Quantitative Biology 2014, 2(2): 80–84
DOI 10.1007/s40484-014-0034-6

MEETING REPORT

Meeting report on YBPW 2014 (the 3rd Young Bioinformatics PIs Workshop)

Mengyi Sun[1,†], Bingyu Yan[1,†], Chengkun Wu[2,3,*] and Xiaole Shirley Liu[4,*]

[1] Key Laboratory of Gene Engineering of Ministry of Education, State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-Sen University, Guangzhou 510275, China
[2] Faculty of Life Sciences, University of Manchester, Manchester M13 9PL, UK
[3] School of Computer, National University of Defense Technology, Changsha 410073, China
[4] Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute/Harvard School of Public Health, Boston, MA 02115, USA
* Correspondence: chengkun.wu@manchester.ac.uk, xsliu@jimmy.harvard.edu

Received November 2, 2014

On Saturday September 20th, Drs. Xionglei He and Jian Ren hosted the 3rd Young Bioinformatics PI Workshop (YBP) at Sun Yat-Sen University. YBP was started in 2012 to encourage the scientific as well as social interactions among young principal investigators (PIs) as more young bioinformaticians start their independent career in China. The first meeting was hosted by Drs. Xiaole Shirley Liu and Yong Zhang in Tongji University with a dozen PIs participating. Last year, Dr. Yi Zhao from the National Academy of Science Computing Institute hosted the second workshop, and the number of participating PIs increased to about 20. The meeting this year attracted over 40 PIs, as well as students and postdocs.

Each YBPW lasts for one day, with a very casual format and the goal to facilitate scientific and social exchanges among the participating PIs. The host institute only pays for the venue and two meals and helps reserve hotel rooms for the participating PIs. Each PI pays for his own travel and is asked to be present at the whole meeting. Presenting PIs should discuss their own work on a single topic with reasonable details for 15–20 min with a 5 min Q&A. Once a PI presents at a YBPW, he will only be invited to speak again after attending two more YBPW without giving a talk. Besides scientific presentations, there are also ample breaks and discussions for participants to interact and socialize. At the end of each meeting, each participating PI is asked to suggest other colleagues who might be interested in participating and presenting in future meetings. A WebChat group has been established for all previously participating PIs.

In this article, we summarize the 13 talks given by the PIs at the third YBPW. They represent a diverse set of bioinformatics topics that Chinese bioinformaticians are working on. We hope the YBPW could serves its purpose to allow young bioinformatics PIs to present their work, learn from each other, and together promote bioinformatics research in China.

THE 13 TALKS

Wenfeng Qian (Institute of Genetics and Developmental Biology, Chinese Academy of Sciences)

Talk title: SL1 *trans*-splicing enhances translational efficiency in nematodes. The correlation between mRNA level and protein expression level is fairly weak, and thus, translational efficiency may play an important role in protein level regulation. So far, scientists have not known all the mechanisms involved in these regulation processes. Dr. Qian showed that the *trans*-spliced leader 1 (SL1) could enhance translational efficiency by replacing the native 5′-UTR. By using whole genome ribosome profiling data to compare the

† These authors contributed equally to this work

translational efficiency of the one-to-one orthologous genes in four nematode species, he showed that SL1 *trans*-splicing could be a very important mechanism to regulate protein translational efficiency in nematodes, after excluding the confounding factors such as Kozak consensus sequence and codon usage bias. Dr. Qian also showed that there is an optimal 5′-UTR length (21-25 bp) for mRNA translation and *trans*-splicing facilitates genes to optimize their 5′-UTR length. Other factors, such as mRNA folding energy may also have its effect because 5′-UTR length does not explain all the translational effects by SL1 *trans*-splicing. Since the high-throughput sequencing technology, such as the ribosome sequencing (Ribo-seq) involved in this study, is increasingly powerful, we are expecting to acquire more and more new viewpoints to understand the biologic process in a quantitative manner.

Ruibin Xi (Peking University)

Talk title: Copy number variation detection based on exome sequencing data. High-throughput sequencing has been widely used in this decade; many scientists have found lots of single nucleotide variations and small indels from the data ocean to study how changes in DNA sequence and RNA expression level affected the biologic process like cancer evolution and cell differentiation. But copy number variation (CNV) detection in high-throughput sequencing data is always having a lot of biases, because of the target sequence capturing bias, GC bias and so on. Dr. Xi introduced an algorithm in his bias correction for exome sequencing data (BIC-seq-Exome) method to correct these biases in tumor genomes. By comparing microarray data and whole genome sequencing data, he found that this algorithm can detect CNV at very high accuracy. And another allele specific CNV (ASCNV) detection method (ASCNV-exome) provided by him based on Hidden Markov Model also shows significant power. These computational methods successfully identified new recurrent CNVs that may be important for tumorigenesis. And these algorithms open a door to our researchers to accurately use high-throughtput sequencing data to find the impact of CNV in many diseases like cancer. These ideas from mathematics and computer science are changing our manner of research.

Shuai Cheng Li (City University of Hong Kong)

Talk title: Quantifying significance of MHC II residues. The major histocompatibility complexes (MHC) are the immune recognition cell surface molecules involved in cell immunity response of all higher vertebrates. They are highly polymorphic. As protein sequence determines structure and function, MHC can be grouped into serotypes according to the specificity of the response. Dr. Li announced that they could link MHC sequence directly with its serotypes. He proposed a linear programming-based approach to find significant residue positions as well as quantify their significance in MHC II DR molecules to classify MHC molecules. The prediction and recognition of this method performs very well (98.4% prediction performance). The methods are available at http://code.google.com/p/quassi/.

Yang Shen (Sun Yat-Sen University)

Talk title: What are microRNAs good for? –From ecosystem to transcriptome stability. The contradiction between the weak repression of gene by miRNA and its strong phenotypic effect has confused people for a long time. Treating gene regulation network as an ecosystem, Dr. Shen showed that the through weak repression of many targets, miRNA help to stabilize the huge gene interaction network, hence a drastically phenotypic effect might be observed while changing the expression of miRNAs. What's more, he found that by specific targeting different modules in the network, miRNAs execute their functions more efficiently. This might help to explain the heterogeneous expression patterns of miRNAs in different tissues. Dr. Shen's work provided a different perspective on viewing gene interaction, and might be extended to the study of many key components in gene regulation system.

Kai Wang (University of Southern California)

Talk title: Bioinformatics approaches for functional interpretation of genome variation. Bioinformatics approaches were developed to interpret the relationship between genotypes and phenotypes. ANNOVAR is a tool that can generate annotations for functionally important variants with respect to genes, genomic regions, public databases and user-compiled datasets [1]. It has been successfully applied in analyzing clinical genome sequencing data [2]. Tools like ANNOVAR can generate a large list of disease candidate genes, but those genes need to be prioritized in order to identify the genuine disease causal genes. Phenolyzer was developed for this purpose (http://phenolyzer.usc.edu). It takes disease or phenotype terms as input and produces a ranked list of genes for certain diseases with normalized scores. It mainly works by integrating prior knowledge on genes and phenotypes. Integrated CAncer GEnome Score (iCAGES) was specifically developed to prioritize cancer driver genes from single cancer genomes. The iCAGES pipeline takes somatic mutations as input. For each mutation, it calculates a predicated score from pre-trained radial

Support Vector Machine(SVM) and generates an iCAGES score by combining the Phenolyzer score for the mutated gene. The iCAGES score can then be used to filter genes with strongest driver mutations for each patient. Those tools together with others form a pipeline that can generate functional interpretation of sequencing data.

Ting Ni (Fudan University)

Talk title: The function of alternative polyadenylation in cellular senescence. Cellular senescence is a complex biological process in cell cycle. Most researchers draw their attention to gene expression levels and cellular network regulation to study the involved factors. Prof. Ting Ni (Fudan University, Shanghai, PRC) treated this problem in a new perspective, mRNA 3′ alternative polyadenylation (APA). He introduced high-throughput paired-end polyadenylation sequencing (PA-seq) to calculate the length of 3′-UTR in senescent cells such as mouse embryonic fibroblast (MEF, in collaboration with Prof. Wei Tao, Peking University) and rat vascular smooth muscle cell (VSMC), and found that mRNAs in senescent cells have longer 3′-UTR globally. Furthermore, Gene Ontology(GO) analysis indicated that these mRNAs' corresponding genes enriched in many cellular senescence relative pathways, such as TGF-beta signaling pathway, ubiquitin mediated proteolysis signaling pathway and lysosome signaling pathway. The genes have longer mRNA tail that involved in cellular senescence can give more information about the regulation in this process. This PA-seq technology gives us an innovative idea and method in cellular senescence and other researches.

Jian Ren (Sun Yat-Sen University)

Talk title: Systematic study of SUMO regulation in Homo sapiens. Small ubiquitin-like modifier (SUMO) regulates a variety of cellular processes, such as protein localization, stability, interactions and physiological functions. But the interplay between covalent and non-covalent regulation of SUMO remains poorly understood. Dr. Ren introduced a Group-based Prediction System (GPS) method to predict new SUMO sites and developed a high-throughput screening platform for the identification of both SUMO substrates and SUMO-binding proteins. With this platform combining the GPS method optimized with particle swarm optimized algorithm and Bimolecular Fluorescence Complementation (BiFC) protein screening library, he could predict and identify SUMOylation sites and SUMO-binding motifs, then analyze the collaborative mechanism between SUMOylation and SUMO-binding and furthermore, explore the complete mechanism of SUMO regulations in Homo sapiens. These combinations of experimental and computational technology shows significant power to promote our research, especially the algorithms in mathematical and computational science really have its vitality in biology.

Zhenhai Zhang (Southern Medical University)

Talk title: Finding needles in the haystack – antibodyomics. Antibody evolution during B cell maturation plays an important role in adaptive immunity. The next generation sequencing allows us to gain more insight into such an intriguing process. Dr. Zhang introduced antibodyomics and his work on tracking the formation of HIV-neutralizing antibody. Using sequences from a slow progressor, Dr. Zhang constructed the phylogenic trees of neutralizing antibodies. Furthermore, He identified and experimentally validated several neutralizing antibodies, some of them fall into new classes. Previous study has showed that neutralizing antibody formation is highly context dependent, particularly, certain key mutations must appear in appropriate time [3]. Dr. Zhang's work helped to elucidate such process, thus might shed light on the development of effective vaccine against HIV.

Chengkun Wu (University of Manchester; National University of Defense Technology)

Talk title: Building large-scale biomedical text mining facility on Tianhe-2 supercomputer. Text mining has become an important step in systematic studies in the biomedical community. It is used to provide comprehensive and in-time knowledge extraction from the available literature, promoting unbiased access to previous results and evidence, and facilitating data-driven hypothesis generation. Tools were developed to identify genes and pathways involved in certain diseases and extract their interacting patterns. Specifically, PathNER [4] was proposed for mining pathway mentions from literature and PWTEES (https://github.com/chengkun-wu/PWTEES) was designed to extractmolecular events involving genes and pathways. Text mining requires significant computational resources and large-scale analysis can take months to finish. Consequently, the Tianhe-2 bioinformatics team aims to set up biomedical text mining infrastructure on top of the Tianhe-2 supercomputer. The facility will be able to provide support for systematic biomedical studies in various aspects including information retrieval, information extraction, knowledge discovery, hypothesis generation, and data integration, etc. they also plan to provide tailored text

mining services, subject to requirements from users and friends.

Yingrui Li (BGI Technology)

Talk title: Trans-omics for personalized translational medicine. Omics research has greatly enriched the central dogma of biology with epigenetics, proteomics, and metabolomics. This has represented one of many evolving trends. For instance, sequencing has successfully expedited Mendelian disease or rare disease research, which are mainly caused by some specific genetic variations. Meanwhile, complex diseases exhibit less explained heritability (more complex disease, more unknown factors). For instance, one pathological tumor might include multiple genomic distinct tumors. Genomics might provide important methods to address such problem. For example, through exome sequencing and protein identification of the tumor tissues, it would be possible identify key mutated genes via bioinformatics prediction. Such methodology has been successfully applied to develop tumor immune therapy by genomics. Besides human genome, human metagenome has also been indicated as healthcare monitor. BGI Tech has developed a transomics platform to support new-generation sequencing, genotyping and protein profiling, boosted by considerable computational resources. However, it is considered that the conventional computational and bioinformatics scheme needs revision. With the availability of large amount of trans-omics data, it is now possible to identify signatures/patterns and correlation directly from raw data or generate reference data structure.

Xiaowo Wang (Tsinghua University)

Talk title: Model-guided quantitative analysis of microRNA-mediated regulation on competing endogenous RNAs using a synthetic gene circuit. Dr. Wang presented his quantitative study on the complex interaction network between miRNAs and their targets, especially the competitive endogenous RNA (ceRNA) effect, caused by competitive binding of multiple different transcripts with their shared miRNA. Combining mathematical stimulation and synthetic miRNA circuit in vitro, Dr. Wang found that the effect was largely determined by the relative abundance and binding energy between miRNAs and their targets. Furthermore, he showed that the different regulation patterns between miRNA and siRNA might be due to the different recycling rates of miRNA and siRNA, respectively. His work may shed light on the understanding of miRNA regulation and the rational design of siRNAs.

Zhang Zhang (Beijing Institute of Genomics, Chinese Academy of Sciences)

Talk title: Big data integration, curation, and analysis. Biocuration involves the translation and integration of information relevant to biology into a database that enables integration of the scientific literature as well as large data sets. It is usually a time-consuming and laborious procedure and it faces essential challenges including omics data explosion, limited number of expert curators and funding cuts. Biological big data curation requires close collaborations among community, curators and journals. As a result, wiki-based community curation that can harness collective intelligence has emerged as a possible solution. A contribution quantification method considering edit quantity and edit quality and automated explicit authorship generation was developed and implemented as an extension (AuthorReward) to MediaWiki (http://cbb.big.ac.cn/software). The method was successfully testified in RiceWiki (http://ricewiki.big.ac.cn), bearing the potential to deal with big data integration and curation. Qomo, a cloud-computing platform was also developed for biological big data analysis. It enables users to customize their own data processing pipelines via a user-friendly interface, with the aim to realize reusability, repeatability and reproducibility, providing integral services in support for big data integration, analysis, sharing and publication.

Zhi Xie (Sun Yat-Sen University)

Talk title: Systems biology study of human immune response after vaccination. Today, huge amounts of data continue to accumulate through high throughput techniques. However, relatively little information has been obtained from the rich data. Systems biology, which focuses on the quantitative relationship between the elements of complex biological system, may be the exit of the maze. To understand human immunity, Dr. Xie reported his work by analyzing immune parameters in depth both at baseline and in response to influenza vaccination. Peripheral blood mononuclear cell transcriptomes, serum titers, cell subpopulation frequencies, and B cell responses were assessed in 63 individuals before and after vaccination and were used to develop a systematic framework to dissect inter- and intra- individual variation and build predictive models of postvaccination antibody responses. Strikingly, independent of age and pre-existing antibody titers, accurate models could be constructed

using pre-perturbation cell populations alone. The data and analytic framework presented provide a useful resource for studying human immunity in health and disease.

REFERENCES

1. Wang, K., Li, M. and Hakonarson, H. (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res., 38, e164
2. Brownstein, C. A., Beggs, A. H., Homer, N., Merriman, B., Yu, T. W., Flannery, K. C., DeChene, E. T., Towne, M. C., Savage, S. K., Price, E. N., et al. (2014) An international effort towards developing standards for best practices in analysis, interpretation and reporting of clinical genome sequencing results in the CLARITY Challenge. Genome Biol., 15, R53
3. Liao, H.-X., Lynch, R., Zhou, T., Gao, F., Alam, S. M., Boyd, S. D., Fire, A. Z., Roskin, K. M., Schramm, C. A., Zhang, Z., et al. (2013) Co-evolution of a broadly neutralizing HIV-1 antibody and founder virus. Nature, 496: 469–476
4. Wu, C., Schwartz, J.-M. and Nenadic, G. (2013) PathNER: a tool for systematic identification of biological pathway mentions in the literature. BMC Syst. Biol., 7, S2