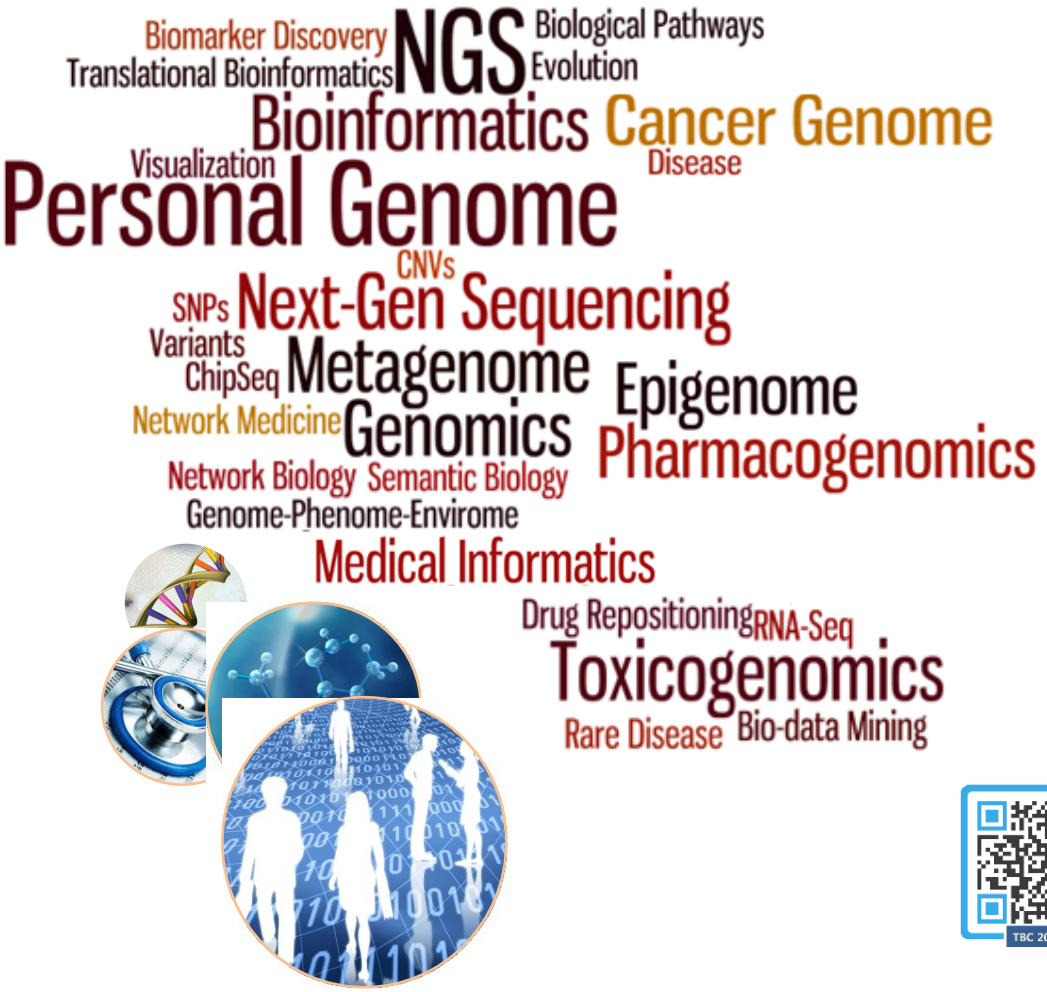
Translational Bioinformatics & Genomics



TIME and **LOCATION** | 13TH – 16TH, October, 2012, Jeju Island, Korea
Regency & Terrace Ballroom, Hyatt Regency Jeju
Conference | http://www.snubi.org/TBC2012/

Korean Society of Bioinformatics and Systems Biology,
Systems Biomedical Informatics Research Center
TBC 2012 / TGC 2012 Organizing Committee
BIOINFO 2012 Organizing Committee

Table of Contents

Greetings

It is my great pleasure to welcome all the delegates to the second annual Translational Bioinformatics Conference 2012 (TBC 2012). Translational bioinformatics is the key discipline to advance the basic research results into clinical applications with the ultimate goal of personalized medicine. It is truly multi-disciplinary area that requires fusion of state-of-the-art technologies in BT, IT, and HT (Health Technology). Recent progress in genome technologies such as GWAS and NGS declares the beginning of personal genome era, and re-assures the importance of integrating biological information with the medical information to deduce clinically applicable hypotheses and results. Translational bioinformatics is at the heart of such advances. Thus, it is quite timely to promote an international conference on the translational bioinformatics and genomics. I am sure that TBC 2012 / BIOINFO 2012 would take the place of central event for exchanging ideas, networking for collaborations, and learning new concepts and technologies in the field of translational bioinformatics. Finally, I would like to thank the organizing committee members, especially Prof. Ju Han Kim, for their efforts to organize the first conference so successfully, and all invited speakers and presenters for active participation in spite of their busy schedule.

Sanghyuk Lee

President, Korean Society for Bioinformatics and Systems Biology

Director, Korean Bioinformation Center (KOBIC)

Greetings

Dear Colleagues and Friends,

Along with the Organizing Committee, I am delighted to welcome you to attend the second annual Translational Bioinformatics Conference (TBC). Empowered by the enormous success of TBC 2011, Seoul, Korea, we all have closely collaborated for true promotion of translational bioinformatics. TBC 2012 is being held as a twin conference with TGC (Translational Genomics Conference) emphasizing the importance of tightly integrating the two disciplines. TBC 2012 / TGC 2012 provide a general forum for disseminating the latest research in genomics, bioinformatics, translational research, and medical informatics.

The revolutions both in bioinformatics and high-throughput genomics will eventually transform the current practice of medicine forever throughout diagnostics, therapeutics, and prognostics. Translational Bioinformatics has successfully translated the flood of high-throughput Translational Genomics data into meaningful biomedical products for cancers, rare diseases, personal genome interpretations, and pharmaco- and toxico-genomes. Translational bioinformatics is now even more powerful than ever.

Thanks to the invited speakers and presenters from all around the world to this conference, who are shaping the future of how future biomedical informatics translates into better practice, we are sure you will find an exciting atmosphere at TBC 2012 with wonderful weather. The future of translational bioinformatics, as William Gibson said, is already here, it's just not widely distributed yet.

I wish all participants of the conference have pleasant and memorable experience. Please enjoy TBC 2012 and the beautiful heart of Jeju Island.

With my best regards,

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

Chair, TBC 2012 Organizing Committee

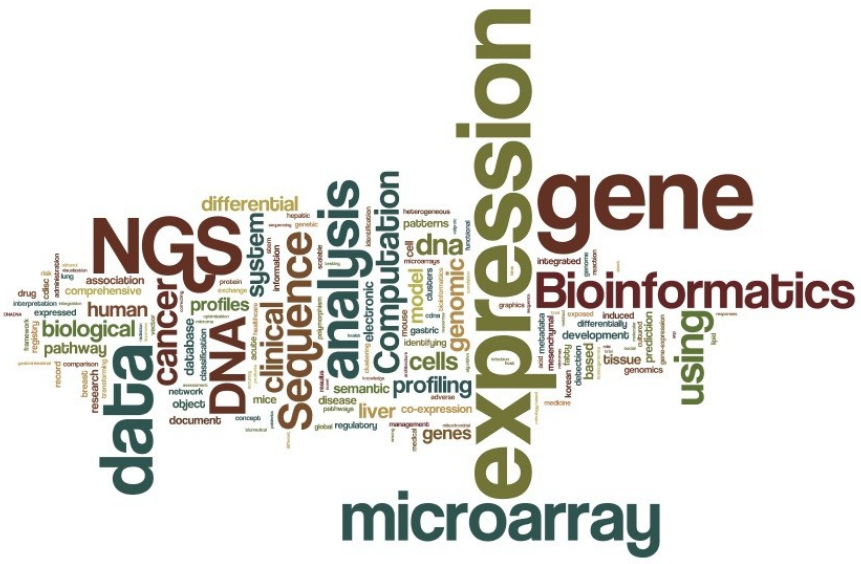
Greetings

Dear Colleagues and Friends,


I cordially welcome all the speaker and participants of TGC 2012. Now accessibility of genomic information is getting easier due to the development of the next generation sequencing (NGS) technologies. Large amount of genomic information will make a huge impact on the field of biotechnology. Moreover, it will be more useful in the diagnosis and treatment of human disease. In particular, the dissemination of the technology is expected to soon be a breakthrough for rare diseases and cancer. Rare Disease Research Center (RCRD) will provide service for rare disease research such as NGS, functional study and structural analysis. To this end, RCRD will held "Translational Genomics Conference" every year to deliver the current technology and information on genomic medicine. I am pleased to invite keynote speakers, Dr. Naomichi Matsumoto and Dr. Steven E. Brenner and eight speakers to the first meeting. I want to share the valuable information on Mendelian disease and cancers with all the participants of TGC2012 and TBC2012. I appreciate your interest and participation. Thank you.

Hae-Il Cheong, M.D., Ph.D. and Woong Yang Park, M.D., Ph.D.

Co-Chairs, TGC 2012 Organizing Committee



Translational Bioinformatics Conference

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 such that the first annual will be held in Seoul, Korea. 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.

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

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)

Program at a Glance

Saturday, Oct. 13, 2012

Welcoming reception for the OC members, TBC 2012
OC Board Meeting with Welcoming Dinner

TGC 2012: Translational Genomics Conference 2012

	Regency Ballroom
9:00~10:00AM	Registration
10:00~10:10AM	Opening of TGC 2012 Chair, Organizing committee **Hae-Il Cheong** (Seoul National University)
10:10~11:10AM	**Keynote I: Exome Sequencing in Mendelian Disorders** Chair, Woong-Yang Park **Naomichi Matsumoto** (Yokohama City University, Japan)
11:10~1:00PM	**Invited Session I : Translational Genomics for Rare Diseases** Chair, Hae-Il Cheong, Director, Research Coordination Center for Rare Diseases, SNUH **Murim Choi** (Yale University, USA) **Woong-Yang Park** (Seoul National University, Korea) **Chang-Suk Ki** (Samsung Medical Center, Korea) **Dong-Sup Kim** (KAIST, Korea)
1:00~ 2:00PM	**Lunch**
2:00~ 3:00PM	**Keynote II: Ultraconserved Nonsense: Gene Regulation by Alternative Splicing & RNA Surveillance** Chair, Jinsoo Kim, **Steven E. Brenner** (Berkley University, USA)
3:00~ 3:30PM	**Coffee Break**
3:30~ 5:30PM	**Invited Session II:** Chair, Yonsoo Lee **Jinsoo Kim** (Seoul National University, Korea) **Ji-Yeob Choi** (Seoul National University, Korea) **Wonshik Han** (Seoul National University, Korea) **Jongil Kim** (Seoul National University, Korea)
6:00~ 6:30PM	Closing Chair, Organizing committee **Cheong HI**

TBC 2012 Day 1
Translational Bioinformatics Conference

	Opening Paper Session, TBC2012 (Terrace Ballroom)
1:00 ~ 3:30PM	Poster Exhibition and Viewing (Day 1)
Session	**S1. Human Genome Sequence Analysis** Chair: Sael Lee (SUNY Korea)
3:30~4:45PM	**S1-1** MiST:Variant-detection through Whole-exome Sequencing *Sailakshmi Subramanian* **S1-2** Improve the Nucleotide Coding Technique, Use Support Vector Machine, Get the Better Accuracy: Survey of Human SpliceSite Prediction *A.T.M. Bari* **S1-3** New Features of MTRAP Alignment and Its Advantage: All-in-one Interface for Sequence Analysis,MSA and the Support for Non-coding RNA *Toshihide Hara*
Session	**S2. Cancer Genome Informatics** Chair: Hyun Goo Woo (Ajou Univ.)
4:45~6:00PM	**S2-1** Computational Methods for Cancer Subtype Classification using Integrated Data *Shinuk Kim* **S2-2** A Combination Algorithm for 5-Year Survivability of Breast Cancer Patient Classification *Kung-Jeng Wang* **S2-3** Gene Interaction-Level Cancer Classification using Gene Expression Profiles *Ashis Saha*

Sunday, Oct. 14, 2012

TBC 2012 Day 2
Translational Bioinformatics Conference
Three day conference with paper competitions
Regency and Terrace Ballrooms, Hyatt Regency in Jeju

	Regency Ballroom	Terrace Ballroom
9:00~ 6:00PM	Registration	
9:30~10:00AM	Opening TBC 2012	
10:00~10:50AM	**Keynote I: Turning Protein Networks into Ontologies** **Trey Ideker** (UC San Diego)	
Session	**S3. Bio/Medical Data Mining** Chair: Hyunjung Shin (Ajou Univ.)	**S4. Semantic Biology/Medicine** Chair: Raewoong Park (Ajou Univ.)
11:00~ 11:25AM	**S3-1** Globally Inferring Targets From Phenotypic Small-Molecule Screens *S. Joshua Swamidass*	**S4-1** Semantic PubMed Searches *Illhoi Yoo*
11:25~ 11:50AM	**S3-2** More Reproducible Results from Small-sample Clinical Genomics Studies by Multi-parameter Shrinkage, with Application to Highthroughput RNA Interference Screening Data *Mark van de Wiel*	**S4-2** Research Domain Grouping and Analysis in Bioinformatics Domain using Text Mining *Junbeom Kim*

11:50~ 12:15PM	**S3-3** Breast Cancer Survivability Prediction with Labeled, Unlabeled, and Pseudo-Labeled Patient Data *Juhyeon Kim*	**S4-3** ICD-9 Tobacco Use Codes are Effective Identifiers of Smoking Status *Laura K. Wiley*
12:15~ 1:10PM	**Lunch**, Poster Exhibition and Viewing (Day 2)	
1:10~ 2:00PM	**Keynote II: Understanding Complex Human Disease through Cell-lineage Specific Networks** **Olga Troyanskaya** (Princeton University)	
Session	**S5. Network Biology/Medicine I** Chair: In Suk Lee (Yonsei Univ.)	**S6. Functional and Structural Modeling** Chair: Dongsup Kim (KAIST)
2:00~ 2:25PM	**S5-1** Extracting of Coordinated Patterns of DNA Methylation and Gene Expression in Ovarian Cancer *Je-Gun Joung*	**S6-1** Diplotyper: Diplotype-based Association Analysis *Sunshin Kim*
2:25~ 2:50PM	**S5-2** Network Models of GWAS Uncover the Topological Centrality of Protein Interactions in Complex Disease Traits *Younghee Lee*	**S6-2** Computational Studies of Posttranslational Modifications *Zexian Liu*
2:50~ 3:15PM	**S5-3** Identification of Multiple Gene- Gene Interactions for Ordinal Phenotypes *Kyunga Kim*	**S6-3** Efficiency of Spatial Model in Assigning Protein Sequences to Protein Families *Hamid Pezeshk*
3:15~ 3:40PM	**S5-4** Key Genes for Modulating Information Flow Play a Temporal Role as Breast Tumor Coexpression Networks are Dynamically Rewired by Letrozole *Nadia Penrod*	**S6-4** Computational Approach for Protein Structure Prediction *Amouda Nizam*
3:40~ 4:10PM	**Coffee Break**, Poster Viewing	
4:10~ 5:00PM	**Keynote III: A Protein-Domain Approach for Analysis of Disease Mutations.** **Maricel Kann** (U. of Maryland Baltimore County)	
5:00~ 6:00PM	**Keynote IV: Computational Intelligence Strategies for Embracing the Complexity of Genetic Architecture** **Jason Moore** (Dartmouth University)	
	Dinner Beach Party, Poster Viewing	

Monday, Oct. 15, 2012

TBC 2012 Day 3

Translational Bioinformatics Conference

Three day conference with paper competitions

Regency and Terrace Ballrooms, Hyatt Regency in Jeju

	Regency Ballroom	Terrace Ballroom
9:00~10:00AM	**Keynote V: Big Data needs Good Tools: Translational Bioinformatics in Cell Innovation Project** **Takashi Gojobori** (National Institute of Genetics)	
Session	**S7. Disease Genome Informatics** Chair: Chaeyoung Lee (Soongsil Univ.)	**S8. Cancer Genome and Disease**

10:00~10:25AM	**S7-1** Revealing Molecular Mechanism of Rare Mental Disorders *Emil Alexov*	**S8-1** Personalized Chemotherapy for Ovarian Cancer by Integrating Genomic Data with Clinical Data *Younchul Kim*
10:25~10:50AM	**S7-2** Comparative Genomics Revealed General Evolutionary Trends of Insulin *Elbashir Abbas*	**S8-2** The Role of Genetic Heterogeneity and Epistasis in Bladder Cancer Susceptibility and Outcome: A Learning Classifier System Approach *Ryan Urbanowicz*
10:50~11:15AM	**S7-3** An Information-Gain Approach to Detecting Three-Way Epistatic Interactions in Genetic Association Studies *Ting Hu*	**S8-3** Multiclass Cancer Classification Using Gene Expression Comparisons *Sitan Yang*
11:15~11:40AM	**S7-4** Rare Variant Analysis_Using Publically Available Biological Knowledge *Carrie Moore*	**S8-4** Curation-Free Biomodules_Mechanisms in Prostate Cancer Predict Recurrent Disease *James L Chen*
11:40~ 1:00PM	**Lunch**, Poster Exhibition and Viewing (Day 3)	
1:00~ 2:00PM	**Keynote VI: New Approaches to Understanding the Genetic Component to Common Human Disease.** **Nancy Cox** (Chicago University)	
Session	**S9. Drug and Biomarker Discovery** Chair: Sun Choi (Ewha Univ.)	**S10. Network Biology/Medicine II** Chair: KiYoung Lee (Ajou Univ.)
2:00~ 2:25PM	**S9-1** Comparison and Validation of Genomic Predictors for Anticancer Drug Sensitivity *Simon Papillon-Cavanagh*	**S10-1** Detection of Pleiotropy through a Phenome-Wide Association Study (PheWAS) in the National Health and Nutrition Examination Surveys (NHANES) *Molly Hall*
2:25~ 2:50PM	**S9-2** Improve Binding Affinity by Twin Adhesive Drugs Mined in-between Docking Bio-mimicry Omega-shape Nona-peptide Agretope on HLA-1 Pit *Chun-Fan Chang*	**S10-2** Analysis of Type 2 Diabetes GWAS Dataset using Expanded Gene Set Enrichment Analysis and Protein-Protein Interaction Network *Chiyong Kang*
2:50~ 3:15PM	**S9-3** Altering Physiological Networks Using Drugs: Steps Towards Personalized Physiology *Adam Grossman*	**S10-3** Integrative Analysis of Congenital Muscular Torticollis: from Gene Expression to Clinical Indication *Shin-Young Yim*
3:15~ 3:40PM	**S9-4** Compensating for Literature Annotation Bias when Predicting Novel Drug-Disease Relationships through Medical Subject Heading Over-representation Profile (MeSHOP) Similarity *Warren Cheung*	**S10-4** Detecting Early-warning Signals of Type 1 Diabetes and Its Leading Biomolecular Networks by Dynamic al Network Biomarkers *Xiaoping Liu*
3:40~ 4:00PM	**Coffee Break**, Poster Viewing	
4:00~ 5:00PM	**Keynote VII: Dynamic Conservation of Gene Co-expression and Oncogene Deciphering** **Yi-Xue Li** (Shanghai Jiao Tong University)	
5:00~ 6:00PM	**Keynote VIII: Informatics to Enable Precision Medicine: Achievements, Obstacles and Opportunities** **Jessica Tenenbaum** (Duke University)	
6:00~ 6:30PM	**Special Remark on Translational Bioinformatics Paper Publications and TBC 2011 & 2012,**	

	by the Chief Editor of JAMIA **Lucila Ohno-Machado** (UC San Diego) **Closing Ceremony**
	Dinner on your own

Tuesday, Oct. 16, 2012

TBC 2012 Day 4

General Assembly, TBC 2012

Satellite meeting, BIOINFO 2012 sponsored by Korean Society of Bioinformatics and Systems Biology

	Regency Ballroom
8:30~09:00AM	Registration and BIOINFO 2012 announcement
9:00~09:10AM	Opening Remark Sanghyuk Lee (President, Korean Society for Bioinformatics and Systems Biology)
Session	**S1 Bio big data processing and integration** Chair: Sun Kim (Seoul Nat. Univ.)
9:10~09:30AM	**S1-1** Unified framework for multi-level biosystem modeling Doheon Lee (KAIST)
9:30~09:50AM	**S1-2** Rapid denoising of pyrosequenced amplicons for metagenomics Sungroh Yoon (Seoul Nat. Univ.)
9:50~10:10AM	**S1-3** Integrative approaches for DNA copy number aberrations in cancer Hyunju Lee (GIST)
10:10~10:30AM	**S1-4** Reference-assisted post-assembly of a de novo assembled genome Jaebum Kim (Konkuk Univ.)
10:30~10:50AM	**Coffee Break**, Poster Viewing
Session	**S2 Next-generation sequencing for next-generation biology** Chair: Jung Kyoon Choi (KAIST)
10:50~11:10AM	**S2-1** Regulation of nucleosome positioning and modification in transcription factor binding regions Jung Kyoon Choi (KAIST)
11:10~11:30AM	**S2-2** Genome-wide decoding of mRNA and miRNA maps Sung Wook Chi (Sungkyunkwan Univ.)
11:30~11:50AM	**S2-3** Unraveling of design principle in bacterial genomes Byung-Kwan Cho (KAIST)
11:50~12:10AM	**S2-4** Application of NGS in improvement of efficiency in radiotherapy Buhyun Youn (Pusan Nat. Univ.)
12:10~13:10PM	**Lunch**, Poster Viewing
Session	**S3 Computational biology-molecular modeling/simulations** Chair: Keun Woo Lee (Gyeongsang Nat. Univ.)
13:10~13:30PM	**S3-1** Protein function prediction by community detection of a PPI network Jooyoung Lee (KIAST)
13:30~13:50PM	**S3-2** Structural and thermodynamic investigation of protein aggregation in water Sihyun Ham (Sookmyung Women's Univ.)

13:50~14:10PM	**S3-3** Single-molecule study on DNA mismatch repair protein Jong-Bong Lee (POSTECH)
14:10~14:30PM	**S3-4** Barriers and wells to ion translocation in the Connexin 26 Hemi-channel Myunggi Yi (Pukyong Nat. Univ.)
14:30~14:50PM	**Coffee Break**
Session	**S4 Systems biology: evolution to translational medicine** Chair: Daehee Hwang (POSTECH)
14:50~15:10PM	**S4-1** Sociology in the genetic world Pan-Jun Kim (APCTP)
15:10~15:30PM	**S4-2** Genome-wide analysis and modeling of CpG methylation in 30 breast cancer cell lines Sun Kim (Seoul Nat. Univ.)
15:30~15:50PM	**S4-3** Opening the systemic analysis of ubiquitination-mediated protein regulation network Gwan-Su Yi (KAIST)
15:50~16:10PM	**S4-4** Alteration of epigenome landscaping is linked to neurodegeneration Hoon Ryu (Seoul Nat. Univ.)
	Closing Ceremony (TBC / BIOINFO)

	Terrace Ballroom
Session	**T1 Challenges in Bioinformatics (Student Session)** Chair: Daehee Hwang (POSTECH)
9:10~09:30AM	**T1-1** Large-scale reverse docking profiles and their applications Minho Lee (KAIST)
9:30~09:50AM	**T1-2** miRGator 3.0: a microRNA portal for deep sequencing, expression, and microRNA target investigation Yukyung Jun (Ewha Univ.)
9:50~10:10AM	**T1-3** Global loss of CG methylation potentiates defense response of Arabidopsis thaliana Daeseok Choi (POSTECH)
10:10~10:30AM	**T1-4** Performance comparison of two GSA methods for GWAS results Ji Sun Kwon (Soongsil Univ.)
10:30~10:50AM	**T1-5** Development of the exon graph for identificaiton of peptides and proteins Hyun Woo Kim (Hanyang Univ.)
Session	**T2 Symposium for the Association of Industry, Academy and Research Institutes** Chair: Jong Eun Lee (DNA Link)
10:50~11:15AM	**T2-1** Overview of Next Generation Sequencing Technology Sujin Kim (DNA Link)
11:15~11:40AM	**T2-2** Applications using Single Molecule Real-Time(SMRT) Sequencing Technology Siddharth Singh (PacBio)
11:40~12:05AM	**T2-3** Introduction to MiSeq Data Analysis and Cloud Computing Kyunga Kim (BMS)
12:05~13:10PM	**Lunch,** Poster Viewing

Keynote Speakers

1:00-2:00 pm (Monday, Oct. 15)

Nancy J. Cox

Chicago University

New Approaches to Understanding the Genetic Component to Common Human Disease

Although genome-wide association studies (GWAS) has enabled us to identify many new loci with highly significant and reproducible associations to common diseases and related quantitative traits, these discoveries have not yet given us much new understanding of the biology underlying disease, nor enabled us to develop accurate predictive risk models. In this talk I will describe a new approach to characterizing the genetic component to common diseases with complex inheritance that promises both a more comprehensive understanding of the biological basis of disease as well as practical utility for predicting risk. Examples of the application of this approach to data on such disparate complex traits as bipolar disorder, schizophrenia, type 2 diabetes and autism illustrate well its value, and demonstrate that it can be applied equally well to data generated through array genotyping or next generation sequencing.

4:00-5:00 pm (Monday, Oct. 15)

Yi-Xue Li

Shanghai Jiao Tong University

Dynamic Conservation of Gene Co-expression and Oncogene Deciphering

Gene expression profiling from patients provide much biological information for oncogene deciphering. Traditional methods, such as the Student's t-test and the clustering methods, identify differentially expressed genes through comparison of adjacent disease stages. These methods take no account of time conservative features and cooperative properties of gene signatures across whole disease stages, and may cause high false positive in finding disease related genes. Some new methods, like the multiclass ordinal analyses, were developed to identify genes involved in cancer development by extracting consistently increasing or decreasing expressed gene signatures in consideration of global changes of gene expression. Because gene expression profiling data are too complicated and heterogeneous, it is still a challenge to mine disease related genes. In fact, by using different methods on same gene expression data, we rarely get consistent results. In term of this, we developed an algorithm to deal with gene expression data in consideration of time serial conservation properties. In our method any specific expressed gene can be ranked with a time conservative score to evaluate its importance in cancer progression and development. Comparing with current methods, our algorithm can effectively and exactly identify functional gene sets by evaluating the global conservative properties of gene expression signatures. According to our approaches, a total of 480 genes in 29 clusters were obtained, only 8 percent of them can be identified by other studies. In a case study, 2 clusters were randomly selected and 9 genes were carefully annotated. All of those genes showed strong functional link with carcinoma occurrences and themselves form a small gene regulatory network mediated by P53, c-Myc, Sp1, IRF1, etc... Thus, to some extent, our evolutionary conservation analyzing based methodology compensates for the inherent weaknesses of current statistics methods and provides a new way for dynamic gene expression profile analyzing.

Keynote Speakers

9:00-10:00 am (Monday, Oct. 15)

Takashi Gojobori

National Institute of Genetics (NIG), Center for Information Biology and DNA Data Bank of Japan (DDBJ)

Big Data needs Good Tools: Translational Bioinformatics in Cell Innovation Project

As we know, the next-generation sequencing (NGS) technologies are changing a paradigm of genomic science so rapidly. First, a huge amount of nucleotide sequence data comes out from medical institutions such as medical schools of the universities and even city hospitals rather than laboratories in basic sciences. Second, how to obtain appropriate DNA or RNA samples timely in a given condition has become more crucial than how to be equipped with expensive sequencing machines, because sequencing itself is no longer a limiting factor of conducting genomic research from both viewpoints of time and cost. Third, when almost all the targets will become sequence-based, we have to deal with not only SNPs but also other types of variations such as CNV, Indels, and other DNA rearrangements. This may urge us to change the present due-course of GWAS, for example. Fourth and finally, development of powerful and accurate bioinformatics tool as well as construction of appropriate database have become essential in analyzing the so-called Big Data in order to produce the significant outcome. This paradigm change should be more emphasized particularly when we focus on translational medical research. In Japan, we conduct the research and development of NGS-sequence-based bioinformatics tools under the name of Cell Innovation Project in collaboration with RIKEN. I would present the current progress of this particular project with special reference to translational bioinformatics.

1:10-2:00 pm (Sunday, Oct. 14)

Maricel Kann

University of Maryland Baltimore County

A Protein-Domain Approach for Analysis of Disease Mutations

Identifying the functional context for key molecular disruptions in complex diseases is a major goal of modern medicine that will lead to earlier diagnosis and more effective personalized therapies. Most available resources for visualization and analysis of disease mutations center on gene analysis and do not leverage information about the functional context of the mutation. In addition, these gene-centric approaches are confounded by the fact that gene products (proteins) may share some functional sub-units or protein domains but not others. I will describe a resource for domain mapping of disease mutations, DMDM, a protein domain database developed by our group in which each disease mutation is aggregated and displayed by its protein domain location. We have also developed a methodology using domain significance scores (DS- Scores) to detect statistically significant disease mutation clusters at the protein domain level. When we applied the DS-Scores to human data, we identified domain hotspots in oncogenes, tumor suppressors, as well as in genes associated with Mendelian diseases. In addition, I will describe recent work on analyzing cancer somatic mutations from individual cancer patient genomes. We found that incorporating information about classification of proteins and protein sites leads to new hypotheses regarding the role of tumor somatic mutations in cancer. Our analysis confirms that the domain-centric approach creates a framework for leveraging structural genomics and evolution into the analysis of disease mutations.

Keynote Speakers

5:00-6:00pm (Sunday, Oct. 14)

Jason Moore

Dartmouth University

Computational Intelligence Strategies for Embracing the Complexity of Genetic Architecture

Given infinite time, humans would progress through modeling complex data in a manner that is dependent on prior knowledge of their domain, computer science and statistics as well as their prior experience working with other data. For example, a human modeler interested in identifying genetic risk factors for type II diabetes might start by examining insulin metabolism genes. We will review extensions and enhancements to an artificial intelligence-based computational evolution system (CES) that has the ultimate objective of tinkering with data as a human would. The key to the CES system is the ability to identify and exploit expert knowledge from biological databases or prior analytical results. Our prior studies have demonstrated that CES is capable of efficiently navigating large and rugged fitness landscapes toward the discovery of biologically meaningful genetic models of disease predisposition.

5:00-6:00 pm (Monday, Oct. 15)

Jessica Tenenbaum

Duke University

Informatics to Enable Precision Medicine: Achievements, Obstacles and Opportunities

The field of translational bioinformatics is at an exciting stage of progression. The past 5-10 years have seen the establishment of TBI as a widely recognized discipline unto itself, and the launch of a number of large-scale initiatives that TBI has enabled. A recent report from the National Academies describes how the recent explosion of molecular data coupled with clinical data on actual patients holds the potential to define an entirely new taxonomy of disease. In this new taxonomy, disease would be classified not solely by macroscopic symptoms many of which have been observed for centuries, but rather based on underlying molecular and environmental causes. This paradigm shift, enabled by novel methods for the generation, storage, analysis, and visualization of "big data" in biology and medicine, promises to do nothing short of rewrite the textbook of medicine moving forward. It will change the way we approach biomedical research and practice across the spectrum of scale, from molecules to populations. As technology continues to advance, assay costs to decrease, and as methods are further refined, the next decade is likely to feature increasingly pervasive examples of applied translational bioinformatics, both in healthcare and other areas of day to day life. In this talk I will highlight success stories and outstanding achievements in, or enabled by, translational bioinformatics. I will describe some important caveats and obstacles we face in this rapidly advancing field, as well as some ideas on how to address those hurdles. Finally, I will explore some of the tremendous opportunities we face in the years ahead.

Keynote Speakers

4:10-5:00 pm (Sunday, Oct. 14)

Olga Troyanskaya

Princeton University

Understanding Complex Human Disease through Cell-lineage Specific Networks

The ongoing explosion of new technologies in functional genomics offers the promise of understanding gene function, interactions, and regulation at the systems level. This should enable us to develop comprehensive descriptions of genetic systems of cellular controls, including those whose malfunctioning becomes the basis of genetic disorders, such as cancer, and others whose failure might produce developmental defects in model systems. However, the complexity and scale of human molecular biology make it difficult to integrate this body of data, understand it on a systems level, and apply it to the study of specific pathways or genetic disorders. These challenges are further exacerbated by the biological complexity of metazoans, including diverse biological processes, individual tissue types and cell lineages, and by the increasingly large scale of data in higher organisms. I will describe how we address these challenges through the development of bioinformatics frameworks for the study of gene function and regulation in complex biological systems and through close coupling of these methods with experiments, thereby contributing to understanding of human disease. I will specifically discuss how integrated analysis of functional genomics data can be leveraged to study cell-lineage specific gene expression, to identify proteins involved in disease in a way complementary to quantitative genetics approaches, and to direct both large-scale and traditional biological experiments.

10:10-11:10 am (Saturday, Oct. 13)

Naomichi Matsumoto

Yokohama City University Graduate School of Medicine

Exome Sequencing in Mendelian Disorders

Disease-related genome analysis (DGA) has been developed and sophisticated together with technology advances. The advent and frequent update of next generation sequencers (NGSs) can attain the appropriate accuracy for mutation analysis and push DGA into the new stages. We now use Illumina Genome Analyzer (GA) IIx and Hiseq2000 which can produce as much as 60-Gb and 600-Gb sequences in one run, respectively. To focus on genes, we utilized exon capture methods such as SureSelect (Agilent). The current NGS protocol uses 100-108-bp pair-end reads and usually produces 8-9 Gb sequences (per one sample) could be enough for analysis of the whole exome: 90 % of exome bait regions are covered by 8-10 reads or more. Sequences are aligned using MAQ, BWA, Novoalign and commercial-based NextGENe software all of which are able to extract nucleotide changes and small insertions/deletions. The most critical step is the priority scheme selecting variants. We have been successful in addressing culprit mutations in several Mendelian diseases. I will present our procedures used in our projects including Coffin-Siris syndrome and others.

Keynote Speakers

2:00-3:00 pm (Saturday, Oct. 13)

Steven E. Brenner

UC Berkeley

Ultraconserved Nonsense: Gene Regulation by Alternative Splicing & RNA Surveillance

Nonsense-mediated mRNA decay (NMD) is a cellular RNA surveillance system that recognizes transcripts with premature termination codons and degrades them. Using RNA-Seq, we discovered large numbers of natural alternative splice forms that appear to be targets for NMD. This coupling of alternative splicing and RNA surveillance can be used as a means of gene regulation. We found that all conserved members of the human SR family of splice regulators have an "unproductive" alternative mRNA isoform targeted for NMD degradation. Preliminary data suggest that this is used for creating a network of auto- and cross-regulation of splice factors. Strikingly, the splice pattern for each SR protein is shared with mouse, and each alternative splice is associated with an ultraconserved or highly-conserved region of ~100 or more nucleotides of perfect identity between human and mouse--amongst the most conserved regions in these genomes. Further, we recently discovering that most ancient known alternative splicing event is in this family and creates an alternate transcript to be degraded by NMD. Despite conservation since the pre-Cambrian, when the genes duplicate they change their regulation, so that nearly every human SR gene has its own distinctive sequences for unproductive splicing. As a result, this elaborate mode of gene regulation has ancient origins and can involve exceptionally conserved sequences, yet after gene duplication it evolves swiftly and often.

10:00-10:50 am (Sunday, Oct. 14)

Trey Ideker

UC San Diego

Turning Protein Networks into Ontologies

Ontologies have been very useful for capturing knowledge as a hierarchy of concepts and their interrelationships. In biology, a prime challenge has been to develop ontologies of gene function given only partial biological knowledge and inconsistency in how this knowledge is curated by experts. I will present a method by which large networks of gene and protein interaction, as are being mapped systematically for many species, can be transformed to assemble an ontology with equivalent coverage and power to the manually-curated Gene Ontology (GO). The network-extracted ontology contains 4,123 biological concepts and 5,766 relations, capturing the majority of known cellular components as well as many additional concepts, triggering subsequent updates to GO. Using genetic interaction profiling we provide further support for novel concepts related to protein trafficking, including a link between Nnf2 and YEL043W. This work enables a shift from using ontologies to evaluate data to using data to construct and evaluate ontologies.

Acknowledgements: This lecture was supported by the National Research Foundation of Korea (NRF) grants funded by the Korea Government, the Ministry of Education, Science & Technology (MEST) (2009-0086964).

Papers

S1-1 Paper

MiST: Variant-detection through Whole-exome Sequencing

Sailakshmi Subramanian[1], Valentina Di Pierro[1], Hardik Shah[1], Ajish George[1], Bruce Gelb[1], Ravi Sachidanandam[1]

[1]*United States Mount Sinai School of Medicine*

Whole-exome sequencing is a promising approach to find causative mutations in human disease, especially for Mendelian disorders. It involves the capture of sequences from exons in genomic DNA using probes from exonic regions of the genome. The captured exonic sequences are deeply sequenced and analyzed for variants from the reference genome. There are several tools to align sequenced reads to reference genomes and call SNPs and variants. We have developed a variant-calling platform, MiST that builds on our previously published tool, Geoseq. The tool mimics the experimental technique, computationally fishing reads from the deep sequencing set using probes from the targeted exons. The captured reads are mapped with great sensitivity to accurately call SNPs and variants. Our pipeline carefully eliminates paralogous read- mapping, which can lead to spurious SNP calls. It also tracks strand-bias and clonality in the sequencing libraries, allowing for more accurate measurements of coverage and variant detection. The platform identifies variant calls that have already been seen in other samples by comparing them to a database of known variants collected from dbSNP, 1000-genomes and private variant collections. A web-based interface allows users to visualize the alignments and other raw data underlying a variant call. The user can rapidly filter calls based on known and predicted functional characteristics. The pipeline is parallelizable and runs over a cluster, allowing the process to be scaled up. It also comes with a web-based interface that allows end-users to explore and visualize the data. We used targeted re-sequencing (Sanger) to confirm the validity of a few of the variants inferred by MiST. In addition, we compare it to variants calls made by the gatk platform and demonstrate the benefits of our approach, as well as the commonalities between the programs.

S1-2 Paper

Improve the Nucleotide Coding Technique, Use Support Vector Machine, Get the Better Accuracy: Survey of Human Splice Site Prediction

A.T.M.Golam Bari[1], Mst.Rokeya Reaz[1], Md.Azam Hossain[1], Ho-Jin Choi[2], Byeong-Soo Jeong[1]

[1]*Kyung Hee University, Dept. of Computer Engineering, 1732 Deokyoungdaero, Giheung-gu, Yongin-si, Gyeonggi-do, 446-701, Republic of Korea*

[2]*Korea Advanced Institute of Science and Technology,335 Guseong-dong, Yuseong-gu, Daejeon 305-701, Republic of Korea*

Splice site prediction in DNA sequence is a basic search problem for finding exon-intron and intron-exon boundary. Removing introns and then joining the exons together forms the coding sequences which are the input of translation process and a necessary step in central dogma of molecular biology. Finding out the exact GT and AG ending sequence among the set of ATCGs sequence and identifying the true and false GT and AG ending sequences are the main task of splice site prediction. In this paper, we survey recent research works on splice site prediction based on support vector machines (SVM)). The basic difference among these works is nucleotide encoding technique - some methods encode sparse way whereas others encode in a probabilistic manner. All these coding sequences serve as input of SVM. The task of SVM is to classify them using its learning model. We observe each coding techniques and classify them according to their similarity. Our survey paper will provide basic understanding of encoding approach for splice site prediction.

S1-3 Paper

New Features of MTRAP Alignment and its Advantage: All-in-one Interface for Sequence Analysis, MSA and the Support for Non-coding RNA

Toshihide Hara[1,2], Keiko Sato[1,2], Masanori Ohya[1,2]

[1]Department of Information Sciences, Tokyo University of Science, [2]Quantum Bio-Informatics Research Division, Tokyo University of Science, 2641 Yamazaki, Noda City, Chiba, Japan

Sequence alignment of proteins or DNA/RNA sequences is one of the most important things in modern bioinformatic analysis. In this field studies start with the comparison of target sequences, and the comparison is realized by constructing the alignment. Under a rapid increase of genome data from the growth of Next Generation Sequencing, the need for high quality alignment becomes more apparent. Although there exists an obvious need, the quality level is not enough. Recently we developed a high quality alignment method called MTRAP. We showed that the significant improvement of sequence alignment can be done by considering the correlation between two consecutive pairs of residues. In the first paper we showed that our method generates good results for protein sequences, but it is not understood whether it works for DNA/RNA sequences or not. In this paper, we show our recent study for non-coding RNA sequences. In addition, we explain the new features of recent version of MTRAP.

S2-1 Paper

Computational Methods for Cancer Subtype Classification using Integrated Data

Shinuk Kim[1,2,3], Taesung Park[2], Mark Kon[1,3]

[1]Bioinformatics program, Boston University, Boston, MA 02215, USA
[2]Department of Statistics, Seoul National University, Seoul 151-747 Republic of Korea
[3]Department of Mathematics and Statistics, Boston University, Boston, MA 02215 USA

MicroRNAs (miRNAs) are known to be strongly involved in cancer pathology through regulation of target messenger RNA (mRNA) molecules. We study a potentially useful methodology based on machine learning (ML) involving integration of separate biomarker classes to improve prediction and separation of ovarian cancer survival times. We use an ML-based protocol for feature selection, integrating information from miRNA and mRNA profiles at the feature level. For prediction of survival phenotypes, we use two classifiers, one a machine learning method (support vector machine, SVM), and the second a novel regression-based method (SVM-based Fisher feature selection together with Cox proportional hazard regression, FSCR). We compared these two methods using three types of cancer tissue features: i) miRNA expression, ii) mRNA expression, and iii) integrated miRNA and mRNA expression information, with features selected either from combined miRNA/mRNA profiles (CFS), or separately from the two feature sets (IFS). The accuracy of survival classification using the combined miRNA/mRNA profiles was 88.64 % using IFS-SVM, and 84.09% using IFS-FSCR in a balanced dataset. These accuracies are higher than those using miRNA alone (81.82%, SVM; 75%, FSCR) or mRNA alone (70.45%, SVM; 72.73%, FSCR). The latter differences indicate sometimes strong interactions between miRNA and mRNA features which are not visible in individual analyses. In addition we focus on the most significant miRNAs obtained by SVM-based feature selection which include hsa-miR-23b, hsa-miR-27b. We predicted 16 target genes of hsa-miR-23b and hsa-miR-27b, by integrating sequence information, and information of gene expression profile which include cancer related genes.

S2-2 Paper

A Combination Algorithm for 5-year Survivability of Breast Cancer Patient

Kung-Jeng Wang[1], Bunjira Makond[1], Kun-Huang Chen[1,2]

[1]*Department of Industrial Management, National Taiwan University of Science and Technology, Taipei 106, Taiwan, R.O.C.*

[2]*School of Dentistry, College of Oral Medicine, Taipei Medical University, Taipei, 110, Taiwan, R.O.C.*

In this study, we have proposed the new algorithm to enhance the effectiveness of classification for 5-year survivability of breast cancer patients which the data set is imbalanced. The algorithm is the combination of Synthetic minority oversampling technique (SMOTE) and Particle swarm optimization (PSO) based decision tree (C5): SMOTE+PSO+C5. G-mean is a metric to evaluate the proposed algorithm for classification; moreover, the proposed algorithm is compared with PSO+C5 and C5. The results show that SMOTE+PSO+C5 algorithm has the highest performance for 5-year survivability of breast cancer patient classification when the data set is imbalanced. This proposed method can classify well for both survival and non-survival cases. In addition, implementation PSO+C5 method to imbalanced data cannot improve the classification performance from using standard classifier solely.

S2-3 Paper

Gene Interaction-Level Cancer Classification using Gene Expression Profiles

Ashis Saha[1] and Jaewoo Kang[1]

[1]*Korea University, Seoul 136713, Korea.*

Recent studies suggest that biological pathways have the power to be stronger biomarkers for cancer than individual genes. The knowledgebase of pathways contains the interactions among the genes. However, it is not necessary for all the genes in a pathway to interact with each other. Closely interacting genes are supposed to have a collective effect to cause cancer or other disease. Here we propose a novel cancer classification method utilizing the collective effect of the set of closely interacting genes which we call Gene Interaction Set (GIS). We first find out the possible strength levels of each gene interaction set using clustering method and then rank all the sets with our proposed entropy metric using the proportion of samples of different classes having same strength level and finally predict the class of a new sample by weighted voting of top k gene interaction sets. The important feature of our method is that the process of causing the disease can easily be figured out. We validate our method comparing with other classification methods known to produce very high accuracy on 7 cancer datasets.

S3-1 Paper

Globally Inferring Targets From Phenotypic Small-Molecule Screens

S. Joshua Swamidass[1,2], Michael Barratt[1], Bradley T. Calhoun[1]

[1]*Division of Laboratory and Genomic Medicine, Department of Pathology and Immunology, Washington University School of Medicine, St. Louis, MO.*
[2]*Chemical Biology/Novel Therapeutics, Broad Institute of Harvard and MIT, Cambridge, MA*

A central challenge in modern drug discovery is the identification of the target proteins and pathways that can be manipulated to modulate disease. Gaps in our understanding of how targets modulate disease are evident in the high rate of Phase II clinical trial failures, when medicines are first tested for efficacy. The high reward for finding novel connections between targets and diseases is evident in several examples where known medicines have been repurposed to treat new diseases. In this study, we present and validate a new way of Globally Inferring protein Targets from Phenotypes (GIPT) by finding patterns in small-molecule screens of medically-relevant, cellular assays. Mining phenotypic, small-molecule screens is a promising strategy because it leverages translatable experimental data and because it is biased towards druggable proteins. We demonstrate that this strategy can both recover known targets and suggest plausible novel targets for several medically-relevant phenotypes - including insulin signaling, amyloid precursor protein expression, and cyclic-AMP levels - with applications in diabetes, Alzhiemer's diease, and depression.

S3-2 Paper

More Reproducible Results from Small-sample Clinical Genomics Studies by Multi-Parameter Shrinkage, with Application to High-throughput RNA Interference Screening Data

Mark A. van de Wiel[1], Renee X. de Menezes[2], Ellen Siebring[2,3], Victor W. van Beusechem[2]

[1]Department of Epidemiology and Biostatistics, [2]RNA Interference Functional Oncogenomics Laboratory (RIFOL), [3]Department of Pulmonary Disease, VU University Medical Center, PO Box 7057, 1007 MB Amsterdam, The Netherlands

High-throughput (HT) RNA interference screens are increasingly used for reverse genetics and drug discovery. These experiments are laborious and costly, hence sample sizes are often very small. Powerful statistical techniques to detect siRNAs that potentially enhance treatment are currently lacking, because they do not optimally use the amount of data in the other dimension, the feature dimension. We introduce ShrinkHT, a Bayesian method for shrinking multiple parameters in a statistical model, where 'shrinkage' refers to borrowing information across features. ShrinkHT is very flexible in fitting the effect size distribution for the main parameter of interest, thereby accommodating skewness that naturally occurs when siRNAs are compared with controls. In addition, it naturally down-weights the impact of nuisance parameters (e.g. assay-specific effects) when these tend to have little effects across siRNAs. We show that these properties lead to better ROC-curves than with the popular limma software. Moreover, in a 3 + 3 treatment vs control experiment with `assay' as an additional nuisance factor, ShrinkHT is able to detect three significant siRNAs with stronger enhancement effects than the positive control. In the context of gene-targeted (conjugate) treatment, these are interesting candidates for further research.

S3-3 Paper

Breast Cancer Survivability Prediction with Labeled, Unlabeled, and Pseudo-Labeled Patient Data

Juhyeon Kim[1] and Hyunjung Shin[1]

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

Prognostic study on breast cancer survivability has been aided by machine learning algorithms which provide prediction on the survival of a particular patient on the basis of historical patient data. A labeled patient record however, is not easy to collect. It takes at least five years to label a patient record as "survived" or "not survived": meanwhile, unguided trials on numerous types of oncology-therapy cost highly. Moreover, it requires confidentiality agreements from both doctors and patients to obtain a labeled patient record. The difficulties in collection of labeled patient data have drawn researchers' attention to Semi-Supervised Learning (SSL), one of the most recent machine learning algorithms, since it is capable of utilizing unlabeled patient data as well which relatively much easier to collect, and therefore is regarded as a pertinent algorithm to circumvent the difficulties. However, the fact is yet valid even on SSL that more labeled data lead

to better prediction. To make up for insufficiency of labeled patient data, one may consider an idea of tagging virtual labels to unlabeled patient data, namely “pseudo-labels”, and using them as if they are labeled. The proposed algorithm, "SSL Co-training", implements the idea based on SSL. SSL Co-training was tested on the surveillance, epidemiology, and end results database for breast cancer (SEER) and achieved avg. 76% accuracy and avg. 0.81 AUC.

S4-1 Paper

Semantic PubMed Searches

Illhoi Yoo[1, 2]

[1]Health Management & Informatics, School of Medicine, [2]Informatics Institute, University of Missouri, Columbia, MO, USA

The Evidence-Based Medicine (EBM) Working Group has defined efficient biomedical literature searching as a core skill required for the practice of the EBM. Although the information obtained from PubMed could significantly improve the quality of health care, physicians typically do not pursue their questions about patient care. This paper discusses the importance of PubMed searches for physicians, identifies the origin of the well-known obstacles to answering physicians’ clinical questions using PubMed, and introduces a novel system called Semantic-oriented MEDLINE search (SoMs) to the original problems to enhance their information retrieval experience in PubMed. Based on the variety of the literature in information retrieval, cognitive science, and medical science, we analyzed widely accepted obstacles to answering physicians’ clinical questions and then identified the origins of the obstacles to provide a technical solution for each obstacle category. Physicians’ information seeking behavior problem is two-fold: a user-side problem and a system-side problem. The user-side problem comes from the user’s emergent information needs and unfamiliarity with MeSH terms and the MeSH Tree, and the system-side problem comes from the fragmented information available from PubMed. We suggest the use of a biomedical semantic network with a concept-filtering tool to address the emergent information need problem, and the Concept-Based PubMed Archive (CBPA) to address the fragmented information problem. The SoMs can concisely answer many clinical questions PubMed cannot.

S4-2 Paper

Research Domain Grouping and Analysis in Bioinformatics Domain using Text Mining

Junbeom Kim[1], Chae-Gyun Lim[1], Sung Suk Kim[1], Dukyong Yoon[2], Rae-Woong Park[2], Ho-Jin Choi[1]

[1]Department of Computer Science, Korea Advanced Institute of Science and Technology, 291 Daehak-ro, Yuseong-gu, Daejeon 305-701, Korea
[2]School of Medicine & Graduate School of Medicine, Ajou University, San 5 Woncheon-dong, Yeongtong-gu, Suwon 443-721, Korea

In this paper, we propose a new information extraction and analysis using a text mining of the research domains for assistance of bioinformatics research. To do this work, we use Term Frequency Inverse Document Frequency method and reference link aggregation which combine each other and induce useful information to analysis the structures and relations of the interest fields. From the information induced from TFIDF and reference link aggregation, useful connections and relations, that generates and finds new information and knowledge, can be obtained. The results help researchers to extract and find more additional knowledge of related domains and fields. To show usefulness of the proposed method, we demonstrate research domain clustering and induced results from the clusters.

S4-3 Paper

ICD-9 Tobacco Use Codes are Effective Identifiers of Smoking Status

Laura K. Wiley[1,2], Anushi Shah[2], Hua Xu[2], William S. Bush[1,2]

[1]Center for Human Genetics Research, [2]Department of Biomedical Informatics, Vanderbilt University School of Medicine, Nashville, TN, USA

With the increased development of clinic-based biorepositories, Electronic Medical Records (EMRs) are being used for genetic epidemiology research. These studies often require identification of and adjustment for clinical covariates, such as smoking status. Unfortunately, a patient's smoking status is often difficult to extract from clinical text. The International Classification of Disease 9th Edition (ICD-9) contains two codes designating tobacco use - one for former and one for current use - but the reliability of these codes for classifying smoking status is often questioned due to their ambiguous use in clinical environments. In this study we evaluated the utility of these codes to identify ever-smokers in general and high smoking prevalence (lung cancer) clinic populations. We assessed potential biases in documentation, and performed temporal analysis relating transitions between smoking codes to smoking cessation attempts. We also examined the suitability of these codes for use in genetic association analyses. We establish that ICD-9 tobacco use codes can precisely identify smokers in a general clinic population (specificity = 1; sensitivity = 0.32), and that there is little evidence of documentation bias. Frequency of code transitions between "current" and "former" tobacco use is significantly correlated with initial success at smoking cessation (p<0.0001). Finally, we illustrate that code-based smoking status assignment is a comparable covariate to text-based smoking status for genetic association studies. Our results support the use of ICD-9 tobacco use codes for identifying smokers in a clinical population, and justify use of this derived status in genetic studies utilizing electronic health records.

S5-1 Paper

Extracting of Coordinated Patterns of DNA Methylation and Gene Expression in Ovarian Cancer

Je-Gun Joung[1,2,3], Dokyoon Kim[1,2], Kyung Hwa Kim[1,2], Ju Han Kim[1,2]

[1]*Seoul National University Biomedical Informatics (SNUBI), Div. of Biomedical Informatics,* [2]*Systems Biomedical Informatics National Core Research Center,* [3]*Institute of Endemic Diseases, Seoul National University College of Medicine, 103 Daehak-ro, Jongno-gu, Seoul 110-799, Korea*

DNA methylation, a regulator of gene expression, plays an important role in diverse biological processes including developmental process, carcinogenesis and aging. In particular, aberrant DNA methylation has been enormously observed in several types of cancers. Currently, it is important to extract disease-specific genesets associated with the regulation of DNA methylation. Here we propose a novel approach to find the minimum regulatory units of genes, co-Methylated and co-Expressed Gene Pairs (MEGPs) that are highly correlated gene pairs between DNA methylation and gene expression showing the co-regulatory relationship. To evaluate whether our method is meaningful to extract disease-associated genes, we applied our method to a large-scale dataset from The Cancer Genome Atlas, extracted significantly associated MEGPs and analyzed their functional correlation. We observed that our many MEGPs are physically interacted each other and show high semantic imilarity with Gene Ontology terms. Furthermore, we performed gene set enrichment tests to identify how they are correlated in a complex biological process. Our MEGPs were highly enriched in the biological pathway associated with ovarian cancers. Our approach can be useful for discovering coordinated epigenetic markers associated with specific diseases.

S5-2 Paper

Network Models of GWAS Uncover the Topological Centrality of Protein Interactions in Complex Disease Traits

Younghee Lee[1,2], Haiquan Li[1,2,3], Jianrong Li[1,2,3], Ellen Rebman[1,3], Kelly Regan[3], Eric R Gamazon[2], James L Chen[1,4], Xinan Yang[1,2], Nancy J Cox[1,2,5], Yves A Lussier[1,2,4,5,6]

[1]*Center for Biomedical Informatics and* [2]*Section of Genetic Medicine, Department of Medicine, The University of Chicago, Chicago, IL 60637*
[3]*Department of Medicine, The University of Illinois at Chicago, Chicago, IL, 60612,*
[4]*Section of Hematology/Oncology, Department of medicine, The University of Chicago, Chicago, IL60637*
[5]*Institute for Genomics and Systems Biology, and* [6]*Computation Institute, The University of Chicago, Chicago, IL 60637*

While Genome Wide Association Studies (GWAS) of complex traits have revealed thousands of reproducible genetic associations to date, these loci collectively confer very little of the heritability of their respective diseases and, in general, have contributed little to our understanding the underlying disease biology. Physical protein interactions have been utilized to increase our understanding of human Mendelian disease loci but have yet to be fully exploited for complex traits. Here, we hypothesized that protein interaction modeling of GWAS findings could highlight important disease-associated loci and unveil the role of their network topology in the genetic architecture of diseases with complex inheritance. Network modeling of proteins associated with the intragenic SNPs of the NHGRI catalog of complex trait GWAS revealed that complex trait associated loci are more likely to be hub and bottleneck genes in available, albeit incomplete, networks (odds ratio=1.59, FET-P value < 2.24×10^{-12}). Network modeling also prioritized novel Type 2 Diabetes(T2D) genetic variations from the Finland-United States Investigation of NIDDM Genetics and the Wellcome Trust GWAS data, and demonstrated the enrichment of hubs and bottlenecks in prioritized T2D GWAS genes. The potential biological relevance of the T2D hub and bottleneck genes was revealed by their increased number of first

degree protein interactions with known T2D genes according to several independent sources (P-value<0.01, probability of being first interactors of known T2D genes). Virtually all common diseases are complex human traits, and thus the topological centrality in protein networks of complex trait genes has implications in genetics, personal genomics, and in therapy.

S5-3 Paper

Identification of Multiple Gene-Gene Interactions for Ordinal Phenotypes

Kyunga Kim[1], Min-Seok Kwon[2], Sohee Oh[3], Taesung Park[2,3]

[1]*Department of Statistics, Sookmyung Women's University, South Korea*
[2]*Interdisciplinary Program in Bioinformatics, Seoul National University, South Korea*
[3]*Department of Statistics, Seoul National University, South Korea*

Multifactor dimensionality reduction (MDR) is a powerful method for analysis of gene-gene interactions and has been successfully applied to many genetic studies of complex diseases. However, the main application of MDR has been limited to binary traits, while traits having ordinal features are commonly observed in many genetic studies (e.g., obesity classification - normal, pre-obese, mild obese and severe obese). We propose ordinal MDR (OMDR) to facilitate gene-gene interaction analysis for ordinal traits. As an alternative to balanced accuracy, the use of tau-b, a common ordinal association measure, was suggested to evaluate interactions. Also, we generalized cross-validation consistency (GCVC) to identify multiple best interactions. GCVC can be practically useful for analyzing complex traits, especially in large-scale genetic studies. In simulations, OMDR showed fairly good performance in terms of power, predictability and selection stability and outperformed MDR. For demonstration, we used a real data of body mass index (BMI) and scanned 1~4-way interactions of obesity ordinal and binary traits of BMI via OMDR and MDR, respectively. In real data analysis, more interactions were identified for ordinal trait than binary traits. On average, the commonly identified interactions showed higher predictability for ordinal trait than binary traits. The proposed OMDR and GCVC were implemented in a C/C++ program, executables of which are freely available for Linux, Windows and MacOS upon request for non-commercial research institutions.

S5-4 Paper

Key Genes for Modulating Information Flow Play a Temporal Role as Breast Tumor Coexpression Networks are Dynamically Rewired by Letrozole

Nadia M. Penrod[1,2] and Jason H. Moore[2,3]

[1]*Department of Pharmacology and Toxicology,* [2]*Department of Genetics,* [3]*Institute for Quantitative Biomedical Sciences, Geisel School of Medicine at Dartmouth College, Hanover, NH, USA*

Genes do not act in isolation but instead as part of complex regulatory networks. To understand how breast tumors react to the presence of the drug letrozole it is necessary to understand how the entire gene network changes as it is perturbed by the drug. Using transcriptomic data generated from sequential tumor biopsy samples, taken at diagnosis and following 10-14 days and 90 days on letrozole, we build temporal gene coexpression networks. Coexpression is determined by a pairwise partial correlation statistic. We find that the breast tumor network is in a continual state of flux maintaining few relationships between time points. This means that the genes integral for maintaining network integrity and controlling information flow are dynamically changing as the network is rewired. By understanding how gene-gene relationships change in the presence of the drug letrozole we can begin to understand causes of drug resistance.

S6-1 Paper

Diplotyper: Diplotype-based Association Analysis

Sunshin Kim[1], KyungChae Park[2], Chol Shin[3], Nam H Cho[4], Jeong-Jae Ko[1], InSong Koh[5], KyuBum Kwack[1]

[1]*Department of Biomedical Science, College of Life Science, CHA University, Seongnam, Korea*
[2]*Department of Family Medicine, CHA Bundang Medical Center, CHA University, Seongnam, Korea*
[3]*Division of Pulmonary and Critical Care Medicine, Department of Internal Medicine, Korea University Ansan Hospital, Ansan, Korea*
[4]*Department of Preventive Medicine, Ajou University School of Medicine, Suwon, Korea, 5Department of Physiology, College of Medicine, Hanyang University, Seoul, Korea*

Diplotyper is a fully automated tool for performing association analysis based on diplotypes in a population. Diplotyper combines a novel algorithm designed to cluster haplotypes of interest from a given set of haplotypes with two existing tools: Haploview, for analyses of linkage disequilibrium blocks and haplotypes (with frequency threshold of 1%), and PLINK, to generate all possible diplotypes from a given population sample and calculate linear or logistic regression. In addition, procedures for generating all possible diplotype groups from the haplotype groups and transforming these diplotypes into PLINK formats were implemented. Diplotyper was tested through association analysis of hepatic lipase (LIPC) gene polymorphisms or diplotypes and levels of high-density lipoprotein (HDL) cholesterol. This analysis identified much more significant signals over single-locus tests.

S6-2 Paper

Computational Studies of Post-translational Modifications

Zexian Liu[1], Jian Ren[2], Yu Xue[3]

[1]*China University of Science and Technology of China*
[2]*China Sun Yat-sen University*
[3]*China Huazhong University of Science and Technology*

Background: Through temporally and spatially modified proteins, post-translational modifications (PTMs) greatly expand the proteome diversity and play critical roles in regulating the biological processes. Identification of site-specific substrates is fundamental for understanding the molecular mechanisms and biological functions of PTMs, while it is still a great challenge under current technique limitations. To date, the accumulation of experimental discoveries makes it available to develop computational tools for prediction of PTMs.

Methods: To predict PTM sites, a previously developed GPS (Group-based Prediction System) algorithm was adopted and improved. Weight training and k-mean clustering methods were introduced for prediction of pupylation sites in prokaryotic proteins and tyrosine nitration sites, respectively. Besides PTMs, GPS algorithm was extended to predict I-Ag7 and HLA-DQ8 epitopes through combination with Gibbs sampling approach. The CPLA database was constructed with manually collected experimental identified lysine acetylation sites from literature. The protein-protein interaction (PPI) information for construction of protein network was collected from five major PPI databases.

Results: The GPS algorithm was improved and employed to implement a series of softwares to predict PTMs including GPS-CCD, GPS-PUP and GPS-YNO2 for prediction of calpain cleavage, pupylation, tyrosine nitration site, respectively. Furthermore, the GPS algorithm was extended to develop predictor of GPS-MBA and GPS-ARM for prediction of MHC Class II Epitopes and APC/C recognition motif, respectively. With the predictive tools and the pipeline, we systematically compared the functional distribution and preference of S-nitrosylation and nitration. The functional diversity of the D-box and KEN-box mediated APC/C recognition and degradation was also statistically exploited. In addition, by integrating existed protein acetylome data, the human lysine acetylation network (HLAN) was firstly modeled and demonstrated, while the triplet relationship among HAT-substrate-HDAC was proposed as the fundamental component of HLAN.

Conclusions: Taken together, since the developed computational tools could provide helpful information with convenience, we anticipated that the combination of computational predictions and experimental verifications will become the foundation of systematically understanding the mechanisms and the dynamics of PTMs.

S6-3 Paper

The Efficiency of Spatial Model in Assigning Protein Sequences to Protein Families

Hamid Pezeshk[1,3], Vahid Rezaei[2,3]

[1]*School of Mathematics, Statistics and Computer Science, College of Science University of Tehran, Iran.*
[2]*Faculty of Mathematical Science, Tarbiat Modares University, Tehran, Iran.*
[3]*Bioinformatics Research Group, School of Computer Science, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran*

In this research we introduce a spatial model on a regular lattice based on multiple sequence alignment (MSA) for assignment of a protein sequence to a protein family. In this model, we assume that both the top and the bottom residues of each amino acid, in a profile of aligned protein sequences, contain useful information due to evolutionary relationship. For this purpose, we use top twenty profiles in the Pfam database to assess the performance of our spatial model in protein assignment to protein families. We then compare our model with profile hidden Markov model (PHMM). Results

show that using spatial model will increase the accuracy of protein sequence assignments considerably.

S6-4 Paper

Computational Approach for Protein Structure Prediction

Amouda Nizam[1], G.Jeyakodi1, C.Manimozhi [1]

[1]*Centre for Bioinformatics, Pondicherry University, India*

Genetic algorithm (GA) is used to solve difficult optimization problem of huge space where little is known in various domain and biological field is no exception. Many variants of Standard GA (SGA) are applied to a complex problem like Protein Structure Prediction (PSP) which is identified as NP-hard problem in molecular biology. Unfortunately SGA requires a special attention by the non-domain experts for the right choice of values for the parameter setting manually to reach a better solution. This research proposes a novel algorithm (SOGA) by blending a self-organizing concepts and GA in order to automate the appropriate choice of the parameter values. The proposed algorithm is developed with the entire knowledge of the problem (PSP) and the selection of different parameters is based on the problem and fitness value acquired in each generation. SOGAPSP is validated by comparing the native and predicted structure of protein. The minimal energy value of predicted protein structure indicates the stability of molecule. The Rampage server result implies the confirmation psi and phi angles of the predicted protein structure are feasible for amino acid residues in protein structure. The RSMD value indicates the similar conformation with the native structure of protein. The efficiency of the proposed algorithm reduces the time requirement for optimizing the parameter values to avoid premature convergence by self organizing the genetic operators of GA. The application of this algorithm to protein structure prediction achieved better results by self organizing the cross-over rates and mutation. Exceptionally there is no requirement of known structure to predict the unknown structure.

S7-1 Paper

Revealing Molecular Mechanism of Rare Mental Disorders

Zhe Zhang[1,2], Shawn Witham[1], Margo Petukh[1], Gautier Moroy[2], Maria Miteva[2], Yoshihiko Ikeguchi[3], Emil Alexov[1]

[1]*Computational Biophysics and Bioinformatics, Department of Physics, Clemson University, Clemson, SC 29634, USA*
[2]*Universite Paris Diderot, Sorbonne Paris Cite, Molecules Therapeutiques In Silico, Inserm UMR-S 973, 35 rue Helene Brion,75013 Paris, France*
[3] *Faculty of Pharmaceutical Sciences, Josai University, Japan*

Intellectual disability (ID) is a disease which is characterized by significant limitations in cognitive abilities and social/behavioral adaptive skills. It is one of the primary reasons for pediatric, neurologic, and genetic referrals. Particularly, with respect to the protein-encoding genes on the X chromosome, it was shown that approximately 10% of them have been implicated in ID, and the corresponding ID is termed X-linked ID (XLID). Although the numbers of mutations and reported families are small and XLID is a rare disease, collectively the impact of XLID is significant, because the patients almost always cannot fully participate in society. Here we report our findings of the effects of missense mutations of wild type properties of proteins and protein complexes involved in XLID. Using various in silico methods we reveal the molecular mechanism of XLID for cases involving proteins with available 3D structure. The 3D structures were used to predict the effect of disease-causing missense mutations on the folding free energy, conformational dynamics, hydrogen bond network and, if appropriate, on protein binding free energy. It is shown that vast majority of XLID mutation sites are outside the active pocket and are accessible from the water phase providing the opportunity that their effect can be altered by binding appropriate small molecules to the vicinity of the mutation site. This observation is used to demonstrate, computationally and experimentally, that a particular case, the Snyder-Robinson Syndrome causing G56S spermine synthase mutation, can be rescued by small molecule binding.

S7-2 Paper

Comparative Genomics Revealed General Evolutionary Trends of Insulin

ElbashirAbbas[1], Junbeom Kim[1], Yan Zhang[2], Luonen Chen[2], Ho-Jin Choi[1]

[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

Since its discovery, the hormone insulin has been associated with several diseases that plague man. The most famous of these is diabetes mellitus. As of last year 346 million people worldwide have been diagnosed with diabetes. No permanent treatment exists, and 80% of deaths are due to an inability in acquiring the chronic treatment. Previous studies have not thoroughly attempted to identify the origins of insulin, and with the recent discoveries and advances in available data it is possible to perform such a study and determine the evolution of this peptide. In addition, comparative studies have identified an overlooked an aspect in insulin that has not been thoroughly investigated. Namely, the new properties attributed to C-peptide, a subunit of the precursor of insulin. In this paper we present a comparative study between vertebrates and invertebrates with regards to the insulin precursor and insulin receptor. Our goal is to determine insulin origins and evolution across vertebrates and invertebrates by performing a comparative study of the insulin precursor and receptor in these species. Phylogenetic trees were constructed to visualize and determine the level of conservation of proinsulin and c-peptide and their respective distribution across different vertebrates. We have determined that both vertebrates and invertebrates contain insulin or insulin like proteins, however there number may differ, the coding patterns differ and the physical composition of C-peptide differs. Also the interacting insulin and insulin receptor residues found in both species classes show that some are conserved among both, but the majority are different.. Further work is required to expand on the results acquired and add to the insights gained.

S7-3 Paper

An Information-Gain Approach to Detecting Three-Way Epistatic Interactions in Genetic Association Studies

Ting Hu[1], Yuanzhu Chen[1,2], Jeff W. Kiralis[1], Ryan L. Collins[1], Christian Wejse[3], Giorgio Sirugo[4], Scott M. Williams[1,5], Jason H. Moore[1, 5]

[1]Department of Genetics, Geisel School of Medicine, Dartmouth College, Hanover, NH, USA
[2]Department of Computer Science, Memorial University, St. John's, NL, Canada
[3]Center for Global Health, School of Public Health, Aarhus University, Skejby, Denmark
[4]Centro di Genetica, Centro di Ricerca Scientifica, Ospedale San Pietro FBF, Rome, Italy
[5]Institute for Quantitative Biomedical Sciences, Dartmouth College, Hanover, NH, USA

Epistasis has been historically used to describe the phenomenon that the effect of a given gene on a phenotype can be dependent on one or more other genes, and is an essential element for understanding the association between genetic and phenotypic variations. Quantifying epistasis of orders higher than two is very challenging due to both the computational complexity of enumerating all possible combinations in genome-wide data and the lack of efficient and effective methodologies. In this study, we propose a fast, non-parametric, and model-free measure for three-way epistasis using information gain. It is able to separate all lower-order effects from pure three-way epistasis. Our method was verified on synthetic data and applied to real data from a candidate-gene study of tuberculosis (TB) in a West African population. In the TB data, we found a statistically significant pure three-way epistatic interaction effect that was stronger than any lower-order associations. Our study provides a methodological basis for detecting and characterizing high-order gene-gene interactions in genetic association studies.

S7-4 Paper

Rare Variant Analysis Using Publically Available Biological Knowledge

Carrie B. Moore[1,2], John R. Wallace[2], Alex T. Frase[2], Sarah A. Pendergrass[2], Marylyn D. Ritchie[2]

[1]Center for Human Genetics Research, Vanderbilt University, Nashville, TN 37232, USA,
[2]Center for Systems Genomics, Pennsylvania State University, University Park, PA 16802, USA

With the recent flood of genome sequence data, there has been increasing interest in rare variants and methods to detect their association to disease. We developed a flexible collapsing method inspired by biological knowledge called BioBin. We also built the Library of Knowledge Integration (LOKI), a repository of data assembled from public databases, which contains resources such as: the National Center for Biotechnology (NCBI) dbSNP and gene Entrez database information, Kyoto Encyclopedia of Genes and Genomes (KEGG), Reactome, Gene Ontology (GO), Protein families database (Pfam),

NetPath -signal transduction pathways, Molecular INTeraction database (MINT), Biological General Repository for Interaction Datasets (BioGrid), Pharmacogenomics Knowledge Base (PharmGKB), Open Regulatory Annotation Database (ORegAnno), and information from UCSC Genome Browser about evolutionary conserved regions (ECRs). BioBin can apply multiple levels of burden testing, including: functional regions, evolutionary conserved regions, genes, and/or pathways. We tested BioBin using simulated data as well as with low coverage data from the 1000 Genomes Project to evaluate bins with simulated causative variants and conducted a pairwise comparison of rare variant (MAF < 0.03) burden differences between Yoruba individuals (YRI) and individuals of European descent (CEU). Lastly, we analyzed NHLBI GO Exome Sequencing Project Kabuki dataset, with sequenced data from individuals with Kabuki syndrome, a congenital disorder affecting multiple organs and often intellectual disability, contrasted with 1000 genomes data as controls. BioBin is proving to be a very useful and flexible tool to analyze sequence data and uncover novel associations with complex disease.

S8-1 Paper

Personalized Chemotherapy for Ovarian Cancer by Integrating Genomic Data with Clinical Data

Youngchul Kim[1], Kian Behbakht[2], Jennifer R. Diamond[2], Dan Theodorescu[2], Jae K. Lee[1]

[1]Department of Public Health Sciences, University of Virginia, PO Box800717, Charlottesville, VA 22908, USA
[2]University of Colorado Cancer Center, University of Colorado Denver, Box 8117, Aurora, CO 80045, USA

Despite multiple standard chemotherapy drugs and novel agents, the overall therapeutic response of advanced Epithelial Ovarian Cancer (EOC) patients has been stagnant over the last two decades. Aggressive tumors such as EOC are highly heterogeneous in their therapeutic responses, so overall therapeutic responses are not likely to be improved much if used without selection. Previous biomarker studies of drug response were limited as it was difficult to develop single drug predictors based on patients treated with multiple drugs. Additionally, outcomes were often confounded with other factors beyond given therapies. By directly combining patients' therapeutic outcome information with the COXEN algorithm based on each drug's cell line activity data, we have developed integrated predictors of three standard chemotherapy drugs in treating EOC: paclitaxel, cyclophosphamide, and topotecan. Our integrated COXEN predictors of the three drugs demonstrated high predictability simultaneously on patients' short-term therapeutic responses and long-term survival outcomes. In particular, when the three drug predictors were hypothetically used for a historical patient cohort, overall survival and progression-free survival of the cohort would have been prolonged more than one year and five months, respectively. When examined for patients with recurrent disease, overall survival was improved more than 21 months. While the current study still remains within analytic potential due to relatively small sample sizes for rigorous evaluation of some of these predictors, the study has shown a possibility that overall therapeutic response and outcome can be dramatically improved by optimally utilizing these integrated predictors for individual patients with EOC.

S8-2 Paper

The Role of Genetic Heterogeneity and Epistasis in Bladder Cancer Susceptibility and Outcome: A Learning Classifier System Approach

Ryan J. Urbanowicz[1], Angeline S. Andrew[1], Margaret R. Karagas[1], Jason H. Moore[1]

[1]*Geisel School of Medicine, Dartmouth College, 1 Medical Center Dr., Lebanon, NH 03756*

Detecting complex patterns of association between genetic or environmental risk factors and disease risk has become an important target for epidemiological research. In particular, strategies that accommodate multifactor interactions or heterogeneous patterns of association can offer new insights in association studies wherein traditional analytic tools have had limited success. In an effort to concurrently address these phenomena, previous work has successfully considered the application of learning classifier systems (LCSs), a flexible class of evolutionary algorithms that distributes learned associations over a population of rules. Subsequent work addressed the inherent problems of knowledge discovery and interpretation within these algorithms, allowing for the characterization of heterogeneous patterns of association. While these previous advancements were evaluated using complex simulation studies, this study applied these collective works to a real world genetic epidemiology study of bladder cancer susceptibility. Notably, we replicated the identification of previously characterized factors that modify bladder cancer risk: i.e. single nucleotide polymorphisms (SNPs) from a DNA repair gene, and smoking. Furthermore, we identified potentially heterogeneous groups of subjects characterized by distinct patterns of association. Cox proportional hazard models comparing clinical outcome variables between the cases of the two largest groups yielded a significant, meaningful difference in survivorship. A marginally significant difference in time to recurrence was also noted. These results support the hypothesis that an LCS approach can offer greater insight into complex patterns of association. This methodology appears to be well suited to the dissection of disease heterogeneity, a key component in the advancement of personalized medicine.

S8-3 Paper

Multiclass Cancer Classification using Gene Expression Comparisons

Sitan Yang [1] and Daniel Q. Naiman[2]

[1,2]*Applied Mathematics and Statistics Department, Johns Hopkins University, Baltimore, Maryland 21218, USA*

As our knowledge of cancer has grown, its heterogeneous nature has become increasingly apparent, and there has been an accompanying tendency to identify and differentiate various cancer subtypes. In this situation, microarray-based cancer classification poses new methodological and computational challenges, and the identification of novel and effective approaches to multiclass classification deserves greater attention. While cancer classification has achieved

considerable success in binary problems, the situation for multiclass problems is not as clear. In this paper, we introduce a new approach to multiclass cancer diagnosis based on gene expression profiles. Our method focuses on detecting a small set of genes whose expression levels have significant changes relative to each other from class to class. For a k-class problem, the decision rule only depends on the relative orderings of expression values of k genes and is transparent enough to be immediately explored for biological discoveries. We demonstrate on five cancer datasets that our method, while simple, is as powerful as many popular but complex classifiers. Furthermore, we show that the decision rules built on these datasets involve some informative genes that are known to have biological relevance for some cancer types, which may help us understand their potential mechanisms.

S8-4 Paper

Curation-Free Biomodules Mechanisms in Prostate Cancer Predict Recurrent Disease

James L. Chen[1], Alexander Hsu[1,2], Xinan Yang[1], Jianrong Li[2], Gurunadh Parinandi[2], Haiquan Li[2], Yves A. Lussier[1,2,3]

[1]Ctr for Biomed. Informatics and Dept. of Medicine, The University of Chicago, Chicago, IL
[2]Depts of Medicine & of Bioengineering, University of Illinois at Chicago, Chicago, IL
[3]University of Illinois Hospital and Health Science System

Motivation: Gene expression-based prostate cancer gene signatures of poor prognosis are hampered by lack of gene feature reproducibility and a lack of understandability of their function. Molecular pathway-level mechanisms are intrinsically more stable and more robust than an individual gene. The Functional Analysis of Individual Microarray Expression (FAIME) we developed allows distinctive sample-level pathway measurements with utility for correlation with continuous phenotypes (e.g. survival). Further, we and others have previously demonstrated that pathway-level classifiers can be as accurate as gene-level classifiers using curated genesets that may implicitly comprise ascertainment biases (e.g. KEGG, GO). Here, we hypothesized that transformation of individual prostate cancer patient gene expression to pathway-level mechanisms derived from automated high throughput analyses of genomic datasets may also permit personalized pathway analysis and improve prognosis of recurrent disease.

Results: Via FAIME, three independent prostate cancer gene expression arrays with both normal and tumor samples were transformed into two distinct types of molecular pathways mechanism and then compared: (i) the curated Gene Ontology (GO) and (ii) dynamic expression activity networks of cancer (Cancer Modules). FAIME-derived mechanisms for tumorigenesis were then identified. Curated GO and computationally generated “Cancer Module” mechanisms overlap significantly and are enriched for known oncogenic deregulations and highlight potential areas of investigation. We further show in two independent datasets that these pathway-level tumorigenesis mechanisms can identify men who are more likely to develop recurrent prostate cancer (log-rank_p=0.019 and 0.04, respectively).

S9-1 Paper

Comparison and Validation of Genomic Predictors for Anticancer Drug Sensitivity

Simon Papillon-Cavanagh[1], Nicolas De Jay[1], Nehme Hachem[1], Catharina Olsen[2], Gianluca Bontempi[2], Hugo Aerts[3], John Quackenbush[4], Benjamin Haibe-Kains[1]

[1]Bioinformatics and Computational Genomics Laboratory, Institut de recherches cliniques de Montreal, University of Montreal, Montreal, Quebec, Canada
[2]Machine Learning Group, Universite Libre de Bruxelles, Bruxelles, Belgium
[3]Department of Radiation Oncology and 4 Department or Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Harvard University, Boston, MA, USA,

An enduring challenge in personalized medicine lies in selecting the right drug for each individual patient. While direct testing of drugs on patients is the only way to assess their clinical efficacy and toxicity, we dramatically lack resources to test the hundreds of drugs that are currently under development. Therefore the use of preclinical model systems has been intensively investigated as this approach enables to test response to hundreds of drugs in multiple cell lines in parallel. Recently two large-scale pharmacogenomic studies screened multiple anticancer drugs on more than 1000 cell lines. Here we propose to combine these datasets to build and robustly validate genomic predictors of drug response. We compared five different approaches for building predictors of increasing complexity. We assessed their performance in cross-validation and in two large validation sets, one containing the same cell lines present in the training set and another dataset composed of cell lines that have never been used during the training phase. Sixteen drugs were found in common between the datasets. We were able to validate multivariate predictors for four out of the sixteen tested drugs, namely Irinotecan, PD-0325901, PLX4720 and Lapatinib. Moreover, we observed than response to 17-AAG, an inhibitor of Hsp90, could be efficiently predicted by the expression level of a single gene, NQO1. Altogether these results suggest that predictors could be robustly validated for specific drugs. If successfully validated in patients' tumor cells, and subsequently in clinical trials, they could act as companion tests for the corresponding drugs and play an important role in personalized medicine.

S9-2 Paper

Improve Binding Affinity by Twin Adhesive Drugs Mined in-between Docking Bio-mimicry Omega-shape Nona-peptide Agretope on HLA-1 Pit

Chun-Fan Chang[1], Chen-Chieh Fan[2,3], Hsueh-Ting Chu[4], Cheng-Yan Kao[2]

[1]Department of Animal Science, Chinese Culture University, Taipei 11114, Taiwan;
[2]Department of Computer Science and Information Engineering, National Taiwan University, Taipei 10617, Taiwan; and
[3]ENT Division, National Taiwan University Hospital, Taipei 10002, Taiwan.
[4]Department of Computer Science and Information Engineering, Asia University, Taichung 41354, Taiwan.

Motivation: The oncogenesis process of nasopharyngeal carcinoma (NPC) may equip proliferation advantage and immune evasion in overcoming efficient host immune clearance mechanisms against Epstein Barr virus (EBV). The proliferation advantage is likely from encoding EBV latent infection phase membrane protein 1 (LMP1) and the immune

evasion is likely from mutating EBV genome for poor immune reactivity at AMI-antigen epitopes and CMI-antigen epitopes/agretopes of LMP1/LMP2 and EBNA upon class I human leukocyte antigen (HLA-1) IIn this work, we developed a structure-based immunoinformatic tool of EBV-LMP1 related omega-shape nona-peptide (LMP1np) design for docking HLA-1 pit towards mining twin adhesive drugs (TAD) with improved binding affinity (BAff).

Results: Our implemented bio-mimicry peptide design algorithm tool (bmPDA tool) designs nona-peptide structures with bulge-side epitope and anchor-side agretope from LMP-1 and NLMP-1 segments for docking HLA-1 of A*0201 and A*0207. The design efficiency of bio-mimicry peptide by bmPDA tool is demonstrated with preliminary reference nona-peptide structure of vasopressin protein. The binding affinity (BAff) between putative agretope and verified HLA1 pit shows notable weakening for likely immune evasion in the cases of A*0207 and NLMP1 at initial amino acid positions of 32, 35, 86, 92, 125, 147, and 166. In that, our algorithm mines twin adhesive drugs (TAD) among FDA-approval list exemplified with Nizatidine, Benzonatate, Entecavir, Famotidine, and Alprostadil for improving BAff between A*0207 pit and weak agretope of NLMP1np structures.

S9-3 Paper

Altering Physiological Networks using Drugs: Steps towards Personalized Physiology

Adam D Grossman, PhD[1], Mitchell J Cohen, MD[2], Geoffrey T Manley, MD, PhD[3], Atul J Butte, MD, PhD[4]

[1]Department of Bioengineering, Stanford University, Stanford, CA, USA
[2]Department of Surgery, University of California San Francisco, San Francisco, CA, USA
[3]Department of Neurosurgery, University of California San Francisco, San Francisco, CA, USA
[4]Department of Pediatrics and the Department of Medicine, Stanford University School of Medicine, Stanford, CA, and Lucile Packard Children's Hospital, Palo Alto, CA, USA.

The rise of personalized medicine has reminded us that each patient must be treated as an individual. One factor in making treatment decisions is the physiological state of each patient, but definitions of relevant states and methods to visualize state-related physiologic changes are scarce. We constructed correlation networks from physiologic data to demonstrate changes associated with pressor use in the intensive care unit. We collected 29 physiological variables at one-minute intervals from nineteen trauma patients in the intensive care unit of an academic hospital and grouped each minute of data as receiving or not receiving pressors. For each group we constructed Spearman correlation networks of pairs of physiologic variables. To visualize drug-associated changes we split the networks into three components: an unchanging network, a network of connections with changing correlation sign, and a network of connections only present in one group. Out of a possible 406 connections between the 29 physiological measures, 64, 39, and 48 were present in each of the three component networks. The static network confirms expected physiological relationships while the network of associations with changed correlation sign suggests putative changes due to the drugs. The network of associations present only with pressors suggests new relationships that could be worthy of study. We demonstrated that visualizing physiological relationships using correlation networks provides insight into underlying physiologic states while also showing that many of these relationships change when the state is defined by the presence of drugs. This method applied to targeted experiments could change the way critical care patients are monitored and treated.

S9-4 Paper

Compensating for Literature Annotation Bias when Predicting Novel Drug-Disease Relationships through Medical Subject Heading Over-representation Profile (MeSHOP) Similarity

Warren A. Cheung[1,2], BF Francis Ouellette[3,4], Wyeth W. Wasserman[1,5]

[1]Centre for Molecular Medicine and Therapeutics at the Child and Family Research Institute, University of British Columbia, Vancouver, BC, Canada, [2]Bioinformatics Graduate Program, University of British Columbia, Vancouver, BC, Canada, [3]Ontario Institute for Cancer Research, Toronto, ON, Canada, [4]Department of Cells and Systems Biology, University of Toronto, Toronto, ON, Canada, [5]Department of Medical Genetics, University of British Columbia, Vancouver, BC, Canada

Medical Subject Heading Overrepresentation Profiles (MeSHOPs) quantitatively summarise the literature associated with biological entities such as diseases or drugs. A profile is constructed by counting the number of times each MeSH term is assigned to an entity-related research publication in the MEDLINE/PUBMED database and calculating the significance of the count relative to a background expectation. Based on the expectation that drugs suitable for treatment of a disease (or disease symptom) will have similar annotation properties to the disease, we successfully predict drug-disease associations by comparing MeSHOPs of diseases and drugs. The MeSHOP comparison approach delivers an 11% improvement over bibliometric baselines. However, novel drug-disease associations are observed to be biased towards drugs and diseases with more publications. To account for the annotation biases, a correction procedure is introduced and evaluated. By explicitly accounting for the annotation bias, unexpectedly similar drug-disease pairs are highlighted as candidates for drug repositioning research.

S10-1 Paper

Detection of Pleiotropy through a Phenome-Wide Association Study (PheWAS) in the National Health and Nutrition Examination Surveys (NHANES)

M.A. Hall[1], A. Verma[1], K.D. Brown-Gentry[2], R. Goodloe[2], J. Boston[2], S. Wilson[2], B. McClellan[2], C. Sutcliffe[2], H.H. Dilks[2,3], N.B. Gillani[2], H. Jin[2], P. Mayo[2], M. Allen[2], N. SchnetzBoutaud[2], D.C. Crawford[2,3], M.D. Ritchie[1], S.A. Pendergrass[1]

[1]Center for Systems Genomics, Department of Biochemistry and Molecular Biology, The Huck Institutes of the Life Sciences, The Pennsylvania State University, University Park, PA, USA;
[2]Center for Human Genetics Research, [3]Department of Molecular Physiology and Biophysics, Vanderbilt University, Nashville TN, USA

Herein we describe the results of a Phenome-wide association study (PheWAS) utilizing the diverse genotypic and phenotypic data that exists for multiple race-ethnicites in the National Health and Nutrition Examination Surveys (NHANES), conducted by the Centers for Disease Control and Prevention (CDC) and accessed by the Epidemiological Architecture for Genes Linked to Environment (EAGLE) study. PheWAS is a novel approach for discovering the complex mechanisms involved in human disease by testing SNPs for association with a large and diverse set of phenotypes. Comprehensive unadjusted tests of association were performed in NHANES III and NHANES 1999-2002 for 575 SNPs with 1009 phenotypes stratified by race-ethnicity. We identified 51 PheWAS associations that were consistent between the two surveys for the same SNP, phenotype-class, direction of effect, and race-ethnicity with $p<0.01$, allele frequency > 0.01, and sample size > 200. Of these, 28 replicated previously reported SNP-phenotype associations, 9 were related to previously reported associations in the literature, and 14 were novel SNP-phenotype associations. We also identified SNPs associated with multiple novel phenotypes. These results demonstrate the utility of phenome-wide association studies for exploring associations between genetic variation and phenotypic variation in a high throughput and comprehensive manner using existing epidemiologic study data. The results of PheWAS promise to expose more of the genetic architecture underlying multiple traits and generate hypotheses about pleiotropic interactions for future research.

S10-2 Paper

Analysis of Type 2 Diabetes GWAS Dataset using Expanded Gene Set Enrichment Analysis and Protein-Protein Interaction Network

Chiyong Kang[1], Hyeji Yu[1], Gwan-Su Yi[1]

[1]*Department of Bio and Brain Engineering, KAIST, Daejeon 305701, Korea*

Genome-wide association studies (GWAS) have been identified approximately 40 type 2 diabetes (T2D) associated SNPs. However, only small fraction of the T2D genetic risk is explained with identified T2D associated SNPs. While pathway enrichment analysis that considers multiple SNPs is suggested to reveal the mechanisms of complex diseases, pathway gene set can cover only small portion of human genes. For the better understanding of biological mechanisms of T2D and T2D causal gene detection, enrichment analysis with expanded gene sets and mapping GWAS based T2D associated gene into protein-protein interaction (PPI) network are proposed. Gene set enrichment analysis (GESA) is applied on WTCCC T2D GWAS dataset with expanded gene sets including pathway, function, TF-target, miRNA-target and complex. From expanded GSEA, 451 T2D associated gene sets are detected with p-value < 0.05 and 441 gene sets out of selected 451 gene sets contain known T2D genes. To find novel T2D gene candidates, 64 GWAS based T2D associated genes which are from 2,960 SNPs with p-value threshold 0.05 in WTCCC T2D GWAS dataset are mapped into integrated PPI network and total 24 novel T2D gene candidates are detected. Among detected T2D gene candidates, GBR2 is the most associated gene with T2D. Expanded GSEA and PPI mapping of GWAS based T2D associated genes showed the possibility of providing insights of T2D mechanisms and detecting novel T2D gene candidates.

S10-3 Paper

Integrative Analysis of Congenital Muscular Torticollis: from Gene Expression to Clinical Indication

Shin-Young Yim, MD, PhD[1], Dukyong Yoon, MD, MS[2], Myong Chul Park, MD, PhD[3], Il Jae Lee, MD, PhD[3], Jang-Hee Kim, MD, MS[4], Myung Ae Lee,PhD[5], Kyu-Sung Kwack, MD, PhD[6], Jan-Dee Lee, MD, PhD[7], Euy-Young Soh, MD, PhD[8], Young-In Na, MS[9], Rae Woong Park, MD, PhD[2], KiYoung Lee, PhD[2], Jae-Bum Jun, MD, PhD[9]

[1]The Center for Torticollis, Department of Physical Medicine and Rehabilitation, Ajou University School of Medicine, Suwon, Republic of Korea
[2]Department of Biomedical Informatics, Ajou University School of Medicine, Suwon, Republic of Korea
[3]Department of Plastic and Reconstructive Surgery, Ajou University School of Medicine, Suwon, Republic of Korea
[4]Department of Pathology, Ajou University School of Medicine, Suwon, Republic of Korea
[5]Brain Disease Research Center, Ajou University School of Medicine, Suwon, Republic of Korea
[6]Department of Radiology, Ajou University School of Medicine, Suwon, Republic of Korea
[7] Department of Surgery, Eulji General Hospital, Seoul, Republic of Korea
[8]Department of Surgery, Ajou University School of Medicine, Suwon, Republic of Korea
[9]Department of Rheumatology, The Hospital for Rheumatic Diseases, Hanyang University College of Medicine, Seoul, Republic of Korea

Congenital muscular torticollis (CMT) is characterized by thickening and/or tightness of the unilateral sternocleidomastoid muscle (SCM), ending up with torticollis. Our aim was to discover differentially expressed genes (DEGs) and novel protein interaction network modules of CMT and to discover the relationship between gene expressions and clinical severity of CMT or protein expressions encoded by DEG. Twenty-three sternocleidomastoid muscle (SCM) of CMT patients and 5 normal SCMs were allocated for microarray, MRI, or imunohistochemical studies. We identified 269 genes as the DEGs in CMT. Gene ontology enrichment analysis revealed that the main function of the DEGs is for extracellular region part during developmental processes. Five CMT-related protein network modules were identified, which showed that the important pathway is fibrosis related with collagen and elastin fibrillogenesis with an evidence of DNA repair mechanism. The expression levels of some meaningful DEGs showed good correlation with the pre-operational MRI color intensities of CMT, indicating clinical severity. Moreover, the protein expressions encoded by the DEGs confirmed the different gene expressions of CMT. We provided an integrative analysis of CMT from gene expression to clinical indication, which showed good correlation with clinical severity of CMT. Furthermore, the CMT-related protein network modules were identified, which provided more in-depth understanding of pathophysiology of CMT.

S10-4 Paper

Detecting Early-warning Signals of type 1 Diabetes and its Leading Biomolecular Networks by Dynamical Network Biomarkers

Xiaoping Liu[1,2], Rui Liu[3,4], Xing-Ming Zhao[2], Luonan Chen[1,2,4]

[1]*Key Laboratory of Systems Biology, SIBS-Novo Nordisk Translational Research Centre for PreDiabetes, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China;*
[2]*Institute of Systems Biology, Shanghai University, Shanghai 200444, China;*
[3]*Department of Mathematics, South China University of Technology, Guangzhou 510640, China;*
[4]*Collaborative Research Center for Innovative Mathematical Modelling, Institute of Industrial Science, University of Tokyo, Tokyo 153-8505, Japan*

Type 1 diabetes is a complex disease and harmful to human health, and most of the existing biomarkers are mainly to measure the disease phenotype after the disease onset (or drastic deterioration). Until now, there is no effective biomarker which can predict the upcoming disease (or pre-disease state) before disease onset or disease deterioration. Further, the detail molecular mechanism for such deterioration of the disease, e.g., driver genes or causal network of the disease, is still unclear. In this study, we detected early-warning signals of type 1 diabetes and its leading biomolecular networks based on serial gene expression profiles of NOD mice by identifying new type of biomarkers, i.e., dynamical network biomarkers which form a specific module for marking the time period just before the drastic deterioration of type 1 diabetes. Specifically, two dynamical network biomarkers were obtained to signal the emergence of two critical deteriorations for the disease, and could be used to predict the upcoming sudden changes during the disease progression. We found that the two critical transitions led to peri-insulitis and hyperglycemia in NOD mices, which are consistent with the experimental results. Hence, the identified dynamical network biomarkers can be used to detect the early-warning signals of type 1 diabetes and predict upcoming disease onset before the drastic deterioration. In addition, we also demonstrated that the leading biomolecular networks are causally related to the initiation and progression of Type 1 diabetes, and provide the biological insight into the molecular mechanism of type 1 diabetes. Experimental data and Functional analysis on DNBs validated the computational results.

Paper

Creating Subnetworks from Transcriptomic Data on Central Nervous System Conditions Informed by a Massive Transcriptomic Network

Yaping Feng[1], Judith A. Syrkin-Nikolau[2], Eve S. Wurtele[1]

[1]Iowa State University, Department of Genetics, Development and Cell Biology, Ames, IA 50011, USA, [2] Macalester College, MN, 55105

We use a human pairwise co-expression matrix derived from a large dataset (>18,000 samples) of high quality publicly available transcriptomic data representing relationships in gene expression across a diverse set of biological conditions (1) as a context network to explore CNS transcriptomics. In oneapproach, we derive a network from within the CNS samples, derive gene clusters, and compare thesignificance of these to the clusters derived from the larger network. In the second approach, we identify genes that characterize individual subsets of samples from within a disease condition. Specifically, differences in gene expression within and between to two designations of glial cancer, astrocytoma and glioblastoma, are evaluated in the context of the broader network. Such related groups of genes, termedoutlier-networks tease out abnormally expressed genes and the particular samples they are associated with. This study identifies a set of 48 subnetworks of outlier genes belong to astrocytoma and glioblastoma.As a case study, we investigate the relationships among the genes of a small astrocytoma-only subnetwork.

[S1] Bio big data processing and integration

Chair: Sun Kim (Seoul Nat. Univ.)

S1-1 9:10 ~ 9:30	"Unified framework for multi-level biosystem modeling" Doheon Lee (KAIST)
S1-2 9:30 ~ 9:50	"Rapid denoising of pyrosequenced amplicons for metagenomics" Sungroh Yoon (Seoul Nat. Univ.)
S1-3 9:50 ~ 10:10	"Integrative approaches for DNA copy number aberrations in cancer" Hyunju Lee (GIST)
S1-4 10:10 ~ 10:30	Reference-assisted post-assembly of a de novo assembled genome Jaebum Kim (Konkuk Univ.)

S1-1 Unified framework for multi-level biosystem modeling

Doheon Lee (KAIST)

Several reasons including robustness, redundancy, and crosstalk of bio-networks, have been reported to explain the limited efficacy and unexpected side-effects of clinical drugs. Many pharmaceutical laboratories have begun to develop multi-compound drugs to remedy this situation, and some of them have shown successful clinical results. Simultaneous application of multiple compounds could increase efficacy as well as reduce side-effects through pharmacodynamics and pharmacokinetic interactions. However, such approaches require overwhelming cost of preclinical experiments and tests as the number of possible combinations of compound dosages increases exponentially. Computer model-based experiments have been emerging as one of the most promising solutions to cope with such complexity. Though there have been many efforts to model specific molecular pathways using qualitative and quantitative formalisms, they suffer from unexpected results caused by distant interactions beyond their localized models. This talk presents an initial design of a unified multi-level framework for whole-body-scale biosystem modeling based on the object-oriented technology and predicate calculus. It demonstrates the expressiveness of the framework with the Type 2 diabetes (T2D) model, which involves the malfunction of numerous organs such as pancreas, circulation system, liver, and muscle.

S1-2 Rapid denoising of pyrosequenced amplicons for metagenomics

Sungroh Yoon (Seoul Nat. Univ.)

Metagenomic sequencing has become a crucial tool for obtaining a gene catalogue of operational taxonomic units (OTUs) in a microbial community. In biomedical research, sequencing-based approaches have been proposed to understand the impact of intestinal microbes on human health by assessing their genetic properties. High-throughput pyrosequencing is a next-generation sequencing technique very popular in microbial community analysis due to its longer read length compared to alternative methods. Computational tools are inevitable to process raw data from pyrosequencers, and in particular, noise removal is a critical informatics step to obtain robust sequence reads. However, the slow rate of existing denoisers has bottlenecked the whole pyrosequencing process, let alone hindering efforts to improve robustness. To address these, we propose a new

approach that can accelerate the denoising process substantially. Our approach can effectively reduce overestimating the number of OTUs, producing more realistic estimates of species-level OTUs than a state-of-the-art alternative under the same condition. Leveraged by our approach, we hope that metagenomic sequencing will become an even more appealing tool for microbial community analysis.

S1-3 Integrative approaches for DNA copy number aberrations in cancer

Hyunju Lee (GIST)

This talk presents integrative approaches for the analysis of DNA copy number aberrations (CNAs) and gene expressions in cancer. CNAs are one of important molecular signatures in cancer initiation, development, and progression.However, these aberrations span through a wide range of chromosomes, so it is hard to distinguish cancer related genes from other genes that are not closely related to cancer but located in the broadly aberrant regions. To address this issue, we developed the wavelet based method to distinguish cancer-driving genes from passenger genes. We also developed the biclustering method for revealing the structural changes in DNA and functional changes in RNA to discover cancer related pathway.

S1-4 Reference-assisted post-assembly of a de novo assembled genome

Jaebum Kim (Konkuk Univ.)

Next-generation sequencing (NGS) technologies together with de novo assembly algorithms have provided us the unprecedented opportunity to unravel the genomes of different species at low cost. This trend will be further accelerated by the launch of several large-scale genome projects such the Genome 10K and i5k projects. However, due to the limitation of read length of NGS and the lack of physical map for most of the target species, identifying the order and orientation of assembled sequence scaffolds from the de novo genome assembly on chromosomes is still a pressing challenge and is largely unresolved. To address this problem, we developed a novel computational method, called RACA, to reliably assemble the sequence scaffolds of a de novo genome assembly into longer chromosomal fragments without using the genetic or physical map. Given the de novo genome assembly of a target species together with a closely related species as reference and one or more outgroup genomes, RACA reconstructs highly probable order and orientation of the scaffolds in the target species based on (i) probabilities that represent the likelihood of scaffold adjacencies by considering the evolution of a genome and (ii) the coverage of paired-end reads from the target species. Simulation results indicated that our approach can be applied to any de novo assembled genome if a good reference assembly of a closely related species is available, achieving 98 % accuracy. We applied our method to the reconstruction of Tibetan antelope chromosomes based on the 1,434 scaffolds assembled by SOAPdenovo using cattle genome as the reference and human genome as the outgroup. Our method was able to further assemble these scaffolds into 105 chromosome fragments, of which 13 chromosome fragments correspond to complete cattle chromosomes. In addition, we were able to identify Tibetan antelope scaffolds that span 46 known evolutionary breakpoint regions, and 130 mis-assembled scaffolds in the de novo assembly. Experimental validation showed that our predictions are highly accurate. As read length of NGS increases by the implementation of new technologies, the chromosome assemblies obtained from our method will become even more accurate. We believe this method will significantly facilitate the study of chromosome evolution and genome rearrangement for large number of genomes sequenced by NGS.

[S2] Next-generation sequencing for next-generation biology

Chair: Jung Kyoon Choi (KAIST)

S2-1 10:50 ~ 11:10	"Regulation of nucleosome positioning and modification in transcription factor binding regions" Jung Kyoon Choi (KAIST)
S2-2 11:10 ~ 11:30	"Genome-wide decoding of mRNA and miRNA maps" Sung Wook Chi (Sungkyunkwan Univ.)
S2-3 11:30 ~ 11:50	"Unraveling of design principle in bacterial genomes" Byung-Kwan Cho (KAIST)
S2-4 11:50 ~ 12:10	"Application of NGS in improvement of efficiency in radiotherapy" Buhyun Youn (Pusan Nat. Univ.)

S2-1 Regulation of nucleosome positioning and modification in transcription factor binding regions

Jung Kyoon Choi (KAIST)

Promoters and enhancers maintain nucleosome-depleted, open chromatin status but simultaneously, require the presence of nucleosomes for specific histone modifications. Based on extensive chromatin datasets from the ENCODE project and those published in human T cells, we investigated the positioning and modification of nucleosomes in regions covering the conserved binding motif of 258 transcription factors (TFs). We found that TF binding sites (TFBSs) are embedded in nucleosome-encoding DNA sequences and buried under nucleosomes in vivo when inaccessible. Upon TF binding, nucleosomes seem to be repositioned in the flanking regions and define the boundaries of accessible chromatin while the central nucleosome covering the TFBS being replaced by a H2AZ-contaning nucleosome carrying activating histone marks such as H3K4me3, H3K27ac, etc. Our genome-wide analyses demonstrate how the orchestration of TF binding is accomplished concertedly by DNA sequences, nucleosome positioning, histone modifications, and histone variants.

S2-2 Genome-wide decoding of mRNA and miRNA maps

Sung Wook Chi (Sungkyunkwan Univ.)
Graduate School, Department of Health Sciences & Technology,
Samsung Advanced Institute for Health Sciences & Technology (SAIHST),
Sungkyunkwan University

The limited number of primary transcripts in the genome has promoted interest in the possibility that much of the complexity in the regulation of gene expression may be determined by RNA regulation controlled by RNA-binding proteins (RNABPs) and/or microRNAs (miRNAs). However, applying biochemical methods to understand such interactions in living tissues is major challenge. Here we developed a genome-wide means of mapping messenger ribonucleoprotein (mRNP) sites in vivo, by high-throughput sequencing of RNA isolated by crosslinking immunoprecipitation (HITS-CLIP), providing genome-wide maps of RNABP-RNA interactions in vivo. Furthermore, HITS-CLIP analysis is extended to the problem of identifying miRNA targets, for which prediction is a major challenge since miRNA activity requires base pairing through only 6-8 "seed" nucleotides. By generating crosslinking of native Argonaute (Ago) protein-RNA complexes in mouse brain, Ago HITS-CLIP produced two simultaneous datasets—Ago-miRNA and Ago-mRNA binding sites—that were combined with bioinformatic analysis to identify miRNA-target mRNA interaction sites. We validated genome-wide interaction maps for miR-124, and generated additional maps for the 20 most

abundant miRNAs present in P13 mouse brain. Not all Ago mRNA clusters correspond to known seed sequence, leading to the discovery of a new rule for miRNA-mRNA interactions, termed “pivot pairing rule”. HITS-CLIP provides a general platform to identify functional mRNP and miRNA binding sites in vivo and a solution to determining precise sequences for targeting clinically relevant sites of RNA regulation.

References

1. Chi SW*, Hannon GJ, Darnell RB*. An alternative mode of miRNA target recognition. ***Nature Structural and Molecular Biology*** 2012, 19: 321-327 (*Co-corresponding Author).
2. Darnell JC, Van Driesche S, Zhang C, Hung KYS, Aldo M, Fraser CE, Stone EF, Chen C, Fak J, Chi SW, Licatalosi D, Richter JD and Darnell RB. 2011. FMRP Stalls Ribosomal Translocation on mRNAs Linked to Synaptic Function and Autism. ***Cell*** 2011, 146 (2): 247-61
3. Chi SW, Zang JB, Mele A, Darnell RB. Argonaute HITS-CLIP decodes miRNA-mRNA interaction maps. ***Nature*** 2009, 460 (7254): 479-86
4. Licatalosi DD, Mele A, Fak JJ, Ule J, Kayikci M, Chi SW, Clark TA, Schweitzer AC, Blume JE, Wang X, Darnell JC, Darnell RB. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. ***Nature*** 2008, 456 (7221): 464-9.

S2-3 Unraveling of design principle in bacterial genomes

Byung-Kwan Cho (KAIST)

Department of Biological Sciences and KI for the BioCentury, KAIST

Over the past decade or so, dramatic developments in our ability to experimentally determine the contents and functions of genomes have taken place. In particular, high-throughput technologies are now inspiring a new understanding of the bacterial genome on a global scale. Bacterial genomes are organized by structural and functional elements, including promoters, transcription start and termination sites, open reading frames, regulatory noncoding regions, untranslated regions and transcription units. In the work presented here, methods have been developed for the genome-wide understanding of the structural and functional elements of bacterial genomes using massive high-throughput sequencing. Furthermore, we iteratively integrate high-throughput, genome-wide measurements to determine the organizational structure of the Escherichia coli K-12 MG1655 genome. Based upon the enormous genomic information, the artificial genome can be designed and synthesized. In an effort to supply only the necessary genes for full functionality, and therefore reduce the complexity of the metabolic and regulatory networks, large blocks of nonessential genes have been deleted. However, removal of the nonessential genes resulted in reduction of the growth rate of the reduced-genome E. coli on M9 minimal medium. To investigate the ability of the reduced-genome strain to overcome the reduction of growth rate, we monitored the acquisition and fixation of mutations that conveyed a selective growth advantage during adaptation to the M9 minimal medium by whole-genome resequencing. This comprehensive genetic information on the reduction of microbial genome will provide the foundation for designing and rewriting an artificial genome.

S2-4 Application of NGS in improvement of efficiency in radiotherapy

Buhyun Youn (Pusan Nat. Univ.)

Department of Biological Sciences, Pusan National University

Radiotherapy plays a critical role in the treatment of non-small cell lung cancer (NSCLC). However, radioresistance has been considered as a main factor restricting efficacy of radiotherapy. Despite several experimental and clinical studies for resistance to radiation, the precise mechanism of radioresistance still

remains unclear. It might be involved in a partial understanding of the cellular radioresistance mechanism at a single molecule level. In this study, we aimed to investigate the radiation-induced alteration of gene expression at entire transcriptome level and to identify the critical radioresistance-related genes in radioresistant NSCLC cells. We suggested that RNA-seq (a massive sequencing-based technique) could be an ideal approach to gain insight into the complex radiation response and to overcome the limitation of previous studies indicating somewhat fragmentary evidences of radioresistance. Using RNA-seq, we conducted quantitative and qualitative analysis of radiation-induced gene expression patterns in radioresistant A549 NSCLC cells. Then, bioinformatic approaches such as Gene Ontology (GO) analysis and Ingenuity Pathway Analysis (IPA) were applied to identify enriched GO categories, and molecular networks and hub genes with their interacting partners, respectively. They provided useful information concerning regulatory factors associated with cellular radioresistance in NSCLC cells. Furthermore, the exact radiation responses focusing on these putative regulators were investigated using biological/biochemical studies and compared in two NSCLC cell lines, A549 and NCI-H460 with different radiosensitivity. As a result, we found a novel radioresistance mechanism through functional orchestration of rpS3, TRAF2 and NF-κB in response to ionizing radiation (IR). We demonstrated that IR-dependent rpS3-TRAF2 complex dissociation by phosphorylation of both rpS3 and TRAF2 is a key control point of radioresistance in NSCLC cells. These results suggested that modulation of rpS3 and TRAF2 could effectively regulate the tumor radioresistance. In our research, combined analysis of transcriptome sequencing, subsequent bioinformatic analysis and biological/biochemical examination was first applied to investigate the radioresistance mechanism of NSCLC cells. It could provide useful information on identification of potential biomarker of radioresistance, help to understand the complex radiation responses, and ultimately contribute to effective radiotherapy of NSCLC.

[S3] Computational biology-molecular modeling/simulations

Chair: Keun Woo Lee (Gyeongsang Nat. Univ.)

S3-1 13:10 ~ 13:30	"Protein function prediction by community detection of a PPI network" Jooyoung Lee (KIAS)
S3-2 13:30 ~ 13:50	"Structural and thermodynamic investigation of protein aggregation in water" Sihyun Ham (Sookmyung Women's Univ.)
S3-3 13:50 ~ 14:10	"Single-molecule study on DNA mismatch repair protein" Jong-Bong Lee (POSTECH)
S3-4 14:10 ~ 14:30	"Barriers and wells to ion translocation in the Connexin 26 Hemi-channel" Myunggi Yi (Pukyong Nat. Univ.)

S3-1 Protein function prediction by community detection of a PPI network

Jooyoung Lee (KIAS)

Juyong Lee[1], Steven P Gross[2] and Jooyoung Lee[1]*

[1]Center for in silico protein science, Korea Institute for Advanced Study, Seoul, Korea [2]University of California, Irvine, USA

In the post-genomic era, we are overwhelmed by a deluge of experimental data, and network science has the potential to become an invaluable method to increase our understanding of large interacting datasets. However, this potential is often unrealized for two reasons: uncovering the hidden community structure of a network—known as community detection—is difficult, and further, even if one has an idea of this community structure, it is not a priori obvious how to efficiently use this information. Here, within the context of protein function prediction, we address both these issues. First, we present a new community detection method that is faster than the current state of the art, and generates better solutions, allowing extraction of additional hidden information. Second, we develop a better approach to use this community information to predict proteins' functions: we determine when and why this community information is important. We show that some classes of prediction benefit from relatively simply local community information, but that others with longer-range interactions benefit from very careful determination of community structure. In such cases, our community-based approach uncovers hidden non-local information allowing improved prediction of protein function. Thus, for the first time, using community information we can predict function better than methods that only use local information.

S3-2 Structural and thermodynamic investigation of protein aggregation in water

Sihyun Ham (Sookmyung Women's Univ.)

Department of Chemistry, Sookmyung Women's University

Hyochangwon-gil 52, Yongsan-gu, Seoul 140-742, Korea

Globular proteins may convert their native conformations into non-native forms due to the intrinsic or external perturbations in microenvironmental conditions. When a protein is in its non-native state, it can further unfold or self-assembly to form amyloid aggregates, which are presumably toxic and consequently can cause various diseases. In this regard, protein aggregation is one of the most actively investigated issues in relation to the development of the therapeutics and medical applications for the cure of protein aggregation diseases. We recently developed novel theory and computational methods to effectively execute thermodynamic quantities

on the dynamical process of protein, especially protein misfolding and aggregation processes. By taking advantage of this novel method, we have successfully provided the molecular basis on the experimentally observed folding, misfolding, and aggregation phenomena on various proteins. Here, I report our recent efforts on the structural and thermodynamic investigations on the folding, misfolding and aggregation processes of various proteins using the unguided, fully atomistic, explicit-water molecular dynamics (MD) simulations as well as the integral-equation theory of liquids. The MD simulated atomic-level structures of various aggregation-prone proteins are provided and verified by the experimentally available structural data. Gibbs free energy and its constituents including protein internal energy, protein configurational entropy, and solvation free energy are reported to elucidate thermodynamic driving forces for the misfolding and aggregation processes of proteins in water. I will address the role of water as well as the role of key residues in initiating the misfolding and aggregation processes of various proteins associated with diseases.

S3-3 Single-molecule study on DNA mismatch repair protein

Jong-Bong Lee (POSTECH)
Department of Physics/School of Interdisciplinary Bioscience & Bioengineering
POSTECH

It has been studied that DNA repair proteins search a target via a 1-dimensional diffusion along naked DNA. However, this single-molecule tracking approach lacked of understanding the catalytic function of the repair proteins and moreover the mechanism of the downstream transactions on the DNA. I will present MutS mechanics on mismatched DNA: the catalytic processes of the mismatch repair initiation protein. Our single-molecule analysis reveals that ATP alters the MutS diffusion mechanics: ATP-bound MutS formed at the mismatch is released from the mismatch and diffuses along the DNA for the downstream signaling, which shows a distinct diffusion mechanism from the MutS that moves along the DNA backbone to find the unpaired nucleotide. The studies propose a solution for a highly controversial issue in DNA mismatch repair.

S3-4 Barriers and wells to ion translocation in the Connexin 26 Hemi-channel

Myunggi Yi (Pukyong Nat. Univ.)
Department of Biomedical Engineering, Pukyong National University, Busan 608-737, Korea

Cells can communicate with outside environment through the membrane proteins such as channels and receptors. Gap junction channels (GJC) are found in most mammalian tissues and mediate the exchange of various ions, nicotinamide adenine dinucleotide (NAD+), cyclic adenosine monophosphate (cAMP), ATP, glutamate, protaglandin etc.. Humans have more than 20 different members of connexins composing the GJC's, and their mutations cause several human diseases, including neurodegenerative diseases, skin diseases, and deafness. Connexin26 participates in the physiology of a variety of organs including liver, kidney, intestine, lung, spleen, testes, brain, cochlea, and skin. Despite the wells of studies, the mechanism of channel gating and ion selectivity remains unclear. In order to characterize the ion conductance of the channel we have performed free energy calculations using all-atom molecular dynamics simulations with cennexin26 hemi-channel and gap junction channel in explicit membrane bilayers. The sources of the potential barriers and wells were identified and described in atomic detail.
This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2012R1A1A1012707) and KISTI super computing center (KSC-2011-C2-44).

[S4] Systems biology: evolution to translational medicine

Chair: Daehee Hwang (POSTECH)

S4-1 14:50 ~ 15:10	"Sociology in the genetic world" Pan-Jun Kim (APCTP)
S4-2 15:10 ~ 15:30	"Genome-wide analysis and modeling of CpG methylation in 30 breast cancer cell lines" Sun Kim (Seoul Nat. Univ.)
S4-3 15:30 ~ 15:50	"Opening the systemic analysis of ubiquitination-mediated protein regulation network" Gwan-Su Yi (KAIST)
S4-4 15:50 ~ 16:10	"Alteration of epigenome landscaping is linked to neurodegeneration" Hoon Ryu (Seoul Nat. Univ.)

S4-1 Sociology in the genetic world

Pan-Jun Kim (APCTP)
Leader of the Junior Research Group, Asia Pacific Center for Theoretical Physics
Adjunct Professor, Department of Physics, POSTECH
TEL: +82-54-279-8678
Email: pjkim@apctp.org OR panjun@postech.ac.kr

Genes in organisms have a number of social interactions with one another in their biological contexts. For example, proteins produced from one gene may interact with other proteins produced from another gene to perform together a particular biological task, and such pair of cooperative genes may often reside together in the same organisms. We analyzed thousands of genes across ~600 bacterial species, and found genes with favored co-occurrence in the same organisms (termed correlogs) or disfavored co-occurrence (termed anti-correlogs). These co-occurrence patterns are significantly reflective of actual biochemical interplays between genes, and distinct cliques of correlogous genes are seamlessly interrelated through anti-correlogous links between the cliques. The 'sociology' of genes inferred by this approach provides useful information on how to engineer a cell, such as for production of a desired bioproduct.

S4-2 Genome-wide analysis and modeling of CpG methylation in 30 breast cancer cell lines

Sun Kim (Seoul Nat. Univ.)
Computer Science and Engineering,
Bioinformatics Institute,
Interdisciplinary Program in Bioinformatics,
Seoul National University, Seoul 151-744, Korea

Aberrant DNA methylation of genomic regions, including CpG islands, CpG shores, and first exons, plays a key role in the altered gene expression patterns characteristic of all human cancers. In 30 breast cancer cell lines representing different tumor phenotypes, we conducted a comprehensive analysis of genome-wide DNA methylation patterns and also an integrated analysis to identify the relationship between DNA methylation in these different genomic regions and expression levels of downstream genes, using MethylCap-seq data (affinity purification followed by next-generation sequencing of eluted DNA) and Affymetrix gene expression microarray data. In this talk, we report genome-wide methylation profiles associated with different molecular

subtypes of human breast cancer (luminal, basal A, and basal B) and the effect of DNA methylation on gene expression in breast cancer phenotypes. On the computational side, we show how to identify differentially methylated regions across multiple cells and how to perform integrated analysis of DNA methylome data and gene expression data for the omics data from multiple cells with different phenotypes.

S4-3 Opening the systemic analysis of ubiquitination-mediated protein regulation network

Gwan-Su Yi (KAIST)

Speaker: Prof. Gwan-Su Yi , KAIST Korea

Ubiquitination is a regulatory process responsible for regulating the degradation of proteins influencing nearly all cellular processes. In the ubiquitination process, ubiquitin-protein ligase (E3) plays a key role by recognizing substrates for ubiquitination. Thus, comprehensive knowledge about the substrate specificity of E3s can enhance the understanding of the regulatory mechanisms of cellular processes by adding the regulation of protein degradation. The existing findings of substrate specificity of E3 are, however, scattered over a number of resources, and thus E3-mediated regulatory mechanisms cannot be analyzed in a systemic or systematic manner. E3Net system has been constructed to provide a comprehensive collection of available E3-substrate relations from available datasets, databases and literature. Currently, E3Net contains 2349 E3s and 5307 substrates in 449 organisms and 1750 E3-substrate specific relations between 525 E3s and 1335 substrates in 48 organisms. On the basis of substantially enhanced E3-substrate relation data in E3Net, we could build a framework for systematic analysis of E3-mediated regulatory networks involved in diverse cellular processes. In this framework, the user can create an integrative view of various function terms for each E3-specific substrate group and multiple E3s correlated with each function term together. This resource will facilitate further in-depth investigation of ubiquitination-dependent regulatory mechanisms.

S4-4 Epigenome alteration is linked to neurodegeneration: Is it a Marker or a Maker?

Hoon Ryu (Seoul Nat. Univ.)

WCU Neurocytomics Group, Department of Biomedical Sciences, Seoul National University Graduate School, Seoul, South Korea

Department of Neurology, Boston University School of Medicine and
VA Boston Healthcare System, Boston, MA 02130, USA

The conventional wisdom is that DNA acts as a master regulator for controlling the destiny of the cell. However, a growing body of evidence indicates that epigenetic modifications are pivotal in organizing when, what, and how genes are turned on and how cellular events those underlie aging processes and neurodegenerative conditions are regulated. Epigenetic changes encompass an array of molecular modifications to both DNA (methylation) and chromatin (acetylation, methylation, phosphorylation, sumoylation, and ubiquitination), including miRNAs. Huntington's disease (HD) is an autosomal dominant neurodegenerative disease caused by an expanded trinucleotide CAG repeat in the gene coding for huntingtin (Htt). Deregulation of chromatin remodeling is linked to the pathogenesis of HD but the mechanism is elusive. In this regard, my lab is studying which “Histone Code” under neurodegenerative conditions deregulates epigenomes. We are performing histone-ChIP genome-wide sequencing combined with RNA-sequencing followed by platform integration analysis. We are analyzing genomes involving neuronal synaptic transmission, cell motility, and neuronal differentiation pathways that were down regulated while their promoter regions were highly occupied with methylated H3K9 in HD. The complexity that can be achieved with a single modification such as methylation of histone H3K9 is remarkable. Taken together, our understanding about mechanisms of epigenetic modification may warrant future investigations for identifying novel biological markers and therapeutic targets for HD including other neurodegenerative disorders

Posters

TBC-1: *Young-Hyun Kim, Jae-Won Huh and Kyu-Tae Chang*

Recently exonized Alu elements in Macaca fascicularis

TBC-2: *Jae-Won Huh and Kyu-Tae Chang*

Genome diversification mechanism between human and chimpanzee

TBC-3: *Kyu-Tae Chnag*

Transcriptome sequencing and gene analyses in the crab-eating macaque

TBC-4: *Yoon-Ho Hong, Malcolm Campbell, Kyungjoon Lee, In-Hee Lee and Sek-Won Kong*

Recent Positive Selection in Human Genes That are Enriched for Disease Mutations, but Limited for Polymorphism

TBC-5: *Sung Han Kim and Seok Soo Byun*

Gene expression changes as resistant markers to cisplatin in a panel of bladder cancer cell lines

TBC-6: *Vitaliy Kuznetsov, Elena Zemlyanskaya and Sergey Degtyarev*

GlaI-qPCR assay — a new instrument for quantitative DNA methylation analysis and its application for tumor suppressor genes study

TBC-7: *Pin-Cian Wang, Liang-Chuan Lai, Mong-Hsun Tsai, Eric Y. Chuang, Cheng-Yan Kao and Pei-Chun Chen*

A Filtering Algorithm for Gene-Gene Interaction using Case-Only Data

TBC-8: *Seon-Kyu Kim, Young-Kyu Park and Seon-Young Kim*

20-gene-based risk score classifier predicts disease recurrence in non-muscle invasive bladder cancer

TBC-9: *Sanghoon Moon, Kwang Su Jung, Young Jin Kim, Miyeong Hwang, Kyungsook Han, Bok-Ghee Han, Jong-Young Lee, Kiejung Park and Bong-Jo Kim*

Genome-wide analysis of CNV and SNP in Koreans

TBC-10: *Kuei-Chung Shih, Hsiao-Chieh Chi and Chuan-Yi Tang*

3D-QSAR Pharmacophore Modeling of Thromboxane A2 Receptor for Discovery New Inhibitors

TBC-11: *Su Yeon Kim and Terry Speed*

Comparison of somatic mutation-calling methods based on DNA sequence from matched tumor-normal pairs

TBC-12: *Nam Hee Kim, Youngdoe Kim, Young Jin Kim, Ji Hee Oh, Mee Hee Lee and Juyoung Lee*

The estimation of heritability analyses for BMI using genotype score based on Korean Cohort

TBC-13: *Ji Hee Oh, Young Jin Kim, Sanghoon Moon, Jong-Young Lee and Yoon Shin Cho*

Genotype instability during long-term subculture of lymphoblastoid cell lines

TBC-14: *Jaeyun Sung, Pan-Jun Kim, Leroy Hood, Donald Geman and Nathan Price*

Multi-study integration of brain cancer transcriptomes reveals organ-level diagnostic signatures

TBC-15: *Vasilina A. Sokolova, Valery A. Chernukhin, Danila A. Gonchar, Elena V. Kileva, Larisa N. Golikova, Vladimir S. Dedkov, Natalya A. Mikhnenkova, Elena V. Zemlyanskaya, Vitaliy V. Kuznetsov and Sergey Kh. Degtyarev*

Methyl-directed Site-specific DNA Endonuclease MteI is a New Instrument for Analysis of CpG Island Methylation

TBC-16: *Lyong Heo, Young Jin Kim, Sanghoon Moon and Jong-Young Lee*

Nonunique SNP problems in association study

TBC-17: *Young Jin Kim, Kwang Joong Kim, Lyong Heo, Yun Kyoung Kim, Sanghoon Moon, Youngdoe Kim, Mi Yeong Hwang, Bong-Jo Kim and Jong-Young Lee*

Exonic variants in Korean population

TBC-18: *Si Ra Kim, Seung Ho Park, Bum Joon Park, Kwang Soo Jang and In Young Choi*

Development of Korea Common Data Model for Adverse Drug Signal Detection based on multi-center EMR systems

TBC-19: *Doo Yang and Ilya Ioshikhes*

Various nucleosome positioning patterns in Drosophila

TBC-20: *Soo-Yong Shin, Yongdon Shin, Yong-Man Lyu, Hyo Joung Choi, Jihyun Park and Jaeho Lee*

Anonymized Patient Chart Review Tool in Asan Medical Center

TBC-21: *Jongkwang Kim, Gao Long and Kai Tan*

Integrate Genomics and Molecular Interactome Data for Brain Tumor Pathway Discovery and Prognosis

TBC-22: *Hyun-Young Kim, Hyeoun-Ae Park, Yul Ha Min and Eun-Joo Jeon*

Development of a Consumer-engaged Obesity Management Ontology based on Nursing Process

TBC-23: *Jee Yeon Heo, Yongjin Choi, Hae-Seok Eo, Youngho Kim, Taesung Park and Hyung-Seok Choi*

Performance of microRNA target prediction algorithms

TBC-24: *Thomas Thorne, Pietro Fratta, Michael Hanna, Elizabeth Fisher and Michael Stumpf*

Graphical modeling of regulatory interactions in sporadic Inclusion Body Myositis

TBC-25: *Eiru Kim and Insuk Lee*

Jiffynet: A web server generating Gene networks for newly sequenced species

TBC-26: *Tak Lee, Jung Eun Shim and Insuk Lee*

Studying Plant Complex Traits Through Network-assisted Systems Genetics of *Arabidopsis Thaliana*

TBC-27: *Nayoung Kim and Sukjoon Yoon*

Systematic analysis of cell line data for the development of novel cancer treatment

TBC-28: *Kang-Hoon Lee, Kyung-Seop Shin, Woo-Chan Kim, Jeongkyu Roh, Seung-Ho Choi, Dong-Ho Cho and Kiho Cho*

Genome Signature Image (GSI): Concise visualization of species/strain-specific profiles of repetitive element occurrences for cataloguing and evolutionary studies

TBC-29: *Mi Yeong Hwang, Sanghoon Moon, Young Jin Kim, Lyong Heo, Yun Kyoung Kim,Youngdoe Kim, Bok-Ghee Han, Jong-Young Lee, and Bong-Jo Kim*

Analysis of copy number variation in exome sequencing data

TBC-30: *Hyunju Ryoo, Minyoung Kong, Younyoung Kim and Chaeyoung Lee*

Identification of functional nucleotide sequence variant in the promoter of CEBPE gene

TBC-31: *Wonhee Jang, Hyunju Ryoo, Jihye Ryu, Jeyoung Woo, Minyoung Kong, Younyoung Kim and Chaeyoung Lee*

Functional promoter nucleotide variants and their haplotypes of the gene encoding CCL21

TBC-32: *Boyoung Kim, Seung-Min Baek and Sunmi Choi*

Development of Web-based Case Report System in Traditional Korean Medicine for Clinic Doctor

TBC-33: *Jehoon Jun, Minjae Yoo and Kwang-Hwi Cho*

ChemTools : Python based Chemoinformatics Toolkit

TBC-34: *Ok Sung Jung, Bong Hun Ji and Kwang-Hwi Cho*

Molecular Dynamic Studies to predicted protein-protein interactions using GPU accelerated AMBER : application to

TBC1 interacting Rab family proteins

TBC-35: *Sung Hee Park and Sangsoo Kim*
A Novel Data Mining Approach for Inferring Phenotypic Association Networks to Discover the Pleiotropic Effects

TBC-36: *Zsolt Boldogkoi and Dora Tombacz*
Transcription Interference Networks are the coordinators of the gene expressions

TBC-37: *Sanghun Bae, Hyunwook Han, Hanwool Kim and Jisook Moon*
Subnetwork-based analysis of human disease in protein complex with housekeeping functions

TBC-38: *Jeyoung Woo, Minyoung Kong, Younyoung Kim and Chaeyoung Lee*
Functional haplotypes in 5' region of RGS14 gene

TBC-39: *Young-Ho Yun, Ye-Ni Choi, Moon-Kyung Shin, Kwang-Choon Kim and Jaegeol Cho*
Health SORA, the Smart Health Care Program for Cancer Survivors

TBC-40: *Jimin Shin, Hyunmin Kim, Chaeyoung Lee and David Bentley*
A computational framework for differential alternative polyadenylation profiles between cancer and normal cells

TBC-41: *Han Wool Kim, Hyun Wook Han, Sang Hun Bae and Ji Sook Moon*
The genetic regulation of aging process and age-related disease

TBC-42: *Jung Eun Shim and Insuk Lee*
Discovery of Pathway Information Content of Protein Domains based on Domain Co-occurrence Network

TBC-43: *Haein Kim, Ensel Oh, Young Kee Shin and Yoon-La Choi*
Identification and Characterization of Gastric Cancer Subtypes using Expression Microarray Data

TBC-44: *Yoonsook Moon, Minyoung Kong, Younyoung Kim and Chaeyoung Lee*
Functional nucleotide polymorphism in the promoter region of WFS1 gene

TBC-45: *Ensel Oh, Yoon-La Choi and Young Kee Shin*
Comparison of Formaldehyde Fixed Paraffin Embedded (FFPE) and Frozen Tissues for Exome Sequencing

TBC-46: *Meehye Kang, Gila Jung, Sung Kim, Wan-Soo Kim and Youn-Ho Lee*
Molecular and biochemical characterization on the artificial hibernation in the olive flounder, *Paralichthys olivaceus*.

TBC-47: *Jihye Ryu and Chaeyoung Lee*
Unraveling selection signatures by composite log likelihood

TBC-48: *Hee-Joon Chung, Byoungoh Kim, Taehun Kim, Keun Bong Kwak and Dongman Lee*
The Health Avatar Platform: development of platform for interacting health agents and personal avatar

TBC-49: *Ningning He and Sukjoon Yoon*
Systematic Analysis of Genotype-dependent Gene Expression Signatures and Drug Sensitivity in NCI60 Datasets

TBC-50: *Jin-Muk Lim, Jung Nyeo Chun, Hong-Gee Kim and Ju-Hong Jeon*
The role of TRP channel interactome in prostate cancer

TBC-51: *Xiaoqi Wang and Sukjoon Yoon*
Using CSSP to predict chameleon peptides

TBC-52: *Haein An, Gila Jung and Chang-Bae Kim*
Transcriptome analysis during the developmental stages for predator induced polyphenism in *Daphnia pulex*

TBC-53: *Junha Shin and Insuk Lee*

Network analysis by phylogenetic profiling revealed domain-specific evolution of cellular pathways

TBC-54: *Minyoung Kong, Younyoung Kim and Chaeyoung Lee*

Functional polymorphism located in the promoter of the coagulation factor XI gene as a putative genetic factor for susceptibility to venous thromboembolism

TBC-55: *SeongBeom Cho, InSong Go, Hyo-Jeong Ban, Hyesun Yoon, Yeunjung Kim, Jaepill Jeon and BokGhee Han*

Temporal gene expression profiles identify genetically determined transcriptional regulation of human leukocytes

TBC-56: *Hyunjung Kang, Sooyoung Cho, Ikjung Choi, Yeongjun Jang, Sanghyuk Lee and Wankyu Kim*

gsGator – an integrated web platform for cross-species gene set analysis

TBC-57: *Taejeong Bae, Kyoohyoung Rho, Yong-Ho In and Sunghoon Kim*

Identification of transcriptional network regulating prognostic gene expression signature of colorectal cancer patients

TBC-58: *Lee Sael and Daisuke Kihara*

Local Similarity Search of Physicochemical Properties in Protein-Ligand Binding Sites

TBC-59: *Meilling Liu, Sanghoon Moon, Youngjin Kim and Sungho Won*

Association analysis of CNV data with linear mixed model

TBC-60: *Young Lee, Suyeon Park, Woojoo Lee and Sungho Won*

Analysis of longitudinal data : Applications of Linear Mixed Model to The Korean Association Resource(KARE)

TBC-61: *Seongwon Cha, Hyunjoo Yu and Jong Kim*

Differential influences of common variants on erythrocyte-related traits according to Sasang constitutional types

TBC-62: *Youngdoe Kim, Jungmin Lim, Donghe Li, Jaemoon Lee and Sungho Won*

Comparing algorithms for genotype imputations in family-based design

TBC-63: *Youngdoe Kim, Yong Ki Jung, Sung Oh Kang, Nam Hee Kim, Young Jin Kim,Juyoung Lee, Sungho Won*

A large-scale genome-wide association study of Korean Family cohorts for genetic variants influencing metabolic syndrome

TBC-64: *Hannah Kim, Ilhak Lee, Ji Yong Park, Sang Hyun Kim and So Yoon Kim*

Ethical, Legal, and Social Frameworks on Issues of Bioinformatics

TBC-65: *Denise Daley, David Zamar, Ben Tripp, Brad Cavanagh and George Ellis*

PATH2: Software for Conducing Gene-Ontology And Pathway Based Analyses using Genome-Wide Association Data

TBC-66: *SoJeong Yi, Sangin Lee, Youngjo Lee, Seonghae Yoon, Inbum Chung, HyeKyung Han, Jae-Yong Chung, Ichiro Ieiri and In-Jin Jang*

Comparison of Genetic Variations in Drug Metabolizing Enzyme and Transporter Genes among Korean, Japanese, and Chinese Population

TBC-1: Recently exonized Alu elements in Macaca fascicularis

Young-Hyun Kim[1,2], Jae-Won Huh[1,2] and Kyu-Tae Chang[1,2]

[1]*National Primate Research Center, Korea Research Institute of Bioscience and Biotechnology, Ochang 363-883, Republic of Korea*
[2]*University of Science & Technology, National Primate Research Center, KRIBB, Ochang 363-883, Republic of Korea*

Crab-eating monkey (*Macaca fascicularis*) and rhesus monkey (*Macaca mullata*) are frequently used and valuable primate model species. Although they most common primate model organism for biomedical approaches, their genetic information is not yet applicable except for rhesus monkey. In this study, we tried to analyze genomic diversity of closely related two macaca species with recently integrated Alu elements. First, the *Macaca fascicularis* mRNA sequences (10221 mRNA) were collected from Genebank database, and 'young' *Alu*-exonized mRNA sequences were sorted by repeatmasker program (216 mRNA). Second, for avoiding the false positive data (avoiding the genomic contaminated cDNA sequences), manual correction were conducted. Third, ten genes were chosen, and eight genes contained young *Alu* element were identified. Finally, for the verification of exonized young *Alu* element, PCR amplification and sequencing procedure were conducted using various human and primate DNA samples. Intriguingly, two genes (C9orf6 and NOLC1 gene) harbor the insertional polymorphic Alu element in their transcript. Although, we did not use the whole genome information of *Macaca fascicularis*, genome wide survey could be a useful tool for understanding the useful primate model organism.

TBC-2: Genome diversification mechanism between human and chimpanzee

Jae-Won Huh[1,2] and Kyu-Tae Chang[1,2]

[1]*National Primate Research Center, Korea Research Institute of Bioscience and Biotechnology, Ochang 363-883, Republic of Korea*
[2]*University of Science & Technology, National Primate Research Center, KRIBB, Ochang 363-883, Republic of Korea*

Chimpanzee is the most closely related living species of human. Human and chimpanzee genome project show that there is only about 1 % genome difference between the two species. Thus, the comparison of gene sequences of two species could show us the genetic components that are related with lineage specific events. We compared and investigated the gene regions between human and chimpanzee using bioinformatic and experimental tools. *In silico* comparison was performed between human and chimpanzee genome. Among the 65248 insertion-deletion (INDEL) loci, 285 genes regions were identified, and 130 gene regions were experimentally validated. Although, 48 gene loci did not show any genetic differences, 32 gene loci showed the lineage specific INDEL events (insertion in human - 12 genes, deletion in chimpanzee – 20 genes). Those INDEL events categorized into five different evolutionary mechanism including retroelements-related (12 genes), homologous recombination and excision (12 genes), tandem repeats variation (5 genes), gene conversion (2 genes), and processed pseudogene formation (1 gene) mechanism. These results suggest that not only simple integration events can drive the genetic differences, but deletion mediated by the recombination event also participate the lineage specific evolutionary events between human and chimpanzee lineage.

TBC-3: Transcriptome sequencing and gene analyses in the crab-eating macaque

Kyu-Tae Chnag[1,2]

[1]*National Primate Research Center, Korea Research Institute of Bioscience and Biotechnology, Ochang 363-883, Republic of Korea*
[2]*University of Science & Technology, National Primate Research Center, KRIBB, Ochang 363-883, Republic of Korea*

As a human mimic, the crab-eating macaque (*Macaca fascicularis*) is an invaluable non-human primate model for biomedical research, but the lack of genetic information on this primate has represented a significant obstacle for its broader use. Here, we sequenced the transcriptome of 16 tissues and identified genes to resolve the main obstacles for understanding the biological response of the crab-eating macaque. From 4 million reads with 1.4 billion base sequences, 31,786 isotigs containing genes similar to those of humans, 12,672 novel isotigs, and 348,160 singletons were identified using the GS FLX sequencing method. Approximately 86% of human genes were represented among the genes sequenced in this study. Additionally, 175 tissue-specific genes were identified, 81 of which were experimentally validated. In total, 4,314 alternative splicing (AS) events were identified and analyzed. Intriguingly, 10.4% of AS events were associated with transposable element (TE) insertions. Finally, investigation of TE exonization events and evolutionary analysis were conducted, revealing interesting phenomena of human-specific amplified trends in TE exonization events. This report represents the first large-scale transcriptome sequencing and genetic analyses of *M. fascicularis* and could contribute to its utility for biomedical research and basic biology.

TBC-4: Recent Positive Selection in Human Genes That are Enriched for Disease Mutations, but Limited for Polymorphism

Yoon-Ho Hong[1], Malcolm Campbell[2], Kyungjoon Lee[2], In-Hee Lee[2] and Sek-Won Kong[2]

[1]*Department of Neurology, Seoul National University Boramae Municipal Hospital, Korea*

[2]*Children's Hospital Informatics Program, Boston Children's Hospital, USA*

Examining the near-full spectrum of genetic variation across the whole human genome is now possible with the advances of high-throughput sequencing technology. This enables population scale analysis of sequence variations, which provides an opportunity to explore characteristics of human disease genes and mutations in the context of molecular evolution. Here, using the whole genome sequence data of 37 putatively healthy unrelated individuals, we investigated the effects of natural selection in shaping the frequency spectrum of genetic polymorphism and disease mutations. We found that a quantitative estimate of evolutionary constraints is significantly higher in genes with lower frequency of polymorphic coding variants. The correlation between polymorphism and natural selection is also supported by 1) population and comparative analyses at the gene level, which revealed a significantly greater spectrum of single nucleotide polymorphisms (SNPs) in genes under positive selection, and 2) analysis in the context of human disease and gene essentiality, which confirmed the limited spectrum of polymorphism in disease genes with greater essentiality. Interestingly, the signature of recent or ongoing positive selection was consistently found in a subset of disease genes that are limited for polymorphism but enriched for disease-linked mutations. This suggests that recent adaptive selection might have acted on evolutionarily conserved genes, increasing the spectrum of disease-linked mutations.

TBC-5: Gene expression changes as resistant markers to cisplatin in a panel of bladder cancer cell lines

Sung Han Kim[1] and Seok Soo Byun[2]

[1]*Seoul National University Hospital, Republic of Korea*

[2]*Seoul National University Bundang Hospital, Republic of Korea*

BACKGROUND: Cisplatin, one of the most effective anticancer drugs for bladder cancer, develops resistance during treatment by a cellular self-defense system of activating or silencing a variety of different genes, resulting in genetic and epigenetic alternations. As a result, the resistance mechanism of cisplatin is one of the most investigated subjects in clinical fields. In order to understand the resistance mechanism and to establish a possible gene candidate, a panel of cisplatin-resistant and general bladder cancer cell lines were used in a combination of microarray and real time-PCR profiling to investigate the possible resistant cisplatin gene expression.

METHOD: The human bladder cancer cell line (T24) obtained from the American Type Culture Collection (ATCC) and the preformed bladder cancer resistant cell line at 2.0μg/ml of cisplatin (T24R2) were used for the microarray analysis to define the different expressions of significant genes resistant to cisplatin. Those upregulated significant genes were compared to tissue assay of bladder cancer resistant to cisplatin chemotherapy by real time PCR using. A fold change≥ 2 with p-value< 0.05 of statistics was considered significant.

RESULTS: Among a list of 488 up-regulated genes and 69 pathways from microarray analysis, a panel of 23 genes was selected for real time-PCR validation from four selected cancer-related pathways (p53, apoptosis, cell cycle, and pathway in cancer). All 23 genes were determined to be significantly different and up-regulated in both the microarray and the RT-PCR with fold change >2.0. They are PRKAR2A and 2B, CYCS, Bcl-2, BIRC3, DFFB, CASP6, CDK6, CCNE1, CUL2, FN1, STEAP3, MCM7, ORC2 and 5, LEF1, ANAPC1 and 7, CDC7 and 27, SKP1, WNT5a and 5b genes. Especially, the fold changes of CUL2, MCM7, WNT5A and 5B, LEF1, Bcl-2, CYCS, and PRKAR2B were greater than 4.0, suggesting high correlation with cisplatin resistance.

CONCLUSIONS: A panel of 23 up-regulated genes including the 5 genes with greater fold changes was determined to be significantly different from cisplatin resistant bladder cancer and bladder cancer cell lines. We propose that their gene expression profiles may play one of the key roles in the resistance mechanism to cisplatin in patients with bladder cancer.

TBC-6: GlaI-qPCR assay — a new instrument for quantitative DNA methylation analysis and its application for tumor suppressor genes study

Vitaliy Kuznetsov[1], Elena Zemlyanskaya[1] and Sergey Degtyarev[1]

[1]*SibEnzyme Ltd., Novosibirsk, Russia, 630117*

De novo DNA methylation in mammals is performed by Dnmt3a and Dnmt3b DNA methyltransferases, which recognize a tetranucleotide 5'-RCGY-3' and modify the inner CG-dinucleotide with formation of 5'-R(5mC)GY-3'/3'-YG(5mC)R-5'[1].

GlaI is a novel methyl-directed site-specific DNA-

endonuclease which recognizes DNA sequence 5'-R(5mC)↓GY-3' and cleaves it as indicated by arrow [2]. Thus, the recognition sequence of GlaI exactly corresponds to a product of DNA methylation with Dnmt3a and Dnmt3b. GlaI cleaves DNA completely and requires no additional cofactors [3]. Recently we have developed GlaI-PCR assay which allows determination of 5'-R(5mC)GY-3' sites in studied DNA region [4]. The method includes DNA hydrolysis with GlaI followed by PCR with primers designed for the DNA region of interest. Earlier we have used GlaI-PCR assay to determine DNA methylation status of regulatory regions of tumor suppressor genes (TSGs) [5]. In this work we perform real time GlaI-PCR assay for quantitative determination (GlaI-qPCR) of 5'-R(5mC)GY-3' sites in studied DNA regions. This assay was applied for study of DNA methylation in regulatory regions of RARB, NOTCH1, DAPK1, SEPT9b, IGFBP3, CEBPD, MGMT and RASSF1A TSGs in malignant cell lines HeLa, Raji, U-937, Jurkat and in the control fibroblast cell line L-68. We received methylation profiles of these genes for each cell line. In correspondence with previous data regulatory regions of TSGs are methylated in malignant cell lines. However, the methylation profiles are different for each cell line. This allows differentiating between different types of cancer cells. The results show that method of GlaI-qPCR assay may be used for quantitative determination of *de novo* DNA methylation.

References

1. Handa V, and Jeltsch A. J. Mol. Biol. 2005; 348, 1103-1112.

2. Tarasova GV et al. BMC Mol. Biol. 2008; 9, 7.

3. Abdurashitov MA et al. BMC Genomics, 2009; 10, 322.

4. SE Scientific Library [http://science.sibenzyme.com/article12_article_53_1.phtml]

5. SE Scientific Library [http://science.sibenzyme.com/article8_article_58_1.phtml]

TBC-7: A Filtering Algorithm for Gene-Gene Interaction using Case-Only Data

Pin-Cian Wang[1], Liang-Chuan Lai[2], Mong-Hsun Tsai[3], Eric Y. Chuang[4], Cheng-Yan Kao[1] and Pei-Chun Chen[5]

[1]*Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, Taiwan*

[2]*Graduate Institute of Physiology, National Taiwan University, Taiwan*

[3]*Graduate Institute of Biotechnology, National Taiwan University, Taiwan*

[4]*Bioinformatics and Biostatistics Core, Research Center for Medical Excellence, National Taiwan University, Taiwan*

[5]*Department of Statistics and Informatics Science, Providence University, Taiwan*

Genome-wide association studies (GWAS) are typical study designs in genetic epidemiology using whole-genome SNP data. Single-locus test is used in most GWAS. However, some researchers have indicated the problems of GWAS using single-locus strategy. Gene-gene interaction becomes a more important issue. Exhaustive search methods such as multifactor dimensionality reduction (MDR) are powerful tools for gene-gene interaction detection. However, the main limitation of MDR is heavy computation. Therefore, the aim of our research was to design a filtering algorithm to select a candidate SNP set for further analysis and that can save computation time and get same prediction, called the deviance of independence (DOI).

DOI describes the level of dependence between two SNPs. In the first step of DOI calculation, the SNP data in control samples was removed because it was hypothesized that the frequency of allele and genotype may be stable in normal population. Next, the frequency of expected two-SNP combination and real two-SNP combination were calculated. The frequency of expected two-SNP combination was derived from the frequency of two individual SNPs according the principle of independence. Finally, DOI values were calculated by the summation of each absolute difference between the frequency of expected and real two-SNP combination. It is expected that the SNP combinations with high DOI have more potential to be the interaction combinations.

We use simulation and real data to examine DOI performance. The simulation results show that DOI values may be used to predict the interaction combinations. In addition, the WTCCC Rheumatoid arthritis (RA) chromosome 22 data and Parkinson's disease (PD) chromosome 20 data were used for real data application. And the results demonstrate that potential interactions can be identified after using DOI value as a filter criterion. In sum, DOI algorithm is a powerful tool to filter a candidate gene set for further interaction analysis.

TBC-8: 20-gene-based risk score classifier predicts disease recurrence in non-muscle invasive bladder cancer

Seon-Kyu Kim[1], Young-Kyu Park[1] and Seon-Young Kim[1]

[1]*Medical Genomics Research Center, Korea Research Institute of Bioscience and Biotechnology, Daejeon, Korea*

Background: Bladder cancer is a genetic disorder driven by the progressive accumulation of multiple genetic changes. While several molecular markers for the recurrence of bladder cancer have been studied, the limited value of current prognostic markers has created the need for new molecular indicators of bladder cancer outcomes. Here, we sought to identify a molecular signature associated with disease recurrence in non-muscle invasive bladder cancer (NMIBC) and to assess its usefulness as a prognostic indicator.

Methods: Microarray gene expression profiling was performed using gene-expression data from 102 primary NMIBC specimens (Korean cohort) to identify a gene expression signature associated with disease recurrence. The prognostic value of the gene expression signature was validated in an independent cohort (European cohort, n=302). A risk score based on the expression data of 20 genes was developed in the Korean cohort and validated in the European cohort. The association between the 20-gene-based risk scoring method and prognosis of NMIBC patients was assessed using Kaplan- Meier plot, the log-rank test, Cox proportional hazards model, and leave-one-out cross validation method.

Results: The determination of gene expression patterns by microarray data analysis identified 822 genes associated with disease recurrence. Of the 822 genes, 20 genes which are highly associated with recurrence free survival were detected by time-dependent ROC analysis. The risk score was developed by using Cox coefficient values of 20 genes in the Korean cohort and its robustness was validated in the European cohort (log-rank test, $P < 0.001$). Multivariate Cox regression analysis revealed that the risk score was an independent strong predictor of disease recurrence (hazard ratio = 6.082, 95% confidence interval = 3.280 to 11.279, $P < 0.001$).

Conclusions: The risk scoring method based on 20 genes represents a promising diagnostic tool to identify NMIBC patients that have a high risk of recurrence.

TBC-9: Genome-wide analysis of CNV and SNP in Koreans

Sanghoon Moon[1], Kwang Su Jung[2], Young Jin Kim[1], Miyeong Hwang[1], Kyungsook Han[4], Bok-Ghee Han[3], Jong-Young Lee[1], Kiejung Park[2] and Bong-Jo Kim[1]

[1]*Division of Structural and Functional Genomics, Center for Genome Science, National Institute of Health, Chungcheongbuk-do, 363-951, Korea*
[2]*Division of Bio-Medical informatics, Center for Genome Science, National Institute of Health, Chungcheongbuk-do, 363-951, Korea*
[3]*Center for Genome Science, National Institute of Health, Chungcheongbuk-do, 363-951, Korea*
[4]*School of Computer Science and Engineering, Inha University, Inchon, 402-751, Korea*

To date, single-marker association analysis in genome-wide association studies (GWAS) has identified a large number of single nucleotide polymorphisms (SNPs) that are highly associated with complex diseases, but only a small portion of genetic heritability is explained by these variants. A copy number variation (CNV) is a physical change of genomic segment ranging from a kilobase to several megabases. CNV may alter disease susceptibility and gene dosage for genetic risk, so is a useful source for finding missing heritability.

Recent studies have reported that 60% of the detected CNVs were called with a single copy-number class, which cannot be tested for association and that well-defined polymorphic CNVs tagged by SNPs are more likely to affect multiple expression traits than frequency-matched variants. CNVs encompassing single genes or a set of genes can be more causative variants of genetic disease than SNPs alone. Therefore, SNPs correlated with CNVs are a valuable resource for GWAS.

Most CNV databases (except SCAN) do not consider polymorphic CNV (multi copy-number class). SCAN database also contains CNV data of Caucasian and Yoruba populations, and does not provide Asian CNV data. Due to the difference in CNVs between distinct ethnic groups, providing polymorphic CNVs and allele frequency of each genotype in Asian populations will help investigate CNV-association with diseases and ethnic differences.

In this study we developed a database called Korean Genomic Variant Database (KGVDB), which provides polymorphic CNV regions and well-tagged SNP information. The data were obtained from 4,700 individuals using two different genotyping platforms and publicly available CNV data. The large data set of KGVDB will provide a rich public resource for the study of CNV and SNP.

TBC-10: 3D-QSAR Pharmacophore Modeling of Thromboxane A2 Receptor for Discovery New Inhibitors

Kuei-Chung Shih[1], Cheng-Yu Ma[1], Hsiao-Chieh Chi[1] and Chuan-Yi Tang[1,2]

[1]*Department of Computer Science, National Tsing Hua University, Hsinchu, Taiwan 30013, R.O.C.*
[2]*Department of Computer Science and Information Engineering, Providence University, Taichung, Taiwan 43301, R.O.C.*

Thromboxane A_2 (TXA_2) is a hormone derived from arachidonic acid (AA) through cyclooxygenases (COX) and thromboxane synthase (TXS), and it is a platelet

aggregator by activating thromboxane A $_2$ receptor (TP) to induce platelet aggregation and cell proliferation. Based on the action of platelet activation, TXA_2 is associated with thrombosis, acute myocardial infarction and many diverse inflammatory diseases. There are some different approaches to achieve antiplatelet therapy through this prostanoid pathway. One strategy is to inhibit COX so that TXA_2 could not be produce from AA, such as the most well-known antiplatelet drug, aspirin. Despite aspirin could resist myocardial infarction and stroke, it may lead to gastrointestinal disorder and allergy. TXS inhibition is one kind of inhibitors for suppressing TXS to generate TXA_2, but it does not work efficiently because other endoperoxides and isoprostanes can also active TP just like TXA_2. Accordingly, the method to directly inhibit TP seems to be attractive. However, TP antagonists include ifetroban, sulotroban, GR32191 and other antithrombotic agents still stay in phase II or III of clinical development due to the safety concerns and efficacy. The previous studies were not proposed available co-complex structure between TP and Thromboxane A_2 (TXA_2) or any of its inhibitors, it is necessary to establish a screening model for rational drug design in silico. Our research is focus on building the TP phaemacophore hypothesis for discovering other potential TP inhibitors. This study report, we developed pharmacophore hypothesis for discovery new TP inhibitors. The best hypothesis has one hydrogen-bond acceptor (A) and three hydrophobic aromatic groups (HYAR), its correlation coefficient of training set and testing set were 0.933 and 0.923, respectively. According to statistical validation and chemical features analysis, our best pharmacophore hypothesis has excellent ability to help medicinal chemists in their efforts to identify or design new TP inhibitors.

TBC-11: Comparison of somatic mutation-calling methods based on DNA sequence from matched tumor-normal pairs

Su Yeon Kim[1] and Terry Speed[1,2]

[1]*University of California at Berkeley, Berkeley 94720, USA*
[2]*Walter and Eliza Hall Institute of Medical Research, Parkville Victoria 3052, Australia*

Somatic mutation-calling based on DNA from matched tumor-normal patient samples is one of the key tasks carried by many cancer genome projects. In particular, The Cancer Genome Atlas (TCGA) is now routinely compiling catalogs of somatic mutations for hundreds of patients for various tumor types. Nonetheless, mutation calling is still a very challenging problem. TCGA benchmark studies reveal that even up-to-date mutation callers from major sequencing centers show substantial discrepancies. For most tumor types, validation data is not yet available, and even when it will be, only a fraction of all candidate mutations are likely to be validated. In order to compare mutation callers without genome-wide gold standard validation data, we have developed an approach using pseudo-positives (presumed somatic mutations) and pseudo-negatives (presumed not somatic mutations) that are defined using another caller. The other callers can be built on using publicly available variant calling methods such as GATK or SAMtools. This approach allows us to give a convenient visualization of the discrepancies between the different mutation call sets, and to summarize each mutation-caller's performance in terms of pseudo-false-positive and pseudo-false-negative rates. Some insights were gained from observing consistent results from two other callers that are not expected to introduce the same biases.

TBC-12: The estimation of heritability analyses for BMI using genotype score based on Korean Cohort

Nam Hee Kim[1], Youngdoe Kim[1], Young Jin Kim[1], Ji Hee Oh[1], Mee Hee Lee[1] and Juyoung Lee[1]

Division of Structural and Functional Genomics, Center for Genome Science, National Institute of Health, Korea Centers for Disease Control and Prevention, Korea

The aim of study was to estimate variation and their heritability for BMI including genotype score and compare BMI to other cohort. We have constructed community and twin-family based on cohort, which is an ongoing prospective studies and surveyed samples were drawn from the Korean Genome and Epidemiology Study and Korea Genome Analysis Project in Korea.

We selected 2,473 subjects in twin-family cohort and surveyed their zygosity using the self-report questionnaires about 2,000 items and genotyped using Affy 6.0. From community-based cohort(KARE; Korea Association REsource), we selected 8,842 subjects and surveyed their self-report questionnaires about 1,400 items and genotyped using Affy 5.0. Including genotype score of BMI estimated heritability for BMI using SOLAR, GCTA, GENABEL.

TBC-13: Genotype instability during long-term subculture of lymphoblastoid cell lines

Ji Hee Oh[1], Young Jin Kim[1], Sanghoon Moon[1], Jong-Young Lee[1] and Yoon Shin Cho[1,2]

[1]*Division of Structural and Functional Genomics, Center for Genome Science, National Institute of Health, Chungcheongbuk-do 363-951, Republic of Korea*

[2]*Department of Biomedical Science, Hallym University, 1 Hallymdaehak-gil, Chuncheon, Gangwon-do 200-702, Republic of Korea*

Epstein-Barr virus (EBV-transformed lymphoblastoid cell lines (LCLs) promise to address the challenge posed by the limited availability of primary cells needed as a source of genomic DNA for genetic studies. However, the genetic stability of LCLs following prolonged culture has never been rigorously investigated. To evaluate genotypic errors caused by EBV integration into human chromosomes, we isolated genomic DNA from human peripheral blood mononuclear cells and LCLs collected from 20 individuals and genotyped the DNA samples using the Affymetrix 500K SNP array set. Genotype concordance measurements between two sources of DNA from the same individual indicated that genotypic discordance is negligible in early-passage LCLs (less than 41 passages) but substantial in late-passage LCLs (more than 40 passages). Analysis of concordance on a chromosome-by-chromosome basis identified genomic regions with a high frequency of genotypic errors resulting from the loss of heterozygosity observed in late-passage LCLs. Our findings suggest that, whereas LCLs harvested during early stages of propagation are a reliable source of genomic DNA for genetic studies, investigations that involve genotyping of the entire genome should not use DNA from late-passage LCLs.

TBC-14: Multi-study integration of brain cancer transcriptomes reveals organ-level diagnostic signatures

Jaeyun Sung[1], Pan-Jun Kim[1], Leroy Hood[2], Donald Geman[3] and Nathan Price[2]

[1]*Asia Pacific Center for Theoretical Physics, Korea*

[2]*United States Institute for Systems Biology, USA*

[3]*Institute for Computational Medicine, Department of Applied Mathematics and Statistics, Johns Hopkins University, USA*

The identification of molecular signatures from either tissues or blood to accurately reflect the major cancers of an organ system would be a significant advance in molecular cancer diagnostics. Towards this goal, we identified comprehensive diagnostic signatures of major cancers of the human brain from a multi-study, integrated transcriptomic dataset. These signatures are based on comparing ranked expression values of gene-pair sets, which are aggregated into a brain cancer marker-panel of 44 unique genes. Many of these genes have established relevance to the brain cancers tested herein, with others having known roles in cancer biology. Phenotype prediction follows a diagnostic hierarchy, and the corresponding hierarchically-structured signatures achieved 90% classification accuracy against a multi-disease alternative hypothesis when training and validation sets were drawn from the same population distribution (cross validation). Despite accurately distinguishing among phenotypes in single-population cross-validation, diagnostic signatures must remain robust even across more heterogeneous populations to justify their broad clinical use. To address this issue, we found that sufficient dataset integration across multiple studies greatly enhanced reproducibility and accuracy in diagnostic performance on truly independent validation sets, whereas signatures learned from one dataset typically had high error on independent validation sets. Looking forward, we discuss our approach in the context of improving blood diagnostics for cancers of organ systems.

TBC-15: Methyl-directed Site-specific DNA Endonuclease MteI is a New Instrument for Analysis of CpG Island Methylation

Vasilina A. Sokolova[1], Valery A. Chernukhin[1], Danila A. Gonchar[1], Elena V. Kileva[1], Larisa N. Golikova[1], Vladimir S. Dedkov[1], Natalya A. Mikhnenkova[1], Elena V. Zemlyanskaya[1], Vitaliy V. Kuznetsov[1] and Sergey Kh. Degtyarev[1]

[1]*SibEnzyme Ltd., Novosibirsk, Russia 630117*

Methyl-directed (MD) DNA endonucleases specifically cleave short methylated DNA sequences and don't cut unmethylated DNA. Biochemical properties of MD endonucleases are similar to those of restriction enzymes, both types of enzymes require only Mg^{2+} ions as a cofactor. Today more than ten MD DNA endonucleases recognizing different sites with 5-methylcytosine are discovered and characterized [1]. Among them MD DNA endonucleases BlsI, BisI, PkrI and Glul have the same recognition site 5'-GCNGC-3', but activity of these enzymes depends on the amount and position of 5-methylcytosines in the recognition sequence.

A new methyl-directed site-specific DNA endonuclease MteI was isolated from Microbacterium testaceum. MteI recognizes a prolonged methylated DNA sequence of nine bases in length with a central pentanucleotide 5'-GCNGC-3'. MteI activity depends on a number of 5-methylcytosines and their positions in the recognition site. MteI cleaves DNA sequence 5'-G(5mC)G(5mC)^NG(5mC)GC-3'/3'-CG(5mC)GN^(5mC)G(5mC)G-5' as indicated by arrows. The enzyme activity is significantly higher if 5'-GC-3' dinucleotides in this site are replaced by 5'-G(5mC)-3' dinucleotides and additional 5'-G(5mC)-3' dinucleotides are present in both DNA strands.

We have developed a method of MteI-PCR assay which allows determining the methylated CpG islands. The method includes DNA hydrolysis with MteI followed by

PCR with primers designed for the DNA region of interest. MteI-PCR assay has been applied to study methylation of CpG islands located in regulatory regions of tumor suppressor genes and revealed different patterns of DNA methylation.

1. http://mebase.sibenzyme.com/md-endonucleases

TBC-16: Nonunique SNP problems in association study

Lyong Heo[1], Young Jin Kim[1], Sanghoon Moon[1] and Jong-Young Lee[1]

[1]*Division of Structural and Functional Genomics, Center for Genome Science, National Institute of Health, Osong, Korea*

In the recent years, genome-wide association study (GWAS) have successfully identified numerous phenotype associated SNPs. In GWAS, SNP is served as a marker indicating a specific genomic region. Chromosomal position of each SNP is well annotated in NCBI dbSNP database. In dbSNP, however, annotation errors have been reported such as a SNP with multiple position, position change, and chromosome change. Doron and Sheweiki reported that 4.2~11.9% of HapMap SNPs were mapped to nonunique genomic region. Since a marker is only valid if it maps to unique region, SNPs mapped at nonunique region would not be adequate for association analysis. In this study, we analyzed nonunique SNPs in two versions of dbSNP database, b130 (hg18) and b135 (hg19). Nonunique rsIDs account for 3.46% and 2.26% of b130 and b135, respectively. Also, position change due to dbSNP build update was 0.39% for b130 and 0.13% for b135. We inquired GWAS catalog for studying the effect of nonunique SNPs. As of August 2012, GWAS catalog included 1355 publications with 8754 SNPs (7131 unique SNPs). Among catalogued SNPs, we found 237 SNPs mapped at nonunique position. Our results indicate that SNPs should be carefully annotated and tested for its validity as a marker in association study.

TBC-P17: Exonic variants in Korean population

Young Jin Kim[1], Kwang Joong Kim[1], Lyong Heo[1], Yun Kyoung Kim[1], Sanghoon Moon[1], Youngdoe Kim[1], Mi Yeong Hwang[1], Bong-Jo Kim[1] and Jong-Young Lee[1]

[1]*Division of Structural and Functional Genomics, Center for Genome Science, KNIH, KCDC*

Recent advancement of high-throughput genotyping technologies has enabled us to carry out a genome-wide association study (GWAS) in a large cohort. The main goal of genome-wide association study is to identify the complex phenotype associated loci. The discovery of the associated loci would lead us to understand the underlying mechanisms of complex traits. Despite the great success of GWAS, however, a limited number of susceptibility variants discovered in the previous GWAS accounts for only a small proportion of phenotypic variance. Missing heritability of the current genome analysis is the bottleneck preventing us from taking a step forward to personal genome, personal medication, disease prediction and prevention. In this context, Next Generation Sequencing (NGS) technology has been gathered much attention due to its usability in accessing genomic data at the base pair level of resolution. In this context, exome sequencing comprising 400 Korean samples facilitated the assessment of full spectrum of allele frequencies including coding altering variants. The analyses of all variants within coding regions would reveal undiscovered possible causal common or rare variants near previously associated loci.

TBC-18: Development of Korea Common Data Model for Adverse Drug Signal Detection based on multi-center EMR systems

Si Ra Kim[1], Seung Ho Park[2], Bum Joon Park[2], Kwang Soo Jang[2] and In Young Choi[1]

[1]*Graduate School of Healthcare Management and Policy, The Catholic University of Korea, Seoul 137701 , Korea*
[2]*Master course of engineering, Hanyang University of Korea, Seoul 133791, Korea*

The adverse drug reaction (ADR) research based on Clinical Data Warehouse(CDW) was getting important in accordance with the electronic clinical information like Electronic Medical Record (EMR) than spontaneous adverse drug reaction (ADR) reporting. The drug safety monitoring based on EMR is able to collect more objective pharmacovigilance and analyze ADR earlier than spontaneous adverse drug reaction (ADR) reporting. We analyzed drug safety surveillance model with three researches; EU-ADR data model of Europe, Mini-Sentinel data model of Food and Drug Administration (FDA) and Observational Medical Outcomes Partnership (OMOP) data model of National Institutes of Health (NIH). Based on the comparison of three data models, we developed the Korea ADR common data model (CDM) for early detection of adverse drug reaction in Korea. This project is called as K-ADR (Korea- Adverse Drug Reaction). The K-ADR consists of eight tables which contain demographic table, drug table, visit table, procedure table, diagnosis table, death table, laboratory table and report-machinery table. Each table consists of 5~12 fields. In addition, terminology standard such as

ICD-10 and WHO-ART will be provided to integrate multiple EMR systems. The K-ADR reflected Korea EMR structures will contributes for pharmacovigilance activity. The pharmacovigilance activity by using EMR is able to accurate signal detection through the diagnosis name and drug prescription information by patient. Also the K-ADR could be detected adverse drug events (ADEs) that contain under-reported ADEs and deficient ADEs. Further efforts for development of the standardized guidelines about procedure code and laboratory code will be needed for multi-institutional pharmacovigilance database system. The pharmacovigilance activity based EMR will be cost-effective method to detect ADR signals.

Acknowledgement: This research was supported by a grant(12172KFDA212) from Korea Food and Drug Administration in 2012.

TBC-19: Various nucleosome positioning patterns in Drosophila

Doo Yang[1,2] and Ilya Ioshikhes[1,2]

[1]*Ottawa Institute of Systems Biology, Canada*

[2]*Department of Biochemistry, Microbiology & Immunology University of Ottawa, Canada*

Nucleosome plays an important role in gene regulation by affecting the accessibility of transcription factors to the DNA. DNA sequence is one of the factors that position nucleosomes.

Finding the nucleosome positioning sequence (NPS) is challenging because the nucleosome binding is not as specific as transcription factor motifs However, some sequence features , such as dinucleotide periodicity, can be observed by analyzing nucleosome sequences collectively.

Drosophila genome sequences of H2A and H2A.Z nucleosomes were analyzed to find a novel NPS and relationship with biological functions.

The nucleosome positions and sequences were obtained from the published Chip-Seq data (Mavrich, el al., 2008, Nature for H2A.Z and Henikoff, et al., 2011, Genes & Devlop.) In order to minimize the noise in sequence pattern, only the +1 nucleosomes sequences were selected and separated into H2A and H2A.Z sequences. Then the dinucleotide patterns were analyzed.

Two novel NPS patterns, WW/SS and RR/YY, are proposed. The WW/SS sequence pattern is similar but not identical to the previously proposed yeast NPS. The *Drosophila* WW/SS NPS has higher content of SS at dyad. The 10 bp periodicity is stronger off the dyad and disrupted near dyad. The RR/YY NPS shows that dinucleotides are more periodic between 25 to 45 bp from dyad than near dyad or outer region. GO analysis of the genes having either WW/SS or RR/YY nucleosomes showed differences in biological functions. It suggested that possible relationship between gene functions and nucleosome sequences.

Comparison of H2A and H2A.Z NPS showed differences in the dinucleotide pattern. The most significant difference is that H2A.Z NPS has stronger peaks at the ± 45 bp from dyad instead of ± 55 bp in H2A. These positions in DNA are close to the protein domain where H2A.Z and H2A histones are different. In yeast, H2A.Z positioning is dependent on SWR1 and is immobile once positioned. H2A.Z is also well phased at the down stream of TSS. Combined with the fact that H2A.Z plays a role in proper gene activation, H2A.Z may serve as a barrier of downstream nucleosomes to maintain the proper binding sites for transcription factors and other proteins.

TBC-20: Anonymized Patient Chart Review Tool in Asan Medical Center

Soo-Yong Shin[1,2], Yongdon Shin[2], Yong-Man Lyu[2], Hyo Joung Choi[2], Jihyun Park[2] and Jaeho Lee[1,2,3]

[1]*Department of Biomedical Informatics, Asan Medical Center, Seoul 138-736, Korea*
[2]*Office of Clinical Research Information, Asan Medical Center, Seoul 138-736, Korea*
[3]*Department of Emergency Medicine, Asan Medical Center, University of Ulsan College of Medicine, Seoul 138-736, Korea*

Asan Medical Center (AMC) has been developing AMC biomedical research infrastructure to improve the efficiency of clinical research as well as to protect privacy of patients. As a first step, AMC developed the anonymized patient chart review tool to protect patients' privacy by complying with government regulations in Korea. The primary purpose of this tool is to decide if a chosen patient should be included or excluded for a proposed study by reviewing the patient's anonymized clinical data. For this purpose, the AMC anonymized patient chart review tool aims to provide the comprehensive clinical data in AMC data warehouse including diagnosis, medication, lab results, pathology/radiology reports, progress notes, admission note, discharge summary, and operative report. Also it tries to provide the easy user interface by implementing the same interface as other AMC medical information systems. To generate the anonymized clinical data, 18 identifiers defined by HIPAA were removed as follows: 1) each patient was assigned to new research ID which is different from hospital patient ID. 2) All structured identifiers stored in EMR database were removed. 3) The remaining identifiers in the narrative texts were masked using the pre-defined regular expressions. As a future work, we have plans to scramble the date in clinical data

and develop one-time research ID method which can generate a different ID each time even for the same patient for stronger protection of patients' privacy. We are also developing a research cohort discovery tool to estimate the approximate number of patients satisfying the research criteria.

TBC-21: Integrate Genomics and Molecular Interactome Data for Brain Tumor Pathway Discovery and Prognosis

Jongkwang Kim[1], Gao Long[1] and Kai Tan[1]

[1]*University of Iowa, Dept. of Internal Medicine, Dept. of Biomedical Engineering, 65536 Iowa city, USA*

Glioblastoma (GBM: grade IV astrocytoma) is the most common and lethal form of brain cancer. Median patient survival time is 15 mo. Few predictive gene markers for prognosis and treatment. This study integrates three types of data: transcriptomic, epigenomic profiles, and protein-protein interactome to find pathway markers that are responsible for long-term survival (LTS) compared to short-term survival (STS). 13 pathway markers were found from the integrated analysis. Pathway markers were tested on 115 GBM patient samples for the classification accuracy into STS and LTS cases. The accuracy (82.2%) is 13.6% higher than using one or two types of data, demonstrating that integration of transcriptomic, epigenomic and interactome data is a more powerful approach to elucidating molecular pathways distinguishing GBM subtypes.

TBC-22: Development of a Consumer-engaged Obesity Management Ontology based on Nursing Process

Hyun-Young Kim[1], Hyeoun-Ae Park[2], Yul Ha Min[2] and Eun-Joo Jeon[2]

[1]*Eulji University, College of Nursing, Deajeon 301-832, Korea*
[2]*Seoul National University, Seoul 110-799, Korea*

The purpose of this study is to develop an ontology to represent the consumer-engaged obesity management process based on clinical practice guidelines. Since life style modification by the consumers is the most important aspect in obesity management, we introduced concepts of consumer's engagement into obesity management process. We also considered data traffic when we developed the ontology.

We developed the ontology by defining the scope of obesity management, selecting a foundational ontology, extracting the concepts, assigning relations among classes, and representing classes and relations with Protégé.

We identified behavioural intervention, dietary advice, and physical activity from the guideline as obesity management strategies. Nursing process was selected as a foundational ontology to represent consumer's engagement in obesity management process. Since, consumers engage in their obesity management when they identify expected. Nursing process is a patient-centered, and goal-oriented method consisting of five phases (assessment, nursing diagnosis, outcome identification, implementation, and evaluation). These phases are repetitive and cyclic in obesity management process. First cycle represents first encounter of obesity management from initial assessment to outcome identification. Second cycle represents second encounter and onward. Two cycles are connected through the assessment in the second cycle being the evaluation of the first cycle. With this approach we were able to minimize data traffic in the obesity management process. We extracted 127 concepts, which included assessment data (such as sex, body mass index, and waist circumference) and the inferred data to represent nursing diagnosis and evaluation (such as degree of and reason for obesity and success or failure in life style modification). Relations linking concepts are "part of", "instance of", "derives from", "derives into", "has plan", "followed by", and "has intention". The concepts and relations were formally represented using the Protégé.

We were able to represent obesity management with consumer's engagement using nursing process as a foundational ontology. Nursing process can be used as a foundational ontology to support development of ontologies representing consumer's behavioural modification.

Acknowledgements: This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (no.2012-012257 and no. 2012- 0000998).

TBC-23: Performance of microRNA target prediction algorithms

Jee Yeon Heo[1], Yongjin Choi[1], Hae-Seok Eo[1], Youngho Kim[1], Taesung Park[2] and Hyung-Seok Choi[1]

[1]*Bio&Health Team, Future IT R&D Laboratory, LGE Advanced Research Institute, Seocho-gu, Seoul 137-724, Korea*
[2]*Department of Statistics, Seoul National University, Gwanak-gu, Seoul 151-747, Korea*

MicroRNAs (miRNAs) are a class of small non-coding RNAs (~22 nt), which regulate gene expression through suppressing mRNA translation or inducing mRNA degradation by binding to their target mRNAs in multiple

biological processes such as cell cycle control, cell growth, cell differentiation, apoptosis, embryo development and so on. Many computational and bioinformatic approaches to predicting target mRNAs of each miRNA have been developed including miRanda, PITA, TargetScan, DIANA-microT, Microcosm and miRDB. Here, we compared the performances of these six above-mentioned miRNA target prediction algorithms. First, 6,901 common pairs (0.003%) were selected from the total 2,842,985 miRNA-target mRNA pairs predicted by all six algorithms. Second, 3,507 validated miRNA-target mRNA pairs were collected from the experimentally validated databases including TarBase, miR2Disease, miRTarBase and miRecords. Among them, 879 pairs (25%) were not predicted by any algorithm and 214 pairs (6%) were predicted by all six algorithms. Finally, Receiver operating characteristic (ROC) curves and area under curve (AUC) values were calculated to compare of the performance of each algorithm. Our comparison results show that DIANA-microT has the highest accuracy (60%) and miRanda has the lowest accuracy (49%) and prediction scores of each miRNA target prediction algorithm are lowly correlated to each other.

TBC-24: Graphical modeling of regulatory interactions in sporadic Inclusion Body Myositis

Thomas Thorne[1], Pietro Fratta[2], Michael Hanna[3], Elizabeth Fisher[2] and Michael Stumpf[1]

[1]*Centre for Bioinformatics and Systems Biology Imperial College London, UK*
[2]*Department of Neurodegenerative Disease, UCL Institute of Neurology, UK*
[3]*National Hospital for Neurology & Neurosurgery, University College London, UK*

Sporadic Inclusion Body Myositis (sIBM) is a disease that causes inflammation of the muscles and progressive weakening and wasting of the muscles, and the mechanisms by which it acts are not currently fully understood. Here we present an analysis of gene expression microarray data from both disease and control cases in an attempt to identify regulatory interactions that may be involved in the disease. To model the regulatory network structure we employ a Gaussian Graphical Model (GGM) formalism, whereby the data are assumed to be generated from a multivariate Normal distribution. In the GGM model a pair of genes will only share an edge if they have a non-zero partial correlation – that is if their correlation cannot be explained by the expression of any of the other genes. Since we are faced with a situation in which there are a significantly larger number of genes than data points, we apply a sparse regression methodology to infer the partial correlations between genes. Here we choose to apply a sparse Bayesian regression method that has been demonstrated to outperform methods such as the Lasso. To perform inference of the model parameters we apply variational inference, a technique whereby the Bayesian posterior distribution is approximated by a factorised set of exponential family distributions.

TBC-25: Jiffynet: A web server generating Gene networks for newly sequenced species

Eiru Kim[1] and Insuk Lee[1]

[1]*Biotechnology Department of Biotechnology, College of Life Science and Biotechnology, Yonsei University, Korea*

Current one of the emerging approaches in studying biological systems is systems biology which is a study field that focuses on complex interactions in biological systems. Since development of next generation sequencing technology, large amounts of sequencing data as diverse species are now available. However, lacking of their genetic analysis, It is no possible to study them systematic approaches. For a biologist who wants to study novel species systematically, we have developed a web server providing draft models of various networks. The draft net, we call this "JiffyNet", which is made from mapping associalogs with well-established existing network such as HumaNet, WormNet, YeastNet, and RiceNet. Associalogs are derived from combining orthologs of two species and their interaction. Through this it is possible to make JiffyNet of user defined species by finding associalog and mapping to base networks. We are making the webserver that enables biologist to build their own JiffyNet. A biologist may upload their sequencing data, the server sends JiffyNet created using the data through e-mail.

TBC-26: Studying Plant Complex Traits Through Network-assisted Systems Genetics of *Arabidopsis Thaliana*

Tak Lee[1], Jung Eun Shim[1] and Insuk Lee[1]

[1]*Department of Biotechnology, College of Life Science and Biotechnology, Yonsei University, 262 Seongsanno, Seodaemun-Gu, Seoul, 120-749, Korea*

As next generation sequencing (NGS) technology develops rapidly, Genome Wide Association Study (GWAS) is being highlighted for searching genes that are associated with certain traits such as disease genes in humans and stress resistant genes in plants By sequencing genomes of organisms and statistically associating sequence variants to certain traits, GWAS is expected to show high performance on the discovery of novel genes.

However, even though GWAS has high cost and requires intensive work, it does not give expected outcomes so far. Here, we present a novel way of analyzing associations between genetic variants and phenotypes of a plant model organism, *Arabidopsis thaliana*, by using a Network guided approach.

Using the *Arabidopsis* functional gene network (AraNet), we develop a unique algorithm that would effectively predict the significant variant-phenotype associations of *Arabidopsis* GWAS. AraNet is constructed by integrating various omics data and predicts functional relationships for 73% of total *Arabidopsis* genome. An algorithm that combines GWAS data and integrated omics data of AraNet, would give more power in predicting genes that have low significance in GWAS but still important in certain phenotypes

TBC-27: Systematic analysis of cell line data for the development of novel cancer treatment

Nayoung Kim[1] and Sukjoon Yoon[1]

[1]*Department of Biological Sciences, Sookmyung Women's University, Seoul 140742, Korea*

An integrative approach of large-scale omics and drug response data on various cell lines enables us to identify the cellular signaling and drug sensitivity in cancer. Here we represent system-level analysis of cell line data for predicting sensitivity and mechanism of targeted drug response based on major genotypes of cancers. Association study with the genotypic classification was performed on drug data and omics data such as transcriptome, proteome, and phosphateome on human cancer cell lines. This approach reproduced the known patterns of mechanism-based drug response in cancers. Furthermore, gene and protein signatures significantly associated with genotype were identified and integrated to drug-centered network. This study provides an integrated approach for omics, drug response data, and cancer mutation types in cancers. Our platform is applicable to generate an accelerated hypothesis and validate the optimized therapeutic window for single or combined anticancer agents.

TBC-28: Genome Signature Image (GSI): Concise visualization of species/strain-specific profiles of repetitive element occurrences for cataloguing and evolutionary studies

Kang-Hoon Lee[1], Kyung-Seop Shin[2], Woo-Chan Kim[2], Jeongkyu Roh[2], Seung-Ho Choi[2], Dong-Ho Cho[2] and Kiho Cho[1]

[1]*Department of Surgery, University of California, Davis and Shriners Hospitals for Children Northern California, USA*

[2]*Division of Electrical Engineering, School of Electrical Engineering and Computer Science, Korea Advanced Institute of Science and Technology, Korea*

The genomes of living organisms, ranging from bacteria to humans, contain diverse populations of repetitive elements (REs). Our recent studies revealed that the RE profile, including RE arrays, of the human genome is unique in comparison to the mouse genome while gene sequences of humans and mice share a homology of ~90%. Also, a preliminary survey of the genomes of various other species demonstrated that genomic RE profiles are species-specific. In this study, we developed a suite of protocols/programs to concisely visualize genome signatures using species/strain-specific RE profiles. Since the genomes of higher eukaryotes, including humans and non-human primates, have not yet been fully decoded, we developed the genome signature technology using complete genome sequences from the domains of Archaea and Bacteria. The genome sequences of 117 Archaea-domain and 1,068 Bacteria-domain members were obtained from the National Center for Biotechnology Information and subjected to a genome-wide survey for the occurrence of 5-nucleotide REs. The top 50 highest frequency REs were then selected from each genome followed by an assembly of the 50 different REs into a RE string of 250 nucleotides, from high to low frequency. The string of high frequency REs now represents a unique signature of each genome. Of note, the two key parameters (number of high frequency REs and RE length) for the generation of genome signature sequences are tuneable. The genome signature sequence was then visualized into an image, named Genome Signature Image (GSI), using a CMYK color scheme. Interestingly, not all members within a pre-established phylogenetic branch shared similar CMYK color patterns and it can be confirmed by examination of the GSIs of the 1,185 microorganisms using different parameters. The tuneable GSIs represent and visualize unique characteristics of any genome and the concise RE string of each genome enables phylogenetic studies involving large sample numbers.

TBC-29: Analysis of copy number variation in exome sequencing data

Mi Yeong Hwang[1], Sanghoon Moon[1], Young Jin Kim[1], Lyong Heo[1], Yun Kyoung Kim[1],Youngdoe Kim[1], Bok-Ghee Han[2], Jong-Young Lee[1], and Bong-Jo Kim[1]

[1]*Division of Structural and Functional Genomics, Center for Genome Science, National Institute of Health, Chungcheongbuk-do, 363-951, Korea*

[2]*Center for Genome Science, National Institute of Health,*

Chungcheongbuk-do, 363-951, Korea

Copy number variation (CNV) has been reported lots of associations with complex diseases such as schizophrenia and obesity. To discover CNVs in the human genome, comparative genome hybridization array (aCGH) and single nucleotide polymorphism (SNP) array have been mainly used. However, CNVs from these array-based platforms have inaccurate breakpoints due to low resolution. Therefore, it is hard to discover exact size of CNV regions. Moreover, small size genomic variants such as less than 500 bp were also rarely detected. Recently, next generation sequencing (NGS) techniques have developed rapidly. In addition, exome sequencing approaches has been regarded as a tool for Mendelian disease gene discovery.

In this study, randomly selected 139 individuals enrolled from population-based cohort were genotyped with Agilent/Hiseq exome sequencing. Much of the detected CNV regions were validated by Agilent 60K aCGH. As a result, we discovered 10,084 from exome sequencing. More than 80% CNVs detected from exome sequencing (8,113/10,084) was less than 300 bp in length. We compared all of the detected CNV regions with previously reported regions and also examined recurrent copy-number deletion regions that might cause loss-of-function.

TBC-30: Identification of functional nucleotide sequence variant in the promoter of CEBPE gene

Hyunju Ryoo[1], Minyoung Kong[1], Younyoung Kim[1] and Chaeyoung Lee[1]

[1]*School of Systems Biomedical Science, Soongsil University, Seoul, Korea*

Research efforts have been made to identify genetic factors for susceptibility to complex acute lymphoblastic leukemia (ALL). ALL has been known as the most common childhood malignancy. Especially, a recent outstanding genomewide association study (GWAS) revealed an association (odds ratio = 1.34, $P = 2.88 * 10^{-7}$) of ALL with the SNP of rs2239633 in a 5'upstream region of the gene encoding CCAAT/enhancer binding protein epsilon (CEBPE) in an English population (907 cases and 2,398 controls). The current study examined promoter activity in the promoter region to see if sequence variants can regulate the expression of the gene and to identify functional variant(s). Three haplotypes were estimated with the rs2239633 and its proximity single nucleotide polymorphisms (SNPs) in strong linkage. The wild haplotype was TGTTTTC (HT1) and second most consisted of the entirely opposite alleles to the wild haplotype (CCACGCT, HT2). Minigene constructs with the haplotypes were utilized to see the luciferase activity. Their luciferase activity revealed the strongest expression with the HT2 and the weakest with the HT1. Further luciferase activity showed that rs2239632 was the functional nucleotide variant which had made the different expression. The promoter activity concurred with our in silico analysis where different transcription factors were predicted with the haplotypes. We concluded that rs2239632 could regulate the expression of the CEBPE gene. This might result in the association in the previous GWAS with the rs2239633 which was strongly linked to the rs2239632 (r^2=0.949). Its risk allele would increase the gene product and lead to leukemogenesis. As a result, person with the allele or the corresponding haplotype would be more susceptible to ALL.

TBC-31: Functional promoter nucleotide variants and their haplotypes of the gene encoding CCL21

Wonhee Jang[1], Hyunju Ryoo[1], Jihye Ryu[1], Jeyoung Woo[1], Minyoung Kong[1], Younyoung Kim[1] and Chaeyoung Lee[1]

[1]*School of Systems Biomedical Sciences, Soongsil University, Seoul, Korea*

Genetic architecture for rheumatoid arthritis (RA) has been quite limitedly known in spite of a great concern on its causal factors. Recent genomewide association studies (GWAS), however, have identified several genetic signals associated with susceptibility to RA. Especially, a meta-analysis of previously published GWAS showed an association ($P = 2.8 \times 10^{-7}$, OR=1.12) with the gene encoding chemokine (C-C motif) ligand 21 (CCL21) using a total of 3,393 cases and 12,462 controls. The sequence variant (rs2812378) identified in the meta-analysis was located in a 5'upstream region of the gene. The current study aimed to identify functional variants in the promoter region in which the association signal was observed. Four nucleotide variants in an estimated linkage disequilibrium block were considered as candidate functional variants. Different transcription factors were predicted by allelic substitutions at all of the variants. Luciferase assay revealed that the minigene construct with wild haplotype (TCGG) had a smaller expression level than that with the haplotype of CCTG which included risk allele of rs2812378 identified in the meta-analysis. We concluded that the haplotype CCTG and the allele C of rs2812378 could overproduce CCL21 comparing to their corresponding wild types. The overexpression of the chemokine would lead to a larger susceptibility to RA considering that the chemokine was involved in ectopic lymphoid structures affected by RA.

TBC-32: Development of Web-based Case Report System in Traditional Korean Medicine for Clinic Doctor

Boyoung Kim[1], Seung-Min Baek[1] and Sunmi Choi[1]

[1]*Korea Institute of Oriental Medicine, Daejeon 305811, Korea*

The paper develops a web-based case report system for Traditional Korean medicine to be provided to Oriental Medicine doctors in local clinics. First of all, we arranges literatures of case report, which are gathering existing papers of case report, based on the STRICTA, and provide them as educational materials. Additionally various types of case report should be standardized to be accessible by web based system. Finally, we can prepare the foundation to practice evidence-based Medicine in Traditional Korean Medicine through the purposed system.

TBC-33: ChemTools : Python based Chemoinformatics Toolkit

Jehoon Jun[1], Minjae Yoo[1] and Kwang-Hwi Cho[1]

[1]*Soongsil University, Korea*

Python based Chemoinformatics Toolkit (ChemTools) has been developed. The development of NMR and X-ray equipment led to the discovery of numerous chemical compound structures. And these chemical structure databases led to *in silico* drug discovery using computers. Among many *in silico* methods, virtual screening is an essential tool which is widely used in most of the pharmaceutical companies and related academic fields. In these drug discovery processes, computational tools for managing, mining, and collecting database are very important. However, accuracy and performance of some of public available tools has limited ability. For this reason, we have developed an chemoinformatics toolkit which include several in- and out-house codes. ChemTools contains modules, such as yaChI(Chemical line notation) , 3DG (3D structure generator from connectivity), conformer generator and filters for eliminating unwanted data from large chemical database, which are useful to treat large chemical database. And, ChemTools can edit molecule atom-by-atom and bond-by-bond using very simple syntax. ChemTools is based on python, so the modules could be combined with any combinations in python script language. The toolkits inherit some modules from Pybel such as SIMLES code generator, InChI code generator, and Energy minimizer. The modules we developed such as yaChI and 3DG are more reliable than any other modules have been released. The performance of in-house codes are presented with their counterparts and shows improved performance. ChemTools would be are very useful tools for researches which treat large chemical database such *in silico* drug discovery or material design.

TBC- 34: Molecular Dynamic Studies to predicted protein-protein interactions using GPU accelerated AMBER : application to TBC1 interacting Rab family proteins

Ok Sung Jung[1], Bong Hun Ji[1] and Kwang-Hwi Cho[1]

[1]*Soongsil University, Korea*

Current advances in computer simulation enable us to perform large scale molecular simulation relatively easily. Especially GPU accelerated AMBER package (AMBER-GPU) shows improved performance, in terms of speed, compared to CPU version. AMBER-GPU has been applied to study TBC1 interacting Rab family proteins. As TBC family proteins function GTPase-activating protein for Rab family proteins, TBC family proteins are considered to have important roles in cell cycle and differentiation in various tissues. And, Rab family proteins are known to be participated in protein transport, membrane traffic, exocytosis, endosomal recycling by taking part in transport from endoplasmic reticulum to Golgi complex. Therefore, knowing the interaction of TBC family proteins with Rab family proteins is very essential for studying transport system.

However, it is time-consuming and expensive to study the interactions between various TBC family and Rab family. So, it is necessary to apply a computational approach to predict the interaction of the complexes prior to the in vitro experiments.

TBC1D4 (also known as AS160) and TBC1D1, are the two RabGAPs integral for the GLUT4 translocations in adipocytes and skeletal myocytes respectively, whose crystal structure have been recently reported (PDBID:3QYE). There are about 60 Rab family proteins and 18 out of them are experimentally treated to investigate the association with GLUT4 vesicles. Among them only a few (four) Rabs have been shown to be potential substrates for TBC1D1 or TBC1D4. Recently, the structures of TBC1D1 and Rab family proteins have been reported and more is coming. Using the structures the experimental result of protein-protein interaction between TBC1 and Rab family proteins are validated with computational method using AMBER. A certain energy cut has been found between binders and non-binders. We are expanding our work to the Rab family proteins which any experiments are not done yet to find possible interacting partners.

TBC-35: A Novel Data Mining Approach for Inferring Phenotypic Association Networks to Discover the Pleiotropic Effects

Sung Hee Park[1] and Sangsoo Kim[1]

[1]*School of Systems Medical Science, Soongsil University, Seoul, Korea*

Pleiotropy is a genetic phenomenon that a single gene has effects on multiple phenotypes. In the human diseases and model organisms, the pleiotropy can imply that different mutations in the same gene cause different pathological effects. Examples of pleiotropic effects have been observed more with an increasing number of variants identified through genome-wide association studies (GWAS). However, current GWAS are performed in a single trait framework without considering genetic correlations between important disease traits. Hence, the general framework of GWAS has limitations in discovering genetic risk factors affecting pleiotropic genes.

This work reports a novel data mining approach to discover patterns of multiple phenotypic associations over 52 anthropometric and biochemical traits in KARE and to infer the phenotypic association networks from the patterns expressed as association rules. This method applied to the GWAS for multivariate phenotype highLDLhighTG derived from the predicted patterns of the phenotypic networks associated with high levels of triglycerides. The patterns of the phenotypic association networks were informative to draw relations between plasma lipid levels with bone mineral density and a cluster of common traits (Obesity, hypertension, insulin resistance) related to Metabolic Syndrome (MS). The 15 variants of six genes (PAK7, C20orf103, NRIP1, BCL2, TRPM3, and NAV1) were identified for significant associations with highLDLhighTG.

Our results suggest that the six pleiotropic genes may play important roles in the pleiotropic effects on lipid metabolism and the MS, which increase the risk of Type 2 Diabetes and cardiovascular disease by analysis of Mouse QTL and PPI interaction Network on top of phenotypic associations discovered. This work provides insights into explaining disease comorbidity when the pleiotropic genes share common etiological pathways.

TBC-36: Transcription Interference Networks are the coordinators of the gene expressions

Zsolt Boldogkoi[1] and Dora Tombacz[1]

[1]*Department of Medical Biology, Faculty of Medicine, University of Szeged, Szeged 6720, Hungary*

Gene expression is mainly controlled at the level of transcription. Non-coding RNAs play very important roles in this process at various levels of genetic regulation, including the control of chromatin organization, transcription, various post-transcriptional processes, and translation. In this study, we report the detection of a genome-wide expression of antisense non-coding RNAs from the genome of pseudorabies virus, which is a neurotropic-herpesvirus. We put forward the Transcription Interference Network (TIN) hypothesis in an attempt to explain the genomic design and the existence of the antisense RNAs in a common interpretation framework. This hypothesis suggests the existence of a novel genetic regulatory layer, which controls the cascade of herpesvirus gene expression at the level of the transcription. The TIN is proposed to represent a mechanism, which plays a central role in the programmed step-by-step switches of transcription between kinetic classes and subclasses of viral genes. The proposed model may be not restricted to the herpesviruses, but might explain the mechanism of an important regulatory system existing in other organisms belonging to various taxonomic classes.

This project is supported by the Swiss Hungarian Contribution and the European Union and co-financed by the European Social Found.

TBC-37: Subnetwork-based analysis of human disease in protein complex with housekeeping functions

Sanghun Bae[1], Hyunwook Han[2], Hanwool Kim[3] and Jisook Moon[1,2,3]

[1]*College of Life Science, Department of Applied Bioscience, CHA University, Seoul, Korea*

[2]*Department of Biomedical Science, CHA University, Seoul, Republic Korea*

[3]*CHA Stem Cell Institute, CHA Health Systems, Seoul, Republic Korea*

Given that proteins in a living system serve the components of protein complexes or molecular machines to achieve a number of cellular processes and aberrant protein inter-relationship contribute to a disorder of molecular system, a comprehensive analysis of protein-protein interaction network (PPIN) is essential for a systemic understanding of human disease.

However, a substantial number and complexity of the entire protein interaction has led to the difficulty of network-based research, which makes analysis of sub-network, otherwise known as small world, necessary because of the greatly reduced number of proteins to be analysed. In this regard, the present study is concerned with the sub-network consisting of components of one protein complex that is responsible for basic cellular

maintenance functions and their interactors, with our aim focused on systemic approach to human disease.

To construct human interactome PPIN as a first step for this study, we extracted binary protein-protein interaction data from eight molecular interaction database: HIPPIE, HPRD, REACTOME, BIOGRID, InnateDB, DIP, MINT and Intact; and integrated them (172,400 interactions) to increase coverage of PPI data. Proteins of interest used as seed-proteins and their neighbours in the integrated PPIN were selected for creating sub-network, the components of which were mapped to OMIM (Online Mendelian Inheritance in Man) data and GAD (Genetic Association Database) data, representative sources of genotype-phenotype correlation. In enrichment analysis (hypergeometric test), certain disease class terms were over-represented in the sub-network. Moreover, Network properties, GO term and pathway enrichment analysis revealed that the sub-network has distinct features that provide a possible explanation for overrepresentation of particular disease categories in the protein complex with housekeeping function.

Our findings suggest that a subnetwork-based, focused analysis can be a practical application for understanding the underlying nature of human disease and allow us to interpret the properties of disease-related genes on a systemic level.

TBC-38: Functional haplotypes in 5' region of RGS14 gene

Jeyoung Woo[1], Minyoung Kong[1], Younyoung Kim[1] and Chaeyoung Lee[1]

[1]*School of Systems Biomedical Science, Soongsil University, Seoul 156-743, Korea*

Limited knowledge has been known for genetic factors on multiple sclerosis (MS) which leads to nerve degeneration in brain and spinal cord. Recently, an outstanding genomewide association study (GWAS) showed that a single nucleotide polymorphism (SNP, rs4075958) confer the risk of MS. The variant was located in the promoter region of the gene encoding regulator of G-protein signaling 14 (RGS14), a GTPase activating protein (GAP). We investigated the promoter activity of the variants in the region to see whether the sequences can regulate expression of the gene and to identify functional variants in the region. Three haplotypes were estimated with the rs4075958 and 4 SNPs in strong linkage. For each haplotype, a minigene was constructed containing the selected SNPs and firefly luciferase gene. Luciferase activity of each haplotype was measured by Dual-Luciferase Reporter Assay system. As a result, promoter activity has been shown different by the haplotypes. Especially, the largest difference was observed between wild haplotype and the haplotype with all the alleles complement to the wild type. This concurred with the previous GWAS in which the SNP conferred the risk of MS. We concluded that the haplotype with the complement alleles could increase expression of the RGS14 gene. The overexpressed product suppresses $G\alpha_{i/o}$ of mGluR4 and thus increases cAMP that activates T_H17. Consequently, the T_H17 would lead to neuroinflammation, and the accumulated neuroinflammation might increase the susceptibility to MS.

TBC-39: Health SORA, the Smart Health Care Program for Cancer Survivors

Young-Ho Yun[1], Ye-Ni Choi[1], Moon-Kyung Shin[1], Kwang-Choon Kim[2] and Jaegeol Cho[2]

[1]*Seoul National University College of Medicine, Korea*

[2]*Samsung DMC R&D Center, Korea*

Although the numbers of cancer-survivors are steadily growing, there are few programs designed to accommodate survivors with Information Technology-based (IT) health promotion. According to previous studies, cancer survivors' Quality of Life (QOL) is significantly lower than general population, yet there is few programs designed for QOL of survivors, and only focus on specific area, such as exercise and nutrition. Realizing the need of comprehensive health-care program, we designed an IT-based program called Health SORA (Smart, Optimizing, Realistic, Authentic health care program) customized for total health care of cancer survivors.

We studied and analyzed strategies and theories in various fields: transtheoretical model (TTM), behavior/health psychology, fundamental principles of coaching and other leadership theories. Combining the theories, program flow chart is developed. Health care categories to be managed are determined by previous publications. Categories cover physical, mental, social, and existential areas for complete health care.

Managed categories are 12 total, which including exercise, nutrition, emotion, physical examination, fatigue, sleep, weight control, family and society, existential well-being, comorbidity and medication, pain, and quit smoking and moderate drinking. Each category is managed by following orders and the cycle repeats weekly for most of them: 1)evaluation, 2)analysis, 3)decision making, 4)planning, 5)acting, and 6)monitoring and receiving feedback. For example, user first assess one's exercise behavior (TTM, amount of exercise, regularity etc.) in evaluation. Next, user reviews one's current exercise status and decides whether to manage it or not. Once decided to manage, user can plan for certain education and activity. After actual performance of activity, user manages the category by reviewing one's status change in

management phase.

This is the first smart and comprehensive prognosis program that includes 12 important health care areas for cancer survivors. We believe that this total health care program can effectively contribute to improve health and QOL of cancer survivors.

TBC-40: A computational framework for differential alternative polyadenylation profiles between cancer and normal cells

Jimin Shin[1,2], Hyunmin Kim[1], Chaeyoung Lee[2] and David Bentley[1]

[1]*Department of Biochemistry and Molecular Genetics, University of Colorado School of Medicine, Aurora, Colorado, USA*
[2]*School of Systems Biomedical Science, Soongsil University, Seoul, Korea*

Alternative polyadenylation of mRNAs is greatly concerned as an important mechanism for post-transciptional regulation in eukaryotic genes. Approximately half of all expressed genes are thought to produce alternatively polyadenylated mRNAs in human. Recent studies showed that alternative polyadenylation in a specific tissue turned out to be important in oncogenesis. For example, mRNA isoforms having longer or shorter UTR lengths were observed in breast cancer cell lines, and a direction of the length changes is cell-type-dependent. This study aimed to overcome limitations of appropriate statistical background models and quantification of changes in the number of polyA sites in the currently available computational analysis of Alternative polyadenylation. We proposed an analysis with a computational framework for evaluation of the differential Alternative polyadenylation profiles between normal and cancer cells. The proposed approach deals with tasks of peak identification and peak comparison. It was to use a nonparametric normalization with LASSO algorithm in order to panelize peak patterns with artifacts. This method is called polyA shifting index (PSI). The PSI has a property of capturing non-linear trends of the changes in the numbers of polyA sites. Furthermore, the corresponding statistic also has an unbiasedness property in the changes over a long distance. The proposed method is needed to be publically available, which would accelerate identification of the differential Alternative polyadenylation profiles.

TBC-41: The genetic regulation of aging process and age-related disease

Han Wool Kim[1], Hyun Wook Han[2], Sang Hun Bae[2] and Ji Sook Moon[1,2]

[1]*CHA Stem Cell Institute, CHA Health Systems, Seoul, Republic Korea*
[2]*Department of Biomedical Science, CHA University, Seoul, Republic Korea*

Aging process is inevitable biological process of all life, and its fundamental mechanism remains unresolved. Recent studies only investigated simple difference of the network properties and disease classification from the relationship between aging genes and genetic disease genes. Further contributing factors such as methylation and miRNA are more important to uncover aging process and pathogenesis of diseases. Here, for further investigation, we compiled and analyzed human disease (OMIM) and aging (GenAGE) genes to investigate the relationship between aging and disease genes. We categorized the genes with three gene groups: disease only genes, aging only genes, and aging-disease genes. Each of these groups was subsequently characterized. Of the 2117 genes, 1856 genes were disease only, 155 genes were aging only, and 106 were aging-disease genes. Interestingly, Analyses of GO (Gene Ontology) enrichment, transcription factor, protein interaction network, and methylation revealed that each gene group is uniquely involved in different functional categories, and show different transcription factors, miRNA, degree centrality, and methylation pattern. Also, from analyses of disease genes, we uncovered that disease only and aging-disease genes are enriched in different disease categories. Our results shed light on elucidating the relationship between the genesis of a various diseases and aging process.

TBC-42: Discovery of Pathway Information Content of Protein Domains based on Domain Co-occurrence Network

Jung Eun Shim[1] and Insuk Lee[1]

[1]*Department of Biotechnology, College of Life Science and Biotechnology, Yonsei University, 262 Seongsanno, Seodaemun-Gu, Seoul, 120-749, Korea*

Identification of functional building blocks, such as proteins, genes, and protein domains, is important for understanding the biological processes of a cell. Protein domain is particularly useful feature, because it is the structural, functional and evolutionary units of proteins. However, domain-based identification of protein function is still quite difficult problem. In this reason, we developed a network-based quantification of domain functions to identify protein domains which play a critical role to drive protein-level functions, using Domain Information Content Score (DomICS). In this framework, we first constructed a gene network by domain co-

occurrence measured in which we give larger weights to rarer domains, and then measured association scores of a specific pathway using the linkage information in our network. Finally, we developed the pathway information content of each domain, meaning the specificity of pathway associated domains.. To evaluate the performance of the proposed method, in a microbe yeast (Saccharomyces cerevisiae) and multi-cellular human (Homo sapiens), we evaluated the predicted pathway information content of each domain by literatures and the enrichment analysis with known domains for Gene Ontology biological process (GO-BP) terms by Interpro2GO.

TBC-43: Identification and Characterization of Gastric Cancer Subtypes using Expression Microarray Data

Haein Kim[1], Ensel Oh[1], Young Kee Shin[1] and Yoon-La Choi[2]

[1]*Laboratory of Molecular Pathology and Cancer Genomics, Seoul National University College of Pharmacy, Seoul, Korea*
[2]*Laboratory of Cancer Genomics and Molecular Pathology, Department of Pathology, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Korea*

Gastric cancer is one of the most common cancer in Korea, and the development of targeted therapies in the treatment of gastric cancer have been accelerated by the emerging understanding of gastric cancer genome. Alike other types of cancer, gastric cancer is highly heterogeneous, and the identification and characterization of gastric cancer subtypes are the first step to search novel targets for anti gastric cancer drugs. We selected 265 genes showing significantly over expressed in gastric tumors by comparing the expression microarray data of 80 paired gastric tumor and matched normal tissues using Significance Analysis of Microarray (SAM) and NetRank with COXPRESdb database. With the selected genes, we identified two subtypes (subtype A and subtype B) of gastric cancer by clustering the independent 200 gastric cancer tissues. According to GO analysis, the 88 genes which showed high expression in subtype A were related to angiogenesis and Wnt-signaling, and the last of the selected genes which showed high expression in subtype B were involved with immune response such as monocyte and leukocyte chemotaxis. We observed that the subtype A included high stage (stageⅢ, IV) tumors more than subtype B, and it seemed to be related with the active angiogenesis and Wnt-signaling in subtype A. In subtype B, high activity of immune response seemed to keep early tumors from developing to higher stage. From the identification of two subtypes of gastric cancer and characterizing each subtype, we could understand the gastric cancer genome more profoundly and the selected genes would provide the clue to find the targets for anti-gastric cancer drugs.

TBC-44: Functional nucleotide polymorphism in the promoter region of WFS1 gene

Yoonsook Moon[1], Minyoung Kong[1], Younyoung Kim[1] and Chaeyoung Lee[1]

[1]*School of Systems Biomedical Science, Soongsil University, Seoul, Korea*

Genomewide association studies have identified common variants of the genetic risk for type 2 diabetes (T2D), especially by several international consortia. A recent meta-analysis has revealed four nucleotide variants including rs4689388 associated with T2D ($P < 2 \times 10^{-8}$). The variant was located in the promoter of Wolfram Syndrome 1 (WFS1) gene. Thus, we investigated promoter activity with 2 haplotypes (ATCGT with the frequency of 0.67, GATCG with the frequency of 0.33) estimated with 5 SNPs (rs4689388, rs4320200, rs13107806, rs13127445, and rs4273545) in strong linkage around the rs4689388. Luciferase assay for reporter-WFS1 haplotype constructs in HEK293 cells showed that the minigene with the wild haplotype showed a larger expression level than that with the minor haplotype ($P < 0.05$). Further analysis revealed that the expression level with the minor haplotype was smaller ($P < 0.05$) than that with the substitution of its first allele (AATCG), but corresponding to that with the wild haplotype ($P > 0.05$). In conclusion, rs4689388 was the functional variant for up-regulation of the WFS1 gene. Its major allele (A) could produce excessive product of the gene, which increases endothelial reticulum (ER) stress. Finally, a considerable ER stress would lead to a large susceptibility to T2D.

TBC-45: Comparison of Formaldehyde Fixed Paraffin Embedded (FFPE) and Frozen Tissues for Exome Sequencing

Ensel Oh[1], Yoon-La Choi[2] and Young Kee Shin[1]

[1]*Laboratory of Molecular Pathology and Cancer Genomics, Seoul National University College of Pharmacy, Seoul, Korea*

[2]*Laboratory of Cancer Genomics and Molecular Pathology, Department of Pathology, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Korea*

Formalin-fixed, paraffin-embedded (FFPE) tissue is the

most widely practiced method for clinical sample preservation and archiving. However, FFPE tissues have been unfavoured for NGS sequencing because its DNA/RNA is likely to be mutated or degraded through the preparation procedure of formaldehyde fixation. We investigated whether the DNA from FFPE tissue was compatible with frozen tissues for exome sequencing. Exome sequencing was performed with two paired FFPE and frozen tissues generated from two dermatofibrosarcoma protuberance (DFSP) cancer tumors. The DNA from the FFPE tissues were severely degraded compared to the frozen tissues, therefore, the insert size of the FFPE tissue was quite shorter than the frozen tissues. However, the sequencing base quality of the FFPE tissues was as good as frozen tissues, and the average coverage of both types of tissues were almost the same as about x100. The rate of properly mapped paired reads were about 90% for frozen tissues and 70% for FFPE tissues, and more than 95% of total targeted exomes were completely covered in both frozen and FFPE tissues. The number of SNPs called from FFPE tissues were similar to from the frozen tissues, and the dbSNP rate and Ti/Tv ratio of SNPs from FFPE tissues were 95% and 2.5 respectively. The number of Indels from FFPE tissues were also similar to from frozen tissues. Tumor specific SNPs were selected by subtracting the SNPs in blood from either the SNPs in FFPE or in frozen tissues, and the FFPE and frozen tissues showed well overlapped lists of SNPs indicating that FFPE is compatible with frozen for exome sequencing. From the results, we conclude that FFPE tissue could be a good resource for cancer genome study using exome sequencing.

TBC-46: Molecular and biochemical characterization on the artificial hibernation in the olive flounder, *Paralichthys olivaceus*

Meehye Kang[1], Gila Jung[1], Sung Kim[1], Wan-Soo Kim[1] and Youn-Ho Lee[1]

[1]*Marine Ecosystem Research Division, Korea Institute of Ocean Science & Technology, Ansan, Korea*

The aim of this study was to understand the molecular and physiological changes in an artificially hibernated olive flounder, Paralichthys olivaceus. At first, biochemical properties of artificially hibernated organism were examined through blood analysis. Serum glucose and triglyceride were significantly increased ($p < 0.05$) during hibernation, while alkaline phosphate (ALP) and glutamic-pyruvic transminase (GPT) had no significant change ($p > 0.05$). Then the genes associated with the artificial hibernation were investigated with the brain tissue using RNA-seq technology. Change of the expressed genes was examined with DEGseq R package, and gene ontology (GO) functional enrichment analysis. A total of 915 differentially expressed genes including 468 up-regulated and 447 down-regulated genes ($p < 0.001$) were identified. The GO of the differentially expressed genes (DEGs) revealed 45 significantly enriched GO terms indicating up and down regulation of genes, most of which were associated with protein binding, transcription factor activity, transcription factor complex, and sequence-specific DNA binding. Several genes such as intestinal fatty acid binding protein (IF), period 4, and somatolactin (SL) showed significant change in the expression level. For IF and SL, the change of expression level was quantitatively confirmed by the real time PCR.

TBC-47: Unraveling selection signatures by composite log likelihood

Jihye Ryu[1] and Chaeyoung Lee[1]

[1]*School of Systems Biomedical Science, Soongsil University, Seoul 156-743, Korea*

Positive selection not only increases beneficial allele frequency but also causes augmentation in allele frequencies of sequence variants in proximity. Signals for the positive selection would be identified by harbouring distribution of the sequence variants around a favourable mutation, and statistical differences from the expected values by chance determines the signals. We introduced a composite log likelihood-based method (CLL) which calculates a composite likelihood of the allelic frequencies observed across sliding windows of 5 adjunct loci and compares the value with the critical statistic estimated by 50,000 times of permutation. We applied the method to identification of selection signatures in Korean cattle. A total of 11,799 nucleotide polymorphism data were used for 71 Korean cattle and 209 foreign beef cattle. As a result, 147 signals were observed between Korean cattle and foreign cattle ($P < 0.01$). The selection signatures with the greatest CLL for each of 30 chromosomes encompassed 148 sequence variants among which 41 variants were located in the region encoding proteins. The signals might be candidate genetic factors for beef quality by which the Korean cattle have been selected.

TBC-48: The Health Avatar Platform: development of platform for interacting health agents and personal avatar

Hee-Joon Chung[1,2], Byoungoh Kim[1], Taehun Kim[1], Keun Bong Kwak[1] and Dongman Lee[1]

[1]*Department of Computer Science, KAIST, Daejeon*

305701, Korea

[2]*Seoul National University Biomedical Informatics (SNUBI), Seoul National University College of Medicine, Seoul 110799, Korea*

eHealth is a field of increasing interest with the potential to revolutionize the way health care and prevention is provided, shifting the balance of power and responsibility from health care professionals to patients and citizens. Health avatar is a user application that provides health information through health agent based on personal medical, genomic and ubiquitous data. The Health Avatar Platform (HAP) is a run-time environment for allowing appropriate intelligent health agents to get "plug-in"ed to a health avatar and providing a data and access grid for heterogeneous clinical and genomic data.

We have completed the first phase of the HAP: a) defining an application programming interface for both avatar and agent developers, b) developing a broker that provides a match-making service between agent and avatar and a communication channel between them, and c) prototyping an obesity management agent application as a showcase of the system capabilities.

TBC-49: Systematic Analysis of Genotype-dependent Gene Expression Signatures and Drug Sensitivity in NCI60 Datasets

Ningning He[1] and Sukjoon Yoon[1]

[1]*Department of Biological Sciences, Sookmyung Women's University, Seoul 140-742, Korea*

Most cell lines recapitulated known tumor-associated genotypes and genetically defined cancer subsets, irrespective of tissue types. Drug treatment on many different cell lines provides an important preclinical model for early clinical applications of novel targeted inhibitors. The NCI60 is a program developed by the NCI/NIH aiming the discovery of new chemotherapeutical agents to treat cancer. Here we present a novel statistical method, CLEA (Cell Line Enrichment Analysis) to quantitatively correlate the genotype with gene expression signatures and drug sensitivity in cancer cell lines. The results provided us new insights on genotype-dependent gene expression signatures, cancer pathways and chemical sensitivity. It will have applications in predicting and optimizing therapeutic windows of anti-cancer agents.

TBC-50: The role of TRP channel interactome in prostate cancer

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

[1]*Biomedical Knowledge Engineering Lab, Seoul National University, Korea*

[2]*Department of Physiology, Seoul National University College of Medicine, Korea*

Transient receptor potential (TRP) channels translate various cellular stimuli into electrochemical signals, leading to changes in membrane potentials and intracellular Ca^{2+} levels. Aberrant regulation of intracellular Ca2+ homeostasis is closely associated with various cancers, particularly prostate cancer: however, the possible involvement of TRP channels in prostate cancer is largely unknown. To explore the role of TRP channels in prostate cancer, in this study, we have attempted to extract and integrate two different datasets: prostate cancer microarray data from the GEO database (accession # GSE3325) and TRP channel interactome data from the TRIP Database 2.0 (http://www.trpchannel.org). We found altered expression pattern of TRP channel interactome components according to tumor stages (benign, primary, and metastatic), which is represented as node-weighted networks using the Cytoscape program. Co-expression correlation analysis identified that certain TRP channel isotypes tend to be co-expressed with their interacting proteins, which can support disease module hypothesis of network medicine. In addition, we performed GO and pathway analyses to identify how certain TRP channels are associated with prostate cancer phenotypes. Our results may help future experimental investigation to understand the role of TRP channel-mediated Ca^{2+} signaling in prostate cancer biology and to develop novel therapeutic strategies for treatment of prostate cancer. [This research was supported by the MKE(The Ministry of Knowledge Economy), IT Convergence Healthcare Research Center support program supervised by the NIPA(National IT Industry Promotion Agency) (NIPA-2012-H0401-12-1001)]

TBC-51: Using CSSP to predict chameleon peptides

Xiaoqi Wang[1] and Sukjoon Yoon[1]

[1]*Department of Biological Sciences, Sookmyung Women's University, Seoul 140-742, Korea*

The sequence potential for non-native β-strand formation and the presence of protein sequences have been investigated extensively from the perspective that such structural features are implicated in protein stability and effectiveness. We demonstrated that calculation of contact-dependent secondary structure propensity (CSSP) is highly sensitive in detecting non-native beta-strand propensities in helical regions of proteins. Beta-sheet formation is the main reason for protein aggregation.

Based on our study, the CSSP method offers an alternative for designing peptide fragments with varied propensity for conformational change between helix and beta-strand.

TBC-52: Transcriptome analysis during the developmental stages for predator induced polyphenism in *Daphnia pulex*

Haein An[1], Gila Jung[2] and Chang-Bae Kim[1]

[1]*Department of Green Life Science, Sangmyung University, Seoul 110743, Korea*

[2]*Marine Ecosystem Research Division, Korea Institute of Ocean Science and Technology, Ansan 426744, Korea*

An invertebrate crustacean *Daphnia pulex* is one of the most suitable models for understanding how organisms adapt and survive to aquatic environmental stresses including predator-induced morphological responses. It has been known that neckteeth formation and maintenance at critical times is a defensive mechanism for D. pulex against the predator *Chaoborus sp*. The genetic mechanism of the defensive morph formation and maintenance for developmental ranges is very little known. To understand its genomic mechanism, we carried out comprehensive transcriptomes at various developmental stages in *D. pulex* by using RNA-seq technique. As the results, 37 Gb raw reads were generated and assembled. The 62,228 unigene clusters were annotated by blastx alignments against NCBI non-redundant (NR), COG, SwissProt, GO, and KEGG databases. According to the searches, 30,495 unigene clusters were matched to at least one database. Gene expression differences among developmental stages were greater than those between the two phases, normal and defensive morph in each stage. Differentially expressed transcripts (DETs) were discovered by measuring and comparing gene expression between the two phases in each stage. The most distinct phase differences in gene expression appeared in adult/egg stage. According to the detailed analyses, the defensive morph in the stage shows lower activity in signalling molecules and interaction, nucleotide metabolism. We identified 68 transcripts as candidates for defensive morph markers, containing insect cuticle protein and receptor transporting protein. This study could contribute to further studies of the candidate genes and epigenetic mechanism for defensive morph formation and maintenance in *D. pulex*.

TBC-53: Network analysis by phylogenetic profiling revealed domain-specific evolution of cellular pathways

Junha Shin[1] and Insuk Lee[1]

[1]*Network Biology Laboratory, Department of biotechnology, College of Life Science and Biotechnology, Yonsei University, Seoul 120749, Korea*

Phylogenetic profiling is a computational method to identify functional associations of genes within one organism, based on the comparisons of evolutionary co-inheritance patterns according to the completely sequenced genomes of other organisms. The composition – both abundance and heterogeneity - of genome set and the scoring scheme for relationship are two important factors to affect to the utility of a profile. Because a profile needs only genome sequence data to be generated, it is a practical bioinformatic technique along with recently advanced sequencing techniques and those exponentially growing sequenced data results. There are several previous reports that this method works optimally with a genome set consisted of bacterial organisms only.

Here we reinvestigated the optimal condition for phylogenetic profiling with increased fully sequenced genomes which were not available in previous studies. We could verify the improvement of prediction performance by grown numbers of genome data; therefore, at now, it could be available not only to discover functional association of genes even in higher eukaryote but also to retrieve human disease genes via investigating the resultant network model. Moreover, co-inherited genes associations show differences in various features between the inherited orientation of prokaryote and eukaryote. Followed by these distinctions, we could find the domain-specific nature and also explain the molecular mechanisms of pathway-level evolution.

TBC-54: Functional polymorphism located in the promoter of the coagulation factor XI gene as a putative genetic factor for susceptibility to venous thromboembolism

Minyoung Kong[1], Younyoung Kim[1] and Chaeyoung Lee[1]

[1]*School of Systems Biomedical Science, Soongsil University, Seoul, Korea*

Several genome-wide association study (GWAS) and meta-analysis of GWAS have been conducted for venous thromboembolism (VTE). A recent MARTHA and FARIVE project was reported the rs3756008 in promoter region of the coagulation factor XI (FXI) gene as nucleotide sequence variant associated with VTE in European ($P = 6.46 \times 10^{-11}$). Coagulation factor XI (FXI) is the zymogen of a plasma serine protease (FXIa) triggered the middle phase of the intrinsic blood coagulation pathway, and its plasma levels were associated with VTE. Thus, we searched the SNPs in strong linkage around the rs3756008, and the rs3756009

was selected. We investigated alteration of luciferase-reporter gene expression by the 2 haplotypes (AA with the frequency of 0.62, TG with the frequency of 0.38) and by the each SNP in HEK293 cells. Wild haplotype-reporter minigene showed a larger expression level than minor haplotype-reporter minigene ($P < 0.001$). Further analysis revealed that nucleotide substitution (A to T) at rs3756008 showed difference for expression level of 2 haplotypes ($P < 0.001$). In conclusion, minor allele (T) at rs3756008 was the regulatory allele for low expression of the FXI gene. Low FXI levels might result in reduced functional activity of activated coagulation factor XII (FXIIa), and blockage of FXIIa activity might be involved in the risk of vessel occlusion. It could not exclude a possibility that low FXI levels might lead to a susceptibility to VTE.

TBC-55: Temporal gene expression profiles identify genetically determined transcriptional regulation of human leukocytes

SeongBeom Cho[1], InSong Go[2], Hyo-Jeong Ban[1], Hyesun Yoon[1], Yeunjung Kim[1], Jaepill Jeon[1] and BokGhee Han[1]

[1]*Center for Genome Science, National Institute of Health, Korea Center for Disease Control, Chungcheongbuk-do, Republic of Korea*

[2]*Department of Physiology, School of Medicine, Hanyang University, Kyungkido, Republic of Korea*

In this study, we investigated genetic markers affecting temporal gene expression in human leukocytes using expression quantitative trait (eQTL) loci analysis. During an oral glucose tolerance test, glucose, insulin levels and gene expressions of leukocytes in peripheral blood were measured at three time points. Through eQTL analysis, we identified relationship between gene expression, genetic component and environmental factors. Association analysis between the gene expressions and SNPs only (marginal model) found *cis* SNPs showing differential allele-specific gene expression. The analysis with the interaction terms (interaction model) identified interactions between SNPs and temporal glucose or insulin levels, or both, which significantly affected gene expression. Functional annotation revealed that the significant SNPs of the marginal model were related to various diseases. Moreover, SNPs of the interaction model showed a strong tendency for transcription factor binding site enrichment. Finally, using a differential allele-specific coexpression (DACE) method, we searched for SNP–pathway pairs that showed molecular networks of significant allele-specific changes of coexpression. The DACE method identified a *trans*-regulatory effect of the SNPs on pathway gene coexpression patterns. In conclusion, we identified tentative genetic markers affecting temporal gene expression change in human leukocytes through a genetic component alone or through interaction with the genetic components, glucose and/or insulin. These results will be resource for studying regulatory components of biological processes that are either determined by genetic component alone or by gene–environment cross talk.

TBC-56: gsGator – an integrated web platform for cross-species gene set analysis

Hyunjung Kang[1], Sooyoung Cho[1], Ikjung Choi[1], Yeongjun Jang[2], Sanghyuk Lee[1,2] and Wankyu Kim[1]

[1]*Department of Life and Pharmaceutical Science, Ewha Womans University, Ewha Research Center for Systems Biology, 52, Ewhayeodae-gil, Seodaemun-gu, Seoul 120-750 Korea*

[2]*Korean Bioinformation Center, Korea Research Institute of Bioscience and Biotechnology, Daejeon 305-806, Korea*

Gene set analysis (GSA) is useful to interpret its biological theme using *a priori* defined gene sets such as gene ontology or pathway. While model organisms are a rich source for inferring the function of human genes, few GSA tools enable to use these information. Here, we developed gsGator, a web-based platform for functional interpretation of gene sets with many useful features such as cross-species GSA, simultaneous analysis of multiple gene sets, and a fully integrated network viewer. An extensive set of gene annotation information is amassed including GO & pathway, genomic annotation, molecular network, miRNA target and phenotype information from various model organisms. gsGator enables virtually fully-automated analysis, providing intuitive understanding of the relations among genes and gene sets using an interactive network viewer. Particularly, gsGator supports cross-species GSA in a user-friendly manner, allowing full utilization of accumulated knowledge e.g. knockout phenotype from model organisms. Cross-species GSA greatly expands the scope of GSA, leading to the discovery of conserved gene modules among different species. (http://gsGator.ewha.ac.kr).

TBC-57: Identification of transcriptional network regulating prognostic gene expression signature of colorectal cancer patients

Taejeong Bae[1,2,2], Kyoohyoung Rho[1], Yong-Ho In[2,3] and Sunghoon Kim[1]

[1]*College of Pharmacy, Seoul National University, Seoul 151-742, Korea*

[2]*Information Center for Bio-pharmacological Network, Seoul National University, Suwon 443-270, Korea*

[3]*Medicinal Bioconvergence Research Center, Advanced*

Institutes of Convergence Technology, Suwon 443-270, Korea
[4]*Korean Bioinformation Center, Daejeon, Korea*
[5]*World Class University Program Department of Molecular Medicine and Biopharmaceutical Sciences, Seoul National University, Seoul 151-742, Korea*

Background: Identification of gene expression signatures in cancer patients has been proven useful to determine the cancer types and stage and also to predict the prognosis of patients. However, expression signature itself does not provide information about the causality of changes of pathological cellular states. Construction of a transcription network that regulates the cancer signature can provide clues to hidden mechanisms of cancer progression.

Results: Here we inferred and analysed the transcriptional network regulating prognostic gene expression signature of colorectal cancer that is known to classify patients to good prognosis and poor prognosis group. To construct a colon cancer-specific regulatory network, we used the ARACNE algorithm followed by a series of filtering algorithms to find significant transcription factors. The inferred network consists of 9 transcription factors (TFs) regulating 75 genes out of 86 genes in colon cancer signature. The following analysis identified 6 TFs (PRRX1, SPDEF, FOSL2, HIF1A, RUNX1 and FOXD1) as master regulators regulating high risk signature genes for poorer prognostic subgroup and 3 others (PLAGL2, ASCL2 and TCF7) as ones regulating low risk signature genes for better prognostic subgroups. The common tumorigenic feature of HIF1A, RUNX1 and FOSL2 suggested that the tumorigenic feature of prognostic gene signature may be involved in metastasis of colorectal cancer while the tumorigenic roles of PRRX1, SPDEF and FOXD1 are unclear.

*Conclusions:*These results showed that the transcriptional network analysis is a powerful tool to reveal the regulatory programs related to prognosis of colorectal cancer patients.

TBC-58: Local Similarity Search of Physicochemical Properties in Protein-Ligand Binding Sites

Lee Sael[1] and Daisuke Kihara[2]

[1]*State University of New York Korea, Korea*
[2]*Purdue University, USA*

Physicochemical similarity search of protein binding site have various applications such as finding the protein binding partners, protein function prediction, and prediction of unintended drug binders. We present two ligand binding pocket comparison methods: Pocket-Surfer (Chikhi R. et al. Proteins, 2010) and Patch-Surfer (Sael L. et al. Proteins, 2012). Pocket-Surfer captures shape and physicochemical properties of a binding site surface globally. In contrast, Patch-Surfer represents a binding site as a combination of segmented surface patches, each of which is characterized by its geometric shape, electrostatic potential, hydrophobicity, and concaveness. By relaxing the constraint put on by rigidity of global binding site structure, local similarities can be captured. This is effective when pocket shapes are slightly different due to structural flexibility but bind to the same ligand type. Both methods encode the surface properties of whole pocket or patches that compose the pockets by the 3D Zernike descriptors, which have been found to be successful in representing protein global surface properties (Sael L., Li B., et al. Proteins, 2008; Sael L., La D. et al. Proteins, 2008). We validated the two proposed method by measuring the prediction accuracy of the ligand binding predictions, i.e., predictions of the types of ligand that can bind to proteins. The performance was evaluated on a data set of 100 non-homologous proteins that bind to either one of nine types of ligands. 84.0% of the binding ligands were predicted correctly within the top three scoring ligands with the shape and pocket size information using the Patch-Surfer and 81.0% when Pocket-Surfer was used. The performance was further improved to 87.0% when surface properties, i.e. electrostatic potential and hydrophobicity, were added in the Patch-Surfer. Overall, we show that proposed methods are powerful in protein binding site similarity analysis even in the absence of homologous proteins in the database.

TBC-59: Association analysis of CNV data with linear mixed model

Meilling Liu[1], Sanghoon Moon[2], Youngjin Kim[2] and Sungho Won[1]

[1]*Dept of Statistics, Chung-Ang University, Korea*
[2]*The Center for Genome Science, Korea National Institute of Health, Korea*

Copy number variation (CNV) has been expected to have an important effect on human genetic diseases. However even though several statistical methods have been proposed for CNV association studies, most of the existing approaches are restricted to the independent individuals. In this manuscript, we provided a new method for the analysis of CNV with related samples and it can also be applied to the unrelated samples under the presence of population substructure. The proposed approach consists of signal model, phenotype model and copy number model where the signal model provides the relationship between the observed intensity and the unknown CNV, and phenotype model explains the

causality of the CNV to the phenotype. In our approach, we considered the correlation structure for both signal and phenotype model, and the multiple probe intensities are incorporated to them. Our simulation studies show that the proposed method outperforms the previous approaches and we illustrate the practical implications of the new analysis method by an application to Alzheimer.

TBC-60: Analysis of longitudinal data : Applications of Linear Mixed Model to The Korean Association Resource(KARE)

Young Lee[1], Suyeon Park[1], Woojoo Lee[2] and Sungho Won[1]

[1]*Dept of Statistics, Chung-Ang University, Seoul, Korea*

[2]*Department of Statistics Inha University, Korea*

Last decade genome-wide association studies (GWAS) has been successfully accomplished and we could find many significantly associated SNPs with phenotypes of interest. However the multiple testing problem is still intractable issues and it becomes more serious for next generation sequencing analysis. In this manuscript, we investigated the analysis of longitudinal data for GWAS. Because genotyping cost is often more expensive than phenotyping, the longitudinal data analysis can be an alternative choice for multiple testing problems. Here the linear mixed model has been applied to the phenotypes with repeated observations in Korean Association REsource (KARE) project and principle component analysis (PCA) has been conducted to adjust for population stratification. We found that the power is proportional to the number of repeated measurements and sample size while it is inversely proportional to the correlation coefficient of repeated observations.

TBC-61: Differential influences of common variants on erythrocyte-related traits according to Sasang constitutional types

Seongwon Cha[1], Hyunjoo Yu[1] and Jong Kim[2]

[1]*Constitutional Medicine & Diagnosis Research Group,* [2]*Vice-President, Korea Institute of Oriental Medicine (KIOM), Daejeon, 305-811, Korea*

Hematological disorders such as anemia and erythrocytosis characterized by measuring erythrocyte-related traits are known to be associated with cardiometabolic diseases. Genetic variants associated with hematological traits have been elucidated in several genome-wide association studies (GWAS). In Sasang constitutional medicine (a Korea-specific type of personalized medicine), human beings are categorized into four types harbouring differential prevalence of cardiometabolic diseases and anemia. In this study, we aimed to investigate whether each constitutional type had differential genetic factors associated with hematological traits. Therefore, we examined the effects of the variants reported to be definitely associated with hematological traits from previous GWAS researches on the same hematological traits according to Sasang constitutional types. We performed multiple linear regression analyses with measurements of RBC, Hb, Hct, MCV, MCH, MCHC, and RDW in two Korean populations: 1,701 and 3,472 subjects recruited from the Korea Constitution Multicenter Study and the Korea Genome and Epidemiology Study, respectively. The Sasang constitutional types were categorized by the Sasang Constitutional Analysis Tool: in total, 2,696 subjects with Taeum type, 1,881 subjects with Soyang type, and 596 subjects with Soeum type. Among initially selected over 30 polymorphisms, we finally found 4 variants in 4 genetic loci (*HBS1L-MYB, TMPRSS6, SPTA1, and ITFG3*) presenting association signals both in the two populations. Two variants of *HBS1L-MYB* and *TMPRSS6* were associated with measurements of RBC, MCV, MCH, MCHC, and/or RDW in total population and two sub-populations with Taeum and Soyang types. The variant of *SPTA1* was associated with MCHC in total populations, and the *ITFG3* variant was associated with Hb in a sub-population with Soeum type. These results showed that the profile of variants associated with hematological traits was different according to Sasang constitutional types, especially between Soeum type and the others.

TBC-62: Comparing algorithms for genotype imputations in family-based design

Youngdoe Kim[1], Jungmin Lim[2], Donghe Li[2], Jaemoon Lee[2] and Sungho Won[2]

[1]*Division of Structural and Functional Genomics, The Center for Genome Science, Korea National Institute of Health, KCDC, Osong, Korea*

[2]*Department of Applied Statistics, Chung-Ang University, Seoul, Korea*

Genotype imputation is now an essential tool in the analysis of genome-wide association scans to handle the missing data, untyped genotypes, etc. However, even though its importance, a few approaches have been proposed for the imputation of genotype in family-based design, and the accuracy for each method has not been confirmed. In this manuscript we compared several methods for genotype imputations with Korean Healthy TWIN cohort. We compared IMPUTE2, BEAGLE, MACH and GHOST, and the accuracy for each software has been calculated. In addition we considered two-stage imputation algorithm. We, first, impute the genotypes

with Mendelian transmission and then haplotype-based imputation algorithm has been conducted. Even though the difference between different software is small, our results show that the two-stage algorithm performs slightly better.

TBC-63: A large-scale genome-wide association study of Korean Family cohorts for genetic variants influencing metabolic syndrome

Youngdoe Kim[1,2], Yong Ki Jung[2], Sung Oh Kang[2], Nam Hee Kim[1], Young Jin Kim[1],Juyoung Lee[1], Sungho Won[2]

[1]*Division of Structural and Functional Genomics, The Center for Genome Science, Korea*

National Institute of Health, KCDC, Osong, Korea

[2]*Department of Applied Statistics, Chung-Ang University, Seoul, Korea*

To identify genetic factors influencing several traits (height, body mass index (BMI), triglycerides (TG), high density lipoprotein (HDL), low deinsity lipoprotein (LDL), diastolic blood pressure (DBP) and systolic blood pressure (SBP)) of metabolic syndrome (MetS), we conducted a genome-wide association study (GWAS) with 1,801 samples from Korean Healthy Twin cohorts and 784 samples from Ansung Family extended cohorts recruited in Korea. In particular we found that the phenotypic distributions for TG were not normally distributed and thus they were log-transformed for GWAS. The linear mixed model with the restricted maximum likelihood (REML) method has been applied to find a significant association. We found that two SNPs were significantly associated with log TG at the genome-wide scale and both SNPs were replicated in the other cohort.

TBC-64: Ethical, Legal, and Social Frameworks on Issues of Bioinformatics

Hannah Kim[1], Ilhak Lee[1], Ji Yong Park[1], Sang Hyun Kim[2] and So Yoon Kim[1,3]

[1]*Department of Health Law and Bioethics, College of Medicine, Yonsei University, Korea*

[2]*Department of Health Law and Bioethics, Graduate School of Public Health, Yonsei University, Korea*

[3]*Centre for ELSI Research, Asian Institute for Bioethics and Health Law, Yonsei University, Seoul 120821, Korea*

Fundamental roles of bioinformatics are to identify the genes and cellular pathways relating to diseases and to link them to the advanced clinical fields such as prevention, diagnosis, and treatment of human diseases. Whereas this field accelerates the progress of development and generalization, it raises various ethical, legal, and social questions focusing on patients or research participants.

Thus, Centre for Ethical, Legal, and Social Issues Research (Centre for ELSI Research) developed frameworks to investigate, analyse, and evaluate the developed issues in the aspects of ethical, legal, and social context. The frameworks are efficient not only to predict the effects of translational bioinformatics and medicine so to make appropriate response or strategies, but also multinational comparative studies. We expect the applicable range of the frameworks is from bioinformatics to other cutting–edge biotechnology area.

Going through the final stage of development of the framework, we are planning next step. It is to address the implications for individuals and society, drawing all prospective ethical, legal, and social issues on each sub-project, as well as reviewing key issues through discussions with researchers and expert panels, as our next step. This article will provide the introduction of the whole schemes for refining them more.

TBC-65: PATH2: Software for Conducing Gene-Ontology And Pathway Based Analyses using Genome-Wide Association Data

Denise Daley[1], David Zamar[1], Ben Tripp[1], Brad Cavanagh[1] and George Ellis[1]

[1]*University of British Columbia, Canada*

Most genome-wide association (GWA) studies lack the power to detect single nucleotide polymorphisms (SNPs) with small effects. However, the aggregate effect of several SNPs working together within a pathway is more easily detectable. Testing for pathway-based association is a promising approach in identifying genes with small additive effects that work together to increase or decrease susceptibility to common complex diseases. Perhaps the most important role performed by pathway-based approaches is in the identification of underlying biological mechanisms leading to disease. Although several algorithms exist for conducting pathway-based analyses, not all of them have been implemented for public usage. We have developed a software package that implements several pathway-based methods and provides an easy to use interface for conducting analyses. Source code and binaries are freely available for download at http://genapha.icapture.ubc.ca/Path2. Our software is implemented in Java, but makes use of both Perl and R and is supported on Linux and Windows. To illustrate its usage, we perform an ontology-based and a pathway-based analysis of the published results from the GABRIEL consortium large-scale genome-wide association study of asthma.

TBC-66: Comparison of Genetic Variations in Drug Metabolizing Enzyme and Transporter Genes among Korean, Japanese, and Chinese Population

SoJeong Yi[1], Sangin Lee[2], Youngjo Lee[2], Seonghae Yoon[1], Inbum Chung[1], HyeKyung Han[1], Jae-Yong Chung[1], Ichiro Ieiri[3] and In-Jin Jang[1]

[1]*Department of Clinical Pharmacology and Therapeutics, Seoul National University College of Medicine and Hospital*
[2]*Department of Statistics, Seoul National University, Seoul, 110-799, Korea*
[3]*Department of Clinical Pharmacokinetics, Graduate School of Pharmaceutical Sciences, Kyushu University, Fukuoka, 812-8582, Japan*

Inter-ethnic difference of genetic polymorphism in genes encoding drug-metabolizing enzymes and drug transporters is one of major factors causing ethnic sensitivity for drug response. In this study, the authors explored genetic differences among 3 major East Asian populations, Korean, Japanese, and Chinese in single nucleotide polymorphisms (SNPs) on genes related with drug absorption, metabolism, disposition, and transport.

Using DMET® plus platform (Affymetrix, USA), the allele or genotype frequencies of 1,936 variants (1,931 SNPs and 5 copy number variations) representing in 225 drug-metabolizing enzyme and transporter genes were determined from 786 healthy male participants (448 Koreans, 208 Japanese, and 130 Chinese). To compare allele or genotype frequencies among 3 ethnic groups in the high-dimensional data, a principal component analysis (PCA) method and regularized multinomial logit model, which is a multi-class classification procedure, were employed.

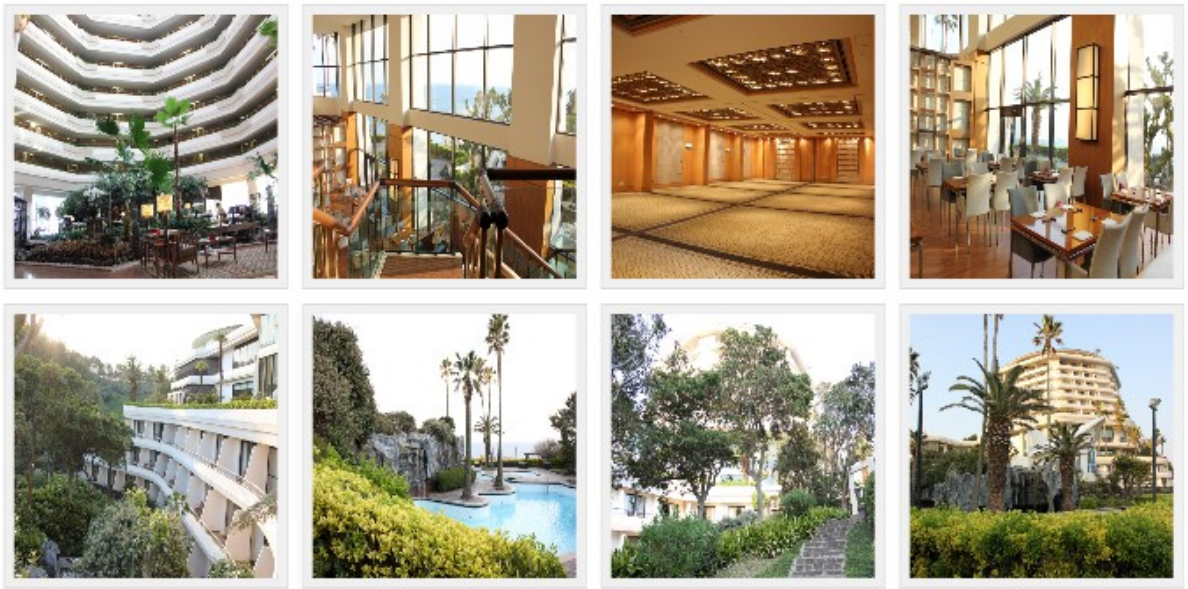
Of the 1,936 variants, 1,071 variants (55.3%) were monomorphic and 127 variants (6.6%) were 'no call', therefore, the rest 738 biallele variants were analysed. The result of PCA showed that Korean, Japanese, and Chinese were not distinguished by first few principal components. However, multinomial logit model via least absolute shrinkage and selection operator (LASSO) could classify three ethnic groups using a model with 105, 98 and 99 selected markers for Korean, Japanese, and Chinese, respectively. The accuracy of prediction model was 87.9%, and misclassification error rate was 12.1%. The most significant genetic variations were EPHX1_16466T>C for Korean (coefficient= -1.24), CYP2A6_1799T>A for Japanese (coefficient = 2.45), and rs17064 on ABCB1 for Chinese (coefficient = 2.37).

In conclusion, this comprehensive genetic variant assessment suggests that genetic differences in genes encoding drug-metabolizing enzymes and drug transporters are very small among Korean, Japanese, and Chinese.

Venue

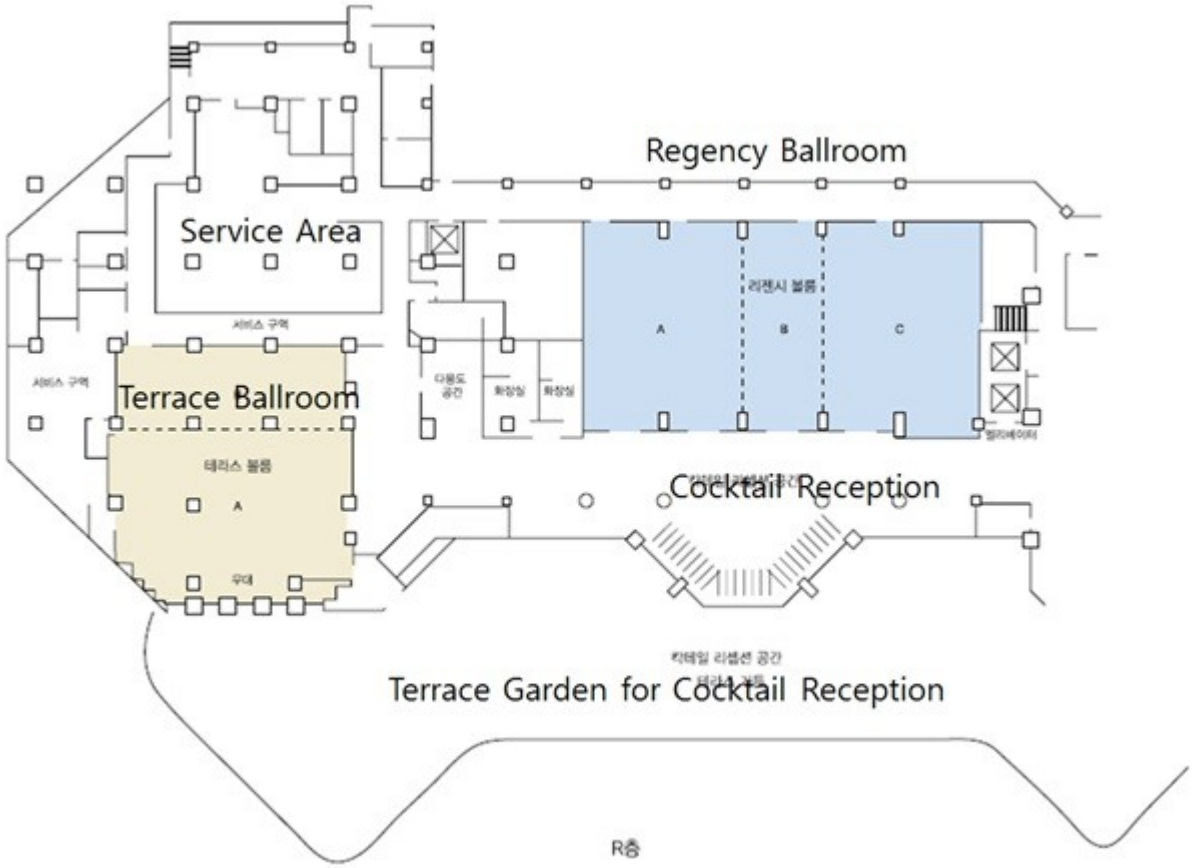
Location

Regency & Terrace Ballroom, Hyatt Regency Jeju, Jeju Island, Korea



Venue

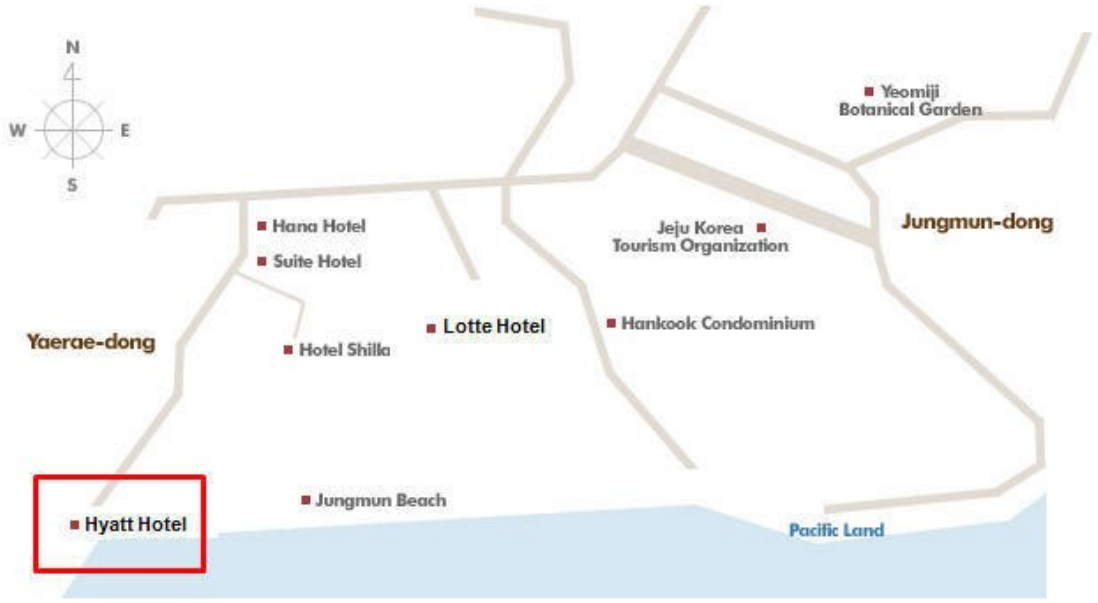
Floor Plan



Terrace & Regency Ballroom

Venue

Map


N
W
E
S
Yeomiji
Botanical Garden
Hana Hotel
Suite Hotel
Jeju Korea
Tourism Organization
Jungmun-dong
Yaerae-dong
Hotel Shilla
Lotte Hotel
Hankook Condominium
Hyatt Hotel
Jungmun Beach
Pacific Land

Tours

Jeju Island - New 7 Wonders of Nature Jejudo is a volcanic island, 130 km from the southern coast of Korea. The largest island and smallest province in Korea, the island has a surface area of 1,846 sqkm. A central feature of Jeju is Hallasan, the tallest mountain in South Korea and a dormant volcano, which rises 1,950 m above sea level. 360 satellite volcanoes are around the main volcano.

Recommended Tour Courses

Seongsan Ilchulbong (Sunrise Peak)

99 rocky peaks surround the crater like a fortress and the gentle southern slope connected to water is a lush grassland.

On the grassland at the entrance of Sunrise Peak, you can enjoy horseback riding. Breathtaking scenic views while taking a rest in the middle of climbing up the peak such as Mount Halla, the deep blues of the ocean, the multi-colored coast line, and the picturesque neighboring villages will become unforgettable memories.

Jeju Olle Tour

"Olle" [Ole] is the Jeju word for a narrow pathway that is connected from the street to the front gate of a house. Hence, "Olle" is a path that comes out from a secret room to an open space and a gateway to the world. If the road is connected, it is linked to the whole island and the rest of the world as well. It has the same sound as "Would you come?" in Korean, so Jeju's "Olle" sounds the same as 'Would you come to Jeju?'.

[Route 8] Wolpyeong to Daepyeong(Port)

This route continues along the seashore through Jusangjeolli which is a formation of stone pillars piled up along the coast. The Jusangjeolli were created when the lava from Mt. Hallasan erupted into the sea of Jungmun. The sight of the abundant pampas grass makes your walk even more enjoyable. A pathway made of numerous rocks on the coastline was built by the marine corps for Jeju Olle, so it is called The Marine Corps Trail. The pathway used to be used only by local divers. The terminus of the route, Daepyeong Pogu (port), sits on the end of a valley and the open fields run towards the sea. Gun San, a mountain in Daepyeong, was created by the son of a sea god thank his master for his merciful attitude, according to local legend.

Halla Mountain

Mount Halla is the mountain of one of the three gods and is a notable mountain. It stands at the center of Jeju Island, spreading east and west. The east face is steep, the north side is gentle, and the east and west form a flat, wide highland. Mount Halla is a dormant volcano created by volcanic activities during the quaternary period of the Cenozoic era. It is primarily covered with basalt. On its top is a crater Baeknok Lake. This mountain is a home to alpine plants and houses as many as 1,800 species of flora. It also boasts luxuriant natural forests and vast grasslands.

In addition, precipitous cliffs and slopes, and unusual rock formations standing along valleys produce magnificent scenic views. The community of azaleas also adds to the beauty of Mount Halla. Mount Hallas autumnal tints and snow-covered scenes have been selected as the best of the best.

It is possible to climb up to Wetse Oreum along Eorimok Trail and Youngsil Trail and to the top along Seongpanak Trail and Kwaneumsa Temple Trail.

Micheon Cave (Ilchul Land)

Filled with underground mystery, Micheon Cave has academic, tourism, and cultural value. Fresh air, crystal clear water, green fields, and a secondary volcanic cone (oreum) are nearby. This underground cave is nature awe-inspiring spot that provides an opportunity for contemplating human nature and the future.

You'll be fascinated with the nature beauty that simply cannot be felt in the city. Micheon Ilchulland Cave, where the new sun is rising, is here to make your tour more enjoyable.

Cheonjiyeon Waterfall

The waterfall falls from a precipice with thundering sounds, creating white water pillars. It has the name Cheonjiyon, meaning 'the heaven and the earth meet and create a pond'. At 22 m in height and 12 min width, the waterfall tumbles down to the pond to produce awe-inspiring scenery. The valley near the waterfall is home to Elaeocarpus sylvestrisvar, ellipticus, which is Natural Monument No. 163, Psilotum nudum, Castanopsis cuspidatavar.sieboldii, Xylosma congestum, Camellia and other subtropical trees. This place is also famous as home to the eel of Anguilla mauritiana, which is Natural Monument No. 27 and is active primarily at night. The Chilsipri Festival is held in every September at the falls.

This is a **Dol-Haru-bang**, and it means a grandfather made of stone. The people in Jeju Island believe that this Dol-Haru-bang is a guardian of the Jejudo to protect it. This souvenir is made a cute feature compare with the original Dor-haru-bang. Therefore, the real Dol-Haru-bang could be seen more fearful a bit rather than this one.

Sponsors

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
Systems Biomedical Informatics National Core Research Center

서울대학교 의료정보학 협동과정
Interdisciplinary Program of Medical Informatics

RCRD
보건복지부지정
Research Center for Rare Diseases
희귀질환 진단치료기술 연구사업단

Sponsors


life
technologies™


DNA Link


MACROGEN


MDxK
Molecular Diagnostics Korea
(주)엠디엑스케이


Sandia
Linux Cluster Expert Group


SK telecom


INSILICOGEN
www.insilicogen.com
㈜ 인실리코젠


BMS
(주)비엠에스


CoreIT


accelrys


kobio
Korean
Bioinformation
Center


KiSTi
www.kisti.re.kr
한국과학기술정보연구원
Korea Institute of Science and Technology Information