From structure to function, how bioinformatics help to reveal functions of our genomes

Xiaoqi Zheng1, Sheng Zhong2

1Department of Mathematics, Shanghai Normal University, Shanghai 200234, China

2Department of Bioengineering, University of California, San Diego, California 92093, USA

\*Corresponding author

Email addresses:

Abstract

A report on the 13th International Bioinformatics Workshop, held in Harbin, China, Aug 5-6, 2017.

**Keywords**: Bioinformatics, non-coding RNAs, RNA editing, Chromosome structure, Transposable elements, Cancer genomes

Introduction

The International Bioinformatics Workshop (IBW), held every other year in China since 2003, has grown into an international forum for the most important breakthroughs in all bioinformatics-related fields. Several topics received central attention from this year’s IBW presentations, including novel functional features of the genomes and the transcriptomes, three-dimensional genome organization, and recent evolutions of mammalian and human genomes.

Mining genomic dark matters

Approximately 30 percent of the human genome is composed of repetitive sequences, also known as transposable elements. Transposons in the human genomes were often ignored in functional analyses, until a recent conceptual change led to the functions of previously so-called “junk” DNA becoming the intriguing “dark matters” of the genomes.

By showing the birth of new exons derived from insertion of *Alu* elements, Yi Xing (University of California, Los Angeles, USA) revealed one of the novel functions of the genome dark matter. *Alu*-derived exons exhibited diverse alternative splicing patterns in various human tissues. In addition to their role in modulation of translation efficiency, some of these *Alu*-originated exons also encode new peptides. How do transposable elements shape human transcriptional networks? Ting Wang (Washington University in St. Louis, USA) gave us an example of AluJb, a subfamily of *Alu* sequences, whose insertion into the genome gave rise to alternative promoters to activate oncogene LIN28B. Knocking out a specific copy of AluJb in a lung cancer cell line suppressed cell growth and migration.

Cryptic functions of RNA shortening and synonymous mutation

Wei Li (Baylor College of Medicine, USA) reported a novel mechanism of 3’ UTR shortening that leads to the repression of tumor suppressor genes through disruption of ceRNA crosstalk. This process is mediated at least in part by a RNA cleavage factor CFIm25, which may be responsible for regulating alternative polyadenylation sites on thousands of messenger RNAs.

Synonymous single-nucleotide variants (sSNVs) are usually left out during the analyses process due to the absence of resulting amino acid changes. However, by using regSNPs-splicing software prioritizing sSNVs associated with RNA splicing, Yunlong Liu (Indiana University, USA) found that disease-causing sSNVs are enriched in protein functional domains. Potential function enrichment for intronic single-nucleotide variants (iSNVs) was also discussed.

Tools to identify novel functional features of transcriptome

Yi Xing (University of California, Los Angeles, USA) presented rMATS-turbo, an updated version of rMATS that allows ultra-fast detection of differential alternative splicing and isoforms from replicated RNA-seq data. Shirley Liu described TRUST (https://bitbucket.org/liulab/trust/), to assemble T cell receptor (TCR) hypervariable-region sequences through assigning informative unmapped reads from tumor RNA-seq data into TCR genes. Wei Wang (University of California San Diego, USA) proposed the Taiji pipeline (https://github.com/kaizhang/Taiji), to construct gene regulatory network and identify key regulators of a specific cell stage by integrating multi-type high-throughput sequencing data including RNA-seq, open chromatin and histone modifications.

Circular RNAs, RNA editing and interactions

Circular RNAs are a type of endogenous non-coding RNAs whose 3’ and 5’ ends are covalently linked to form a backsplicing structure. They are tissue specific and evolutionally conserved, suggesting a potential functional role. Fangqing Zhao (Beijing Institutes of Life Science, CAS, China) reported the prevalence of alternative splicing within circRNAs and their tissue specific expression patterns. To help explore this line of work, a new method was developed by the Zhao group based on back splicing and reverse overlap features of circRNA, which could recover ~80% full transcripts of circRNAs in cells (unpublished data). Li Yang (CAS-MPG Partner Institute for Computational Biology, China) discussed the species-specific expression of circRNAs from an evolutionary perspective. He reported that the fast-evolved SINEs (short interspersed nuclear repetitive DNA elements), especially *Alu* elements in human, play a role in the biogenesis of circRNAs.

RNA editing is a type of post-transcriptional alteration of RNA sequences, which could lead to changes in RNA structure or protein products through substitution, deletion or insertion of nucleotides. Han Liang (The University of Texas MD Anderson Cancer Center, USA) reported A-to-I RNA editing on microRNAs and its function in cancer. Through a pan-cancer analysis of TCGA transcriptomic data, Liang and his colleagues identified miR-200b, whose editing level showed distinct patterns of patient survival times from those of the primary miRNA. Specifically, unedited miR-200b served as an inhibitor of epithelial-mesenchymal transition (EMT) and suppressed tumor metastasis. In contrast, with only a single nucleotide modification in the mature region, the edited miR-200b could promote the migration and invasion of cancer cells by retargeting a new set of genes including a key metastatic suppressor, LIFR. This striking example highlights the importance of RNA editing in cancer development.

Deciphering the interactions between different types of RNAs or RNAs and chromatin is a key step to understand their functions. Sheng Zhong (University of California San Diego, USA) reported two techniques, MARIO and MARGI, for massively detecting RNA-RNA and RNA-chromatin interactions *in vivo*. Zhong also introduced the plans of the 4D Nucleome consortium (www.4dnucleome.org) for revealing genome architecture and nuclear organization.

Visualization and analysis of genomic interaction data

New methods and tools have been developed to analyze and present genomic interaction datasets. Yun Li (University of North Carolina, Chapel Hill) proposed HUGIn (http://yunliweb.its.unc.edu/HUGIn/), a unified web browser for visualizing and annotating Hi-C data from human primary tissues and cell lines. Zhihua Zhang (Beijing Institute of Genomics, CAS, China) proposed a new 3D genome visualization tool, Delta, and a new method DeDoc for topologically associated domain (TAD) calling through merge and combine the structural coding tree. Notably, DeDoc could stably detect TAD with a few single cell Hi-C data.

Another common task for analyses of genomic interaction data is to identify long-range genomic interactions. Jian Ma (School of Computer Science, Carnegie Mellon University, USA) presented PEP to predict enhancer-promoter interactions using only sequence-based features.

Genome evolution, adaption, and personal variations

Wenfeng Qian (Institute of Genetics and Developmental Biology, CAS, China) discussed the relationship between genetic interactions (epistasis) and the eukaryotic gene order on a chromosome. By extending population genetics theories, Qian proposed a hypothesis that genetic interaction networks could drive the evolution of gene order. In support of his hypothesis, an anti-correlation between epistasis and gene distance was indeed observed with the analysis of the global genetic interaction network published recently in the budding yeast, partially because genes exhibiting positive epistasis tend to translocate close to each other on a chromosome during evolution.

Two speakers investigated the genetic basis of high altitude adaptation (HAA) for human and mastiff in Tibet Plateau through genome wide association study. Shuhua Xu (CAS-MPG Partner Institute for Computational Biology, China) investigated the genetic origin of high altitude adaption in Tibetans using deep-sequenced whole genome data. Using ArchaicSeeker developed by their own group, he suggested that Tibetans are admixture of multiple populations and derived their ancestry from both archaic and modern human groups. Xu also proposed a “fitness-borrow” hypothesis to explain altitude adaptation mechanism of Tibetans and Sherpas. Yixue Li investigated the genetic basis of hypoxia adaptation of Tibetan mastiff. He identified two loci of genes EPAS1 and HBB that are associated with tolerance of hypoxia and this trait possibly originated from Tibet gray wolfs.

Ge Gao (Peking University, China) reported COPE, an annotation tool of genomic variations accounting for the accumulative effects of multiple variants within the same loci. It identifies multiple function-changing variants that are neglected by conventional tools from 1000 Genomes dataset. Kai Ye (Xi’an Jiaotong University, China) presented Pindel-C (https://github.com/genome/pindel), for detecting complex indels and structural variations from next generation sequencing data. It detected complex indels in 285 cancer genes that are missed in previous TCGA studies. As an application to whole genome sequencing data from 250 trio-families in the genome of the Netherlands project, he found an interesting phenomenon that majority of *germline* mutations are of paternal origin.

Conclusion

As commented by Xiaole Shirley Liu (Harvard University, USA), , bioinformatics emerged from an auxiliary tool in biomedical research and grown into an independent discipline at the forefront of biological discoveries and applications.. IBW participants have become increasingly diverse over the years, perhaps reflecting the increasing levels of collaborations between computational and experimental biologists as well as biomedical practitioners.

# Abbreviations

3C: Chromosome conformation capture technology; ceRNA: Competing endogenous RNA; ChIA-PET: Chromatin Interaction Analysis by Paired-End Tag Sequencing; Hi-C: High-throughput chromosome conformation capture; sSNV: Synonymous single-nucleotide variant; TAD: Topologically associated domain; TCGA: The Cancer Genome Atlas

# Funding

This project was supported by the National Natural Science Foundation of China 61572327 (XZ) and NIH U01CA200147 and DP1HD087990 (SZ).

# Authors’ contributions

XZ wrote the original draft. SZ edited the manuscript. Both authors read and approved the final manuscript.

# Competing interests

Sheng Zhong is a cofounder of Genemo Inc.