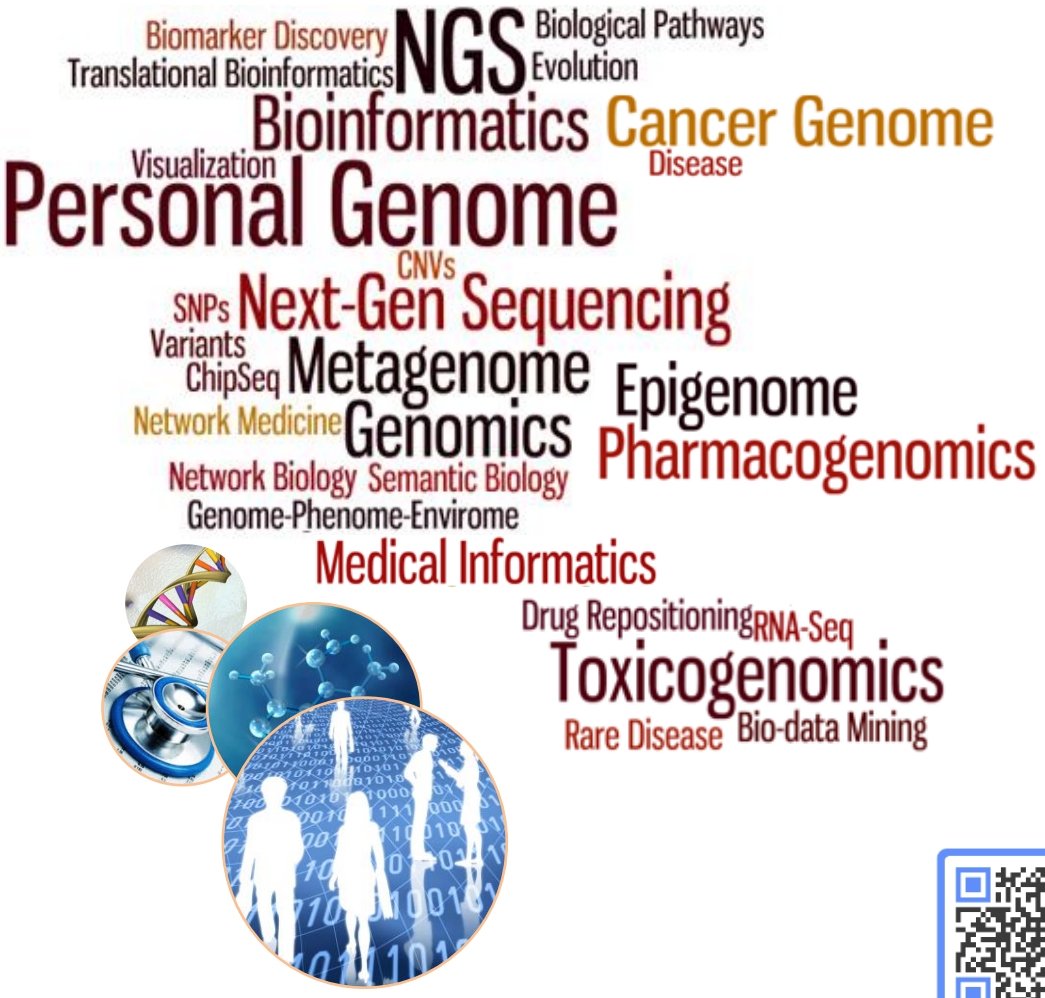
The Third Annual Translational Bioinformatics Conference

TBC 2013.10.02

Translational Bioinformatics Conference

Translational Bioinformatics



TBC 2013

TIME and | 2nd – 4TH, October, 2013
LOCATION | JW Marriott Hotel, Seoul, Korea
Conference | http://www.snubi.org/TBC2013/

The Korean Society of Medical Informatics
Systems Biomedical Informatics Research Center
TBC 2013 Organizing Committee

Table of Contents

Welcome Messages

Translational Bioinformatics Conference (TBC) will aim to highlight the multi-disciplinary nature research field and provide an opportunity to bring together and exchange ideas between translational bioinformatics researchers. TBC puts its initial emphasis on promoting translational bioinformatics research activities initiated in Asia-Pacific region. Translational bioinformatics is a rapidly emerging field of biomedical data sciences and informatics technologies that efficiently translate basic molecular, genetic, cellular, and clinical data into clinical products or health implications. Translational bioinformaticians with a mix of computer scientists, engineers, epidemiologists, physicists, statisticians, physicians and biologists come together to create the unique intellectual environment of our meeting.

Learning Objectives

Major topic areas of this year are focused on infra-technological innovations from bench to bedside, with a particular emphasis on clinical implications

- To present and exchange the latest progresses in translational bioinformatics.
- To identify the current challenges, to find research and funding opportunities, and develop future perspectives.
- To demonstrate how genomic data-driven informatics approaches can facilitate clinical research, genomic medicine, and healthcare
- To facilitate trans-disciplinary interactions among computational biology, genomics, bio-data sciences, translational medicine, and healthcare.
- To provide educational opportunities for the rapidly growing new comers.
- To develop and deploy platform for resource and problem sharing among nation-wide biomedical informatics initiatives.

Greetings

Dear Colleagues and Friends,

Along with the Organizing Committee, I am delighted to welcome you to attend the third annual Translational Bioinformatics Conference (TBC 2013) jointly hosted by ISCB-Asia in Seoul, the home place of TBC. The Human Genome Variation meeting (HGV 2013) is also aligned with TBC in a serial manner for the whole week. TBC / ISCB-Asia 2013 provides a general forum for disseminating the latest research in genomics, bioinformatics, translational research, and biomedical informatics. What a remarkable growth for the last three years!

As you may agree, 'Translational Bioinformatics' is now very well accepted throughout most of the biomedical communities. We all have closely collaborated for promoting the field of translational bioinformatics. We have decided that the fourth annual TBC 2014 will be held in Qingdao, Sandong, China hosted by the Computational Systems Biology Society of ORS China. According to our tentative plan, TBC 2015 will be again in Jeju Island, Korea and TBC 2016 may be in Goa, India. We welcome proposals for better TBCs.

Thanks to the invited speakers and presenters from all around the world to this conference, who are shaping the future of translational bioinformatics and genomics, I am sure that you will find an exciting atmosphere at TBC / ISCB-Asia 2013. All participants are the ones who shape the future of translational bioinformatics. As Alan Kay said, the best way to predict the future of translational bioinformatics is to invent it at TBC.

I wish all participants of the conference have pleasant and memorable experience. Please enjoy TBC / ISCB-Asia 2013 and the beautiful weather of Seoul.

With my best regards,

Ju Han Kim, M.D., Ph.D., M.S.

Chair, TBC / ISCB-Asia 2013 Organizing Committee

Organizing Committee Members

Ju Han Kim, M.D., Ph.D. (Korea)
Professor and Chair, Div. of Biomedical Informatics
Director, Systems Biomedical Informatics Research Center
Seoul National University College of Medicine

Atul Butte, M.D., Ph.D. (U.S.A.)
Stanford Center for Biomedical Informatics Research
Stanford University School of Medicine

Luonan Chen, Ph.D. (China)
Key Laboratory of Systems Biology,
Shanghai Institute for Biological Sciences, China

Indira Ghosh, Ph.D. (India)
Dean and Professor, School of Informatics Technology,
Jawaharlal Nehru University, New Delhi, India

Maricel Kann, Ph.D. (U.S.A.)
University of Maryland Baltimore County

Yves A. Lussier, M.D. (U.S.A.)
Director, Center for Biomedical Informatics
of the Institute of Translational Medicine, University of Chicago

Lucila Ohno-Machado, M.D., Ph.D. (U.S.A.)
Founding Chief, Division of Biomedical Informatics, UC San Diego
Director, Biomedical Research Informatics for Global Health Program

Marylyn DeRiggi Ritchie, Ph.D. (U.S.A.)
Assistant Professor, Biochemistry and Molecular Biology
Pennsylvania State University

Tomohiro Sawa, M.D., Ph.D. (Japan)
Chief Information Officer, Headquarters, Teikyo University
Dept. of Anesthesiology, Teikyo University

Sangsoo Kim, Ph.D. (Korea)
Professor & Director, School of Systems Biomedical Sciences,
Soongsil University

Youngju Kim, Ph.D. (Korea)
Principal Researcher, Genome Resource Center,
Korea Research Institute of Bioscience & Biotechnology (KRIBB)

Kiejung Park, Ph.D. (Korea)
Director, Div. of Bio-Medical Informatics
National Institute of Health, Korea

Raewoong Park, M.D., Ph.D. (Korea)
Director, Medical & Bio Informatics lab,
Dept. of Medical Informatics, School of Medicine, Ajou University

Hyunjung Shin, Ph.D. (Korea)
Professor, Ajou University Datamining Lab,
Dept. of Industrial and Information Systems Engineering

Sanghyuk Lee, Ph.D. (Korea)
Director, Korean Bioinformation Center
Professor, Dept. of Life Sciences, Ewha Womans University
Director, Ewha Research Center for Systems Biology

Hojin Choi, Ph.D. (Korea)
Professor, Dept. of Computer Science
Korea Advanced Institute of Science and Technology (KAIST)

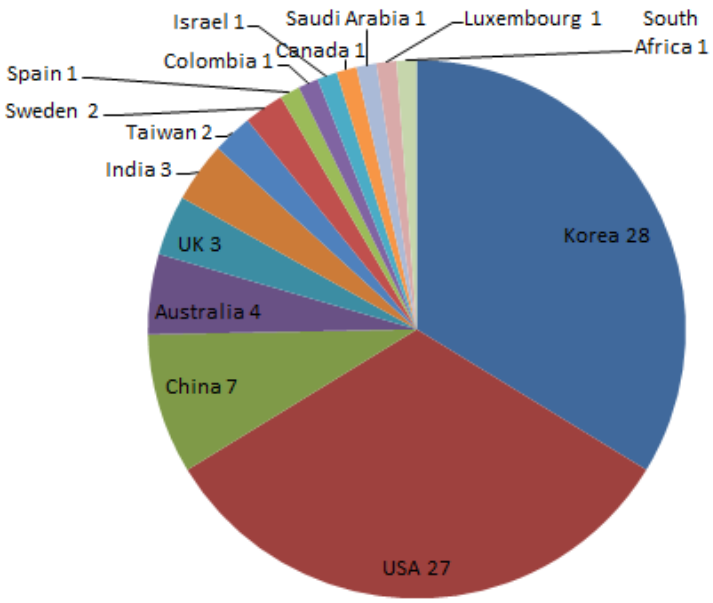
TBC 2013 Participants

The third annual Translational Bioinformatics Conference (TBC/ISCB-Asia 2013) will be held at the JW Marriot Hotel, Seoul, Korea. Seoul is the regional hub of Northeast Asia and a large city with a sizeable research community in biology, information technology and medicine. It has more than 40 universities, major research institutes and a number of companies. The world-class information technology and medical technology of Korea attracts researchers to Seoul.

A total of 15 sessions provided attendees an opportunity to share the results of translational bioinformatics. Forty-seven selected papers were presented on the cutting edge research regarding clinical applications, cancer bioinformatics, proteoimformatics, multi-omic applications, linking phenotypes, post-GWAS, biomedical big data and new technologies. A number of selected papers will be published in JAMIA, BMC Medical Genomics and other informatics journals. The program committees selected a total of 34 high quality posters that will be presented at the conference.

Participating nations

- **A total of 81 researchers from 15 nations will give talks to attendees**
- **Fifteen nations:** Korea, United States of America, China, Australia, United Kingdom, India, Taiwan, Sweden, Spain, Colombia, Israel, Canada, Saudi Arabia, Luxembourg, South Africa



Pre-Congress Tutorials

Wednesday, Oct. 2, 2013

Welcoming reception for the OC members, TBC / ISCB-Asia 2013
OC Board Meeting with Welcoming Dinner

Pre-Conference Tutorial: **Translational Bioinformatics, Real World Applications**
- Please refer to PLoS Computational Biology: Translational Bioinformatics -

Tutorial - I. 2:30PM ~ 6:30PM

- **Introduction to Translational Bioinformatics**

  Atul Butte (Stanford University)

- **Phenome-Wide Association Study (PheWAS) and GWAS for Complex Disease**

  Marylyin D. Ritchie (Penn State University)

- **Cancer Genomics and Epigenomics**

  Sun Kim (Seoul National University)

Tutorial - II. 2:30PM ~ 6:30PM

- **Disease & drugs models with ENCODE (Encyclopedia of DNA Elements)**

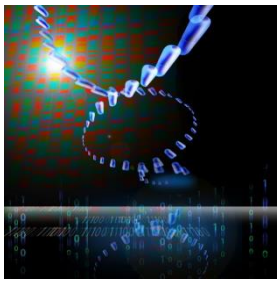
  Yves A. Lussier (University of Illinois)

- **Text-mining for Translational Bioinformatics**

  Larry Hunter (University of Colorado)

- **Molecular approaches for Translational Bioinformatics**

  Maricel G. Kann (University of Maryland)



"Translational bioinformatics is an emerging field that addresses the current challenges of integrating increasingly voluminous amounts of molecular and clinical data. Its aim is to provide a better understanding of the molecular basis of disease, which in turn will inform clinical practice and ultimately improve human health."

Program at a Glance

Day 1, Thursday, Oct. 3, 2013

	Grand Ballroom A	Grand Ballroom B	Grand Ballroom C
07:00~ 18:00	Registration Opens		
08:30~ 09:30	ISCB Open Business Meeting		
09:30~ 10:00	**Grand Opening, TBC/ISCB-Asia 2013**		
10:00~ 10:20	**Coffee Break**		
10:20~ 11:10	**Keynote I:** Translational Analytics: Semantic Computing at Genomic Scale. **Larry Hunter,** University of Colorado		
11:10~ 12:00	**Keynote II:** Translating a trillion points of data into therapies, diagnostics, and new insights into disease**.** **Atul Butte**, Stanford University		
12:00~ 13:00	**Lunch**		
Session	**S1. Clinical Applications**	**S2. Cancer Bioinformatics**	**H1. Highlight Research**
13:00~ 13:25	**S1-1** Concordance of deregulated mechanisms unveiled in underpowered experiments: PTBP1 knockdown case study *Vincent Gardeux*	**S2-1** Integrative analysis reveals disease-associated genes and biomarkers for prostate cancer progression *Yin Li*	**H1-1** Heart Attacks: Leveraging A Cardiovascular Systems Biology Strategy To Predict Future Outcomes *Carlo Vittorio Cannistraci*
13:25~ 13:50	**S1-2** Predicting different phenotypes of asthma and eczema using machine learning *Mattia Prosperi*	**S2-2** A Coupling Approach of a Predictor and a Descriptor for Breast Cancer Prognosis *Hyunjung Shin*	**H1-2** Bridging Cancer Biology with the Clinic: A Novel Personalized Prognostic Indicators for Breast cancer *Xinan Yang*
13:50~ 14:15	**S1-3** Comparison of warfarin therapy clinical outcomes following implementation of an automated mobile phone-based critical laboratory value text alert system *Yufeng Jane Tseng*	**S2-3** Identifying Potential Subtypes of Melanoma based on Pathway Activity Profiles *Sungwon Jung*	**H1-3** Interpreting individuals' genomes: Practical applications from newborn screening, and findings from CAGI 2013—the Critical Assessment of Genome Interpretation *Steven Brenner*
14:15~ 14:40	**S1-4** Automatic detection and resolution of measurement-unit conflicts in aggregated data *Soroush Samadian*	**S2-4** Identifying multi-biomarkers to distinguish malignant from benign colorectal tumors by a mixed integer programming *Luonan Chen*	
14:40~ 15:00	**Coffee Break**		
Session	**S3. Proteoinformatics**	**S4. Mult-Omic Applications**	**H2. Highlight Research**
15:00~ 15:25	**S3-1** Integrated analysis of genome-wide DNA methylation and gene expression profiles in molecular subtypes of breast cancer *Je-Keun Rhee*	**S4-1** Integrated Analysis of microRNA-target Interactions with Clinical Outcomes for Cancers *Je-Gun Joung*	**H2-1** The SADI Personal Health Lens: A Web Browser-Based System for Identifying Personally Relevant Drug Interactions *Soroush Samadian*
15:25~ 15:50	**S3-2** Prediction of C-peptide Like Family using Multiple Predictive Models and Feature Encodings *Junbeom Kim*	**S4-2** "N-of-1-pathways" unveils personal deregulated mechanisms from a single pair of RNA-seq samples: towards precision *Vincent Gardeux*	**H2-2** Correlation network-guided novel key gene identification *Feng He*
15:50~ 16:15	**S3-3** Derivative Component Analysis for Mass Spectral Serum Proteomic Profiles *Henry Han*	**S4-3** Knowledge Boosting: A graph-based integration with multi-omics data and genomic knowledge for cancer clinical outcome prediction *Dokyoon Kim*	**H2-3** Imbalanced network biomarkers for traditional Chinese medicine Syndrome in gastritis patients *Rui Li*
16:15~ 16:45	**Coffee Break**		
16:45~ 17:35	**Keynote III:** The function and evolution of structural genetic variants in human phenotypic variability. **Charles Lee,** Harvard Medical School		

Program at a Glance

17:35~ 18:30	**Keynote IV:** Sharing and connecting knowledge, rather than data. **Anthony Brookes,** University of Leicester
18:30~	**Conference Dinner**

Day 2, Friday, Oct. 4, 2013

	Grand Ballroom A	Grand Ballroom B	Grand Ballroom C
Session	**S5. Linking Phenotypes**	**H3. Highlight Research**	**ISCB Scientific Session**
08:30~ 08:55	**S5-1** The Multiscale Backbone of the Human Phenotype Network based on Biological Pathways Network *Christian Darabos*	**H3-1** Computational Studies of Ubiqutin and Ubiquitin-like Conjugation *Gao Tianshun*	**I-1** Balanced Nucleo-cytosolic Partitioning Defines a Spatial Network for Coordination of Circadian Physiology in Plants *Daehee Hwang*
08:55~ 09:20	**S5-2** Relative impact of multi-layered genomic data on gene expression phenotypes in serous ovarian tumors *Kyung-Ah Sohn*	**H3-2** Mechanisms of PDGFRα promiscuity and PDGFRβ specificity in association with PDGFB *Janneth González*	**I-2** NGS sequence analysis for regulation and epigenomics *Tim Bailey*
09:20~ 09:45	**S5-3** ATHENA: Identifying interactions between different levels of genomic data associated with cancer clinical outcomes using grammatical evolution neural network *Dokyoon Kim*	**H3-3** DisplayHTS: a R package for visualizing high-throughput screening data and results *Xiaohua Zhang*	**I-3** Statistical significance of combinatorial regulations *Jun Sese*
09:45~ 10:10	**S5-4** Topological Analysis of Statistical Epistasis Networks Reveals Pathways Associated with Alzheimer's Disease *Qinxin Pan*	**H3-4** Methylerythritol phosphate pathway to isoprenoids: Kinetic modeling and in silico enzyme inhibitions in P. falciparum *Indira Ghosh*	**I-4** An integrative characterization of recurrent molecular aberrations in glioblastoma genomes *Chen-Hsiang Yeang*
10:10~ 10:40	**Coffee Break**		
10:40~ 11:30	**Keynote V:** Personalome: clinically actionable 'Omics. **Yves Lussier,** University of Illinois at Chicago		
11:30~ 12:20	**Keynote VI:** Translational bioinformatics for identification of prostate cancer biomarkers. **Bairong Shen,** Soochow University		
12:20~ 13:20	**Lunch**		
Session	**S6. Post-GWAS**	**S7. Biomedical Big Data**	**H4. Highlight Research**
13:20~ 13:45	**S6-1** Practical issues for screening and variable selection method in a Genome-Wide Association Analysis *Yong-gang Kim*	**S7-1** Health Monitoring System based on Lifelog Analysis *Yongjin Kwon*	**H4-1** A protein domain-centric approach for the comparative analysis of human and yeast phenotypically relevant mutations *Maricel Kann*
13:45~ 14:10	**S6-2** IGENT: Efficient Entropy based Algorithm for Detecting Genome-wide Gene-Gene Interaction Analysis *Min-Seok Kwon*	**S7-2** Differentially Private Distributed Logistic Regression using Public and Private Biomedical Datasets *Xiaoqian Jiang*	**H4-2** Yin and Yang of reciprocally scale-free biological networks between lethal genes and disease genes *Ju Han Kim*
14:10~ 14:35	**S6-3** Identification of novel therapeutics for complex diseases from genome-wide association data *Mani P. Grover*	**S7-3** Effectively processing medical term queries on the UMLS Metathesaurus by Layered Dynamic Programming *Yang Xiang*	**H4-3** Comparative analysis using K-mer and K-flank patterns provides evidence for CpG island sequence evolution in mammalian genomes *Heejoon Chae*
14:35~ 14:50	**Coffee Break**		
Session	**S8. New Technologies**	**H5. Highlight Research**	**H6. Highlight Research**
14:50~ 15:15	**S8-1** GAMUT: GPU Accelerated MicroRNA analysis to Uncover Target genes through CUDA-miRanda *Xiaoqian Jiang*	**H5-1** Arpeggio: harmonic compression of ChIP-seq data reveals protein-chromatin interaction signatures *Kelly Stanton*	**H6-1** Statistical epistasis networks reduce the computational complexity of searching three-locus genetic models *Jason H. Moore*

Program at a Glance

15:15~ 15:40	**S8-2** A Novel Multi-scale Visualization Software for Data-driven Biomedical Data Exploration *Gang Su*	**H5-2** TrAp: a tree approach for fingerprinting subclonal tumor composition *Yuval Kluger*	**H6-3** PhenDisco (Phenotype Discoverer): a New Information Retrieval System for the database of Genotypes and Phenotypes *Hyeoneui Kim*
15:40~ 16:05	**S8-3** In Silico Cancer Cell versus Stroma cellularity index computed from species-specific human and mouse transcriptome of xenograft models: towards accurate stroma targeting therapy assessment *Xinan Yang*	**H5-3** Variants Affecting Exon Skipping Contribute to Complex Traits *Younghee Lee*	**H6-2** Phenome-Wide Association Study (PheWAS) for Detection of Pleiotropy within the Population Architecture using Genomics and Epidemiology (PAGE) Network *Sarah Pendergrass*
16:05~ 16:20	**Coffee Break**		
16:20~ 17:10	**Keynote VII:** Distal enhancers and their role in mediating genotype-phenotype associations. **Sridhar Hannenhalli,** University of Maryland		
17:10~ 18:00	**Keynote VIII:** Co-methylation. **Terry Speed,** UC Berkeley		
18:00~ 18:30	Special Remark on Translational Bioinformatics Paper Publications and TBC 2011 ~ 2013, by the Chief Editor of JAMIA **Lucila Ohno-Machado,** UC San Diego		
18:30~	**Closing Ceremony with Traditional TBC Lottery**		

Keynote Speakers

10:20-11:10 (Wednesday, Oct. 3)

Larry Hunter

University of Colorado

Translational analytics: Semantic computing at genomic scale

Cheap sequencing technology has transformed biomedical research, but translating the results of genome-scale assays into advances in medical care remains difficult. A key challenge is contextualizing genome-scale experimental results in light of all relevant prior knowledge. However, knowledge of biomedicine is very large, heterogeneous in character, and often accessible only in the texts of journal articles. To overcome these challenges, computational tools need to be designed to address several mutually constraining requirements: representing and reasoning about the knowledge relevant to an experimental result (data analytics), building the relevant knowledge-base from databases and the literature (text analytics), and presenting the right amount of information in the right way and at the right time to many different kinds of users (visual analytics). These analytics depend on semantic computing approaches, which do not necessarily map well onto existing supercomputing architectures. Furthermore, unlike for physicians, characterization of the information gathering needs of translational researchers remains poor. I will present some illustrative examples of translational analytics in action, and speculate about potentially productive approaches to addressing these challenges.

11:10-12:00 (Wednesday, Oct. 3)

Atul Butte

Stanford University

Translating a trillion points of data into therapies, diagnostics, and new insights into disease

There is an urgent need to translate genome-era discoveries into clinical utility, but the difficulties in making bench-to-bedside translations have been well described. The nascent field of translational bioinformatics may help. Dr. Butte's lab at Stanford builds and applies tools that convert more than a trillion points of molecular, clinical, and epidemiological data -- measured by researchers and clinicians over the past decade -- into diagnostics, therapeutics, and new insights into disease. Dr. Butte, a bioinformatician and pediatric endocrinologist, will highlight his lab's work on using publicly-available molecular measurements to find new uses for drugs including drug repositioning for inflammatory bowel disease, discovering new treatable inflammatory mechanisms of disease in type 2 diabetes, and the evaluation of patients presenting with whole genomes sequenced.

Keynote Speakers

16:45-17:35 (Wednesday, Oct. 3)

Chalres Lee

Harvard Medical School

The function and evolution of structural genetic variants in human phenotypic variability

17:35-18:30 (Wednesday, Oct. 3)

Anthony Brookes

University of Leicester, United Kingdom

Sharing and connecting knowledge, rather than data.

Data alone (i.e., without full context and consent) are virtually 'useless', and carry many risks and complications - and yet extensive discussion and many efforts are underway to try to promote ubiquitous data sharing. Truly fundamental problems with widespread data sharing are perhaps being under-estimated. For example, by their very nature many sorts of omics data cannot be anonymised, and so directly allow the identification of research subjects who were promised privacy. Data also have strategic value in terms of collaborative discussions and career progression. Perhaps this focus on data sharing is missing the point - in that the 'knowledge' revealed by the data is the truly 'useful' commodity, as it alone allows one to make predictions from observations. Critically, as well as being more 'useful', knowledge is also far simpler to deal with in terms of ownership, sharing and exploitation. So we need to think beyond the 'data sharing' roadblock, and embrace new ideas such as data 'discovery', remote data analysis, data integration portals, automated authorisation, and database-journals. In short, we must focus on converting the data sharing challenge to a knowledge use imperative. Details and examples will be given.

Keynote Speakers

10:40-11:30 (Thursday, Oct. 4)

Yves A. Lussier

University of Illinois at Chicago

Personalome: clinically actionable 'Omics

This presentation will highlight how multiple scales and analytes of deregulated 'omics measures can be interpreted on a single patient basis, combined and prioritized for predicting clinical outcome. While identification of deregulated polymorphisms are readily assessed on personal DNA-sequencing data, such signal derived over personal transcriptomes remains an unmet challenge. Further the mechanistic interpretation of non-coding polymorphisms and of transcriptomes to yield deregulated personal is also mostly unexplored. We will review leading edge approaches solving these issues using ENCODE data as well as strategies to derive interpretable and clinically actionable personal mechanism profiles. While the post-genome era lasted a decade, the post-ENCODE period is poised to rapidly be trumped by the advent of the "personalome" powering precision therapy.

11:30-12:20 (Thursday, Oct. 4)

Bairong Shen

Soochow University, China

Translational bioinformatics for identification of prostate cancer biomarker

With the accumulation of high throughput biological data, especially the next generation sequencing data, it becomes possible to integrate these biological data with medical data to identify important molecular features for the early diagnosis, prognosis and treatment of complex diseases. But for the integration and modeling, we may face many challenges, such as 1) the modeling and simulation of the heterogeneous data at a systems network level; 2) the collection of paired biological and personalized medical data to reconstruct personalized models for the precise diagnosis, prognosis and treatment of complex diseases. To take the advantage of these biomedical data and overcome these challenges as well as promote the application of informatics to translational research, we take prostate cancer as a case study and applied novel bioinformatics methods to the integration and analysis of prostate cancer associated data which include gene expression profiling, micro-mRNA interaction and protein-protein interaction, to identify the putative prostate cancer biomarkers for diagnosis and prognosis, the identified biomarkers were further validated by experimental or computational analysis. We concluded with our findings and the future perspectives on translational prostate cancer research.

Keynote Speakers

16:20-17:10 (Thursday, Oct. 4)

Sridhar Hannenhalli

University of Maryland

Distal enhancers and their role in mediating genotype-phenotype associations

The ENCODE project, via generation of unprecedented transcriptomic and epigenomic profiles, has revealed a complex layer of transcriptional regulation mediated by distal regulatory enhancers distributed throughout the human genome. These data open up more questions than they answer. I will talk about our attempts to identify distal enhancers based on epigenomic marks, and then present our ongoing efforts to predict the gene targets of the distal enhancers. The broader context and the motivation for these studies are to interpret eQTL and GWAS results in light of enhancer-mediated regulatory networks. I will then present a first look at correlated networks of enhancers in the human genome.

17:10-18:00 (Thursday, Oct. 4)

Terry Speed

University of California, Berkeley

Co-methylation

CpG methylation is a mitotically heritable epigenetic mark on DNA which plays a key role in genomic imprinting, X-inactivation, transcriptional regulation, tissue specificity and carcinogenesis. What is co-methylation? Loosely, it is the persistence of the methylated (M) or unmethylated (U) state along a chromosome. Slightly more precisely, it is the association between the methylation state at nearby CpGs, as a function of their separation. "Co-" here is meant to bring to mind correlation. This talk will summarize some results concerning co-methylation we obtained by analysing publicly available sequence data on whole genome bisuplhite-treated DNA. Our immediate goal was to see whether we can simulate whole-genome methylation data that is indistinguishable from the real thing. I'll explain why we want to do this. It turns out to be quite hard (for us).

ISCB Scientific Session

ISCB Scientific Session

Room: Grand Ballroom C
Date: Friday, Oct. 4, 08:30 - 10:10

I-1: Balanced Nucleo-cytosolic Partitioning Defines a Spatial Network for Coordination of Circadian Physiology in Plants

Daehee Hwang, *Posttech, Korea*

Abstract

Biological networks consist of a defined set of regulatory motifs. Subcellular compartmentalization of regulatory molecules can provide a further dimension in implementing regulatory motifs. However, spatial regulatory motifs and their roles in biological network have rarely been explored. Here, we show, using experimentation and mathematical modeling, that spatial segregation of GIGANTEA (GI), a critical component of plant circadian systems, into nuclear and cytosolic compartments leads to differential functions as positive and negative regulators of the circadian core gene, LHY, forming an incoherent feedforward loop to regulate LHY. This regulatory motif formed by nucleo-cytoplasmic partitioning of GI confers, through the balanced operation of the nuclear and cytosolic GI, strong rhythmicity and robustness to external and internal noises to the circadian system. Our results show that spatial and functional segregation of a single molecule species into different cellular compartments provides a unique means to extend the regulatory capabilities of biological networks.

I-2: NGS sequence analysis for regulation and epigenomics

Tim Bailey, *University of Queensland, Australia*

I-3: Revisiting statistical significance for finding combinatorial effects.

Jun Sese, *Tokyo Insitute of Technology, Japan*

Abstract

To understand complex associations between genotypes and phenotypes, a first step is to list up statistically significant combinations of the features. However, the discovery is not only computationally non-trivial but also extremely unlikely due to multiple testing correction. The exponential growth of the number of tests forces us to set a strict limit to the maximum arity. In this talk, we introduce an efficient branch-and-bound algorithm named Limitless Arity Multiple testing Procedure (LAMP) to count the exact number of testable combinations and calibrate the Bonferroni factor to the smallest possible value. LAMP lists up significant combinations without any limit, while the family-wise error rate is rigorously controlled under the threshold. We applied LAMP to the discovery of combinatorial regulations of transcription factors. From human breast cancer transcriptome, LAMP discovered statistically significant combinations of as many as eight binding motifs. This method may contribute to uncover pathways regulated in a coordinated fashion and find hidden associations in heterogeneous data.

I-4: An integrative characterization of recurrent molecular aberrations in glioblastoma genomes

Chen-Hsiang Yeang, *Academica Sinica, Taiwan*

Abstract

Glioblastoma multiforme (GBM) is the most common and malignant primary brain tumor in adults. Decades of investigations and the recent effort of the Cancer Genome Atlas (TCGA) project have mapped many molecular alterations in GBM cells. Alterations on DNAs may dysregulate gene expressions and drive malignancy of tumors. It is thus important to uncover causal and statistical dependency between "effector" molecular aberrations and "target" gene expressions in GBMs. A rich collection of prior studies attempted to combine copy number variation (CNV) and mRNA expression data. However, systematic methods to integrate multiple types of cancer genomic data -- gene mutations, single nucleotide polymorphisms, copy number variations, DNA methylations, mRNA and microRNA expressions, and clinical information -- are relatively scarce.

We proposed an algorithm to build "association modules" linking effector molecular aberrations and target gene expressions and applied the module-finding algorithm to the integrated TCGA GBM datasets. The inferred association modules were validated by six tests using external information and datasets of central nervous system tumors: (1)indication of prognostic effects among patients, (2)coherence of target gene expressions, (3)retention of effector-target associations in external datasets, (4)recurrence of effector molecular aberrations in GBM, (5)functional enrichment of target genes, and (6)co-citations between effectors and targets. Modules associated with well-known molecular aberrations of GBM -- such as chromosome 7 amplifications, chromosome 10 deletions, EGFR and NF1 mutations -- passed the majority of the validation tests. Furthermore, several modules associated with less well-reported molecular aberrations -- such as chromosome 11 CNVs, CD40, PLXNB1 and GSTM1 methylations, and mir-21 expressions -- were also validated by external information. In particular, modules constituting trans-acting effects with chromosome 11 CNVs and cis-acting effects with chromosome 10 CNVs manifested strong negative and positive associations with survival times in brain tumors. By aligning the information of association modules with the established GBM subclasses based on transcription or methylation levels, we found each subclass possessed multiple concurrent molecular aberrations. Furthermore, the joint molecular characteristics derived from 16 association modules had prognostic power not explained away by the strong biomarker of CpG island methylator phenotypes. Functional and survival analyses indicated that immune/inflammatory responses and epithelial-mesenchymal transitions were among the most important determining processes of prognosis. Finally, we demonstrated that certain molecular aberrations uniquely recurred in GBM but were relatively rare in non-GBM glioma cells. These results justify the utility of an integrative analysis on cancer genomes and provide testable characterizations of driver aberration events in GBM.

Scientific Paper Sessions

S1. Clinical Application

Room: Grand Ballroom A
Date: Thursday, Oct. 3, 13:00 - 14:20

S1-1: Concordance of deregulated mechanisms unveiled in underpowered experiments: PTBP1 knockdown case study.

Vincent Gardeux[1,2,3,§], Ahmet Dirim Arslan[4,5,§], Ikbel Achour[1,2,§], Tsui-Ting Ho[4,6,§], William T. Beck[4,10,*], Yves A. Lussier[1,2,4,7,8,9,10,*]

1 Institute for Translational Health Informatics, University of Illinois at Chicago, Illinois, USA.
2 Department of Medicine, University of Illinois at Chicago, Chicago, Illinois, USA.
3 Department of Informatics, School of Engineering, EISTI (Ecole Internationale des Sciences du Traitement de l'Information), Cergy-Pontoise, France.
4 Department of Biopharmaceutical Science, College of Pharmacy, University of Illinois at Chicago, Ill., USA.
5 Robert H. Lurie Comprehensive Cancer Center, Northwestern University, Chicago, Illinois, USA.
6 Cancer Institute, University of Mississippi Medical Center, Jackson, Mississippi, USA.
7 Department of Bioengineering, University of Illinois at Chicago, Chicago, Illinois, USA.
8 Computation Inst. & Inst. For Genomics & Systems Biol, Argonne National Lab. & Un. of Chicago, Ill., USA.
9 Institute for Personalized Respiratory Medicine, University of Illinois at Chicago, Illinois, USA.
10 University of Illinois Cancer Center, Chicago, IL, USA.
§ Equal contribution

Abstract

Background: Genome-wide transcriptome profiling generated by microarray and RNA-Seq often provides deregulated genes or pathways applicable only to larger cohort. On the other hand, individualized interpretation of transcriptomes is increasely pursued to improve diagnosis, prognosis, and patient treatment processes. Yet, robust and accurate methods based on a single paired-sample remain an unmet challenge.

Method: "N-of-1-pathways" translates gene expression data profiles into mechanism-level profiles on single pairs of samples (one p-value per geneset). It relies on three principles: I)statistical universe is a single paired sample, which serves as its own control; ii) statistics can be derived from multiple gene expression measures. We analyzed deregulated mechanisms associated with the depletion of the alternative splicing protein, PTBP1. Using a single paired neuronal cell line RNA-Seq transcriptomes (Gold Standard), our method predicts mechanisms that were compared to those of breast and ovarian cancer cell lines (mRNA expression microarray data).

Results: N-of-1-pathways predictions outperform those of GSEA and Differentially Expressed Genes enrichment (DEG-enrichment), within- and cross-datasets. N-of-1-pathways uncovered concordant PTBP1- dependent mechanisms across datasets (Odds-Ratios >= 13, p-values <= 1x10-5), such as RNA splicing and cell cycle. In addition, it unveils tissue-specific mechanisms of alternatively transcribed PTBP1-dependent genesets. Furthermore, we demonstrate that GSEA and DEG-Enrichment preclude accurate analysis on single paired samples.

Conclusion: N-of-1-pathways enables robust and biologically relevant mechanism-level classifiers with small cohorts and one single paired samples that surpasses conventional methods. Further, it identifies unique sample/ patient mechanisms, a requirement for precision medicine.

Software: http://Lussierlab.org/publication/N-of-1-pathways.

S1-2: Predicting different phenotypes of asthma and eczema using machine learning

Mattia C.F. Prosperi[1,2,*], Susana Marinho[2], Angela Simpson[2], Iain Buchan[1], Adnan Custovic[2]

1 Centre for Health Informatics, Institute of Population Health, Faculty of Medical and Human Sciences, University of Manchester, Manchester, United Kingdom
2 Centre for Respiratory Medicine and Allergy, Institute of Inflammation and Repair, University of Manchester, Manchester, United Kingdom

Abstract

Asthma is the most common chronic disease in the developed countries, with a relatively modest drug armamentarium. There is increasing recognition that asthma is a heterogeneous disease with similar clinical manifestations (phenotypes), but different underlying pathophysiological causes (endotypes).

We investigate here the predictive ability of linear/non-linear machine learning models (from logistic regression to random forests, validated via extra-sample bootstrapping) in an unselected population, with respect to different operational definitions of asthma, wheeze, and eczema, using a large heterogeneous set of attributes (demographic, clinical, laboratory features, genetic profiles, environmental exposures). The aim is to identify to which extent such heterogeneous information contributes and combines towards specific clinical manifestations.

Our study population included 554 adults, 42% male, 38% previous or current smokers. Proportion of asthma, wheeze, and eczema diagnoses was 16.7%, 12.3%, and 21.7%, respectively. Models were fit on 223 non-genetic variables plus 215 single nucleotide polymorphisms. In general non-linear models achieved a better sensitivity/specificity trade-off as compared to other methods, more markedly when considering asthma and wheeze, less with respect to eczema (area under the curve 84%, 76% and 64%, respectively). Findings confirm the relevant contribution of allergen sensitisation combined with lung function markers (but not for eczema, for which new predictors like whole body impedance are found). Predictive ability of genetic markers alone is limited.

Looking forward to a longitudinal extension as well as increasing the amount of information processed, this study marks the grounds for a better understanding of disease mechanisms towards the development of personalized diagnostic tools.

S1-3: Comparison of warfarin therapy clinical outcomes following implementation of an automated mobile phone-based critical laboratory value text alert system

Shu-Wen Lin[1,2,3], Wen-Yi Kang[4], Dong-Tsamn Lin[5], James Chao-Shen Lee[6], Fe-Lin Lin Wu[1,2,3], Chuen-Liang Chen[7], Yufeng J. Tseng[3,4,7,*]

1 Graduate Institute of Clinical Pharmacy, College of Medicine, National Taiwan University
2 School of Pharmacy, College of Medicine, National Taiwan University
3 Department of Pharmacy, National Taiwan University Hospital
4 Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University
5 Department of Pediatrics and Laboratory Medicine, College of Medicine, National Taiwan University
6 Department of Pharmacy Practice, College of Pharmacy, University of Illinois at Chicago
7 Department of Computer Science and Information Engineering, National Taiwan University

Abstract

Objective: To evaluate clinical outcomes of patients on warfarin therapy following implementation of a Personal Handy-phone System-based (PHS) alert system capable of generating and delivering text messages to communicate critical prothrombin time (PT) / international normalized ratio (INR) laboratory results to practitioners' mobile phones in a large tertiary teaching hospital.

Design: A retrospective analysis was performed comparing patient clinical outcomes and physician prescribing behavior following conversion from a manual laboratory result alert system to an automated system.

Measurements: Clinical outcomes and practitioner responses to both alert systems were compared. Complications to warfarin therapy, warfarin utilization, and PT/INR results were evaluated for both systems, as well as clinician time to read alert messages, time to warfarin therapy modification, and monitoring frequency.

Results: No significant differences were detected in major hemorrhage and thromboembolism, warfarin prescribing patterns, PT/INR results, warfarin therapy modification, or monitoring frequency following implementation of the PHS text alert system. In both study periods, approximately 80% of critical results led to warfarin discontinuation or dose reduction. Senior physicians' follow-up response time to critical results was significantly decreased in the PHS alert study period compared to the manual notification study period (P=0.015). No difference in follow-up response time was detected for junior physicians.

Conclusions: Implementation of an automated PHS-based text alert system did not adversely impact clinical or safety outcomes of patients on warfarin therapy. Approximately 80% immediate recognition of text alerts was achieved. The potential benefits of an automated PHS alert for senior physicians were demonstrated.

S1-4: Automatic detection and resolution of measurement-unit conflicts in aggregated data

Soroush Samadian[1], Bruce McManus[1] and Mark Wilkinson[2]

1 UBC James Hogg Research Center, Institute for Heart + Lung Health, Room 166 - 1081 Burrard Street, St. Paul's Hospital Vancouver, BC, Canada, V6Z 1Y6

2 Centro de Biotecnolog⁄a y Genómica de Plantas, Universidad Politécnica de Madrid, Madrid, España

Abstract

Motivation: Measurement-unit conflicts are a perennial problem in integrative research domains such as clinical meta-analysis. As multi-national collaborations grow, as new measurement instruments appear, and as Linked Open Data infrastructures become increasingly pervasive, the number of such conflicts will similarly increase. We propose a generic approach to the problem of (a) encoding measurement units in datasets in a machine-readable manner, (b) detecting when a dataset contained mixtures of measurement units, and (c) automatically converting any conflicting units into a desired unit, as defined for a given study.

Results: We utilized existing ontologies and standards for scientific data representation, measurement unit definition, and data manipulation to build a simple and flexible Semantic Web Service-based approach to measurement-unit harmonization. A cardiovascular patient cohort in which clinical measurements were recorded in a number of different units (e.g., mmHg and cmHg for blood pressure) was automatically classified into a number of clinical phenotypes, semantically defined using different measurement units.

Conclusion: We demonstrate that through a combination of semantic standards and frameworks, unit integration problems can be automatically detected and resolved.

Scientific Paper Sessions

S2. Cancer Bioinformatics

Room: Grand Ballroom B
Date: Thursday, Oct. 3, 13:00 - 14:20

S2-1: Integrative analysis reveals disease-associated genes and biomarkers for prostate cancer progression

Yin Li[1,§], Wanwipa Vongsangnak[1,§], Luonan Chen[2], Bairong Shen[1,*]

1 Center for Systems Biology, Soochow University, Suzhou, 215006, China
2 Key Laboratory of Systems Biology, Chinese Academy of Sciences, Shanghai, 200031, China
§ Co-first authors

Abstract

Background Prostate cancer is one of the most common complex diseases with high leading cause of death in men. Identification of prostate cancer associated genes and biomarkers is thus essential as it can gain insights into the mechanisms underlying disease progression and advancing for early diagnosis and developing effective therapies.

Methods In this study, we presented an integrative analysis of gene expression profiling and protein interaction network at systematic level to reveal candidate disease-associated genes and biomarkers for prostate cancer progression. We first reconstructed the human prostate cancer protein-protein interaction network (HPC- PPIN) and then the network was integrative analyzed with the prostate cancer gene expression data to identify modules related to different phases in prostate cancer. At last, the candidate module biomarker was validated by its predictive ability of prostate cancer progression.

Results Different phases-specific modules were identified for prostate cancer. Among these modules, transcription Androgen Receptor (AR) nuclear signaling and Epidermal Growth Factor Receptor (EGFR) signaling pathway were shown to be the pathway targets for prostate cancer progression. The identified candidate disease-associated genes showed better predictive ability of prostate cancer progression than those of published biomarkers. In context of functional enrichment analysis, interestingly candidate disease- associated genes were enriched in the nucleus and different functions were encoded for potential transcription factors, for examples key players as AR, Myc, ESR1 and hidden player as Sp1 which were considered as potential biomarkers for prostate cancer.

Conclusions The successful results on prostate cancer samples demonstrated that the integrative analysis is powerful and useful approach to detect candidate disease-associate genes and modules which can be used as the potential biomarkers for prostate cancer progression. The data, tools and supplementary files for this integrative analysis are deposited at http://www.ibio-cn.org/HPC- PPIN/.

S2-2: A Coupling Approach of a Predictor and a Descriptor for Breast Cancer Prognosis

Hyunjung Shin[1,*] and Yonghyun Nam[1]

1 Department of Industrial Engineering, Ajou University, Wonchun-dong, Yeongtong-gu, Suwon 443-749, South Korea

Abstract

Background In cancer prognosis research, diverse machine learning models have applied to the problems of cancer susceptibility (risk assessment), cancer recurrence (redevelopment of cancer after resolution), and cancer survivability, regarding an accuracy (or an AUC--the area under the ROC curve) as a primary measurement for the performance evaluation of the models. However, in order to help medical specialists to establish a treatment plan by using the predicted output of a model, it is more pragmatic to elucidate which variables (markers) have most significantly influenced to the resulting outcome of cancer or which patients show a similar patterns.

Proposed Method In this study, a coupling approach of two sub-modules--a predictor and a descriptor--is proposed. The predictor module generates the predicted output for the cancer outcome. Semi-supervised learning Co-training algorithm is employed as a predictor. On the other hand, the descriptor module post- processes the results of the predictor module, mainly focusing on which variables are more highly or less significantly ranked when describing the results of the prediction, and how patients are segmented into several groups according to the trait of common patterns among them. Decision trees are used as a descriptor.

Results The proposed approach, 'predictor-descriptor', was tested on the breast cancer survivability problem based on the surveillance, epidemiology, and end results database for breast cancer (SEER). The results present the performance comparison among the established machine leaning algorithms, the ranks of the prognosis elements for breast cancer, and patient segments.

S2-3: Identifying Potential Subtypes of Melanoma based on Pathway Activity Profiles

Sungwon Jung[1], Seungchan Kim[1]

1 Integrated Cancer Genomics Division, Translational Genomics Research Institute, 445 North 5th Street, Phoenix, Arizona 85004, USA

Abstract

Identifying subtypes of complex diseases such as cancer is the very first step toward developing highly customized therapeutics on such diseases, as their origins significantly vary even with similar physiological characteristics. There have been many studies to recognize subtypes of various cancer based on genomic signatures, and most of them rely on approaches based on the signatures or features developed from individual genes. However, the idea of network-driven activities of biological functions has gained a lot of interests, as more evidence is found that biological systems can show highly diverse activity patterns because genes can interact differentially across specific molecular contexts. In this study, we proposed a method to compute the dissimilarity between two patient samples based on their pathway profiles, where pathway profiles are evaluated by computing the likelihoods of genetic networks and silenced interactions within pathways. By using the proposed dissimilarity measure between sample pathway profiles in clustering melanoma gene expression data, we identified two potential subtypes of melanoma with distinguished pathway profiles, where the two groups of patients showed significantly different survival patterns. We also investigated selected pathways with distinguished activity patterns between the two groups, and the result suggests hypotheses on the mechanisms driving the two potential subtypes.

S2-4: Identifying multi-biomarker to distinguish malignant from benign colorectal tumours by a mixed integer programming

Meng Zou[1,§], Peng-Jun Zhang[2,§], Xin-Yu Wen[2], Luonan Chen[3,*], Ya-Ping Tian[2,*] and Yong Wang[1,*]

1 National Center for Mathematics and Interdisciplinary Sciences, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100080, China.
2 Department of Clinical Biochemistry, Chinese PLA General Hospital, Beijing, 100853, China
3 Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200233, China
§ Joint first author

Abstract

Biomarkers serve as useful tools to aid in the early diagnosis and ultimately battle the complex disease. For many malignancies, multi-biomarker from clinical data plays an important role in patient management and has been actively studied. Specifically, serum-based diagnosis to distinguish colorectal cancers (CRC) from benign colorectal tumours is very challenging.

Here, we develop a novel mixed integer programming based multi-biomarker diagnostic method. This method allows us to select the best subset of clinical markers by maximizing the accuracy to distinguish case and control samples given the number of selected biomarkers. We then generated serum profiling data for 101 CRC patients and 96 benign colorectal disease patients and analyzed 61 clinical features measured in serum individually and further their combinations. Four features were identified as our optimal small multi-biomarker panel, including known colon cancer biomarkers CEA and IL-10, as well as novel biomarkers IMA and NSE. Single feature analysis shows that CEA has the area under the curve (AUC) of receiver operating characteristic (ROC) 0.6995, followed by NSE (0.6643), IMA (0.6521), and IL-10 (0.6165). While the combined multi-biomarker panel greatly improved predictive leave-one-out cross-validation (LOOCV) accuracy to 0.7857 by nearest centroid classifier and an AUC 0.8438 by an independent three fold cross validation by support vector machines (SVMs). When we extend our optimal selection to a larger multi-biomarker panel with 13 features, the LOOCV reaches 0.8673 and AUC gets 0.8437. In addition to accuracy, our method is efficient in computational time. When compared with the exhaustive search method to select 2, 3, and 4 markers with SVM, our method dramatically reduced the searching time by 1000 folds while achieving high accuracy. Furthermore, our method can efficiently select multi- biomarker panel with more than 5 features when the exhaustive methods fail.

In conclusion, we propose a novel model to select the best multi-biomarker panel. Our method takes less running time and improves the clinical interpretability, and can serve as a useful tool for other complex disease studies.

S3. Proteoinformatics

Room: Grand Ballroom A
Date: Thursday, Oct. 3, 15:00 - 16:15

S3-1: Integrated analysis of genome-wide DNA methylation and gene expression profiles in molecular subtypes of breast cancer

Je-Keun Rhee[1], Kwangsoo Kime[2], Heejoon Chaee[3], Jared Evanse[4], Pearlly Yane[5], Byung-Tak Zhange[1,6], Joe Graye[7], Paul Spellmane[7], Tim Huange[8], Kenneth Nephewe[9,10] and Sun Kim[1,2,6,*]

1 Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul 151-742, Korea

2 Bioinformatics Institute, Seoul National University, Seoul 151-744, Korea
3 School of Informatics and Computing, Indiana University, Bloomington, IN 47408, USA
4 Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, MN 55905, USA
5 The Ohio State University Comprehensive Cancer Center Nucleic Acid Shared Resource-Illumina Core, Columbus, OH 43210, USA
6 School of Computer Science and Engineering, Seoul National University, Seoul 151-742, Korea
7 OHSU Knight Cancer Institute, Portland, OR 97239, USA
8 Department of Molecular Medicine/Institute of Biotechnology, The University of Texas Health Science Center at San Antonio, San Antonio, TX 78229-3900, USA
9 Medical Sciences, Indiana University School of Medicine, Bloomington, IN 47405, USA
10 Department of Cellular and Integrative Physiology, Indiana University School of Medicine, Indianapolis, IN 46202, USA

Abstract

Aberrant DNA methylation of CpG islands, CpG island shores and first exons is known to play a key role in the altered gene expression patterns in all human cancers. To date, a systematic study on the effect of DNA methylation on gene expression using high resolution data has not been reported. In this study, we conducted an integrated analysis of MethylCap-sequencing data and Affymetrix gene expression microarray data for 30 breast cancer cell lines representing different breast tumor phenotypes. As well-developed methods for the integrated analysis do not currently exist, we created a series of four different analysis methods. On the computational side, our goal is to develop methylome data analysis protocols for the integrated analysis of DNA methylation and gene expression data on the genome scale. On the cancer biology side, we present comprehensive genome-wide methylome analysis results for differentially methylated regions and their potential effect on gene expression in 30 breast cancer cell lines representing three molecular phenotypes, luminal, basal A and basal B. Our integrated analysis demonstrates that methylation status of different genomic regions may play a key role in establishing transcriptional patterns in molecular subtypes of human breast cancer.

S3-2: Prediction of C-peptide Like Family using Multiple Predictive Models and Feature Encodings

Elbashir Abbas[1], Ho-Jin Choi[1], Yan Zhang[2], Luonen Chen[2]

1 Knowledge Engineering and Collective Intelligence Lab.(KECI), Dept., of Computer Science, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 305-701, Korea
2 Key Laboratory of Systems Biology, Shanghai Institutes for Biological Sciences(SIBS), Chinese Academy of Sciences, Shanghai 200233, China

Abstract

Since its description in 1967, C-peptide has been historically thought to be an inert and biologically non-active peptide. That is, no physiological roles or functions were attributed to it other than connecting A and B chains' and aiding in proper folding of mature insulin. An increasing body of experimental evidence has challenged this view and purports the notion that C-peptide is bioactive, evidenced by observed signaling characteristics from in vitro experimental studies. The most pronounced is the ameliorated effect it has on diabetes induced renal and nerve dysfunction. Accordingly, the past decade has witnessed a renewal in C-peptide research aimed at providing a complete physiological characterization of the peptide. In this paper we provide the initial steps in addressing this endeavor computationally. We performed an investigative study on C-peptide that spanned 75 organisms, of which the

physiochemical properties and compositional makeup C-peptide denoted its most pronounced aspect. This was used in developing a framework composed of different predictive models and feature encodings for predicting C-peptide like family.

S3-3: Derivative Component Analysis for Serum Proteomics Data

Henry Han[1,2]

1 Department of Computer and Information Science, Fordham University, New York NY 10023 USA
2 Quantitative Proteomics Center, Columbia University, New York 10027 USA

Abstract

A new machine learning algorithm: derivative component analysis (DCA) is proposed for high dimensional proteomics data. Unlike conventional feature selection approaches, DCA aims at capturing subtle data behaviors in addition to unveiling global data behaviors through multi-resolution analysis. Compared with classic PCA and ICA methods that view each feature an indecomposable information unit in a single resolution way, DCA examines each feature in a multi-resolution approach by seeking its derivatives to capture latent data characteristics and conduct de-noising. We demonstrate DCA's advantages in disease phenotype discrimination and meaningful biomarker discovery by comparing it with state-of-the-art algorithms on benchmark data. Our results show that high-dimensional proteomics data are actually linearly separable under derivative component analysis. As a novel multi-resolution feature selection algorithm, DCA not only overcomes the weakness of the traditional methods in latent data behavior discovery, but also provides new techniques and insights in translational bioinformatics and machine learning.

S4. Multi-Omic Applications

Room: Grand Ballroom B
Date: Thursday, Oct. 3, 15:00 - 16:15

S4-1: Integrated Analysis of microRNA-target Interactions with Clinical Outcomes for Cancers

Je-Gun Joung[1,2,3], **Dokyoon Kim**[1,2,4], **Su-Yeon Lee**[1,2], **Hwa Jung Kang**[5], **Ju Han Kim**[1,2]

1 Seoul National University Biomedical Informatics (SNUBI), Div. of Biomedical Informatics, Seoul National University College of Medicine, Seoul 110-799, Korea
2 Systems Biomedical Informatics National Core Research Center, Seoul National University College of Medicine, Seoul 110-799, Korea
3 Institute of Endemic Diseases, Seoul National University College of Medicine, 103 Daehakro, Jongno-gu, Seoul 110-799, Korea
4 Center for Systems Genomics, Pennsylvania State University, University Park, Pennsylvania, USA
5 Translational Bioinformatics Lab., Samsung Genome Institute, Samsung Medical Center, Seoul Korea

Abstract

Clinical statement alone is not enough to predict the progression of disease. Instead, the gene expression profiles have been widely used to forecast clinical outcomes. Many genes related to survival have been identified, and recently miRNA expression signatures predicting patient survival have been also investigated for several cancers. However, miRNAs and their target genes associated with clinical outcomes have remained largely unexplored. Here, we demonstrate a survival analysis based on the regulatory relationships of miRNAs and their target genes. The patient survivals for the two major cancers, ovarian cancer and glioblastoma multiforme (GBM), are investigated through the integrated analysis of miRNA-mRNA interaction pair. We found that there is a larger survival difference between two patient groups with an inversely correlated expression profile of miRNA and mRNA. It supports the idea that signatures of miRNAs and their targets related to cancer progression can be detected via this approach, and subsequent therapeutic targets can in turn be identified.

S4-2: "N-of-1-pathways" unveils personal deregulated mechanisms from a single pair of RNA-Seq samples: towards precision medicine

Vincent Gardeux[1,2,3,§], Ikbel Achour[1,2,§], Mark Maienschein-Cline[1], Gurunadh Parinandi[1,4], Jianrong Li[1,2], Neil Bahroos[1], Haiquan Li[1,2], Joe G.N. Garcia[2,4,6,7], Yves A. Lussier[1,2,4,5,6,8,9,*]

1 Institute for Translational Health Informatics, University of Illinois at Chicago, Illinois, USA.
2 Department of Medicine, University of Illinois at Chicago, Chicago, Illinois, USA.
3 Department of Informatics, School of Engineering, EISTI (École Internationale de Sciences du Traitement de l'Information), Cergy-Pontoise, France
4 Department of Bioengineering, University of Illinois at Chicago, Chicago, Illinois, USA.
5 Computation Institute, Argonne National Laboratory & University of Chicago, Illinois, USA.
6 Inst. for Personalized Respiratory Medicine, University of Illinois at Chicago, Illinois, USA.
7 Department of Pharmacology, University of Illinois at Chicago, Chicago, Illinois, USA.
8 Dept. Biopharmaceutical Science, College of Pharmacy, Un. of Illinois at Chicago, Ill, USA.
9 Inst. For Genomics and Systems Biology, The University of Chicago, Chicago, Illinois, USA.
§ Equal contribution

Abstract

Background: In the groundbreaking genomic era, the emergence of precision medicine ushered in the opportunity to incorporate individual molecular data into patient care. Indeed, DNA-sequencing predicts somatic mutations of individual patients. However, these genetic features are static and overlook dynamic epigenetic and phenotypic response to therapy. Meanwhile, accurate personal transcriptome interpretation remains an unmet challenge. Further, N-of-1 (single subject) efficacy trials are increasingly pursued. However, they are not powered for molecular marker discovery.

Method: "N-of-1-pathways" translates gene expression data profiles into pathway-level profiles on single patient paired samples (one p-value per geneset). Using RNA-Seq data of 55 TCGA lung adenocarcinoma patients, it predicts individually deregulated pathways. Pooling patient-level predictions together, we then compare these pathways to those of three independent lung adenocarcinoma studies (microarray gold standards).

Results: The precision-recall curves of N-of-1-pathways predictions are comparable to those of GSEA and DEG enrichment from both internal and three external evaluations. We further show that >99.7% of 362 biological processes found in cross-patient studies are predicted by N-of-1-pathways, which also unveils 89 additional mechanisms unrelated to the gold standard shared by 1 to 40 patients. Moreover, a heatmap illustrates deregulated pathways at the single patient-level and highlights both individual and shared mechanisms ranging from molecular to organ-systems levels (e.g. DNA repair, signaling, immune response, organ development, etc.)

Conclusion: N-of-1-pathways provides a robust statistical and relevant biologic interpretation of individual response to therapy that were overlooked by cross-patient studies. Further, it enables mechanism-level classifiers with smaller cohorts as well as N-of-1-studies.
Software: https://Lussierlab.org/N-of-1-pathways

S4-3: Knowledge Boosting: A graph-based integration with multi-omics data and genomic knowledge for cancer clinical outcome prediction

Dokyoon Kim[1,2], Je-Gun Joung[1,3], Kyung-Ah Sohn[1,4], Hyunjung Shin[5], Marylyn D. Ritchie[2], Ju Han Kim[1,6,*]

1 Seoul National University Biomedical Informatics (SNUBI), Div. of Biomedical Informatics, Seoul National University College of Medicine, Seoul 110799, Korea
2 Center for Systems Genomics, Pennsylvania State University, University Park, Pennsylvania, USA
3 Translational Bioinformatics Lab (TBL), Samsung Genome Institute (SGI), Samsung Medical Center, Seoul, Korea
4 Department of Information and Computer Engineering, Ajou University, Suwon, Korea
5 Department of Industrial & Information Systems Engineering, Ajou University, San 5, Wonchun-dong, Yeoungtong-gu, 443-749, Suwon, Korea
6 Systems Biomedical Informatics Research Center, Seoul National University, Seoul 110799, Korea

Abstract

Cancer is a complex disease, which can be dysregulated through multiple mechanisms. Thus, no single level of genomic data fully elucidates tumor behavior since there are many genomic variations within/between levels in a biological system such as copy number alterations, DNA methylation, alternative splicing, miRNA regulation, post translational modification, etc. Nowadays, a number of heterogeneous types of data have become more available from the Cancer Genome Atlas (TCGA), generating multiple molecular levels of omics dimensions from genome to phenome. Given multi-omics data, information from one level to another may lead to some clues that help to uncover an unknown biological knowledge. Thus, integration of different levels of data can aid in extracting new knowledge by drawing an integrative conclusion from many pieces of information collected from diverse types of genomic data. Previously, we have proposed a graph-based framework that integrates multi-omics data including copy number alteration, DNA methylation, gene expression, and miRNA expression, for cancer clinical outcome prediction. Genomic features do not act in isolation, but rather interact with other genomic features in complex signaling or regulatory networks since cancer is caused by the deregulation of alteration in pathways or complete processes. Thus, it would be desirable to incorporate genomic knowledge when integrating multi-omics data for cancer clinical outcome prediction. Here, we proposed a new graph-based framework for integrating different levels of genomic data and genomic knowledge at hand in order to improve the predictive power and provide an enhanced global view on the interplay between levels. To highlight the validity of our proposed framework, we used an ovarian cancer dataset from TCGA for the stage, grade, and survival outcome prediction. Integrating multi-omics data with genomic knowledge to construct pre-defined features results in higher performance in clinical outcome prediction and higher stability. With integration of multi-omics data and genomic knowledge, understanding the molecular pathogenesis and underlying biology in cancer is expected to provide better guidance for improved diagnostic and prognostic indicators and effective therapies.

Scientific Paper Sessions

S5. Linking Phenotypes

Room: Grand Ballroom A
Date: Friday, Oct. 4, 08:30 - 10:10

S5-1: The Multiscale Backbone of the Human Phenotype Network based on Biological Pathways

Christian Darabos[1], Marquitta J. White[1,2], Britney E. Graham[1], Derek Leung[1], Scott Williams[1], and Jason H. Moore[1]

1 Department of Genetics, Institute for Quantitative Biomedical Sciences, Dartmouth College, Hanover, USA
2 Center for Human Genetics Research, Vanderbilt University, Nashville, USA

Abstract

Networks are commonly used to represent and analyze large and complex systems of interact- ing elements. We built pathway-based human phenotype network (PHPN) of over 800 physical attributes, diseases, and behavioral traits; based on about 2,300 genes and 1,200 biological pathways. Using GWAS phenotype-to-genes associations, and pathway data from Reactome, we connect human traits based on the common patterns of human biological pathways, detecting more pleiotropic effects, and expanding previous studies from a gene-centric approach to that of shared cell-processes. The resulting network has a heav- ily right-skewed degree distribution, placing it in the scale-free region of the network topologies spectrum. We extract the multi-scale information backbone of the PHPN based on the local densities of the network and discarding weak connection. Using a standard community detection algorithm, we construct phenotype modules of similar traits without applying expert biological knowledge. These modules can be assimilated to the disease classes. However, we are able to classify phenotypes according to shared biology, and not arbitrary disease classes. We present examples of expected clinical connections identified by PHPN as proof of principle. Furthermore, we highlight an unexpected connection between phenotype modules and discuss potential mechanistic connections that are obvious only in retrospect. The PHPN shows tremendous poten- tial to become a useful tool both in the unveiling of the diseases' mmon biology, and in the elaboration of diagnosis and treatments.

S5-2: Integrative approach for modeling the association of multi-layered genomic data with gene expression traits

Kyung-Ah Sohn[1,§], Dokyoon Kim[2,3,§], Jaehyun Lim[2,4], Ju Han Kim[2,4,*]

1 Department of Information and Computer Engineering, Ajou University, Suwon 443749, Korea
2 Seoul National University Biomedical Informatics (SNUBI), Div. of Biomedical Informatics, Seoul National University College of Medicine, Seoul 110799, Korea
3 Department of Biochemistry and Molecular Biology, Pennsylvania State University, University Park, Pennsylvania, USA
4 Systems Biomedical Informatics Research Center, Seoul National University, Seoul 110799, Korea
§ These authors contributed equally to this work

Abstract

Large-scale multi-layered genomic datasets have been emerging through collaborative efforts such as TCGA and this

provides valuable opportunities to deepen the knowledge of the molecular basis of cancer. Although many approaches have been proposed for the integrative analysis of such multi-layered data, few approaches address the problem of elucidating gene expression traits with more than two types of genomic features such as SNP, copy number alteration, methylation levels or miRNA expressions. In this work, we present a statistical framework for modeling the association of multi-layered genomic data with gene expression traits. A high-dimensional integrative genomic feature vector is constructed using multiple types of genomic features and then each gene expression trait is regressed on the integrative feature vector in a sparse regression framework. As a result, a small number of significant associations between genomic features and gene expression traits can be obtained with the corresponding association strengths. This approach allows systematic investigation of the relative contribution of different types of genomic data to gene expression traits. We demonstrate our approach on the real data of TCGA ovarian cancer patients. Our analysis shows that the integrative genomic features have greater predictive power for gene expression traits than each single type of genomics features.

S5-3: ATHENA: Identifying interactions between different levels of genomic data associated with cancer clinical outcomes using grammatical evolution neural network

Dokyoon Kim[1], Ruowang Li[1], Scott M. Dudek[1], Marylyn D. Ritchie[1,*]

1 Center for Systems Genomics, Pennsylvania State University, University Park, Pennsylvania, USA

Abstract

Gene expression profiles have been broadly used in cancer research as a diagnostic or prognostic signature for the clinical outcome prediction such as stage, grade, metastatic status, recurrence, and patient survival, as well as to potentially improve patient management. However, emerging evidence shows that gene expression-based prediction varies between independent data sets. One possible explanation of this effect is that previous studies were focused on identifying genes with large main effects associated with clinical outcomes. Thus, non-linear interactions without large individual main effects would be missed. The other possible explanation is that gene expression as a single level of genomic data is insufficient to explain the clinical outcomes of interest since cancer can be dysregulated by multiple alterations through genome, epigenome, transcriptome, and proteome levels. In order to overcome the variability of diagnostic or prognostic predictors from gene expression alone and to increase its predictive power, we need to integrate multi-levels of genomic data and identify interactions between them associated with clinical outcomes. Here, we proposed an integrative framework for identifying interactions within/between multi-levels of genomic data associated with cancer clinical outcomes using the Grammatical Evolution Neural Networks (GENN). In order to demonstrate the validity of the proposed framework, ovarian cancer data from TCGA was used as a pilot task. We found not only interactions within a single genomic level but also interactions between multi-levels of genomic data associated with survival in ovarian cancer. Notably, the integration model from different levels of genomic data achieved 72.89% balanced accuracy and outperformed the top models with any single level of genomic data. Understanding the underlying tumorigenesis and progression in ovarian cancer through the global view of interactions within/between different levels of genomic data is expected to provide guidance for improved prognostic biomarkers and individualized therapies.

S5-4: Topological Analysis of Statistical Epistasis Networks Reveals Pathways Associated with Alzheimer's Disease

Qinxin Pan[1], Ting Hu[1,2], Li Shen[3], Andrew J. Saykin[3] and Jason H. Moore[1,2,*]

1 Department of Genetics, Geisel School of Medicine, Dartmouth College, Hanover, NH, USA
2 Institute for Quantitative Biomedical Sciences, Dartmouth College, Hanover, NH, USA
3 Department of Radiology and Imaging Science, Center for Neuroimaging, Indiana University School of Medicine, Indianapolis, IN, USA

Abstract
Most pathway analysis approaches rely on main effects of genes and do not take gene-gene interactions into account. Gene-gene interactions, i.e., epistasis, are believed to account for a portion of the presumed missing heritability. Moreover, conventional methods treat each pathway independently whereas in reality they cooperate and work together as an intertwined system. In this study, we construct statistical epistasis networks (SEN) un- derlying Alzheimer's disease (AD) and infer risk-associated pathways from their topological structures. We test for pathways that possess central positions in the SENs and characterize the interactions among pathways. We find that pathway glycosphingolipid biosynthesis ganglio series, which has been hypothesized to be involved in AD pathobiology, holds central positions in the SENs and is actively interacting with a high number of other pathways. Other central pathways include alpha linolenic acid metabolism, sphingolipid metabolism, peroxisome, ether lipid metabolism, primary bile acid biosynthesis etc. In addition to central pathways, we identify a few pathways that are frequently interacting with other pathways. The pathways identified in our study should be further investigated, especially in the context of epistasis.

S6. Post-GWAS

Room: Grand Ballroom A
Date: Friday, Oct. 4, 13:20 - 14:35

S6-1: Practical issues for screening and variable selection method in a Genome- Wide Association Analysis

Sungyeon Hong[1], Yongkang Kim[1], Taesung Park[1,2,*]

1 Department of Statistics, Seoul National University, Seoul 151-747, Korea
2 Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul 151-747, Korea

Abstract
Variable selection plays an important role in high dimensional statistical modeling analysis. Computational cost and estimation accuracy are two main concerns for statistical inference of high dimensional data. Recently, many high dimensional data have been generated in biomedical science such as microarray data and single nucleotide polymorphism (SNP) data. Especially, the genome-wide association studies (GWAS) which focus on identifying SNPs associated with a disease of interest, have produced ultra-high dimensional data. Numerous methods have been proposed to handle GWAS data. Most statistical methods have adopted a two-stage approach: (1) pre- screening for dimensional reduction, (2) variable selection for identification of causal SNPs. The pre-screening step selects SNPs in terms of their p-values or absolute value of regression coefficients in single SNP analysis. Penalized regression such as

Ridge, Lasso, adaptive Lasso and Elastic-net are commonly used for the variable selection step. In this paper, we investigate which combination of prescreening method and penalized regression performs best on quantitative phenotype via real GWA data containing 327,872 SNPs from 8842 individuals.

S6-2: IGENT: Efficient Entropy based Algorithm for Detecting Genome-wide Gene-Gene Interaction Analysis

Min-Seok Kwon[1], Mira Park[2] and Taesung Park[1,3,*]

1 Interdisciplinary program in Bioinformatics, Seoul National University, Seoul, Korea
2 Department of Preventive Medicine, Eulji University, Korea
3 Department of Statistics, Seoul National University, Seoul, Korea

Abstract

With the development of high-throughput genotyping and sequencing technology, there are growing evidences of association with genetic variants and complex traits. In spite of thousands of genetic variants discovered, such genetic markers have been shown to explain only a very small proportion of the underlying genetic variance of complex traits. Gene-gene interaction (GGI) analysis is expected to unveil a lot of portion of unexplained heritability of complex traits. In this work, we propose IGENT, Information theory-based GEnome- wide gene-gene iNTeraction method. IGENT is an efficient stepwise algorithm for identifying genome-wide gene-gene interactions (GGI) and gene-environment interaction (GEI). For detecting significant GGIs in genome-wide scale, it is important to reduce computational burden significantly. Our method uses information gain (IG) and evaluates its significance without resampling. Through 70 simulation data sets, the power of the proposed method is shown to be nearly equivalent to the power of the proposed method. The proposed method successfully detected GGI for age-related macular degeneration (AMD). The proposed method is implemented by C++ and available on Windows, Linux and MacOSX.

S6-3: Identification of novel therapeutics for complex diseases from genome-wide association data

M. P. Grover[1], S. Ballouz[2], K. A. Mohanasundaram[1], R. A. George[3], C. D. H. Sherman[1], T. M. Crowley[4,5], M. A. Wouters[1,4]

1 Life and Environmental Sciences, Deakin University, Geelong, Victoria, Australia.
2 Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, United States.
3 Victor Chang Cardiac Research Institute, 405 Liverpool St, Darlinghurst, 2010, NSW, Australia.
4 School of Medicine, Deakin University, Geelong, Victoria, Australia.
5 Australian Animal Health Laboratory, CSIRO Animal, Food and Health Sciences, Portarlington Road, Geelong, Victoria, Australia.

Abstract

Background: Human genome sequencing has enabled the association of phenotypes with genetic loci, but our ability to effectively translate this data to the clinic has not kept pace. Over the past 60 years, pharmaceutical companies have successfully demonstrated the safety and efficacy of over 1,200 novel therapeutic drugs [1] via costly clinical studies. Integration of drug-target data with candidate gene prediction systems can identify novel phenotypes which may benefit

from current therapeutics. Such a drug repositioning tool can save valuable time and money spent on phase I clinical trials.
Results: We adopted a simple approach to integrate drug data with candidate gene predictions at the systems level. We previously used Gentrepid (www.gentrepid.org) as a platform to predict 1,805 candidate genes for the seven complex diseases considered in the WTCCC genome wide association study, namely Type 2 Diabetes (T2D), Bipolar Disorder (BD), Crohn's Disease (CD), Hypertension (HT), Type 1 Diabetes (T1D), Coronary Artery Disease (CAD) and Rheumatoid Arthritis (RA) [2, 3]. Using the publicly available drug databases, Therapeutic Target Database (TTD), PharmGKB and DrugBank (DB) as sources of drug-target association data, we identified a total of 390 (22%) candidate genes as novel therapeutic targets for the phenotype of interest and 2,132 drugs feasible for repositioning against the predicted targets.
Conclusions: By integrating genetic, bioinformatic and drug data, we have demonstrated that currently available drugs may be repositioned as novel therapeutics for the seven diseases studied here, quickly taking advantage of prior work in pharmaceutics to translate ground breaking results in genetics to clinical treatments.

S7. Biomedical Big Data

Room: Grand Ballroom B
Date: Friday, Oct. 4, 13:20 - 14:35

S7-1: Health Monitoring System based on Lifelog Analysis

Yongjin Kwon[1], Kyuchang Kang[1], Changseok Bae[1]

1 Human Computing Section, Software Research Laboratory, Electronics and Telecommunications Research Institute, Daejeon, Korea

Abstract
Health status is closely related to daily routines. To monitor a patient's health status, it is important to track and interpret routine data continuously. Despite the effectiveness of everyday activity information, however, both collecting and analyzing the routine data are a difficult task. To collect and analyze routine data in a seamless and accurate way, it is required to build a system that incorporates a variety of sensors, data man- agement techniques, lifelog analysis algorithm, and summarization techniques. This paper introduces a health monitoring system based on lifelog analysis. Triaxial acceleration and angular velocity data are considered as lifelog data, which are measured by the accelerometer in smartphones. A smartphone collects lifelog data con- tinuously and transfers them into a server in a secure and reliable way. The lifelog data are interpreted by our activity recognition engine in the server, and the results are used as routine information to help practitioners or other vendors provide enhanced services.

S7-2: Differentially Private Distributed Logistic Regression using Hybrid datasets

Zhanglong Ji[1], Xiaoqian Jiang[1], Shuang Wang[1], Li Xiong[2], Lucila Ohno-Machado[1]

1 Division of Biomedical Informatics, University of California, San Diego, CA, USA
2 Department of Mathematics and Computer Science, Emory University, Atlanta, GA, USA

Abstract

Differential privacy is a state-of-the-art framework for data privacy research. It offers provable privacy against attackers who have auxiliary information. However, differentially private methods sometimes introduce too much noise and make outputs less useful. We hypothesized that this situation could be alleviated in an environment where public and private data sets for the same study are available for analysis. In biomedical settings, for example, some patients are willing to sign an open-consent agreement to make their data (publicly) available for research (such as the individuals who are contributing to the 1,000 Genome Project), but others are not. Therefore, hybrid models that leverage public data while rigorously protecting private data can be developed. In this paper, we propose a novel distributed logistic regression model to be built from many data sets, including public and private ones, in a differentially private way. We showed that our algorithm has advantage over: (1) a logistic regression model based on only public data, and (2) differentially private distributed logistic regression models based on private data under various scenarios.

S7-3: Effectively processing medical term queries on the UMLS Metathesaurus by Layered Dynamic Programming

Kaiyu Ren[1,2], Albert M. Lai[1], Kun Huang[1], Aveek Mukhopadhyay[2], Raghu Machiraju[2], Yang Xiang[1]

1 Department of Biomedical Informatics,
2 Department of Computer Science and Engineer, The Ohio State University, Columbus, OH 43210, USA

Abstract

Mapping medical terms to standardized UMLS concepts is a basic step for leveraging biomedical texts in data management and analysis. However, available methods and tools have major limitations in handling queries over the UMLS Metathesaurus that contain inaccurate query terms, which frequently appear in real world applications. To provide a practical solution for this task, we propose a layered dynamic programming mapping (LDPMap) approach, which can efficiently handle these queries. Our empirical study shows that LDPMap has much higher accuracies in mapping inaccurate medical terms to UMLS concepts, in comparison with the UMLS Metathesaurus Browser and MetaMap.

S8. New Technologies

Room: Grand Ballroom A
Date: Friday, Oct. 4, 14:50 - 16:05

S8-1: GAMUT: GPU Accelerated MicroRNA analysis to Uncover Target genes through CUDA-miRanda

Shuang Wang[1], Jihoon Kim[1], Xiaoqian Jiang[1], Stefan F Brunner[2], and Lucila Ohno-Machado[1]

1 Division of Biomedical Informatics, University of California, San Diego, CA, USA
2 Biomedical Informatics, University of Applied Sciences Upper Austria, Hagenberg, Austria

Abstract

Non-coding sequences such as microRNAs have important roles in disease processes. Computational microRNA target identi- fication (CMTI) is becoming increasingly important since traditional experimental methods for target identification pose many difficulties. These methods are time-consuming, costly, and often need guidance from computational methods to narrow down candidate genes anyway. However, most CMTI methods are computationally very demanding, since they need to handle not only several million query microRNA and reference RNA pairs, but also several million nucleotide comparisons within each given pair. Thus, the need to perform microRNA identification at such large scale has increased the demand for parallel computing. Although most CMTI programs (e.g., the miRanda algorithm) are based on a modified Smith-Waterman (SW) algorithm, the existing parallel SW implementations (e.g., CUDASW++ 2.0/3.0, SWIPE) are unable to meet this demand in CMTI tasks. We present CUDA-miRanda, a fast microRNA target identification algorithm that takes advantage of massively parallel computing on Graphics Processing Units (GPU) using NVIDIA's Compute Unified Device Architecture (CUDA). CUDA-miRanda specifically focuses on the local alignment of short (i.e., < 32 nucleotides) sequences against longer reference sequences (e.g., 20K nucleotides). Moreover, the proposed algorithm is able to report multiple alignments (up to 191 top scores) and the corresponding traceback sequences for any given (query sequence, reference sequence) pair. Speeds over 5.36 Giga Cell Updates Per Second (GCUPs) are achieved on a server with 4 NVIDIA Tesla M2090 GPUs. Compared to the original miRanda algorithm, which is evaluated on an Intel Xeon E5620@2.4 GHz CPU, the experimental results show up to 166 times performance gains in terms of execution time. In addition, we have verified that the exact same targets were predicted in both CUDA-miRanda and the original miRanda implementations through multiple testing datasets. Furthermore, GPUs are inexpensive compared to high performance compute (HPC) environments in which miRanda would have to run to achieve similar performance. We offer an alternative to HPC that can be developed locally at a relatively small cost. The community of GPU developers in the biomedical research community, particularly for genome analysis, is still growing. With increasing shared resources, this community will be able to advance CMTI in a very significant manner. Our source code is available at http://dbmi-engine.ucsd.edu/cudaMiranda.

S8-2: A Novel Multi-scale Visualization Software for Data-driven Biomedical Data Exploration

Gang Su[1,2], Barbara Mirel[3], Anuj Kumar[1,4], Charles F Burant[5], Brian Athey[2] and Fan Meng[1,2]

1 The Molecular and Behavioral Neuroscience Institute,
2 Department for Computational Medicine and Bioinformatics,
3 School of Education,
4 Molecular, Cellular, Developmental biology,
5 Department of Molecular and Integrative Physiology, University of Michigan, Ann Arbor Michigan 48105, USA

Abstract

Inspired by hierarchical clustering heatmaps, CoolMap is a general-purpose, multi-scale, flexible and extensible software application for visual exploration of big biomedical datasets. To overcome the difficulty of interactive data-driven analysis of large datasets such as omics experiments and clinical trial data, CoolMap offers the capability of aggregate rows and columns in a tabular dataset at various concept levels, defined either by external data structures such as Gene Ontology (GO), KEGG pathways or experiment sample groups, or computed groups such as clustering results. Aggregated values at different concept levels, such as the data mean or standard deviation, can then replace the individual data points at the intersection of these concept terms to reduce the size and complexity of the original data and provide a high-level overview for data-driven pattern discovery and hypothesis generation. Once a 'hotspot' is identified, the researcher may drill down for additional details by expanding a concept to its child entities for fine details, while maintaining the surrounding context in coarser overview. Data could be visualized using a variety of ways, such as color, shape, text or summary plots using our custom developed high performance and extensible rendering engine. Many other auxiliary functions, such as data filtering, searching, multi-view linking, data rows/columns sorting, rearrangement, resizing, etc., were also developed to facilitate the knowledge discovery process. CoolMap was also developed using modular design so that it can be extended to for general tabular data visualization or augment other visualization software such as Cytoscape for network analysis. Compared with a variety of classic heatmap tools, CoolMap is significantly more efficient for big data exploration and analysis.

S8-3: In Silico Cancer Cell versus Stroma cellularity index computed from species-specific human and mouse transcriptome of xenograft models: towards accurate stroma targeting therapy assessment

Xinan Yang[1], Yong Huang[1], Younghee Lee[1], Vincent Gardeux[2,3,4], Ikbel Achour[2,4], Kelly Regan[2,4], Ellen Rebman[2,4], Haiquan Li[2,4], Yves A. Lussier[1,2,4,5,*,§]

1 Ctr for Biomed Inform and Sect of Genetic Medicine, Dept. of Medicine, Un. of Chicago, USA.
2 Institute for Translational Health Informatics, University of Illinois at Chicago, Illinois, USA.
3 Department of Informatics, School of Engineering, EISTI, Cergy-Pontoise, France.
4 Dept. of Medicine, University of Illinois at Chicago, Chicago, IL, USA.
5 Comprehensive Cancer Ctr and Ludwig Ctr for Metastasis Research; Un. of Chicago, IL, USA.
6 Depts of Bioengineering & of Pharmaceutical Science, Un. of Illinois at Chicago, IL, USA.
7 Computation Institute & Institute For Genomics and Systems Biology, Argonne National Laboratory and The University of Chicago, IL, USA.
8 Inst. for Personalized Respiratory Medicine, Un. of Illinois at Chicago, Chicago, Illinois, USA.
§ This work was conducted in part while at The University of Chicago

Abstract

Background: The current state of the art for measuring stromal response to targeted therapy requires burdensome and rate limiting quantitative histology. Transcriptome measures are increasingly affordable and provide an opportunity for developing a stromal versus cancer ratio in xenograft models. In these models, human cancer cells are transplanted into mouse host tissues (stroma) and together co-evolve into a tumour microenvironment. However, profiling the mouse or human component separately also remains problematic. Indeed, laser captured- microdissection is labour intensive. Moreover, gene expression using commercial microarrays introduces significant and underreported cross-species hybridization errors that are commonly overlooked by biologists. Method: We developed a customized dual-species array, H&M array, and performed cross-species and species- specific hybridization measurements. We validate a new methodology for establishing the stroma vs cancer ratio using transcriptomic data.

Results: In the biological validation of the H&M array, cross-species hybridization of human and mouse probes was significantly reduced (4.5 and 9.4 fold reduction, respectively; $p<2x10\text{-}16$ for both, Mann-Whitney test). We confirmed the capability of the H&M array to determine the stromal to cancer cells ratio based on the estimation of cellularity index of mouse/human mRNA content in vitro. This new metrics enable to investigate more efficiently the stroma-cancer cell interactions (e.g. cellularity) bypassing labour intensive requirement and biases of laser capture microdissection.

Conclusion: These results provide the initial evidence of improved and cost-efficient analytics for the investigation of cancer cell microenvironment using species-specificity arrays specifically designed for xenografts models.

Highlight Research Tracks

Highlight Research 1.

Room: Grand Ballroom C
Date: Thursday, Oct. 3, 13:00 - 14:20

H1-1: Heart Attacks: Leveraging a cardiovascular systems biology strategy to predict future outcomes

Carlo Vittorio Cannistraci[1], Timothy Ravaasi[1] and Enrico Ammirati[1]

1 Integrative Systems Biology Laboratory, Division of Biological and Environmental Sciences and Engineering, Division of Applied Mathematics and Computer Science and Engineering, Computational Bioscience Research Center, King Abdullah University for Science and Technology, Thuwal, Kingdom of Saudi Arabia

Abstract
Inflammation is likely involved in ST-elevation acute myocardial infarction (STEMI), and patients with STEMI can present with high levels of circulating interleukin-6 (IL6) at the onset of symptoms. We used machine learning techniques to identify characteristic inflammatory cytokine patterns in the blood of emergency-room patients with STEMI, and observed two functional modules characterizing the reciprocal behaviours of the cytokines in patients with high IL6 levels. Next, exploiting reverse engineering techniques, we inferred which cytokines were crucial inside the respective modules. Combining them together with IL6 in a unique formula yielded a risk-index - a kind of composed-biomarker - that outperformed any single cytokine and classical prognostic factors in the prediction of cardiac dysfunction at discharge and death at six months. Our methodology was considered a translational research innovation for the definition of composed-inflammatory-markers in cardiology, while our findings have potential implications for risk-oriented patient stratification and design of immune-modulating therapies.

H1-2: Bridging cancer: biology with the clinic: a novel personalized prognostic indicators for breast cancer

Xinan Yang[1,*], Prabhakaran Vasudevan[1], Vishwas Parekh[1], Aleks Penev[1], John M. Cunningham[1]

1 Section of Hematology/Oncology, Department of Pediatrics, Comer Children's Hospital, The University of Chicago, Chicago, Illinois, United States of America

Abstract
Identification and characterization of crucial gene target(s) that will allow focused therapeutics development remains a challenge. We have interrogated the putative therapeutic targets associated with the transcription factor Grainy head-like 2 (GRHL2), a critical epithelial regulatory factor. We demonstrate the possibility to define the molecular functions of critical genes in terms of their personalized expression profiles, allowing appropriate functional conclusions to be derived. A novel methodology, relative expression analysis with gene-set pairs (RXA-GSP), is designed to explore the potential clinical utility of cancer-biology discovery. Observing that Grhl2-overexpression leads to increased metastatic potential in vitro, we established a model assuming Grhl2-induced or -inhibited genes confer poor or favorable prognosis respectively for cancer metastasis. Training on public gene expression profiles of 995 breast cancer patients, this method prioritized one gene-set pair (GRHL2, CDH2, FN1, CITED2, MKI67 versus CTNNB1 and CTNNA3) from all 2717 possible gene-set pairs (GSPs). The identified GSP significantly dichotomized 295 independent patients for metastasis-free survival (log-rank tested p = 0.002; severe empirical p=0.035). It also showed evidence of clinical prognostication in another independent 388 patients collected from three studies (log-rank tested p = 3.3e-6). This GSP is independent of most traditional prognostic indicators, and is only significantly associated with the histological grade of breast cancer (p

= 0.0017), a GRHL2-associated clinical character (p = 6.8e-6, Spearman correlation), suggesting that this GSP is reflective of GRHL2-mediated events. Furthermore, a literature review indicates the therapeutic potential of the identified genes. This research demonstrates a novel strategy to integrate both biological experiments and clinical gene expression profiles for extracting and elucidating the genomic impact of a novel factor, GRHL2, and its associated gene-sets on the breast cancer prognosis. Importantly, the RXA-GSP method helps to individualize breast cancer treatment. It also has the potential to contribute considerably to basic biological investigation, clinical tools, and potential therapeutic targets.

H1-3: Interpreting individuals' genomes: Practical applications from newborn screening, and findings from CAGI 2013—the Critical Assessment of Genome Interpretation

Steven E. Brenner[1], John Moult[2], CAGI Participants

1 University of California, Berkeley, CA, USA
2 IBBR, University of Maryland, Rockville, MD, USA

Abstract

The Critical Assessment of Genome Interpretation (CAGI, 'kā-jē) is a community experiment to objectively assess computational methods for predicting the phenotypic impacts of genomic variation. In the experiment, participants are provided genetic variants and make predictions of resulting phenotype. These predictions are evaluated against experimental characterizations by independent assessors. A long-term goal for CAGI is to improve the accuracy of phenotype and disease predictions in clinical settings.

The third CAGI experiment (concluded in July 2013) consisted of ten diverse challenges. CAGI deliberately extends challenges from previous years, with the continuity allowing measurement of progress. For example, in the second CAGI, in a challenge to predict Crohn's disease from exomes, one group was able to identify 80% of affected individuals before the first false positive healthy person. In the third CAGI experiment, this challenge used an improved dataset, and several groups performed remarkably well, with one group achieving a ROC AUC of 0.94. The experiment also revealed important population structure to Crohn's disease in Germany. For three years, CAGI has posed a challenge with Personal Genome Project (PGP) genome data. This year, two groups were able to successfully map a significant number of complete genomes to their corresponding trait profiles submitted by PGP participants. In the expanded challenge to predict benign versus deleterious variants in DNA double-strand break repair MRN genes-Rad50 (from last year), Mre11, and Nbs1-as determined by those that appear in a breast cancer case versus healthy control, predictions show how methods differ sharply in their effectiveness even amongst proteins in the same complex.

A new challenge this year was to use exomes from families with lipid metabolism disorders. In the case of hypoalphalipoproteinemia (HA), a company made predictions which showed how understanding the problem structure and employing an extensive knowledgebase led to remarkably good results. Another related challenge revealed a twist wherein real-world data differed sharply from theoretical models.

The other challenges were to predict which variants of BRCA1 and BRCA2 are associated with increased risk of breast cancer; to predict how variants in p53 gene exons affect mRNA splicing; to predict how well variants of a p16 tumor suppressor protein inhibit cell proliferation; and to identify potential causative SNPs in disease-associated loci.

Overall, CAGI revealed that the phenotype prediction methods embody a rich and diverse representation of biological knowledge, and they are able to make predictions that are highly statistically significant. However, we also found the accuracy of prediction on the phenotypic impact of any specific variant was unsatisfactory and of questionable clinical utility. The most effective predictions came from methods honed to the precise challenge, including the specific genes of interest as well as the problem context. Prediction methods are clearly growing in sophistication, yet there are extensive opportunities for further progress.

Complete information about CAGI may be found at http://genomeinterpretation.org.

Highlight Research Tracks

Highlight Research 2.

Room: Grand Ballroom C
Date: Thursday, Oct. 3, 15:00 - 16:15

H2-1: The SADI PersonThe SADI Personal Health Lens: A Web Browser-Based System for Identifying Personally Relevant Drug Interactions

Ben Vandervalk[1], E Luke McCarthy[1], José Cruz-Toledo[2], Artjom Klein[3], Christopher J O Baker[3], Michel Dumontier[2], Mark D Wilkinson[4]

1 James Hogg Research Centre, Heart & Lung Institute, University of British Columbia, Vancouver, BC, Canada
2 Department of Biology, Carleton University, Ottawa, ON, Canada
3 Department of Computer Science and Applied Statistics, University of New Brunswick, Saint John, NB, Canada
4 Centro de Biotecnología y Genómica de Plantas, Universidad Politécnica de Madrid, Pozuelo de Alarcón (Madrid), Spain

Abstract

Background: The Web provides widespread access to vast quantities of health-related information that can improve quality-of-life through better understanding of personal symptoms, medical conditions, and available treatments. Unfortunately, identifying a credible and personally relevant subset of information can be a time-consuming and challenging task for users without a medical background.

Objective: The objective of the Personal Health Lens system is to aid users when reading health-related webpages by providing warnings about personally relevant drug interactions. More broadly, we wish to present a prototype for a novel, generalizable approach to facilitating interactions between a patient, their practitioner(s), and the Web.

Methods: We utilized a distributed, Semantic Web-based architecture for recognizing personally dangerous drugs consisting of: (1) a private, local triple store of personal health information, (2) Semantic Web services, following the Semantic Automated Discovery and Integration (SADI) design pattern, for text mining and identifying substance interactions, (3) a bookmarklet to trigger analysis of a webpage and annotate it with personalized warnings, and (4) a semantic query that acts as an abstract template of the analytical workflow to be enacted by the system.

Results: A prototype implementation of the system is provided in the form of a Java standalone executable JAR file. The JAR file bundles all components of the system: the personal health database, locally-running versions of the SADI services, and a javascript bookmarklet that triggers analysis of a webpage. In addition, the demonstration includes a hypothetical personal health profile, allowing the system to be used immediately without configuration. Usage instructions are provided.

Conclusions: The main strength of the Personal Health Lens system is its ability to organize medical information and to present it to the user in a personalized and contextually relevant manner. While this prototype was limited to a single knowledge domain (drug/drug interactions), the proposed architecture is generalizable, and could act as the foundation for much richer personalized-health-Web clients, while importantly providing a novel and personalizable mechanism for clinical experts to inject their expertise into the browsing experience of their patients in the form of customized semantic queries and ontologies.

H2-2: Correlation network-guided novel key gene identification

Feng He[1,2,§], Hairong Chen[1,§], Michael Probst-Kepper[3], Robert Geffers[4], Serge Eifes[2], Antonio del Sol[2], Klaus Schughart[1], An-Ping Zeng[5,6] and Rudi Balling[2,*]

1 Department of Infection Genetics, Helmholtz Centre for Infection Research (HZI), University of Veterinary Medicine Hannover, Braunschweig, Germany
2 Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, Esch-sur-Alzette, Luxembourg

3 Institute of Microbiology, Immunology and Hospital Hygiene, Städtisches Klinikum Braunschweig GmbH, Braunschweig, Germany
4 Department of Cell Biology, Helmholtz Centre for Infection Research, Braunschweig, Germany
5 Group of Systems Biology, Helmholtz Centre for Infection Research, Braunschweig, Germany
6 Institute of Bioprocess and Biosystems Engineering, Hamburg University of Technology, Hamburg, Germany
§ These authors contributed equally to this work

Abstract
Human FOXP3+, CD25+, CD4+ regulatory T cells(Tregs) are essential to the maintenance of immune homeostasis. Several genes are known to be important for murine Tregs, but for human Tregs the genes and underlying molecular networks controlling the suppressor function still largely remain unclear. Here, we describe a strategy to identify the key genes directly from an undirected correlation network which we reconstruct from a very high time-resolution (HTR) transcriptome during the activation of human Tregs/CD4+ T-effector cells. We show that a predicted top-ranked new key gene PLAU (the plasminogen activator urokinase) is important for the suppressor function of both human and murine Tregs. Further analysis unveils that PLAU is particularly important for memory Tregs and that PLAU mediates Treg suppressor function via STAT5 and ERK signaling pathways. Our study demonstrates the potential for identifying novel key genes for complex dynamic biological processes using a network strategy based on HTR data, and reveals a critical role for PLAU in Treg suppressor function.

H2-3: Imbalanced network biomarkers for traditional Chinese medicine Syndrome in gastritis patients

Rui Li[1,§], Tao Ma[1,§], Jin Gu[1], Xujun Liang[1] and Shao Li[1,*]

1 Bioinformatics Division and Center for Synthetic and Systems Biology, TNLIST/Department of Automation, Tsinghua University, Beijing 100084, China.
§ Co-first authors

Abstract
Cold Syndrome and Hot Syndrome are thousand-year-old key therapeutic concepts in traditional Chinese medicine (TCM), which depict the loss of body homeostasis. However, the scientific basis of TCM Syndrome remains unclear due to limitations of current reductionist approaches. Here, we established a network balance model to evaluate the imbalanced network underlying TCM Syndrome and find potential biomarkers. By implementing this approach and investigating a group of chronic superficial gastritis (CSG) and chronic atrophic gastritis (CAG) patients, we found that with leptin as a biomarker, Cold Syndrome patients experience low levels of energy metabolism, while the CCL2/MCP1 biomarker indicated that immune regulation is intensified in Hot Syndrome patients. Such a metabolism-immune imbalanced network is consistent during the course from CSG to CAG. This work provides a new way to understand TCM Syndrome scientifically, which in turn benefits the personalized medicine in terms of the ancient medicine and complex biological systems.

Highlight Research 3.

Room: Grand Ballroom B
Date: Friday, Oct. 4, 08:30 - 10:10

H3-1: Computational Studies of Ubiqutin and Ubiquitin-like Conjugation

Tianshun Gao[1,*], Yu Xue[1]

1 Huazhong University of Science and Technology, Wuhan, P. R. China, 430074

Abstract

Justification: The 2004 Nobel Prize in Chemistry was awarded for the discovery of ubiquitin and ubiquitin proteasome system that is a highly-specific, ATP-dependent pathway responsible for targeting specific proteins for degradation and regulating nearly all of cellular processes. Since the investigation has advanced with thousands of experimental efforts, an integrative and comprehensive data resource is still not available. Besides, systematic identification of ubiquitinated proteins with modified sites has emerged to be another hot topic, at least 10 prediction programs for ubiquitination sites have been developed, but none of them were able to predict ubiquitin ligases for substrates' ubiquitination sites. Identification of protein ubiquitination sites with their cognate ubiquitin ligases (E3s) has been critical for understanding the complete ubiquitination and potential relationships between ubiquitinaton and other important cellular processes. We have developed UUCD database and GPS-PLUB predictor to overcome the two difficulties.

Methods: From the scientific literature, 26 E1s, 105 E2s, 1003 E3s and 148 deubiquitination enzymes (DUBs) were collected and classified into 1, 3, 19 and 7 families, respectively. 981 ubiquitination sites with E3 information were also collected and 1154 site-E3 pairs were integrated. Furthermore, there were 965 sites without redundancies in the 8 main E3 families and 1126 sites in the 87 single E3s. To computationally characterize potential enzymes in 70 eukaryotic species, we constructed 1, 1, 15 and 6 hidden Markov model (HMM) profiles for E1s, E2s, E3s and DUBs at the family level, separately. Moreover, the ortholog searches were conducted for E3 and DUB families without HMM profiles. All experimentally identified enzymes were taken as the benchmark dataset to evaluate the prediction performance and robustness of the HMM identifications. We first classified E3-associated proteins into two classes as E3 activity and E3 adaptor. Besides, we adopted the GPS (Group-based Prediction System) algorithm, and developed a useful tool for predicting E3-specific ubiquitination sites for 87 E3s in hierarchy. Especially, a reasonable approach used in the predictor was able to successfully estimate the theoretically maximal false positive rates (FPR). Taking APC/C family as an example, the training sensitivity and specificity were 100% and 94.74% respectively with the FPR of 2%. The predictor showed a great performance and a significant accuracy.

Results: a database UUCD (Ubiquitin and Ubiquitin-like Conjugation Database) was developed with 738 E1s, 2937 E2s, 46 631 E3s and 6647 DUBs of 70 eukaryotic species. Besides, a useful tool GPS-PLUB (Prediction of ubiquitin Ligase-based Ubiquination sites) was designed for predicting E3-specific ubiquitination sites for 87 E3s in hierarchy.

Conclusions: Taken together, we developed a family-based database (http://uucd.biocuckoo.org) for ubiquitin and ubiquitin-like conjugation, through a similar E1 (ubiquitin-activating enzyme)-E2 (ubiquitin-conjugating enzyme)-E3 (ubiquitin-protein ligase) enzyme thioester cascade and a predictor for identification of E3-specific ubiquitination sites. We believe that they can lead users to generate a comprehensive view of the ubiquitination modification and also serve as a useful resource for further researches.

H3-2: Mechanisms of PDGFRα promiscuity and PDGFRβ specificity in association with PDGFB

Daniel Torrente[1], Ricardo Cabezas[1], Marco Fidel[1], Francisco Capani[2], Yuly Sanchez[1], Ludis Morales[1], George E. Barreto[1,*], Janneth González[1,*]

1 Departmento de Nutrición y Bioquímica, Facultad de Ciencias, Pontificia Universidad Javeriana, Bototá D.C., Colombia

2 Laboratorio de Citoarquitectura y Plasticidad Neuronal, Instituto de Investigaciones Cardiológicas "Prof. Dr. Alberto C. Taquini", UMA-CONICET, Buenos Aires, Argentina

Abstract

Platelet-derived growth factor (PDGF) receptor α interacts with PDGFA, B, C and AB, while PDGFRβ just binds to PDGFB and D, suggesting that PDGFRα is more promiscuous than PDGFRβ. The structural analysis of PDGFRα-PDGFA and PDGFRα-PDGFB complexes and a molecular explanation for the promiscuity of PDGFRα and the specificity of PDGFRβ remain unclear. In the present study, we modeled the three extracellular domains of PDGFRα using a previous crystallographic structure of PDGFRβ as a template. Additionally, we analyzed the interacting residues of PDGFRα-PDGFA and PDGFRα-PDGFB complexes using docking simulations. The validation of the resulting complexes was evaluated by molecular dynamics simulations. Structural analysis revealed that changes of non-aromatic amino acids in PDGFRα to aromatic amino acids in PDGFRβ (ILE 139PHE, PRO267PHE and ASN204TYR) may be

involved in the promiscuity of PDGFRα. Indeed, substitution of amino acids with low probability of rotamer changes in PDGFRβ (MET133ALA, ASN163GLU and ASN179SER) and energy stability due the formation of hydrogen bond in PDGFRβ could explain the specificity of PDGFRβ. These results could be used as an input for a better and more specific drug design for diseases related with the malfunction of PDGFs and PDGFRα such as cancer and atherosclerosis.

H3-3: DisplayHTS: a R package for visualizing high-throughput screening data results

Xiaohua Douglas Zhang[1,*] and Zhaozhi Zhang[2]

1 Early Development Statistics, BARDS, Merck Research Laboratories, West Point, PA 19486, USA
2 Central Bucks South, Warrington, PA 18976, USA

Abstract

RNA interference (RNAi) research has been used to elucidate gene function, to identify novel drug targets, and to reveal the molecular biological system. RNAi high-throughput screening (HTS) study allows genome-wide loss-of-function screening. One of the major advantages of RNAi HTS is its ability to simultaneously interrogate thousands of genes. With the ability of generating a large amount of data per experiment, RNAi HTS has led to an explosion in the rate of data generated in recent years. Consequently, one of the most fundamental challenges in RNAi HTS experiments is to glean biological significance from mounds of data, which relies on the development and adoption of suitable statistics/bioinformatics methods. Recently, we have been developing novel analytic methods specifically for quality control and hit selection in RNAi HTS experiments. We published a R package displayHTS in Bioinformatics in 2013. This package implements recently developed methods and figures for displaying data and hit selection results in HTS experiments. It generates useful distinctive graphics. In this presentation, I will describe related statistical methods, elaborate the R package and demonstrate how to use the R package in HTS experiments.

3-4: Methylerythritol phosphate pathway to isoprenoids: Kinetic modeling and *in silico* enzyme inhibitions in P. falciparum

deoxy-D-xylulose 5-phosphate reductoisomerase (DXR) as drug target from the systemic perspective, and additional target identification, using metabolic con- trol analysis and inhibition studies. In addition to DXR, 1-deoxy-D-xylulose 5-phosphate synthase (DXS) can be targeted because it is the first enzyme of the pathway and has the highest flux control coefficient followed by that of DXR. *In silico* inhibition of both enzymes caused large decre- ment in the pathway flux. An added advantage of targeting DXS is its influence on vitamin B1 and B6 biosynthesis. Two more potential targets, 2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase and 1-hydroxy-2-methyl-2-(E)-butenyl 4-diphosphate synthase, were also identified. Their inhibi- tion caused large accumulation of their substrates causing instability of the system.

Highlight Research 4.

Room: Grand Ballroom C
Date: Friday, Oct. 4, 13:20 - 14:35

H4-1: A protein domain-centric approach for the comparative analysis of human and yeast phenotypically relevant mutations

Thomas A Peterson[1], DoHwan Park[2], Maricel G Kann[1,*]

1 Department of Biological Sciences, University of Maryland, Baltimore County, Baltimore, MD, USA.
2 Department of Mathematics and Statistics, University of Maryland, Baltimore County, Baltimore, MD, USA.

Abstract

Background: The body of disease mutations with known phenotypic relevance continues to increase and is expected to do so even faster with the advent of new experimental techniques such as whole-genome sequencing coupled with disease association studies. However, genomic association studies are limited by the molecular complexity of the phenotype being studied and the population size needed to have adequate statistical power. One way to circumvent this problem, which is critical for the study of rare diseases, is to study the molecular patterns emerging from functional studies of existing disease mutations. Current gene-centric analyses to study mutations in coding regions are limited by their inability to account for the functional modularity of the protein. Previous studies of the functional patterns of known human disease mutations have shown a significant tendency to cluster at protein domain positions, namely position-based domain hotspots of disease mutations. However, the limited number of known disease mutations remains the main factor hindering the advancement of mutation studies at a functional level. In this paper, we address this problem by incorporating mutations known to be disruptive of phenotypes in other species. Focusing on two evolutionarily distant organisms, human and yeast, we describe the first inter-species analysis of mutations of phenotypic relevance at the protein domain level.
Results: The results of this analysis reveal that phenotypic mutations from yeast cluster at specific positions on protein domains, a characteristic previously revealed to be displayed by human disease mutations. We found over one hundred domain hotspots in yeast with approximately 50% in the exact same domain position as known human disease mutations.
Conclusions: We describe an analysis using protein domains as a framework for transferring functional information by studying domain hotspots in human and yeast and relating phenotypic changes in yeast to diseases in human. This first-of-a-kind study of phenotypically relevant yeast mutations in relation to human disease mutations demonstrates the utility of a multi-species analysis for advancing the understanding of the relationship between genetic mutations and phenotypic changes at the organismal level.

H4-2: Yin and Yang of reciprocally scale-free biological networks between lethal genes and disease genes

Hyun Wook Han[1,2,3], Jung Hun Ohn[1,3], Jisook Moon[2,*] and Ju Han Kim[1,3,*]

1 Division of Biomedical Informatics, Seoul National University Biomedical Informatics (SNUBI), Seoul National University College of Medicine, Seoul 110799, Korea
2 College of Medicine, CHA General Hospital, CHA University, Seoul 135081, Korea
3 Systems Biomedical Informatics Research Center, Seoul National University, Seoul 110799, Korea

Abstract

Biological networks often show a scale-free topology with node degree following a power-law distribution. Lethal genes tend to form functional hubs, whereas non-lethal disease genes are located at the periphery. Uni-dimensional analyses, however, are flawed. We created and investigated two distinct scale-free networks; a protein-protein interaction (PPI) and a perturbation sensitivity network (PSN). The hubs of both networks exhibit a low molecular evolutionary rate ($P < 8 \times 10^{-12}$, $P < 2 \times 10^{-4}$) and a high codon adaptation index ($P < 2 \times 10^{-16}$, $P < 2 \times 10^{-8}$), indicating that both hubs have been shaped under high evolutionary selective pressure. Moreover, the topologies of PPI and PSN are inversely proportional: hubs of PPI tend to be located at the periphery of PSN and vice versa. PPI hubs are highly enriched with lethal genes but not with disease genes, whereas PSN hubs are highly enriched with disease genes and drug targets but not with lethal genes. PPI hub genes are enriched with essential cellular processes, but PSN hub genes are enriched with environmental interaction processes, having more TATA boxes and transcription factor binding sites. It is concluded that biological systems may balance internal growth signaling and external stress signaling by unifying the two opposite scale-free networks that are seemingly opposite to each other but work in concert between death and disease.

H4-3: Comparative analysis using K-mer and K-flank patterns provides evidence for CpG island

sequence evolution in mammalian genomes

Heejoon Chae[1], Jinwoo Park[2,3], Seong-Whan Lee[4], Kenneth P. Nephew[5] and Sun Kim[2,3,*]

1 Department of Computer Science, School of Informatics and Computing, Indiana University, Bloomington, IN, USA
2 Department of Computer Science and Engineering, Bioinformatics Institute, Seoul National University, Seoul, Korea
3 Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, Korea
4 Department of Brain and Cognitive Engineering, Korea University, Seoul, Korea
5 Medical Sciences Program, Indiana University School of Medicine, Indiana University, Bloomington, IN, USA

Abstract
CpG islands are GC-rich regions often located in the 50 end of genes and normally protected from cytosine methylation in mammals. The important role of CpG islands in gene transcription strongly suggests evolutionary conservation in the mamma- lian genome. However, as CpG dinucleotides are over-represented in CpG islands, comparative CpG island analysis using conventional sequence analysis techniques remains a major challenge in the epigenetics field. In this study, we conducted a comparative analysis of all CpG island sequences in 10 mammalian genomes. As sequence similarity methods and character composition techniques such as information theory are particularly difficult to conduct, we used exact patterns in CpG island sequences and single character discrepancies to identify differences in CpG island sequences. First, by calculating genome distance based on rank cor- relation tests, we show that k-mer and k-flank patterns around CpG sites can be used to correctly reconstruct the phylogeny of 10 mammalian genomes. Further, we used various machine learning algorithms to demonstrate that CpG islands sequences can be characterized using k-mers. In addition, by testing a human model on the nine different mammalian genomes, we provide the first evidence that k-mer signatures are consist- ent with evolutionary history.

Highlight Research 5.

Room: Grand Ballroom B
Date: Friday, Oct. 4, 14:50 - 16:05

H5-1: Arpeggio: harmonic compression of ChIP-seq data reveals protein-chromatin interaction signatures

Kelly Patrick Stanton[1], Fabio Parisi[1], Francesco Strino[1], Neta Rabin[2], Patrik Asp[3] and Yuval Kluger[1,4,*]

1 Department of Pathology, Yale University School of Medicine, 333 Cedar Street, New Haven, CT 06520, USA
2 Department of Exact Sciences, Afeka - Tel-Aviv Academic College of Engineering, Tel-Aviv 69107, Israel
3 Department Of Liver Transplant, Montefiore Medical Center, Albert Einstein College of Medicine, Bronx, NY 10467, USA
4 NYU Center for Health Informatics and Bioinformatics, New York University Langone Medical Center, 227 East 30th Street, New York, NY 10016, USA

Abstract
Researchers generating new genome-wide data in an exploratory sequencing study can gain biological insights by comparing their data with well-annotated data sets possessing similar genomic patterns. Data compression techniques are needed for efficient comparisons of a new genomic experiment with large repositories of publicly available profiles. Furthermore, data representations that allow comparisons of genomic signals from different platforms and across species enhance our ability to leverage these large repositories. Here, we present a signal processing approach that characterizes protein-chromatin interaction patterns at length scales of several kilobases. This allows us to efficiently compare numerous chromatin-immuno- precipitation sequencing (ChIP-seq) data sets consisting of many types of DNA-binding proteins collected from a variety of cells, conditions and organisms. Importantly, these interaction patterns broadly reflect

the biological properties of the binding events. To generate these profiles, termed Arpeggio profiles, we applied harmonic deconvolution techniques to the autocorrelation profiles of the ChIP-seq signals. We used 806 publicly available ChIP-seq experiments and showed that Arpeggio profiles with similar spectral densities shared biological properties. Arpeggio profiles of ChIP-seq data sets revealed characteristics that are not easily detected by standard peak finders. They also allowed us to relate sequencing data sets from different genomes, experimental platforms and protocols. Arpeggio is freely available at http://sourceforge. net/p/arpeggio/wiki/Home/.

H5-2: TrAp: a tree approach for fingerprinting subclonal tumor composition

Francesco Strino[1], Fabio Parisi[1], Mariann Micsinai[1,2] and Yuval Kluger[1,2,3,*]

1 Department of Pathology, Yale University School of Medicine, New Haven, CT 06520, USA 2 NYU Center for Health Informatics and Bioinformatics, New York University Langone Medical Center, 227 East 30th Street, New York, NY 10016, USA
3 Yale Cancer Center, New Haven, CT 06520, USA

Abstract
Revealing the clonal composition of a single tumor is essential for identifying cell subpopulations with metastatic potential in primary tumors or with resistance to therapies in metastatic tumors. Sequencing technologies provide only an overview of the aggregate of numerous cells. Computational approaches to de-mix a collective signal composed of the aberrations of a mixed cell population of a tumor sample into its individual components are not available. We propose an evolutionary frame- work for deconvolving data from a single genome-wide experiment to infer the composition, abundance and evolutionary paths of the underlying cell subpopulations of a tumor. We have developed an algorithm (TrAp) for solving this mixture problem. *In silico* analyses show that TrAp correctly deconvolves mixed subpopulations when the number of subpopulations and the measurement errors are moderate. We demonstrate the applic-ability of the method using tumor karyotypes and somatic hypermutation data sets. We applied TrAp to Exome-Seq experiment of a renal cell carcinoma tumor sample and compared the mutational profile of the inferred subpopulations to the mutational profiles of single cells of the same tumor. Finally, we deconvolve sequencing data from eight acute myeloid leukemia patients and three distinct metastases of one melanoma patient to exhibit the evolutionary relationships of their subpopulations.

H5-3: Variants Affecting Exon Skipping Contribute to Complex Traits

Younghee Lee[1,*], Eric R. Gamazon[1], Ellen Rebman[1], Yeunsook Lee[2], Sanghyuk Lee[3], M. Eileen Dolan[1], Nancy J. Cox[1,*], Yves A. Lussier[1,4,*]

1 Department of Medicine, The University of Chicago, Chicago, Illinois, United States of America
2 Bioinformatics and Computational Biology Program, Iowa State University, Ames, Iowa, United States of America
3 Departments of Life Sciences, Ewha Womans University, Seoul, Korea
4 Departments of Medicine and of Bioengineering, University of Illinois at Chicago, Chicago, Illinois, United States of America

Abstract
DNA variations affect alternative splicing, an overlooked mechanism present in 40% of complex human diseases. A commonly held hypothesis asserts that, in complex human traits, altered splicing patterns might be more important than expression changes in determining disease-associated risks. Furthermore, the therapeutic potential of using single nucleotide polymorphisms (SNPs) to cause alternative splicing of exons has been experimentally demonstrated in models of human disease. The precise mechanism by which SNPs regulate this process remains to be fully elucidated. In this study, we develop an integrative approach that utilizes sequence-based analysis and genome-wide expression profiling to identify genetic variations that may affect alternative splicing. We also provide the first proof of their enrichment among validated disease-associated variations. Our study provides insights into the functionality of these variations and emphasizes their importance for complex human traits and diseases.

Highlight Research Tracks

Highlight Research 6.

Room: Grand Ballroom C
Date: Friday, Oct. 4, 14:50 - 16:05

H6-1: Statistical epistasis networks reduce the computational complexity of searching three-locus genetic models

Ting Hu[1], Angeline S. Andrew[2], Margaret R. Karagas[2], Jason H. Moore[3,*]

1 Department of Genetics, Geisel School of Medicine Dartmouth College, Hanover, NH 03755, USA
2 Department of Community and Family Medicine, Geisel School of Medicine Dartmouth College, Hanover, NH 03755, USA
3 Institute for Quantitative Biomedical Sciences Departments of Genetics and Community and Family Medicine, Geisel School of Medicine Dartmouth College, Hanover, NH 03755, USA

Abstract
The rapid development of sequencing technologies makes thousands to millions of genetic at- tributes available for testing associations with various biological traits. Searching this enormous high-dimensional data space imposes a great computational challenge in genome-wide association studies. We introduce a network-based approach to supervise the search for three-locus models of disease susceptibility. Such statistical epistasis networks (SEN) are built using strong pairwise epistatic interactions and provide a global interaction map to search for higher-order interactions by prioritizing genetic attributes clustered together in the networks. Applying this approach to a population-based bladder cancer dataset, we found a high susceptibility three-way model of ge- netic variations in DNA repair and immune regulation pathways, which holds great potential for studying the etiology of bladder cancer with further biological validations. We demonstrate that our SEN-supervised search is able to find a small subset of three-locus models with significantly high associations at a substantially reduced computational cost.

H6-2: PhenDisco (Phenotype Discoverer): a New Information Retrieval System for the database of Genotypes and Phenotypes

Hyeoneui Kim[1], Son Doan[1], Lucila Ohno-Machado[1]

1 Division of Biomedical Informatics, University of California San Diego, La Jolla, CA, USA

Abstract
The database of Genotypes and Phenotypes (dbGaP) developed by the National Center for Biotechnology Information (NCBI) contains information on phenotypes, genotypes and study protocols from various Genome Wide Association Studies (GWAS). Although dbGaP is a critical resource that can facilitate new exploratory research or cross-study validation, lack of standardization in a way that phenotype information is presented becomes a major barrier to accurate and complete retrieval of the studies with a phenotype of interest. As a solution to this challenge, we developed an NLP-based process to standardize phenotype variables, and an information retrieval tool that processes user queries, and then displays the results in the order of relevance. These processes were implemented in PhenDisco. In a preliminary evaluation, PhenDisco showed better retrieval performance than dbGaP, as well as superior acceptance by users,, showing that it fills an important gap in this area.

H6-3: Phenome-Wide Association Study (PheWAS) for Detection of Pleiotropy within the Population Architecture using Genomics and Epidemiology (PAGE) Network

Highlight Research Tracks

Sarah A. Pendergrass[1], Kristin Brown-Gentry[2], Scott Dudek[2], Alex Frase[1], Eric S. Torstenson[2], Robert Goodloe[2], Jose Luis Ambite[3], Christy L. Avery[4], Steve Buyske[5,6], Petra Bůžková[7], Ewa Deelman[3], Megan D. Fesinmeyer[8], Christopher A. Haiman[9], Gerardo Heiss[4], Lucia A. Hindorff[10], Chu-Nan Hsu[3], Rebecca D. Jackson[11], Charles Kooperberg[8], Loic Le Marchand[12], Yi Lin[8], Tara C. Matise[5], Kristine R. Monroe[9], Larry Moreland[13], Sungshim L. Park[12], Alex Reiner[8,14], Robert Wallace[15], Lynn R. Wilkens[12], Dana C. Crawford[2,16], Marylyn D. Ritchie[1,*]

1 Center for Systems Genomics, Department of Biochemistry and Molecular Biology, The Pennsylvania State University, Eberly College of Science, The Huck Institutes of the Life Sciences, University Park, Pennsylvania, United States of America
2 Center for Human Genetics Research, Vanderbilt University, Nashville, Tennessee, United States of America
3 Information Sciences Institute, University of Southern California, Marina del Rey, California, United States of America
4 Department of Epidemiology, University of North Carolina, Chapel Hill, North Carolina, United States of America
5 Department of Genetics, Rutgers University, Piscataway, New Jersey, United States of America
6 Department of Statistics, Rutgers University, Piscataway, New Jersey, United States of America
7 Department of Biostatistics, University of Washington, Seattle, Washington, United States of America
8 Division of Public Health, Fred Hutchinson Cancer Research Center, Seattle, Washington, United States of America
9 Department of Preventive Medicine, Keck School of Medicine, University of Southern California/Norris Comprehensive Cancer Center, Los Angeles, California, United States of America
10 National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland, United States of America
11 Ohio State University, Columbus, Ohio, United States of America
12 Epidemiology Program, University of Hawaii Cancer Center, Honolulu, Hawaii, United States of America
13 University of Pittsburgh, Pittsburgh, Pennsylvania, United States of America
14 Department of Epidemiology, University of Washington, Seattle, Washington, United States of America
15 Departments of Epidemiology and Internal Medicine, University of Iowa, Iowa City, Iowa, United States of America
16 Department of Molecular Physiology and Biophysics, Vanderbilt University, Nashville, Tennessee, United States of America

Abstract

Using a phenome-wide association study (PheWAS) approach, we comprehensively tested genetic variants for association with phenotypes available for 70,061 study participants in the Population Architecture using Genomics and Epidemiology (PAGE) network. Our aim was to better characterize the genetic architecture of complex traits and identify novel pleiotropic relationships. This PheWAS drew on five population-based studies representing four major racial/ethnic groups (European Americans (EA), African Americans (AA), Hispanics/Mexican-Americans, and Asian/Pacific Islanders) in PAGE, each site with measurements for multiple traits, associated laboratory measures, and intermediate biomarkers. A total of 83 single nucleotide polymorphisms (SNPs) identified by genome-wide association studies (GWAS) were genotyped across two or more PAGE study sites. Comprehensive tests of association, stratified by race/ethnicity, were performed, encompassing 4,706 phenotypes mapped to 105 phenotype-classes, and association results were compared across study sites. A total of 111 PheWAS results had significant associations for two or more PAGE study sites with consistent direction of effect with a significance threshold of p,0.01 for the same racial/ethnic group, SNP, and phenotype-class. Among results identified for SNPs previously associated with phenotypes such as lipid traits, type 2 diabetes, and body mass index, 52 replicated previously published genotype-phenotype associations, 26 represented phenotypes closely related to previously known genotype-phenotype associations, and 33 represented potentially novel genotype-phenotype associations with pleiotropic effects. The majority of the potentially novel results were for single PheWAS phenotype-classes, for example, for CDKN2A/B rs1333049 (previously associated with type 2 diabetes in EA) a PheWAS association was identified for hemoglobin levels in AA. Of note, however, GALNT2 rs2144300 (previously associated with high-density lipoprotein cholesterol levels in EA) had multiple potentially novel PheWAS associations, with hypertension related phenotypes in AA and with serum calcium levels and coronary artery disease phenotypes in EA. PheWAS identifies associations for hypothesis generation and exploration of the genetic architecture of complex traits.

Posters Session

TBC-1: *Firdous Khan and Ashley Pretorius*
The implications of RBBP6 in various types of cancer

TBC-2: *Hye Hyeon Kim, Soo Youn Lee, Su Youn Baik, Kye Hwa Lee and Ju Han Kim*
MELLO: Medical Life-Log Ontology

TBC-3: *Junbeom Kim, Jun Hyuk Kang and Ho-Jin Choi*
Investigation on Gene Expression Patterns of Cardiac Myocyte Hypertrophy using Coexpression Network Analysis

TBC-4: *Frida Belinky, Gil Stelzer, Simon Fishilevich, Shahar Zimmerman, Marilyn Safran and Doron Lancet*
Tell me your pathways

TBC-5: *Dmitriy Shin, Gerald Arthur, Mihail Popescu, Dmitry Korkin and Chi-Ren Shyu*
Computational Morphoproteomics: Inferring Biological Relationships from Resource Description Framework

TBC-6: *Qing Zhang, Xiaodan Fan and Dianjing Guo*
A quantitative mixture model for transcriptome prediction

TBC-7: *Won-Hyong Chung, Namshin Kim, Kyung-Tai Lee and Tae-Hun Kim*
De novo genome sequencing project of Korean native pig:Current status of genome assembly and annotation

TBC-8: *Punit Kaur, Parul Sharma, Sujata Sharma and T. P. Singh*
Systems Biology Integrative approach uncovers newer molecular targets in Metachromatic Leukodystrophy

TBC-9: *Chol-Hee Jung, Gianluca Severi, Melissa Southey, Dallas English, Andrew Lonie, Helen Tsimiklis, John Hopper, Graham G Giles and Laura Baglietto*
Measuring DNA methylation in large epidemiological prospective studies: an example of a nested case-control

TBC-10: *Seunghwan Jung, Soobok Joe and Hojung Nam*
Analysis of Functional Impacts on Massive Cancer Mutation Data

TBC-11: *Tun-Wen Pai*
Identifying Cross-Species Simple Sequence Repeat Biomarkers

TBC-12: *Parul Sharma, Sujata Sharma, T.P Singh and Punit Kaur*
Systematic and integrative analysis of large gene/protein interaction network for Rett syndrome

TBC-13: *Haein An and Chang-Bae Kim*
Expression profiling using RNA-seq for identifying developmentally regulated genes in Daphnia pulex

Posters Session

TBC-14: *Jun Kang, Hee Jin Lee, Ho Yun Lee, Jeong Hee Lee, Hajeong Lee, Guhyun Kang and Joon Seon Song*
Loss of the Heterochromatic X Chromosome in High Grade Ovarian Serous Carcinoma

TBC-15: *Dukyong Yoon, Dong Ki Kim, Eun-Young Jung, Sean Hennessy, Hyung Jin Choi, Ju Han Kim and Rae Woong Park*
How should we normalize laboratory results from multiple institutes to combine clinical data for unbiased

TBC-16: *Jee Yeon Heo, Hae-Seok Eo, Yong-Jin Choi and Hyung-Seok Choi*
miSeqaid: A pipeline for the analysis of microRNA sequencing data

TBC-17: *Hyun-Hwan Jeong, Sangseob Leem and Kyubum Wee*
High-order epistatic interaction detection using clique finding algorithm in genome-wide association studies

TBC-18: *Andrea Ganna, Donghwan Lee, Erik Ingelsson and Yudi Pawitan*
Rediscovery rate estimation for assessing the validation of significant findings in high-throughput studies

TBC-19: *Setia Pramana, Stefano Calza, Chen Suo, Fredrik Jonsson and Yudi Pawitan*
Molecular Subtyping of Breast Cancer using RNA-Sequence Data

TBC-20: *Alexander Ivliev, Marina Bessarabova and Yuri Nikolsky*
Application of pathway descriptors to detect similarities between human diseases

TBC-21: *Kyung-Sik Ha, Jin-Muk Lim and Hong-Gee Kim*
Development of microarray analysis automation system in Cytoscape plugin

TBC-22: *Petroula Proitsi, Richard Dobson, Cristina Legido-Quigley and John Powell*
Plasma metabolites as Alzheimer's Disease (AD) biomarkers

TBC-23: *Dhanusha Yesudhas, Suresh Panneerselvam, Shaherin Basith and Sangdun Choi*
Insight into the Binding Mode Analysis of Combinatorial Cancer Drugs with Cytochrome P450

TBC-24: *Je-Keun Rhee, Honglan Li, Byoung-Tak Zhang, Kyu-Baek Hwang and Soo-Yong Shin*
Phasing haplotype of a single individual by evolutionary algorithm

TBC-25: *Sol Moe Lee, Myungguen Chung, Kyu Jam Hwang, Young Ran Ju, Jae Wook Hyeon, Jun Sun Park, Chi-Kyeong Kim, Sangho Choi, Jeongmin Lee and Su Yeon Kim*
Biological network inference for allelic differences between familial CreutzfeldtJakob disease (fCJD) patients with E200K and Healthy individuals

TBC-26: *Chulbum Park, Se Mi Lee, Seong Eun Park and Ji-Yeob Choi*
A meta-analysis of pharmacogenetic studies of ABC and SLC transporters among cancer patients

Posters Session

TBC-27: *Fabian Buske, Phillippa Taberlay and Susan Clark*
The value of controls in peak calling from ChIP-seq experiments

TBC-28: *Fabian Buske, Susan Clark and Denis Bauer*
NGSANE - A HPC Processing Framework for Terabyte-scale Sequencing Data

TBC-29*: Hans-Joachim Sonntag, Mattia Prosperi, Iain Buchan, Angela Simpson and Adnan Custovic*
Evolution of IgE Sensitization Profiles for Timothy Grass and House Dust Mite Allergens

TBC-30*: Darius Coelho and Lee Sael*
Gene Expression Similarity between Breast and Prostate Cancer

TBC-31*: Yukyung Jun, Kyungsun Choi, Sanghyuk Lee and Wankyu Kim*
Systematic discovery of disease-associated miRNAs using an integrated network approach

TBC-32*: Jin-Muk Lim, Hong-Gee Kim and Ju-Hong Jeon*
FUT8 play an important role as a glucose metabolic agent in EML4-ALK Fusion NSCLC

TBC33*: Taehyung Kim, Leonardo Salmena and Zhaolei Zhang*
Landscape of ceRNAs in human genome and their potential role in cancer

TBC-34: *Geoff Macintyre*
Integrated genomics for lethal prostate cancer: exploring the subclonal evolution of metastatic prostate cancer and identification of hormone driven structural rearrangements

Posters Session

TBC-1: The implications of RBBP6 in various types of cancer

Firdous Khan[1,*] and Ashley Pretorius[1,*]

1 University of the Western Cape, South Africa

Abstract
Background: The 250 kDa RBBP6 protein was found to bind both p53 and Rb1 tumor suppressor proteins. In addition, RBBP6 has been associated with multiple biological functions, such as mitosis, mRNA processing, translation and ubiquitination.
Objectives: Using an in silico approach to identify RBBP6 binding partners (BPs). The information will be used to investigate the relation between RBBP6 and its bps and to further probe their in various cancer types.
Materials & Methods: RBBP6 was used as input to identify its BPs. This was followed by expression profiling across several cancer experiments. Lastly promotor content analyses was carried out to establish gene regulatory networks based on functional annotation (FA) and de novo motif prediction.
Results: In the current study 20 bps were identified for RBBP6. Expression profiling revealed RBBP6 and its are BPs differentially expressed in 14 cancers. Whilst FA analyses indicated that they are involved in similar biological processes such as regulation of apoptosis, programmed cell death etc.. . De novo motif discovery revealed 10 regulatory elements present in the promoters of RBBP6 and its BPs.
Discussion: Differential expression in many cancers and the association with the aforementioned FAs indicates a strong implication in cancer progression. The regulatory elements identified are directly linked to the FAs identified, validating the co-expression relationship between RBBP6 and its BPs.
Conclusions: The study showed that RBBP6 and its BPs share FAs, and common regulatory elements, inference can thus be made that they are highly involved in the progression of cancer

TBC-2: MELLO: Medical Life-Log Ontology

Hye Hyeon Kim[1], Soo Youn Lee[1], Su Youn Baik[1], Kye Hwa Lee[1] and Ju Han Kim[1,*]

1 Seoul National University Biomedical Informatics, Seoul National University College of Medicine, Seoul 110799, Korea

Abstract
Expectation of utilizing quantified-self data in medicine is increasing as accurate and reliable medical monitoring is possible through many body tracking devices. It has led to the development of more lifelogging devices, but also has caused uncontrolled generation of lifelogging term even for the devices having the same function. Computational analysis of lifelogging data has been hampered by lack of adequate and integrated lifelogging terms and related ontology so far. Therefore, we developed a MEdical Life-Log Ontology (MELLO) with over 500 terms to overcome this problem. Based on the core 50 data sets extracted from 25 body tracking devices and related mobile apps, first we searched and extracted enriched lifelogging terms from SNOMED-CT, having scattered lifelogging. Then we classified them manually into 7 major categories with hierarchical structure by three curators. We completed the MELLO as annotating each term with synonyms from UMLS, and definitions from Wikipedia. Our ontology was successfully validated by applying it to two different devices performing the same function. We show that the MELLO is able to integrate the different lifelogging terms with the same semantic for personal lifelogging data analysis.

TBC-3: Investigation on Gene Expression Patterns of Cardiac Myocyte Hypertrophy using Coexpression Network Analysis

Junbeom Kim[1], Jun Hyuk Kang[1] and Ho-Jin Choi[1,*]

1 KAIST, Republic of Korea

Abstract
Heart failure is a complex and multifactorial disease, which threatens one's life. Normal hearts are usually triggered by insults and derived to hypertrophy and heart failure. In this paper, extracting different gene expression patterns between normal, hypertrophy, and heart failure hearts using coexpression network analysis is performed. Generally, differentially expressed gene method is used for this kind of problem. However, conventional method compare between only individual gene, while coexpression network analysis consider correlation between genes and compare between the modules which are sets of genes. The contributions of this work are: 1) Applying coexpression network analysis framework to different target disease, heart failures; 2) Defining a new scheme for identifying and validation of modules. The coexpression network analysis framework is originally applied to hepatocellular carcinoma to extract differentially expressed genes during development of the disease. Here, the new scheme is proposed to distinguish the stages of heart failure.

TBC-4: Tell me your pathways

Frida Belinky[1,*], Gil Stelzer[1], Simon Fishilevich[1], Shahar Zimmerman[1], Marilyn Safran[1,2] and Doron Lancet[1]

1 Departments of Molecular Genetics, Weizmann Institute of Science, Rehovot, Israel
2 Departments of Biological Services, Weizmann Institute of Science, Rehovot, Israel

Abstract
A key annotation facet for a gene is the list of biological pathways it belongs to. However, the flat pathway list is of limited utility, due to a high degree of intra- and inter-source redundancy and inconsistency.
Striving to convey an integrated, internally consistent view of biological pathways per gene, we have clustered

3840 pathways from 12 sources into ~1600 super-pathways including singleton pathways.
This resulted in a collection of manageable super-pathways, each with no more than 80 members, and with optimal inter-cluster orthogonality.
Pathway expression, based on averaging gene expression binary vectors, reveals the super-pathway pattern of expression across 16 tissues.
Pathway evolution, inferred from genes orthology, reveals that most of the human pathways evolved mainly in three evolutionary time points:
(1) In the last universal common ancestor (LUCA).
(2) In the ancestor of eukaryotes.
(3) In the ancestor of Metazoa (animals).
Interestingly, super-pathways that are highly expressed in the liver are enriched in group (1), while super-pathways that are highly expressed in the testes are enriched in group (2).

TBC-5: Computational Morphoproteomics: Inferring Biological Relationships from Resource Description Framework Networks

Dmitriy Shin[1,*], Gerald Arthur[1,3], Mihail Popescu[2,3,4], Dmitry Korkin[3,4] and Chi-Ren Shyu[3,4]

1 University of Missouri, School of Medicine, Department of Pathology and Anatomical Sciences, Columbia, MO 65212, United States
2 University of Missouri, School of Medicine, Department of Health Management and Informatics, Columbia, MO 65212, United States
3 University of Missouri, Graduate School, MU Informatics Institute, Columbia, MO 65211, United States
4 University of Missouri, College of Engineering, Department of Computer Science, Columbia, MO 65211, United States

Abstract
Morphoproteomics is an emerging field aimed at systems-level identification of protein circuitries in a personalized medicine setting. Morphoproteomics is based on comprehensive analysis of immunohistochemical protein expression patterns in individual patient cases. A number of morphoproteomic studies have demonstrated better clinical outcomes and potential to improve therapeutics and diagnostics, also known as theranostics. A standard morphoproteomics practice, however, is heavily dependent on the expert knowledge and is therefore prone to inter- and intra- observer variability, which can undermine its widespread usage.
We propose a computational approach to improve traditional morphoproteomics by utilizing vast amounts of curated biological knowledge. First, we transform this knowledge into Resource Description Framework (RDF) knowledge networks using description logic inference and biological ontologies. Second, inspired by the ideas from the probabilistic causal theory, we introduce a method to traverse these networks and infer the biological mechanisms relevant to the case. Finally, the inferred information is presented in the form of diagrams for clinical decision-making. As a proof-of-concept, we have applied the formalism to the data from a clinical case of Acute Lymphoblastic Leukaemia performed using traditional morphoproteomics. The diagram inferred by our method shows high level of concordance with the human derived morphoproteomic diagram. The correlated expression of AKT, NF-kappa-B (nuclear) and BCL-2 proteins and the activation of an anti-apoptotic mechanism were noted by the experts in this case. The same flow of events was inferred by our computational approach.
We, therefore, conclude that our approach could provide the important advancements to the clinical implementation of morphoproteomics. A comprehensive assessment of the approach with more experimental data will be conducted to further explore its clinical utility.

TBC-6: A quantitative mixture model for transcriptome prediction

Qing Zhang[1], Xiaodan Fan[1] and Dianjing Guo[1,*]

1 The Chinese University of Hong Kong, Hong Kong

Abstract
Although many computational methods have been widely adopted to infer the transcription regulatory networks (TRNs), quantitative models that accurately predict the dynamic behavior of genes based on gene expression data are still in need by both wet-lab experimental design and synthetic biology.
In the present work, we propose a quantitative mixture model for transcriptome inference under a wide range of experimental conditions. Using cross-validation on a E.coli transcriptome data, the prediction power of the proposed model was estimated under various system perturbations, such as, gene knock-out, gene over-expression, and network rewiring. By linking a new experimental condition to the known conditions, the model can be used to reveal the possible functional relationships between different conditions. In addition, the model can also be extended to generate benchmark synthetic transcriptome data for the evaluation of TRN inference algorithms. The good performance of this method allows its wide application in synthetic biology system redesign and in biological experimental design.

TBC-7: De novo genome sequencing project of Korean native pig:Current status of genome assembly and annotation

Won-Hyong Chung[1,*], Namshin Kim[1], Kyung-Tai Lee[2] and Tae-Hun Kim[2]

1 Korean Bioinformation Center, Korea Research Institute of Bioscience and Biotechnology, Korea 2 National Institute of Animal Science, Rural Development Administration, Korea

Abstract
De novo genome sequencing of a Korean native pig: current progress and future works Starting from the panda genome sequencing project, genome sequencing using only NGS technology has been widely adapted to

mammalian genome sequencing projects. The pig is one of the most important food sources and one of the oldest forms of livestock. Even though two pig breeds, Duroc swine and Mini-pig, were sequenced previously, it is important to sequence local breeds to elucidate their genomic features. Korean native pig assumed to be come to Korea via north China around 2000 years ago. It has long black coarse hair, long straight nose, and small body weight (approximately 70 kg at adult). Here we report the progress of the de novo genome sequencing project to make a reference genome sequence of Korean native pig. We sequenced various kinds of libraries (170 bp, 300 bp, 400 bp, 500 bp and 600 bp insert paired-end; 2 Kbp, 5Kbp, 7Kbp and 10Kbp insert mate-pair) to approximately 136x coverage using Illumina GA IIx and HiSeq 2000 platforms. Small amount of 20Kb long insert library (~0.1x) was sequenced using 454 GS FLX platform. De novo sequence assembly was performed on these sequence sets using AllPaths-LG. After filling gaps using GapCloser in SOAPdenovo2 package, we applied RACA pipeline which curates and rearranges scaffolds using comparative genomic information. This resulted in 357 scaffolds totaling 2.52 Gb with a mean scaffold size of 7 Mb and N50 size of 17.2 Mb covering over 95% of the Duroc swine genome (excluding unplaced scaffolds). The assembly result is superior to the Mini-pig genome's (1,138,136 scaffolds, N50 size of 5.4 Mb). Korean pig draft genome will be a good resource for genome-wide comparative genomics between pig breeds or novel gene identification.

TBC-8: Systems Biology Integrative approach uncovers newer molecular targets in Metachromatic Leukodystrophy

Punit Kaur[1,*], Parul Sharma[1], Sujata Sharma[1] and T. P. Singh[1]

1 Department of Biophysics, All India Institute of Medical Sciences, New Delhi, India

Abstract

Metachromatic Leukodystrophy (MLD) is a neurological disorder caused by deficiency of the enzyme arylsulfatase A (ARSA). This disease impairs the growth or development of the myelin sheath. Mutation or absence of ARSA may cause the accumulation of sulfatides in many tissues of the body, eventually destroying the myelin sheath of the nervous system. The molecular interaction analysis of MLD was carried out using Cytoscape tool and its plug-ins. The functional modules and potential drug targets were identified as highly interconnected sub-graphs in the network. Molecular functions (gene ontology) of these genes were studied using BiNGO implemented in Cytoscape 2.8. DAVID, an online bioinformatics tool, was used for Pathway and disease enrichment analysis to get deeper insight into the molecular mechanism of MLD. The highest ranking sub-network was found to have 126 genes, and in addition to ARSA, four more genes were found to be highly inter-connected, namely SMAD9, PSAP, BMPR2 and UBE3A, which may play a major role in pathogenesis for this disease. The genes which were found to be potential drug targets for this disorder are TAF1, SMAD2, BRCA1, HNF4A, AR, SMAD9, CDC2, RB1, UBC, CDK2, UBB, PSAP, CDC23, MYC, MNAT1, CCNH, CDK7. The MLD initiating genes and other important candidate genes were found to be mostly involved in the Binding process and Catalytic activity. This analysis has lead to the identification of experimentally inadequately explored genes which are currently not reported in MLD physiopathology. Additionally major pathways likely to be affected in MLD include sphingolipid metabolism, lysosome, proteolysis, arrhythmogenic right ventricular cardiomyopathy (ARVC). Thus through a disease enrichment analysis we corroborated that MLD is not only associated with neuronal degeneration but also has probable links with cancer, metabolism and immune system.

TBC-9: Measuring DNA methylation in large epidemiological prospective studies: an example of a nested case-control study of breast cancer using the Illumina Infinium 450k BeadChip array

Chol-Hee Jung[1], Gianluca Severi[2], Melissa Southey[3], Dallas English[4], Andrew Lonie[5], Helen Tsimiklis[3], John Hopper[4], Graham G Giles[2] and Laura Baglietto[2,*]

1 Life Sciences Computation Centre, Victorian Life Sciences Computation Initiative, Carlton, Victoria, 3010, Australia 2 Cancer Epidemiology Centre, Cancer Council of Victoria, Melbourne, Australia 3 Genetic Epidemiology Laboratory, Department of Pathology, The University of Melbourne, Australia 4 Centre for Molecular, Environmental, Genetic and Analytic Epidemiology, School of Population Health, University of Melbourne, Australia 5 Life Sciences Computation Centre, Victorian Life Sciences Computation Initiative, Carlton, Victoria, 3010, Australia

Abstract

DNA methylation is a key epigenetic mechanism that regulates gene expression and is known to be involved in many human diseases including cancer. The development of new technologies to measure genome-wide DNA methylation makes it possible to conduct large epidemiological studies to test multiple hypotheses of association between methylation and disease. The major challenges posed by this type of study include the use of DNA from archival biospecimens of different type (e.g. dried blood spots, lymphocytes, buffy coats), handling missing values and controlling for batch effects. In this paper we discuss these challenges using the example of a prospective case-control study of breast cancer nested within the Melbourne Collaborative Cohort Study.

TBC-10: Analysis of Functional Impacts on Massive Cancer Mutation Data

Seunghwan Jung[1,§], Soobok Joe[1,§] and Hojung Nam[1,*]

1 School of Information and Communications, Gwangju Institute of Science and Technology, 123, Cheomdangwagi-ro, Buk-gu, Gwangju, 500-712, Republic of Korea
§ Equal Contribution

Abstract
A genetic mutation is a change of the nucleotide sequence of the genome of an organism. Mutation can result in several different types of change in sequences: (i) a change in one DNA base pair that results in the substitution of one amino acid for another in the protein made by a gene (missense mutation), (ii) a change in one DNA base pair that makes DNA sequence prematurely signals the cell to stop building a protein (nonsense mutation), (iii) changes the number of DNA bases in a gene by adding a piece of DNA (insertion), and (iv) changes the number of DNA bases by removing a piece of DNA (deletion), and so on.
The importance of the genetic mutations as factors of human diseases has been known for many years. Especially, mutations have a major role in initiation and development of cancer. In general, a common model view defines two classes of mutations in cancer, driver and passenger mutations. A driver mutation is causally implicated in oncogenesis. It has conferred growth advantage on the cancer cell and has been positively selected in the microenvironment of the tissue in which the cancer arises. In the other hand, a passenger mutation has no contribution to cancer development. In this sense, discovering functionally important mutations, including clear 'drivers' is one goal of genome resequencing studies. Thus, in this work, we analysis massive cancer mutation data sets by using the conventional analysis tools to give statistics of how many mutations detected in cancer could have potential to be classified into driver mutations, and their patterns in various types of cancer. Here we used cancer mutation information collected from the COSMIC database, The Cancer Cell Line Encyclopedia (CCLE), and The Cancer Genome Atlas (TCGA) project.

TBC-11: Systematic and integrative analysis of large gene/protein interaction network for Rett syndrome

Parul Sharma[1], Sujata Sharma[1], T.P Singh[1] and Punit Kaur[1,*]

1 Department of Biophysics, All India Institute of Medical Sciences, New Delhi, India

Abstract
Rett syndrome is a neurodevelopmental disorder of the grey matter of the brain that exclusively affects females. Mutations in the methyl-CpG-binding protein 2 gene (MECP2) found on the X-chromosomes is the major cause of Rett Syndrome. Very few drugs with low efficacy have been reported in the literature for Rett syndrome. Additionally, there exists a complete lack of knowledge about its gene Co-Expression network and pathogenesis. System networks are a central paradigm in biology which help in identifying new drug targets which in turn can generate a greater in-depth understanding of the mechanism of diseases. In an effort to explore drug targets, we have implemented a computational platform that integrates gene-gene interactions, differentially expressed genome and literature mining data to build comprehensive networks for drug-target identification. We used Cytoscape and its various plugins for prediction of the probable drug targets, to study the expression of genes in various biological processes and to identify highly interconnected clusters of genes. We have not only confirmed the well known relationship between this syndrome and neurodevelopmental disorder but also identified statistically significant relationships with other biological processes such as cell apoptosis, metabolic processes and many signalling pathways that affect the nervous system, musculo-skeletal system, respiratory system, excretory system and circulatory system. These multi-system complex thus play crucial role in the onset and pathogenesis of Rett Syndrome. Gene Ontology (GO) enrichment analysis was performed in all the obtained clusters. GO analysis exposed the significant molecular functions such as histone deacetylase binding, transcription factor binding and transcriptional co-repressor activity which were found to be associated with the genes that are known to play an important role. It also revealed various important biological functions associated with the highly interconnected hubs in the network. We succeeded in detecting some well known related genes such as MECP2, HDAC1, SIN3A, DNMT1, RCOR1 and NTNG1 together with we also identified GD1, TNF, PAK1, ADIPOQ and CAP2 that have been poorly explored or unknown in the current state of art of Rett syndrome.

TBC-12: Identifying Cross-Species Simple Sequence Repeat Biomarkers

Tun-Wen Pai[1,*]

1 Department of Computer Science and Engineering, National Taiwan Ocean University, Keelung 20224, Taiwan

Abstract
Simple sequence repeats (SSRs) are DNA segments with continuously repeated basic pattern of length from one to six nucleotides. SSRs are not only used as genetic markers in evolutionary studies but also play an important role in gene regulatory activities. Reports have revealed that SSR mutation or expansion may cause earlier symptoms of genetic diseases and lead to serious illness. Therefore, identifying and predicting functional SSRs through cross-species comparison are helpful for understanding the evolutionary mechanisms and associations between genes and functional SSRs, and the identified important biomarkers could be applied to further studies in genetic diseases, gene therapy, and breeding for various species. Due to the abundant number of SSRs, it is difficult to identify functional SSRs by featuring the only information of length and basic pattern from a single genome dataset. Hence, this study proposed a cross-species comparative approach and integrated with a tag cloud visualization technique for SSR biomarker identification. Tag Cloud representation utilizes different font sizes and colors for displaying the relationships

between genes and retrieved SSRs. Here, the SSR database was established by selecting 12 frequently used model species which are clustered into mammal and marine species clusters. Users are required to provide a set of genes or simply input keywords for gene selection from the designed system automatically. The proposed system could identify those extra conserved or unique SSRs through cross-species orthologous genes comparison. To demonstrate system performance, four testing gene sets were applied: (1) all orthologous genes from 12 model species and each gene possessing sequence identities higher than 80% compared with human genome; (2) 17 skeletal development related genes among mammal and marine species clusters; (3) a functional related gene set from a GO term of "embryonic cranial skeleton morphogenesis"; (4) a gene set of all well-known genetic diseases associated with SSR biomarkers. From these testing gene datasets, the system provided effective and efficient approaches for identifying conserved and exclusive SSR biomarker candidates through a friendly designed interface. Besides, the last testing dataset successfully demonstrated that the well-known genetic diseases were indeed associated with the retrieved ultra-conserved SSR biomarkers. Through statistical analysis and enhanced tag cloud representation on functional related gene sets and cross-species clusters, it can be noticed that the patterns, loci, colors, and sizes of identified SSR tags possess high correlations with gene functions, SSR pattern qualities and the numbers of conserved species.

TBC-13: Expression profiling using RNA-seq for identifying developmentally regulated genes in Daphnia pulex

Haein An[1] and Chang-Bae Kim[1,*]

1 Department of Life Science, Sangmyung University, Seoul 110743, Korea

Abstract

To identify genes controlling the developmental stages of Daphnia pulex, we determined gene expression profiles in three developmental stages, late embryo, 1st~3rd instars and 4th~5th instars by using RNA-seq technique. Gene expressions in 1st~3rd instars were more similar to 4th~5th instars than to the late embryo. We suggested that the most distinct stage in the developmental process was late embryo. Differentially expressed genes (DEGs) were discovered by comparing gene expressions of the late embryo with those of the post-embryonic stages. 3,562 genes were up-regulated and 3,332 genes were down-regulated in the embryonic stage. A hierarchical clustering of the DEGs generated two clusters: up-regulated genes and down-regulated genes in the embryonic stage. The DEGs were enriched with GO categories. Late embryo had higher activity in synapse, transcription regulator activity and molecular transducer activity. In the post-embryonic stages, membrane-enclosed lumen, envelop, reproduction and others were highly expressed. Genomic studies from multiple developmental stages are needed for elucidating developmental mechanisms.

TBC-14: Loss of the Heterochromatic X Chromosome in High Grade Ovarian Serous Carcinoma

Jun Kang[1,*], Hee Jin Lee[1], Ho Yun Lee[1], Jeong Hee Lee[1], Hajeong Lee[1], Guhyun Kang[1] and Joon Seon Song[1]

1 Training Program of Certified Physicians in BioMedical Informatics (CPBMI), Korea

Abstract

Introduction: Loss of the heterochromatic X chromosome occurs in certain breast and ovarian cancers. Mitotic segregation errors was thought to be most common mechanism. However, genome-wide deficits in heterochromatin maintenance and dysfunction of BRCA1 were suggested as alternative mechanisms of loss of the heterochromatic X chromosome. We investigated the correlation between the status of loss of the heterochromatic X chromosome and genome wide methylation status and BRCA mutation in ovarian high grade serous carcinoma.

Methods: We analysed X choromosome heterochromatin indicators including XIST and methylation at X chromosome in 164 ovarian high grade serous carcinoma of TCGA data (normalized RPKM of IlluminaHiSeq_RNASeqV2 for XIST level and beta value of Illumina Human Methylation 27k for methylation value). Genome wide methylation status and BRCA mutation were analysed with X chromosome heterochromatic status.

Results: XIST RNA varies in ovarian high grade serous carcinoma. After sorted by RPKM of XIST, X chromosome methylation pattern was vaguely divided into two groups at the level of 1592 RPKM of XIST. Low XIST RNA group accompanied hypomethylation of X chromosome, but not somatic chromosomes. There is no differences BRCA1 mutation between the two groups.

Conclusion: Some of high grade ovarian serous carcinomas have loss of heterochromatic X chromosome. Genome-deficits in heterochromatic maintenance or BRCA1 dysfunction seem not main mechanisms of loss of the heterochromatic X chromosome.

TBC-15: How should we normalize laboratory results from multiple institutes to combine clinical data for unbiased analysis?

Dukyong Yoon[1], Dong Ki Kim[2], Eun-Young Jung[3], Sean Hennessy[4], Hyung Jin Choi[5], Ju Han Kim[6] and Rae Woong Park[1,*]

1 Department of Biomedical Informatics, Ajou University School of Medicine, Suwon 443749, Korea
2 Department of Internal Medicine, Seoul National University College of Medicine, Seoul 110799, Korea
3 Centre for u-Healthcare, Gachon Univ. Gil Hospital, Incheon

405760, Korea
4 Center for Clinical Epidemiology and Biostatistics, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA 19104, USA
5 Department of Internal Medicine, Chungbuk National University College of Medicine, Cheongju 361711, Korea
6 Seoul National University Biomedical Informatics (SNUBI), Seoul 110799, Korea

Abstract
Combining clinical data, including laboratory results, from multiple institutions enables large-scale epidemiological studies. Although the increased adoption of electronic health record systems has facilitated this, there is no method for normalizing the combined clinical data from multiple institutes. Since the patient population at each hospital might differ, simply combining the data without considering population characteristics can lead to biased results. This study applied an age-stratification strategy to compensate for differences in age structure. To demonstrate the effect of age stratification, clinical laboratory results collected at two Korean tertiary teaching hospitals over a 5-year period were used. The hemoglobin level, which decreases with age, was selected for study. The hemoglobin readings were stratified by age from 0 to 79 years at 1-year intervals according to when the patient was tested. The results for patients 80 years old or older were aggregated as one group. For each group, the test results were normalized by standardization and then the data were recombined. The degree of normalization (distance) was measured using Kullback-Leibler divergence and the results were compared with normalized data standardized without using an age-stratification strategy. Misclassifications count, changes from a normal/abnormal state after normalization was also compared. As a result, the distance of the hemoglobin level distribution between the two hospitals was closer for the age-stratified data than for the normalized non-age-stratified data in both males and females (males, 0.051 vs. 0.345; females, 0.010 vs. 0.16). There were also fewer misclassifications in the age-stratified data: 167,737 (8.3%) vs. 205,554 (10.2%) for males and 223,683 (12.7%) vs. 234,921 (13.3%) for females. The difference in the laboratory data distribution between the two hospitals was normalized well when the population characteristics of the hospitals were considered. The consideration of characters in addition to age will provide us with more elaborately normalized distributions.

TBC-16: miSeqaid: A pipeline for the analysis of microRNA sequencing data

Jee Yeon Heo[1], Hae-Seok Eo[1], Yong-Jin Choi[1] and Hyung-Seok Choi[1,*]

1 BioIT Team, Future IT R&D Laboratory, LGE Advanced Research Institute, Seocho-gu, Seoul 137-724, Korea

Abstract
Small non-coding RNAs (ncRNAs) are functional RNA molecules and have a variety of processes from cell development and differentiation, stress responses to carcinogenesis by regulating gene expression. Currently, next generation sequencing (NGS) has been extensively used for small ncRNA profiling, especially for microRNAs (miRNAs), and several tools have been developed for analysing miRNAs expression profiles and predicting novel miRNAs. Here, we present a novel standalone tool, miSeqaid, for analysing miRNAs expression profiles and predicting novel miRNAs from NGS data. miSeqaid consists of four steps - quality control, read mapping, expression analysis and novel miRNA prediction. In step 1, 3'/5' adaptor sequences and contaminated sequences are trimmed and low quality and short reads are removed. In step 2, cleaned sequences are mapped to sequences of several categories (miRBase, RNA, Rfam, Repeat, Genome) using the Bowtie and Blast program. In step 3, miRNA expression values are normalized by RPM (Reads per Million) or quantile normalization. Subsequently, differentially expressed miRNAs are identified using Fisher's exact test or Wilcoxon-Mann-Whitey (WMW) test. P-values were adjusted for solving the multiple testing probleum using the Flase discovery Rate (FDR) and Bonferroni correction. In step 4, sequences that could be mapped to the reference genome but not assigned to the known miRNAs are used in the prediction of novel miRNAs. For the prediction of novel miRNAs, RNAfold, which have showed best performances on the calculation of secondary structures, is adapted. miSeqaid generates various result files, such as summary reports and analysis images including length distribution, read classification, Genome mapping, Repeat mapping, Rfam mapping, RNA mapping, expression analysis, novel miRNA prediction and so on. This tool is implemented by using PERL and R languases and Gnuplot was used to plot the analysis image.

TBC-17: High-order epistatic interaction detection using clique finding algorithm in genome-wide association studies

Hyun-Hwan Jeong[1], Sangseob Leem[1] and Kyubum Wee[1,*]

1 Department of Information and Computer Engineering, Ajou University, Suwon, S. Korea

Abstract
In recent years many studies have been proposed to detect association between multiple SNPs and complex diseases in case-control studies. However, most of the studies are not competent in detecting high-order epistatic interactions in genome-wide association studies (GWAS). Those methods are either only for two-way interaction or unable to cope with heavy computational burden of processing large-scale genotype data for detecting interactions of degree 3 or higher.
We propose a new method to find high-order epistatic interaction using clique-finding algorithm in a graph. The method runs as follows: (1) From every possible pair of SNPs, collect the pairs of SNPs that has significant mutual information value. Mutual information is between a pair of SNPs and the disease status. (2) Construct a graph from the collection of pair of SNPs. The vertices represent SNPs, and the edges represent the collected pairs of SNPs. (3) Find every possible clique in the graph and compute mutual information value of the SNPs in the clique. (4) Finally, sort the list of cliques that are found in step (3) by the mutual information value.
Our proposed method shows better performance than previous methods on simulated data. We also show that the method is feasible for large-scale genotype case-control data in real world.

The method detects several instances of significant high-order epistatic interaction for coronary artery disease (CAD) case-control data that is provided from Wellcome Trust Case Control Consortium (WTCCC).

TBC-18: Rediscovery rate estimation for assessing the validation of significant findings in high-throughput studies

Andrea Ganna[1], Donghwan Lee[1], Erik Ingelsson[2] and Yudi Pawitan[1,*]

1 Karolinska Institutet, Sweden 2 Uppsala University, Sweden

Abstract
It is common and advised practice in biomedical research to validate experimental or observational findings in a population different from the one where the findings were initially assessed. This practice increases the generalizability of the results and decreases the likelihood to report false positive findings. However, the question of what constitutes a successful validation has not been addressed rigorously.
We introduce a new measure called rediscovery rate (RDR) that quantifies the proportion of significant findings from a training sample that are replicated in a validation sample, and illustrate the benefits of using this measure for planning and assessing validation studies. In high-throughput studies, we show that the RDR is a function of false positive rate and power in both the training and validation samples. We derive its estimate based on the training data, assuming that the test statistics follow a mixture distribution. Furthermore, we explain how the RDR is connected to the power of the validation study in the single hypotheses testing and to the Winner's curse bias problem. We foresee two main applications. First, if the validation study has not yet been performed, the RDR can be used to decide the optimal combination between the proportion of findings taken forward to validation and the size of the validation study. Second, if a validation study has already been done, the RDR estimated using the training data can be compared to the observed RDR from the validation data: hence assess the success of the validation study. We use simulated data and real examples from metabolomics experiments in two large studies to illustrate the application of the RDR concept in high-throughput data analyses

TBC-19: Molecular Subtyping of Breast Cancer using RNA-Sequence Data

Setia Pramana[1,*], Stefano Calza[1,2], Chen Suo[1], Fredrik Jonsson[1] and Yudi Pawitan[1,*]

1 Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm 2 Molecular and Translational Medicine, University of Brescia, Italy

Abstract
Molecular classification of breast cancer into clinically relevant subtypes would help to improve diagnosis and adjuvant-treatment decisions. Given that more and more women are diagnosed with early-stage cancers, better specificity in treatment decision would save many of them from unnecessary side effects of the adjuvant treatment. However, cancer classification is still a big challenge. Rapid improvements in molecular analysis, e.g. by the application of next generation sequencing of cancer genomes, have the potential to bring deeper understanding as well as new biomarkers discoveries of the disease. The aim of this study is to use RNA-sequence data to classify breast-cancer patients into known molecular subtypes and get a deeper understanding of the disease.
RNA-seq data were generated from 329 breast cancer samples obtained from The Cancer Genome Atlas (TCGA) project. Based on the gene-level FPKM, a supervised classification for the RNA-seq data was performed by using k-nearest neighbour (k-NN) approach with Swedish breast-cancer data obtained from a classical microarray platform as training data (n=369). We selected 107 genes providing highly significant concordance rate (87%) between the supervised k-NN classification and the K-means unsupervised clustering within the TCGA samples. Most of the samples were classified as luminal subtypes (luminal A 35% and luminal B 21%), and the rest are basal (15%), ERBB2 (16 %) and normal like (13 %) subtypes. Our study shows that we can integrate gene expression from different platforms for molecular subtype discovery. The subtype assigned can be used later to obtain novel subtype-associated genes based on the RNA-seq data using all genes.

TBC-20: Application of pathway descriptors to detect similarities between human diseases

Alexander Ivliev[1], Marina Bessarabova[1] and Yuri Nikolsky[1,*]

1 Thomson Reuters, IP & Science. 5901 Priestly Dr., Carlsbad, CA 92008, USA

Abstract
Identification of molecular alterations (biomarkers) shared between distinct diseases is important for understanding of the underlying mechanisms and diversity of human pathologies and such applications as patients' stratification, translational research, precision medicine and drug repositioning. Standard computational approaches for detection of disease-to-disease similarities largely focus on gene level information, i.e. identification of genes and variants shared by the diseases. Unfortunately, clinically similar diseases or even individual cases of the same disease (e.g. cancer patient samples) can be strikingly different in gene expression and genetic alterations patterns. Nevertheless, one can detect higher-order similarities between human clinical phenotypes at the level of biological pathways. Detection of disease similarities at the level of pathways represents a broad field for data mining which largely remains unexplored by the previous studies.

TBC-21: Development of microarray analysis

automation system in Cytoscape plugin

Kyung-Sik Ha[1,*], Jin-Muk Lim[1] and Hong-Gee Kim[1]

1 Biomedical Knowledge Engineering Lab, Seoul National University, Korea

Abstract
In this study, we made a Cytoscape plugin that allows users to handle more easily the analysis of a microarray data. This plugin has been to automate the process of selecting only the probe with a significant value from microarray raw data. And this process have been made use the packages provided by R. Plugin that by using the protein-protein interaction database, it can now be represented as a network of relationships with other genes and gene the user has been selected. This whole process has been developed on the JAVA platform. This plugin can be accessed easier analysis of microarray data. Then, the user expected to be able to easily draw the network relationship of genes. This research was supported by MSIP (the Ministry of Science, ICT and Future Planning), Korea, under the IT-CRSP(IT Convergence Research Support Program) (NIPA-2013-H0401-13-1001) supervised by the NIPA(National IT Industry Promotion Agency)

TBC-22: Plasma metabolites as Alzheimer's Disease (AD) biomarkers

Petroula Proitsi[1,*], Richard Dobson[1], Cristina Legido-Quigley[2] and John Powell[1]

1 King's College London, Institute of Psychiatry, United Kingdom 2 King's College London, Institute of Pharmaceutical Science, United Kingdon

Abstract
Introduction: There is a need for a better understanding of the biological mechanisms underlying AD and the identification of biomarkers for early clinical diagnosis, progression and conversion. Metabolites are the final product of interactions between gene expression, protein expression, and the cellular environment and represent a more accurate approximation of the phenotype of an organism and complex biological processes.
Aims: The aim of this project is 1) to characterise the plasma metabolic profiles of AD patients, subjects with mild cognitive impairment (MCI), and controls and to utilize these metabolic profiles in order to identify diagnostic, conversion and progression biomarkers; 2) to integrate metabolic profiles with genetic, transcriptomic and proteomic data in order to improve classification/prediction.
Methods: Ultra Performance Liquid Chromatograpy/Mass Spectrometry was performed on plasma samples from 35 AD, 43 MCI & 45 controls (MassLynx- Waters). Samples were divided into Train (2/3 sample) and Test (1/3 sample) datasets. Machine learning approaches were used to classify AD, MCI & CTL.The analytes which predicted disease were identified and investigated further.
Results and conclusions: 1878 analytes were extracted and raw values normalized to the whole area mean. Following removal of analytes with <80% data, transformation and imputation, 573 analytes were analysed. The train dataset was used to tune the parameters of L(1)-L(2)-regularized regression (elastic net) using internal crossvalidation and the model was evaluated on the independent test set. A set of 34 analytes predicted AD with accuracy >75%. Including APOE, the most established AD risk gene, in the model increased accuracy to >83%. Logistic regression analyses showed that some of the individual analytes were associated with AD with p<10-4. Most analytes were associated with changes in lipid metabolism. Results for AD-MCI and MCI-CTL classifier's showed lower accuracy (<75%). Data integration using genetic, expression and protein data will improve the classifier performance.

TBC-23: Insight into the Binding Mode Analysis of Combinatorial Cancer Drugs with Cytochrome P450

Dhanusha Yesudhas[1,§], Suresh Panneerselvam[1,§], Shaherin Basith[1] and Sangdun Choi[1,*]

1 Department of Molecular Science & Technology, Ajou University, Suwon, 443-749, Republic of Korea
§ Equal contribution

Abstract
Combinatorial drug therapy is becoming a promising strategy in the treatment of cancer. However, the patient has an increased risk of suffering from an adverse drug-drug interaction (DDI). DDI is a situation where one drug inhibits the metabolism of another drug, thereby leading to an increased plasma concentration of either drug. The poor metabolism of the drug molecules by cytochrome P450 is one of the reasons for the drug-drug interaction in the combinatorial therapy. However, we have limited knowledge about the interaction of drug molecules with cytochrome P450. Hence, we have utilized computational docking to predict the drug-binding mode and assessed its stability using molecular dynamic simulation studies. Nine cancer drugs which are used in combinatorial therapy were selected from National Cancer Institute (NCI) Database. Previous studies have shown that CYP3A4 isoform metabolizes these drug molecules. Therefore, we performed docking for the selected drug molecules with CYP3A4. One hundred docking structures were generated for each drug molecule. Hydrogen bond analyses of molecular dynamic simulations were used to confirm the selected binding mode of drug molecules. The predicted binding modes of the drugs were found to have good correlations with the available experimental data. These studies will be useful for new drug development and also provide valuable insights in the metabolism of cancer drugs.

TBC-24: Phasing haplotype of a single individual by evolutionary algorithm

Je-Keun Rhee[1], Honglan Li[2], Byoung-Tak Zhang[1,3], Kyu-Baek Hwang[2] and Soo-Yong Shin[4,*]

1 Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul 151-742, Korea
2 School of Computer Science and Engineering, Soongsil University, Seoul 156-743, Korea
3 School of Computer Science and Engineering, Seoul National University, Seoul 151- 744, Korea
4 Department of Biomedical Informatics, Asan Medical Center, Seoul 138-736, Korea

Abstract
Although lots of genetic variations have been identified successfully, haplotype information which is a combination of alleles at adjacent locations on the chromosome can provide much crucial knowledge for whole-genome association studies. Previously, the haplotype were inferred from genotype information of population. Recently, with development of high-throughput sequencing (HTS) technologies, the approach to find haploid of a single individual have been drawn attention. Here, we present an evolutionary algorithm to assemble the haplotype of a single individual by combining its sequence reads. Based on heterozygous single nucleotide polymorphisms (SNPs), the haplotype phasing problem can be considered as a combinatorial optimization problem, and the evolutionary algorithm can effectively solve the computationally complex problem. We applied the proposed method to real whole-genome sequencing datasets from NA12878. The experimental results show our proposed approach can practically reconstruct the haplotype.

TBC-25: Biological network inference for allelic differences between familial Creutzfeldt-Jakob disease (fCJD) patients with E200K and Healthy individuals

Sol Moe Lee[1,§], Myungguen Chung[2,§], Kyu Jam Hwang[1], Young Ran Ju[1], Jae Wook Hyeon[1], Jun Sun Park[1], Chi-Kyeong Kim[1], Sangho Choi[1], Jeongmin Lee[1] and Su Yeon Kim[1,*]

1 Division of Zoonoses, Center for Immunology and Pathology, National Institute of Health, Korea Centers for Disease Control and Prevention, Cheongwon-gun, Chungcheongbuk-do 363-700, Republic of Korea
2 Division of Bio-Medical Informatics, Center for Genome Science, National Institute of Health, Korea Centers for Disease Control and Prevention, Cheongwon-gun, Chungcheongbuk-do 363-700, Republic of Korea
§ Equal contribution

Abstract
The human prion diseases are caused by an abnormal accumulation of misfolded prion protein in the brain. Inherited prion diseases including familial Creutzfeldt-Jakob disease (fCJD) are associated with the mutations of prion protein gene (PRNP). The glutamate to lysine substitution at codon 200 (E200K) in PRNP is the most common pathogenic mutation causing fCJD in the world, and a few cases with E200K have been reported annually in Korea. E200K pathogenic mutation alone is not regarded sufficient to cause prion diseases and unidentified necessary factors have been proposed to explain penetrance of E200K-dependant fCJD. In our previous study, a total 19 genes showed significant differences of genotypes between fCJD patients with E200K and non-CJD individuals. In this study, 19 genes were analyzed to identify biological pathways and relationship among the proteins encoded by the genes. Protein-protein interactions (PPIs) among the proteins encoded by 19 genes through the exome sequencing study were identified using the Michigan Molecular Interaction (MiMI) database and Prion Disease Database (PDDB), and then visualized using Cytoscape v2.8.3. Biological interactions were identified among 8 genes. All of them were linked not by direct interaction, but by 14 interactomes (PLG, TAF1, FRS3, etc). Biological interactions identified by PPI network were about complement and coagulation cascades, lysine degradation, neurodegenerative diseases, and so on. Our results implied that there are a possible co-regulation mechanism and candidate necessary factors of fCJD with E200K. These biological network data can be used for further investigation on the mechanisms of the genetical prion diseases.

TBC-26: A meta-analysis of pharmaco-genetic studies of ABC and SLC transporters among cancer patients

Chulbum Park[1], Se Mi Lee[2], Seong Eun Park[3] and Ji-Yeob Choi[1,*]

1 Seoul National University, Korea
2 Chonnam National University, Korea
3 Duksung women's University, Korea

Abstract
Membrane transporters can be major determinants of pharmacokinetic profiles of anticancer drugs. A meta-analysis was conducted to investigate the association of ATP-binding cassette (ABC) and solute carrier (SLC) transporter genetic polymorphisms with pharmacogenetic outcomes until Jan, 2012. Eligible studies involved cancer patients and compared genetic variants in the ABC and SLC transporters with information anticancer drugs and reported one of the following outcomes: overall survival, progression-free survival, response rate or efficacy, drug toxicity and pharmacokinetic parameters. A total of 158 publications were identified, of which 33 were deemed eligible for inclusion. For efficacy, 6 genes (ABCB1, ABCC1, ABCC2, ABCG2, SLC28A1 and SLC28A2) with 31 polymorphisms were analyzed and any gene was not significantly associated with the response. When stratified by cancer sites or anticancer drugs, ABCB1 variants decreased the risk of resistant rate among colorectal cancer (OR=0.67, 95% CI=0.47~0.96 for 5 reports) and among patients treated with nucleotide analogue (5FU or gemcitabine) (OR=0.67, 95% CI=0.52~0.86 for 11 reports). For toxicity, 5 genes (ABCB1, ABCC2, ABCC4, ABCG2, SLCO1B1) with 17

polymorphisms were analyzed and variants of ABCB1 were significantly associated with drug toxicity overall (OR=0.86, 95% CI=0.74~0.99). Any variants of ABC transporters also decreased the risk of drug toxicity, especially for GI related toxicity (OR=0.79, 95% CI=0.67~0.94 for 19 reports). For survival, patients with any variants of ABCB1 showed poor progression free survival compared to patients with wild types (HR=1.86, 95% CI=1.06~3.25). For pharmacokinetics, patients with ABCB1 variant homozygotes showed higher AUC (SMD=1.75, 95% CI=-0.01~3.51, p=0.051) and lower clearance (SMD=-4.46, 95% CI=-7.05~-1.86, p=0.001) compared to patients with wild types. Variants of ABC transporters were significantly associated with improved pharmacogenetic outcomes of anticancer drugs in a meta-analysis of multiple cancer sites.

TBC-27: The value of controls in peak calling from ChIP-seq experiments

Fabian Buske[1,*], Phillippa Taberlay[1] and Susan Clark[1]

1 Garvan Institute of Medical Research, Australia

Abstract
ChIP-seq is the method of choice for interrogating the DNA occupancy of proteins involved in gene regulation. Reliable assessment of ChIP enrichment (peak calling) requires sequencing of a matched control library (e.g. input DNA) to compensate for biases (copy-number alterations, sequence content, chromatin structure, antibody quality). However, for economic reasons matched control libraries are often sequenced at lower depth than the ChIP enriched sample or are not sequenced at all. It is therefore important to address if input sequenced controls are required for the accurate interpretation of ChIP-seq data and if so should input data be from matched control libraries or will unmatched input libraries suffice.
We investigated the effect of input sequencing controls on peak calling by contrasting matched controls with libraries generated from unmatched biological replicates or obtained from ENCODE project using the same cell lines. We considered the peaks generated from matched controls as the gold standard and assessed the accuracy of unmatched controls from the same cell lines to call the equivalent enriched regions with at least 50% overlap. We observe for all three interrogated histone marks (H3K9K14ac, H3K4me1, H3K27ac) that high accuracy can be achieved with unmatched input controls depending on the sequencing depth of the control library (Peakranger accuracy of 0.99 vs 0.97 vs 0.79 using 21, 13 or 5.5 mil. mapped reads, respectively).
Furthermore, investigating the base pair overlap of the enriched regions, we observe that the algorithm of choice has a greater impact than utilizing an unmatched control (average Jaccard similarity coefficient of 0.24 between Peakranger, Homer and Chromablocks using matched controls and 0.76 between matched versus unmatched control libraries using the same algorithm).
We therefore conclude that it is reasonable to use an unmatched control even from public data if there is high sequencing coverage. This has important ramifications in the processing of ChIP data using different antibodies to interrogate the same cell type.

TBC-28: NGSANE - A HPC Processing Framework for Terabyte-scale Sequencing Data

Fabian Buske[1,*], Susan Clark[1] and Denis Bauer[2,*]

1 Epigenetics Program, Cancer Research Division, Garvan Institute of Medical Research, Kinghorn Cancer Centre, Darlinghurst City, NSW 2010, Australia 2 Computational Informatics, CSIRO, North Ryde, NSW 2113, Australia

Abstract
The first steps of analysing sequencing data (2GS,NGS) have entered a transitional period where analysis steps can be automated in standardised pipelines. With constantly evolving technology, academic software will remain the methods of choice for cutting-edge data analysis. This makes setting-up and maintaining analysis pipelines labour intensive, as most tools do not comply with good software-development practice (i.e. good documentation, legacy support).
Many GUI-enabled tools, like Galaxy, address this issue but are commonly tailored to cater for biologist with only small numbers of experiments. However, with increasing study sizes, the capability of leveraging high performance compute clusters and processing libraries in parallel is paramount.
NGSANE is a lightweight, Linux-based, HPC-enabled framework that minimizes overhead for set-up and processing of new projects yet maintains full flexibility of custom scripting when processing raw sequence data. The framework separates project specific data from commonly used annotation files, scripts and software suites. NGSANE supports Sun-Grid-Engine and Portable-Batch-System job scheduling and can be operated in different modes for development and production thus enabling efficient and flexible processing of NGS data. It currently includes pipelines for adapter trimming, read mapping, peak calling, motif discovery, transcript assembly, variant calling and chromatin conformation analysis by tapping into various

TBC-29: Evolution of IgE Sensitization Profiles for Timothy Grass and House Dust Mite Allergens

Hans-Joachim Sonntag[1], Mattia Prosperi[2,*], Iain Buchan[2], Angela Simpson[2] and Adnan Custovic[2]

1 University of York, United Kingdom 2 University of Manchester, United Kingdom

Abstract
The study of immune responses to allergens has been revolutionised by the routine availability of component resolved diagnostics (ImmunoCAP ISAC) that measure

the specific IgE response towards many allergen components, including timothy grass and house dust mite. Using latent class analysis on data from the population-based Manchester asthma & allergy study (1,186 children followed up from birth, 899 undergoing ISAC IgE testing, 235 with full longitudinal information at age 5, 8 and 11 years), we confirmed the hypothesis of a "molecular spreading" pathway for timothy grass allergens, where sensitisation to the lead allergen Phl p 1 precedes a progression towards a full sensitisation to other Phl p components, with serum concentration increasing over time. Conversely, in the case of house dust mite, we found different pathways related to two distinct allergen groups (Der f 1 & Der p 1) and (Der f 2 & Der p 2). Longitudinally, from age 5 to 11 years, we could either observe a co-development trajectory of sensitisation towards these two groups or stabilisation towards a single group over time. Logistic regression was employed to demonstrate that all house dust mite sensitisation trajectories are significantly (0.05 level) associated with an increased risk of asthma at age 11 (odds ratios ranging from 4.5 to 8.6 as compared to the non-sensitisation pathway). Interestingly, there was a significant difference between the Der f/p 1 pathway and the Der f/p 2 pathway as predictors of eczema, with the former having a significant odds ratio of 3.2 [95%CI 1.2-8.0] as compared to the non-sensitisation pathway. Regression analysis with house dust mite exposure from early ages (<2 years) confirmed the association with longitudinal allergen trajectories, while house dust mite concentration at later ages showed no significant association with any of the three house dust mite sensitization trajectories.

TBC-30: Gene Expression Similarity between Breast and Prostate Cancer

Darius Coelho[1,2] and Lee Sael[1,2,*]

1 Department of Computer Science, State University of New York , Incheon 406840, Korea
2 Department of Computer Science, Stony Brook University, Stony Brook, NY 11794-4400

Abstract
Epidemiologic and phenotypic evidences indicate that breast and prostate cancer have high pathological similarities. Genes that are affected by both breast and prostate cancer are investigated to gain knowledge of the similarity between their pathology. Gene expression data extracted from RNA-seq experiment for breast invasive carcinoma (BRCA) and prostate adenocarcinoma (PRAD) retrieve from TCGA database (http://tcga-data.nci.nih.gov/) were analyzed. Iterative SVM-based ensemble gene selection method was used to select genes that discriminate cancer samples from normal samples. Iterative SVM-based gene selection methods enable correlated gene expressions to be considered simultaneously and ensemble approach stabilizes the selection. The selected gene sets were able to achieve classification accuracy of 90\% for BRCA and 93\% for PRAD. However, only two genes, Transglutaminase 4 (TGM4) and complement component 4A (C4A), were common in the BRCA and PRAD gene set. Based on the Ingenuity Pathways Analysis, although there are no specific associations known to the breast or prostate cancer, TGM4 has known association with the adenocarcinoma in general. Also, C4A do not have known association with the breast or prostate cancer. However, both genes have directly and/or indirectly association with multiple types of cancer and possibilities of being a drug target for both breast and prostate cancer could be found through guilt by association in pathway analysis. Although this information can be important in itself, since the two genes may likely be common genes associated with various types of cancer, further study is needed to confirm that breast and prostate cancer have high pathological similarity.

TBC-31: Systematic discovery of disease-associated miRNAs using an integrated network approach

Yukyung Jun[1], Kyungsun Choi[2], Sanghyuk Lee[1,*] and Wankyu Kim[1,*]

1 Ewha Research Center for Systems Biology (ERCSB), Ewha Womans University, Korea 2 Bio and Brain Engineering, KAIST, Korea

Abstract
miRNAs are thought to be promising diagnostic and therapeutic targets due to their frequent dysregulation in many human diseases. We develop a method which predicts disease-miRNA associations systematically, based on standard gene sets analysis (GSA) using an extensive series of gene signatures for 2,078 human diseases and 1,432 miRNAs. More than 30 types of independent evidences are integrated by performing ~24 million GSA comparisons. As a result, a generic disease-miRNA association network is constructed with >40,000 associations between 956 diseases and 772 miRNAs. It includes many human diseases such as rheumatoid arthritis, muscular dystrophy, Parkinson disease as well as various cancers, where miRNAs may have a critical role in pathogenesis.
As a validation of our model, the influence on cell proliferation is tested for ten candidate miRNAs using cell lines of glioblastoma multiforme (GBM), the most malignant form of brain cancer. Five miRNAs (50%) show a significant decrease in proliferation in multiple cell lines (eight miRNAs (80%) in at least one cell line). Also, some of the miRNAs show a significant correlation with proliferation rate, cell morphology and patient survival. It suggests that disease-associated miRNAs can be identified with a reasonable accuracy, overcoming our limited knowledge on miRNA targeting that is a major hurdle in miRNA functional studies. Our disease-miRNA network provides a foundation to elucidate the functional role of miRNAs in a wide range of human diseases.

TBC-32: FUT8 play an important role as a glucose metabolic agent in EML4-ALK Fusion NSCLC

Jin-Muk Lim[1], Hong-Gee Kim[1] and Ju-Hong Jeon[2]

1 Biomedical Knowledge Engineering Lab, Seoul National University College of Medicine
2 Department of Physiology, Seoul National University College of Medicine

Abstract
The EML4 (echinoderm microtubule-associated protein-like 4)-ALK (anaplastic lymphoma kinase) fusion-type tyrosine kinase is an oncoprotein found in 4 to 5% of non-small-cell lung cancers, and clinical trials of specific inhibitors of ALK for the treatment of such tumors are currently under way. However, patients with these cancers invariably relapse, typically within 1 year, because of the development of drug resistance. Herein, we compare affymetrix microarray of ALK positive set(n=11) and Triple Negative(EGFR/KRAS/ALK) set(n=68). Machine Learning, statistical method, GO analysis and network analysis method are used. FUT8 is found as ALK positive set specific target. FUT8 play an important role as a glucose metabolic agent in this cancer. Our results may help future experimental investigation to understand the signal process of ALK, FUT8 inhibitor combination therapy in non-small-cell lung cancer. [This research was supported by MSIP (the Ministry of Science, ICT and Future Planning), Korea, under the IT-CRSP(IT Convergence Research Support Program) (NIPA-2013-H0401-13-1001) supervised by the NIPA(National IT Industry Promotion Agency)]

TBC-33: Landscape of ceRNAs in human genome and their potential role in cancer

Taehyung Kim[1,2/sup>, Leonardo Salmena5,* and Zhaolei Zhang1,2,3,4,*]

1 Department of Computer Science, University of Toronto, Toronto, ON, Canada
2 The Donnelly Centre, University of Toronto, Toronto, ON, Canada
3 Department of Molecular Genetics, University of Toronto, Toronto, ON, Canada
4 Banting and Best Department of Medical Research, University of Toronto, Toronto, ON, Canada
5 Princess Margaret Hospital, Department of Medical Biophysics, University of Toronto, Toronto, ON, Canada

Abstract
MicroRNAs are small non-coding RNAs that govern many cellular processes by triggering RNA degradation or by inhibiting translation. Recently, miRNAs have been identified as key components of a novel RNA-RNA crosstalk mechanism, where RNAs influence each other's expression level by competing for a limited pool of miRNAs. This phenomenon, termed competing endogenous RNA (ceRNA), results in a positive correlation of expression between the competing transcripts. Despite its potential importance in global gene regulation, there have been very few efforts to systematically study the functional relevance of ceRNAs, especially in cancer. Herein, we aim to extend this limited knowledge by identifying novel ceRNA pairs or networks through the use of the latest microRNA target prediction methods and high-throughput sequencing technology. In particular, we are investigating in the role of pseudogenes in modulating the expression of their parental genes via the ceRNA mechanism. Pseudogenes, previously dismissed as "junk DNA", are genomic loci that resemble protein-coding genes but have lost any ability to code for a functional protein. By coordinating these two ideas, we hypothesize that gene-pseudogene (and gene-gene) regulation in cancer can be achieved through a ceRNA mechanism and this phenomenon includes, but extends well beyond the PTEN-PTENP1 paradigm. Here, we have identified a number of pseudogenes and proteincoding genes that have perturbed expression in tumour samples as compared to control tissue specimens. These perturbations will be evaluated for ceRNA potential, and their role in cancer progression. This work will not only demonstrate the existence of novel ceRNAs, but also extend our understanding on the origins and progress of cancer. A list of confirmed genes and miRNAs contributing as ceRNA networks implicated in cancer may serve as new therapeutic targets, and thereby allow development of a new means to modulate the expression of key cancer genes and to slow down cancer progression

TBC-34: Integrated genomics for lethal prostate cancer: exploring the subclonal evolution of metastatic prostate cancer and identification of hormone driven structural rearrangements

Geoff Macintyre[1,*]

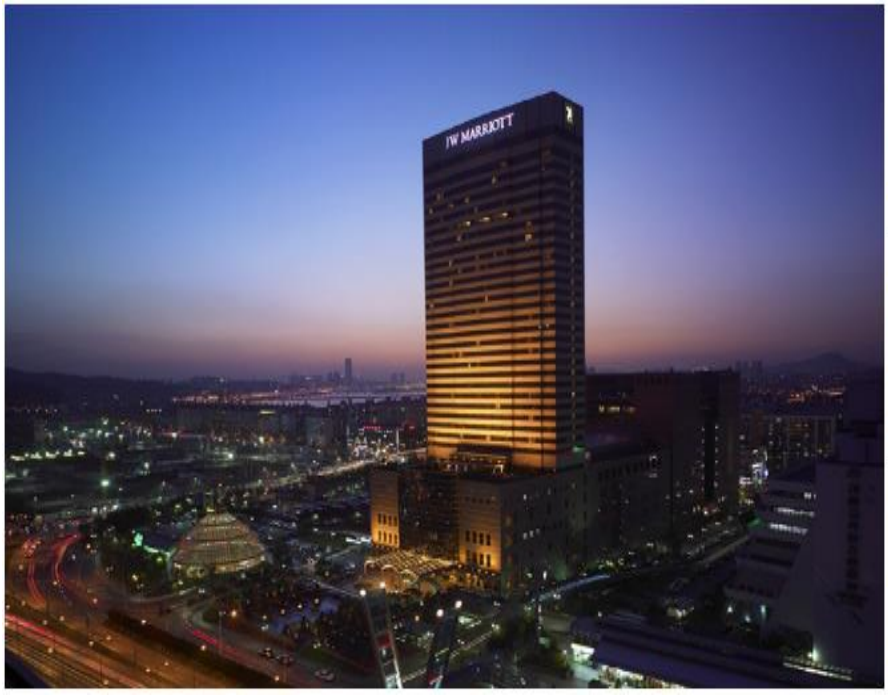
1 NICTA, The University of Melbourne, Australia

Abstract
Prostate cancer is the most diagnosed internal malignancy in the western world. While the majority of prostate cancers are non-lethal, there is currently no reliable approach to distinguish lethal from non-lethal prostate cancer at an early, curable stage. To help understand the molecular mechanisms drving lethal (metastatic) prostate cancer, we carried out molecular profiling of primary tumours and metastases from 7 patients using whole-genome sequencing, RNA-SEQ, Illumina 2.5M SNP Chip, and Illumina 450K methylation chip. These data have allowed us to model the subclonal evolution of these tumours and observe some interesting trends regarding structural rearrangements. In the case of structural rearrangements, we observed that a large proportion of the breakpoints were in close proximity to androgen receptor binding sites. Furthermore, when we looked at breakpoints in 11 other cancers from TCGA and ICGC projects, only breakpoints in those cancers which were hormone dependent showed an associated with androgen receptor. We observed the same results when we looked at estrogen receptor binding sites. Our results suggest that steroid hormone receptors may play a role in the formation of cancer driving structural rearrangements. In terms of subclonal evolution of metastatic disease, we currently only have preliminary results. However, we are already able to observe the effects of surgery and treatment on disease progression.

Venue

Location

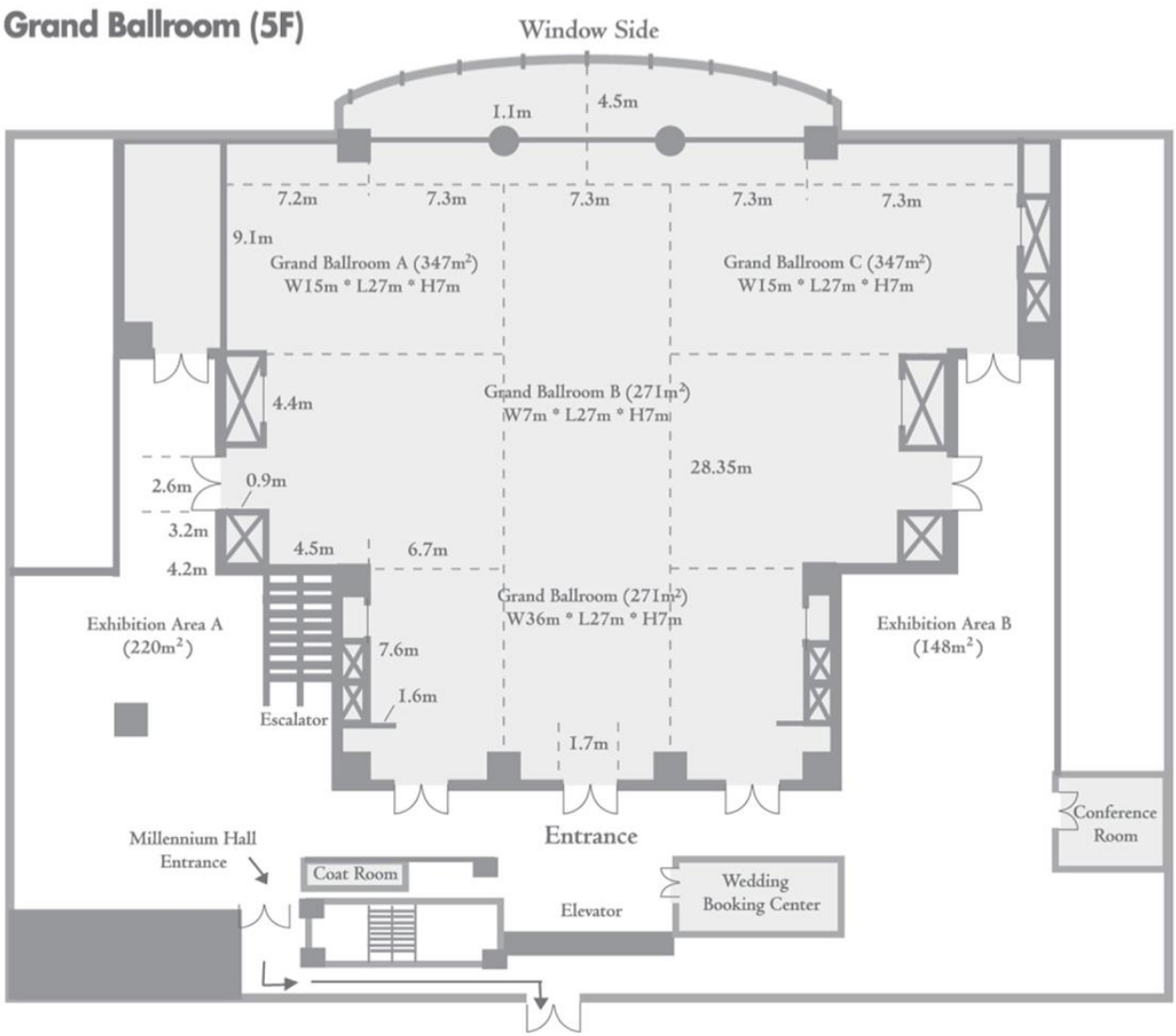
JW Marriott Hotel, Seoul, Korea



Map

Venue

Floor Plan



Tours

Recommended Tour Courses

Naksan Park

This course next brings you up to Naksan Park, a lovely patch of nature in the midst of the crowded city that offers some fine views. The ancient fortress-city of Seoul was laid out according to the traditional Pungsu-jiri-seol theories, which were adapted to Korea's circumstances by National Buddhist Master Doseon-guksa in the 9th Century from the principles of ancient Oriental Geomancy.

Changgyeonggung (Palace)

Built as "Suganggung" by King Seiong the great for his father, King Taejong, this palace was renovated and enlarged by King Seonjong in 1483, for his grandmother, mother and an aunt. Together with Changdeokgun, which is separated from Changgyeonggung by just a stone wall, the palace used to be called "Donggwol (eastern palace)". During the colonial period, the Japanese built a zoo, botanical garden, and museum on this site, and disgracefully called it Changgyeongwon (Changgyeong Garden). In 1983, however, the zoo and botanical garden were removed, and it regained its original name, Changgyeonggung. Myeongjeongjeon (Hall), Myeongjeongmun (Gate), and Honghwamun (Gate) show us the architectural styles popular in 17th century Joseon. With one ticket, visitors can access Jongmyo (Royal Shrine) Shrine via the overpass, as well.

Gyeongbokgung (Palace)

Gyeongbokgung (Palace), built in the 4th year of King Taejo's reign (1395), is where the Joseon Dynasty originated. It is the oldest and central palace of Joseon. It is for this reason that Gyeongbokgung is considered the most beautiful and biggest of the 5 palaces in Seoul, and praised as demonstrating the very height of architectural technology from the medieval period of Northeast Asia.

Parts of the palace were burnt down during "Imjin Waeran", the Japanese invasion of Korea (1592-1598). Since 1610 Changdeokgung played the role of jeonggung instead, until Heungseon Daewongun (father of King Gojong) restored Gyeongbokgung in 1865. Although the palace was seriously damaged once again, under Japanese Colonial rule, the current Korean government is continuing its predecessors' efforts to restore it to its original glory. Major attractions inside the palace include Geunjeongjeon (the main hall of the palace), Gyeonghoeru pavilions, Hyangwonjeong pavilion and the Amisan chimneys, all of which are valuable cultural and historical assets illustrating the essence of the traditional architectural design of the Joseon Dynasty.

There are several historical sites and artifacts of great value that can be found in every corner of Gyeongbokgung. Inside the palace are the National Palace Museum of Korea and the National Folk Museum of Korea to take you on a trip back into Korean history.

Tours

Insa-dong

Almost all foreign visitors to Korea make a trip to Insa-dong. This is because it is the place where “HAN brand,” the traditional culture of Korea as a whole, can be found. At the heart of HAN brand is the intangible nature of culture accumulated over a long course of history. For instance, HANok in Gyeongin Art Museum, HANji (traditional Korean paper) culture whose beauty resonates through your heart, HANbok (traditional Korean clothes) that captures the essence of Korean clothing culture and spirit, and HANsik (traditional Korean food) that exploited the science behind eco-friendly organic foodstuffs hundreds of years ago all make up HAN brand. In Insa-dong, you can learn about and truly enjoy all aspects of HAN brand. In Insa-dong, every alley and store has something traditionally Korean to show off. Known initially as an antique district, Insa-dong was designated as a bastion of traditional culture in 1988, when an array of modern art galleries joined the area. Since 1997, Insa-dong has been "a vehicle-free street" every Sunday and from 10 a.m. to 10 p.m., with no cars in sight, Insa-dong turns into a huge venue for outdoor festivals. Here you can simply indulge in a spot of window shopping, or better yet, purchase paintings, sculptures, ceramics and installation art works by some of the country’s most celebrated (not to mention the up and coming) artists. Unhyeongung(Palace), Jogyesa(Temple), Tapgol Park and Bosingak(Belfry) are all within walking distance from Insa-dong.

N Seoul Tower

Namsan is a witness to the 600 years of Seoul as the capital of Korea. Indeed, Namsan holds Seoul’s history. The Seoul Fortress Wall that crosses Namsan and the Smoke-signal Station on Namsan demonstrate that Namsan was once a long serving strategic military point. When the parts of the Seoul Fortress Wall and Bongsudae that have been lost are rebuilt, Namsan will become an even more historically valuable place.

Here, one can find the perhaps the best panoramic view of Seoul. An observation deck, restaurants and gift shops are located within the N Seoul Tower. In the evenings, a special show of lights, “Electronic Fire,” takes place.

Conference App

Download now the TBC / ISCB-Asia 2013 app for iOS and Android.

http://www.snubi.org/app/tbc2013/

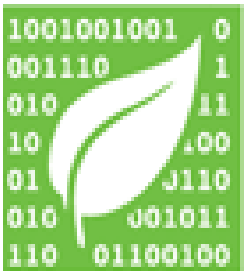
TBC 2013

Informatics Journals Supporting TBC

- **JAMIA** (Journal of American Medical Informatics Association), IF: 3.974 (The first in the Medical Informatics Category)
- **BMC Medical Genomics,** IF: 3.466
- **BioData Mining**
- **Healthcare Informatics Research**
- **Genomics & Informatics**

JAMIA

BMC Medical Genomics



Sponsors


iSCB
INTERNATIONAL
SOCIETY FOR
COMPUTATIONAL
BIOLOGY


Co-sponsoring with AMIA Joint Summits
TBI
CRI
AMIA JOINT SUMMITS ON TRANSLATIONAL SCIENCE
by
AMIA


SEOUL NATIONAL UNIVERSITY
KAIST
ETRI
한국전자통신연구원
AJOU UNIVERSITY
대한의료정보학회
THE KOREAN SOCIETY OF
MEDICAL INFORMATICS
한국생물정보시스템생물학회
Korean Society for Bioinformatics
and Systems Biology
시스템 바이오 정보의학 국가핵심연구센터


KMA
KOREAN MEDICAL
ASSOCIATION


life
technologies™
BMJ


BMS
Discover the Future
(주)비엠에스
bmskorea.co.kr