Large-scale analysis of life data (Theme 4, Satoru Miyano, GL)

In order to focus on more K computer-oriented subthemes which should mutually enhance and create synergy, the group was reorganized. Satoru Miyano remained as the group leader and two PI's Yutaka Akiyama (Tokyo Institute of Technology) and Hideo Matsuda (Osaka University) continued their subthemes. Until March of 2013, Miyano's mission was to manage this group, but from April of 2013, Satoru Miyano joined and started the research. Our biomedical targets are cancer, obesity, and human metagenome. Recent studies revealed that all these three targets are mutually related in our diseases (Yoshimoto S et al. Obesity-induced gut microbial metabolite promotes liver cancer through senescence secretome. ***Nature***. 499(7456):97-101, 2013). The group is organized as shown in Figure 1. We summarize the development and contributions of 2013 fiscal year.

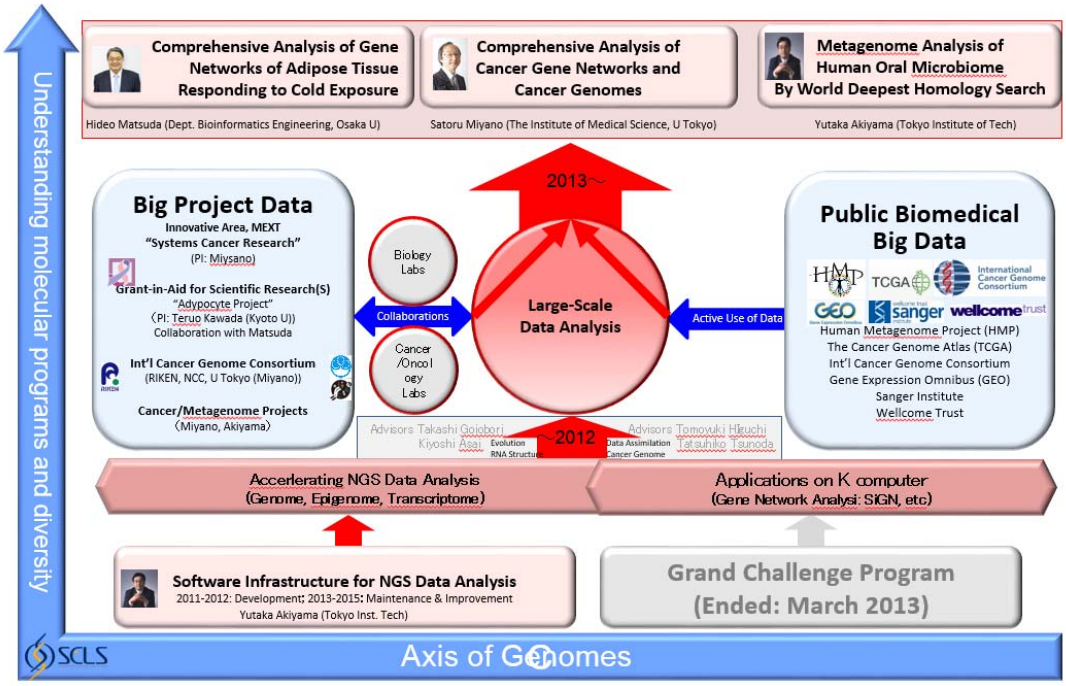

Figure 1: Group organization and subthemes.

1. Large-Scale Comprehensive Analysis of Systems Disorders in Cancer: From Genomes to Gene Networks (2013 -) Satoru Miyano (The University of Tokyo)

In the study of rug resistance in Cancer, we analyzed 728 cancer cell line gene expression profile data related to IC50 of 101 drugs by the software application SiGN-L1. About 600 cell lines showed clear differences in drug resistance/sensitivity. In this very large-scale drug resistant/sensitivity gene network analysis, we computed about 70,000 gene networks each of which has about 13,000 genes as nodes, and revealed 402 dmillion gene-gene causalities in 600 cancer cell

lines for 101 drugs. Then we developed a series of mathematical methods to mine this huge amount of causalities with respect to drug resistance/sensitivity. The analysis suggested diversity of cancers against drugs including candidate target genes (Figures 2 and 3). This study also revealed a weak point (handling of outliers) of SiGN-L1. With a new idea, we got into the process of developing a new software towards 2014. Furthermore, the following issues are in progress: We extended "The Cancer Network Galaxy" (TCNG) (http://tcng.hgc.jp) by using SiGN-BN HC Bootstrap for an EGF-related gene set (1,520 genes) together with 30,261 samples. Finally we finished computing of 250 gene networks. This is being expanded. The next is the development of Genomon-Fusion on K computer (Figure 4) for comprehensive detection of fusion-genes from RNA-seq data. 780 samples of CCLE (Broad Institute, Cancer Cell Line Encyclopedia) were completely analyzed by Genomon-Fusion on K computer and the Human Genome Center supercomputer toward implementation on K computer. A useful tool to extract expression modules from evelopment of EEM (Extraction of Expression Module) was implemented on K computer (Niida A. Gene set-based module discovery decodes cis-regulatory codes governing diverse gene expression across human multiple tissues. ***PLoS One***. 5(6):e10910, 2010). It was applied to gene expression profile data of esophageal cancer and discovered modules regulated by *NFE2L2* (Figure 5). We also developed a cancer evolutions simulation system BEP and obtained the initial observations on intra tumor heterogeneity. The observations give interesting insights on multi-regional colon cancer whole exome analyses (Figure 6). Software tool for biovisulization "Multilayer Heatmap" was also developed.

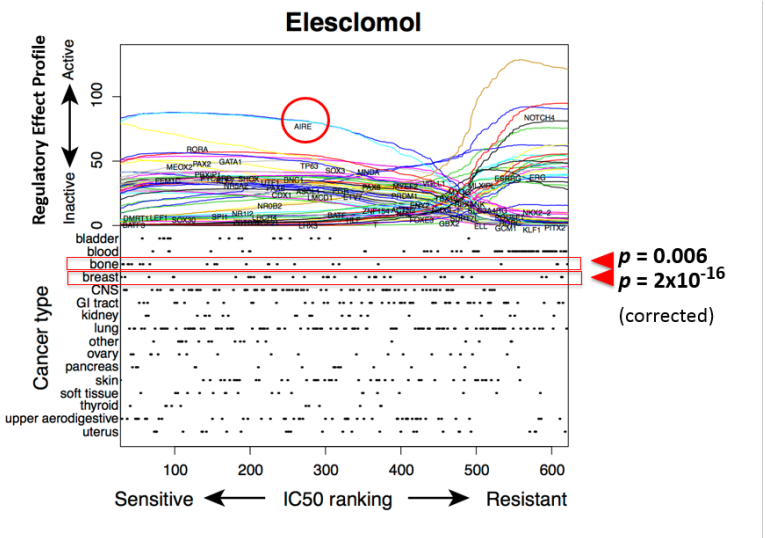

Figure 2: This is a mapping of cancer cell lines on the IC50 score line. For example, gene ARIE has a strong regulatory effect in Elesclomol sensitive cell lines, but it becomes weak for drug resistant cancer. The lower panel shows differences among cancer types. Bladder cancer is mostly sensitive to Elesclomol while blood caner is resistant. But we do not see such difference for other cancer types. Such analyses are going now.

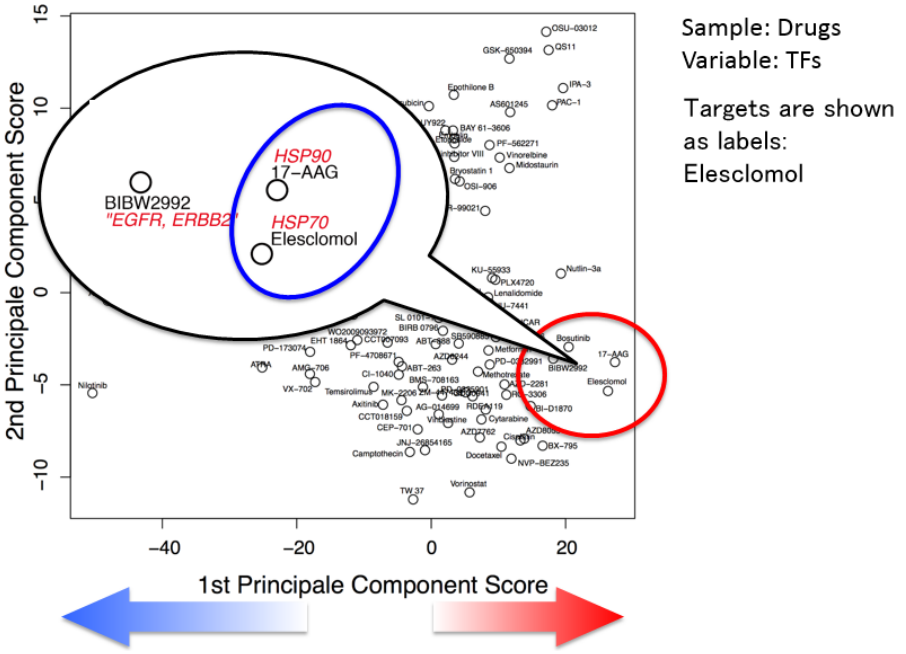

Figure 3: This is a projection of 101 drugs on the first and second principal component space, where regulator genes correspond to variables and drugs correspond to samples. Then we see that these 3 drugs have high 1st component values. Elesclomol is an approved anticancer drug whose target is a heat shock protein HSP70.

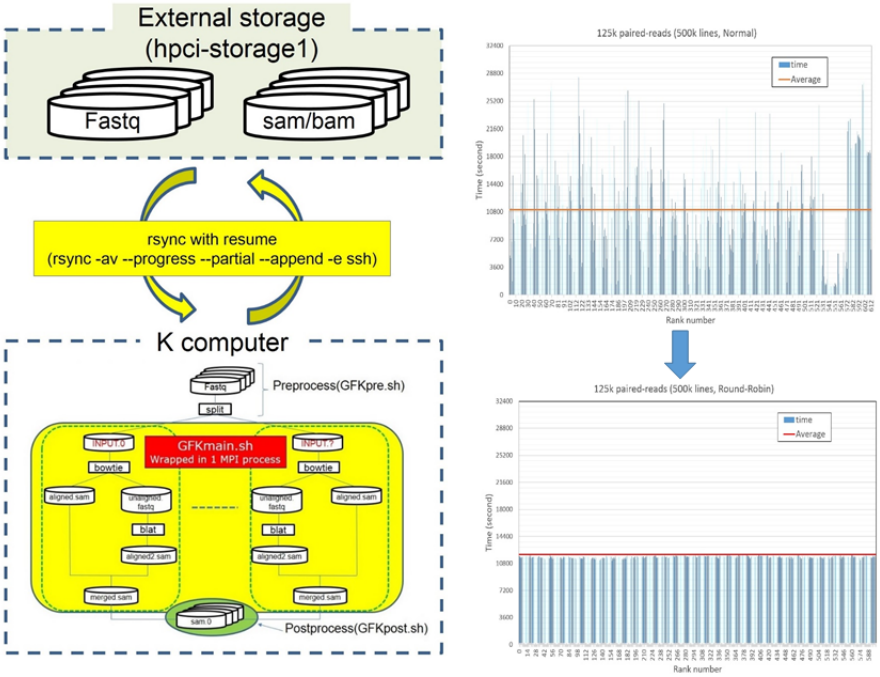

Figure 4: Overview of Genomon-Fusion for K computer (left). Equalization by round-robin (right).

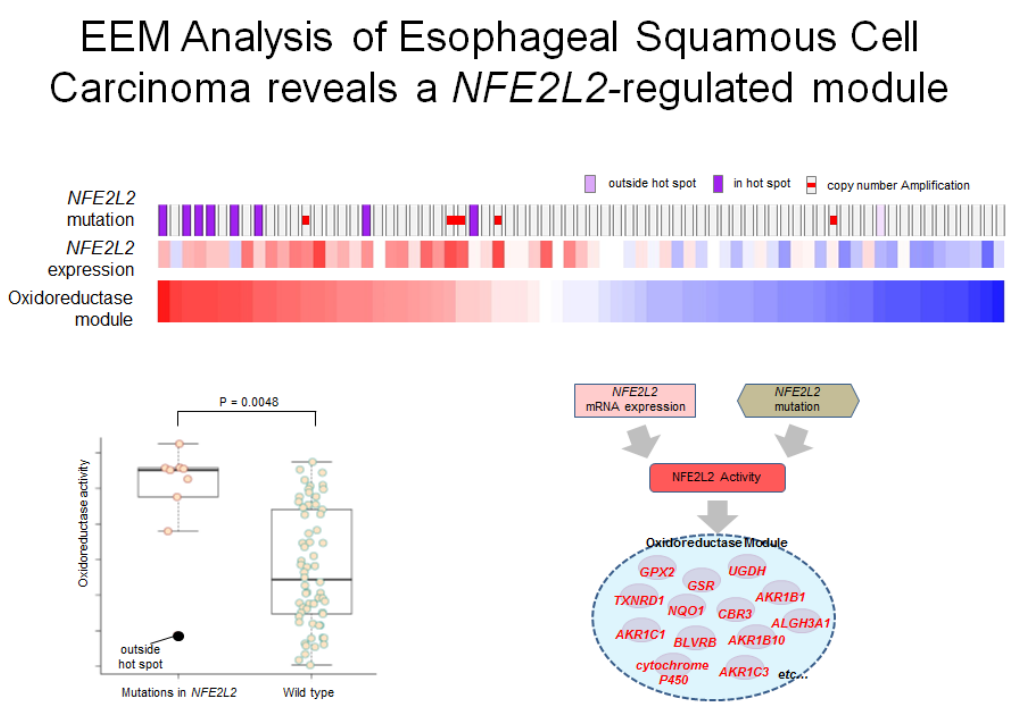

Figure 5: EEM identifiedmModules regulated by *NFE2L2.*

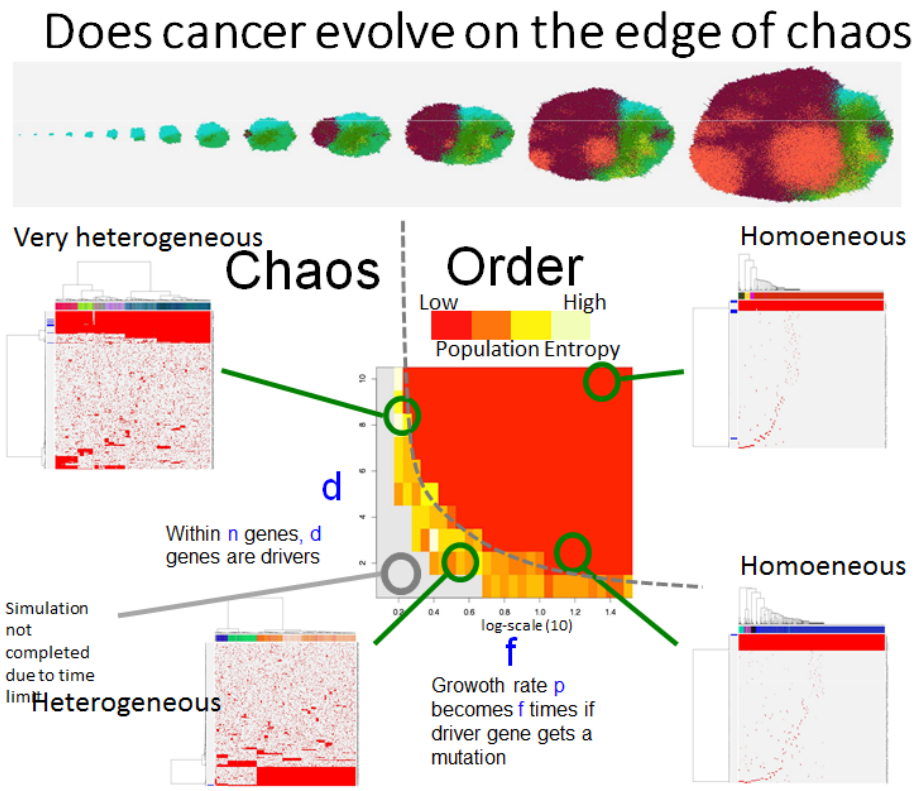

Figure 6: Simulation of cancer evolution.

2. Comprehensive Analysis of Gene Networks of Adipose Tissues Responding to Cold Exposure (2011 -) Hideo Matsuda (Osaka University)

In this study, we investigate the cold exposure responses of three kinds of adipocyte tissues; white adipocyte tissues under exposure to 4℃, white adipocyte tissues, brown adipocyte tissues. For this purpose, we developed a software BENIGN devising large-scale biomolecular network analysis techniques, and applied it to comprehensive analysis of gene networks of adipose tissue responding to cold exposure in collaboration with Professor Teruo Kawada, Graduate School of Agriculture, Kyoto University. The homosapiens has two kinds of adipose tissues. One is white adipose tissue whose function is energy storage. The other is brown adipose tissue that appears in a very small amount in adult human but in some amount in baby. Recent studies revealed that BAT is identified in adult humans by PET-CT scans. The function of BAT is heat production and its heat production is 100 times as much as skeletal muscle. It is known that this heat production is activated by UCP1 in mitochondrion. It is considered that the decrease of BAT is the one of the reason for obesity. WAT and BAT are in different cell lineages. Recently, the third avatar of adipose tissue was found. That is called “beige adipose tissue” (IWAT). It is observed that cold exposure causes transdifferentiation from WAT to beige adipose tissue (Figure 7). The aim of this study is to identify the molecular mechanism as gene networks that controls this transdifferentiation.

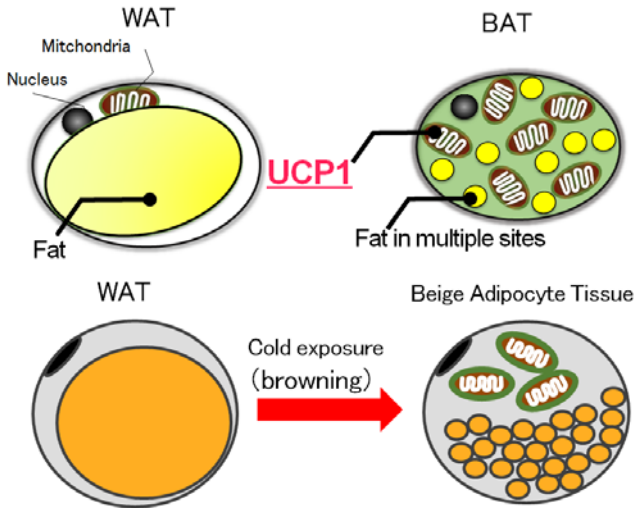

Figure 7: White adipocyte tissue (WAT), brown adipocyte tissue (BAT), and transdifferentiation from WAT to Beige Adipocyte Tissue

Kawada produced gene expression profiles of mice after 2, 8 and 16 days from cold exposure (4 degrees Celsius) stimulus and mice without stimulus and gene networks starting from the adipocyte marker genes (Seed Networks) were estimated with BENIGN. Network analysis revealed that regulatory edges (activate (red) and inhibit (blue)) around UCP1 (yellow) seem to be very different between brown and beige adipose tissues (Figure 8). This also suggested a possible new pathway to induce UCP1 in beige adipose tissue. As a result, a new induction mechanism of thermogenesis in beige adipose tissue is strongly suggested:

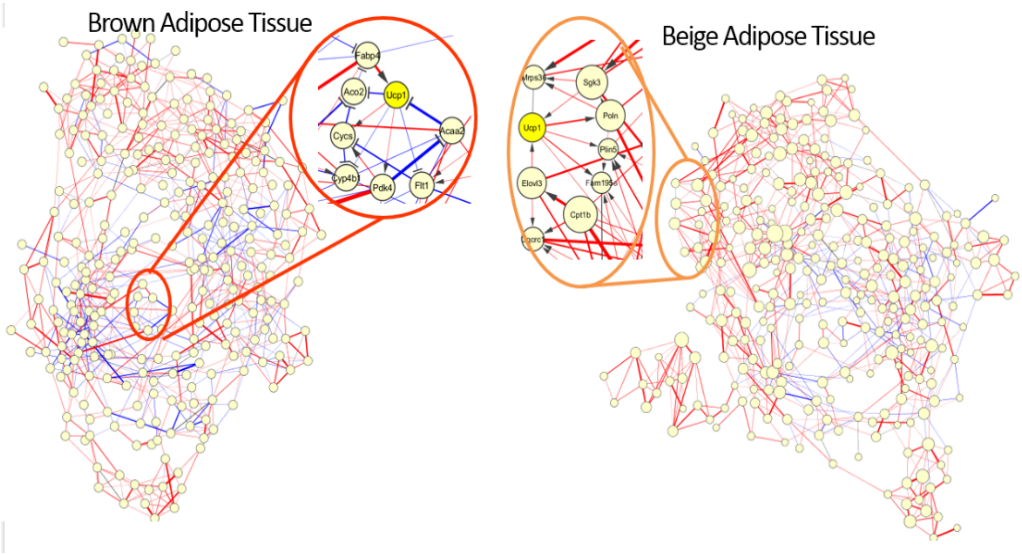

Figure 8: A new induction mechanism of UCP1 in beige adipose tissue is strongly suggested.

There have been published more than 80 papers on brown adipocyte energy production mechanisms and more than 2,400 papers have been published on inflammation mechanisms and pathways (Figure 9). However, there is no paper connecting these two mechanisms. In this study a novel mechanism connecting these two mechanisms. The details are being thoroughly validated (Figure 10). We are now investigating the effects of microRNAs to this mechanism.

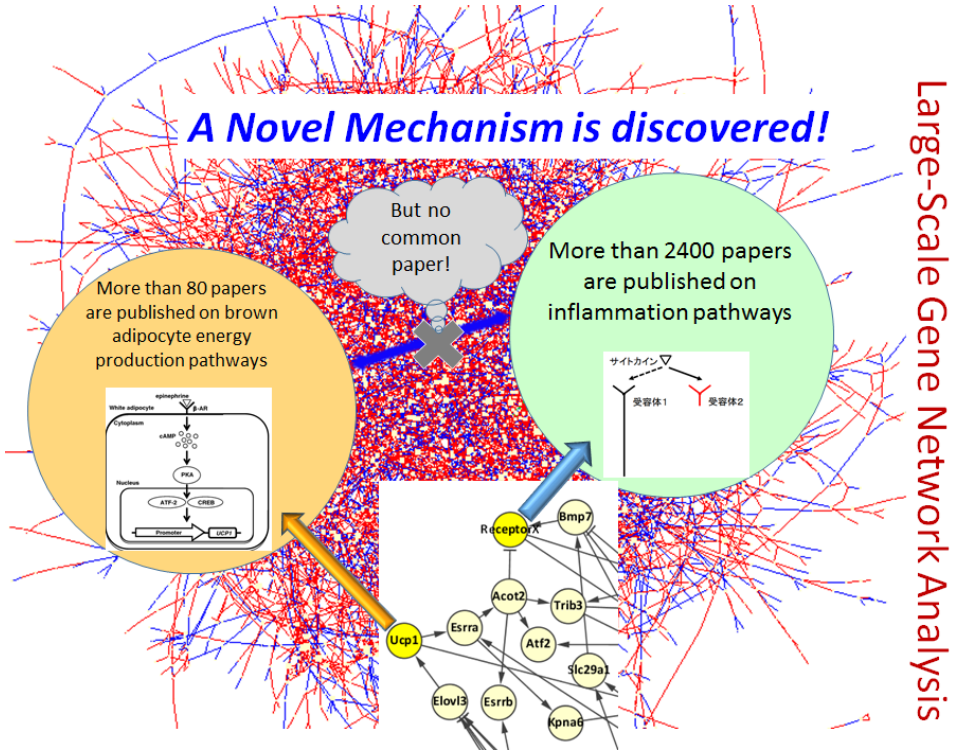

Figure 9: No research has been conductive connecting adipocyte energy production and inflammation mechanisms. Large-scale gene network analysis successfully discovered the connection.

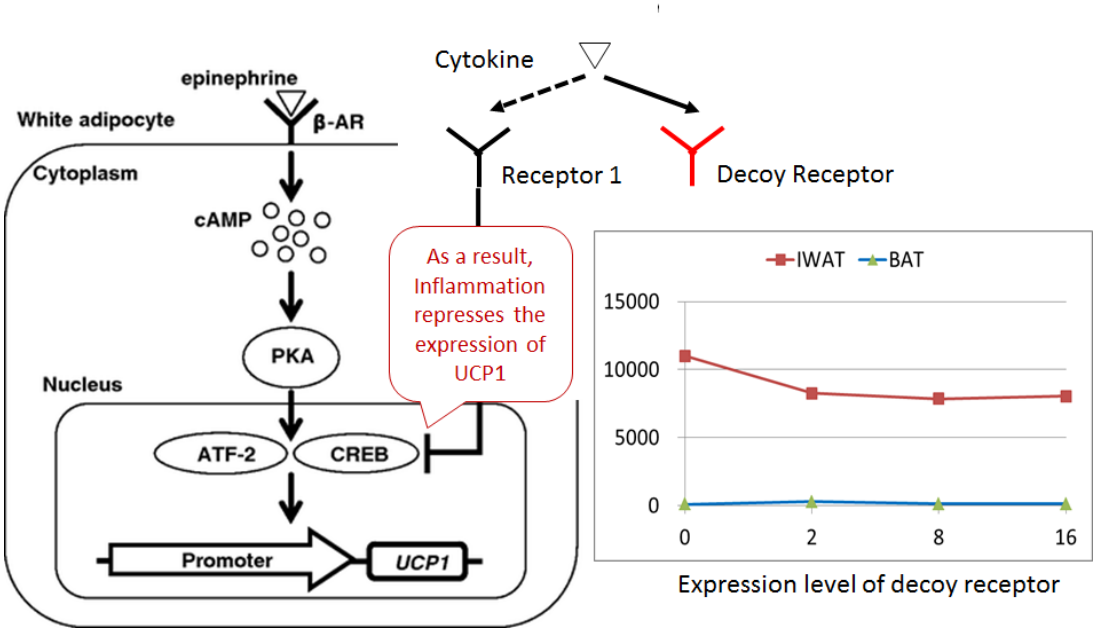

Figure 10: A new mechanism of repression of UCP1.

3. Metagenome Analysis of Human Oral Microbiomes by World Deepest Homology Search (2011 -) Yutaka Akiyama (Tokyo Institute of Technology)

The first contribution is the development of GHOST-MP, a parallel software application for large-scale metagenome analysis on K computer. The performance of GHOST-MP was tested using 80,000 nodes of K computer. Compared with the performance of the last year it is increased by 37% this year. Then we analyzed with GHOST-MP the data from Human Microbiome Project, especially, human oral flora metagenome data (Table 1). 2,610,000,000 reads sequenced from 418 samples ranging over 9 oral sites, were analyzed with GHOST-MP and the results are interpreted using KEGG Orthology. Comparative analysis revealed the distribution of oral flora with respect to similarity and dissimilarity.

Table 1 Metagenome data of oral flora by Human Metagnome Project

Oral Site	# of samples	read counts ($x10^6$)
Attached Keratinized Gingiva	6	361
Buccal Mucosa	121	7478
Hard Palate	1	54
Palatine Tonsils	6	373
Saliva	5	278
Subgingival Plaque	8	517
Supragingival Plaque	128	7965
Throat	7	393
Tongue Dorsum	121	8708
Total	418	26131

A notable achievement in the development of GHOST-MP is a drastic reduction of I/O load. As shown in Figure 11, the scalability is good by 10,368 nodes but from 20,736 nodes the scaling became slower. We need to investigate the reason of this fact. We also investigated the trade-off-probem “data transfer to K computer” and “data analysis on K computer” for metagenome data analysis. As a conclusion, it is shown that the workflow of “transferring data to K computer then analysis” is rational. Finally, transplantation of Genomon-exome to K computer is undergoing and we found some problems to tackle.

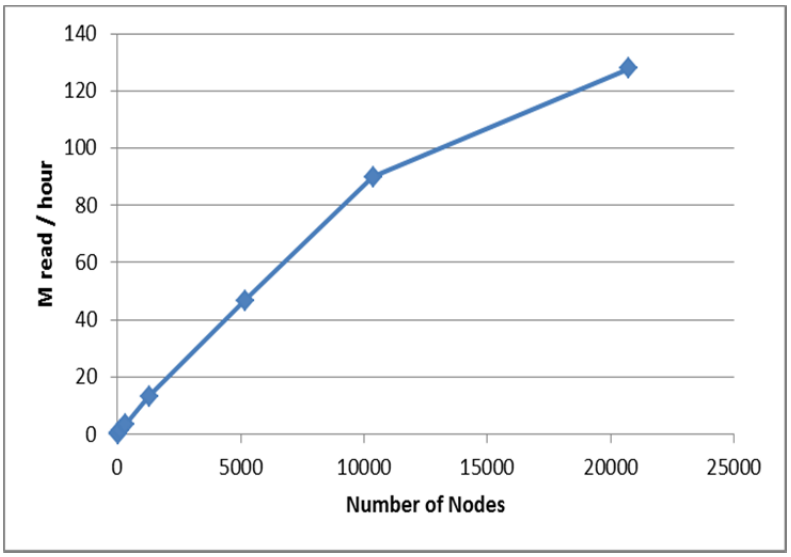

Figure 11: Scaling data of GHOST-MP