sci. and Tech. Beijing, China
[2]Stanford University, USA

Using this analysis, we were able to assess the OTUs' ability to regulate its functional partners in the community, typically not manifested in the pairwise correlation patterns. For example, we identified Flavobacteria as a multifaceted player in the marine microbial ecosystem, and its clades were involved in mediating other OTU pairs. By contrast, SAR11 clades were not active mediators of the community, despite being abundant and highly correlated with other OTUs. Our results suggested that Flavobacteria are more likely to respond to situations where particles and unusual sources of dissolved organic material are prevalent, such as after a plankton bloom. On the other hand, SAR11s are oligotrophic chemoheterotrophs with inflexible metabolisms, and their relationships with other organisms may be less governed by environmental or biological factors.

By integrating liquid association with local similarity analysis to explore the mediated co-varying dynamics, we presented a novel perspective and a useful toolkit to analyze and interpret time series data from microbial community. Our augmented association network analysis is thus more representative of the true underlying dynamic structure of the microbial community. The analytic software in this study was implemented as new functionalities of the ELSA (Extended local similarity analysis) tool, which is available for free download (http://bitbucket.org/charade/elsa).

Paper 113

Branching out to speciation with a birth-and-death model of fractionation: the malvaceae

Yue Zhang[1]
Chunfang Zheng[1]
Sindeed Islam[1]
Yong-Min Kim[2]
David Sankoff[1]

[1]Korea Res. Inst. of Biosci. & Biotech., South Korea
[2]University of Ottawa, Canada

Fractionation is the genome-wide process of losing one gene per duplicate pair following whole genome doubling (WGD). An important type of evidence for duplicate gene loss is the frequency distribution of similarities between paralogous gene pairs in a genome or orthologous gene pairs in two species. We extend a birth-and-death model for fractionation, originally accounting for paralog similarities, to encompass the distribution of ortholog similarities, after multiple rounds of whole genome doubling and fractionation, with the speciation event occurring at any point. We estimate the fractionation rates during all the inter-event periods in each lineage of the plant family Malvaceae. We suggest a major correction of the phylogenetic position of the durian sub-family, and discover a new triplication event in this lineage.

Paper 117

Sorting signed permutations by inverse tandem duplication random losses

Tom Hartmann
Max Bannach
Martin Middendorf

University of Leipzig, Germany

Gene order evolution of unichromosomal genomes, for example mitochondrial genomes, has been modelled mostly by four major types of genome rearrangements: inversions, transpositions, inverse transpositions, and tandem duplication random losses. Generalizing models that include all those rearrangements while admitting computational tractability are rare. In this paper we study such a rearrangement model, namely the inverse tandem duplication random loss (iTDRL) model, where an iTDRL duplicates and inverts a continuous segment of a gene order followed by the random loss of one of the redundant copies of each gene. The iTDRL rearrangement has currently been proposed by several authors suggesting it to be a possible mechanisms of mitochondrial gene order evolution. We initiate the algorithmic study of this new model of genome rearrangement by proving that a shortest rearrangement scenario that transforms one given gene order into another given gene order can be obtained in quasilinear time. Furthermore, we show that the length of such a scenario, i.e., the minimum number of iTDRLs in the transformation, can be computed in linear time.

Paper 124

Large-scale 3D chromatin reconstruction from chromosomal contacts

Yanlin Zhang[1]
Weiwei Liu[1]
Yu Lin[2]
Yen Kaow Ng[3]
Shuaicheng Li[1]

[1]City U of Hong Kong, China
[2]Australian Nat'l Univ., Australia
[3]Univ. Tunku Abdul Rahman, Malaysia

Recent advances in genome analysis have established that chromatin has preferred 3D conformations, which bring distant loci into contact. Identifying these contacts is important for us to understand possible interactions between these loci. This has motivated the creation of the Hi-C technology, which detects long-range chromosomal interactions. Distance geometry-based algorithms, such as ChromSDE and ShRec3D, have been able to utilize Hi-C data to infer 3D chromosomal structures. However, these algorithms, being matrix-based, are space- and time-consuming on very large datasets. A human genome of 100 kilobase resolution would involve ~30,000 loci, requiring gigabytes just in storing the matrices.

We propose a succinct representation of the distance matrices which tremendously reduces the space requirement. We give a complete solution, called SuperRec, for the inference of chromosomal structures from Hi-C data, through iterative solving the large-scale weighted multidimensional scaling problem.

SuperRec runs faster than earlier systems without compromising on result accuracy. The SuperRec package can be obtained from http://www.cs.cityu.edu.hk/~ shuaicli/ SuperRec.

Paper 126

Meta-network: optimized species-species network analysis for microbial communities

Pengshuo Yang[1]
Shaojun Yu[1]
Lin Cheng[2]
Ning Kang[1]

[1]Trinity College, USA;
[2]Huazhong U. of Sci. and Tech., China.

Hence, in this work, we propose the Meta-Network framework to lucubrate the microbial communities. Rooted in loose definitions of network(two species co-exist in a certain samples rather than all samples) as well as association rule mining (mining more complex forms of correlations like indirect correlation and mutual information), this framework outperforms other methods in restoring the microbial communities, based on two cohorts of microbial communities: (a) the loose definition strategy is capable to generate more reasonable relationships among species in the species-species co-occurrence network; (b) important species-species co-occurrence patterns could not be identified by other existing approaches, but could successfully generated by association rule mining.

Results have shown that the species-species co-occurrence network we generated are much more informative than those based on traditional methods. Meta-Network has consistently constructed more meaningful networks with biologically important clusters, hubs, and provides a general approach towards deciphering the species-species co-occurrence networks.

Paper 127

Automatic hierarchy classification in venation networks using directional morphological filtering for hierarchical structure traits extraction

Yangjing Gan[1]
Yi Rong[1]
Fei Huang[1]
Lun Hu[1]
Xiaohan Yu[1]
Pengfei Duan[1]
Shengwu Xiong[1]
Haiping Liu[2]
Jina Peng[1]
Xiaohui Yuan[3]

[1] Wuhan Univ. of Tech., China
[2] Tibet Academy of Agricultural and Animal Husbandry Sci., China
[3] Chinese Academy of Sci. China

The extraction of vein traits from venation networks is of great significance to the development of a variety of research fields, such as evolutionary biology. However, traditional studies normally target to the extraction of reticulate structure traits (ReSTs), which is not sufficient enough to distinguish the difference between vein orders. For hierarchical structure traits (HiSTs), only a few tools have made attempts with human assistance, and obviously are not practical for large-scale traits extraction. Thus, there is a necessity to develop the method of automated vein hierarchy classification, raising a new challenge yet to be addressed. We propose a novel vein hierarchy classification method based on directional morphological filtering to automatically classify vein orders. Different from traditional methods, our method classify vein orders from highly dense venation networks for the extraction of traits with ecological significance. To the best of our knowledge, this is the first attempt to automatically classify vein hierarchy. To evaluate the performance of our method, we prepare a soybean transmission image dataset (STID) composed of 1200 soybean leaf images and the vein orders of these leaves are manually coarsely annotated by experts as ground truth. We apply our method to classify vein orders of each leaf in the dataset. Compared with ground truth, the proposed method achieves great performance, while the average deviation on major vein is less than 5 pixels and the average completeness on second-order veins reaches 54.28%.

Paper 129

Genome-wide analysis of epigenetic dynamics across human developmental stages and tissues

Xia Zhang
Shanghai University, China
Yanglan Gan
Donghua University, China
Guobing Zou
Shanghai University, China
Jihong Guan
Tongji University, China
Shuigeng Zhou
Fudan University, China

Epigenome is highly dynamic during the early stages of embryonic development. Epigenetic modifications provide the necessary regulation for lineage specification and enable the maintenance of cellular identity. Given the rapid accumulation of genome-wide epigenomic modification maps across cellular differentiation process, there is an urgent need to characterize epigenetic dynamics and reveal their impacts on differential gene regulation.

We proposed DiffEM, a computational method for differential analysis of epigenetic modifications and identified highly dynamic modification sites along cellular differentiation process. We applied this approach to investigating 6 epigenetic marks of 20 kinds of human early developmental stages and tissues, including hESCs, 4 hESC-derived lineages and 15 human primary tissues.

We identified highly dynamic modification sites where different cell types exhibit distinctive modification patterns, and found that these highly dynamic sites enriched in the genes related to cellular development and differentiation. Further, we correlated the dynamics scores of epigenetic modifications with the variance of gene expression, and compared the results of our method with those of the existing algorithms. The comparison results demonstrate the power of our method in evaluating the epigenetic dynamics and identifying highly dynamic regions along cell differentiation process.

Paper 134

Prediction of FMN binding sites in electron transport chains based on 2-D CNN and PSSM profiles

Nguyen-Quoc-Khanh Le
Nanyang Tech. Univ., Singapore
Binh P. Nguyen
Victoria Univ. of Wellington, New Zealand

Flavin mono-nucleotide (FMN) is a cofactor which is involved in electron transport chains for carrying and transferring electrons in cellular respiration. Without the interaction from FMN, energy cannot be produced and most of the cellular processes cannot be performed. Therefore, creating a precise model to identify its functions is a crucial problem in order to understand human diseases and design drug targets. We proposed a deep learning model using a two-dimensional convolutional neural network and position specific scoring matrices profiles that could identify FMN interacting residues with the sensitivity of 83.7%, specificity of 99.2%, accuracy of 98.2%, and Matthews correlation coefficients of 0.85 for an independent dataset containing 141 FMN binding sites and 1,920 non-FMN binding sites. Our method outperformed other related works in all typical measurement metrics. Throughout the study, we provided an effective tool for prediction of FMN binding sites in electron transport chains, and our achievement could promote the use of deep learning in bioinformatics and computational biology.

Paper 139

Discovery of perturbation gene targets via free text metadata mining in gene expression omnibus

Djordje Djordjevic[1]
Joshua Y. S. Tang[1]
Yun Xin Chen[1]
Shu Lun[1]
Shannon Kwan[1]
Raymond W. K. Ling[1]
Gordon Qian[1]
Chelsea Y. Y. Woo[1]
Samuel J. Ellis[1]
Joshua W. K. Ho[123]

Victor Chang Cardiac Res. Inst., U of New South Wales, Australia; University of Hong Kong, China

There exists over 2.5 million publicly available gene expression samples across 101,000 data series in NCBI's Gene Expression Omnibus (GEO) database. Due to the lack of the use of standardised ontology terms in GEO's free text metadata to annotate the experimental type and sample type, this database remains difficult to harness computationally without significant manual intervention.

In this work, we present an interactive R/Shiny tool called GEOracle that utilises text mining and machine learning techniques to automatically identify perturbation experiments, group treatment and control samples and perform differential expression. We present applications of GEOracle to discover conserved signaling pathway target genes and identify an organ specific gene regulatory network.

GEOracle is effective in discovering perturbation gene targets in GEO by harnessing its free text metadata. Its effectiveness and applicability have been demonstrated by cross validation and two real-life case studies. It opens up new avenues to unlock the gene regulatory information embedded inside large biological databases such as GEO. GEOracle is available at https://github.com/VCCRI/GEOracle.

Paper 140

Constructing optimal energy functions for protein structure prediction using reverse Monte Carlo sampling

Dongbo Bu[1]
Chao Wang[1]
Haicang Zhang[1]
Shiwei Sun[1]
Yi Wei[1]
Wei-Mou Zheng[2]

[1]Institute of Computing Tech., [2]Institute of Theoretical Physics, CAS, China

We present a framework to construct effective energy functions for protein structure prediction. Unlike existing energy functions only requiring the native structure to be the lowest one, we attempt to maximize the attraction-basin where the native structure lies in the energy landscape. The underlying rationale is that each energy function determines a specific energy landscape together with a native attraction-basin, and the larger the attraction-basin is, the more likely for the Monte Carlo search procedure to find the native structure. Following this rationale, we constructed effective energy functions as follows: i) To explore the native attraction-basin determined by a certain energy function, we performed reverse Monte Carlo sampling starting from the native structure, identifying the structural conformations on the edge of attraction-basin. ii) To broaden the native attraction-basin, we smoothened the edge points of attraction-basin through tuning weights of energy terms, thus acquiring an improved energy function. Our framework alternates the broadening attraction-basin and reverse sampling steps (thus called BARS) until the native attraction-basin is sufficiently large. We present extensive experimental results to show that using the BARS framework, the constructed energy functions could greatly facilitate protein structure prediction in improving the quality of predicted structures and speeding up conformation search.

Using the BARS framework, we constructed effective energy functions for protein structure prediction, which could improve the quality of predicted structures and speed up conformation search as well.

Paper 141

A secure SNP panel scheme using homomorphically encrypted k-mers without SNP calling on the user side

Sungjoon Park[1]
Minsu Kim[1]
Seokjun Seo[2]
Seungwan Hong[1]
Kyoohyung Han[1]
Keewoo Lee[1]
Jung Hee Cheon[1]
Sun Kim[1]

[1]Seoul National University, Seoul, South Korea;
[2]Hyperconnect Inc., South Korea

In this paper, we propose a secure SNP panel scheme using homomorphically encrypted K-mer without requiring SNP calling on the user side and without revealing the panel information to the user. Use of the powerful homomorphic encryption technique is desirable, but there is no known algorithm to efficiently align two homomorphically encrypted sequences. Thus, we designed and implemented a novel secure SNP panel scheme utilizing the computationally feasible equality test on two homomorphically encrypted K-mer. To make the scheme work correctly, in addition to SNP in the panel, sequence variations at the population level should be addressed. We designed a concept of Point Deviation Tolerance (PDT) level to address the false positives and false negatives. Using the TCGA BRCA dataset, we demonstrated that our scheme works at the level of over a hundred thousand somatic mutations. In addition, we provide a computational guideline for the panel design, including the size of K-mer and the number of SNPs.

The proposed method is the first of its kind to protect both the user's sequence and the hospital's panel information using the powerful homomorphic encryption scheme. We demonstrated that the scheme works with a simulated dataset and the TCGA BRCA dataset. We have shown only the feasibility of the proposed scheme and much more efforts should be done to make the scheme usable for clinical use.

Paper 143

Boolean network modeling of beta-cell apoptosis and insulin resistance in type 2 diabetes mellitus

Pritha Dutta[1]
Lichun Ma[2]
Yusuf Ali[1]
Peter M.A. Sloot[1]
Jie Zheng[3]

[1]Nanyang Technogical Univ., Singapore;
[2]National Inst. of Health, USA;
[3]ShanghaiTech University, China

Major alteration in lifestyle of human population has promoted Type 2 diabetes mellitus (T2DM) to the level of an epidemic. This metabolic disorder is characterized by insulin resistance and pancreatic beta-cell dysfunction and apoptosis, triggered by endoplasmic reticulum (ER) stress, oxidative stress and cytokines. Computational modeling is necessary to consolidate information from various sources in order to obtain a comprehensive understanding of the pathogenesis of T2DM and to investigate possible interventions by performing in silico simulations.

In this paper, we propose a Boolean network model integrating the insulin resistance pathway with pancreatic beta-cell apoptosis pathway which are responsible for T2DM. The model has five input signals, i.e. ER stress, oxidative stress, tumor necrosis factor alpha (TNFalpha), Fas ligand (FasL), and interleukin-6 (IL-6). We performed dynamical simulations using random order asynchronous update and with different combinations of the input signals. From the results, we observed that the proposed model made predictions that closely resemble the expression levels of genes in T2DM as reported in the literature.

The proposed model can make predictions about expression levels of genes in T2DM that are in concordance with literature. Although experimental validation of the model is beyond the scope of this study, the model can be useful for understanding the aetiology of T2DM and discovery of therapeutic intervention for this prevalent complex disease.

Paper 145

De Novo glycan structural identification from mass spectra using tree merging strategy

Fusong Ju[1]
Jingwei Zhang[1]
Dongbo Bu[1]
Yan Li[2]
Jinyu Zhou[2]
Hui Wang[1]
Yaojun Wang[1]
Chuncui Huang[2]
Shiwei Sun[1]

[1]Institute of Comput. Tech.,
[2]Institute of Biophysics,
[3]University of Chinese Academy of Sciences, CAS, China;

In this study we propose an efficient and reliable approach to glycan structure identification using tree merging strategy. Briefly, for each MS peak, our approach first calculated monosaccharide composition of its corresponding fragment ion, and then built a constraint that forces these monosaccharides to be directly connected in the underlying glycan tree structure. According to these connecting constraints, we next merged constituting monosaccharides of the glycan into a complete structure step by step. During this process, the intermediate structures were represented as subtrees, which were merged iteratively until a complete tree structure was generated. Finally the generated complete structures were ranked according to their compatibility to the input mass spectra. Unlike the traditional enumerating followed by filtering strategy, our approach performed deisomorphism to remove isomorphic subtrees, and ruled out invalid structures that violates the connection constraints at each tree merging step, thus significantly increasing efficiency. In addition, all complete structures satisfying the connection constraints were enumerated without any missing structure. Over a test set of 10 N-glycan standards, our approach accomplished structural identification in minutes and gave the manually-validated structure first three highest score. We further successfully applied our approach to profiling and subsequent structure assignment of glycans released from glycoprotein mAb, which was in perfect agreement with previous studies and CE analysis.

Paper 146

A fast and efficient count-based matrix factorization method for detecting cell types from single-cell RNAseq data

Shiquan Sun
Yabo Chen
Yang Liu
Xuequn Shang

Northwestern Polytechnical University, China

Single-cell RNA sequencing (scRNAseq) data always involves various unwanted variables, which would be able to mask the true signal to identify cell-types. More efficient way of dealing with this issue is to extract low dimension information from high dimensional gene expression data to represent cell-type structure. Several powerful matrix factorization tools have been developed for scRNAseq data, such as NMF, ZIFA, pCMF and ZINB-WaVE. But the existing approaches either are unable to directly model the raw count of scRNAseq data or are really time-consuming when handling a large number of cells (e.g. n > 500).

In this paper, we developed a fast and efficient count-based matrix factorization method (single-cell negative binomial matrix factorization, scNBMF) based on the TensorFlow framework to infer the low dimensional structure of cell types. To make our method scalable, we conducted a series of experiments on three public scRNAseq data sets, brain, embryonic stem and pancreatic islet. The experimental results show scNBMF is more powerful to detect cell types and 10-100 folds faster than the scRNAseq bespoke tools.

In this paper, we proposed a fast and efficient count-based matrix factorization method, scNBMF, which is more powerful for detecting cell type purposes. A series of experiments were performed on three public scRNAseq data sets. The results show that scNBMF is a more powerful tool in large-scale scRNAseq data analysis. scNBMF was implemented in R and Python, and the source code are freely available at https://github.com/sqsun.

Paper 148

Identification of Hürthle cell cancers: solving a clinical challenge with genomic sequencing and a trio of machine learning algorithms

Yangyang Hao[1]
Quan-Yang Duh[2]
Richard Kloos[1]
Joshua Babiarz[1]
R. Mack Harrell[3]
S. Thomas Traweek[4]
S.Y. Kim[1]
G, Fedorowicz[1]
P. Sean Walsh[1]
Peter Sadow[5]
Jing Huang[1]
Giulia Kennedy[1]

[1]Veracyte Inc.,
[2]U of California, San Francisco
[3]Memorial Center for Integrative Endocr. Surgery,
[4]Thyroid Cytopathology Partners,
[5]Harvard Medical School, USA.

We sought to overcome this low-specificity limitation by expanding the feature set for ML using next-generation whole transcriptome RNA sequencing and called the improved algorithm the Genomic Sequencing Classifier (GSC). The Hürthle identification leverages mitochondrial expression and we developed novel feature extraction mechanisms to measure chromosomal and genomic level loss-of-heterozygosity (LOH) for the algorithm. Additionally, we developed a multi-layered system of cascading classifiers to sequentially triage Hürthle cell-containing FNAB, including: 1. presence of Hürthle cells, 2. presence of neoplastic Hürthle cells, and 3. presence of benign Hürthle cells. The final Hürthle cell Index utilizes 1,048 nuclear and mitochondrial genes; and Hürthle cell Neoplasm Index leverages LOH features as well as 2,041 genes. Both indices are Support Vector Machine (SVM) based. The third classifier, the GSC Benign/Suspicious classifier, utilizes 1,115 core genes and is an ensemble classifier incorporating 12 individual models.

The accurate algorithmic depiction of this complex biological system among Hürthle subtypes results in a dramatic improvement of classification performance; specificity among Hürthle cell neoplasms increases from 11.8% with the GEC to 58.8% with the GSC, while maintaining the same sensitivity of 89%.

Paper 149

Detecting virus-specific effects on post-infection temporal gene expression

Quan Chen
Jun Zhu

Icahn School of Medicine at Mount Sinai, USA

Different types of viruses have different envelope proteins, and may have their shared or distinctive host-virus interactions which result in various post-infection effects in humans and animals. These effects often do not appear at once but take time to unfold. To characterize the virus-specific effects, we applied a Multivariate Polynomial Time-dependent Genetic Association (MPTGA) method, previously proposed for detecting differences in temporal gene expression traits, to test for the differences in mouse lung transcriptome response to infection of different subtypes of influenza A viruses.

We compared two methods: the Multivariate Polynomial Time-dependent Genetic Association (MPTGA) method, and the conventional modified t-test, to study the virus-specific effects on mouse lung gene expression. Both methods found H3N2 to be the most different virus among the three viruses tested, with the largest number of genes with H3N2-specific effects. However, the MPTGA method demonstrated much higher power of detection, and the detected genes with virus-specific effects showed better biological relevance.

Transcriptome response to virus infection is dynamic. MPTGA which leverages temporal gene expression traits showed increased power in detecting biologically relevant virus-specific effects comparing conventional t-test method.

Paper 150

Microbiota in the apical root canal system of tooth with apical perodontitis

Wenhao Qian[1]
Ting Ma[2]
Mao Ye[1]
Zhiyao Li[1]
Yuanhua Liu[2]
Pei Hao[2]

[1]Shanghai Xuhui District Dental Center, Chian;
[2]Institut Pasteur of Shanghai, Chinese Academy of Sciences, China

This study aims to uncover the composition and diversity of microbiota associated to the root apex to identify the relevant bacteria highly involved in AP, with the consideration of root apex samples from the infected teeth (with/without root canal treatment), healthy teeth as well as the healthy oral.

Four groups of specimens are considered, the apical part of root from diseased teeth with and without root canal treatment, and wisdom teeth extracted to avoid being impacted (tooth healthy control), as well as an additional healthy oral control from biofilm of the buccal mucosa. DNA was extracted from these specimens and the microbiome was examined through focusing on the V3-V4 hypervariable region of the 16S rRNA gene using sequencing on Illumina MiSeq platform. Composition and diversity of the bacterial community were tested for individual samples, and between-group comparisons were done through differential analysis to identify the significant changes.

We observed reduced community richness and diversity in microbiota samples from diseased teeth compared to healthy controls. Through differential analysis between AP teeth and healthy teeth, we identified 88 OTUs significantly down-regulated as well as 22 up-regulated OTUs for AP.

This study provides a global view of the microbial community of the AP associated cohorts, and revealed that AP involved not only bacteria accumulated with a high abundance, but also those significantly reduced ones due to microbial infection.

Poster 151

An algorithm to predict feature-based gene equivalence between mouse strains

Rex Asabor
Harvey Mudd College, USA

Joel Richardson
Jackson Laboratory, USA

Richard Baldarelli
Jackson Laboratory, USA

The advent of low-cost high-throughput sequencing commenced the generation of full genome annotations for multiple inbred strains of mice. As the number of sequenced mouse strains continue to rise, new technology must be developed to expedite the discovery of gene function. In this report, we discuss a new approach to the alignment of genomes, which analyzes the genomic features from the sequence annotations and predicts cross-strain gene equivalence based on a calculated dissimilarity index, opposed to conventional techniques that compare nucleotide sequences. The development of this algorithm included the exploration of different parameters to represent and measure the similarity between genomic features. The final algorithm utilizes a logistic regression model as a metric and was assessed by comparing its predictions of associations between strain genes to the MUSCLE multiple sequence alignment of those same features. Despite its ultimate overall accuracy of 90% in predicting new associations between all strains and 97% in predicting associations between the newer, non-reference strains within a selected chromosomal region, there still exist areas for development. This algorithm for genomic feature comparison could be tweaked to be able to associate genes of organisms that lack a high quality annotated reference genome. Future enhancements of this algorithm will enable it to be an even more useful tool to biologists as they investigate the genetic origins of phenotypic variations.

Poster 152

Identification and classification of long noncoding RNAs in cucumis sativus

Yao Lin
Kun Lin

Beijing Normal Univ., China

Recently, more and more studies have shown compelling biological roles of diverse groups of long noncoding RNAs (lncRNAs). In this work, we present an ongoing study about identification and classification of putative lncRNAs in cucumber (C. satavus), a widely cultivated plant of the Cucurbitaceae family.

With a set of 105 RNA-Seq samples with read lengths of 150 bp taken from five tissues at seven time points, in total 48,953 transcripts were assembled with HISAT2, StringTie and TACO. Constrained by the alignment with known protein-coding sequences and the ORF length, 7,778 transcripts were selected as lncRNA candidates. Of them, 6,372 were labeled as “noncoding” by a tool called LncADeep (Yang et al., 2018) which calculates a coding potential for each lncRNA candidate. In future, we will predict those lncRNAs containing conserved short open reading frames (sORFs) by means of comparative genomics, those

lncRNAs acting as endogenous target mimic of miRNAs identified from small RNA-seq dataset, those regulating target protein-coding genes, and will classify other lncRNA candidates according to the profiles of short motifs or k-mers.

Taken together, we will identify thousands of lncRNAs in cucumber, increasing the number of the cucumber ncRNAs. We hope our analysis may allow a greater level of detail in exploring the lncRNAs in cucumber.

Poster 153

Identification and profiling of drug-induced cellular phenotypes in human monocytes using machine learning

Shaista Hussain[1]
Binh P. Nguyen[1]
Mardiana Marzuki[2]
Shuping Lin[3]
Graham Wright[3]
Amit Singhal[2]

[1]Inst. of High Perform. Comput., Singapore;
[2]Singapore Immunol. Network, Singaproe;
[3]Inst. of Medical Biology Singapore

Image-based profiling and high-performance computational analysis can be used to characterize the biological effects of thousands of compounds in the drug discovery process. Similar pipeline can also be used to identify the undesired and anomalous biological effects of drug(s), which can limit its safe use in respective therapeutical application(s).

The developed high-content computational pipeline was used to analyze the effects of 225 FDA approved drugs on human monocyte cell line (THP-1) based on 12 image features. Control modeling step found that 22 drugs significantly induced different effects on the monocytes as compared to the untreated control cells. GMM based phenotypic profiling of these cellular alterations resulted in 8 clusters, which were used to generate phenotypic signatures for the drug treatments. This mapping of phenotypic clusters to the drug signatures was used to identify similarities between drug compounds. One such drug cluster included Tamoxifen citrate, Metformin and other drugs. Functional assays further showed that Tamoxifen and Metformin can synergize in restricting the growth of triple-negative tumor cells.

Our high-content screening in conjunction with machine learning approach identified the toxic effect of known FDA approved drugs on human monocytes. This can have broad implications on the use of these drugs for their respective therapeutic conditions or repurposing of the drugs for different clinical interventions.

Poster 155

Prediction of cancer related single amino acid variation by machine learning method

Hsiao-Wei Wu
Chih-Hao Lu

China Medical University, ROC

Single-nucleotide polymorphisms (SNPs) are the natural DNA variations caused by changes in the single-base nucleotide. Some evidence has showed that SNPs might be related to some diseases, including cancer. Though, most genomic variations might be silent, the coding change would affect the transcription product, protein, which directly involved in the biological process. Therefore, analyzing the single amino acid variation (SAV) in protein sequence becomes increasingly important. The aim of this study is to develop a prediction model which can identify whether the SAV is cancer related or not. First, we collected the data from CanProVar 2.0., a human Cancer Proteome Variation Database. The prediction model is based on the protein structure characteristics. All of property factors of SAV generated from protein structure will be converted into feature vectors and fed into Support Vector Machine (SVM) for model training and result testing. The feature selection procedure will be performed by the genetic algorithm. Then, the structure model will be integrated for final prediction model for cancer related SAV prediction.

Poster 157

Topology data analysis to visualize and reconstruct complex trajectories of cell development for large-scale scRNA-seq data

Ziwen Chen[1]
Shaokun An[1]
Xiangqi Bai[1]
Fuzhong Gong[1]
Liang Ma[2]
Lin Wan[1]

[1]*Academy of Mathematics and Systems Science, CAS, China;*
[2]*Beijing Institute of Genomics, CAS, China*

The rapid advances of single-cell RNA sequencing (scRNA-seq) technologies provide unprecedented opportunities to reveal the mechanisms of cell fate decisions. However, visualizing and reconstructing the complex trajectories of cell development for single-cell snapshot data remains a great challenge. Here, we develop a novel algorithm based on topology data analysis to visualize and reconstruct the underlying cell developmental trajectories for large-scale scRNA-seq data. The algorithm powerfully handles the high-dimensional and heterogeneous scRNA-seq data by (1) revealing the intrinsic structures of data based on the nonlinear dimensionality reduction algorithm, elastic embedding; and (2) extracting high density level-set clusters of representative cell states (RCSs) from the single-cell multimodal density landscape from topological perspectives. The algorithm reconstructs the cell state-transition path by finding the geodesic minimum spanning tree of RCSs on density landscape, establishing an underlying constrain of the minimum-transition-energy of cell fate decisions. We demonstrate that the proposed algorithm is powerful to visualize and reconstruct complex tree-like trajectories of cell development for large-scale scRNA-seq data, while maintaining computational efficiency. It has high accuracy in pseudotime calculation and branch assignment on real scRNA-seq, as well as simulated datasets. Moreover, the algorithm is robust for parameter choices and permutations of data.

Poster 159

Pathway-based approaches for analysis of candidate genes associated with neuropsychiatric disorders

Ying Hu
Tianjin Normal Univ., China

Ju Wang
Tianjin Medical Univ., China

As we know, one of the most widely used pathway enrichment analysis approach, over-representation analysis (ORA), ignores the function non-equivalence of genes in candidate gene set and may have low discriminative power in identifying some dysfunctional pathways. To overcome such drawbacks, we first assigned a function weighting score to each candidate gene based on their correlation with disease, and then by incorporating the function weighting scores of candidate genes into standard ORA pathway analysis, we proposed a novel pathway identification approach. Finally, we applied this approach to nicotine dependence and identified the dysfunctional pathways involved in this disorder.

In a biological process, pathways are non-independent to each other; instead, they usually work together in a highly orchestrated fashion and the function of two pathways can be cooperative, compensatory or alternative. Focusing on the complex relationship among pathways, we proposed a method to measure the relationship between pathways based on their distribution in the protein-protein interaction network. For the pathways in the KEGG database, a total of 2143 pathway pairs with close connections were identified. Further, we analyzed the pathway relationship and identified the major pathways related to Parkinson's disease via this method.

Poster 160

A survey on cellular RNA editing activity in response to HBV infections of human hepatocellular carcinoma patients

Ting Ma
Yiyi Jiang
Yuanhua Liu
Pei Hao

Institut Pasteur of Shanghai, CAS, China

Adenosine-to-Inosine (A-to-I) RNA editing is mediated by the family of adenosine deaminase acting on RNA (ADAR) proteins. Previous studies have found massive differentials in A-to-I RNA editing events between normal and tumor tissues of hepatocellular carcinoma (HCC). However, A-to-I RNA editing activity in response to hepatitis B virus (HBV) infection has not been fully characterized in HCC patients. This work was designed to investigate the differences of A-to-I RNA editing activity between HCC and HBV-related HCC.

Results: Using a bioinformatics pipeline called SPRINT, we characterized the profiles of A-to-I RNA editing events in twenty-eight paired HCC tumors and matched adjacent tissues. We totally identified 2,225,112 A-to-I RNA editing sites (A-to-I RESs) and 11,720,998 A-to-I RESs in 9 HCC patients and 19 HBV-related HCC patients, respectively. And most A-to-I RESs in HCC patients were enriched in introns, intergenic regions of genes and 3'-UTR. Then we found both the editing rate and the expression levels of ADAR1 were higher in tumors than adjacent normal tissues in two types of HCC patients. However, there was no significant difference in the editing levels of common A-to-I RESs between tumor and normal tissues. Compared with HCC patients, the editing levels of specific A-to-I RESs in HBV-related HCC patients showed higher diversity.

Our study represents the first comprehensive analysis of A-to-I RNA editing activity in HCC and HBV-related HCC patients. However, HBV infection did not show significant changes in A-to-I RNA editing activity in HCC patients.

Poster 161

The detection of positive selection signals in human populations

Xinrui Lin
Kui Lin
Ning Zhang
Erli Pang

Beijing Normal Univ., China

The 1000 Genomes Project established a deep catalogue of human genetic variation, The finale analysis, publish in October 2015, incorporated 2,504 individuals from 26 populations. Therefore, such data is appropriate to detect signatures of recent natural selection in modern humans.

We used three signatures: long haplotype, high frequency derived alleles and highly differentiated alleles. For each signature, (1) we calculated a raw score and defined a P value separately for each SNP; (2) we integrated results for all SNPs within each gene to identify the candidate genes of selection. Then, we combined the three signals to identify the candidate genes of selections.

We gained 71 positive selected genes and 87 involved pathways finally. Populations expect in Africa were able to detect positive selected genes. Pathway and process enrichment analysis was applied to gave 8 clusters.

Our approach gave a composite of multiple signals to detect positive selection genes with extendibility enabled the similar research even in non-model organisms.

Poster 162

Microbial interaction extraction based on feature vector and SVM

Ran Zhong
Xingpeng Jiang
Central China University, China

Thousands of microorganisms form a complex microbial community that dynamically regulates the body's ecological environment which is essential to human health. As a large number of experimentally validated microbial interactions have been published in the medical literature, the retrieving of these knowledge resources has become more and more difficult and cannot be systematically used. It is also hard to detect the potential interactions in the massive biomedical text through manual way. Text mining provides an automatic and effective way for microbial interaction extraction (MIE) to organize the intricate microbial interactions in vast amounts of texts into available Database or Knowledge Graph. In this paper, a Microbial Interaction Corpus (MICorpus) with a total of 1005 abstracts including 2,193 sentences and 7483 of the interaction relationships was manually labeled to provide a valuable data resource for the MIE task. Based on the corpus, we propose a supervised learning method based on feature vector and SVM to extract the microbial interaction from biomedical literature. The results are F-score1=86.32%, F-score2=78.21% based on the document and sentence-level corpus, respectively.

Poster 163

Detecting circular RNA from high-throughput sequence data with De Bruijn graph

Xin Li
Yufeng Wu
University of Connecticut, USA

Circular RNA is a type of non-coding RNA, which has a circular structure. Many circular RNAs are stable and contain exons, but are not translated into proteins. Circular RNA has important functions in gene regulation and plays an important role in some human diseases. Several biological methods, such as RNase R treatment, have been developed to identify circular RNA. Multiple bioinformatics tools have also been developed for circular RNA detection with high-throughput sequence data. In this paper, we present circDBG, a new method for circular RNA detection with bidirectional de Bruijn graph. We conduct various experiments to evaluate the performance of CircDBG on both simulated and real data. Our results show that CircDBG finds more reliable circRNAs with low bias, is more efficient, and performs better in balancing accuracy and sensitivity than existing methods. As a byproduct, we also introduce a new method for classifying circular RNAs based on reads alignment. Finally we report a potential chimeric circular RNA that is found by CircDBG on real sequence data. CircDBG can be downloaded from https://github.com/lxwgcool/CircDBG.

Poster 164

GRaPe 2.0: Accelerating the construction of kinetic models for cell metabolism

Fu Yap
Simon Hubbard
Jean-Marc Schwartz
University of Manchester, UK

We present the GRaPe 2.0 software package which is aimed at streamlining the process of generating large-scale metabolic models. Major innovations include the use of convenience kinetics to generate rate equations, enabling modelling of reactions containing any number of substrates and products without need for manual definition of complex kinetic functions; the addition of regulatory interactions such as allosteric inhibitions or activations; and the simultaneous use of several types of omics data, such as metabolomics and proteomics, to inform parameter values.

The software was used to build a kinetic model of the trehalose metabolic pathway in Saccharomyces cerevisiae. The model was parametrized using quantitative proteomics data measured under standard growth conditions, and validated by testing its capability to predict data under heat stress conditions. The completed model was used to investigate factors related to the rise in flux during heat stress and affecting the production of trehalose, a compound of industrial interest valued for its protective properties. The model revealed that feed-forward activation of pyruvate kinase by fructose 1,6-bisphosphate during heat stress contributes to the increase in metabolic flux. We also demonstrate that overexpression of enzymes involved in the production and degradation of trehalose can lead to higher trehalose yield in the cell. Together these results show that a modelling approach based on generic equations and parameter estimation using omics data can provide an efficient solution to create kinetic models of metabolic systems with predictive properties. The GRaPe 2.0 software and source code are freely available from https://github.com/chuanfuyap/GRaPe2 .

Poster 166

Enhanced knowledge generation of rate-changes in transcriptional regulation using ontology rules, sentence structure and deep neural networks

Wenting Liu
University of California at Los Angeles, USA

Yilei Zhang
Nanjing Medical Univ., China

Considering the complexity of biosystems, it's very challenging to automatically and accurately generate biological knowledge from bio-literatures. The existing biological databases do not record temporal information of gene regulations, which are very important to understand the underlying mechanism of many diseases and biological processes. We previously constructed a corpus of time-delays related to the transcriptional regulation (bio-events) of yeast from the PubMed abstracts, summarized the knowledge rules of the bio-events as rate-changes in transcriptional regulation ontology, and obtained 86% accuracy by using the decision tree classifier with the ontology rule features. Deep neural networks (DNN) achieve great success in many machine learning applications including knowledge generation. The word2vec model learned the word embedding features from documents can achieve 50-70% accuracy on many text classification tasks. However, the sentence structure and domain knowledge are rarely considered in DNNs of document classification. We proposed to combine word2vec features, sentence structure, and our ontology rule features to improve DNNs for knowledge generation of rate-change in transcriptional regulation. Experimental results show that on predicting transcriptional regulation, the word2vec in DNN model achieves 73% accuracy, while our combined features in DNN with same parameters achieves 96% accuracy; on predicting the rate-changes in transcription regulation events, word2vec in DNN achieves only 59% accuracy, and our combined features in DNN achieves 90% accuracy. This shows the power of ontology rules and sentence structure features with DNN in knowledge generation.

Poster 167

A new model for predicting metagenomic functional profile from 16S rRNA gene amplicon sequencing data

Shion Hosoda
Michiaki Hamada

Waseda University, Japan

Next generation sequencing enabled investigating environmental microbes using metagenomic analysis.There are two ways of sequencing for metagenomic analysis, which are whole metagenome shotgun (WMS) sequencing and 16S rRNA gene amplicon (16S) sequencing.While WMS sequencing can obtain both of microbial compositions and functional profiles, 16S sequencing can only obtain microbial compositions but has lower experimental cost. Therefore, predicting functional profiles from 16S sequencing data is an important task. However, previous software cannot predict these information with high accuracy.

In this study, we proposed a method that can predict functional profiles accurately based on hierarchical Bayes model. Our method predicts functional profiles by estimating ortholog table, which indicates each ortholog gene copy number of each microbe, as model parameters. In this model, each functional profile is generated by the Dirichlet distribution whose mode is a product of microbial composition and ortholog table. In addition, each element of ortholog table is generated by the Poisson distribution.

Our method was applied to Human Microbiome Project data that are sequenced by both of WMS and 16S, and it was compared with conventional software.

Our method outperformed existing method by more than 20\% in terms of correlation between samples and predictions, and ortholog table estimated by our method characterized microbe's evolutionary distance.

Poster 168

Identification of the therapeutic targets for selective killing of HBV-positive hepatocytes

Chien-Jung Huang
Yu-Chao Wang

National Yang-Ming Univ., ROC

Hepatitis B virus (HBV) infection is a global healthcare problem. According to the report of World Health Organization in 2015, an estimated 257 million people were living with chronic HBV infection globally. Most infected people will become lifelong carriers of HBV because there is currently no cure for HBV infection. In addition, HBV infection is a major risk factor of cirrhosis and hepatocellular carcinoma. Therefore, how to control or even cure HBV infection would be an essential public health issue. At present, there are two types of available agents (interferon and nucleotide analogues) for the treatment of HBV infection. Nevertheless, these drugs can only control the disease, achieving functional cure (loss of hepatitis B surface antigen), but not complete cure (elimination of infected hepatocytes, including cells with integrated HBV DNA). Therefore, in this study, we would like to identify the selective killing target genes which can be used to develop new therapies to treat HBV infection.

A computational framework was developed to identify the selective killing targets of HBV-positive hepatocytes. These genes which are essential for cell survival in HBV-positive hepatocytes but are not essential for cell survival in HBV-negative cells may be considered as potential drug targets to achieve the ultimate goal of complete cure for hepatitis B. Further biological experiments should be conducted to validate these identified selective killing targets to confirm the reliability of our method. In addition, this proposed method can be applied to diseases other than hepatitis B (such as cancer) to identify new therapeutic targets for other diseases.

Poster 170

A comprehensive comparison study between genes associated with different drug sensitivity of cancer cells based on connectivity map

Xinhua Liu
Yuan Zhou
Qinghua Cui
Peking University, China

It is widely accepted that aberrant gene expression could contribute significantly to drug sensitivity in various diseases, including cancer. However, our understanding on the association between genes and drug sensitivity and drug resistance is still limited. Here, we systematically compared genes regulated by different drugs/small bioactive molecules in five types of cancer cells, including MCF7, ssMCF7, PC3, HL60 and SKMEL5, by examining their genomic distribution, evolution, interaction network topology, subcellular localization, as well as their functions through exploring genome-wide expression profiles deposited in Connectivity Map (CMAP). Based on the threshold of top 15% genes regulated by more drugs (TGenes) and the rest 85% genes (LGenes), we found a more clustering tendency among TGenes than that of LGenes in genome. What's more, LGenes were found to be more conservative on evolution and have higher protein-protein interaction network degree than TGenes. Strikingly, TGenes were significantly enriched in cancer related pathways, while, LGenes were mainly associated with development of nervous system diseases. Taken together, our study should shed new light on mechanisms of drug resistance in cancers, but extensive researches are still needed in other types of diseases.

Poster 171

Learning protein structural fingerprints under the label-free supervision of domain knowledge

Yaosen Min
Shang Liu
Xuefeng Cui
Tsinghua University, China

Finding homologous proteins is the indispensable first step in many protein biology studies. Thus, building highly efficient ``search engines'' for protein databases is a highly desired function in protein bioinformatics. As of August 2018, there are more than 140,000 protein structures in PDB, and this number is still increasing rapidly. Such a big number introduces a big challenge for scanning the whole structure database with high speeds and high sensitivities at the same time. Unfortunately, classic sequence alignment tools and pairwise structure alignment tools are either not sensitive enough to remote homologous proteins (with low sequence identities) or not fast enough for the task. Therefore, specifically designed computational methods are required for quickly scanning structure databases for homologous proteins.

Here, we propose a novel ContactLib-DNN method to quickly scan structure databases for homologous proteins. The core idea is to build structure fingerprints for proteins, and to perform alignment-free comparisons with the fingerprints. Specifically, the fingerprints are low-dimensional vectors representing the contact groups within the proteins. Notably, the Cartesian distance between two fingerprint vectors well matches the RMSD between the two corresponding contact groups. This is done by using RMSD as the domain knowledge to supervise the deep neural network learning. When comparing to existing methods, ContactLib-DNN achieves the highest average AUROC of 0.959. Moreover, the best candidate found by ContactLib-DNN has a probability of 70.0% to be a true positive. This is a significant improvement over 56.2%, the best result produced by existing methods.

Tutorial title: RNA-seq and single-cell data analysis

Cheng Li (李程)
Principal Investigator
School of Life Sciences and Center for Statistical Science
Peking University, China
Email: cheng_li@pku.edu.cn

Abstract:
Contemporary life sciences and medicine are moving towards the era of large data as represented by high-throughput sequencing. How to model, analyze and interpret genomic data will determine whether we can quickly and accurately discover new biological phenomena and rules, and provide accurate medical care for patients. This tutorial will introduce RNA-seq and single-cell sequencing data types in genomics, and statistical analysis and plotting methods commonly used in data analysis, including exploratory data analysis, normalization and clustering. The tutorial will discuss related literature and data examples, and use the R programming environment for data analysis and plotting exercises.

Related course material:
http://202.205.131.32/forum/upload/forum.php?mod=viewthread&tid=323

Cheng Li's biography:

Dr. Cheng Li studied computer science at Beijing Normal University (BS, 1995) and statistics at University of California, Los Angeles (PhD, 2001). He has worked at Harvard School of Public Health and Dana-Farber Cancer Institute as an Assistant Professor since 2002 and Associate Professor since 2008. Dr. Cheng Li's group has developed many novel gene expression and SNP microarray analysis and visualization methods, and implemented and maintained highly-cited genomics analysis software such as dChip and batch effect adjustment software ComBat (2600 citations). He has worked at Peking University, School of Life Sciences since 2013 and now focuses on 3D genomics experiments, analysis and applications to cancer.

Analysis and interpretation of metagenome data from microbiomes and complex microbial communities

Rohan B.H. Williams
Singapore Centre for Environmental Life Sciences Engineering (SCELSE)
Nanyang Technological University and National University of Singapore

The investigation of microbiomes and complex microbial communities is now a major research direction in biology, health and medicine. *The ability to understand the biology of such microbial consortia, and develop their translational potential, is underpinned by the computational analysis of metagenome data.*

In this tutorial we will cover the major approaches to analysing and interpreting metagenome data obtained from microbiomes and complex microbial communities.

I will commence with a short historical survey of the field, focusing on the limitations of culture techniques for isolating environmental bacteria, which drove the development of non-culture based methods such as 16S amplicon sequencing.

My major focus will be on analysis and interpretation of whole community shotgun metagenomics, obtained with short read sequencing technologies (Illumina). I will describe how these complex data can be analysed using different approaches, namely *1*) read level analysis making use of reference sequence databases; *2*) via direct use of reference genomes and *3*) by using metagenome assembly methods as a way of recovering draft genomes of the member species of the community. I will discuss both taxonomic and functional analysis of these data. Emphasis will be placed on the strengths and limitations of each approach, and the practical issues related to their use and interpretation (including command line usage examples and review of specific workflows).

I will conclude by describing several new approaches that are gaining importance, including HiC-metagenomics, long read metagenomics and single cell based methods using microfluidics.

Recommended reading:
Quince, C., Walker, A.W., Simpson, J.T., Loman, N.J., Segata, N. (2017), Shotgun metagenomics, from sampling to analysis, *Nature Biotechnology* **35**: 833-84.

Speaker biography
Rohan Williams is Head of the Integrative Analysis Unit at the Singapore Centre for Environmental Life Sciences Engineering (SCELSE), an autonomous research centre in microbial eccology and biofilm biology co-hosted by the National University of Singapore and Nanyang Technological University. Following undergraduate studies in Physics, Williams obtained PhD in Medicine from the University of New South Wales (Sydney, Australia) in 2003. From 2004-2007 he was an NHMRC Peter Doherty Fellow at UNSW and then a Group Leader at the Australian National University (2007-2011) prior to taking up his present appointment. His interests and expertise lie in statistical bioinformatics, design, analysis and interpretation of experiments using genomic technologies, systems microbiology and the analysis of complex microbial communities using multi-omics approaches.

Title：Gene regulatory network inference: from correlation to causality

Abstract:

Gene regulatory network inference aims to address the fundamental question: how gene expression is regulated? i.e., when, where, and how much a gene is controlled by the sequence of DNA bases located in the regulatory region of the gene and further interpreted by proteins called transcription factors that bind to those regions and increase or decrease gene expression. This induces a series of follow-up questions: Who is the regulators (cis-element, trans-element, cis-element interaction, trans-element interaction …)? What is the quantitative function? How the tissue/condition specificity is achieved?

In this tutorial, we will treat the gene regulatory network inference as the modeling process to find out the best data generation network (knowledge or mechanism) to explain the observed high throughput data. We will review the ideas, models, and algorithms for network inference driven by the three data generation stages:

1. High-throughput gene expression experiments
2. Chromatin immunoprecipitation-based genome-wide mapping
3. Genome-wide measurement of chromatin accessibility

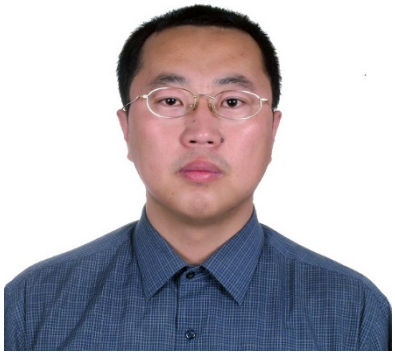
Reference

Modeling gene regulation from paired expression and chromatin accessibility data. Duren, Zhana; Chen, Xi; Jiang, Rui; Wang, Yong; Wong, Wing Hung. Proceedings of the National Academy of Sciences, vol. 114 no. 25 E4914-E4923 (2017).

Modeling the causal regulatory network by integrating chromatin accessibility and transcriptome data. Y Wang, R Jiang, WH Wong. National Science Review 3 (2), 240-251 (2016).

Y Wang, XS Zhang, L Chen. A network biology study on circadian rhythm by integrating various omics data. OMICS A Journal of Integrative Biology 13 (4), 313-324 (2009)

Y Wang, T Joshi, XS Zhang, D Xu, L Chen. Inferring gene regulatory networks from multiple microarray datasets. Bioinformatics 22 (19), 2413-2420 (2006)



Brief Bio:

Yong Wang is Professor at the Academy of Mathematics and Systems Science (AMSS), Chinese Academy of Sciences (CAS). He is also a joint faculty member in the National Center for Mathematics and Interdisciplinary Sciences (NCMIS), Chinese Academy of Sciences and Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences. He received his Ph.D. degree in Operations Research and Control Theory from AMSS of CAS in 2005, his Master's Degree in Operations Research and Control Theory from the Dalian University of Technology in 2002 and his Bachelor's Degree in Mathematics and Physics from the Inner Mongolia University in 1999. He visited the Bioinformatics Program in Boston University as a research associate. From 2010. 10 to 2011. 4, he served as the Research Staff in The Computational Biology Research Center (CBRC) of National Institute of Advanced Industrial Science and Technology (AIST) in Japan. From 2012. 11 to 2016. 2, he is a visiting scholar in Stanford University. His current interest is in Bioinformatics and Systems Biology. He developed diverse computational methods ranging from theory, model, and algorithm to elucidate the relationship between sequence variant, regulatory element, regulator, gene expression, and evolution of biomolecular systems, to probe design principles of biological regulations and networks, and to investigate systems biology mechanisms of complex traits (http://wanglab.amss.ac.cn).

Introduction to Wuhan

Wuhan, composed of the three towns of Wuchang, Hankou, and Hanyang, is the capital of Hubei Province. The three towns, separated by the Yangtze and the Han River, are linked by bridges, and because these municipalities are so closely connected by waterways, Wuhan is also called the "city on rivers." Being the largest inland port on the middle reaches of the Yangtze River and a major stop on the Beijing-Guangzhou Railway, Wuhan is one of China‘s most important hubs of water and rail transportation and communications.

Wuhan has an old history and rich cultural traditions. It began to prosper as a commercial town about two thousand years ago, when it was called Yingwuzhou (Parrot beach). From the first century to the beginning of the third century, the towns of Hanyang and Wuchang began to take shape. During the Song Dynasty (960-1279), the area became one of the most prosperous commercial centers along the Yangtze River. By the end of the Ming Dynasty (1368-1644), Hankou had become one of the four most famous cities in China. Today, Wuhan is the political, economic, and cultural center of Central China. It boasts of one of China‘s leading iron and steel complexes -- the Wuhan Iron and Steel Corporation. Wuhan is also a city with a strong revolutionary tradition.

Located in Ziyou Road, Wuchang District of Wuhan City, Hubu Xiang (户部巷) is one of the most famous streets in Wuhan with over 400 years of history. The 150 meter long street is crowded with various kinds of suppliers, including snack stalls, shopping booths and entertainment venues. It has become Wuhan's official snack street, containing 160 stores that sell 170 types of breakfast foods and snacks. In here, you can try some authentic Re Gan Mian (热干面), Fried Doupi (豆皮) and shaomai (Steamed Pork Dumplings).

Yellow Crane Tower, located near the Southern end of Yangtze River Bridge, is one of China's most famous towers. Originally used as a watch tower in ancient times, it inspired many ancient poets and artists. Well known poet Li Bai wrote during the Tang Dynasty: "My old friend, bids farewell to me in the west at Yellow Crane Tower. Amid April's mist and flowers, he goes down to Yangzhou. His distant image disappears in blue emptiness. And all I see is the long river flowing to the edge of the sky." As the foremost symbol of Wuhan, the original Yellow Crane Tower is said to exist as early as the 3rd Century. It has been destroyed and rebuilt several times. The current reconstruction was completed in 1986.

The Yangtze River Bridge called "ten thousand miles of Yangtze river bridge," the first is the first bridge across the Yangtze River Bridge. The bridge in 1955 to start in September and October 15, 1957. The bridge was formally opened in the construction of the soviet government, the experts to help bridge design and construction provides a lot of instruction. The bridge was built after three Boroughs of Wuhan even with great promote the development of Wuhan.

Introduction to Huazhong Agricultural University

Located in Wuhan, Huazhong Agricultural University (HZAU) is a national key university of "Project 211" directly under the Ministry of Education. HZAU enjoys a history of 120 years, tracing back to Hubei Farming School founded in 1898 by Zhang Zhidong, the governor of Hubei and Hunan provinces at the time.

Surrounded on three sides by tranquil lakes and backed by verdant hills, the campus covers an area of 495 hectares with well-spaced teaching blocks and lab buildings, making it an ideal place for teaching and research. HZAU consists of 18 colleges and departments, with more than 2600 faculty and staff, and over 26,000 of enrollment. Featuring life sciences and giving prominence to agricultural disciplines, HZAU has achieved coordinated development of multi-disciplines of agriculture, science, engineering, humanities, law, economics, management, art, and etc.

Over the past few years, HZAU has stabilized the annual employment rate of its college graduates at a high of around 94%. HZAU has produced notable alumni, including academicians of the two Academies, senior provincial government officials, well-known entrepreneurs and the role models of college students like Xu Benyu, who was awarded Touching China Figures for his dedication in volunteer works.

In the dawn of the new age, we as HZAUers, under our school motto “Learn and Practice; Achieve and Help Achieve” and the guideline of “educating the mind while cultivating academic atmosphere”, are striving for creating a top-ranking, world-famous research university with distinctive characteristics.

Scientific Research

Since 2011, HZAU has undertaken 4825 research projects. 129 projects have won awards on provincial and ministerial levels and above, of which 9 are national scientific and technological achievements awards. 919 projects have been granted patents, 688 of which are patents for invention and 222 are patents for utility model.

HZAU has made a number of important achievements in a variety of fields of research including plant and animal genetics and breeding, genomics and molecular biology, animal preventive veterinary medicine, pomology, horticulture and vegetable science, animal and plant disease prevention and control and safety evaluation, plant and animal production technology, food engineering and agricultural product processing, efficient use of agricultural resources and development of biomass energy, agricultural economics and land management, generating remarkable social and economic profits.


P1 West Gate
P2 International Acdemic Exchange Center
P3 Main Building
P4 Specimen
P5 Yifu Building
P6 Gymnasium
P7 Library
P8 The 24 Solar Terms column
P1
P2
P3
P4
P5
P6
P7
P8
LIBRARY

Introduction to the College of Information

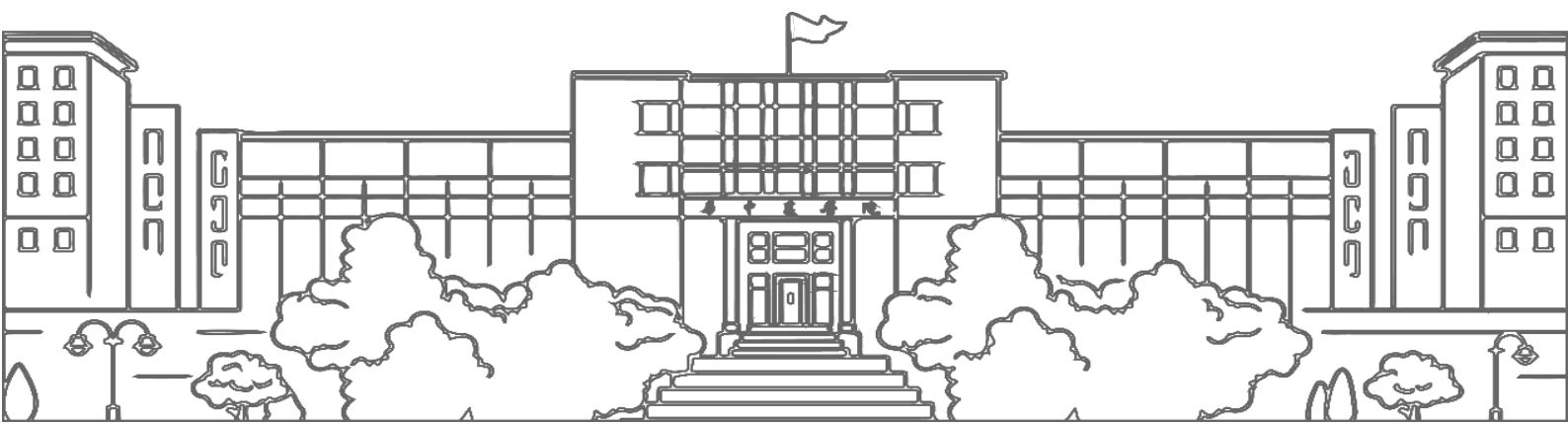
Huazhong Agricultural University (HZAU) Information College was established in April 2014. The college presently consists of four departments: Department of Computer Science, Department of Bioinformatics, Department of Data Science and Big Data Technology and Computer Public Teaching Department.

The college has built the Computer and Information Technology Experimental Teaching Center and the Key Laboratory of Agriculture Biological Information of Hubei province. The college possesses not only second-level doctoral program and master's degree in agricultural information engineering and bioinformatics, computer science and technology first-level master degree, but also Computer Science and Technology, Bioinformatics, Data Science and Big Data Technology, among which the bioinformatics major has been approved as the strategic emerging pillar industry talent training undergraduate program in Hubei province, ranking the first in the country for four times. Besides, the college owns one Innovative Group of Natural Science Foundation of Hubei Province, two undergraduate programs: the Computer Science and Technology and the Bioinformatics and one excellent young and young science and technology innovation team of Hubei's university of higher learning.

The college attaches importance to the cultivation of students' comprehensive ability and has achieved certain results。In December 2017, our students won the first prize in the application challenge competition of the fourth national mobile

Internet application development and innovation competition of colleges and universities; In March 2018, our students won the first prize in ASC18 world university supercomputer competition. Besides the IGEM team instructed by our college Professor Ma Bin Guang captured the gold award for the fourth time in International Genetically Engineered Machine Competition and received a single Best nomination.

The college adheres to the vision of running a school: "college as a college, teacher as a teacher"; holds on the idea of education: "there are students in the teachers' heart, there is a world in students' heart"; follows the path of strengthening the college: "integrate into the industrial system, join the international cycle"; and using modern information technology to promote advantageous disciplines; striving to build into a new growth point of school development.



ANNOROAD | GENOME.cn

Introduction of Annoroad Gene Tech.

Annoroad Gene Technology was established in 2012, focusing on new-generation sequencing (NGS) technology with industrial application on human medical health and life science research. It is a renowned company among China's Genome industry, and recognized as a national level high-tech enterprise. Annoroad is one of the largest providers in NGS service, with comprehensive platforms including **PromethION, Pacbio Sequel, HiSeq X-ten, NovaSeq 6000, Miseq and NextSeq 550AR.**

AnnoGene, as a registered brand under Annoroad, carries out extensive R&D collaborations with higher education institutions and research institutes worldwide, in multi-omics levels such as genomics, transcriptomics, and epigenetics, with featured products in gene expression analysis, single-cell techniques and chromosome Hi-C technology.

With the rapid development in recent years, Annoroad now has opened a branch in Yiwu, Zhejiang province under the great support of local government, remain committed to providing excellent technical solutions for life science research.

3D-Genomics

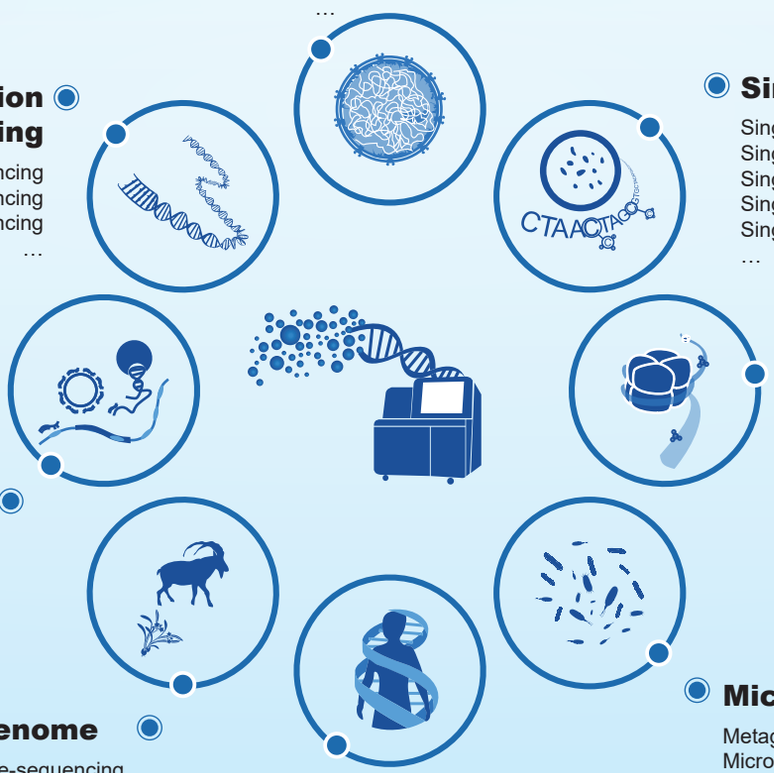
Animal Hi-C
Plant Hi-C
Single Cell Hi-C
Capture Hi-C
…

The Third Generation Sequencing

Full –Length-Isoform Sequencing
Microbial Whole Genome Sequencing
Human Whole Genome Sequencing
…

Single Cell Sequencing

Single Cell Genome Sequencing
Single Cell Exome Sequencing
Single Cell Transcriptome Sequencing
Single Cell Bisulfite Sequencing
Single Cell G&T-Seq
…



Epigenomics

WGBS MeDIP
ChIP-seq
6mA-IP seq
Target-BS …

Transcriptomics

RNA Sequencing
lncRNA Sequencing
circRNA Sequencing
…

Microbial Genome

Metagenomic Sequencing
Microbial *De novo*
Precision Meta-genomics
…

Plant & Animal Genome

Plant & Animal Genome Re-sequencing
Plant & Animal *De novo* sequencing
…

Human Genome

Whole Genome Re-sequencing
Whole Genome Exome Sequencing
Targeted Sequencing
ctDNA Sequencing
…

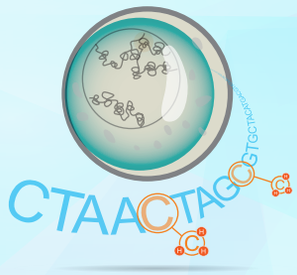
GENOME.cn
+86-4008-986-980 info@genome.cn www.genome.cn
Building B1, Yard 88, Kechuang 6 Rd, Beijing Economic-Technological Development Area, Beijing, 100176, China
annuogene

ANNOROAD | GENOME.cn

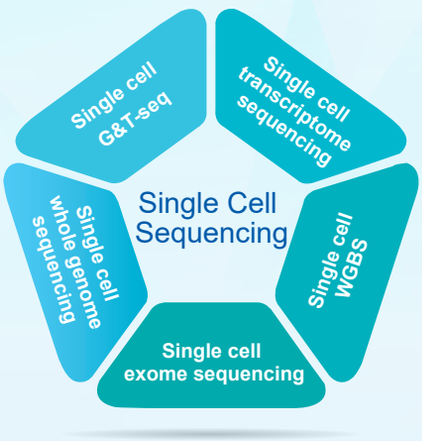
Featured products | Single Cell Sequencing&Hi-C Sequencing

Single-cell Sequencing: From “Individual” to “Cell”

Single cell sequencing solves the problems of limited samples or cell heterogeneity, it is broadly used in research of tumor mechanism, cancer diagnosis, embryonic development, stem cells, etc. Annoroad’s Single cell multi-omic services provide Singel-cell Sequencing of Genomics, Transcriptomics and Epigenomics.



Service types

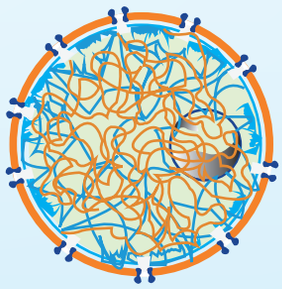


Application Areas

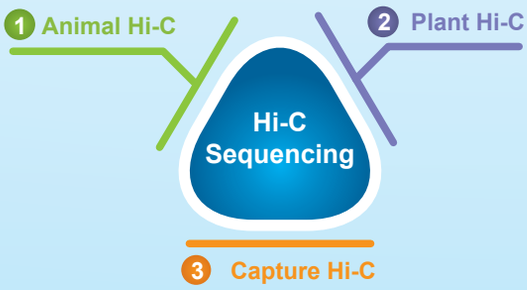
1. Cancer research
2. Embryonic development
3. Heterogeneity of cells
4. Pre-implantation genetic diagnosis(PGD)
5. Genetic recombination in germ cells
6. Single cell epigenetics

Hi-C Sequencing: Genomic Research from “1D” to “3D”

Hi-C sequencing is a method to study three-dimensional architecture of genomes and interactions of chromatin. On 2015, Annogene started to provide Hi-C and single cell Hi-C sequencing services, helped our clients discovering genomics on 3D level. The Hi-C sequencing services from Annogene captures more valid fragments, reaches higher resolution and makes it possible to study 3D genome on resolution level of 1K(Highest among reported so far) .



Service types



Application Areas

1. Studies chromatin interaction
2. Maps genome 3D structure
3. *De novo* genome assembly
4. Reveals molecular background of disease by chromatin interactions
5. Studies the biological regulatory mechanism with multi-omics

GENOME.cn
+86-4008-986-980 info@genome.cn www.genome.cn
Building B1, Yard 88, Kechuang 6 Rd, Beijing Economic-Technological Development Area, Beijing, 100176, China
annuogene

```
static {
        complTable['A'] = 'T';
        complTable['C'] = 'G';
        complTable['G'] = 'C';
        complTable['T'] = 'A';
        complTable['N'] = 'N';

        complTable['a'] = 't';
        complTable['c'] = 'g';
        complTable['g'] = 'c';
        complTable['t'] = 'a';
        complTable['n'] = 'n';
    }
```

```
private static String revComplement(String seq) {
    char[] list = new char[] { 'A', 'C', 'G', 'T', 'N', 'a', 'c', 'g', 't', 'n'};
    StringBuilder result = new StringBuilder(seq);
    result.reverse();
    for (int i = seq.length() - 1; i >= 0; i--) {
        for (int j = 0; j < list.length; j++) {
            if (result.charAt(i) == list[j]) {
                result.setCharAt(i, complTable[result.charAt(i)]);
                break;
            }
        }
    }
    return result.toString();
}
```