”What is Life ?　Catalogue (Omics) and Build (Synthetic biology) “

Date: January 8th, 2020

1. Prof. Jacques Corbeil, Laval University

‘High throughput mass spectrometry and machine learning approaches for biomarker discovery’

2. Prof. Minoru Kanehisa, Kyoto University

‘Toward understanding the origin and evolution of cellular organisms’

3. Prof. Shigeyuki Yokoyama, RIKEN

‘Structural Genomics and Proteomics’

4. Prof. Masayuki Yamamura, Tokyo Institute of Technology

‘A prospect on the appropriate usage of mathematical models in new biology’

5. Prof. Chieh-Chen Huang, National Chung Hsing University

‘Reverse TCA cycle: from the origin of life to establish cell factory’

6. Prof. Yasunori Aizawa, Tokyo Institute of Technology

‘Exploring Dark Matter of the Human Genome　~ Introduction of SYNTHETIC GENOMICS ~ ’

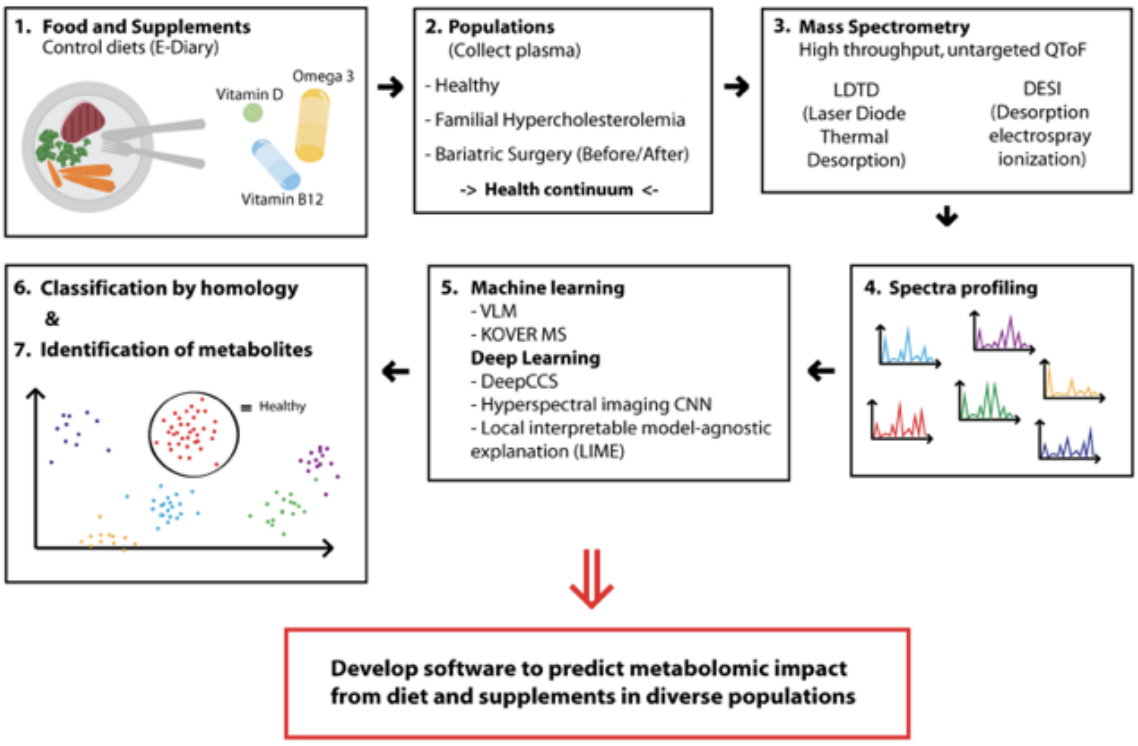
High throughput mass spectrometry and machine learning approaches for biomarker discovery

Jacques Corbeil Ph. D.
Department of Molecular Medicine, Big Data Research Centre, Laval University

The food and supplements we ingest are important determinants of our health because they modulate many key physiological processes including our microbiome and provide essential nutrients. A systematic and comprehensive metabolomic analysis of the impact of food and supplements on our health is highly warranted. What is the fate of the food we eat and to what extent can it be monitored in our plasma as a function of health or disease markers or weight loss? Can we evaluate and gain some insight on the effects of therapies and diets on specific metabolomic diseases and states using plasma metabolites? Can we effect meaningful changes in the health status by providing certain foods and supplements? How is it different between men and women? All these questions will be addressed in this presentation.

Metabolomics is an emerging field of "omics" research specializing in the near-global analysis of small molecule metabolites found in living organisms. Its applications are already being seen in a broad range of disciplines. There are 2,891 endogenous small molecule metabolites that were detected and quantified in human blood (Wishart et al. 2018 and http://www.hmdb.ca) but this estimate is growing as methods become more sensitive and bioinformatics algorithm predict additional ones (Zamboni et al. 2015). The full spectrum of endogenous metabolites potentially expected in the human blood is estimated to be 23,913 (Wishart et al. 2018). Moreover, metabolites, being the small molecules of life, are the real effectors of the phenotypes regulating our physiology. In this sense, it is easier to predict behaviour when directly measuring them instead of proxies such as genes, transcripts or proteins. Obviously, adding these additional omics can make the predictor more robust even more so if we have additional metadata associated with the participants.

Figure 1: Schematic representation of the research process.



High throughput mass spectrometry and machine learning approaches for biomarker discovery

We have already conducted a study on 211 participants that were fed a traditional Mediterranean diet or remained on a controlled North American diet for a 4-week period in a crossover design. All foods and drinks were prepared by food technicians and were provided to participants. Plasma samples were collected after the end of each 4-week diet. Following mass spectrometry peak processing and analysis using only LDTD in the positive ionization mode, 21644 ions were detected. 37 peaks (m/z) had higher relative abundances in samples of participants on the Mediterranean diet. We found that 14 of these 37 features were undetectable in participants following a North American diet. From the 14 features, we putatively identified using FoodDB (http://foodb.ca), four metabolites respectively found in fish, broad bean, nutmeg, and various spices, that correlated with foods comprised specifically in the Mediterranean diet. We can ascertain compliance with a 93% confidence level.These biomarkers identified through machine learning algorithms implementation and eventually, with more data using unsupe rvised deep learning techniques, could be employed to assess diet compliance to a Mediterranean diet in an unbiased manner. Moreover, a closer examination could lead to individual profiling of features essentially evaluating the effect at the participant level as opposed to a cohort effect or a mean effect in a tested sample. We used artificial intelligence (AI) and specifically decision tree and random forest algorithms to evaluate these features for class prediction.

This methodology will bring a completely new dimension to the evaluation of the benefits of certain foods, diets and supplements in both health and disease states. Being able to monitor in an unbiased fashion the food intake of patients provides a tremendous gain in patients' management for both diets and mitigating diseases impact. Our approach is presently focused on food intake and diseases associated with severe metabolomic imbalances. However, the methodology can be applied to all diseases were metabolomic balance is a key component of homeostasis and health but dysregulated in diseases and states such as inflammation, infection, autoimmune diseases and cancer. Another important aspect is to monitor patients in a longitudinal fashion adding another vector to the analysis. In conclusion, we have a robust process to investigate disease states and how to compensate and mitigate their impacts. The longitudinal survey of metabolites in plasma of patients coupled with deviation for normalcy algorithms is the ultimate personalized medicine procedure to detect early diseases. Food intake, lifestyle (level of exercises), environment and genetic are the major determinants of our health. It is imperative to be able to better monitor the impact of our diet on our health. There is a great deal of confusions on the benefits of diets, supplements and specific foods on our health. We propose to measure the levels of thousands of metabolites to more precisely characterize and correlate their presence to specific health status.

Wishart DS, Feunang YD, Marcu A, Guo AC, Liang K, Vázquez-Fresno R, Sajed T, Johnson D, Li C, Karu N, Sayeeda Z, Lo E, Assempour N, Berjanskii M, Singhal S, Arndt D, Liang Y, Badran H, Grant J, Serra-Cayuela A, Liu Y, Mandal R, Neveu V, Pon A, Knox C, Wilson M, Manach C, Scalbert A. HMDB 4.0: the human metabolome database for 2018. Nucleic Acids Res. 2018 Jan 4;46(D1):D608-D617. Doi: 10.1093/nar/gkx1089. PubMed PMID: 29140435; PubMed Central PMCID: PMC5753273.

Zamboni N, Saghatelian A, Patti GJ. Defining the metabolome: size, flux, and regulation. Mol Cell. 2015 May 21;58(4):699-706. Doi: 10.1016/j.molcel.2015.04.021. Review. PubMed PMID: 26000853; PubMed Central PMCID: PMC4831058.

Toward understanding the origin and evolution of cellular organisms

Minoru Kanehisa
Institute for Chemical Research, Kyoto University, Uji, Kyoto, Japan

In the traditional view, the genome is a blueprint of life containing all necessary information that would make up an organism. In my view, the genome specifies only the molecular building blocks, while the cell, the basic unit of life, contains information about how they interact and react to form a system. What we inherit is not just the genome, but the entire cell, and there is a cellular continuity of the germline leading to the origin of life. Unless we uncover the underlying information systems, how they have been developed in the cell and how they have been transmitted along the germlines of different species, we will be unable to decipher the genome. From this perspective, knowledge of cellular functions and other high-level features of organisms has been captured from experimental observations reported in published literature and organized in the KEGG database. By integrating the molecular wiring diagrams encoded in the cell with the molecular building blocks encoded in the genome for all available cellular organisms, KEGG has become a reference resource for biological interpretation of genome sequences and other high-throughput data. In addition to enhancing such practical values, my goal is to develop KEGG into a knowledge base toward understanding basic principles of biological systems, such as coevolution of genomes and information systems. I will discuss our attempt to understand coevolution of genomes and metabolic networks.

Kanehisa, M., Toward understanding the origin and evolution of cellular organisms. Protein Sci. 28:1947-1951 (2019). https://doi.org/10.1002/pro.3715

Structural Genomics and Proteomics

Shigeyuki Yokoyama, RIKEN

The molecular functions of proteins are based on their three-dimensional structures. The international projects of structural genomics and proteomics started in 2000 to establish a comprehensive view of the variety of protein structures as a basis to understand protein functions (*1*). Japan had promoted the structural genomics and proteomics projects, by elucidating the large-scale mapping of the protein structure space. The first effort was conceptualized as early as in 1995, and began with the Protein Folds Project and the Structurome Project (the Whole Cell Project by Seiki Kuramitsu) at RIKEN in 1997 (*2*,*3*). In response to the success of genome sequencing projects, the worldwide cooperation in structural genomics project program was initiated. Launched in 2002, the National Project on Protein Structural and Functional Analyses (NPPSFA or the “Protein 3000” Project) was organized by the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan. The basic aim of the project was to obtain deep insight into the biological network by solving over 3000 protein structures and functions of biological and medical importance. It is estimated that the project result lays the foundations for biological, medical, and pharmaceutical researches and has a critical impact in underpinning any life science researches, “What is life?” in general today.

To analyze structures of a large number of proteins, technology development was necessary. The bottleneck of the structural genomics/proteomics existed in the step of production of protein samples suitable for structure determination either by X-ray crystallography or NMR spectroscopy. Therefore, many protein expression methods were developed and tested for the large-scale protein production. Typically, cell-free protein production was much improved during the structural genomics era, and was used for automated protein production, in particular for stable-isotope labeling. On the other hand, the high-throughput beamlines were developed at the synchrotron radiation facilities, SPring-8 and Photon Factory, and were systematically used, and automated data collection started. Furthermore, a large-scale NMR facility was established, and used to perform high-throughput determination of protein domain structures.

The Whole Cell Project is a structural and functional genomics of a model organism. The extreme thermophile *Thermus thermophilus* HB8 was selected to interpret the whole biological phenomena of the cell as the small-genome microorganism constitutes 'minimum protein sets' for cells. Its genome size is about 2 Mbp, and the number of its open reading frames (ORFs) is about 2,200 (http://www.thermus.org). Two-thirds of the ORFs are common to the genomes of most organisms including the human, and one-third of the ORFs are hypothetical proteins. As the thermophile proteins are more thermostable in general than their

mesophile homologues, they are suitable for protein production, crystallization, and structure determination by X-ray crystallography. For the hypothetical proteins, structural analysis was used to predict the function with the success rate of about 60%. The function of the rest of the hypothetical proteins was estimated from transcriptome analysis, metabolome analysis, gene disruption experiments, etc.

In contrast to the proteins conserved from bacteria to humans, evolutionarily new proteins from higher eukaryotes, including humans, mostly consist of functional domains, which are rather independent from each other with respect to structure and function. The sizes of such independent functional domains are mostly smaller than 20 kDa, and NMR spectroscopy was more suitable than X-ray crystallography for their structure determination. In Japan, excellent libraries of full-length cDNA clones from humans were utilized for the structural genomics/proteomics of human proteins, which was a great advantage of the Japanese projects. The cell-free protein production was automated for production of stable-isotope labeled human functional domain samples. The NMR data were collected by using the NMR facility equipped with the high-field NMR instruments at 600–900 MHz. Spectral analyses and structure calculation were also automated, and more than one thousand structures were determined.

Most of the structure types, or the fold families, were found by the structural genomics era. Structural analyses of proteins by X-ray crystallography and/or NMR spectroscopy are much more performed by many research groups of biological studies than before the structural genomics/proteomics. The structural data in the Protein Data Bank are used to perform homology modeling of proteins of interest. Functional studies are guided by the experimentally determined or homology modeled structures of the proteins. Chemical compounds developed by the aid of protein structural information are now useful to probe protein functions in cells. Protein production technology has advanced to cover difficult targets, such as huge protein complexes and integral membrane proteins, which are now suitable targets for single-particle reconstruction by cryo-electron microscopy. Drug discovery based on protein structures is expected to be one of the major approaches of drug developments.

1. Stevens, R. C., Yokoyama, S., Wilson, I. A. “Global efforts in structural genomics”. *Science* **294**, 89-92, 2001.

2. Yokoyama, S., Matsuo, Y., Hirota, H., Kigawa, T., Shirouzu, M., Kuroda, Y., Kurumizaka, H., Kawaguchi, S., Ito, Y., Shibata, T., Kainosho, M., Nishimura, Y., Inoue, Y., Kuramitsu, S. “Structural genomics projects in Japan”. *Prog. Biophys. Mol. Biol.* **73**, 363-376, 2000.

3. Yokoyama, S., Hirota, H., Kigawa, T., Yabuki, T., Shirouzu, M., Terada, T., Ito, Y., Matsuo, Y., Kuroda, Y., Nishimura, Y., Kyogoku, Y., Miki, K., Masui, R., Kuramitsu, S. “Structural genomics projects in Japan”. *Nat. Struct. Biol.* **7** Suppl., 943-945, 2000.

A prospect on the appropriate usage of mathematical models in new biology

Masayuki YAMAMURA

School of Computing, Tokyo Institute of Technology

Abstract

Mathematical models in an appropriate level of abstraction can be useful tools to reproduce and predict any levels of biological phenomena. We show a portion of clever usages in different biological issues. As fundamentals, enzyme reactions in cells are expressed by one-pod non-linear ordinary differential equations in the form of Michaelis-Menten equations or their variation called Hill's equations with allosteric effects. Elowitz et al. showed a good example of stability analysis by using linearized differential equations in their epoch-making paper of "repressirator" in Nature 2000. They showed the combination of frequently used transcription factors will not cause an oscillation. They realized an alternative.

Firstly, We will show a model for the immunologic tolerance including non-monotonic non-linear behavior which is often hard to express by typical equations. B cells will be activated according to the amount of antigen stimuli. It is known that B cells will not be activated under very weak antigen stimuli. We compared different candidate components in the downstream of antigen-bound IgM. If we assign Hill's equation, the best fit Hill's coefficient becomes 4.8. It meets the common view that the required density or the congestion degree of antigen should be around 4 or 5 [1].

We will also show an example of stochastic fluctuation analysis by regarding one-pod non-linear ordinary differential equations as a stochastic process. We designed new genetic

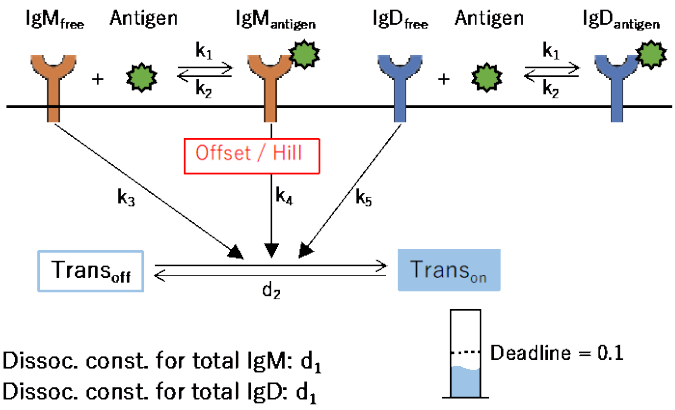


Fig.1 antigen response scheme for B cells

circuit called “diversity generator” where we added an intercellular communication with quorum sensing molecules into the “toggle switch” circuit by Collins et al. in Nature 2000. We showed how we can adjust the switching ratio with simulations by standard Gillespie’s method. [2].

Finally, we will show extending one-pod ordinary differential equations into spatially arranged multi-pod ones. In early embryo stage of C. elegance, it is known that the functional differentiation of each cells will be determined by the spatial pattern which is driven by physical interaction between divided cells restricted within a hard eggshell. We modeled moving cells by using multi-pod ordinary differential equations of mechanical devices arranged in a spatial network. We could reproduce the spatial pattern of divided cells[3].

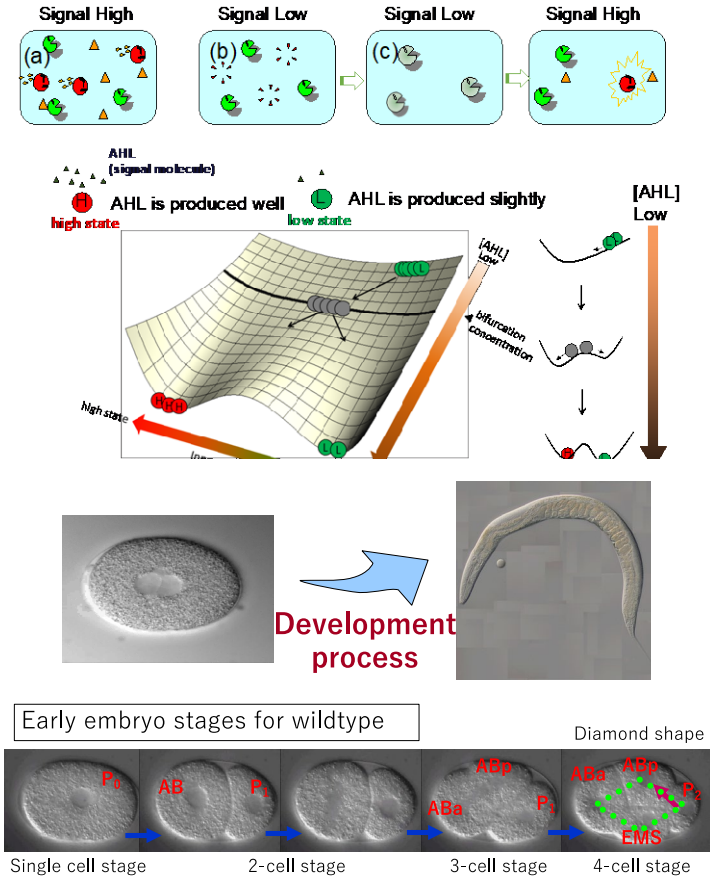


Fig.3 Early embryo development of C. elegance

References

[1] Shoya Yasuda, Yang Zhou, Yanqing Wang, Qing Lu, Masayuki Yamamura, Ji-Yang Wang, “IgD attenuates B cell response by inhibiting IgM-induced survival in mature B cells”, International Immunology, 30 (7): 311-318 (2018).

[2] Ryoji Sekine, Masayuki Yamamura, Shotaro Ayukawa, Kana Ishimatsu, Satoru Akama, Masahiro Takinoue, Masami Hagiya, and Daisuke Kiga, Tunable synthetic phenotypic diversification on Waddington’s landscape through autonomous signaling, Proceedings of the National Academy of Sciences of the United States of America (PNAS), Vol.108, No.44, 17969-17973 (2011).

[3] Atsushi Kajita, Masayuki Yamamura, and Yuji Kohara, Computer Simulation of the Cellular Arrangement in Early Cleavage of the Nematode C.elegans, Vol.19, No.6, 704-716, Bioinformatics (2003).

Reverse TCA cycle: from the origin of life to establish cell factory

Chieh-Chen Huang

Department of Life Sciences, National Chung Hsing University, Taiwan

The increasing of atmospheric carbon dioxide has been considered as a major contribution to greenhouse effect and plays an important role in global climate change. How to establish a cell factory that could produce carbon neutral products has become an emerging topic for studying. The reductive tricarboxylic acid (rTCA) cycle has been considered as an ancient metabolic core in the reducing environment of early earth that provides precursor organic compounds for the synthesis of all major classes of biomolecules. It is basically the reverse running of oxidative tricarboxylic acid cycle and leads the fixation of CO_2 for autotrophic growth. Among six known kinds of CO_2 fixation pathways, the rTCA cycle is considered as the most energy-efficient. Notably, both oxidative and reductive TCA cycle would operate in these microorganisms under mixotrophic (heterotrophic and autotrophic at the same time) growth. On the other hand, the oxidative TCA cycle occurs among mostly heterotrophic organisms, the metabolic pathway has already driven to replenish metabolites in the cycle, known as anaplerotic reactions. These characteristics make the machinery of reductive TCA cycle shown great potential for carbon fixation in heterotrophic host cells.

In our study, synthetic biology logic and experiments are used to convert *Escherichia coli* from its familiar organotrophic mode to a chemolithotrophic mode where CO_2 provides carbon source and energy is provided by inorganic H_2 (as electron

donor) to NO_3^- (as electron recipient) respiratory chain energy generation. Genes for 4 central steps/enzymes of the tricarboxylic acid cycle (TCA) were supplemented with the equivalent genes for reversible TCA cycle steps from the photoautotrophic bacterium *Chlorobaculum tepidum*, so that the reversible (reductive for CO_2 fixation) TCA cycle could function in *E. coli* host cells. The new *E. coli* substrain performed chemolithotrophic CO_2 fixation, while further studies by $^{13}CO_2$-metabolites-labelling analysis revealed that α-ketoglutarate:ferredoxin oxidoreductase was the only foreign enzyme required to form a novel non-native four-step carbon fixation cycle in *E. coli*. In silico carbon pathway analysis supported and explained results. The gene expression profile of chemolithotrophic living *E. coli* was attempted to shape the heritage of genetic system from LUCA (Last Universal Common Ancestor)

Exploring Dark Matter of the Human Genome
~ Introduction of SYNTHETIC GENOMICS ~

Yasunori Aizawa

School of Life Science and Technology, Tokyo Institute of Technology
yaizawa@bio.titech.ac.jp

One of my scientific goals is to understand causality between genomic information and cell function, especially in human cellular system. To achieve this, we have recently introduced into our research roadmap the concept of synthetic biology, “What I cannot create, I do not understand.” And now, our group is aiming at creating artificial cells where human protein machinery works under the control of a synthetic genome that is much simpler than the wildtype human genome, so-called, minimum human genome.

With the recent achievements and progresses in genome synthesis for bacteria and yeast, synthetic genomics now extends its targets for synthesis to much more complex genomes including the human genome. The human genome has born paralog gene expansion and transposon propagation, resulting in the redundant cell networks and the vast noncoding genomic regions. The ENCODE Project has assigned biochemical activities to the majority of these noncoding regions. It however remains unclear whether these observed activities are functionally irrelevant. Therefore, to understand the relationship between genomic information and redesign and synthesize the human genome, it is critical to evaluate the dispensability of genomic segments residing outside of the annotated open reading frames (ORFs). In this respect, genome design always precedes genome synthesis.

Even before synthetic genomics approaches were taken, my group has been exploring the functional relevance of the noncoding segments in the human genome such as long noncoding RNA genes[1,2], untranslated regions of mRNAs[3, 4] and retrotransposon-derived regions[5]. This led us to unexpected findings including the one that small translatable ORFs in the noncoding genomic regions code functional proteins that had been unrecognized by modern biology. And more recently, joining two major international consortiums of synthetic genomics

(Genome Project-write (GP-write)[6] and Sc2.0[7]), we initiated the method development for large-scale genome engineering to evaluate the dispensability of large chunks of noncoding regions. In this presentation, I will share our latest data concerning the significance of noncoding regions of the human genome and introduce our latest roadmap toward the minimum human genomes.

1. Transcripts of unknown function in multiple-signaling pathways involved in human stem cell differentiation. NAR. 37, 4987-5000, 2009.
2. Hypothetical gene C18orf42 encodes a novel protein kinase A-binding protein. Genes Cells. 20, 267-80, 2015.
3. Transposable elements shape the human proteome landscape via formation of cis-acting upstream open reading frames. Genes Cells. 23, 274-284, 2018.
4. Autonomous functionality of an upstream open reading frame in polycistronic mammalian mRNAs. bioRkiv. https://doi.org/10.1101.325571. 2018.
5. Establishment of a genome-wide and quantitative protocol for assessment of transcriptional activity at human retrotransposon L1 antisense promoters. Genes Genet Syst. 92, 243-9, 2018.
6. GP-write; https://engineeringbiologycenter.org/
7. Synthetic Yeast 2.0; http://syntheticyeast.org/sc2-0/