NIG-OIST
Symposium on

# *Evolutionary Bioinformatics*

August 10th-12th, 2015 A.D.

OIST Seaside House
Okinawa, Japan

Hosted by
National Institute of Genetics (NIG), Mishima, Japan
and
Okinawa Institute of Science and Technology (OIST), Okinawa, Japan

Financially supported by
National Institute of Genetics (NIG), Mishima,
Okinawa Institute of Science and Technology (OIST), Okinawa,
and The Motoo Kimura Trust Foundation for the Promotion of Evolutionary Biology

The 13th Japan-China-Korea Bioinformatics Symposium

# Welcome to Okinawa, Japan!

## Mensôre!

(This means “welcome” in Ryukyu language, which is closely related to Japanese language)

We are pleased to host NIG-OIST symposium on evolutionary bioinformatics as the 13th Japan-China-Korea Bioinformatics Symposium. We are grateful to financial supports from National Institute of Genetics (NIG), Okinawa Institute of Science and Technology (OIST), and The Motoo Kimura Trust Foundation for the Promotion of Evolutionary Biology. We thank Dr. Seungwoo Hwang of Korean Bioinformation Center and Professor Xie Lu of Shanghai Center for Bioinformation Technology for taking care of Korean and Chinese participants. We also thank Ms He Qinwen of Shanghai Center for Bioinformation Technology, Ms Teruya Tomomi of Satoh Lab at OIST, and Mrs. Mizuguchi Masako and Mrs. Iida Ai of Saitou Lab at NIG for their all efforts in many kinds of secretarial works.


**Local Organizing Committee**

Gojobori Takashi, Distinguished Professor, Computational Bioscience Research Center (CBRC), King Abdullah University of Science and Technology (KAUST), Saudi Arabia

Ikeo Kazuho, Associate Professor, Laboratory for DNA Data Analysis, National Institute of Genetics (NIG), Japan

Saitou Naruya, Professor, Division of Population Genetics, National Institute of Genetics (NIG), Japan (Chair)

Satoh Noriyuki, Professor, Marine Genomics Unit, Okinawa Institute of Science and Technology (OIST), Japan

# PROGRAM

## Day 1: Monday, August 10, 2015

13:00-13:30 Registration & Poster preparation

13:30-13:40 Introduction by Saitou (NIG) & Satoh (OIST)

13:40-14:15 Invited Talk 1 by Chen Runsheng, China

Noncoding sequence and precision medicine

14:15-14:50 Invited Talk 2 by Kim Sanguk, Korea

Network analysis to understand the process of metazoan evolution

14:50-15:25 Invited Talk 3 by Satoh Noriyuki, Japan

Horizontal gene transfer, when and how: a case study of tunicates

15:25-16:00 Invited Talk 4 by Zhao Guoping, China

Big BioData vs Small BioCuration

16:00-16:30 Tea Break

16:30-17:05 Invited Talk 5 by Kim Sangsoo, Korea

Predicting the splicing quantitative trait loci

17:05-17:40 Invited Talk 6 by Gojobori Takashi, Japan

Comparative Metagenomics: Diversities of Marine Microorganisms between Japan and Saudi Arabia

17:40-18:15 Invited Talk 7 by Li Yixue, China

Data mining: driven by functional studies for biology

18:15-18:50 Invited Talk 8 by Kim Ryan Woonbong, Korea

Building infrastructure for Korean bioresource information management and cloud-based omics data analysis platform

19:00-20:00 Dinner at Chura Hall, OIST Seaside House 3rd floor

20:00-21:00 Poster Presentation 1

## Day 2: Tuesday, August 11, 2015

9:00- 9:20 Selected Short Talk 1 by Kimura Ryosuke, Japan

A search for genetic variants associated with 3D facial morphology in Japanese

9:20- 9:40 Selected Short Talk 2 by Park Chan Young, Korea
Structural variants from whole genome sequencing of Korean population
9:40-10:00 Selected Short Talk 3 by Wang Zhen, China
Whole genome sequencing of six dog breeds from continuous altitudes reveals adaptation to high-altitude hypoxia
10:00-10:35 Invited Talk 9 by Nakamura Yasukazu, NIG, Japan
Towards better genome annotation
10:35-11:10 Invited Talk 10 by Simakov Oleg, OIST, Japan
Dynamics of genome architecture evolution across metazoans
11:10-11:45 Invited Talk 11 by Xie Lu, China
Identification of protein variants influencing protein abundance levels in breast cancer
11:45-14:00 Lunch
14:00-14:35 Invited Talk 12 by Inoue Ituro, NIG, Japan
Is "thrifty genotype hypothesis" by Neel still valid?
14:35-15:10 Invited Talk 13 by Akashi Hiroshi, NIG, Japan
Inferring ancestral DNA sequences under non-stationary models: methodology and applications
15:10-15:45 Invited Talk 14 by Kim Jaebum, Korea
Ancestral genome reconstruction
15:45-16:20 Invited Talk 15 by Sinclair Robert, OIST, Japan
This is not relevant
16:20-16:40 Tea Break
16:40-18:30 Poster Presentation 2
19:00-21:00 Banquet at Chura Hall, OIST Seaside House 3rd floor

## Day 3: Wednesday, August 12, 2015

9:00- 9:20 Selected Short Talk 4 by Urbanczyk Henryk, Japan
Systematics of "Harveyi clade" bacteria (family Vibrionaceae), using evolutionary bioinformatics
9:20- 9:40 Selected Short Talk 5 by Kim DongHyo, Korea

The change of Genotype-Phenotype relationship can be explained by rewiring of functional module

9:40-10:00 Selected Short Talk 6 by Shi Yi, China
The Building Blocks of the Genome 3D Structure

10:00-10:20 Selected Short Talk 7 by Seo JiEun, Korea
Targeted sequencing using pooled DNA samples to maximize variant discovery power for alzheimer disease susceptibility prediction

10:20-10:40 Selected Short Talk 8 by Li Ying, China
Multi-scale RNA Comparison

10:40-11:05 Invited Talk 16 by Ikeo Kazuho, NIG, Japan
Platform for Drug Discovery from genome sequence to protein structure

11:05-11:40 Invited Talk 17 by Saitou Naruya, NIG, Japan
Diversity of conserved noncoding sequence evolution among eukaryotes

11:50-14:00 Lunch & Closing Ceremony

Removal of posters

15:00-16:00 Visit to OIST (optional)

Symposium will be held at Seminar room of OIST Seaside House 1st floor.

Talk time for invited speakers is 35 min. including questions and discussions.

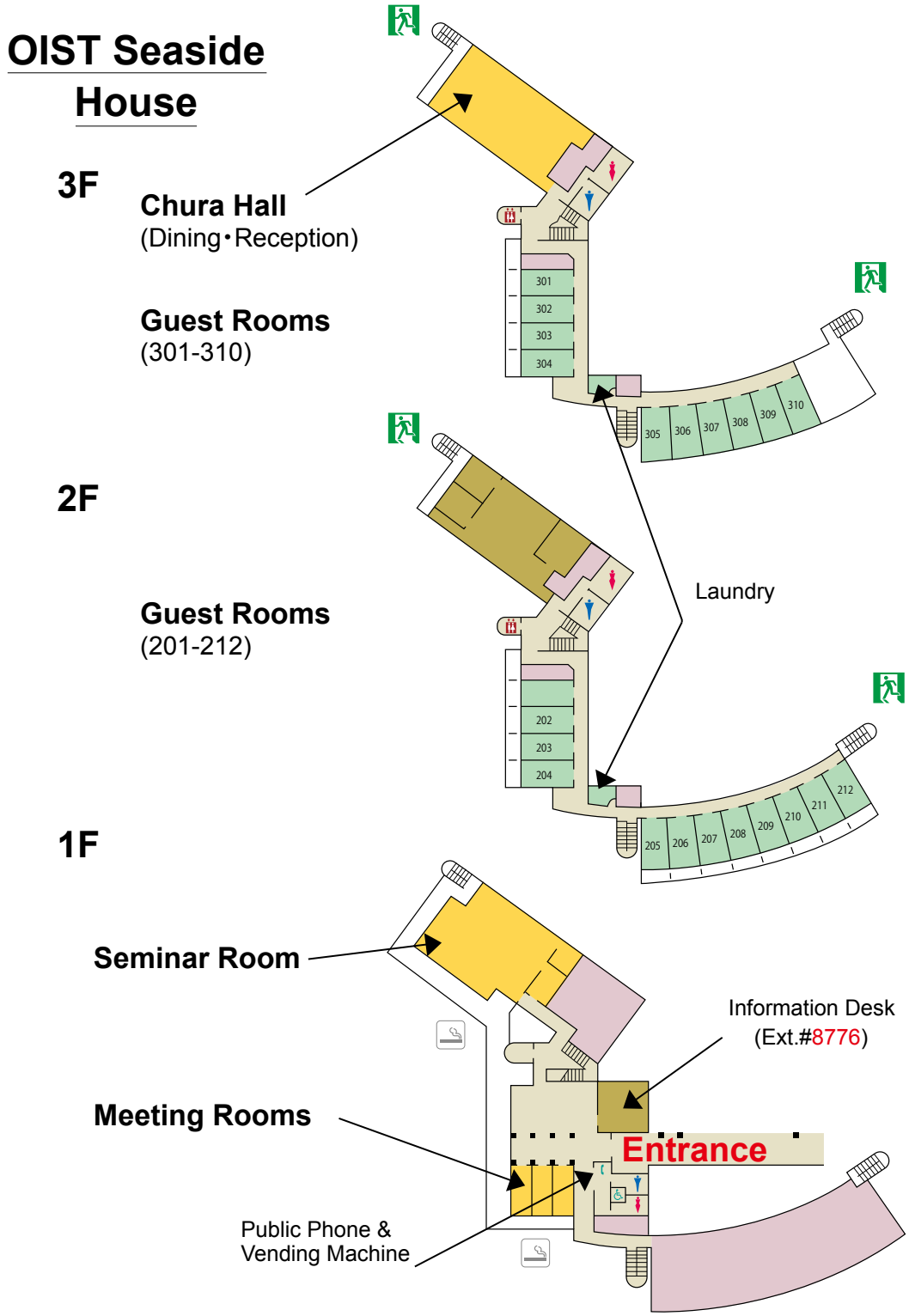
Talk time for selected young investigators is 20 min. including questions and discussions.

Posters will be put at OIST Seaside House first floor. Size of poster is W880mm X H1,170mm.

In this abstract book, all names follow East Asian style; family name (surname) is first, then given name.

Invited speakers will stay at Costa Vista Okinawa Hotel & Spa, and young attendants will stay at OIST Seaside House.

# OIST Seaside House



# PARTICIPANTS

## Invited Speakers

AKASHI Hiroshi, Professor, Division of Evolutionary Genetics, National Institute of Genetics (NIG), Japan

CHEN Runsheng, Professor, Institute of Biophysics, Chinese Academy of Sciences, China

GOJOBORI Takashi, Distinguished Professor, Computational Bioscience Research Center (CBRC), King Abdullah University of Science and Technology (KAUST), Saudi Arabia, and Project Professor, National Institute of Genetics (NIG), Japan

IKEO Kazuho, Associate Professor, Laboratory for DNA Data Analysis, National Institute of Genetics (NIG), Japan

INOUE Ituro, Professor, Division of Human Genetics, National Institute of Genetics (NIG), Japan

KIM Jaebum, Assistant Professor, Konkuk University, Korea

KIM Ryan Woonbong, Director, Korean Bioinformation Center (KOBIC), Korea

KIM Sangsoo, Professor, Soongsil University, Korea

KIM Sanguk, Professor, Pohang University of Sciecen and Technology (POSTECH), Korea

LI Yixue, Professor, Institute of Biochemistry and Cell Biology, Shanghai Institutes of Biological Sciences (SIBS), Chinese Academy of Sciences, China

NAKAMURA Yasukazu, Professor, Genome Informatics Laboratory, National Institute of Genetics (NIG), Japan

SAITOU Naruya, Professor, Division of Population Genetics, National Institute of Genetics (NIG), Japan

SATOH Noriyuki, Professor, Marine Genomics Unit, Okinawa Institute of Science and Technology (OIST), Japan

SIMAKOV Oleg, Molecular Genetics Unit, Okinawa Institute of Science and Technology (OIST), Japan

SINCLAIR Robert, Associate Professor, Mathematical Biology Unit, Okinawa Institute of Science and Technology (OIST), Japan

XIE Lu, Professor and Vice Director, Shanghai Center for Bioinformation Technology (SCBIT), China

ZHAO Guoping, Professor, Institute of Plant Physiology and Ecology, Shanghai Institutes of Biological Sciences (SIBS), Chinese Academy of Sciences, China

# Selected young speakers (*also presents poster)

KIM DongHyo, Graduate student, Pohang University of Science and Technology, Pohang, Korea

KIMURA Ryosuke, Associate Professor, The University of Ryukyu Medical School, Okinawa, Japan

LI Ying, Associate Professor, Jilin University, Jilin, China

PARK Young Chan, Graduate student, Hanyang University, Seoul, Korea

SEO Ji Eun, Graduate student, Seoul National University, Seoul, Korea

SHI Yi, Assistant Professor, Shanghai Jiaotong University, Shanghai, China

URBANCZYK Henryk, Associate Professor, Faculty of Agriculture, Miyazaki University, Miyazaki, Japan (*)

WANG Zhen, Assistant Professor, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China

# Poster presenters

CHO Kwang Hoon, Graduate student, Korean Bioinformation Center, Korea

EUN Seok Chan, Researcher, Seoul National University, Korea

KAWASHIMA Kent Diel, Graduate student, SOKENDAI (Graduate University for Advanced Studies) / National Institute of Genetics (NIG), Japan

LEE Moses, Graduate student, Seoul National University, Korea

LI Hong, Assistant Professor, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, China

LUO Lan, Graduate student, Institute of Computing Technology, Chinese Academy of Sciences, China

MA Lili, Student, College of Life Science and Technology, Huazhong University of Science and Technology, China

MISHRA Neha, Graduate student, SOKENDAI/ NIG, Japan

PARK SunHye, Graduate student, Hanyang University, Korea

QIN Guangrong, Assistant Professor, Shanghai Center for Bioinformation Technology, China

SHOKAT Shayire, Graduate student, SOKENDAI/ NIG, Japan

SOHN Sumin, Graduate student, Ulsan National Institute of Science and Technology, Korea

WANG Yan, Associate Professor, Jilin University, China

YOON Hyejun, Graduate student, Ulsan National Institute of Science and Technology, Korea

YOO Taekyeong, Graduate student, Seoul National University, Korea

UCHIZONO Shun, Graduate student, Kyushu University Graduate School of Systems Life Sciences, Fukuoka, Japan

Invited Talk 1

## Noncoding Sequence and Precision Medicine

CHEN Runsheng

Institute of Biophysics, Chinese Academy of Sciences
Beijing 100101, China

In recent year Translational Medicine, Individualized Drug Therapy and Precision Medicion have became very hot field. A formidable challenge we face in these fields is to understand how genetic information results in the concerted action of gene products in time and space to generate function or to lead to disease. In the case omics data of human such as genome, transcriptome, proteome and so on should be obtain and analyse.

Just as scientists were getting a grasp of large scale genome data, they found that the regions of DNA encoding protein (*i.e.*, what is generally known as 'genes') only occupied a small fraction, not exceeding 3 % of the genome. The remaining 97 % or so of the DNA sequence was mostly without any clear function. At first, researchers were wont to label this DNA as 'noncoding DNA' or simply as 'junk DNA'. After whole genome comparisons were carried out it was discovered that lower organisms such as viruses and bacteria only contained small amounts of 'junk DNA', whereas in higher animals and plants, 'junk DNA' might even make up the major part of the genome. This is to say, as one move from 'lower' to 'higher' organisms, from simple to complex, from cells with low information content to organisms with high information content, the amount of noncoding DNA in the genomes increases. This implies that 'junk DNA' may contain information related to the complexity of organisms. Whole genome comparisons have shown that the number of genes (*i.e.*, genes encoding proteins) in the fly and nematode (14000 to 20000) are only 2-3 times that of yeast (approximately 6000), and only slightly lower than the gene number in human and mouse (approximately 24000). Thus, the increase in gene numbers does not reflect the increase in biological complexity.

The results of large scale transcriptional analyses in recent years suggest that sequences in the noncoding parts of the genome may be expressed in the form of noncoding RNAs, and accumulating evidence suggest that noncoding RNA has important functions. The research on microRNA is the most prominent example. In addition to the 21-24 nucleotide long microRNAs, a number other types of noncoding RNA have been discovered. In the talk some results of noncoding research in our Lab would be introduced.

Invited Talk 2

## Network analysis to understand the process of metazoan evolution

KIM Sanguk

Department of Life Sciences, Pohang University of Science and Technology (POSTECH), Pohang, Korea

A central question in animal evolution is how multicellular animals evolved from unicellular ancestors. Modular architecture is important for the evolution of cellular systems since modular rearrangements facilitate functional innovations and modular insulations provide robustness to perturbations. We investigated domain-domain interactions (DDIs) and domain-linear motif interactions (DLIs) during the course of network evolution and find that DLIs mediate between-module interactions, and that their relative abundance has dramatically increased in metazoan species. Linear motifs have been identified as evolutionary interaction switches since subtle amino acid changes can cause the short sequences in linear motifs to appear and disappear. Our results suggest that subtle changes in linear motifs have contributed to the rewiring of functional modules and, consequently, to functional innovations in metazoan species.

Moreover, we analyzed the evolution of membrane protein domains and find that membrane proteins frequently recruit domains from soluble proteins in metazoan species. Newly incorporated soluble domains became particularly important players in intercellular PPI network. Especially, they are enriched in functions critical for multicellularity, such as cell-adhesion, immune and developmental processes. Our results suggest that domain sharing between membrane and soluble proteins was a major mechanism for generating the panoply of proteins required for cellular cooperation in metazoans.

Invited Talk 3

## Horizontal gene transfer, when and how: a case study of tunicates

SATOH Noriyuki

Marine Genomics Unit, Okinawa Institute of Science and Technology (OIST), Okinawa, Japan

Although horizontal gene transfer (HGT) appears common in bacterial genomes, few examples of HGT have been shown in metazoans. Here we discuss one of typical cases of HGT in metazoans, namely transfer of bacterial gene encoding cellulose synthase into tunicates (urochordates). Tunicates are only animal group that can produce cellulose by themselves. Tunicates are covered by an outer coat, named tunic, most of which components are cellulose. Searching genomes of *Ciona intestinalis* and *C. savignyi* demonstrates the presence of a gene encoding cellulose synthase (CesA). Interestingly, Ci-CesA is composed of both synthase domain and cellulase domain. A mutant called "*swimming juveniles* (*sj*)" was isolated, in which the mutation occurs in CesA, indicating the function of the gene. We are now studying how bacterial CesA gene became functional in ascidians.

Invited Talk 4

# Big BioData vs Small BioCuration

ZHAO Guo-Ping

Shanghai Institutes for Biological Sciences, CAS
Chinese National Human Genome Center at Shanghai
Email: gpzhao@sibs.ac.cn

Since the Human Genome Project, rapid development of high-throughput technologies in Omics along with the massive global efforts in BioBanking and Translational Medicine have pushed Biology into the era of Big Data, more accurately, the **Big BioData**. We have spent decades struggling to collect "enough" biological and biomedical data to understand the mechanism of life. However, when **the Big BioData** overwhelms us today with its huge **Volume** readily expanding from TB to PB and approaching EB without any sign of slowing down the pace and its rapid increase in **Velocity beating the Moore's Law** and reaching **"Real-Time"** with the aid of mobile and non-invasive physiological data detection, are we ready to face the challenge? In fact, we do have the opportunity. Although the **Big BioData** are extremely **Complex** for its **Variety with serious problems in Veracity**, it certainly has its own characteristics significantly different from that of the general Big Data. Among which, we have to emphasize that the intrinsic structure of BioData with respect to both immediately connected information and distantly related datasets, which directly leads it to certain level of **systematic correlation** and **controllable quantitation**.

Obviously, life sciences today need more robust, expressive, computable, quantitative, accurate and precise ways to handle the **Big BioData**. There have been numerous efforts inhandling the complexity of information, integrating the data from very heterogeneous resources, and setting up effective standards to be adopted when facing the **Big BioData.** However, the very basic foundation of these efforts, *i.e.*, the "old" and "boring" job of **Biocuration** seems seriously dragged and even neglected, to certain extent. Therefore, I fully support the comments and proposals made by Doug Howe, Seung Yon Rhee *et al.** in 2008, *i.e.*, *Biocuration,* the field that links biologists and their data urgently needs structure, recognition and support.

Based on and closely related to **Biocuration, Data Mining/Deep Learning** employing **Big BioData** driven by biological questions or functional studies should be the key or at least one of the key directions of future **Bioinformatics**. Thus, **Computational Systems Biology,** which integrates the computational tools with systems biology experimentation must play a central role in **Standardization** of the Big BioData. Guided by these principles and standards, tools and techniques for analyzing Big BioData have been under development and should eventually enable people to convert the massive amount of information into a better understanding about the basic mechanisms of biology/medicine and more precision/effective in disease prevention and treatment.

Invited Talk 5

## Predicting the splicing quantitative trait loci

KIM Sangsoo

Department of Bioinformatics & Life Science, Soongsil University
Seoul, Korea

Genome-wide association (GWA) studies have identified numerous genetic loci associated various traits or diseases. However, most of the GWA signals are non-coding and their molecular underpinning is often elusive. There have been various efforts that have the potential to help understand GWA signals. For example, the ENCODE project has generated a map of regulatory regions, which helps to rationalize the regulatory roles of these non-coding signals. The expression quantitative trait loci (eQTLs) that have been characterized genome-wide are also useful in genetically relating the expression level of a candidate gene. While the eQTL approach is concerned with the overall expression level of a gene, the splicing quantitative trait loci (sQTLs) approach focuses on the relative abundances of transcript variants of a given gene. With the advent of the next-generation sequencing, it is now possible to both discover and characterize alternative splicing forms using a RNA-seq method. If genome-wide genotypes and transcriptome expression levels of the same samples are available, such sQTLs can be sought via a statistical procedure.
Seonggyun Han, a graduate student in my lab, with the help from other members, has developed an R/Bioconductor package called IVAS, which identifies genetic variants affecting alternative splicing. Using IVAS, we looked for sQTLs in three different European populations (CEU, GBR, and FIN) in the 1000 Genomes Project. By combining the results of the three studies in a meta-analysis, we were able to identify a number of sQTLs. Some of them are either found in or overlapped by linkage disequilibrium blocks of known signals in the GWAS Catalog. Their molecular aspects will be presented in this presentation.
This work has been supported by the Industrial Strategic Technology Development Program 10040231, “Bioinformatics platform development for next-generation bioinformatics analysis”, funded by the Ministry of Trade, Industry & Energy.

Invited Talk 6

# Comparative Metagenomics: Diversities of Marine Microorganisms between Japan and Saudi Arabia

GOJOBORI Takashi[1,2]

1. Computational Bioscience Research Center, BESE, KAUST, KSA
2. National Institute of Genetics, Mishima, Japan

Marine metagenomics is a genomic approach in which genomic fragments of any species contained in environmental samples such as a bottle of sea water and a cup of sediment soil are sequenced without morphological identification of those species. This approach is taken to observe ecological and genetic features of a diversity of microorganisms. When this kind of metagenomics is applied for comparative studies between two different locations at different time points, diversities of marine microorganism diversity such as species composition of microorganisms can be used for environmental evaluation of the sea water in addition to understanding of dynamic features of marine microorganism diversity.

Here, we present our on-going projects in which comparative studies of marine metagenomics have been conducted for understanding of marine microorganism diversity between the sea surrounding the Japan islands and the Red Sea in Saudi Arabia. Our preliminary results of the big data analysis in the present studies show that species distributions of microorganisms are substantially different between the two sea regions, manifesting significant characteristics of their respective environmental and ecological situations.

Invited Talk 7

## Data Mining: Driven by Functional Studies for Biology

LI Yixue

Institute of Biochemistry and Cell Biology, Shanghai Institutes of Biological Sciences, Chinese Academy of Sciences, Shanghai, China

Since the Human Genome Project, rapid development of high-throughput technology and the massive global efforts in BioBanking have pushed Biology into the era of Big Data, more accurately, the **B**iology/**O**mics **B**ig **D**ata **(BOBD)**. Its property is similar but not exactly same as that of the common Big Data and actually one can divide these comparative properties into two simple categories. First, it is Not that Big yet! For its **Velocity**, it is fast but not reaching the "real-time" level yet. For its **Volume**, it is just approaching EB level but not too much above that yet. On the other hand, it is Higher in Dimension! For its **Variety**, it is extremely complex ranging from atoms/ molecules to global ecology but with certain level of **systematic correlation**. For its **Veracity**, it is highly vulnerable but with certain level of **controllable quantitation**.

We have spent decades struggling to collect "enough" biological and biomedical data to understand the mechanism of life. However, when **BOBD** overwhelms us, are we ready to face the challenge? The new bottleneck to this problem in biology is how to reveal the essential mechanisms of biological systems by analyzing the **BOBD**.

Life sciences today need more robust, expressive, computable, quantitative, accurate and precise ways to handle the big data, and actually, recent works on this area have already brought remarkable advantage and opportunities, which implies the central roles of bioinformatics and bioinformaticians to the future of the biological and biomedical research fields.

In the **BOBD** era of the life sciences, data is presented in multiple dimensions, which represent the information at various levels of biological systems, including data about genomes, transcriptomes, epigenomes, proteomes, metabolomes, molecular image, molecular pathways, human populations, clinic and medical records. Along with the growing of the BOBD (from Petabyte (PB) to ExaByte (EB)), if scientists can overcome many challenges related to its complexity and its high dimension properties, nobody doubts that it will create huge values. **The core challenges are: how to handle the complexity of information, how to integrate the data from very heterogeneous resources, what kind of principles or standards to be adopted when facing the BOBD.** In that direction, BOBD curation driven by biological questions or functional studies should be the key or at least one of the key measures. Thus, **Computational Systems Biology,** which integrates the computational tools with systems biology experimentation must play a central role in **Standardization** of the BOBD. Guided by these principles and standards, tools and techniques for analyzing BOBD can eventually developed and enable us to translate the massive amount of information into a better understanding of the basic biomedical mechanisms.

As an example, for the "human genome project", it utilized the expertise, infrastructure, and people from 20 institutions and took 13 years of work with over $3 billion to determine the whole genome structure of approximately three billion nucleotides. But now we can sequence a whole human genome for $1,000 and within three days. And, we need only a couple of weeks for analyzing and comparing whole structure of over thousands human genomes supported by current supercomputer.

Invited Talk 8

# Building Infrastructure for Korean Bioresource Information Management and Cloud-based Omics Data Analysis Platform

KIM Ryan Woonbong

Korean Bioinformation Center
Korea

The 21st century is predicted to be the era of bioindustry by futurologists and OECD authorities. In particular, driven by remarkable advances in next-generation sequencing and bioinformatics technology, human genome sequencing projects are being expanded nationwide and personalized genomics is quickly becoming a reality. In accordance with this trend, bioinformatics is being established as key technology for a wide range of bioindustry as well as for genome-based basic research in health, medicine, environment, and pharmaceutics.

An important core for the era of bioeconomy is the acquisition and utilization of biological resources, which provide the primary material for R&D and technological advances. Korean Bioinformation Center (KOBIC) has been designated as the National Center for Biological Research Resource Information in 2009 and since then, we have been working toward establishing an infrastructure to efficiently archive and distribute biological information across various ministries. This has led to the creation of Korean Bioresource Information System (KOBIS).

As the national center for managing biological information, our mission is to provide capabilities and resources to manage and standardize the explosively growing amount of data on bioresource, genome, and biological research data from national R&D grants by developing a systematic and integrative approach. To effectively utilize such data, we construct state-of-the-art infrastructure that is equipped with a variety of analysis tools and workflows. We also ensure that such infrastructure is accessible to all qualified researchers and the general public by carrying out outreach activities, such as Korea Biological Resource Centers Alliance, workshops and seminars, bioinformatics training course, research support service, and newsletters. We are here to build infrastructure to provide innovative web portals and analysis tools, thus fulfilling our principal objective to foster advances in biological research and technology.

Invited Talk 9

# Towards better genome annotation

NAKAMURA Yasukazu

Genome Informatics Laboratory
National Institute of Genetics, Mishima, Japan

Advances in Next-Gen-Sequencing technologies provide a way to obtain comprehensive information about genomes at low cost. More and more nucleotides are being sequenced with the NGS machines, and the bulky data must be processed at full throttle to archive and share for the scientific community. INSDC's traditional core service is sequence archive, however, under present conditions, an outrageous amount of data from NGS's need to be managed by support system provided large-scaled data centers. Such a Next-Gen-System will consist of versatile data analysis pipelines, rich and accurate reference datasets, and large data archive with use of huge storage system. In 2009, DDBJ started a new archive: DDBJ Sequence Read Archive (DRA), a database for row data from NGS. DRA collaboratively exchanges data with the Sequence Read Archive (SRA) of NCBI and ENA [http://trace.ddbj.nig.ac.jp/dra/]. In order to support raw sequence data submission and analysis, we developed the DDBJ read annotation pipeline [http://p.ddbj.nig.ac.jp]. This pipeline provides number of de facto standard mapping and assembly tools. In this talk, I will introduce the current status, updates, problems and future plans of DDBJ and INSDC. Also, I would like to touch on current status of CyanoBase and TogoAnnotation project [http://genome.microbedb.jp/cyanobase], which is a community-based high throughput genome annotation system for the NGS era.

Invited Talk 10

## Dynamics of genome architecture evolution across metazoans

Oleg SIMAKOV

OIST, Okinawa, Japan

In the talk I will describe recent advances in the field of metazoan comparative genomics, the insights into the ancestral metazoan/bilaterian genomes, as well as their diversification. Particular focus will be on the variation in synteny and transposable elements and their effect on development in the recently sequenced genomes of a cephalopod *Octopus bimaculoides* and the hemichordates *Saccoglossus kowalevskii* and *Ptychodera flava*.

Cephalopods (octopus, squid, cuttlefish, nautilus) belong to one of the most species-rich, yet genomically under-sampled, ancient metazoan phylum, the lophotrochozoans. Their convergently-evolved camera eyes, epibolic gastrulation, as well as complex behavior have fascinated evolutionary biologists. Through comparison to other lophotrochozoan genomes [1], we find that *Octopus* lineage has undergone a significant syntenic reshuffling, for example losing well-known linkages, such as Wnts, Forkhead, or Hox. The loss of synteny correlates with distinct (in the time domain) expansions of transposable elements (TEs). We differentiate between different classes of TEs contributing to either genome reshuffling or evolution of novel regulation and tissue-specific expression. In contrast to the high rate of synteny diversification in cephalopods, hemichordates, located at the base of the deuterostomes, show much better gene family, synteny conservation. We show examples of such conservation and its impact on our understanding of the deuterostome ancestor.

[1] Simakov, O., F. Marletaz, S. J. Cho, et al. "Insights into bilaterian evolution from three spiralian genomes." Nature 2013

Invited Talk 11

## Identification of protein variants influencing protein abundance levels in breast cancer

ZHANG Menghuan and XIE Lu

Shanghai Center for Bioinformation Technology
Shanghai 201203, China

Proteins are central to cellular processes. Differences in protein function or abundance may be responsible for phenotypic differences and be associated with human diseases. A large number of studies in human are about genetic variants influencing the transcriptome. In contrast, studies about genetic variations influencing the proteome in human are very few. Here we present the analysis of protein variants affecting protein abundance (pQTLs) with mass spectrometry (MS) data of 36 breast cancer samples from CPTAC. Unlike other studies of pQTLs, here variations were directly identified in protein level. A total of 21,772 tryptic digested peptides representing 8,061 proteins were identified and quantified, among which 223 variations of 137 proteins were found. Then, we identified 80 variations influencing the expression level of protein themselves, and 87 variations influencing the expression level of direct interacting partners in breast cancer. Furthermore, protein domain regions were mapped. We hope to provide a better understanding of the genetic impact on the proteome in breast cancer, for evaluating potential biomarkers and therapeutic agents.

Invited Talk 12

## Is "thrifty genotype hypothesis" by Neel still valid?

INOUE Ituro

Division of Human Genetics, National Institute of Genetics
Mishima, Japan

Genetic basis of common diseases is very complicated because both environmental factors including personal life-styles and genetic factors play substantial and interactive roles in the etiology. One of the examples is type 2 diabetes (T2D) and other metabolic diseases including obesity. James Neel, a pioneer of human genetics, called T2D as 'geneticist's nightmare" because of its complexity also proposed the "thrifty genotype hypothesis" claiming genetic etiologies of diabetes mellitus and other metabolic diseases according to population history more than 60 years ago. The basic concept of the hypothesis is that the genetic background leading to disadvantageous T2D today was advantageous in the past environments. This hypothesis is plausible to understand disease causality regarding natural selection, therefore seems to be positively accepted by the research community.

The recent explosive genetic studies of diabetes identified more than 70 loci with sufficient sample size, which enable to re-examine the "thrifty" genotype hypothesis. In this workshop, I will demonstrate the most updated concept of common diseases after accumulating of the genetic components.

Invited Talk 13

# Inferring ancestral DNA sequences under non-stationary models: methodology and applications

Akashi Hiroshi[1], Matsumoto Tomotaka[1], Yang Ziheng[2]

[1] Division of Evolutionary Genetics, National Institute of Genetics (NIG) Mishima, Japan

[2] Department of Genetics, Evolution and Environment, University College London, London, United Kingdom

Inferred sequences of ancestral species provide opportunities to test many hypotheses of genome evolution. However, reconstructed ancestral sequences may yield spurious results from systematic biases caused by inconsistencies between reconstruction models and actual evolutionary processes. Here, we explore new methods to study complex scenarios of molecular evolution, including fluctuating nucleotide composition, in large genomic datasets. We implement methods to correct for biases and use computer simulation to evaluate inference reliability when substitution processes are not at steady-state.

We simulated base composition evolution on a gene tree with G+C content fluctuations among lineages and compared nucleotide substitutions counts during simulation with inferred counts. Large, systematic biases resulted from (i) parsimony inference or likelihood using single best reconstructions, (ii) the assumption of stationarity when base composition fluctuates, and (iii) the use of likelihood models that do not adequately describe the substitution process. In the scenarios examined, a non-stationary general time reversible (GTR) model accurately inferred substitution counts, even in cases of complex, lineage-specific processes. Parameter-rich models may be limited to large data sets. However, because genome-scale data are becoming increasingly available for close relatives of model organisms, we hope that our method will facilitate a number of research areas (*e.g.* genome-wide compositional changes involved in temperature adaptation). We are currently exploring how to employ ancestral inference for simultaneous analyses of within and between species genetic variation.

Invited Talk 14

# Ancestral genome reconstruction

KIM Jaebum

Department of Animal Biotechnology, Konkuk University
Seoul 05029, Korea

Since the first introduction in *The Origin of Species* (1859) by Charles Darwin, it has been a long belief for many biologists that allied species share a common ancestor. However, the genome organization of species descended from the same ancestor is highly variable. Therefore, in order to better understand the relationships among species, it is necessary to find the answers of questions, such as how genomes have been reorganized and what kinds of genome rearrangements have appeared during evolution. 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 eventually lead to the accumulation of a huge volume of genome assemblies of many species, which is crucial for solving puzzles such as the shape of ancestral genomes, and the mode and extent of their evolution.

Here we will discuss the state-of-the-art approaches for studying genome evolution, such as Zoo-FISH technique and ancestral genome reconstruction algorithms in the fields of molecular cytogenetics and comparative genomics respectively. Zoo-FISH, or cross-species chromosome painting, uses painted probes specific for whole chromosomes to detect conserved blocks among species. And many computational approaches for reconstructing ancestral genomes have been developed by taking advantage of the accumulation of genome sequences of many species, which was enabled by next-generation sequencing technologies together with genome assembly algorithms. We will also discuss what have been found about the changes of genome organization during evolution in the vertebrate lineage by using the aforementioned approaches.

Invited Talk 15

## This is not relevant

SINCLAIR Robert

Mathematical Biology Unit, OIST, Japan

Genomic and metagenomic analysis often involves di-, tri-, tetra-nucleotide or general k-mer frequency analysis. I show that there are exact mathematical (combinatorial) relationships between these frequencies for circular genomes, plasmids or molecules, whether composed of DNA or RNA. As a rule of thumb, it seems that 25% of the frequencies for any class of k-mer are dependent upon the others. This has consequences for statistical analysis, since the number of degrees of freedom usually expresses the number of independent variables. Also, I suggest this knowledge will be of value in more cleanly extracting phylogenetic signal from composition analysis, since the mathematical relations I present are most certainly not influenced by any physical, chemical or evolutionary process. In other words, these exact relationships tell us what is not relevant, and allow us to focus on what is.

Invited Talk 16

# Platform for Drug Discovery from genome sequence to protein structure

IKEO Kazuho

Laboratory for DNA Data Analysis, National Institute of Genetics (NIG)
Mishima, Japan

New world of research in the field of life science was opened by the next Generation sequencing (NGS). The application range is wide from variation study of genome to complicated phenomena like population dynamics etc. It also brings a big change of the style in the style of research; bioinformatics becomes basic research technology in the field at now.

The project titled "Platform for Drug Discovery, Informatics, and Structural Life Science" inherits the outcomes of the previous projects in structural biology, consolidates the technology and promotes sharing those core technologies for drug discovery and related life sciences. Through these activities, the project aims to realize a revolutionary process connecting drug and medical seeds to pharmaceutical products.

For this purpose, the project is proceeded in the following three centers; (1) Analysis Center, (2) Regulation Center, and (3) Information Center.

We are now developing and providing two systems under this activity. We, as the data analysis center, provide an integrative platform to analyze data from NGSs. This platform provides data analysis systems (Maser and various viewers (Genome Explorer, gGraph, iGene) which we developed originally. On Maser systems, the data analysis pipelines are constructed by the workflows that made of a combination of existing NGS data analysis tools.

For protein structure and function or omics, VaProS, VAriation effect on PROtein Structure and function, is a new data cloud for Structural Life Science and is the core technology to lead the collaboration between the discipline in Structural Biology and the whole Life Sciences. VaProS has been developed around the Integrated Structural Biology Database at Institute for Protein Research in Osaka University, together with the selected outcomes from Protein 3000 Project,Targeted Proteins Research Program, Genome Network Project and Cell Innovation Project.

Invited Talk 17

# Diversity of conserved noncoding sequence evolution among eukaryotes

HETTIARACHCHI Nilmini and SAITOU Naruya
Division of Population Genetics, National Institute of Genetics (NIG) ,
Department of Genetics, Graduate University for Advanced Studies (SOKENDAI)
Mishima, Japan

Conserved noncoding sequences (CNSs) are enriched in regulatory sequence elements. We conducted a whole genome analysis on plant CNSs and identified them to be GC rich (Hettiarachchi et al. 2014). Babarinde and Saitou (2013) reported mammalian CNSs to be GC poor. This heterogeneity in GC content might be related to varying sequence features of regulatory elements in different lineages. Since animals and fungi are sister groups, we investigated the features of fungi lineage common CNSs in order to determine the evolutionary origin of low GC content of mammalian CNSs. This investigation was further extended to discover the sequence features of lineage common CNSs of invertebrates, non-mammalian vertebrates with the intention to answer varying regulatory features of different lineages. Currently we have identified that plant, fungi, invertebrate lineage CNSs are predominantly GC rich whereas vertebrates are GC poor. We also found that this GC content feature is directly related to their location in the genome. High GC CNSs showed a tendency to be found in heterochromatin regions whereas low GC CNSs shows a tendency to locate in open chromatin. The sudden transition of high GC content of CNSs from plant, fungi and invertebrates to low GC content in vertebrates and does the structural architecture of CNSs have a direct correlation with its function are some of the questions we intend to answer in the near future.

Hettiarachchi N., Kryukov K., Sumiyama K., and Saitou N. (2014) Lineage specific conserved noncoding sequences of plant genomes: their possible role in nucleosome positioning. Genome Biology and Evolution, vol. 6, no. 9, pp. 2527–2542.

Babarinde I. and Saitou N. (2013) Heterogeneous tempo and mode of conserved noncoding sequence evolution among four mammalian orders. Genome Biology and Evolution, vol. 5, pp. 2330-2343.

Selected Short Talk 1

## A search for genetic variants associated with 3D facial morphology in Japanese

KIMURA Ryosuke

Graduate School of Medicine, University of the Ryukyus
Okinawa, Japan

To identify genetic factors associated with facial morphology, we performed a genome-wide association study using 704 Japanese individuals living in Okinawa. We obtained 3D images of facial surface using a hand-held 3D scanner. To accomplish a fine phenotyping, we conducted homologous modeling, which generated corresponding points between facial data based on a polygon model containing 2,596 semi-landmarks. Then, principal component analysis (PCA) was performed on the facial surface data. Genome-wide SNP typing was carried out using DNA microarrays. PCA was also performed on genome-wide SNP data and, based on the genomic PCs, samples were divided into mainland Japanese and Ryukyuans. Association analyses using a linear regression model were performed for these two populations separately, where each facial PC was used as the dependent variable and sex, age, BMI, genomic PC1 and PC2 were as covariates. The results for the two populations were integrated using meta-analysis. Finally, we found SNPs that were significantly ($P < 5\times10^{-8}$) associated with a facial PC representing facial flatness. Further replication studies will be needed to obtain a robust conclusion.

Selected Short Talk 2

## Structural variants from whole genome sequencing of Korean population

PARK Young Chan

Hanyang University
Seoul, Korea

Deep whole genome sequencing is now widely used for understanding how structural variants are different among individuals and between different populations as the cost of the DNA genome sequencing is plummeted compared to the era of 1000 Genomes Project (1KGP). Although, The 1000 Genomes Project has penetrated wide perspective into aggregating characters of 2,500 individual genomes from 25 populations from world using statistical methods and algorithms, the limitation of early version of short reads NGS technology and the high cost, such as deprecated ABi solid and illumine GA, hinder to generate high coverage data which brings about coverage bias of genetic variants inadvertently. Moreover, the population in the Korean peninsula, known as originating from Mongol and Sothern Asia historically, was not reflected in the 1KGP in any. The Korean Personal Genome Project (KPGP) were performed for whole genome sequencing of 62 Korean individuals by above 30x coverage reclaiming gaps between Korean specific variants and other populations to complement the results from low-coverage sequencing of 1KGP and the deep coverage data contribute to detecting functional variants in high accuracy. We analyzed 62 individuals that is maintaining by Genome Research Foundation using current computational tools.

Selected Short Talk 3

## Whole genome sequencing of six dog breeds from continuous altitudes reveals adaptation to high-altitude hypoxia

WANG Zhen

Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences
Shanghai, China

The hypoxic environment imposes severe selective pressure on species living at high altitude. To understand the genetic bases of adaption to high altitude in dogs, we performed whole-genome sequencing of six dog breeds living at continuous altitudes along the Tibetan plateau from 800 to 5,100 m. Comparison of the breeds from different altitudes reveals the strongest signals of population differentiation at the locus of EPAS1. Especially, four novel non-synonymous mutations specific to high-altitude dogs are identified at EPAS1, one of which occurred at a quite conserved site in the PAS domain. The association testing between EPAS1 genotypes and blood-related phenotypes on additional high-altitude dogs reveals that the homozygous mutation is associated with the decreased blood flow resistance, which may help to improve hemorheologic fitness. Interestingly, EPAS1 was also identified as a selective target in Tibetan highlanders, though no amino acid changes were found. Thus, our results not only indicate parallel evolution of humans and dogs in adaption to high-altitude hypoxia, but also provide a new opportunity to study the role of EPAS1 in the adaptive processes.

Selected Short Talk 4

## Systematics of "Harveyi clade" bacteria (family *Vibrionaceae*), using evolutionary bioinformatics

URBANCZK Henryk[1], OGURA Yoshitoshi[2], HAYASHI Tetsuya[2]

[1] Faculty of Agriculture, University of Miyazaki, Japan
[2] Faculty of Medical Sciences, Kyushu University, Japan

In this study, we describe the use of evolutionary bioinformatics to analyze diversity of bacteria in the so-called "Harveyi clade" (family *Vibrionaceae*, Gammaproteobacteria). The clade consists of 14 closely related species, which are difficult to accurately classify using classic bacterial taxonomy approaches. In order to resolve taxonomic ambiguities within the "Harveyi clade" draft genome sequences of 27 representative strains were determined and analyzed. The sequencing included type strains of seven species within the "Harveyi clade." Whole genome sequence data of additional 35 strains were obtained from public databases. Analysis of the genome sequence data revealed a clear case of synonymy between *Vibrio owensii* and [*V. communis*], and provided evidence that strains of the so-called '*beijerinckii*' lineage should be taxonomically classified as *V. jasicida*. Whole genome sequence data was also used in taxonomic description of a novel species in the "Harveyi clade", *V. hyugaensis* sp. nov.

In addition to resolving taxonomic ambiguities in the "Harveyi clade", we also estimated the number of interspecies recombination events occurring within core genomes of six species in the clade, as well as the number of intraspecies recombination events occurring between 11 strains of *V. campbelli*. Bacteria used in this analysis were selected from a collection of strains isolated in the last 90 years, from various environments worldwide. We found that the number of detected interspecies recombination events was low among all six species. The low frequency of interspecies recombination events was evident when analyzing strains isolated over 80 years apart, from different hemispheres, or from different ecologies, as well as in strains isolated from the same geographic location within a short time frame. In contrast, the number of identified intraspecies recombination events between some (but not all) strains of *V. campbellii* was disproportionally high. The high number of recombination events was detected between *V. campbellii* strains that have significant temporal (over 18 years) and geographical (over 10,000 km) differences in their origins of isolation. Results of this study reveal a remarkable stability of "Harveyi clade" species, suggest that ecology of bacteria had little influence over the frequency of interspecies recombination in the core genomic regions, and give clues about the origins and persistence of species in the clade.

Selected Short Talk 5

## The change of Genotype-Phenotype relationship can be explained by rewiring of functional module

KIM DongHyo

Pohang University of Science and Technology, Pohang, Korea

Phenotype-genotype mapping is crucial to understand the relationship between human disease symptoms and associated genes. But, it is difficult to understand the phenotype consequence of human disease mutations because molecular level study of human subjects cannot be easily done because of ethical reasons. Therefore, we rely on phenotype transfer of gene functions from model organisms to human symptoms based on orthology-function conjecture. It is assumed that phenotypes of uncharacterized genes in human have been predicted from the phenotypes of well annotated organism, such as mouse. However, there are many orthologous genes that have drastically different phenotypes between species.

In this study, we hypothesized that the phenotype conservation and diversification are originated from 'modular' genetic evolution based on a notion that a phenotype is determined by the group of genes rather than single gene. To prove this, we mapped orthologs of mouse and human gene, and measured the genotype-phenotype relationship based Online Mendelian Inheritance in Man (OMIM) and Human Phenotype Ontology (HPO) for human genes and Mouse Genome Informatics (MGI) for mouse genes. A quantitative phenotype conservation score was designed based on phenotype ontology comparisons across the species.

In our surprise, we found that many orthologues have a quite different phenotype ontologies that hinder direct phenotype transfer between orthologues genes between two species. We further investigated the genes with phenotypical divergent orthologues and found that they have changed many partner genes in a same functional module. Specifically, gene in metabolic process tended to change phenotypes between two species. Our study highlights that genes change phenotype during evolution based on genetic modularity rather than sequence divergence which give an important lesson to decipher phenotype annotation transfer between different species.

Selected Short Talk 6

## The Building Blocks of the Genome 3D Structure

SHI Yi

Shanghai Jiaotong University
Shanghai, China

Genome-wide proximity ligation assays allow the identification of chromatin contacts at unprecedented resolution. Several studies reveal that mammalian chromosomes are composed of topological domains (TDs) in sub-mega base resolution, which appear to be conserved across cell types and to some extent even between organisms. Identifying topological domains is now an important step towards understanding the structure and functions of spatial genome organization. However, current methods for TD identification demand extensive computational resources, require careful tuning, and/or encounter inconsistencies in results. In this work, we propose an efficient and deterministic method, TopDom, to identify TDs, along with a set of statistical methods for evaluating their quality. TopDom is much more efficient than existing methods and depends on just one intuitive parameter, a window size, for which we provide easy-to-implement optimization guidelines. TopDom also identifies more and higher quality TDs than the popular directional index algorithm. The TDs identified by TopDom provide strong support for the cross-tissue TD conservation. Finally, our analysis reveals that the locations of housekeeping genes are closely associated with cross-tissue conserved TDs.

Selected Short Talk 7

## Targeted sequencing of pooled DNA samples to maximize variant discovery power for Alzheimer disease susceptibility prediction

SEO Jieun[1], MOOK-JUNG Inhee[1], CHOI Choi[1]

[1]Department of Biomedical Sciences, Seoul National University College of Medicine, Seoul, Korea

Alzheimer's disease (AD), the most common form of dementia, is often classified on the basis of the age at onset; late-onset AD (LOAD) is defined as onset of more than 65 years. Due to the increasing life expectancy, medical, social and economic burden that LOAD imposes on modern societies raises serious concerns. Therefore, understanding the genetic complexity of LOAD will provide valuable insights into the pathogenesis of AD. Attempts to identify common genetic variants that are associated with LOAD susceptibilities have generated a number of genes with questionable functional implications. Therefore, we set up a targeted sequencing of these genes to determine if they carry previously uncharacterized rare variants. To maximize the study power, each of 3 samples from 3,113 controls, 606 mild cognitive impairment cases and 845 LOAD cases were pooled and sequenced together by Ion torrent panel sequencing at a mean coverage of >500. Prior to the production phase, we optimized the variant calling pipeline through a serious of pilot experiments and developed a pooled sequencing data based rare variant calling platform that offers ~95.4% sensitivity and ~99.7% specificity. This study will enable efficient and powerful targeted-genome analysis on common complex diseases.

Key word: Alzheimer’s disease, Late onset AD, pooled DNA sequencing, target sequencing

Selected Short Talk 8

## Multi-scale RNA Comparison

LI Ying

Jilin University, Jilin, China

In recent years, RNAs play more and more important biological roles. Computing the similarity between two RNAs contribute to better understanding the functional relationship between them. But due to the long-range correlations of RNA, many efficient methods of detecting protein similarity do not work well. In order to comprehensively understand the RNA's function, the better similarity measure among RNAs should be designed to consider their structure features (base pairs). Current methods for RNA comparison could be generally classified into alignment-based and alignment-free. In this short talk, we introduce a novel multi-scale RNA comparison which can capture the local and global difference between the information of sequence and structure of RNAs. Compared with the popular RNA comparison approaches in term of the non-coding RNA families classification and RNA mutation analysis, the multi-scale RNA comparison method is validated to be effective.

Poster 02

## Business Insight on Bioinformatics

EUN Seok-Chan, MD, PhD, MBA

Department of Plastic and Reconstructive Surgery, Seoul National University College of Medicine, Korea

Bioinformatics is an interdisciplinary field that is used in the development and storage of data, thus helping in the analysis, organization and retrieval of the biological data. The bioinformatics sector is segmented into medical bio informatics, animal bioinformatics, agriculture bioinformatics, academics and others. The Medical bioinformatics was and is forecast to be the highest revenue generating market through 2020. This was due to the increasing application of drug discovery and development using bioinformatics tools, which were also used in the preventive medicine and gene therapy. The geography segment is segmented into North America, Europe, Asia Pacific and etc.

The key strategies opted by the market players are Product Launch, Agreement and Collaboration. The market analysis is done based on Porter's five forces model and Value chain analysis. According to the Porter's five forces model the bargaining power of the supplier's is low due to the presence of many bioinformatics service providers and the threat from internal substitutes of this market is moderate. Asia's investment in sequencing will allow the country to build a valuable intellectual property portfolio because new discoveries will be made at a furious pace. Asia's sequencing power has the potential to tip the balance in innovation.

Poster 03

**Simulating Pathogen Molecular Evolution in Epidemics: A Graph Approach**

KAWASHIMA Kent Diel

Department of Genetics, SOKENDAI (The Graduate University for Advanced Studies) and Division of Evolutionary Genetics, National Institute of Genetics, Mishima, Japan

The emergence of new pathogen genotypes is a product of interactions at two levels: pathogen molecular evolution occurring within the host during an infection and transmission of pathogens to susceptible hosts in the population. To examine the underlying relationship between these two overlapping processes, forward-time simulations based on standard population genetic and ecological models have been developed. However, these models assume random mating and a well-mixed population, ignoring the importance of host connectedness and heterogeneity, and disconnect the effects of within-host with between-host evolution.

Here I present a graph-based discrete-time model that simulates overall pathogen molecular evolution by integrating within-host and between-host dynamics using autonomous agents connected as a network. Each node represents a single host that takes on one of three states – susceptible, infected, or removed – based on whether a pathogen has traversed the node or not. When a node becomes “infected”, pathogen sequences replicate to carrying capacity, with subsequent generations selected based on a fitness value for each genotype. While in the infected state, a random sample of pathogens can be transmitted to adjacently connected susceptible hosts based on a per-capita susceptibility probability. Thus the collection of nodes represents the entire host population being modeled while the graph of all nodes and edges at a given point in time shows a snapshot of the host contact network.

I plan to use my simulation to test whether the underlying host contact network configuration influences phylogenetic inference and tests for selection. Because some hosts are more connected to the network than others, I speculate that pathogen molecular evolution occurring at these “superspreaders” distorts genotype distributions affecting phylogenetic reconstruction and detection of natural selection.

Poster 08

## Lineage-specific genome evolution in *Drosophila melanogaster* subgroup

MISHRA Neha

Department of Genetics, SOKENDAI (The Graduate University for Advanced Studies) and Division of Evolutionary Genetics, National Institute of Genetics, Mishima, Japan

Evolutionary forces such as mutation, drift, and selection contribute to long-term evolution but the time-scale and magnitude of their variation is not well understood. We used patterns of codon bias as a model to study the effects of these forces on a genome. We studied lineage-specific genome evolution in seven *D. melanogaster* subgroup species. We inferred ancestral states and assigned substitutions to 10 lineages. Several lineages showed strong departures from the equilibrium codon bias. Variability among genes in most lineages is consistent with changes in selection intensity and/or background substitution patterns. Several lineages showed strong departures from the equilibrium codon bias. Variability among genes in most lineages is consistent with changes in selection intensity and/or background substitution patterns. These findings suggested that magnitude of forces governing base composition at synonymous sites may have varied frequently in a lineage-specific manner. To identify the underlying evolutionary mechanisms of lineage-specific changes in base composition, we plan to compare lineage-specific changes in base composition of introns to that of synonymous sites.

Poster 09

## Whole genome based variation analysis study in 62 Koreans

PARK SunHye, PARK YoungChan, PARK Kiejung, KOH InSong

Department of Biomedical Informatics, Hanyang University, Seoul, Korea

The completion of the Human Genome Project (HGP) in 2003 and the advent of new DNA sequencing technology, called next generation sequencing technology (NGS) have led the development of human genome based research. The cost-effective, faster and large-scale NGS technology facilitates population-scale international human genome projects such as the 1000 Genomes Project, UK10K, the 100,000 Genomes Project. These projects contribute to more comprehensive understanding of human genomic variations.

In order to investigate population specific variants, more specific population variation reference set is needed. However, these projects contain only a limited number of human populations (not including Koreans). The goal of this study is to make a Korean specific variation reference set required for Korean specific whole genome based association research.

Here, we have discovered Korean variation information by analyzing 62 Korean whole genomes from the Korean Personal Genome Project (KPGP) initiated by the Genome Research Foundation (PGI). The KPGP whole genome data of 62 individuals are publicly available and can be downloaded from the KOBIC FTP site (ftp://ftp.kobic.re.kr/pub/KPGP/). We utilize the method of mapping on reference genome (GRCh37). Our goal is to identify not only single nucleotide polymorphism (SNP) and Insert/Deletion (InDel) but also structural variation (SV) and copy number variation (CNV). We expect that the result of the present study, i.e. the Korean variation reference information, helps to study Korean specific genomic medicine more effectively.

Poster 11:

## Genetic diversity of Kazakhstan camel population and its evolutionary relationship with the Arabian camel breed

SHOKAT Shayire

Department of Genetics, SOKENDAI (The Graduate University for Advanced Studies) and Division of Population Genetics, National Institute of Genetics, Mishima, Japan

The genus Camelus contains two species: one-hump camel (*Camelus dromedary*) which inhabits the Arabia and Africa, and two-hump camel (*Camelus bacterianus*) which inhabits the Central Asia. However one-hump camels are not only found in Afro-Arabia, but are also inhabitants of Kazakhstan in central Asia. Although it is believed that one-hump camels originated in Arabia, there has been no in-depth study on the comparison of one-hump camel in Arabia and central Asia. To investigate the possible origin of one hump camels as well as the evolutionary relationship between Arabian and Central Asian populations, we determined the sequences of mitochondrial D-loop regions of 17 Kazakhstan camels, including 11 one-hump, 3 two-hump, and 3 hybrid camels. The sequences were analyzed together with the available camel sequences. Our phylogenetic study supports that the Arabian one-hump camels were the ancestral population. Also, the phylogenetic tree shows that the Kazakhstan one-hump camels donot form a single cluster, which probably suggests that Kazakhstan one-hump camel populations are not homogenous. In addition, we confirmed that wild camel (camel bactrianus ferus) and domestic two-hump camel (bactrian camel) are separate lineages. Furthermore, previous studies have demonstrated the usefulness of camelid microsatellite loci as a genetic tool for the study of one-hump and two-hump camels. We would like to further study the genetic diversity and relationships among Kazakhstan camel populations using microsatellite DNA markers.

Poster 16

## Characterizing polymorphism in sugar taste sensitivity in a natural population of *Drosophila melanogaster*

UCHIZONO Shun

Graduate School of Systems Life Sciences, Kyushu University
Hakozaki, Fukuoka 812-8581, Japan

Animals have to evaluate food nutrients essential for survival and taste sensitivity is an important primary factor. Genetic changes in taste sensitivity might drive speciation of feeding habitat, which is one of the key steps in evolution. Here we examined genetic variation in taste sensitivity to sugars, which are essential nutrients for flies, among inbred lines derived from a natural population of *Drosophila melanogaster* (DGRP). Two-choice preference tests that paired glucose with fructose, sucrose, or trehalose revealed the extensive variation in sugar preferences among the inbred lines and that sugar taste sensitivity is polygenic in the wild-derived inbred population. We then selected two strains that showed opposing preferences for glucose and fructose and performed proboscis extension reflex tests for tarsus chemosensilla of a foreleg and electrophysiological recordings from labellar taste sensilla in response to glucose and fructose. Results indicated that sensitivity to fructose is responsible for the opposing preferences and revealed the existence of an insensitive strain rather than merely low responsive to fructose in l-typed labellar chemosensilla. Genetic analyses showed that high sensitivity to fructose is autosomal dominant over low sensitivity and that multiple loci on chromosomes 2 and 3 might control fructose sensitivity. We thus focused on eight gustatory receptor (*Gr*) genes on chromosome 2 or 3, which are candidates for sugar receptor genes, and confirmed if the genes are involved in variation of fructose sensitivity by complementation tests. Results suggest that one or some members of the *Gr64* gene family are implicated in fructose sensitivity. These studies should also contribute to understand the molecular mechanism of the sugar reception in *Drosophila* and may lead to reveal the evolutionary significance of variation in taste sensitivity.

## Poster Presentations: No abstract

Poster 01: CHO Kwang Hoon, Korean Bioinformation Center, Korea

Poster 04: LEE Moses, Seoul National University, Korea

Poster 05: LI Hong, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, China

Poster 06: LUO Lan, Institute of Computing Technology, Chinese Academy of Sciences, China

Poster 07: MA Lili, College of Life Science and Technology, Huazhong University of Science and Technology, China

Poster 10: QIN Guangrong, Shanghai Center for Bioinformation Technology, China

Poster 12: SOHN Sumin, Ulsan National Institute of Science and Technology, Korea

Poster 13: WANG Yan, Jilin University, China

Poster 14: YOON Hyejun, Ulsan National Institute of Science and Technology, Korea

Poster 15: YOO Taekyeong, Seoul National University, Korea

======= Poster of Selected Short Talk Speaker ========

Poster 17 (Selected Short Talk 4): URBANCZYK Henryk, Faculty of Agriculture, Miyazaki University, Miyazaki, Japan